

不動産市場とテック

Author1, Author2, Author3, Author4

xxx, 2019

目次

第 1 章 不動産市場とテクノロジー	1
1.1 AI と不動産業	1
1.2 不動産のマッチング	3
1.3 不動産テックによる社会課題解決	6
1.4 良質なデータ資源の重要性:Garbage in garbage out	8
参考文献	10
第 2 章 不動産市場分析の理論	11
2.1 ヘドニック・アプローチによる不動産価格分析	11
2.2 ヘドニック価格関数の推定	12
2.3 不動産価格の分解と予測	15
2.4 不動産価格の実際の推計	17
参考文献	19
第 3 章 不動産テックにおける機械学習の数理	20
3.1 不動産市場分析と機械学習	20
3.2 勾配降下法	21
3.3 線形回帰	24
3.4 分類（ロジスティック回帰）	27
3.5 ニューラルネットワーク	28
3.6 ノーフリーランチ定理	31
参考文献	32
第 4 章 不動産市場分析における統計・機械学習の利用	33
4.1 不動産市場分析における統計・機械学習の手法	33
4.2 線形回帰モデル	34
4.3 分位点回帰	38
4.4 ニューラルネットワーク	40

4.5	その他の手法	43
4.6	手法の適用	43
参考文献		51
第 5 章 不動産市場への機械学習の適用		52
5.1	不動産市場分析の実際	52
5.2	予測モデルのための不動産価格データの用意	53
5.3	推計手法の選択肢	54
5.4	不動産価格の予測モデル	62
5.5	不動産市場における介入効果の測定	66
5.6	傾向スコアを用いた実証分析の事例	68
5.7	不動産市場分析における機械学習の応用と課題	72
参考文献		74
第 6 章 不動産市場分析における GIS の活用		76
6.1	不動産市場分析と GIS	76
6.2	GIS の活用	76
6.3	空間集計における基本操作	80
6.4	空間データの相関と補間	83
6.5	空間特性に配慮した不動産価格構造の推定	86
6.6	空間構造の取り扱い	87
6.7	実データを用いた推計例	90
6.8	推計結果	92
6.9	不動産市場分析の発展可能性	93
参考文献		96
第 7 章 GIS を用いたエリア指標の開発		98
7.1	エリア指標と不動産テック	98
7.2	不動産の価値評価	99
7.3	「日本版 WalkScore (仮称)」の開発	101
7.4	日本版 WalkScore 研究の発展可能性	107
参考文献		109
第 8 章 不動産間取り図の認識と応用		110
8.1	市場探索行動における不動産間取り図	110
8.2	関連研究	110
8.3	間取り画像のグラフ化手法	113

8.4	実験	116
8.5	まとめ	126
参考文献		129
第 9 章 不動産物件情報の流通と活用を支えるデータベース・情報アクセス技術		131
9.1	データベース・情報アクセス技術の発展	131
9.2	不動産物件情報へのデータベース・情報アクセス技術の応用	133
9.3	RDBMS の仕組み	134
9.4	不動産物件画像への深層学習の適用	135
9.5	質の高い不動産情報データベースの構築	138
参考文献		139
第 10 章 官民ビッグデータを用いた空き家分布把握手法の開発		141
10.1	わが国における空き家の増加とその問題背景	141
10.2	既存の空き家分布把握の手法	142
10.3	空き家分布把握に有用なデータ	142
10.4	鹿児島県鹿児島市の事例	143
10.5	群馬県前橋市の事例	148
10.6	空き家は予測できるのか?	151
参考文献		153
第 11 章 不動産金融市场における不動産テック		155
11.1	不動産投資信託（REIT）市場におけるデータ資源	155
11.2	REIT 市場データと REIT 研究の動向	156
11.3	長期的な資産入替の分析	158
11.4	REIT 情報を用いた不動産市場分析の方向性	164
参考文献		165

第1章

不動産市場とテクノロジー

1.1 AIと不動産業

AI(Artificial Intelligence)・ビッグデータ・IOT(Internet of Things)といった新しい技術は、新しい産業を創出するとともに、従来の産業の在り方を大きく変容させようとしている。国民経済計算(GDP)の10%以上を占める不動産業においても同様であり、不動産業を取り巻き大きな変化が起ころうとしている。

不動産業において、AI等のテクノロジーが注目されるきっかけとなったのが、米国のZillow社が、Zesimate((<https://www.zillow.com/zestimate/>))と呼ばれるサービスを世に出したことに始まったと言ってもいいであろう。同サービスでは、全米のすべての不動産に対してリアルタイムに価格の相場が表示されるとともに、その相場の変化をも示している。不動産市場は、とりわけ売りたいと思ったときにいくらで売れるのか、買いたいと思ったときに、提示された価格は適切であるのかといったことがわかりづらいことから、しばしば不透明な市場であると言わされてきたことから、市場において、大きなインパクトがあったと言ってもいいであろう。

わが国の不動産業の法人数は、2010年には30万社を超え、通年を通じてGDPに占める割合とほぼ同じ全産業の法人数の10%を超える水準にあり、120万人程度の就業者数を抱える。また、不動産業といつても幅が広く、①建物売業、土地売業、②不動産代理業・仲介業、③不動産賃業、④家業、間業、⑤不動産管理業と、多くの業態を持つ。

一般に不動産業と聞くと、不動産代理業・仲介業を想像することが多いのではないか。全国どこに行っても、駅を降りると不動産屋さんの看板を見る。それらの多くが、不動産代理業、仲介業と呼ばれるものであり、不動産、とりわけ不動産を売りたい方、貸したい方と、買いたい方、借りたい方のマッチングをしている。そのような行為の中で、最も重要な仕事の一つで、売れる価格、または買う際の適正な価格を決定するということである。

しかし、前述のように、不動産市場は情報が不完全であり、そのために不透明な市場であるといった指摘を受けることが多い。一般に市場がその資源配分機能を十分に発揮するためには、取引対象となる財の質と価格についての情報が市場における取引参加者に十分にいきわたっていること、そして適切な取引対象(相手)を見出し、取引を実現するための特別な費用が存在していないという条件が求められる。しかし、多くの市場においては、情報は完全ではなく、取引を実現するための機会費用も含めてさまざまな費

用が発生していると考えてもいいであろう。そのために、不動産業が十分に機能しないために、または、そのような情報の非対称性が不動産仲介業の介在価値の一つとして考えられていることから、不動産業に対する誤解が生まれ、不動産業者が情報を囲い込んでいて、その社会的費用を増幅させているようなゴシップ記事などが週刊誌を賑わすことも少なくない。

しかし、不動産業者においても、実は正しい価格はわからないと思ったほうがよい。不動産市場は、「同質の財が存在しない」といった特殊性を有しているために、他の市場財と比較して、正しい市場価格を探し出すことは、極めて困難なのである。さらに、不動産市場においては、わが国だけでなく、ほとんどの国で市場情報が不足している。すべての不動産が常に取引がされているわけではなく、不動産の価格を決定している、不動産や土地の品質を含むすべての属性情報やその不動産を取り巻くエリアの環境情報が揃っているわけではない。

つまり、多くの市場財では、価格を決定している品質に関わる情報の種類が少なく、また、市場で取引されている取引価格を容易に知ることができるものの、不動産市場においては、取引価格に関する情報を知ることができないことが多い。このような社会的な問題を解決するために、多くの国で不動産鑑定業というものも存在しており、不動産価格を「**不動産鑑定士**」と呼ばれる専門家が、市場に代わって決定している。

また、品質情報についても、構造偽装問題や欠陥不動産に象徴されるように、開示されている情報そのものに信頼が置けないといった問題も指摘されてきた。つまり、不動産市場において流通している情報の中には、情報の量的問題のほかに、情報の正確度 (accuracy) といった質的問題が存在しており、なかでも財の品質に関する不確実性が高い。

このような情報が不足しているという問題は、特に不動産市場で顕著であるが、多くの市場財において、問題の程度の差こそあれ、共通に抱える問題である。市場に出回っている多くの市場財では、使用目的が同じであったとしても、性能や機能面で多くの差別化が図られている。仮に規格や設備が同じであっても使用後の時間が異なれば、質の劣化の程度が異なり同質のものではなくなる。このような性能や機能面での質の違いはその商品の市場価格に反映される。同時に、その市場財、つまり商品独自の性能や機能に対する消費者の評価もまた市場で決まる価格に反映されているといえる。

そのような中で、米国の Zillow 社が、Zesimite と呼ばれるサービスを通じて、全米の不動産価格をビッグデータと AI を用いて査定し、それを公開されたことは、大きな挑戦であったともいえる。

このサービスを実現するための技術とは、不動産価格を形成する特性に応じた価格付けを統計的な手法によって分解し、その結果を用いて測定したい対象の不動産の特性と掛け合わせることで価格を予測しようとするものである。

データマイニング・ビッグデータ・機械学習・AI などといった言葉の変遷はあるが、不動産価格を予測するということにおいては、その本質的な技術は 1970 年代において**ヘドニック理論**(第 2 章参照) が登場してから理論的にも、実証的にも大きな進化はあったが、そのような理論が登場する前から、研究分野では実施してきた。

そのような不動産の価格形成要因を、統計的な技術を用いて分解し、予測することが実用化されるようになってきたのは、コンピューター技術の進化によって可能になってきた。計算能力が飛躍的に向上し始めると、従来は小サンプルで実験してきた研究が、実用化される道筋が立ち始めてくるのである。とりわけ 1990 年代に入ってから 2000 年代初頭において、回帰分析と呼ばれる伝統的な統計分析を含み、

データマイニングと呼ばれる様々な技術が注目を浴びた。その統計的手法の中核は、回帰分析と共に、クラスター分析、ニューラルネットワークや回帰木(Regression Tree)などであり、さらにはマーケットバスケット分析、記憶ベース推論(MBR)、リンク分析や遺伝的アルゴリズムなど、現在の機械学習と呼ばれる分野で利用される多くの手法が既に含まれていた。

当時は、コンピューターの計算能力が整ってくる一方で、「データ資源」の脆弱性から、データマイニングのブームが大きな広がりを見せることはなかった。また、金融ブームが訪れることで、金融工学などが注目されるようになる。そのような中では、データマイニング分野で活躍していた技術者や研究者も、金融の世界へと参入していった。しかし、リーマンショックに端を発した経済危機に陥る中で、金融工学は衰退し、最近では金融とAIとを掛け合わせたFinTec(Financial Technology)と呼ばれる一つの産業へと変容してきた。

不動産業においても、不動産テック、RealTec(Real Estate Technology)と呼ばれる産業が成長しようとしている。その背景には、米国の成功もあるが、不動産に関する情報の整備が進み、その利用可能性が合法・違法関わらず容易になってきたためである。統計的な技術の進化としては、ニューラルネットワークの進化形である深層学習(Deep Learning)や回帰木が容易に利用できるようになったことも、それを後押ししていると言ってもいいであろう。

不動産価格を予測する技術の応用は、もともとは1990年代に米国における不動産ローンの二次市場の発達の技術的なけん引力となった自動不動産価格システムの開発時に進化した。具体的には、かつてのFederal Home Loan Mortgage CorporationのFreddie Macは、アメリカ全土の不動産に対する価格査定を自動的に行うことができるLoan Prospectorと呼ばれる製品開発を既に開発していた。そして、そのエンジンには、HNC,Incのニューラルネットワークに基づいていた。さらに、IBMにおいても同様の技術で不動産評価の予測に関する研究開発が行われていた。わが国では、1997年に筆者らの研究チームで同様のシステムを開発し、2000年代初頭から、大手メガバンクの不動産ローン、アパートローンの自動審査システムとして利用されており、現在もそのシステムは20年余り活用され続けている。不動産価格の予測という分野は、機械学習またはAIとよばれる技術が古くから適用され、実用化されてきているのである。

不動産テックと呼ばれる産業において、このような不動産価格の予測にとどまらない、多くの技術の応用が登場してきている。本章では、不動産仲介業に焦点を当てて、現在において登場してきているサービスの裏側を支えて技術について整理することを目的とする。

1.2 不動産のマッチング

1.2.1 不動産仲介業の役割

ここで、不動産業の中でも、就業者数といった意味で、最も大きなシェアを持つ産業の一つである不動産仲介業に注目してみよう。消費者が不動産を購入しようとした際には、その情報収集から開始する。不動産情報のポータルサイトなどを通じて、不動産の販売広告を見る。しかし、その広告には、その不動産周辺の住環境、維持管理の経歴、外見からでは判断できない構造強度など、必要な情報が全て網羅されているわけではない。これらの欠けている情報のなかには、実際に物件およびその周辺を注意深く見ること

で初めてわかるものも多い。このように消費者は最も良い物件を求めて自分で探索作業を行わざるを得ないが、それには多くの時間と手間が必要となる。

このような費用を節約するために、不動産仲介業が介在し、またはテクノロジーがそれを軽減するよう利用されている。ここでは、買い手がどのような費用を支払っていると考えたらいいのか、といったことから整理しよう。

まず、消費者が売られている物件を全て見て回るなどということは不可能である。そのことは逆に売り手には物件の「本来の価値」(情報が完全で誰も市場支配力をもたない場合の価格)よりも高い価格で不動産を売却することができる可能性がある。強気で臨む売り手は、買い急いでいる消費者が現れることを期待して、不動産に高めの値段をつけて売ろうとするかもしれない。逆に、売り急いでいる売り手は本来の価値よりも低い価格をつけて売ろうとするかもしれない。それでも売却に至るまでにある程度の時間を待たねばならないかもしれない。このように情報の不完全性のもとでは、不動産価格は本来の価値から乖離する可能性があり、消費者は時間と手間がかかっても探索活動をやめるわけには行かないである。

続いて、売り手である。売り手においても、売却を希望してから契約にいたるまでの時間が必要以上に多くかかったり、契約にいたることができなかったりするといったことが発生する。当然、売り手にとっては、すこしでも高い価格で、かつできるだけ早く売却したい。しかし、売却の目的にもよるが、しばしば期待が高すぎ、またローンの借り換えの都合等なんらかの理由で、当初売り手側が高すぎる下限価格が設定している場合も多い。その場合は、買い希望者が登場したとしても、契約が成立しないケースが発生する。

以上のように、不動産市場では、売り手・買い手ともに高い費用が発生しているため、その社会的な費用を解消するために、不動産仲介業が存在していると言ってもいいであろう。

1.2.2 買い手のコストと売り手のコスト

それでは、「買い手」の探索コストを考えてみよう。情報が完全な市場ならば、不動産の「本来の価値」と売値が一致する。しかし、買い手は真の価格を知ることができないために、探索(サーチ)をしないといけない。このような費用を測定するためには、サーチ理論が有用である。しかし、伝統的なサーチ理論では、同質的な(homogeneous)財を対象とし(つまり財の品質については情報の不完全性はない)、誰がどんな価格をついているかを知らないという形でモデルを組み立てる(例えば、Turnbull and Sirmans(1993)[1])。しかし現実の不動産市場では価格情報はインターネットで得ることができるが、不動産の品質は均一ではなく、その情報は実際に訪問精査しない限りわからない。

Shimizu, Nishimura and Asami(2004)[2]では、サーチ理論の「価格」を「品質調整済価格」と置き換えて、「価格」は確かに不動産情報誌などを見ればわかるが、「品質調整済価格」は実際に物件を訪問精査しなければわからないと考え、「品質調整済価格」についてサーチ理論を援用した。不動産の広義の品質となる立地条件や建物構造については、買い手はすべて共通の認識を持つとした。従って同一の品質の物件に対しては、もし情報が完全ならすべての買い手が同一の「品質調整済の価格」の値付けをするはずである。

また、現実には不動産品質と同様に、それを探す買い手も異質的(heterogeneous)であり、サーチのコストも本来ならばそれぞれの買い手のサーチコスト合計を考えなければならない。しかし買い手のサーチ

コストの異質性を正面から取り上げるには、膨大な情報と費用がかかる。このような視点を踏まえてマーケティングを考えいかなければならない。また、品質調整済みの不動産価格は、第2章において整理する。そうすると、**品質調整済み価格**がわかっているとすれば、実際の売値と理論価格との乖離、つまり残差項の分布（**ヘドニック式の残差**）が、超過価格を表すことを意味する。つまり、買い手は、超過価格が存在する限り、実際の市場価格よりも、高い価格で不動産を買わないといけなくなってしまうのである。そのため、買い手は、品質に応じた適切な価格で買うことができる物件を探し続けることになる。

この「超過価格」について、標準的なサーチ理論の仮定に従って、標準的な買い手は、個別の物件単位の超過価格は知らないものの、その確率分布は知っているものと仮定し、その分布を分布関数 F 、確率密度関数 f とすると、買い手のサーチコストを推計することが可能となる。

買い手は、すでにいくつかの物件を見て、その時の最低超過価格が y であったとする。その時に、次の物件をサーチすると、サーチ費用として s がかかるものとする。次のサーチを実行すると、 y 以下の超過価格の物件が見つからない（最低超過価格が変わらない）確率は、 $1 - F(y)$ である。それ以外の場合では、 $x (< y)$ の超過価格が見つかる確率密度が $f(x)$ である。従って、次のサーチを行った場合の超過価格の期待値 $X(y)$ は、

$$X(y) = \int_{-\infty}^y xf(x)dx + y[1 - F(y)] \quad (1.1)$$

となる。よって、次のサーチを行う純便益 $B(y, s)$ は、

$$B(y, s) = \{y - X(y)\} - s = \int_{-\infty}^y (y - x)f(x)dx - s \quad (1.2)$$

となる。これは、 x について単調増加関数であり、 f についてのデータがあれば、 $B(y, s) = 0$ となる y の値を求められる。この y を以下、 y^* で表す。そうすると、最適なサーチ戦略は、サーチをして得られた超過価格が y^* 以下になるまでサーチを続けるというものになる。

この戦略に基づいて、初回のサーチでサーチをやめる確率は $F(y^*)$ となる。また、2回目でやめる確率は、初回でサーチをやめない確率 $[1 - F(y^*)]$ に、その回にサーチをやめる確率 $F(y^*)$ をかけたものに等しいから $F(y^*)[1 - F(y^*)]$ となる。一般に、ちょうど n 回だけサーチを行う確率 $Q(n)$ は、

$$Q(n) = F(y^*)[1 - F(y^*)]^{n-1} \quad (1.3)$$

であるから、サーチ回数 n の期待値を求めるとき、

$$\sum_n nQ(n) = \sum_n nF(y^*)[1 - F(y^*)]^{n-1} = \frac{F(y^*)}{1 - F(y^*)} \sum_n n[1 - F(y^*)]^n \quad (1.4)$$

となるが、無限等比級数の公式から

$$\sum_n n[1 - F(y^*)]^n = \frac{1 - F(y^*)}{F(y^*)^2} \quad (1.5)$$

が知られているので、それを代入して整理するとサーチ回数の期待値は $1/F(y^*)$ となる。そこで、買い手のサーチによるコストは、 $s/F(y^*)$ となる。問題は、いかに s を求めるかであるが、これは、平均して1つの不動産に訪れる時に要する時間コスト（訪問に要する時間に賃金率を乗じたもので代用）に加えて、1回訪れるに要するその他の費用（交通費用、情報交換費用など）を加えれば良い。

売り手のコストは、もう少し簡単に考えることができる。「売り手」は売買成立までにある程度の時間がかかるため、その間は住戸資産を生産的用途に用いることができず、いわば無駄に所有していることになる。仮に情報が完全な市場では、売り手は自らの物件に関する情報は十分に認識していることから、その品質情報に応じた「本来の価値」で、すぐに売却することが可能となる。その意味で、物件売却までにかかる機会費用は情報の不完全性に伴う売り手のコストとなる。

そこで、売り手の損失としては、その間の機会費用を計上することができる。その計上方法としては、市場滞留時間を T 、レンタル価格、つまり賃料を $Rent$ とすれば、

$$Rent \times T \quad (1.6)$$

となる。

また、新古典派の資本理論に基づけば、貸し借りが完全に自由にできる資本ストック市場における均衡では、レンタルコストが資本コストに等しくなる。そこでレンタルコストの代わりに資本コストで機会費用を表すことができる。最終売価格を P 、利子率を r とすると、その物件の市場価値は最終売価格に近いと考えられることから、機会費用を

$$P \times r \times T \quad (1.7)$$

で近似できる。

1.3 不動産テックによる社会課題解決

以上のような買い手・売り手の費用が発生しているということは、その費用を節約するために、不動産仲介サービスを利用しようとする。その費用を不動産仲介業者に支払ってきたわけであるが、テクノロジーの進化は、その業務を一層高度化・効率化することを可能とする。

まず、買い手にとっても売り手にとっても、市場で成立している価格や家賃は、サーチ行動または売却行動にとって最も重要な情報となる。適正な市場価格がわかれば、売り手はすぐに売却が可能となり、買い手は住宅購入の意思決定を容易にする。

このような品質調整済みの予測価格 (P) を生成することを実現するためには、**ヘドニック・アプローチ**と呼ばれる経済理論的な背景を伴いながら発達してきた理論と分析手法を用いて、不動産の価格・家賃に関するビッグデータが整備され、コンピューターの計算技術と機械学習と呼ばれる計算手法が進化する中では、多くのサービスが実用化されてきている。このようなサービスが一層進化していくためには、データの発生プロセスをも踏まえた経済理論的な背景の深い理解と、不動産のエリア情報をも含む属性データが、一層整備されていくことが求められている。本書では、経済理論的な背景を第2章において、そして、機械学習の数学的な基礎を第3章において、それぞれ整理した。また、第4章では、伝統的な回帰分析から出発し、不動産の価格予測をするための機械学習を含む推計方法の特徴を示した。

第5章では、観察可能となった情報を用いた実際の推計方法について具体的なデータと推計手法の適用方法を紹介するとともに、それぞれの手法の特徴を紹介する。不動産市場分析においては、不動産価格を予測する、または不動産価格を形成する各種要因がどの程度の価格効果を持っているのかを識別するといった二つの大きな目的がある。前者の目的においては、解釈性が保証されないモデルにおいても予測精

度が高ければよいが、後者においては解釈性が求められる。このような目的に応じて、どのようなモデルで推計していくべきかを具体例をもって示している。

また、不動産の価格予測をしていく手続きにおいては、立地情報やエリア情報の入手は極めて重要になる。とりわけデータ資源を生成していく技術は、不動産テックのみならず全ての統計分析において欠かせない技術である。不動産分析においては、不動産情報は、住所と建物属性だけが入手できる場合が多く、その場合には、不動産の特徴量の中でも重要な要因となる、「最寄り駅」の特定やそこまでの「距離」、または周辺の商店の集積や学校への近接性などといった情報が必要となる。そのような情報を生成する最も有力なデータ生成技術として、地理情報システム、またはGIS(Geographic Information System)の活用が挙げられる。第6章では、不動産テックを推進するためのデータ生成技術としてのGIS技術とそれに付随する統計分析の基礎的な知識を提供する。また、このようなGISの技術を使い、エリア特性を指標化したエリア指標については、第7章に紹介する。加えて、第6章では、不動産価格の分析として、近年において発達してきた二つの手法を紹介する。

第一は、ビルダーズモデルと呼ばれる推計技術である。不動産テックと呼ばれる産業で開発されているサービスでは、価格モデルが欠如していることが多い。**ビルダーズモデル**は、生産関数から出発したモデルであり、そのモデルを適用することにおいて、不動産価格を土地と建物に分離することが可能となる。また、GISを用いたアドレスマッチングサービスが進化する中では、不動産の特性としての緯度・経度といった位置座標が利用できるようになってきた。そのような空間的な位置によっても、不動産価格は差別化される。また、その価格差は、建物価格には発生することがなく、土地価格においてのみ発生する。そこで、ビルダーズモデルをさらに発展させ、座標位置を用いて、空間的な特性を入れたモデルへの拡張方法を紹介する。

これらの一連のテクノロジーを活用することで、不動産価格を予測することができ、売り手と買い手、または仲介業者もピッグデータとテクノロジーを融合させて、市場価格(P)を得ることができるようになる。それを情報提供サービスとして提供することができるようになってきたが、これらのサービスにより、「売り手」サイドの空室の機会費用削減とともに、「買い手」サイドでは物件を探索するための機会費用の低減させることができるようにになったと言つてもいい。また、家賃を対象として価格を予測したサービスも生まれてきたことから、式(1.6)の $Rent$ もまた、市場で容易に観察ができるようになった。

このような情報とともに、消費者に対しては、住宅選択が可能な具体的な情報を提供していかなければならない。例えば、部屋の「間取り」情報であったり、「エリア情報」であったりと、不動産を取り巻く包括的な情報を正しく生成し、消費者へと届けなければならないのである。第8章では、機械学習の手法を用いて「間取り」を認識する技術を、第7章では、エリア情報の生成手法について整理した。米国の不動産テック企業が配信を始めた「Walk Score」と呼ばれるものが注目されているが、第7章では、日本で入手可能なエリアに関わるデータ資源を用いて、日本版のWalk Scoreを推計していくうえでの技術的な背景を紹介した。

さらに、消費者に対しては、市場に流通している不動産物件に関する網羅性が高く正確な不動産物件データベースを提供していかなければ、買い手・売り手のコストも低減させることができない。第9章では、データベース・情報アクセスの基礎技術について解説するとともに、深層学習を不動産物件画像に適用する取り組みの事例についても紹介する。

このような情報提供だけでは、まだまだ残された課題も多い。

第2章で整理するように、買い手の特性は一様ではないためである。単身者であったり、子育て世帯であったり、または高齢者の夫婦だったりすることもある。そのような場合には、それぞれの特性に応じた情報にだけ意味を持つ。例えば、子育てのしやすさといった「保育環境」「教育環境」、日常の買い物のしやすさといった「商業集積」なども、すべてが価格に反映されているわけではない。また、水害が発生する頻度や災害に対する地盤強度、さらには、大気等の環境汚染が健康に害を与えることも多く、居住地選択とは無関係ではない。このような情報は、探索して初めてわかる情報であり、後者などについては住んでみてはじめてわかることが多い。さらには、維持管理の経験や外見からでは判断できない構造強度などは、高度に専門的な知識が必要となるため、探索しても十分に判断できない場合が多い。

様々なテクノロジーの進化が不動産のマッチング効果を高め、市場の潜む費用を低下させるように機能し始めているところであるが、今後のデータ資源の整備とテクノロジーの発達が待たれる分野も多く残されており、今後の発展に期待するとともに、研究開発を進めていかなければならないと考えている。

不動産テックは、上記のような不動産流通分野に限らず、より広範囲で利用されるようになってきた。第10章では、官民が持つビッグデータを用いた空き家の予測モデルを紹介した。高齢化の進展に加え、人口減少が進む地方都市では、都市全体が縮退していく中で、所有者がわからない土地が発生するとともに、空き家も増加してきている。しかし、その数を正確に把握することもまた困難であり、さらに将来に発生する空き家を予測するといったことは極めて困難である。しかし、このような問題に対しても、新しいデータ資源の活用が可能となり、テクノロジーが進化する中では、可能としてくる可能性は高まっている。

さらには、不動産市場は、金融市場と融合する中で、不動産投資信託をはじめとする不動産金融市場もまた、新しい産業として、21世紀に入ってから大きく成長してきている。第11章では、不動産投資信託市場で入手可能なデータ資源を紹介するとともに、応用例を示した。

1.4 良質なデータ資源の重要性:Garbage in garbage out

近年において、不動産分野においても、AIの活用可能性が注目され、不動産TecまたはReal-Tecなどともてはやされるようになってきた。しかし、実用化されているモデルを見ると、不動産市場の専門家として看過できないようなモデルも少なくない。

例えば、個別物件単位における売りの仲介として不動産市場に出現していくと考えられる対象物件をあぶりだすためのモデルに、様々な内外のマクロ経済指標を学習させているようなケースもある。また、同様のマクロ経済の集計量で、企業単位での土地の放出を予測するようなモデルなども提案されてたりする。不均一性を考慮する必要があるものの、データ収集の容易さだけで、個別性の強い不動産市場のモデルを構築しているのである。

これは、機械学習に対する過度の期待による弊害であり、これらモデルは汎用性や一般性はないことは専門家であれば判断できる。このような事態をもたらしている原因としては、AIの活用を担当する企業の担当者と開発を担当するエンジニアの双方に、不動産市場分析技術・統計技術の知識のいずれか、または両方が欠如しているためと考える。

統計学では、「いくらゴミを学習させてもゴミしか出てこない(Garbage in garbage out)」という言葉がある。AIなどの科学技術の進歩は、必ず市場を進化させる。しかし、誤った技術の使い方は、市場の

進化を阻害するだけでなく、その技術の評価を低下させてしまうことにもつながる。不動産市場を一層進化させていくためには、不動産市場と AI 等の技術にも精通した高度不動産人材を育成していくことが急務である。

また、市場分析に欠かせないデータ資源の権利保護と収集手続きも重要な課題である。このような問題は、本著の範囲を超える分野となるものの、そのようなデータ資源への精通もまた、不動産テックを習得していくうえで、極めて重要な知識となってくることも、留意していただきたい。

参考文献

- [1] Turnbull, G. K. and C. F. Sirmans, (1993), “Information, Search, and House Prices”, *Regional Science and Urban Economics*, Vol.23, pp. 545-557.
- [2] Shimizu, C., K. G. Nishimura and Y. Asami (2004), “Search and Vacancy Costs in the Tokyo housing market: Attempt to measure social costs of imperfect information,” *Regional and Urban Development Studies*, 16(3), 210-230.

第2章

不動産市場分析の理論^{*1}

2.1 ヘドニック・アプローチによる不動産価格分析

不動産テックに関する、わが国の論文や著書を見ると、「ヘドニック・アプローチ」によって価格予測を行ったという記述を見ることが多い。しかし、多くの場合において、単なる回帰分析や機械学習を用いて不動産価格の価格形成要因を分解し、それを束ねる形で価格予測をしているだけであり、厳密な意味でのヘドニック法を適用しているものは極めて少ない。ヘドニック・アプローチとは、ある商品の価格をさまざまな性能や機能の価値の集合体（属性の束）とみなし、統計学における回帰分析や機械学習のテクニックを利用して商品価格を推定する方法である。経済的な理論の裏付けをもって関数形を設定しながら商品価格は分析していく手続きであるために、属性の束からなる方程式で表現され、このような式をヘドニック価格関数とよぶ。ヘドニック・モデルでは、その方程式を解いていくことから、厳密な意味での関数形を設定しない機械学習によって推計されたモデルが、ヘドニック・アプローチによって推計されたという表現は、正しいのかどうかは疑わしいところである。また、ヘドニック価格関数を具体化することは、消費者が個々の機能や性能に対してどの程度の価値を見出しているかを明らかにすることと同じであるために、その推計値によって、各商品が持つ属性の効果を識別していくことができる。

伝統的な価格理論との大きな相違は、一般的な市場財では一物一価の法則が市場分析を行う上での有効な仮定となるが、Lancaster(1966)[3] が分析しているように、この仮定は差別化された商品を扱う上で理論的にも、または実証分析を行う上でも不都合となる。そのような中で、Rosen(1974)[4] はこのような属性の束としての商品価格データが、どのような市場メカニズムで発生するのかを理論的に解明するとともに、実際の価格関数の推計手順も提案した最初の研究であった。不動産市場分析の文脈で考えたときに、Rosen(1974)[4] 以前においても、不動産価格を回帰分析の技術を使って分解する研究が存在していたが、データ発生プロセスをどのように記述するか、識別問題や一致性をどのように考えたらいいのかといった観点から見て、ヘドニック価格関数は正しく理解されていなかった。つまり、現在において、不動産テックと呼ばれる産業界において活用されている、不動産に関わる価格や家賃のデータを単に収集し、機械学習の手法を用いて開発されている多くのシステムにおいては、Rosen(1974)[4] 以前の技術を使って推計しているだけであると言っても言い過ぎではないであろう。そのために、実際の活用において、多くの問

^{*1} 本章は、清水千弘・唐渡広志 (2017)[7]『不動産市場の経済経済分析』朝倉書店および清水千弘 (2017)[6]、「ビッグデータで見る不動産価格の決まり方」不動産学会誌、120号、45-51をもとに、整理したものである。

題を引き起こしてしまっている。

Rosen の研究は, Tinbergen(1959)[11] の提起による差別化された生産物の市場均衡理論を発展させたものである。商品供給者のオファー関数 (offer function), 商品需要者の付け値関数 (bid function) およびヘドニック価格関数の構造との間の関係を厳密に検討し, 商品の市場価格を消費者および生産者の行動から特徴づけている。実際に実証分析を行ってはいないものの, 計量経済学的な推定手順についての概略も示している。Witte, Sumka and Erekson (1979)[12] は Rosen 理論を元に具体的に実証分析した研究である。

Rosen 理論では, 単純化されたケース (生産者を同質に扱うケース) においてすら, ヘドニック価格関数から選好や技術の構造を識別するためには非常に複雑な解析を必要とする。Epple(1987)[1] は多数の消費者と生産者を想定した上で, Rosen 理論を発展させた計量経済モデルを定式化している。Rosen 理論の問題点は, 需要と供給からなる構造方程式において, 同時性バイアスが生じるケースを排除できない点である。もし, 重要な属性が観察されておらず, それらが観察された属性と相關している場合には, 均衡におけるヘドニック価格関数の観察された属性の係数推定量には不偏性もなければ一致性もない。ヘドニック・アプローチを不動産市場に適用しようとした場合には, 不動産価格の複雑さから, この問題は深刻な問題になってしまふ。具体的には, 不動産価格は, 立地や建物といった特性だけでなく, 周辺環境といったエリア特性までもが価格に影響をもたらすことから, それら変数も含めて考慮しなければならないためである。そのため, 分析者は常に必要な属性を観察できるわけではなく, 利用できる変数が限定的になってしまふという問題は, ヘドニック・アプローチの利用上最も注意すべき問題点の一つである。

この点に関して, Epple のモデルは観測誤差を正確に処理できるヘドニック価格関数を提起するアプローチとなっている。ただし, このアプローチは効用関数に次の先見的な仮定をおいた上で, 閉じた市場均衡におけるヘドニック価格関数を導き出し, 推定を行うことになる。

- 効用関数の関数型はすべての消費者について同質である。ただし, 選好パラメータが正規分布に従う (共分散は非対角要素が 0 の対角行列)。
- 消費者の効用関数は属性変数が加法分離的で 2 次形式である。
- 差別化された商品の供給が外生的に与えられている。

上記は経済主体間の相互作用がないこと, および市場均衡におけるヘドニック価格関数が描写できるよう実現可能な関数型を想定しており, 決定的な強い仮定である。

2.2 ヘドニック価格関数の推定

2.2.1 付け値関数

ヘドニック・アプローチの理論的枠組みを Rosen(1974)[4] および Epple(1987)[1] にしたがって整理する。ここでは, $K \times 1$ の属性ベクトル X (属性の束) からなる不動産の需要を考える。ここでいう属性とは, 不動産価格を差別化している, 「最寄り駅までの距離」, 「都心までの距離」, 「大きさ」, 「建築後年数」などである。

属性の束で示される不動産の市場価格関数を $P(X)$ としよう。消費者の効用関数を $u(c, X; A)$ と書

く。ここで、 c は価格が 1 に基準化された価値尺度財（スカラー）、 \mathbf{A} は消費者個人を特徴付ける選好パラメータのベクトルである。消費者の所得を I とするとき、予算制約式は $I = P(\mathbf{X}) + c$ となる。消費者の所得と選好の分布を確率密度関数で考え、これを結合確率密度関数 $f(I, \mathbf{A})$ で表わす。

与えられた予算制約のもとで、 (c, \mathbf{X}) について効用を最大化するとき、次の最適化条件が得られる。

$$\frac{\frac{\partial}{\partial \mathbf{X}} u(I - P(\mathbf{X}), \mathbf{X}; \mathbf{A})}{\frac{\partial}{\partial c} u(I - P(\mathbf{X}), \mathbf{X}; \mathbf{A})} = P_{\mathbf{X}}(\mathbf{X}) \quad (2.1)$$

ここで、 $P_{\mathbf{X}}$ は属性の 1 階微分を示している。すなわち、最適な属性の選択は合成財に対する個々の属性の限界代替率が不動産市場価格の限界的価値に等しいところで決定される。不動産市場価格の限界的価値は需要者がその属性に対して支払ってもよい（willingness to pay）と考える属性の価値に等しくなっている。したがって、個々の属性価値を調べるために、市場価格関数 $P(\mathbf{X})$ における各属性の微係数を知る必要がある。

需要者が不動産に対して支払ってもよいと考える最大の価格のことを付け値（bid price）とよぶ。これを θ という記号で定義する。いま、ある一定の効用水準 u^* のもとで選択された属性の束が \mathbf{X}^* であるとき

$$u(I - P(\mathbf{X}^*), \mathbf{X}^*; \mathbf{A}) = u^* = u(I - \theta, \mathbf{X}; \mathbf{A}) \quad (2.2)$$

である。したがって、付け値と属性の関係を示す付け値関数は、この効用関数のもとで $\theta = \theta(\mathbf{X}, I, u; \mathbf{A})$ と陽表的に示すことができる。すると、効用が最大化されるとき、任意の $f(I, \mathbf{A})$ のもとで

$$P_{\mathbf{X}}(\mathbf{X}^*) = \frac{\partial}{\partial \mathbf{X}} \theta(\mathbf{X}^*; u^*, I, \mathbf{A}) \quad (2.3)$$

でなければならない。このことは、市場価格関数の勾配が所得の限界効用に対する属性の限界効用に等しいだけでなく、付け値関数の勾配にも等しくなっていることを示している。ヘドニック・アプローチとは、不動産価格を不動産のさまざまな属性に回帰させたモデルを推定することによって、各属性の価値を予測する手法である。ヘドニック価格関数を 1 次近似すると

$$P(\mathbf{X}) \cong \tilde{P} + \sum_k \frac{\partial P(\tilde{\mathbf{X}})}{\partial X_k} X_k \quad (2.4)$$

であるから、ヘドニック価格関数はさまざまな属性の限界的価値の線型結合式とみなせる。例えば、第 i 属性ベクトル $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{iK})$ に不動産市場価格 P_i を回帰させた古典的な線型回帰モデルは

$$P_i = \beta_0 + \sum_{k=1}^K \beta_k X_{ik} + u_i \quad (i = 1, 2, \dots, n) \quad (2.5)$$

と表現される。ここで、 $\beta_1, \beta_2, \dots, \beta_K$ は不動産属性の限界的価値を示す未知パラメータであり、 u_i は攪乱項である。しかしながら、この線型近似式だけでは多数の消費者の選好を反映したヘドニック価格関数かどうかを識別する手がかりはない。生産者の行動も考慮に入れてモデルを閉じて、均衡状態のヘドニック価格関数を描写する必要がある。

2.2.2 市場均衡とヘドニック価格関数

(2.5) の左辺は不動産市場の需給均衡で決まる市場価格であるから，生産者の行動も描写しなければモデルを閉じることができない。不動産のように差別化された商品の費用関数を $C(\mathbf{X}, M; \mathbf{B})$ とする。ここで， M は建設される不動産の数を示しており， \mathbf{B} は各生産者を特徴づけるパラメータ・ベクトルである。 \mathbf{B} の分布は確率密度関数 $g(\mathbf{B})$ で与えられているものとする。生産者は不動産市場価格を所与として，次の利潤を最大化する属性の束を決定する。

$$\pi = P(\mathbf{X})M - C(\mathbf{X}, M; \mathbf{B}) \quad (2.6)$$

生産者の行動は，短期か長期かによっても異なり，Rosen が示したように短期には 2 パターンの状況を想定できる。

- 生産者にとって M だけが可変的な短期経済
- M および \mathbf{X} のどちらも可変的な短期経済

長期の経済では固定資本（費用関数に明示されていない）も可変的になり，参入・退出の自由が認められる。ここでは，二つめの短期経済を想定して，次の最適化条件を得る。

$$P_{\mathbf{X}}(\mathbf{X}) = \frac{1}{M} \cdot \frac{\partial}{\partial \mathbf{X}} C(\mathbf{X}, M; \mathbf{B}) \quad (2.7)$$

$$P(\mathbf{X}) = \frac{\partial}{\partial M} C(\mathbf{X}, M; \mathbf{B}) \quad (2.8)$$

(2.7) より各生産者は属性の限界的価値が不動産 1 単位あたりの属性の限界費用に等しく，そして，(2.8) より与えられた属性の束のもとで，不動産の市場価格は任意の生産技術をもつ生産者の不動産生産限界費用に等しくなければならぬ。このとき達成される最大利潤はパラメータ \mathbf{B} によって異なる。

ある一定の利潤 π^* のもとでの最適な属性の束 \mathbf{X}^* と生産個数 M^* を選択しているものとしよう。このとき，生産者が提示できる最低の価格（オファー価格）を φ という記号で表わす。すなわち，

$$\varphi M - C(\mathbf{X}, M; \mathbf{B}) = \pi^* = P(\mathbf{X}^*)M^* - C(\mathbf{X}^*, M^*; \mathbf{B}) \quad (2.9)$$

である。この式は，一定の π^* のもとで φ が (\mathbf{X}, M) とどのような関係を持つのかを示している。(2.9) より， $\varphi = \partial C(\mathbf{X}, M; \mathbf{B}) / \partial M$ であるから，これを M について解き，利潤定義式に代入すると， $\pi^* = \varphi \tilde{M}(\mathbf{X}, \varphi; \mathbf{B}) - C(\mathbf{X}, \tilde{M}(\mathbf{X}, \varphi; \mathbf{B}); \mathbf{B})$ が得られる。すなわち，この関係より，オファー関数は $\varphi = \varphi(\mathbf{X}; \pi^*, \mathbf{B})$ と書くことができる。(2.7) より，利潤が最大化されているとき

$$P_{\mathbf{X}}(\mathbf{X}^*) = \frac{\partial}{\partial \mathbf{X}} \varphi(\mathbf{X}^*; \pi^*, \mathbf{B}) \quad (2.10)$$

でなければならない。

\mathbf{X} に対応したあらゆるタイプの不動産の需要と供給とが等しくなるところで市場均衡が成立し，市場価格 $P(\mathbf{X})$ が得られる。(2.3) と (2.7) より，属性の付け値関数とオファー関数との接線の軌跡として均衡における市場価格 $P(\mathbf{X})$ を表わすことができる。すなわち，市場をクリアする価格関数は消費者の付

け値関数と生産者のオファー関数との包絡線でなければならない。図 2.1 は第 1 番目の属性 X_1 に関する付け値関数とオファー関数の接線上に市場価格が成立していることを示している。曲線 $P(X_1, \mathbf{X}_{-1}^*)$ は、 X_1 以外の属性ベクトル \mathbf{X}_{-1} が \mathbf{X}_{-1}^* において最適化されているとき、さまざまな消費者と生産者との間で成立する市場価格の軌跡を示している。

Epple(1987)[1] が指摘したように、市場をクリアするヘドニック価格関数は消費者の所得と選好の確率分布 $f(I, A)$ と生産者のパラメータ分布 $g(B)$ に依存して決まる。もし、生産者が 1 タイプしか存在しなければ、限界費用関数そのものが市場価格関数になる。限界費用と付け値関数の傾きとが等しくなるところで市場がクリアするので、その包絡線は 1 生産者の限界費用関数に一致するからである。

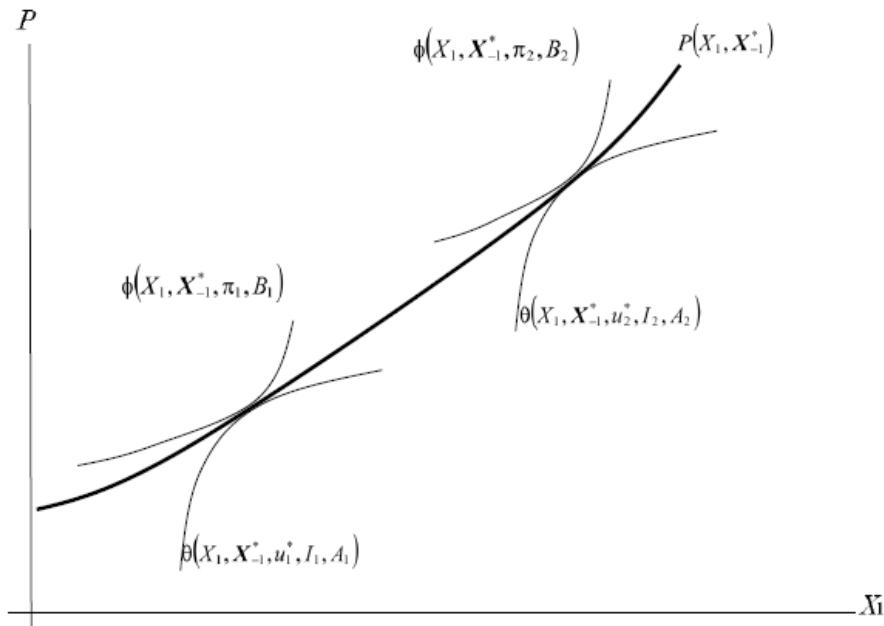


図 2.1 属性 X_1 に関する付け値関数、オファー関数、市場価格関数

2.3 不動産価格の分解と予測

不動産テックの分野では、不動産価格を予測するサービスが提供されるようになってきている。そのようなサービスの裏側では、広い意味でのヘドニック・アプローチを用いて、回帰分析や機械学習のテクニックを用いて、市場で観察された不動産の価格・家賃と不動産の価格を決定する属性データを用いて、価格を分解し、そして予測したい不動産の属性データを取得したうえで予測していく。それでは、市場で観察されている不動産情報とはどのように観察することができるのでしょうか。

不動産価格については、しばしば新聞などでも、その変化が報道されることがある。前年と比較して相場が上がった、または下がったとか、ある地域と比較して、分析対象の地域が、どの程度高いまたは低いといったことがしばしばみられる。そのような相場の水準や変化は、本来はどのように比較しないといけないのであろうか。

不動産は、同一の財が存在しないという特性を持つことから、また、すべての不動産がいつも取引されているわけではないことから、異なる場所で異なる属性を持つ不動産が取引されている。不動産市況が好調にもかかわらず、市場で観察される不動産価格が下落しているということもあるが、その背後には、良質な不動産が市場で取引されてしまい、質の低い不動産が、その後に取引されれば、市況が好調であったとしても、実際に取引される不動産の価格の平均は低くなってしまうことがある。

そのようななかで、不動産の専門家は、相場の水準・変化を見るために、異なる物件の相違や異なる二時点の二つの不動産価格に関しての分布を頭の中で想定している。他の経済市場の財・サービスとの対比において、不動産の専門家が考慮しないといけないのが、前述のように、不動産価格の分布は不動産の性能や属性によって変化するということである。不動産価格は、最寄り駅からの距離などの交通利便性や、同じ場所に立地する不動産であったとしても大きさや建築後年数、または構造によって価格が変化するため、そのような相違を定量的に制御しなければならない。二つの異なる空間または時間の価格の相違を見ようとした場合には、品質を調整しなければならないのである。それでは、時間に関しての統計的な品質の調整方法を説明しよう。ここで時間を空間と読み替えれば、異なる空間間で比較することができるこことを意味する。

まず、 $F_1(p)$ を第1期の価格 (P_1) の累積分布関数 (CDF) とすると、不動産属性 (z) といった条件付きの価格の分布は $F_1(p | z)$ と表すことができる。この時、価格 $F_1(p)$ と属性 $F_1(p | z)$ の関係は (2.11) 式のようになる。

$$F_1(p) = \int_{-\infty}^{\infty} F_1(p | z) u_1(z) dz \quad (2.11)$$

$u_1(z)$ は不動産価格を構成する属性 z の分布である。同様に、 $F_2(p)$ および $F_2(p | z)$ を第2期の不動産価格の属性 $u_2(z)$ に対応した不動産価格の累積分布関数とする。そうすると、 $F_1(p)$ から $F_2(p)$ の価格分布の変化は、(2.12) 式のようになる。

$$F_1(p) - F_2(p) = \int_{-\infty}^{\infty} [F_1(p | z) - F_2(p | z)] u_1(z) dz + \int_{-\infty}^{\infty} F_2(p | z) [u_1(z) - u_2(z)] dz \quad (2.12)$$

(2.12) 式の右側をみれば、第一項が不動産属性 z のもとでの品質調整済み不動産価格の差を表し、第二項がそれぞれの時点の不動産属性の相違を意味する。つまり、一般的に市場で観察される二次点の価格の分布の相違は、「価格の変化」+「属性の変化」といった二つの要素から観察されることになる。つまり、実際には第1期から第2期に対して価格が下落していたとしても、その変化は価格が変化したわけではなく、最寄り駅から遠い物件や築年が古い物件などが中心に取引され、属性の変化によって価格分布が下落しているように見えることもある。そうすると、二つの不動産価格の分布、つまり「相場」を比較しようとした場合には、この第二項である不動産属性の相違を取り除いたうえで価格を比較していくなければならないことがわかる。つまり、(2.13) 式のようと同じ属性 z のもとでの価格の相違を見なければならない。

$$\int_{-\infty}^{\infty} [F_1(p | z) - F_2(p | z)] u_1(z) dz \quad (2.13)$$

近年、内外において実用化されている価格予測システムは、実際には、その価格特性 z に対応した価格ベクトルを推計しているにすぎない。また、価格の変化については、品質調整済みの価格分布の中央値または平均値を示している。具体的には、次の手続きで計算することができる。

$Q_i^\theta(p | z)$ を価格の累積分布 ($F_i(p | z)$) の第 θ -番目の分位点とする ($\theta \in (0, 1)$)。これを次のように条件付き分位 (conditional quantiles) として定義する。

$$Q_i^\theta(p | z) = z\beta_i(\theta) \quad (2.14)$$

条件付き分位は、様々な不動産属性の加重平均として考えられる。ここでは属性価格 $\beta_i(\theta)$ は、 θ 点の価格水準に依存するものと考えればよいため、各分位点の回帰係数と考える。まず、第一期の価格 (P_1) の $\beta_1(\theta)$ を推計するために、 P_1 を用いた分位点ごとの回帰を行うことで、推定統計量の $\hat{\beta}_1(\theta)$ を得る。そうすると、不動産属性 z を所与とすれば、 $p = z\hat{\beta}_1(\theta)$ によって $F_1(p | z)$ が計算される。 $F_1(p | z)$ の推計値を $\hat{F}_1(p | z)$ とする。同様の方法で第二期の価格 P_2 の条件付き価格の累積分布 $F_2(p | z)$ の推計値は、 $\hat{F}_2(p | z)$ となる。そうすると z に関して積分することによって、次のように表現できる。(Shimizu, Nishimura and Watanabe(2016)[10])

$$\begin{aligned} \hat{F}_1(p) &\equiv \int_{-\infty}^{\infty} \hat{F}_1(p | z) u_1(z) dz; \\ \hat{F}_2(p) &\equiv \int_{-\infty}^{\infty} \hat{F}_2(p | z) u_2(z) dz \end{aligned} \quad (2.15)$$

そうすると、実際の計算においては、数式 (2.12) は、次のように書き換えることができる。

$$\hat{F}_1(p) - \hat{F}_2(p) = \int_{-\infty}^{\infty} [\hat{F}_1(p | z) - \hat{F}_2(p | z)] u_1(z) dz + \int_{-\infty}^{\infty} \hat{F}_2(p | z) [u_1(z) - u_2(z)] dz \quad (2.16)$$

相場の水準や変化を予測するサービスは、それぞれの不動産に関する属性に対する係数または重みを推計し、時間の変化に関する係数などを用いて、それをわかりやすい形で消費者に提供されているだけである。

2.4 不動産価格の実際の推計

不動産価格に対して、交通利便性や規模・建築後年数等の相違などの特性に応じた係数または重みを計算する技術は、回帰分析に代表されるように古くから活用されてきた。近年において、再度注目されている機械学習と呼ばれる手法の中でも、中心的な役割を担い、多くの分野で実用化されているのがニューラルネットワークとよばれる技術が進化した深層学習 (Deep Learning) や回帰木と呼ばれる手法である。そのような手法は、どの程度の予測力を持つのであろうか。

ニューラルネットワークや回帰木と呼ばれる推計手法は、そもそも伝統的な回帰分析などと発達するモデルベーションが異なっていた。それらの技術は、「機械は人間を超えることができるのか?」という問い合わせ意識して進化してきた。この研究は、1940 年代に本格化し、フォン・ノイマンによる直列処理による計算機とともに、並列処理のそれに対する研究という大きな二つの流れのなかで、計算技術の問題として行われてきたのである。

ニューラルネットワークとは、人間の脳神経細胞の並列処理システムを模して開発がすすめられた。人間の脳神経細胞 (neuron: ニューロン) は、細胞体・樹状突起・軸索の 3 つの部分から構成されており、さらに軸索の末端にはシナプス (synapse) という部位があり、ここを通じて各細胞の情報伝達が行われる。

そして、脳神経細胞の情報伝達は、軸索・樹状突起の結合部で行われてあり、それはシナプス結合と呼ばれている。このような脳神経細胞の情報伝達を模して開発されたのがニューラルネットワークなのである。その詳細は、清水(2016)[5]または、本書の第3章、4章、6章をご覧いただきたい。

また、このようなクロスセクショナルな価格を予測するという行為とともに、価格の時間的な変化をリアルタイムに測定したいということも多い。一般的に、不動産専門家が相場の水準や変化を大きく見誤るのは、市場の転換点である。そして、従来は、市場の変化を公示地価に代表されるような不動産鑑定士によって決定された価格によって観察されることが多かったために、しばしば見間違えることがあった。Shimizu and Nishimura(2006)[8]における統計実験では、実際の取引価格と公示地価といった不動産鑑定価格とのかい離を時間的に調べている。その結果を見ると、1990年代のバブル期には、不動産鑑定価格は実際の市場価格の半分から6割程度であり、バブル崩壊後には、2割程度高い価格がつけられていたことを示している。その最も大きな原因が、相場を決定する技術よりも、情報の入手速度と選択技術に起因していると結論付けている。

具体的には、従来、不動産鑑定士がその価格決定において利用可能な取引事例と呼ばれるデータ基盤は、アンケート調査に基づくことから3か月から半年程度遅れて入手される。さらには、その網羅性も3割程度と低い。そのため、高い価格決定技術を有していても、データ基盤の脆弱性からその専門性を生かすことができていない。そのような情報入手の時間的なラグや網羅性の低さから、系統的な誤差が生まれてしまうのである。

一方で、近年におけるIOT技術の発達によって、一週間、または数日、場合によっては数時間の時間的な粒度の中で不動産価格に関する大規模データ基盤が更新することができるようになってきた。また、地理情報基盤の整備も進化していることから、実際に人間によって調査を行わなくても、不動産に関する情報だけでなく、地域情報も入手が可能である。また、サイトに対するログも利用することができるようになってきたことで、消費者が求めている各属性別の選好を、その解析を通じて得ることも可能となった。

そうすると、そのように時間的に粒度の細かいデータ基盤をどのように学習させ、リアルタイムに価格を決定することができる技術があるかどうかということが課題になる。その問題に対応するために、Hill, Scholz, Shimizu and Steurer(2018)[2]では、東京とシドニーのデータを用いて、一定の精度を担保しつつ週次単位で予測していくことが可能であることが示された。具体的には、その週単位で得られたデータを用いた価格推計は極めてボラティリティが高くなり、実用化が困難であるものの、一定の期間を重複されることで、時間的にも安定した価格査定が実現できることを示している。

実際の推計手順を示せば、ある週単位での期間 $1, 2, \dots, T$ 期のうちの r 期からはじまる τ 期間を $[r, r + \tau - 1]$ のように表すと、 $[1, \tau]$ のようにそれぞれの期のデータを用いて推計するのではなく、 $[2, \tau + 1] \sim [r, r + \tau - 1] \sim [T - \tau + 1, T]$ といったように、一定の期間を重複させつつ逐次的に適用することで時間的な安定性を持つことが示されている。つまり、移動平均のように、一定の期間の情報を共有しつつ、新しく出現する情報を追加しながらモデルを推計していくのである。これにより市場構造の逐次的な変化をパラメータに反映させることができるとなる(Shimizu, Nishimura and Watanabe(2010)[9])。

これは、一つの手法に過ぎないが、タイムリーかつ大規模なデータ基盤の整備によって、従来において不動産の専門家が直面していた課題が克服され、より正確な相場の決定やその粒度の細かい相場の時間的な変化を、機械学習などによって推計することができるようになってきたのである。時系列的な不動産価格の測定方法の詳細は、第8章を参照されたい。

参考文献

- [1] Epple, D., (1987), “Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products”, *Journal of Political Economy*, Vol.95, pp.58-80.
- [2] Hill, R, M. Scholz, C. Shimizu and M. Steurer (2018) “An Evaluation of the Methods Used by European Countries to Compute their Official House Price Indices,” *Economie et Statistique* n 500-501-502, 221–238.
- [3] Lancaster, K., (1966), “A new approach to consumer theory”, *Journal of Political Economy*, Vol.74, pp.132-157.
- [4] Rosen, S., (1974), “Hedonic Prices and Implicit Markets, Product Differentiation in Pure Competition”, *Journal of Political Economy*, Vol.82, pp.34-55.
- [5] 清水千弘 (2016) ,「市場分析のための統計学入門」朝倉書店.
- [6] 清水千弘 (2017) ,「ビッグデータで見る不動産価格の決まり方」不動産学会誌 , 120 号,45-51.
- [7] 清水千弘・唐渡広志 (2017) 『不動産市場の経済経済分析』朝倉書店.
- [8] Shimizu, C. and K.G.Nishimura, (2006), “Biases in Appraisal Land Price Information: The Case of Japan”, *Journal of Property Investment and Finance*, Vol.26, No.2, pp.150-175.
- [9] Shimizu, C., K. G. Nishimura and T. Watanabe (2010), “House Prices in Tokyo - A Comparison of Repeat-sales and Hedonic Measures-,” *Journal of Economics and Statistics*, 230 (6), 792-813.
- [10] Shimizu,C, K.G.Nishimura and T.Watanabe(2016), “House Prices at Different Stages of Buying/Selling Process ,”*Regional Science and Urban Economics*, 59, 37-53.
- [11] Tinbergen, J., (1959), “On the theory of income distribution”, in: L.M.K.L.H. Klaasen and H.J. Witteveen, eds, *Selected Paper of Jan Tinbergen* (North-Holland, Amsterdam).
- [12] Witte, A. D., H. Sumka and J. Erikson, (1979), “An Estimate of a Structural Hedonic Price Model of the Housing Market: An Application of Rosen’s Theory of Implicit Markets”, *Econometrica*, Vol.47, pp.1151-72.

第3章

不動産テックにおける機械学習の数理

3.1 不動産市場分析と機械学習

不動産市場分析における、不動産の経済的な価値を予測する、間取り図などの図面や不動産を取り巻く周辺環境をセグメンテーション化する、将来において空き家になる住宅を予測するなどといった目的を達成するために、機械学習と呼ばれる技術は、極めて有効に活用することができる。ここでは、そのような機械学習と呼ばれる技術の基本的な考え方を概観することを目的とする。機械学習の数理についてはすでに Bishop (2011)[1] , Murphy (2012)[4] , Goodfellow et al. (2016)[2] のような定評のある成書やその日本語訳も出版されている。本章はそういった本格的な本を読む前に、機械学習の「あらすじ」を把握する目的で読んでいただければ幸いである

データに潜んでいる法則性を見つけ出し、それに基づいた予測や判断を行うためのアルゴリズムの総称を機械学習という。Mitchell (1997)[3] は機械（コンピュータプログラム）が学習をするとは、経験 E を通じてタスク T のパフォーマンス P が向上すること、と定義している。この定義によれば、機械学習の基本的な構成要素はタスク T 、パフォーマンス尺度 P 、経験 E ということになる。タスク T とは例えば、不動産の価格予測、画像認識であり、経験 E とは過去に取得された不動産価格やその不動産の属性データ、画像データである。パフォーマンス尺度 P は例えば、予測された不動産価格と真の値との差や画像認識の正答率のような、誤差を計測する尺度である。機械学習にはいくつかの分類があるが、ここでは教師あり学習、教師なし学習の2つを紹介する。

教師あり学習では、まず N 個のデータのペア $\{(x_i, y_i)\}_{i=1,2,\dots,N}$ が与えられる^{*1}。これらを訓練データと呼ぶ。ここで x_i は特徴量と呼ばれる入力データ、 y_i は x_i に対応する正解（またはラベル）と呼ばれる出力データを表す^{*2}。教師あり学習では訓練データから特徴量 x と正解 y の対応関係を表す関数 f ($y = f(x)$) を推定することが目標となる。 y が連続的なデータのとき、このタスクを回帰という。例えば、 y が不動産価格で、 x が最寄り駅からの距離や建築年数などの属性データであるとき、 x から y を推定する問題が回帰である。一方、 y が離散的なデータ、例えば y が 0 または 1 の値をとるとき、このタスクを分類という。例えば、 x が動物の画像データであり、 y が犬か ($y = 1$)、そうでないか ($y = 0$) を表すような場合である。機械学習の多くのアルゴリズムでは関数 f を直接推定するのではなく、パラ

^{*1} 数学的表記については章末を参照のこと。

^{*2} 正解 y_i はベクトル y_i であってもよい。

メータ化された関数 $f(\mathbf{x}; \mathbf{w})$ を考え、パラメータ \mathbf{w} を推定する問題に帰着させる。本章でもパラメータ化された関数の推定のみを扱う。

パフォーマンス尺度 P は何らかの誤差によって定義される。この誤差を表す関数を損失関数と呼ぶことにしよう^{*3}。パラメータ化された関数 $f(\mathbf{x}; \mathbf{w})$ を考えるときには、損失関数はパラメータ \mathbf{w} の関数 $L(\mathbf{w})$ となる。モデルを $y = f(\mathbf{x}; \mathbf{w})$ としたときの、個々の訓練データ (\mathbf{x}_i, y_i) の誤差を関数 l を用いて $l(\mathbf{x}_i, y_i; \mathbf{w})$ と表現すれば、損失関数は

$$L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N l(\mathbf{x}_i, y_i; \mathbf{w}) \quad (3.1)$$

と書くことができる。例えば、誤差を二乗誤差により定義すると、

$$L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \{y_i - f(\mathbf{x}_i; \mathbf{w})\}^2 \quad (3.2)$$

が損失関数となる。そして、最適化問題

$$\min_{\mathbf{w}} L(\mathbf{w})$$

の解 \mathbf{w}^* によって得られた関数 $f(\mathbf{x}; \mathbf{w}^*)$ が推定されたモデルとなる。新たな特徴量 \mathbf{x}' が得られたとき、 $\hat{y} = f(\mathbf{x}'; \mathbf{w}^*)$ を正解の推定値として予測や判断を行う。ここで、損失関数 (3.1) はモデル $y = f(\mathbf{x}; \mathbf{w})$ の訓練データへの当てはまりのよさを計測していることに注意する。機械学習の本来の目的は未知のデータを含むあらゆるデータに対してよいパフォーマンスをもつモデルを推定することにある。すなわち、

$$L_{\text{true}}(\mathbf{w}) = E[l(\mathbf{x}, y; \mathbf{w})] \quad (3.3)$$

を最小にする \mathbf{w} を求めることにある。ここで、(3.3) における $E[\cdot]$ は (\mathbf{x}, y) を生成している確率分布のもとでの期待値である。(3.3) によって計測する誤差を汎化誤差という。一方、(3.1) によって計測する誤差は訓練誤差と呼ばれる。機械学習の本来の目的は汎化誤差の最小化である。しかし (\mathbf{x}, y) を生成している確率分布を厳密に知ることは実際には不可能であり、したがって汎化誤差を計測することも実際には不可能である。よって、訓練誤差を汎化誤差の近似として最小化を行わざるを得ない。しかし、汎化誤差を意識することは過学習を防ぐためにも重要である。過学習については、後の節で触れる。

本章は教師あり学習に焦点を当てるが、教師なし学習についても簡単に触れておく。教師なし学習では、はじめに N 個のデータ $\{\mathbf{x}_i\}_{i=1,2,\dots,N}$ が与えられる。教師あり学習と違い、教師なし学習では正解データが与えられない。ここでのタスクはデータ \mathbf{x}_i の背後に潜んでいる構造を探し出すことである。例えば、教師なし学習の問題には、データを類似したいくつかのグループに分類するクラスタリング、高次元データを可視化するための次元削減などがある。

3.2 勾配降下法

機械学習においては損失関数 $L(\mathbf{w})$ の最小化問題

$$\min_{\mathbf{w}} L(\mathbf{w}) \quad (3.4)$$

^{*3} 誤差関数やコスト関数と呼ぶこともある。

を解くことが鍵となる。この節では最小化問題の解法について述べる。

もし \mathbf{w}^* が (3.4) の解であるならば

$$\nabla L(\mathbf{w}^*) = 0 \quad (3.5)$$

を満たす。ここで $\nabla L(\mathbf{w})$ は勾配ベクトルで

$$\nabla L(\mathbf{w}) = \left(\frac{\partial L}{\partial w_1}(\mathbf{w}), \frac{\partial L}{\partial w_2}(\mathbf{w}), \dots, \frac{\partial L}{\partial w_D}(\mathbf{w}) \right)^T$$

により定義される。ここで、 D はベクトル \mathbf{w} の要素数を表す。 \mathbf{w} が 1 次元の場合は通常の微分 $L'(w)$ である。しかし、ある \mathbf{w} が (3.5) を満たしたとしても、それが最小化問題 (3.4) の解になるとは限らず、局所的最適解になっている可能性がある（図 3.1(a)）。局所的最適解に対して、(3.4) の解を大域的最適解と呼ぶ。関数 L が凸関数であるときには、条件 (3.5) を満たす \mathbf{w} は大域的最適解になる。関数 L が条件

$$L(t\mathbf{w}_1 + (1-t)\mathbf{w}_2) \leq tL(\mathbf{w}_1) + (1-t)L(\mathbf{w}_2), \quad \forall t \in [0, 1], \quad \forall \mathbf{w}_1, \mathbf{w}_2$$

を満たすとき、凸関数であるという（図 3.1(b)）。いくつかのモデルでは損失関数は凸関数となるが、ニューラルネットワークなどでは損失関数が必ずしも凸にならないので、注意が必要となる。非凸関数に対して (3.4) を解くことは一般には難しく、通常は局所的最適解を求めて満足することが多い。以下でも、局所最適解を求める手法のみを考える。

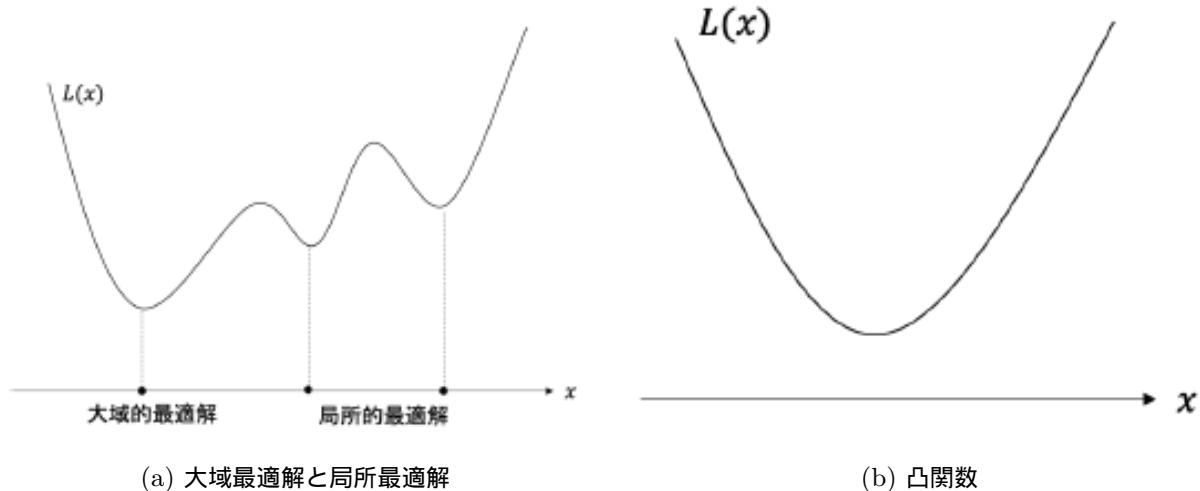


図 3.1 最小化問題の解

線形回帰の場合など $\nabla L(\mathbf{w}) = 0$ が解析的に解ける、すなわち $\mathbf{w} = \dots$ という表現を得られることがある。しかし、機械学習ではデータの次元が非常に大きく、解析的表現による解法は実用的ではないことが多い。そこで、反復法を用いて $\nabla L(\mathbf{w}) = 0$ を満たす \mathbf{w} を近似することを考える。勾配ベクトル ∇L の情報をもとに (3.5) の解を近似する手法を勾配降下法という。一般に実数値関数 $f(\mathbf{w})$ について、ベクトル \mathbf{d} に対して、ある実数 $\eta > 0$ が存在して $f(\mathbf{w} + \eta\mathbf{d}) < f(\mathbf{w})$ となるとき、ベクトル \mathbf{d} を \mathbf{w} における降下方向とよぶ。 $\nabla f(\mathbf{w})^T \mathbf{d} < 0$ を満たす \mathbf{d} は降下方向である。なぜなら、十分小さく $\eta > 0$ を取れば、テイラーの定理より

$$f(\mathbf{w} + \eta\mathbf{d}) \approx f(\mathbf{w}) + \eta \nabla f(\mathbf{w})^T \mathbf{d} \quad (3.6)$$

であるから、 $\nabla f(\mathbf{w})^T \mathbf{d} < 0$ より $f(\mathbf{w} + \eta \mathbf{d}) < f(\mathbf{w})$ とすることができる。 $\mathbf{d} = -\nabla f(\mathbf{w})$ が \mathbf{w} における降下方向になることはすぐに分かる。 \mathbf{w} が 1 次元の場合、 $-f'(x)$ が降下方向であることは明らかであろう（図 3.2(b)）。

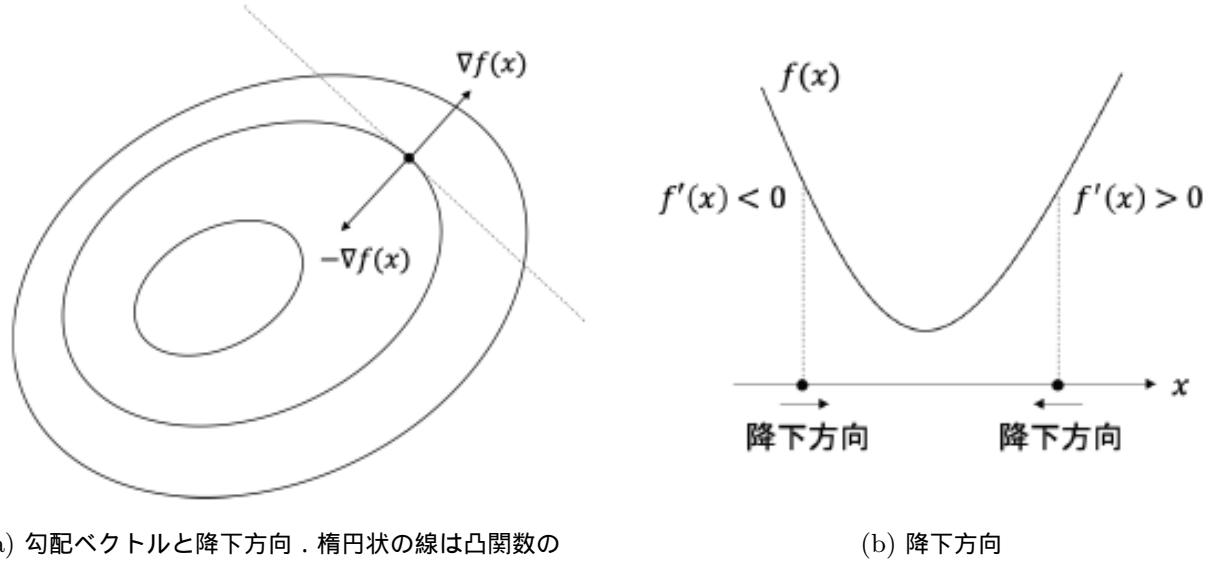


図 3.2 勾配降下法

勾配降下法^{*4}はつぎのような反復法である。

勾配降下法

ステップ 1. 適当な初期点 $\mathbf{w}(0)$ を選び、 $k = 0$ とする。

ステップ 2. $L(\mathbf{w}(k)) = 0$ ならば解を $\mathbf{w}^* = \mathbf{w}(k)$ として、反復法を終了する^{*5}。

ステップ 3. 適当なステップ幅 $\eta(k) > 0$ に対して

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \eta(k) \nabla L(\mathbf{w}(k))$$

とする。

ステップ 4. $k := k + 1$ (k の値を 1 増やす) としてステップ 2 に戻る。

すなわち、勾配降下法とは点 $\mathbf{w}(0)$ から関数 L が減少する方向 $-\nabla L(\mathbf{w}(0))$ に沿って少し進み点 $\mathbf{w}(1)$ に到達、そこで $-\nabla L(\mathbf{w}(1))$ に方向を変えてさらに少し進んで点 $\mathbf{w}(2)$ に到達、そこで $-\nabla L(\mathbf{w}(2))$ に方向を変えて、という手続きを $\nabla L(\mathbf{w}(k)) = 0$ となるまで続ける、というアルゴリズムである。

各ステップで設定するステップ幅 $\eta(k)$ を機械学習では学習率と呼ぶ。機械学習において分析者があらかじめ設定すべきパラメータをハイパー-パラメータと呼び、学習率はハイパー-パラメータの一つである。学習率をステップ k によらず一定値 η とすることもある。学習率の設定法については問題に応じていろいろな手法が開発されているが、決定的な方法は存在せず、試行錯誤により設定せざるを得ない。 η を小

^{*4} 最急降下法と呼ぶこともある。

^{*5} 実際にはあらかじめ決められた十分小さな ϵ に対して、 $\|\nabla L(\mathbf{w}(k))\| < \epsilon$ となったところで終了する。

さく取りすぎると解への収束が遅くなる一方、 η を大きく取りすぎると(3.6)の議論から分かるように、関数 L の値が減少しない点に $\mathbf{w}(k)$ が到達する可能性がある。この学習率の設定が解を求めるための鍵となる。

前節で述べたように機械学習では最小化する関数が

$$L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N l(\mathbf{x}_i, y_i; \mathbf{w})$$

という形をしている。機械学習においては訓練データ数 N は非常に大きい。したがって、損失関数の勾配の式

$$\nabla L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \nabla l(\mathbf{x}_i, y_i; \mathbf{w})$$

から分かるように、勾配降下法において損失関数の勾配を求めるたびに N 個の勾配 $\{\nabla l(\mathbf{x}_i, y_i; \mathbf{w})\}_{i=1, \dots, N}$ の計算が必要になり計算負荷が非常に大きくなってしまう。そこで N 個のデータをすべて使わずに、1つまたは少数のデータのみを使って勾配ベクトルを計算する、確率的勾配降下法という反復法を用いる。確率的勾配降下法では勾配降下法のステップ 3 がつぎのようになる：

確率的勾配降下法

ステップ 3. (ランダムに) 選んだ訓練データ (\mathbf{x}_i, y_i) を用いて $\nabla l(\mathbf{x}_i, y_i; \mathbf{w}(k))$ を計算し、適当なステップ幅 $\eta(k) > 0$ に対して

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \eta(k) \nabla l(\mathbf{x}_i, y_i; \mathbf{w}(k))$$

とする。(他のステップは勾配降下法と同じ)

勾配の計算に用いる訓練データ (\mathbf{x}_i, y_i) はランダムに選んでもよいし、何らかの順序に従って選んでもよい。また、訓練データがつぎつぎと到着する状況において、訓練データを 1 つ観測する度にパラメータ \mathbf{w} を更新する場合にも確率的勾配降下法が利用できる。

3.3 線形回帰

ここからは機械学習のアルゴリズムを具体的に見ていくことにする。まずは線形回帰を扱う。線形回帰では特徴量 \mathbf{x} と正解 y の間の関係をパラメータ $\mathbf{w} = (w_0, w_1, \dots, w_D)^T$ を用いて

$$y = w_0 + \sum_{i=1}^D w_i x_i$$

とモデル化する。機械学習では w_0 をバイアスと呼ぶ。線形回帰モデルは統計学や計量経済学においてもよく用いられるモデルである。しかし、統計学や計量経済学においては回帰係数 w_i の統計的有意性や特徴量（説明変数）と正解（独立変数）の間の相関関係または因果関係に主な関心がある一方、機械学習においては汎化誤差の最小化、すなわち精度のよい予測をすることに主な関心がある。また、線形回帰モデルを用いると機械学習のさまざまな概念を比較的分かりやすく説明できるため、線形回帰モデルは機械学習の教科書で取り上げられることが多い。

より一般的に

$$y = w_0 + \sum_{j=1}^M w_j \phi_j(\mathbf{x}) \quad (3.7)$$

というモデルを考えることもできる。ここで $\{\phi_j\}_{j=1,2,\dots,M}$ は基底関数と呼ばれる非線形関数である。非線形関数を用いることでモデルの表現能力が上昇する、すなわち訓練誤差の小さいパラメータを求めやすくなる。特徴量 \mathbf{x} が 2 次元 ($\mathbf{x} = (x_1, x_2)^T$) のときは

$$\phi_j(\mathbf{x}) = x_1^{3-j} x_2^{j-1}, \quad j = 1, 2, 3$$

が基底関数の一例となる。非線形関数を使うものの、モデル (3.7) はパラメータ \mathbf{w} に関して線形であるため線形回帰モデルと呼ばれる。また、 $\phi_0(\mathbf{x}) = 1$ と定義することで、(3.7) を $y = \mathbf{w}^T \phi(\mathbf{x})$ と簡潔に表現することができる。ここで、 $\mathbf{w} = (w_0, w_1, \dots, w_M)^T$ 、 $\phi(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \dots, \phi_M(\mathbf{x}))^T$ である。

線形回帰では平均二乗誤差をパフォーマンス尺度とし、損失関数は

$$L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \{y_i - \mathbf{w}^T \phi(\mathbf{x}_i)\}^2 \quad (3.8)$$

となる。損失関数 $L(\mathbf{w})$ は凸関数であるため、最小化問題 $\min_{\mathbf{w}} L(\mathbf{w})$ の解は $\nabla L(\mathbf{w}) = 0$ を解くことによって求めることができる。そこで

$$\nabla L(\mathbf{w}) = -\frac{1}{2N} \sum_{i=1}^N (y_i - \mathbf{w}^T \phi(\mathbf{x}_i)) \phi(\mathbf{x}_i)$$

であることを用いると、最適解は

$$\mathbf{w}^* = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} \quad (3.9)$$

と求まる。ここで $\mathbf{y} = (y_1, y_2, \dots, y_N)^T$ 、

$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \dots & \phi_M(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \dots & \phi_M(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \dots & \phi_M(\mathbf{x}_N) \end{pmatrix}$$

である。数学的には簡潔な解を求めることができるが、機械学習ではデータ数 N が非常に大きいことが通常であり、その場合 (3.9) の逆行列 $(\Phi^T \Phi)^{-1}$ を計算することは現実的でなく^{*6}、実際には $\nabla L(\mathbf{w}) = 0$ の解を（確率的）勾配降下法を使って近似する。

さて、ここで単純な線形回帰モデルを用いて過学習について見ておく。（機械学習では実際にはありえないが）特徴量が 1 つ ($D = 1$) の場合で、基底関数を $\phi_j(x) = x^j$ とした線形回帰モデル

$$y = \sum_{j=0}^M w_j x^j \quad (3.10)$$

^{*6} 逆行列が存在しないこともある。

を考える。ここで M はハイパーパラメータとなる。 M を大きく取ると訓練誤差を小さくできるが、モデルが訓練データにフィットしすぎてしまい、汎化誤差が大きくなる、すなわち未知のデータに対する予測能力が落ちてしまう現象が起きる。この現象を過学習と呼ぶ。訓練データの数が少ないとパラメータが多い複雑なモデルを用いると過学習が起きやすくなる。極端な例ではあるが、異なる訓練データの数が $(N + 1)$ 個のときに最大次数 $M = N$ の多項式でモデル化をすると、訓練誤差をゼロにするパラメータ $\{w_j\}_{j=0,1,\dots,M}$ を求めることができる（図 3.3）。しかし、そういうモデルの未知のデータに対する予測力は怪しい。一方で M を小さく取りすぎると訓練誤差と汎化誤差が大きくなる過小学習という現象が起きる可能性がある。

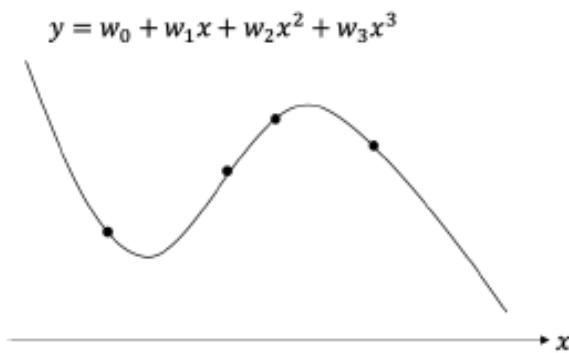


図 3.3 過学習。データ数が 4 のときに 3 次関数でモデル化。

過学習を防ぐ方法の一つとして正則化がある。正則化とは損失関数に正則化項と呼ばれるパラメータ w が大きくなることに対するペナルティ項を加えて、最小化問題を解く手法である。正則化学習では例えば最適化問題

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^N l(\mathbf{x}_i, y_i; \mathbf{w}) + \lambda \sum_{j=1}^M w_j^2$$

を解く。ここで $\sum_{j=1}^M w_j^2$ が L^2 正則化項と呼ばれる正則化項、 $\lambda > 0$ は正則化の強さを表すハイパーパラメータである。正則化項 $\sum_{j=1}^M w_j^2$ を最小化問題に組み込むことで、なるべくゼロに近いパラメータ $\{w_j\}_{j=1,2,\dots,N}$ が求められるようになり、過学習を防ぐことができる。ほかにもさまざまな正則化項が考えられており、 L^1 正則化項 $\sum_{j=1}^M |w_j|$ もよく用いられる。 L^1 正則化項を用いると、ゼロとなるパラメータ $\{w_j\}_{j=1,2,\dots,N}$ の数が多くなるように最小化問題が解かれる。

ここまでいくつかのハイパーパラメータが出てきた。ハイパーパラメータはどのように選択すればよいのだろうか。当然汎化誤差を小さくするように選ぶべきであるが、実際に汎化誤差を計測することはできない。そこで、交差検証と呼ばれる技術を使う。交差検証とは、はじめに与えられたデータを訓練データと検証データに分割し、訓練データを用いてパラメータを推定し、検証データによって誤差を計測する手法である。訓練データと検証データへの分割の仕方を何回か変えて、それぞれの回での誤差を求め、それらを平均することでモデルのよさを検証する。ハイパーパラメータをいろいろに変化させ交差検証を行うことで、最も適当なハイパーパラメータを選ぶことができるようになる。

3.4 分類（ロジスティック回帰）

正解 y が離散値を取る分類問題をロジスティック回帰と呼ばれるアルゴリズムを使って解くことを考える。2 クラスに分類する問題では、正解 y は 0 または 1 の値をとり、例えば $y = 1$ のときクラス 1, $y = 0$ のときクラス 2 と定義する。 $K (> 2)$ クラスに分類する多クラス分類問題では、クラス k に属することを、 k 番目の要素が 1, それ以外の要素はすべて 0 である K 次元ベクトルで表現する。例えば、クラス 3 に属することを $y = (0, 0, 1, 0, \dots, 0)^T$ という K 次元ベクトルで表現する。

まずは 2 クラス分類問題、すなわち正解 y が $y \in \{0, 1\}$ となる場合を扱う。2 クラス分類問題では識別関数 $c(\mathbf{x}; \mathbf{w})$ を用意し $c(\mathbf{x}; \mathbf{w}) > 0$ ならばクラス 1 ($y = 1$), $c(\mathbf{x}; \mathbf{w}) < 0$ ならばクラス 2 ($y = 0$) となるように、訓練データからパラメータ \mathbf{w} を推定することが目標となる。ここでは線形識別関数

$$c(\mathbf{x}; \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D = \mathbf{w}^T \mathbf{x} \quad (3.11)$$

を考える。ここで $\mathbf{w} = (w_0, w_1, \dots, w_D)^T$, $\mathbf{x} = (1, x_1, \dots, x_D)^T$ とした。線形回帰の場合と同様に、基底関数を使って識別関数を $c(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \phi(\mathbf{x})$ としてもよい。線形識別関数によって 2 クラスを分類するということは、2 つのクラスを超平面（直線）で区分することになる（図 3.4(a)）。ロジスティック回帰では特徴量 \mathbf{x} と正解 y の関係を、シグモイド関数 σ を作用させた $\sigma(\mathbf{w}^T \mathbf{x})$ を用いて

$$y = \begin{cases} 1, & \sigma(\mathbf{w}^T \mathbf{x}) \geq 0.5 \\ 0, & \sigma(\mathbf{w}^T \mathbf{x}) < 0.5 \end{cases} \quad (3.12)$$

とモデル化する。シグモイド関数 σ （ロジスティック関数とも言う）とは

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

によって定義される単調増加関数である（図 3.4(b)）。シグモイド関数は $[0, 1]$ に値を取る関数であり、 $\mathbf{w}^T \mathbf{x}$ にシグモイド関数を作用させることで $\sigma(\mathbf{w}^T \mathbf{x})$ を確率とみなすことができるようになる。そこで、特徴量 \mathbf{x} を観測したときにそれがクラス C_k ($k = 1, 2$) に含まれる確率 $P(C_k | \mathbf{x})$ を

$$P(C_1 | \mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x})$$

と定義する。すると、(3.12) は 2 つの確率 $P(C_1 | \mathbf{x})$ と $P(C_2 | \mathbf{x}) = 1 - P(C_1 | \mathbf{x})$ を比較して値の大きい方のクラスに特徴量 \mathbf{x} を分類するというルールになっている。なお、(3.12) は非線形関数によって識別をしているように見えるが、 $\sigma(\mathbf{w}^T \mathbf{x}) = 0.5$ を解くと $\mathbf{w}^T \mathbf{x} = 0$ であり、線形識別関数による識別であることが分かる。

パラメータ \mathbf{w} は最尤法を用いて推定する。 $\pi(\mathbf{x}_i; \mathbf{w}) = P(C_1 | \mathbf{x}_i)$ とおく。訓練データ $\{(\mathbf{x}_i, y_i)\}_{i=1,2,\dots,N}$ が観測される確率は

$$P(\mathbf{w}) = \prod_{i=1}^N \pi(\mathbf{x}_i; \mathbf{w})^{y_i} (1 - \pi(\mathbf{x}_i; \mathbf{w}))^{1-y_i}$$

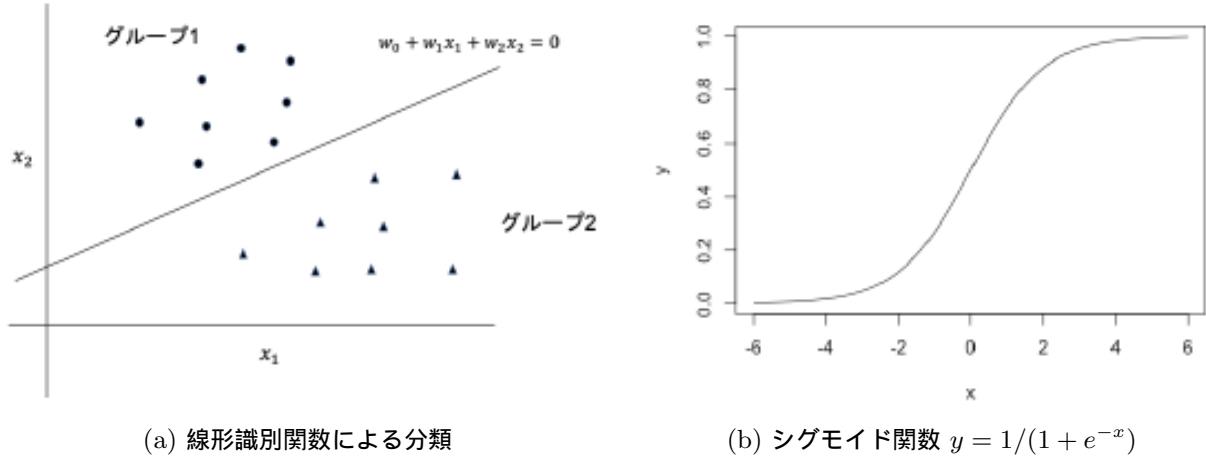


図 3.4 ロジスティック回帰

となる。この $P(\mathbf{w})$ は尤度関数と呼ばれる。最尤法とは尤度関数 $P(\mathbf{w})$ を最大にするパラメータ \mathbf{w}^* を推定値とする手法である。簡単のため $P(\mathbf{w})$ を最大化するのではなく、 $-\log P(\mathbf{w})$ すなわち

$$L(\mathbf{w}) = - \sum_{i=1}^N \{y_i \log(\pi(\mathbf{x}_i; \mathbf{w})) + (1 - y_i) \log(1 - \pi(\mathbf{x}_i; \mathbf{w}))\}$$

の最小化問題を解く。この $L(\mathbf{w})$ がロジスティック回帰における損失関数となる。あとは

$$\nabla L(\mathbf{w}) = \sum_{i=1}^N (\pi(\mathbf{x}_i; \mathbf{w}) - y_i) \mathbf{x}_i$$

であることを使って、確率的勾配降下法などを用いて最適解 \mathbf{w}^* を求めればよい。

K クラス分類の場合には K 個の線形関数 $\mathbf{w}_k^T \mathbf{x}$ を用いて識別をする。ここで、 $\mathbf{w}_k = (w_{k0}, w_{k1}, \dots, w_{kD})^T$ である。この場合には、シグモイド関数の代わりにソフトマックス関数と呼ばれる関数を使って

$$P(C_k | \mathbf{x}) = \frac{\exp(\mathbf{w}_k^T \mathbf{x})}{\sum_{j=1}^K \exp(\mathbf{w}_j^T \mathbf{x})}$$

として、最尤法を用いてパラメータ $\{\mathbf{w}_k^*\}_{k=1,\dots,K}$ を推定すればよい。そして、特徴量 \mathbf{x} を \mathbf{w}_k^* を用いて計算した $P(C_k | \mathbf{x})$ が最も大きくなるクラスに分類すればよい。

3.5 ニューラルネットワーク

機械学習(教師あり学習)とは訓練データ $\{(\mathbf{x}_i, y_i)\}_{i=1,2,\dots,N}$ から特徴量 \mathbf{x} と正解 y の関係 $y = f(\mathbf{x}; \mathbf{w})$ を推定することであった。前節までは

$$y = g \left(w_0 + \sum_{i=1}^D w_i x_i \right) \quad (3.13)$$

とモデル化し、パラメータ w を求める問題を主に扱った。関数 g は線形回帰では恒等関数 $g(u) = u$ であり、ロジスティック回帰では(3.12)によって定義した。この節ではより柔軟で表現力の高いモデルを構築できるニューラルネットワークを紹介する。

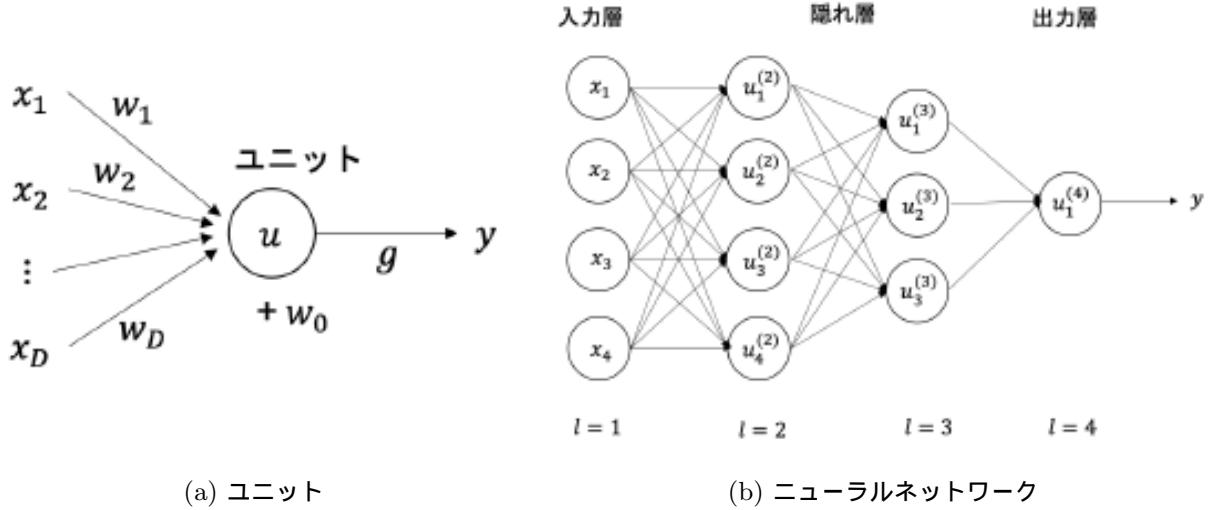


図 3.5 ニューラルネットワーク

(3.13)を図3.5(a)のように表現してみる。この図は、入力 x_1, x_2, \dots, x_D に重み w_1, w_2, \dots, w_D を掛けて足し合わせた $u = \sum_{i=1}^D w_i x_i$ をユニットが受け取り、バイアス w_0 を加えた上で関数 g による変換を施して $y = g(w_0 + u)$ を出力する、と読む^{*7}。このユニットを図3.5(b)のように層状に並べたものをニューラルネットワークとよぶ。とくに、入力から出力に向けて一方向に情報が流れるニューラルネットワークを順伝播型ニューラルネットワークと呼ぶ。また、ニューラルネットワークやそのバリエーション、それらを解くための方法などを総称して深層学習(ディープラーニング)と呼ぶ。本節では順伝播型ニューラルネットワークの基本を概観する。

l を層のインデックス、 L を層数として、最初の層($l = 1$)を入力層、最後の層($l = L$)を出力層、それ以外を隠れ層と呼ぶ。層数と各層のユニット数は任意に決めることができる。第 l 層 i 番目のユニットへの入力を $u_i^{(l)}$ 、第 l 層 i 番目のユニットからの出力を $z_i^{(l)}$ と表す。入力層 $l = 1$ においては $u_i^{(1)} = x_i$ であり、出力層 $l = L$ においては $z_i^{(L)} = y$ である。さらに、第 l 層 i 番目のユニットから第 $(l+1)$ 層 j 番目のユニットへの出力の重みを $w_{ji}^{(l+1)}$ 、バイアスを $w_{j0}^{(l+1)}$ と書くことにすれば、

$$u_j^{(l+1)} = w_{j0}^{(l+1)} + \sum_i w_{ji}^{(l+1)} z_i^{(l)} \quad (3.14)$$

となる。ここで、 $z_j^{(l+1)} = u_j^{(l+1)}$ をしてしまっては、入力 x と出力 y の関係は線形となってしまい、表現力がまったく改善しない。ニューラルネットワークでは活性化関数と呼ばれる関数 $h^{(l)}$ を用いて

$$z_j^{(l+1)} = h^{(l+1)}(u_j^{(l+1)}) \quad (3.15)$$

^{*7} g がヘビサイドの階段関数のとき、このモデルを(単純)パーセプトロンと呼ぶ。

とする。活性化関数としてはシグモイド関数や正規化線形関数 $h(u) = \max\{0, u\}$ などが用いられる。また、出力層における活性化関数として、回帰では恒等関数 $h(u) = u$ を使えばよく、2クラス分類ではシグモイド関数を用いた(3.12)を使えばよい。

(3.14)(3.15)はまとめて

$$\mathbf{z}^{(l+1)} = h^{(l+1)}(\mathbf{u}^{(l+1)}), \quad \mathbf{u}^{(l+1)} = \mathbf{W}^{(l+1)} \mathbf{z}^{(l)}, \quad l = 1, 2, \dots, L-1$$

と書くことができる。ここで

$$\mathbf{W}^{(l+1)} = \begin{pmatrix} w_{10}^{(l+1)} & w_{11}^{(l+1)} & w_{12}^{(l+1)} & \dots \\ w_{20}^{(l+1)} & w_{21}^{(l+1)} & w_{22}^{(l+1)} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad \mathbf{z}^{(l)} = \begin{pmatrix} 1 \\ z_1^{(l)} \\ z_2^{(l)} \\ \vdots \end{pmatrix}$$

である。 $\mathbf{W}^{(l+1)}$ は(第 $(l+1)$ 層のユニット数) \times (第 l 層のユニット数+1)行列、 $\mathbf{z}^{(l)}$ は(第 l 層のユニット数+1)次元ベクトルである。また、活性化関数 $h^{(l)}$ の引数がベクトル \mathbf{u} のときは $h^{(l)}(\mathbf{u}) = (h^{(l)}(u_1), h^{(l)}(u_2), \dots)^T$ と定義する。この表記を使うと、入力 \mathbf{x} と出力 y の関係は

$$y = h^{(L)} \left(\mathbf{W}^{(L)} h^{(L-1)} \left(\dots h^{(3)} \left(\mathbf{W}^{(3)} h^{(2)} \left(\mathbf{W}^{(2)} \mathbf{x} \right) \right) \right) \right) \quad (3.16)$$

と書ける。すなわち、入力 \mathbf{x} からスタートして線形変換 $\mathbf{W}^{(l)}$ と活性化関数 $h^{(l)}$ による変換をつぎつぎと作用させて、出力 y に至る。よって、かなり複雑ではあるが、入力 \mathbf{x} と出力 y の間の関係をパラメータ \mathbf{w} をもつ関数 f を用いて $y = f(\mathbf{x}; \mathbf{w})$ と表現するという機械学習の基本的な考え方は保たれている。表現力の高い非線形関数(3.16)によってモデル化することにより、例えば線形識別関数では分類できない問題も解ける可能性が出てくる。あとは、解きたい問題に応じて損失関数

$$L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N l(\mathbf{x}_i, y_i; \mathbf{w})$$

を定義して、最小化問題 $\min_{\mathbf{w}} L(\mathbf{w})$ を確率的勾配降下法などによって解けばよい。しかし、ニューラルネットワークにはいくつかの困難がある。

まず、(確率的)勾配降下法には勾配 $\nabla L(\mathbf{w})$ の計算が必要になるが、 $\nabla L(\mathbf{w})$ の計算には(3.16)によって定義される y のパラメータ w_{ji} に関する偏微分 $\partial y / \partial w_{ji}$ の計算が必要になる。単純に数値微分を適用すると、合成関数の微分を何度も繰り返すことになるため、計算量が莫大になり実用的ではない。しかし、誤差逆伝播法と呼ばれる微分計算がこの問題を解決する。誤差逆伝播法については第4章や参考文献を参照してほしいが、ニューラルネットワークを含む深層学習が実用化されたようになった理論的背景の一つは誤差逆伝播法の発明とその改良にある。つぎに、ニューラルネットワークにおける損失関数 $L(\mathbf{w})$ は一般には凸関数にはならないという問題がある。したがって、確率的勾配降下法を使って $\nabla L(\mathbf{w}) = 0$ の解を求めたとしても、それが大域的最適解になっている保障はない。また、ニューラルネットワークでは多くのパラメータを持つ非線形関数により入力 \mathbf{x} と出力 y の関係をモデル化することから、過学習が起こりやすことが容易に想像できる。したがって、深層学習においては正則化の技術が一層重要となる。また、層数や各層のユニット数をいくつにすればよいのかについても一般論は知られていない。

このようにさまざまな問題があるにも関わらず、ニューラルネットワークを含め深層学習が画像認識をはじめさまざまな問題に対して高いパフォーマンスを発揮している。しかし、なぜ深層学習がそこまでうまく機能するのかについては、理論的にはまだ分かっていないことが多い。

3.6 ノーフリーランチ定理

ここまでいくつかの機械学習のアルゴリズムを紹介してきた。ここで紹介したアルゴリズム以外にもさまざまなものが知られており、同じ問題に対して複数のアルゴリズムが適用できる場合はたくさんある。では、結局どのアルゴリズムを用いるのが一番よいのだろうか。最後に Wolpert (1996)[5] や Wolpert and Macready (1997)[6] によるノーフリーランチ定理を紹介する。これらの論文の記述は数学的であるが、簡単に述べれば「データに関して何の前提条件も設けなければ、あらゆる問題について最高のパフォーマンスをもつアルゴリズムは存在しない」もしくは「アルゴリズム A がある問題について最もよいパフォーマンスをもっていたとしても、それとは別の問題が存在して、そこでは別のアルゴリズム B がアルゴリズム A よりよいパフォーマンスをもつ」という定理である。すなわち、あらゆる問題を効率的に解く万能なアルゴリズムは存在しないということである。したがって、機械学習においてはデータや問題の特性にあわせたアルゴリズムを選択することが重要ということになる。

数学的表記の注意

本章ではイタリック体の小文字 x はスカラー（実数）を表し、太字のローマン体小文字 \mathbf{x} はベクトルを表す。すべてのベクトルは列ベクトルとする。 T は転置を表す記号であり、文中で列ベクトルを表記する際には、 \mathbf{x}^T という表記を使う。例えば、

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix}$$

のとき、 $\mathbf{x} = (x_1, x_2, \dots, x_D)^T$ と表記する。 $\mathbf{x}^T \mathbf{y}$ はベクトルの内積を表す。すなわち、 $\mathbf{y} = (y_1, y_2, \dots, y_D)^T$ として、

$$\mathbf{x}^T \mathbf{y} = \sum_{i=1}^D x_i y_i$$

である。行列は太字のローマン体大文字 \mathbf{A} で記す。また添字の解釈に注意する。本章では N で訓練データの数、 D で特徴量の数を表している。 \mathbf{x} が特徴量であるとき $\mathbf{x} = (x_1, x_2, \dots, x_D)^T$ を意味する。また、 \mathbf{x}_i は i 番目の訓練データの特徴量を表し、成分表示すれば $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iD})^T$ となる。また、ベクトル \mathbf{x} のノルム $\|\mathbf{x}\|$ を $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^D x_i^2}$ と定義する。

参考文献

- [1] Bishop, C. M. (2011). *Pattern Recognition and Machine Learning*. Springer. (元田浩ら訳,『パターン認識と機械学習(上)(下)』,丸善出版,2007年)
- [2] Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press. (岩澤有祐ら監訳,『深層学習』,アスキードワンゴ,2018年)
- [3] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill .
- [4] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
- [5] Wolpert, D. H. (1996). “The lack of a priori distinctions between learning algorithms,” *Neural Computation*, 8(7), 1341–1390.
- [6] Wolpert, D. H. and W. G. Macready (1997) “No free lunch theorems for optimization,” *IEEE Transactions on Evolutionary Computation*, 1(1), 67-82

第4章

不動産市場分析における統計・機械学習の利用

4.1 不動産市場分析における統計・機械学習の手法

不動産市場分析において、実際に一連の機械学習の技術を適用して、不動産価格を予測しようとした場合には、様々な困難に直面する。不動産価格の予測は、経済市場の中で取引される資産価格や財・サービス市場で取引される財の価格の測定の中でも、最も困難な対象の一つであると言われている。その理由としては、2章で指摘しているように、個別性が強く、たとえ同じ場所に立っていたとしても、建物の材質や建築後年数が異なれば価格は変わり、同じハウスメーカーの住宅であったとしても、異なる場所に立つていれば、周辺の環境や交通利便性に応じて価格が変化してしまう。

不動産市場のような個別性や不均一性が強い市場に対して、第3章で紹介したような線形回帰分析に始まる一連の伝統的な手法を適用し、再現または予測しようとした場合には、市場分割や非線形性の適用など高度な市場分析技術が求められたために、一部の専門家でなければ十分な予測性を保つことができなかつた。そのような中で、ニューラルネットワーク等の新しい技術の登場によって、多くの実務家や不動産市場に精通しない技術者に対して、市場参入の門徒を開くことができたと言っても良いであろう。しかし、実際に不動産データを用いて、不動産価格等を分析しようとすると、様々な困難にぶつかることになる。

本章では、第3章の数理的な背景をもとに、不動産分野で利用される統計・機械学習の手法とその特性について概説する。不動産分野における統計・機械学習の利用可能性は多岐にわたるが、本章では特に、不動産価格関数の推定と予測に用いられる回帰モデルに絞って説明を行う。4.1節では、回帰モデルの仕組みとその推定法に関して説明する。基本的な線形回帰モデルから解説を始め、発展的な手法にまで言及するが、特に線形回帰モデルに関しては基底関数を導入することで、幅広い関数の近似に応用できるように解説した。4.6節では、4.1節で述べた手法を実際の分析に適用する際の注意点・問題点について概説をする。理論的な背景はより専門的な統計の書籍に譲り、本章ではコンピュータによる数値シミュレーションを用いて問題点を具体的に示すことにする。

統計的な検定や推定値の理論的な性質、空間モデルについては取り扱わないので、これらに関する詳細は清水・唐渡(2007)[9]の3・4章を参照されたい。また各モデルとその学習方法に関しては、Bishop

(2012)[1] でより詳細に検討・解説されている。

本章で使用したプログラムのソースコードは、<https://github.com/hayato-n/AsakuraBook> で公開予定である。

4.2 線形回帰モデル

4.2.1 線形回帰

さて、不動産物件のデータが N 件集まっていたとする。また、各物件 $n = 1, \dots, N$ に対して、その価格 y_n と 不動産の特性 x_n が分かっているとしよう。このとき、新しい物件 $*$ の不動産の特性 x_* から、未知の価格 y_* を予測するモデルを構成しよう。このようなモデルを回帰モデルという。

線形回帰モデルの枠組みでは、不動産価格は不動産の特性とその効果量の線形結合で与えられる。すなわち、効果量を β とすると、

$$y_n \approx \beta^T \phi(x_n) \quad (4.1)$$

によって、不動産価格を近似する。ここで ϕ は不動産特性ベクトルを入力としてそれに何らかの変換を施したベクトルを返す関数であり、基底関数と呼ばれる。最も単純な基底関数の一例として、定数項を加えるものが考えられる。また、線形回帰モデルにおいて β は回帰係数とよばれる。

基底関数の導入は、線形回帰モデルの枠組みの中で非線形関数の近似を実現する。例えば、基底関数として多項式変換を考えてみよう。入力としては、1 次元の x を仮定する。このとき D 次の多項式基底関数を以下のように定義する。

$$\phi(x) = (x^0, x^1, \dots, x^{D-1})^T \quad (4.2)$$

ここで $x^0 = 1$ であるから、2 次元の多項式変換は単純に定数項を加える処理に相当する。

図 4.1 は多項式基底関数を用いた非線形関数の近似例である。線形回帰モデルの枠組みの中で、確かに非線形関数の関数を学習できていることがわかる。ただしモデルの推定法によって、学習した関数の形状に多少の差異が見られる。

以降では、線形回帰モデルの推定法について述べる。

4.2.2 最小二乗法

式 (4.1) が精度の良い近似になるように回帰係数 β を学習することを考える。このような、学習される値のことをパラメータとよぶ。パラメータを学習するための 1 つの方法は、「精度の悪さ」を何らかの方法で数値化した損失関数を定義し、それを最小化するようにパラメータを最適化する、というものである。このとき、損失関数はパラメータを入力とする関数である。最もよく使われる損失関数として、実測値と予測値の差の二乗を採用するものがある。具体的には、損失関数として以下のものを採用する。

$$E_{OLS} = \sum_n \varepsilon_n^2 = \sum_n (y_n - \beta^T \phi(x_n))^2 \quad (4.3)$$

ここで ε_n は物件 n の予測誤差である。二乗損失を最小化するようなパラメータの推定法を、最小二乗法 (OLS: Ordinary Least Squares method) と呼ぶ。

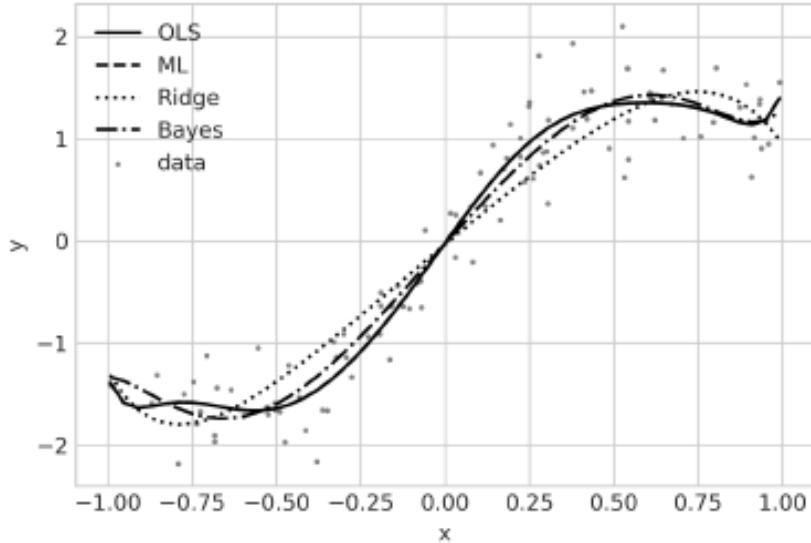


図 4.1 線形回帰モデルによる非線形関数の近似

幸運にも，二乗損失 E_{OLS} を最小化するパラメータ β は解析的に求めることができる。 E_{OLS} を行列を用いて書き直すと，以下のようになる。

$$E_{OLS} = (\mathbf{y} - \Phi\beta)^T(\mathbf{y} - \Phi\beta) \quad (4.4)$$

ただし， $\mathbf{y} = (y_1, \dots, y_N)^T$, $\Phi = (\phi(x_1), \dots, \phi(x_N))^T$ とおいた。 Φ は計画行列とよばれることがある。

このような定式化の下で， E_{OLS} はベクトル β で解析的に微分可能である。ここから最適なパラメータ β が満たすべき条件は，

$$\frac{dE_{OLS}}{d\beta} = -2\Phi^T\mathbf{y} + 2\Phi^T\Phi\beta = 0 \quad (4.5)$$

となる。ここで， $\Phi^T\Phi$ が正則行列（逆行列を持つ行列）であると仮定すれば，

$$\beta = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y} \quad (4.6)$$

が得られる。

4.2.3 最尤推定法

最小二乗法は，二乗損失が最小化されるようにパラメータを調整する方法であった。別のアプローチとして，線形回帰モデルを確率的なモデルと捉える方法がある。具体的には，誤差 ε_n が独立同分布な平均ゼロのガウス分布に従うと仮定する。このとき， \mathbf{y} の $\mathbf{X} = (x_1^T, \dots, x_N^T)^T$ による条件付き分布は

$$\begin{aligned} p(\mathbf{y} | \mathbf{X}) &= \prod_n p(y_n | x_n) \\ &= \prod_n \mathcal{N}(y_n | \beta^T \phi(x_n), \sigma^2) \end{aligned} \quad (4.7)$$

となる。ただし、 $\mathcal{N}(\cdot | \mu, \sigma^2)$ は平均 μ ・分散 σ^2 のガウス分布の確率密度関数を表す。

この条件付き分布をパラメータ $\{\beta, \sigma^2\}$ の関数と見なしたものを、尤度関数 $L(\beta, \sigma^2)$ とよぶ。尤度関数が大きくなるようなパラメータを設定すると、データの条件付き確率が大きくなるので、これを最大化するようにパラメータを調整するという方法が考えられる。これを最尤推定 (MLE: Maximum Likelihood Estimation) とよぶ。

最尤推定を行うとき、通常は尤度関数そのものではなく、その対数 $\ell(\beta, \sigma^2) = \log L(\beta, \sigma^2)$ を使うことが多い。この理由としては、対数尤度の方が尤度そのものよりも数学的に取り扱いやすい場合があること、尤度は個々の確率密度の積であることから、値として極めてゼロに近くなり、コンピュータ上での数値的な誤差が大きくなる場合があること、などが挙げられる。なお、対数は単調増加関数なので、対数尤度の最大化は尤度の最大化と等価である。

さて、線形回帰モデルにおける対数尤度関数の具体的な値を見ていこう。

$$\begin{aligned}\ell(\beta, \sigma^2) &= \sum_n \log p(y_n | \mathbf{x}_n) \\ &= \sum_n \left\{ -\frac{1}{2\sigma^2} (y_n - \beta^T \phi(\mathbf{x}_n))^2 - \frac{1}{2} \log(2\pi\sigma^2) \right\} \\ &= -\frac{\sigma^{-2}}{2} E_{OLS} - \frac{N}{2} \log 2\pi + \frac{N}{2} \log \sigma^{-2}\end{aligned}\tag{4.8}$$

このように、誤差の分布としてガウス分布を採用すれば、対数尤度関数の内部に二乗損失が現れることがわかる。二乗損失 E_{OLS} はパラメータ β のみに依存するので、対数尤度関数が最大化となるときにパラメータが満たすべき条件は以下のようになる。

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma^{-2}} &= -\frac{1}{2} E_{OLS} + \frac{N}{2} \sigma^2 = 0 \\ \frac{\partial \ell}{\partial \beta} &= -\frac{\sigma^{-2}}{2} \frac{dE_{OLS}}{d\beta} = 0\end{aligned}\tag{4.9}$$

よって、 β の最尤推定値は最小二乗推定値と一致し、かつ分散の推定値は $\sigma^2 = \frac{1}{N} E_{OLS}$ となる。

4.2.4 正則化

ここまで最小二乗法と最尤推定法を紹介してきたが、これらはあくまで得られたデータ $\{y_n, \mathbf{x}_n\}_{n=1,\dots,N}$ に対する当てはまりを良くするものであり、新しく得られたデータ $\{y_*, \mathbf{x}_*\}$ に対しても良く当てはまるとは限らない。得られたデータに対してモデルが過剰に適合してしまう状況のことを過学習というが、これについては次節で詳しく確認していく。ここでは過学習を防ぐための手法の1つとして、正則化を紹介する。

正則化の枠組みでは、回帰係数 β が過剰に大きな値をとることに対してペナルティを与える。このペナルティの関数を $E_\beta(\beta)$ とおくと、正則化を行った損失関数 E_r は以下のようになる。

$$E_r = E_{OLS} + \alpha E_\beta(\beta)\tag{4.10}$$

ここで α はペナルティの大きさを決める係数である。

よく使われる正則化の方法として，Ridge 回帰は

$$E_{\beta}(\beta) = \sum_d \beta_d^2 = \beta^T \beta \quad (4.11)$$

LASSO 回帰は

$$E_{\beta}(\beta) = \sum_d |\beta_d| \quad (4.12)$$

と正則化項をおく。ここで， $\beta = (\beta_1, \dots, \beta_D)^T$ とした。

Ridge 回帰の場合，最小二乗法と同様に解析的に β の値を求められる。このときの β は，式 (4.6) 内の $\Phi^T \Phi$ を $\Phi^T \Phi + \alpha \mathbf{I}$ におきかえて，

$$\beta = (\Phi^T \Phi + \alpha \mathbf{I})^{-1} \Phi^T \mathbf{y} \quad (4.13)$$

となる。式 (4.6) の導出では $\Phi^T \Phi$ が正則であることを仮定したが，これが正則でない場合も，Ridge 正則化の下ではこれを正則にできることがわかる。

一方で LASSO は Ridge に比べて求解が難しいものの，回帰係数のうちいくつかのものがゼロとなるように推定されるというメリットがある。

4.2.5 ベイズモデル

最尤推定法の限界として，最尤推定値以外での尤度関数の情報をすべて捨てているという点が挙げられる。例えば，尤度関数が大きくなる推定値が複数あったとする。もし尤度関数がパラメータの「尤もらしさ」を示すものであると解釈するならば，最尤推定法は無数に存在するパラメータの候補の中から「最も尤もらしい」ものだけを採用して，それ以外はすべて捨ててしまう。しかし，尤度関数はパラメータに関する多くの情報を含むので，最尤推定値以外の尤度も全て参照して，モデルを構成する方法があってもよい。

ベイズモデルでは，パラメータに事前分布をおくことで，このようなモデル構成を実現する。例えば，回帰係数 β に事前分布 $p(\beta)$ をおき，分散 σ^2 には事前分布をおかげにこれをハイパーパラメータとしよう。このとき，条件付き確率分布 $p(\mathbf{y} | \mathbf{X}, \sigma^2)$ は，

$$p(\mathbf{y} | \mathbf{X}, \sigma^2) = \int p(\mathbf{y} | \mathbf{X}, \beta, \sigma^2) p(\beta) d\beta \quad (4.14)$$

となる。 β の関数としての尤度関数 $L(\beta) = p(\mathbf{y} | \mathbf{X}, \beta, \sigma^2)$ を考えたとき，この条件付き分布は（事前分布のサポートがある範囲における）あらゆる β に対する尤度関数の情報を参照している。このような方法をベイズ推論という。

条件付き確率分布 $p(\mathbf{y} | \mathbf{X}, \sigma^2)$ のことを周辺尤度という。周辺尤度はハイパーパラメータ σ^2 や事前分布 $p(\beta)$ を決定するハイパーパラメータ^{*1}の関数としても見なせるので，周辺尤度を最大化するようにハイパーパラメータを選択することもできる。このような方法を第二種の最尤推定法や経験ベイズ法とよ

^{*1} ここでは明示していないが，例えば事前分布がガウス分布で与えられているとするならば，その平均や分散がハイパーパラメータとなる

ぶことがある。しかし完全にベイズ的な枠組みでは、ハイパーパラメータに対してさらに事前分布を置いて推論を行う。

回帰係数 β の値を推論したい場合は、ベイズの定理を用いることによって、形式的には以下のようにパラメータの事後分布が得られる。

$$\begin{aligned} p(\beta | \mathbf{y}, \mathbf{X}, \sigma^2) &= \frac{p(\mathbf{y} | \mathbf{X}, \beta, \sigma^2)p(\beta)}{p(\mathbf{y} | \mathbf{X}, \sigma^2)} \\ &= \frac{p(\mathbf{y} | \mathbf{X}, \beta, \sigma^2)p(\beta)}{\int p(\mathbf{y} | \mathbf{X}, \beta, \sigma^2)p(\beta)d\beta} \end{aligned} \quad (4.15)$$

これを用いれば、新規データに関する不動産特性 x_* が得られたときの価格 y_* の予測分布は以下のように構成できる。

$$p(y_* | x_*, \mathbf{y}, \mathbf{X}, \sigma^2) = \int p(y_* | \beta, x_*, \sigma^2)p(\beta | \mathbf{y}, \mathbf{X}, \sigma^2)d\beta \quad (4.16)$$

ベイズ法ではこれまで見てきた手法とは異なり、 β の推定値はある 1 つの値としてではなく、確率分布として与えられる。もし β を点推定したい場合、いくつかの手法が考えられるものの、よく用いられるのは MAP(Maximum a posterior) 推定法である。MAP 推定ではその名の通り、事後確率を最大化する値をパラメータの推定値として採用する。すなわち、回帰係数の MAP 推定値 β_{MAP} は、

$$\begin{aligned} \beta_{MAP} &= \arg \max_{\beta} p(\beta | \mathbf{y}, \mathbf{X}, \sigma^2) \\ &= \arg \max_{\beta} \log p(\beta | \mathbf{y}, \mathbf{X}, \sigma^2) \\ &= \arg \max_{\beta} \{\log p(\mathbf{y} | \mathbf{X}, \beta, \sigma^2) + \log p(\beta)\} \end{aligned} \quad (4.17)$$

で与えられる。なお、式 (4.17)において、事前分布 $p(\beta)$ を平均ゼロのガウス分布とするならば、MAP 推定が Ridge 回帰と一致することが確認できる^{*2}。

4.3 分位点回帰

4.3.1 分位点

線形回帰モデルにおける最尤推定法の議論から分かるように、前述した線形回帰モデルは不動産価格の期待値（平均値）に着目したモデルである。しかし、関心がある統計量は必ずしも期待値であるとは限らず、価格分布上の他の値を知りたい場合もあるだろう。このような要求に応える手法として、分位点回帰法 (Koenker & Bassett 1978[3]) を紹介する。

図 4.2 は分位点回帰法を用いて推定した条件付き分位点の例である。確かに 2.5% 点と 97.5% 点の間にほとんどのデータが収まっていることが確認でき、分位点回帰による分布の予測が実現されていることがわかる。

^{*2} このことは Ridge 回帰における係数 α を経験ベイズ法で選択できることを示唆するが、 β を周辺化することが前提となっている

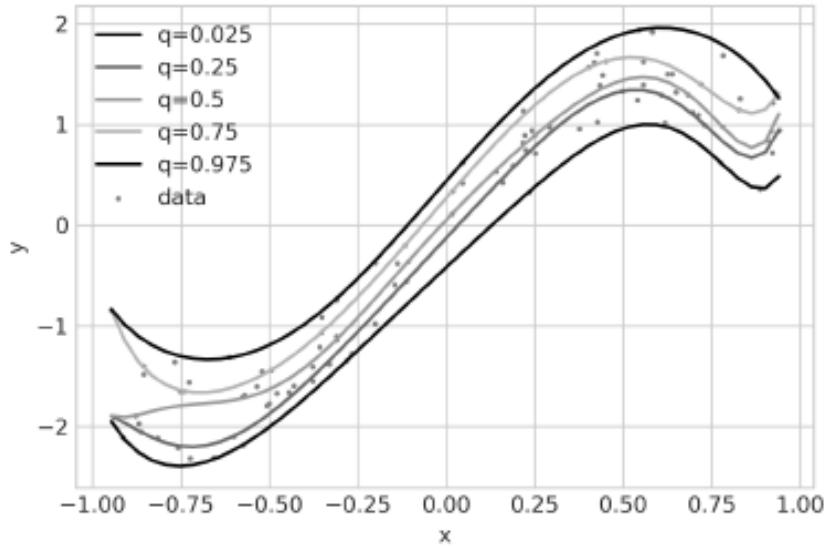


図 4.2 分位点回帰

分位点回帰を理解するために、まずは分位点 (Quantile) について概説しよう。 A なる事象が起こる確率を $P(A)$ と表すとする。 q 分位点 ($0 < q < 1$) とは、单变量の分布 $p(x)$ について、 $P(x \leq \theta_q)$ と $P(\theta_q < x)$ が $q : 1 - q$ となるように分布 $p(x)$ を分割する点 θ_q のことである。特に $q = 0.5 = 50\%$ のとき、この点を中心値 (median) とよぶ。

もしデータ $x = (x_1, \dots, x_n)^T$ が与えられているならば、このデータに対する θ_q の推定値は、損失関数

$$\sum_{n \in \{n: \theta_q \leq x_n\}} q|x_n - \theta_q| + \sum_{n \in \{n: x_n < \theta_q\}} (1 - q)|x_n - \theta_q| \quad (4.18)$$

を最小化する θ_q として与えられる。もしあらゆる q に対応する θ_q が求められているなら、これらの分位点から分布全体の形状がわかることになる。

4.3.2 分位点回帰

分位点回帰 (QR: Quantile Regression) では、不動産価格の条件付き分布 $p(y_n | x_n)$ を考える。このとき、条件付き q 分布点 $\theta_q | x_n$ が線形モデル $\theta_q | x_n = \beta_q^T \phi(x_n)$ で与えられるとすると、 β_q は損失関数

$$E_{QR} = \sum_{n \in \{n: \beta_q^T \phi(x_n) \leq y_n\}} q |y_n - \beta_q^T \phi(x_n)| + \sum_{n \in \{n: y_n < \beta_q^T \phi(x_n)\}} (1 - q) |y_n - \beta_q^T \phi(x_n)| \quad (4.19)$$

を最小化する β_q である。分位点のときと同様に、もしあらゆる q に対応する β_q が求められているなら、条件付き分布の形状全体がわかることになる。つまり、ある不動産の特性 x_* を持つ不動産の集合の価格が、どのような分布に従っているかが予測できる。

$q = 0.5$ としたとき、分位点回帰モデルは中央値を予測する。このモデルは通常の線形回帰モデルと類似しているように思われるが、重要な特長として、外れ値に対してより頑健であることが挙げられる。

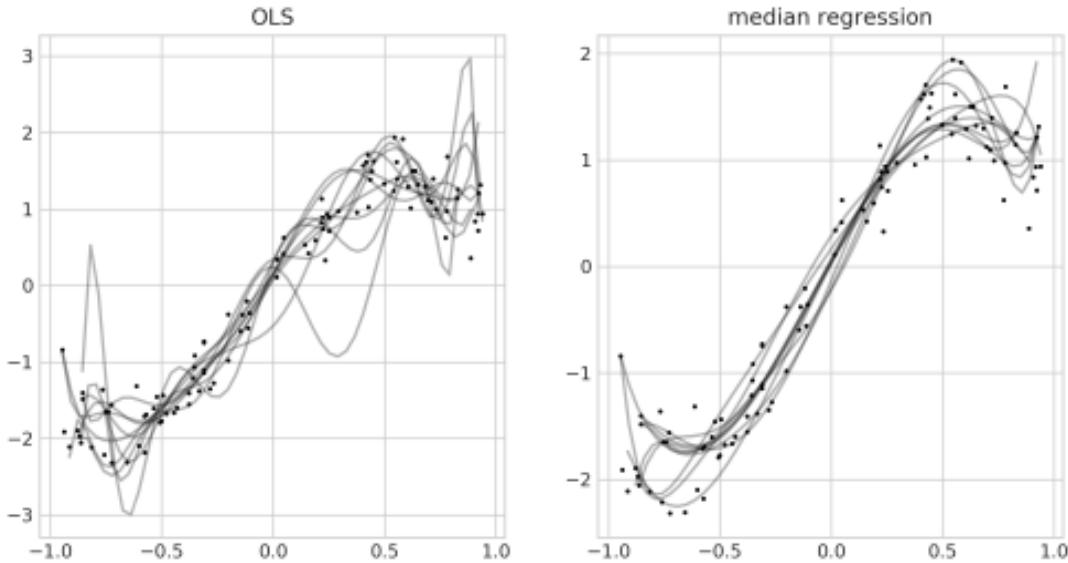


図 4.3 最小二乗法と中央値回帰

図 4.3 は最小二乗法による線形回帰モデルと 中央値回帰モデルを比較したものである。この図は 100 個のサンプルからランダムに 30 個のサンプルを選んで、それらを用いて回帰モデルを学習している。このとき外れ値として 1 つのサンプルにノイズを加えた。このような試験を繰り返し行うこと で、モデルが学習する回帰曲線が安定しているかを確認できる。図を見ると、最小二乗法に比べて中央値回帰の学習した曲線は蛇行が少なく、安定していることが読み取れる。このようなシミュレーションからも、中央値回帰の安定性が確認できる。

4.4 ニューラルネットワーク

4.4.1 基底関数の学習

線形回帰モデルでは、固定された基底関数を用いて回帰を行った。この自然な拡張として、基底関数の形状自体をデータから学習したいという要求が考えられる。ニューラルネットワークはこうした基底関数の学習を実現する手法の 1 つである。ニューラルネットワークのパラメータ推定は解析的にはできないが、その再帰的な構造を利用すれば、パラメータによる損失関数の勾配が効率的かつ厳密に計算できる。そのため、勾配法を用いた最適化がパラメータ推定に広く用いられている。また、確率的勾配降下法を導入すれば、大規模データに適した推定アルゴリズムが得られる。

図 4.4 はニューラルネットワークを用いて大規模データから学習した近似関数の例である。ここでは 10,000 個のサンプルからモデルを学習している。例では関数の形状があまり複雑ではないのでニューラルネットワークの恩恵はあまり感じられないかもしれないが、複雑な関数であっても大規模データの力をすれば近似できてしまうのが、ニューラルネットワークの強みである。

ニューラルネットワークは高い表現力を持つ反面、その学習は難しい。そのため、数多くの学習法やテ

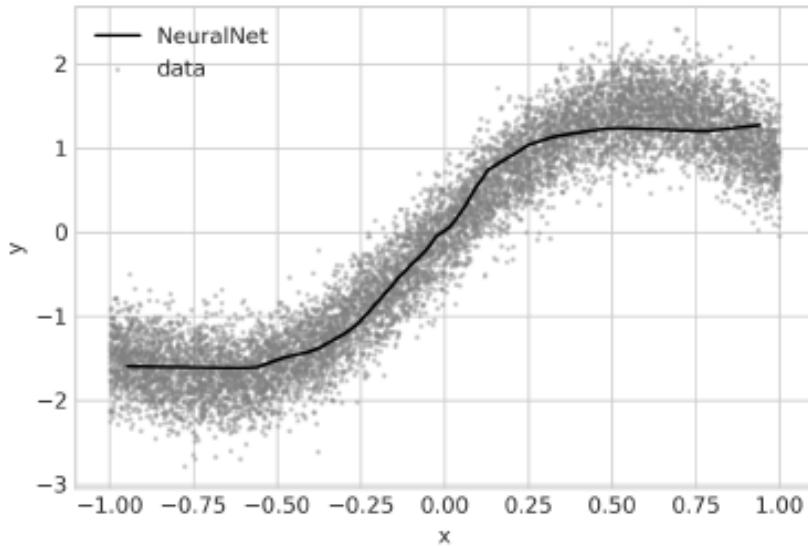


図 4.4 ニューラルネットワーク

クニックが日々開発されている。こうした研究をすべて追うことは困難であるが、基本的な部分については Bishop (2012)[1] や斎藤 (2016)[4] を参照するとよい。

線形回帰モデルによる予測 $\beta^T \phi(x_n) = t_{D,1}$ を下のように書き直す。

$$t_{D,1} = \sum_j w_{D-1,1,j} h_{D-1,j} + b_{D-1,1} \quad (4.20)$$

ここで $h_{D-1,j}$ は x_n の関数（パラメータを持つ基底関数による変換）である。この予測も $h_{D-1,j}$ に関する線形システムであるから、ニューラルネットワークは線形回帰モデルの拡張と見なせる。

ここで、 $h_{d,i} (d = 0, \dots, D-1)$ は以下のように再帰的に定義される。

$$\begin{aligned} h_{d,i} &= \sigma(t_{d,i}) \\ t_{d,i} &= \sum_j w_{d-1,i,j} h_{d-1,j} + b_{d-1,i} \end{aligned} \quad (4.21)$$

ただし $h_0 = (h_{0,1}, h_{0,2}, \dots)^T = x_n$ であり、 σ は $t_{d,i}$ に非線形変換を施す活性化関数である^{*3}。再帰的な定義から、 D はニューラルネットワークの“層”的深さであると解釈できる。

層の深さと各層の幅は任意であり、特に深いネットワークを深層学習（Deep Learning）と呼ぶことがある。ここでは全結合のネットワーク構造のみを紹介したが、タスクに応じてネットワーク構造を工夫することも可能である。例えば、近年画像認識で目覚ましい成果を上げているネットワーク群は、その構造から畳み込みニューラルネットワークとよばれている。

4.4.2 誤差逆伝播

^{*3} 活性化関数は層ごとに異なっていても良い

ニューラルネットワークのパラメータ $\{w_{d,i,j}, b_{d,i}\}$ は非常に数が多く、解析的に最適化することはできない。このため、勾配法（勾配降下法）を用いた数値的な最適化を行う。勾配法を簡単に説明しよう。損失関数を最小化するようなプロセスを考えたとき、もし損失関数をパラメータで微分した勾配が与えられていれば、どの方向にパラメータを動かせば損失が減少するかがわかる。よって、この勾配情報を使ってパラメータを少しづつ動かし、損失関数を最小化するパラメータを探索するのが勾配法である。

勾配法の適用には損失関数の勾配が要求されるが、ニューラルネットワークの場合、その再帰的な構造を用いることで、勾配を効率的に（しかも厳密に）求めることができる。以下で説明するこのような勾配計算のプロセスは、誤差逆伝播（Error Backpropagation）とよばれる。

回帰モデルの場合、ニューラルネットワークの損失関数を二乗損失 E_{OLS} とすれば、

$$E_{OLS} = \sum_n (y_n - t_{D,1}(\mathbf{x}_n))^2 = \sum_n E_n \quad (4.22)$$

と書ける。ここで $t_{D,1}$ は x_n の関数であり、ニューラルネットワークの任意のパラメータを θ とすれば、

$$\frac{\partial E_{OLS}}{\partial \theta} = \sum_n \frac{\partial E_n}{\partial \theta} \quad (4.23)$$

である。よって、各データに対して $\frac{\partial E_n}{\partial \theta}$ を計算して、それをすべて足せば損失関数の勾配が求められる。

パラメータ $\{w_{d,i,j}, b_{d,i}\}$ による偏微分は、以下のような再帰的な手続きによって求められる。

- [1] まず、最終層のパラメータ $\{w_{D-1,1,j}, b_{D-1,1}\}$ による偏微分を求める。予測値 $t_{D,1}$ による偏微分は

$$\frac{\partial E_n}{\partial t_{D,1}} = -2(y_n - t_{D,1}) \quad (4.24)$$

であるが、これと微分の連鎖律（Chain Rule）を用いれば、パラメータによる偏微分が以下のように求められる。

$$\frac{\partial E_n}{\partial w_{D-1,1,j}} = \frac{\partial E_n}{\partial t_{D,1}} \frac{\partial t_{D,1}}{\partial w_{D-1,1,j}}, \quad \frac{\partial t_{D,1}}{\partial w_{D-1,1,j}} = h_{D-1,j} \quad (4.25)$$

$$\frac{\partial E_n}{\partial b_{D-1,1}} = \frac{\partial E_n}{\partial t_{D,1}} \frac{\partial t_{D,1}}{\partial b_{D-1,1}}, \quad \frac{\partial t_{D,1}}{\partial b_{D-1,1}} = 1 \quad (4.26)$$

また、次の偏微分を [2] で使用する。

$$\frac{\partial E_n}{\partial h_{D-1,j}} = \frac{\partial E_n}{\partial t_{D,1}} \frac{\partial t_{D,1}}{\partial h_{D-1,j}}, \quad \frac{\partial t_{D,1}}{\partial h_{D-1,j}} = w_{D-1,1,j} \quad (4.27)$$

- [2] $\frac{\partial E_n}{\partial h_{d,i}}$ が分かっているとき、以下の偏微分が求められる。これにより、再帰的にすべてのパラメー

々で偏微分が実行できる。

$$\frac{\partial E_n}{\partial t_{d,i}} = \frac{\partial E_n}{\partial h_{d,i}} \frac{\partial h_{d,i}}{\partial t_{d,i}}, \quad \frac{\partial h_{d,i}}{\partial t_{d,i}} = \frac{\partial}{\partial t_{d,i}} \sigma(t_{d,i}) \quad (4.28)$$

$$\frac{\partial E_n}{\partial w_{d-1,i,j}} = \frac{\partial E_n}{\partial t_{d,i}} \frac{\partial t_{d,i}}{\partial w_{d-1,i,j}}, \quad \frac{\partial t_{d,i}}{\partial w_{d-1,i,j}} = h_{d-1,j} \quad (4.29)$$

$$\frac{\partial E_n}{\partial b_{d-1,i}} = \frac{\partial E_n}{\partial t_{d,i}} \frac{\partial t_{d,i}}{\partial b_{d-1,i}}, \quad \frac{\partial t_{d,i}}{\partial b_{d-1,i}} = 1 \quad (4.30)$$

$$\frac{\partial E_n}{\partial h_{d-1,j}} = \sum_i \frac{\partial E_n}{\partial t_{d,i}} \frac{\partial t_{d,i}}{\partial h_{d-1,j}}, \quad \frac{\partial t_{d,i}}{\partial h_{d-1,j}} = w_{d-1,i,j} \quad (4.31)$$

以上から、解析的に微分可能な活性化関数を採用すれば、ニューラルネットワークの偏微分は誤差逆伝播の手続きによって解析的に求められることがわかる。

4.5 その他の手法

大規模データを用いて学習したいとき、勾配 $\frac{\partial E_n}{\partial \theta}$ をすべてのデータについて計算するのは計算コストが大きい。そこで、式 (4.23) を確率的に近似推定する方法を考える。

もし N 個のデータの中から偏りなく 1 つのデータをランダムサンプリングしたとすると、 $N \frac{\partial E_n}{\partial \theta}$ を用いて式 (4.23) の勾配を推定できる。これを用いた勾配降下法を、確率的勾配降下法とよぶ。複数のサンプルを抽出すれば、勾配の推定はより安定する。

確率的勾配降下法の他の特色として、勾配の推定量に常にノイズが入るために、通常の勾配法とは異なり損失関数が常には減少しない点が挙げられる。これにより、浅い局所最適解を回避できるとされている。推定された勾配を使用した具体的なパラメータの更新方法は、様々なものが提案されている。

他の重要な手法として、アンサンブル学習が挙げられる。アンサンブル学習では、多数の回帰モデルを組み合わせることによって、予測の安定化と精度向上を図る。代表的なモデルとして、ランダムフォレスト (Breiman 2001)[2] などが挙げられる。アンサンブル学習では決定木ベースのモデルがよく用いられる。それらの手法については、第 6 章を参照されたい。

4.6 手法の適用

4.6.1 過少定式化バイアス

回帰モデルの典型的な応用例として、各不動産の特性が不動産価格に与える影響を検討するというものが挙げられる^{*4}。例えば、線形回帰モデルで不動産価格関数が充分に近似できたとすると、 d 番目の特性 $\phi_d(x)$ の効果量は β_d なので、特性 $\phi_d(x)$ が 1 単位増加すると β_d だけ価格が上昇すると予想される。すなわち、推定された回帰係数 β から各特性が不動産価格に与える影響を検討できる。

しかしながら、不動産価格に影響する特性をすべてデータとして取得するのは、現実には困難である。この場合、仮に不動産価格が本当に線形回帰モデルから生成されている場合であっても、入手できなかつ

^{*4} 通常の回帰モデルの枠組みの中では、変数間の因果関係に関する知見は本来得られない。ここで「影響」といっているのは、あくまで不動産価格がその特性によって決定されるという仮定のもとで導かれるものである。

た特性の存在によって、線形回帰モデルの回帰係数の推定量にバイアスが生じる可能性がある。これを過少定式化バイアスとよぶ。入手できなかった特性が価格に強く影響している場合や、また入手できなかった特性と入手できた特性の間に強い相関がある場合、過少定式化バイアスが発生する懸念は強くなる。

過少定式化バイアスの発生を、数値的なシミュレーションを行って確認してみよう。不動産特性と価格を以下のように生成する。まず2つの特性を、これらが相関を持つように疑似乱数で生成する。その後、適当に生成した回帰係数を持つ線形回帰モデルを用いて、これらの特性から不動産価格を生成する。このようにすれば、真の回帰係数が分かっているから、線形回帰モデルを推定したときの回帰係数の推定誤差を検証できる。

図4.5は500個のサンプルを用いて10,000回推定を行い、観測された推定バイアスをヒストグラムに整理したものである。ここでは、価格に影響する不動産特性は両方分かっていると仮定している。図から分かるように、推定誤差は0を中心として分布することがわかる。このように完全なデータを用いれば、平均的には正しく回帰係数が推定できることがわかる^{*5}。

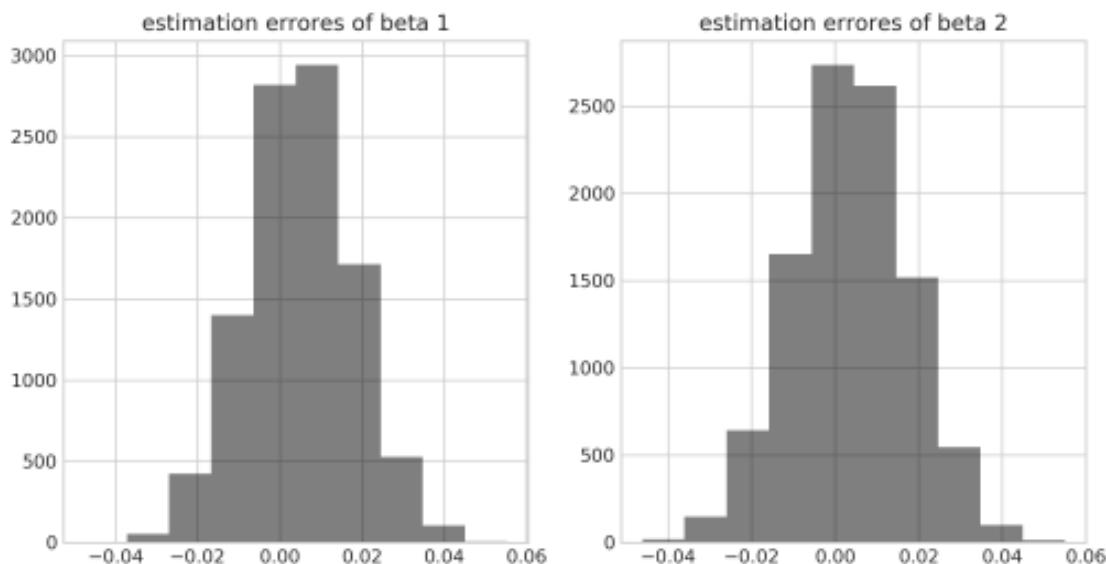


図4.5 欠測のないデータを使用したときの推定誤差

一方で図4.6は、一方の特性の回帰係数を推定する際には、もう一方の特性が分からないという仮定のもので推定した結果である。つまり、過少定式化バイアスが発生するような状況をシミュレーションしている。この場合は、完全なデータを用いたときと比べて推定誤差のはらつきが大きいだけでなく、分布の中心も0になっていないことがわかる。つまり、線形回帰モデルを推定したとき、平均的に見ても回帰係数の推定に誤差が生じることが期待される。

実際の不動産分析においては、以下のような状況が過少定式化バイアスの例として想定しうる。まず自然公園などの公共空間に近いと、不動産価格が少し上昇するとしよう。このとき、自然公園までの距

^{*5} ただし、ここでは価格を生成する真の構造が線形回帰モデルであると分かっていることが仮定されていることに注意されたい。実際の不動産価格は線形回帰モデルから生成されているわけではないので、実際に推定される価格関数はあくまで近似に過ぎない。

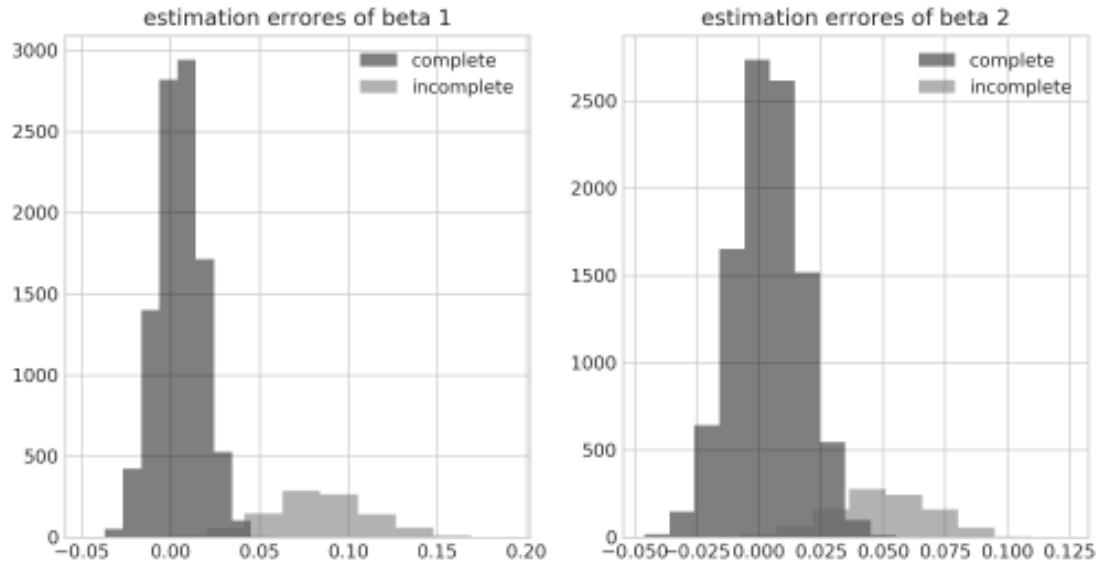


図 4.6 欠測があるデータを使用したときの推定誤差

離は不動産価格に負の効果を持つ（近いほど価格が上がる）。一方で、駅までの距離は不動産価格により大きな影響があると考えられる。しかし分析者の手違いで、回帰モデルに駅までの距離を含めるのを忘れてしまったとする。自然公園が駅の近くに立地していることは稀なので、自然公園までの距離と駅までの距離は負の相関を持つ。よって、このとき自然公園までの距離の小ささは駅までの距離の大きさとして解釈され、結果として自然公園までの距離が不動産価格に正の効果があるかのように推定されてしまう（つまり、遠いほど価格が上がると推定される）。これが過少定式化バイアスである。

このような過少定式化バイアスを防ぐためには、価格に強く影響していると思われる不動産特性をあらかじめ見極め、それらを適切にモデルに組み込むことが求められる。

4.6.2 過学習

近年は機械学習とよばれる非常に柔軟かつ複雑なモデル群が、その予測精度の高さを喧伝されている。本章で紹介した中では、ニューラルネットワークがそれに該当する。これらの手法は様々な関数を近似することができるが、その近似能力の高さは過学習とよばれる問題も引き起こす。

回帰モデル $f(x)$ の学習は、データ $\{y_n, x_n\}_{n=1,\dots,N}$ を用いて損失 E を計算し、それを最小化することで達成される。二乗損失などの通常の損失関数であれば、全体の損失は不動産ごとの予測損失 E_n の和で表され、かつそれは実際の価格 y_n と予測価格 $f(x_n)$ によって決定される。つまり、 $E = \sum_n E_n(y_n, f(x_n))$ である。ここから、学習済みのモデルは学習に使用したデータに関して、平均的な予測損失を最小化していると考えられる。

ここで、もしサンプルサイズに比べて圧倒的に複雑なモデルを採用したとする。このようにモデルが過剰に柔軟な場合、得られたデータをすべて“丸暗記”できてしまう可能性がある。もしデータを丸暗記したならば、データに対する予測損失は 0 となる。

しかしながら、モデルの学習における目的の1つは、いまだ得られていないデータ $\{y_*, x_*\}$ に関して、良い予測をすることである。すなわち、新規データに関する予測損失 $E_*(y_*, x_*)$ を（平均的に）最小化するようにモデルを学習したい。上記で示したような“丸暗記”したモデルは、この目的に適しているとは限らない。というのも、現実の不動産価格予測は暗記テストではないので、丸暗記したモデルの予測はむしろ大きく外れるということがありうるからである。

このような状態を、過学習とよぶ。

図4.7は、線形回帰モデルの学習において、過剰に高次の多項式基底関数を採用した場合の学習例である。この例では回帰曲線がデータの近傍を通るために、過剰に蛇行している。その結果、新しくデータが得られた際におよそあり得なさそうな価格をモデルが予測してしまっていることがわかる。

これは過学習の典型的な例である。

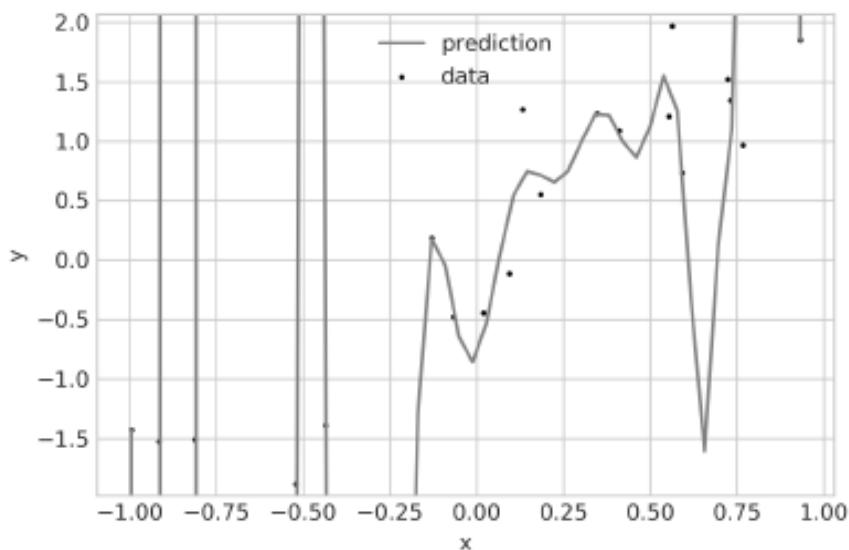


図4.7 過学習の例

過学習を避ける方法としては、データを学習用と検証用に分割する方法が考えられる。つまり、学習用データで学習したモデルの精度を検証用データを用いて確認するのである。また、Ridge回帰のような正則化も有効である。線形回帰モデルのような情報量規準が容易に計算できるモデルならば、それを用いて変数選択を行うこともできる。例えば今回の場合であれば、変数選択によって多項式基底関数の次数を適切に削減することで、過学習を回避できる。

4.6.3 バイアスとバリアンス

ここでは過学習が起こる原因について、予測のバイアスとバリアンスの観点から、シミュレーションを用いた直感的な説明を通して整理する。この議論に関する数学的な詳細は Bishop (2012)[1] を参照されたい。

もしこの世に存在するすべての不動産のデータを持っていれば、（このような条件で価格予測が必要

なのかという疑問はあるものの）予測誤差が最も小さくなるような理想的な関数を作ることができそうだ。しかし実際に入手可能なデータは、世の中に存在するうちの一部の不動産に関するものだけなので、このような一部のデータから（データとして取得できなかつた不動産の価格についても）予測誤差の小さい関数を学習したい。

注意すべきなのは、この世に存在するすべての不動産から一部の不動産のデータを得る際に、どの不動産のデータが得られるか、という点については偶発的な要因が絡むという点である。

あり得ない仮定であるが、データの取得とモデルの学習を何度も繰り返し行うことができるでしょう。データの取得には偶発的な要因が絡むので、そのとき「たまたま」得られたデータに依存して、そのとき学習されたモデルの関数形も異なってくると予想される。

図 4.8 はこのような仮定をシミュレーションを用いて実践した結果である。ここでは 100 個のデータから 30 個のサンプルをランダムにとって、それを用いてモデルを学習する、という手続きを 10 回繰り返している。ここでは多項式基底関数による線形回帰モデルを最小二乗法で学習した。多項式の次数は、1, 3, 10, 100 の 4 通りで試している。

この結果を見ると、以下の興味深い知見が得られる。まず明確にわかるることは、推定される曲線の形状が次数が高くなるほどにはばらついているということである。このようなばらつきの大きさは、データのとり方が学習される回帰モデルの形状に大きく影響するということを示唆しており、望ましくない。このようなばらつきのことを、予測のバリアンスという。これは言い換れば、偶然得られたデータに回帰モデルが必要以上に適合してしまっているということであるので、バリアンスの大きさは過学習のしやすさと強い関係があるといえる。

一方でよく見ると、高次のモデルであっても多数ある曲線の平均をとれば、データ全体をうまく予測できていそうに見える。このように、多数のモデルを学習できたときに、それらの予測を平均値を全体としての予測として取り扱うことを考えよう。このような平均値による予測の悪さを、予測のバイアスという。高次のモデルはバリアンスが大きいので予測が悪く見えるが、バイアスの観点で見ると、実は予測が良い。一方で次数が低すぎるモデル（ここでは次数 1 のモデル）は実際の価格形成の構造に対してモデルの柔軟性が足りていないので、どの回帰モデルの予測もさほど高くなっていない。この結果、これらの平均をとっても予測は悪く、バイアスが大きい。

この例から分かるように、バイアスとバリアンスはトレードオフの関係にある。「データの取得」という偶発的な要因に影響されずに良い予測モデルを作るためには、バイアスもバリアンスも適度に小さくなるような、ちょうどいい柔軟性を持つモデルを選ぶことが重要であることがわかる。

高いバリアンスは過学習の原因があるので、これを適度に抑制する方法が望まれる。すでに紹介したモデルの正則化は、バリアンスを低減するのに有効な手法の 1 つである。図 4.9 は推定法を最小二乗法から Ridge 回帰に変えて、全く同じシミュレーションを行った結果である。ここでは正則化の強さは緩めに設定しているが、バリアンスが大幅に低減され、100 次元のモデルでも関数が安定していることがわかる。

しかしながら、バリアンスの過剰な抑制はバイアスの増大を招くので、正則化を行う場合でもその強さは適切に調整する必要がある。

今回のシミュレーションでは「データの取得」というプロセスを繰り返し実行できたので、図からどの程度バリアンスを抑制すればよいのか判断することができた。しかし実際には、データの取得は一度しか

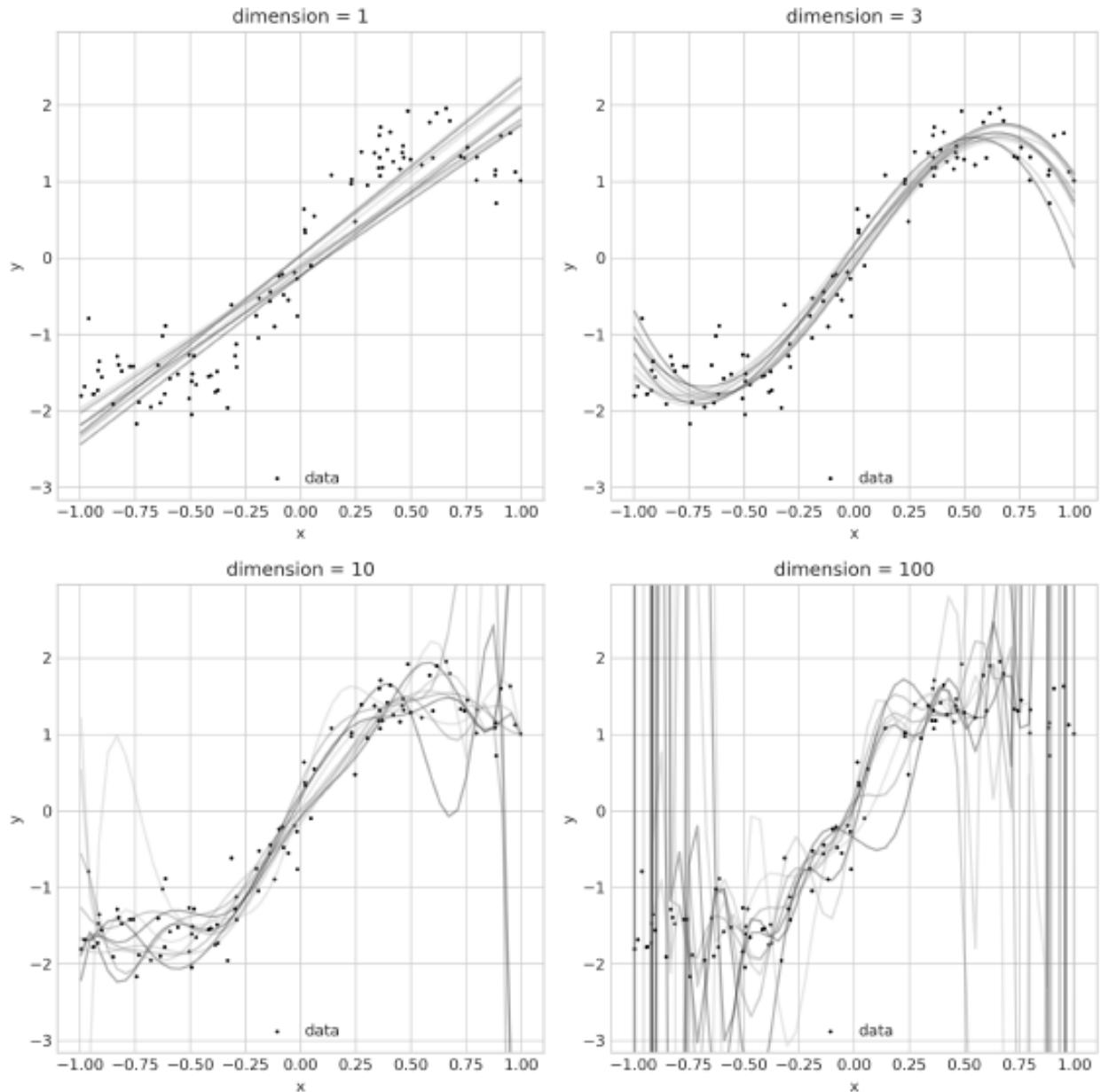


図 4.8 学習されたモデルの偶発的なばらつき

できないことが多いので、与えられた 1 セットのデータから適度な複雑さを持ったモデルを選んだり、正則化の強さを適切に調整したりしなければならない。すでに述べたように、データを学習用と検証用に分割する方法は、この課題を解決する 1 つの方法である。しかしデータが限られている場合は、手持ちのデータすべてを使ってモデルを学習させたい場合もあるだろう。

ここまで議論を踏まえると、より良い予測モデルを作るためのいくつかの方法を新たに考えることができる。

まず、バイアスのみを見れば、複雑なモデルの予測が良さそうであることに着目しよう。複雑なモデル

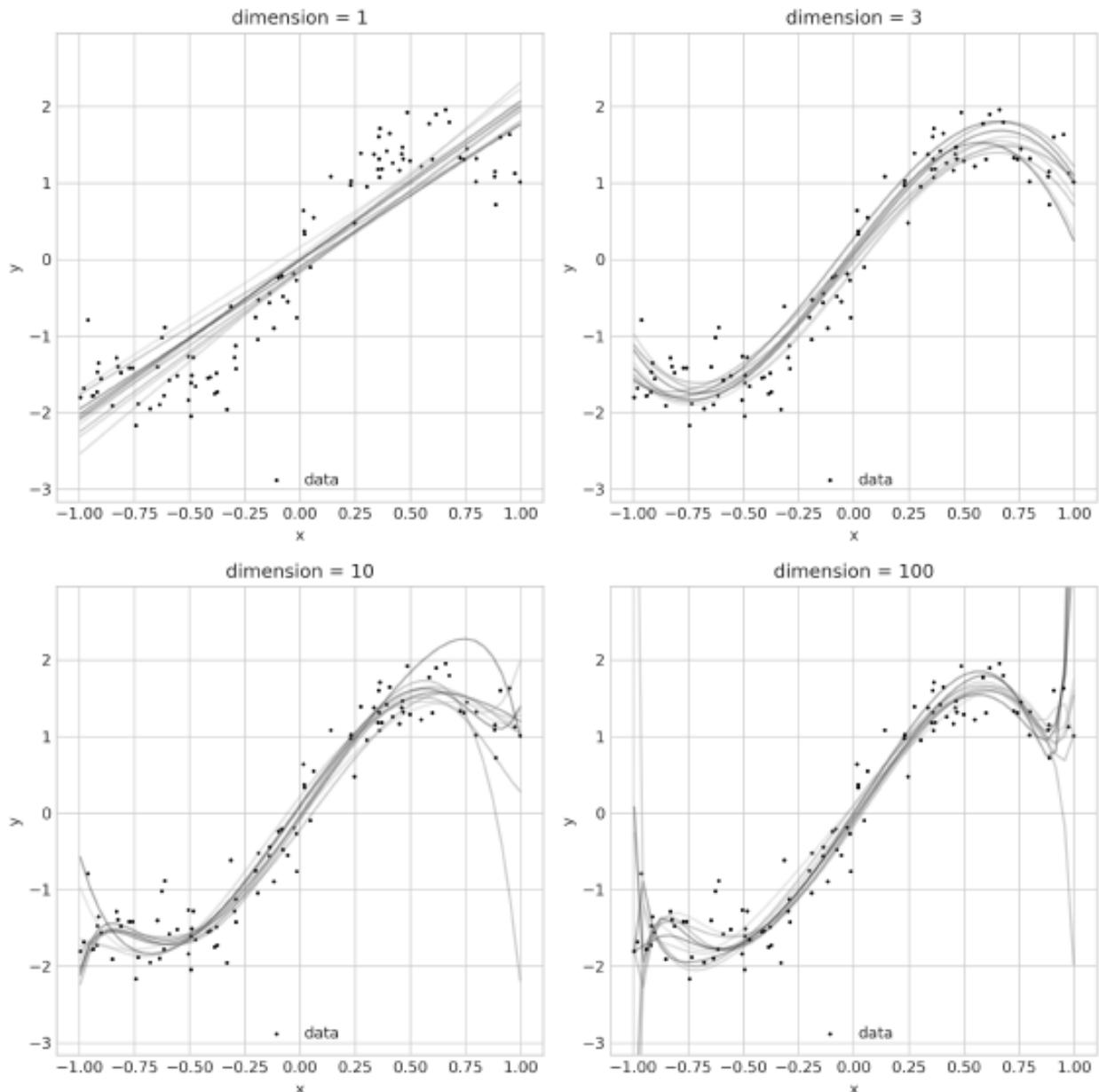


図 4.9 正則化によるバリアンスの低減

の問題点はバリアンスが大きいことであるが、バイアスは小さいので、多数のモデルの平均をとれば、良い予測ができるようである。アンサンブル学習はこのようなアイデアを実現する方法である。具体的な手順の一例としては、得られたデータからランダムにサンプルを抽出し、モデルに学習させる。これを何度も繰り返すと多数のモデルが得られる。これらのモデル群すべてに予測を行わせ、その平均を全体の予測とするのである。

もう 1 つの方法として、予測のバリアンスが大きいことは、推定されたパラメータがばらついていることに由来している、という点に着目することが考えられる。多数のモデルを平均化することは、

様々なパラメータについてモデルの予測を構築してその平均をとっているのであるから，究極的にはありとあらゆるパラメータを使って予測を行い，それを平均化すればよい。ここでベイズ推論の式を見ると，ベイズ予測では不動産価格の条件付き確率に事前分布や事後分布をかけて積分しているので，これはあらゆるパラメータを用いた予測を事前分布や事後分布で重みづけして平均化していることに相当することがわかる。

参考文献

- [1] Bishop, C. (2012), 『パターン認識と機械学習上：ベイズ理論による統計的予測』, 丸善.
- [2] Breiman, L. (2001), ‘Random forests’, *Machine Learning* 45(1), 5-32.
- [3] Koenker, R. & Bassett, G. (1978), ‘Regression Quantiles’, *Econometrica* 46(1), 33.
- [4] 斎藤康毅 (2016), 『ゼロから作る Deep Learning Python で学ぶディープラーニングの理論と実装』, オライリー・ジャパン.
- [5] 清水千弘・唐渡広志 (2007), シリーズ：応用ファイナンス講座 4 『不動産市場の計量経済分析』, 朝倉書店.

第5章

不動産市場への機械学習の適用^{*1}

5.1 不動産市場分析の実際

実際に不動産市場を対象として、データを収集し分析しようとした場合には、どのようにデータ資源にアクセスしたらいいのか、入手できた不動産情報をどのように分析用のデータへと加工したらいいのか、そして、どのようにモデルを設定し、機械学習の各種手法を応用したらいいのかといった問題に直面する。実は、不動産市場分析に限らず、機械学習の各種手法を固有分野に応用しようとすると、適切なデータ資源の選択やその収集 (Shimizu, Nishimura and Watanabe(2016)[17])、分析用データへの変換といったことは、最も大切な技術であり、現在の科学技術の進歩をもってしても依然として解決ができない問題なのである。ここに、分析者の経験や専門性が強く差別化されることになる。本章では、第3章および第4章で整理された機械学習の各種手法を、不動産価格データへの応用していく手順や推計例を紹介する。

不動産市場を対象として、統計的な分析手法を適用しようとした場合には、不動産価格や家賃を予測する、または不動産の市場価値を差別化している要素を取り出し、その効果の大きさを測定するといったことなどに活用される。予測については容易に想像ができる。過去から現在にかけて収集されたデータを用いて、不動産の価格形成構造を再現し、個別の不動産の価格や家賃を予測することで、市場で決定されるであろう市場価格を予測するという技術である。不動産市場に参加している専門家や売り手や買い手の消費者は、常に適正な価格を知りたいという思うものが多いために、機械学習のこのような領域への応用は、内外を問わず積極的に行われている。

後者についてもまた、市場参加者または政策当局ともに関心が高い分野である。新しい性能を不動産に追加した場合、どの程度の価格の差別化が行われるのか、つまり高く売れるのか、または、一年経過するたびに不動産の価値の減耗はどの程度であるのかといったことは、民間市場の参加者だけでなく、政策当局も高い関心を寄せるところである。適用事例を見れば、長期優良住宅であればどの程度の価格プレミアムが付くのか、新耐震と旧耐震でどの程度の価格差が生まれているのか、オートロックとそうでない建物で家賃は違うのか、断熱性能を高めたら高い家賃が取れるのか、といった問題に対して、不動産データを

^{*1} 本章は、高橋祐介・清水千弘 (2020)、「不動産市場データを用いた機械学習の評価」mimeo[18] , Deng, Y., J. Onishi , C. Shimizu and S.Zheng (2018)[2], “The Economic Value of Environmental Consideration in the Tokyo Office Market,”CSIS Discussion paper 155, The University of Tokyo. を要約したものである。

分析して、その効果を抽出している事例が散見される。このような問題は、介入の因果効果の推定問題であり、この問題を取り扱うのが統計的因果推論である。このような応用をしようとすると、解釈性の高いモデルでの推定が求められることから、交絡因子による影響の除去が重要な課題となってきた。

Fisher が提唱した実験計画法は、ランダムな処置の割り当てによって処理群と潜在的に等しい属性をもつ対照群を定義することで、交絡因子の影響を排除した。しかし、社会科学が扱うほぼ全てのデータは観察データであるため、実験計画法によって交絡因子の影響を排除することはできない。そのような中で、第 2 章で整理したヘドニック・アプローチは有力な手法として活用されてきているが、単純にデータを収集して、学習させるだけでは、このような課題には対応ができないために、予測問題とは異なる知識が必要となるのである。

本章では、「不動産市場への機械学習の適用」として、不動産価格を予測する事例と併せて、近年において注目されている環境性能の高い不動産とそうでない不動産との価格がどの程度発生しているのかといった事例をもとに、どのような点に注意していくかなければならないのかを示すことを目的とする。

5.2 予測モデルのための不動産価格データの用意

予測モデルを構築するにしても、介入効果を測定するにしても、不動産価格データを用意しなければならない。しかし、不動産の価格データには、様々な種類があり、その性質は入手ルートによって大きく異なるために、どのような情報をアクセスするのか、その情報源は信頼できるものであるのかどうかといった点は、注意深く判断していかなければならない。例えば、住宅価格を例にとれば、その情報源がインターネット上で公開される広告情報である場合には、その価格情報は市場価格ではなく、売り手が売りたいと思っている希望価格である。しかし、実際の市場では、その希望した価格で売れるることは稀であるために、実際の市場価格よりも情報のバイアスがかかっていることが多い (Shimizu and Nishimura (2006)[8], Shimizu, Nishimura and Watanabe(2016)[17])。

一方、日本を除く多くの国では、実際に成約した市場価格が入手できるために、日本の市場は閉鎖的で不透明な市場であると言われることがある。しかし、そのような国では、登記簿に取引価格が記載されていることが一般的であり、情報としては土地と建物の面積と取引価格が取引日または登記記載日といった時点情報とともに掲載されているだけである。第 2 章で整理しているように、不動産の価格は、交通利便性や周辺環境、建物の構造などによって差別化するために、そのような情報の記載がないデータでは、予測力の高いモデル構築ができないことがわかる。

このようなデータ資源の性質は、①どのようなモデルを設定するのか、②どのような手法によって推計するのか、③推計されたモデルをどのように改善していくのかといった問題に影響を与えることになる。本来は、第 2 章でみたように、消費者または生産者の行動からモデルを設定し、第 3 章で整理されたような数学的な特性に裏付けられた推計手法を適切に選択していくこととなる。その選択においては、第 4 章で示されたような、各推計手法の特性を踏まえながら、データサイエンティストは判断していかなければならない。

ここでは、国土交通省が構築し、公開している取引価格情報を用いた分析例を紹介する。データは、東京 23 区の住宅市場を対象としている。まず、モデルを構築していくうえで、何を目的変数にとり、説明変数とするのかといったか選択が重要となる。このような設定には、固有分野への精通が重要となる。以

表5.1 国土交通省土地・建設産業局・不動産価格指標（住宅）の作成方法（2016）

不動産の特性	説明変数	住宅地	戸建住宅	マンション (区分所有)
広さ	面積			
	部屋の地上階数			
	建物延床面積			
	建物総回数			
近さ	最寄り駅からの直線距離			
	都道府県内主要駅からの直線距離			
新しさ	築年数			
	改修済み			
地域性	所在する市区町村			
	用途地域区分			
	南向きか否か			
取引条件	取引主体の属性			

下の表5.1は国土交通省が示している不動産の特性である。この表からも、経験からもわかるように、不動産価格は、その広さであったり、新しさであったり、最寄り駅までの距離といった要因が、価格を差別化していると予想される。

このような不動産価格と不動産特性との関係は、Hill, Scholz, Shimizu and Steurer (2018)[5] が指摘するように、時間的にも変化してしまう。全ての特性が時間を通じて一定で価格に影響を与えるわけではなく、また、国や地域によってもその影響の度合いは変化してしまう。

このような価格形成要因を参考としつつも、入手できた情報すべてをカバーできているわけではなく、これら以外の情報も分析者が生成することもできる。地理情報システムなどを利用して近隣の公共施設や病院、スーパーといった不動産の周辺環境を扱ったり、Harrison and Rubinfeld (1978)[10] などのように大気汚染レベルを説明変数に追加したりすることもある。このような情報生産が、実は、不動産市場分析において最も重要になってくるのである。このように、情報が入手できたのちには、どのような手法を選択していくのかといった問題に直面する。

5.3 推計手法の選択肢

5.3.1 推計手法の概要

住宅価格の分析を行った従来の研究では、多くが OLS を用いた線形モデルとして推定している。その選択理由として、シンプルなモデルであればモデルの推計価格に対して明確な説明が可能であるという点が大きい。しかし、専有面積や築年数などでは、住宅価格に与える影響が区間によって異なり、非線形になると考えられ、従来の研究でもその存在が示されている。

また、線形モデルとして推定する場合、住宅価格を決定する上で主要な要因である築年数と建築年代といった多重共線性を持つ説明変数を同時に投入してしまうと、モデルの推計値が不安定になってしまふことも挙げられる。そこで、先行研究のうち Shimizu, Nishimura and Karato (2014)[16] や Shimizu, Nishimura and Watanabe(2016)[17] では、非線形モデルとして推計するため、ノンパラメトリックなモデルである連続量 dummy モデル (DmM) と、AIC を評価指標とした Switching Regression Model(SWR)、一般加法モデル (GAM) が用いられ、非線形性を考慮した推計が行われている。非線形モデルを扱う課題として、モデルが複雑になり再現性が低下することが考えられる。

これらのことを考えると、推計手法の選択を行なう際は、推定や予測に対する説明性を担保できるような状態を保ちつつ、線形回帰モデルの持つ課題を克服したり、非線形性を考慮することが重要そうである。ここでは不動産価格の推計手法の選択肢として、代表的な機械学習の手法とその特徴を解説する。

5.3.2 決定木（回帰木）

決定木は、クラス分類と回帰タスクに広く用いられているモデルである。決定木では、Yes/No で答えられる質問で構成された階層的な木構造を学習する。不動産のデータであれば例えば、単身向けの住宅か、ファミリー向けの住宅か、新耐震基準を満たす物件か、旧耐震基準で建てられた物件か、といった分類をしていく。図 5.1 は延床面積で決定木を作成した例である。

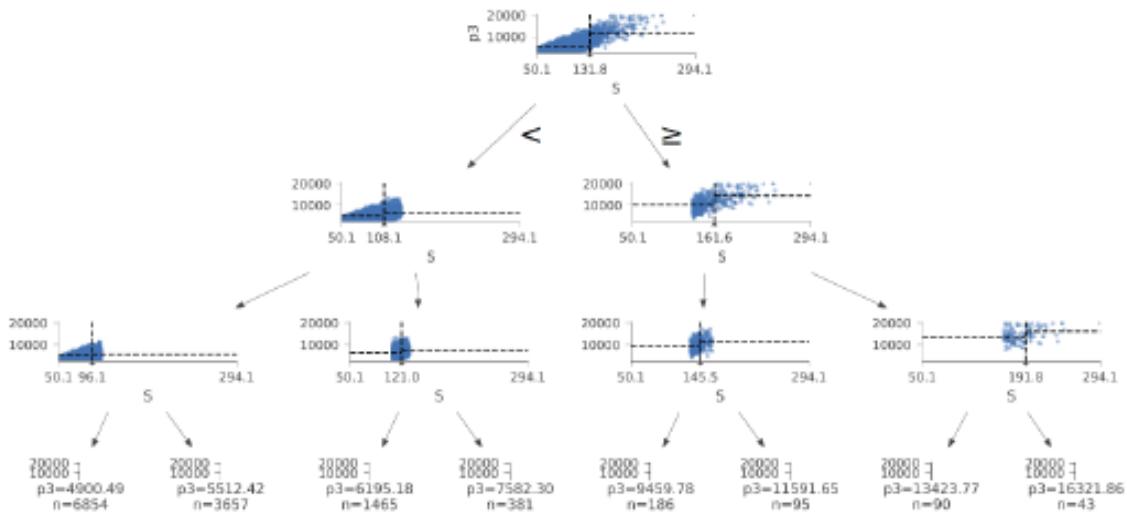


図 5.1 決定木によるデータの分割

この図 5.1 からわかるように、木のノードはどのような基準で分割されているかであり、終端ノードは答えを表す。これを機械学習の用語で表現すると、延床面積を特徴量として扱い、分割の後、延床面積から住宅価格を予測したということになる。

決定木は先に述べたような分割を各特徴量に対して行なうアルゴリズムであるが、分割をやりすぎると容易に過学習の状態となる。3 章と 4 章でも述べられているが、機械学習は柔軟かつ複雑なモデルを推定できることにより、学習データ（訓練データ）に対して過剰に適合してしまい、未知のデータに対応できな

い状態を指す。この問題に対処するために、ホールドアウト検証という、学習に用いるデータと検証に用いるデータを分けることや、学習に関する条件や正則化を導入するパラメータを適切に設定してあげる必要がある。決定木におけるチューニングしなければいけないパラメータとしては、決定木の特徴量の数、決定木の深さ、木を分割する際の最小のサンプル数あたりである。

5.3.3 決定木の発展的な手法

決定木から出発した手法として、ランダムフォレストと呼ばれる手法がある。ランダムフォレストは決定木にアンサンブル法を取り入れたアルゴリズムである。アンサンブル法とは、複数の機械学習モデルを組み合わせることで、より強力なモデルを構築する手法である。機械学習の文献にはこのカテゴリの手法がたくさん存在するが、さまざまなデータセットに対するクラス分類や回帰に関して有効であることがわかっている。その一つがランダムフォレストである。

さらに、勾配ブースティングと呼ばれる手法も出てきている。勾配ブースティングは複数の決定木を組み合わせてより強力なモデルを構築するもう1つのアンサンブル手法である。ランダムフォレスト同様、分類問題にも回帰問題にも対応できる。ランダムフォレストとは対象的に、勾配ブースティングでは、1つ前の決定木の誤りを次の決定木が修正するようにして、決定木を順番に作っていく。勾配ブースティングは機械学習のコンペティション kaggle 等でも広く用いられている。ランダムフォレストに比べるとパラメータ設定の影響を受けやすいが、パラメータチューニングさえうまく行えればこちらの方が性能がよいことが知られている。

勾配ブースティングには、木の深さや木の数を設定するパラメータの他に、learning_rate（学習率）という重要なパラメータがある。これは、現在作成している決定木が、それまでの決定木の間違えをどれくらい強く補正しようとするかを制御するパラメータである。学習率を大きくすると、それ以前に作成した木の結果をそれ以降に作成する決定木が強く補正を行おうとして、モデルは複雑になる。決定木までは構成された木を可視化してどのように分割されて計算されたかを紐解くことができるが、勾配ブースティングのような木構造は一部を取り出してみることはできるものの、推計結果にたどり着くプロセスのすべてを容易に可視化することはできない。

5.3.4 ニューラルネットワーク（ディープラーニング）

ニューラルネットワークは、3章、4章で詳細に説明しているが、ここでも、再度説明しておく。最近では、「ディープラーニング」という名前で再度注目を集め、多くの領域のタスクで結果を残しているが、ディープラーニングの多くは特定のタスクに最適化されて作られたものである。

ニューラルネットワークの元となる考え方には、1957年に考案されたパーセプトロンというアルゴリズムがある。単純パーセプトロンは式(5.1)で表され、複数の信号を入力として受け取り、ひとつの信号を出力するアルゴリズムである。パーセプトロンは入力信号、出力信号、重みで構成され、信号は0か1の二値の値である。入力信号は、ニューロンに送られる際に、それぞれに固有の重みが乗算され、その値が閾値を超えることで1を出力するという動作を行う。パーセプトロンは、複数ある入力信号のそれぞれに

固有の重みを持ち，その重みは各信号の重要性をコントロールする要素として働く。

$$f(x) = \begin{cases} 1, & |w_1x_1 + w_2x_2| \leq \theta \\ 0, & |w_1x_1 + w_2x_2| > \theta \end{cases} \quad (5.1)$$

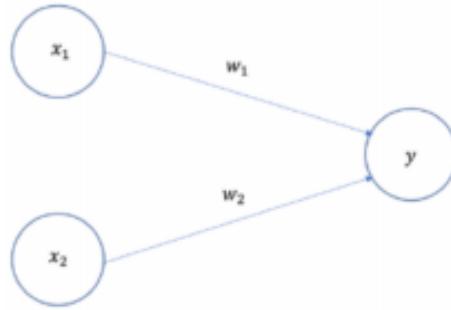


図 5.2 単純パーセプトロン

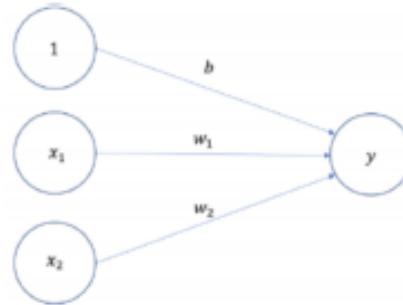


図 5.3 バイアス

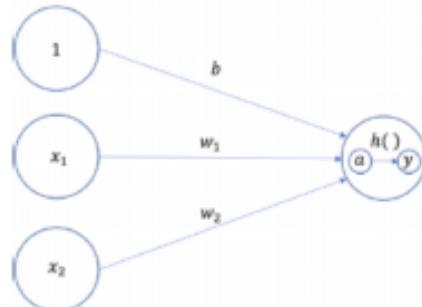


図 5.4 活性化関数

ニューラルネットワークは図 5.2 で示した単純なパーセプトロンをいくつも接続して多層化したものである。ここでは、単純パーセプトロンで扱われるバイアスと各層ごとに設定される活性化関数について触

れる。ニューラルネットワークにおける重みは各信号の重要性をコントロールする値であったが、一方で、図5.3で示すbのように、バイアスはニューロンの発火のしやすさをコントロールするものである。従って、式(5.2)のように各重み付きの入力信号とバイアスの和を計算し、その値が閾値を超えていれば出力信号として1が出力されることになる。図5.4で示す活性化関数は入力信号の総和をどのように出力信号に変換するのかを決める関数である。

$$a = w_1x_1 + w_2x_2 \quad (5.2)$$

式(5.3)で示す活性化関数は、入力信号を受けて次の層へ渡す出力や最終的な出力を制御するために用いられる。基本的には分類問題や回帰問題といった解きたいタスクに合わせて設定する必要がある。代表的な活性化関数の例として、シグモイド関数やソフトマックス関数がある。

$$y = h(a) \quad (5.3)$$

損失関数はニューラルネットワークモデルの性能の悪さを示す指標である。現在のニューラルネットワークモデルが学習データをどれだけよく表しているのかという指標になる。損失関数は任意の関数を用いることができるが、ニューラルネットワークでは一般的に2乗和誤差や交差エントロピー誤差が用いられる。以下の式(5.4)と式(5.5)に、代表的な2乗和誤差と交差エントロピー誤差を示す。

2乗和誤差

$$E = \frac{1}{2} \sum_k (y_k - t_k)^2 \quad (5.4)$$

交差エントロピー誤差

$$E = - \sum_k t_k \log y_k \quad (5.5)$$

ニューラルネットワークにおける学習の目的とは、損失関数の値を可能な限り小さくするようなパラメータを見つけることである。この最適なパラメータを見つけることをパラメータの最適化という。ニューラルネットワークにおけるパラメータ空間は非常に複雑であり、最適なパラメータを見つけることは非常に難しい問題である。

ニューラルネットワークにおける最適なパラメータの探索は、パラメータの勾配を手がかりにパラメータを勾配方向に更新することを繰り返すことで、徐々に最適なパラメータに近づけることで達成される。そこでニューラルネットワークにおける最適化手法として基本的な確率的勾配降下法(SGD)が広く用いられてきた。確率的勾配降下法は単純で実装が簡単である一方、問題によっては非効率な場合がある。そこで最近の研究では確率的勾配降下法の欠点を克服した最適化手法がいくつか提案されている。ここでは基本的な最適化手法である確率的勾配降下法の説明を行い、それらに改良を加えてニューラルネットワークに広く用いられるようになった最適化手法を紹介する。

5.3.5 最適化関数

SGD(stochasticgradientdescent)は、確率的に無作為に選びだしたデータに対して、その勾配を計算し、勾配方向にパラメータを更新することを繰り返す手法である。

$$W \leftarrow W - \eta \frac{\partial L}{\partial W} \quad (5.6)$$

式 5.6 では、更新する重みパラメータを W 、 W に関する損失関数の勾配を $\partial L / \partial W$ とする。 η は学習係数を表し、0.01 や 0.001 といった値を設定し、勾配に対し、どの程度パラメータを更新するかを制御する。

SGD に運動量の考え方を導入して、最適なパラメータへ早く近づく手法が、MomentumSGD である。

$$v \leftarrow \alpha v - \eta \frac{\partial L}{\partial W} \quad (5.7)$$

$$W \leftarrow W - v \quad (5.8)$$

また、ニューラルネットワークの学習において、学習係数の値が重要である。学習係数が小さすぎると学習に時間がかかりすぎてしまい、学習係数が大きすぎると発散して正しい学習が行えない。この学習係数に関する有効な手法として、学習係数の減衰を取り入れた AdaGrad と呼ばれる最適化手法が提案されている。これは学習が進むごとに学習係数を徐々に小さくしていくという手法である。学習の序盤では学習を多めにし、次第に学習を少なくしていくことが、ニューラルネットワークの学習ではよく用いられている。

また、式(5.7)、式(5.8)で表される MomentumSGD の運動量の考え方と、AdaGrad の学習が進むに連れて学習係数を調整する考え方との 2 つを組み合わせた手法として、Adam と呼ばれる最適化手法が提案されている。また Adam の特徴としてハイパーパラメータの偏り補正が行われている。この章の実験でもデータを効率よく学習できることから Adam を最適化手法として採用している。

ニューラルネットワークでは重みやバイアスといったパラメータとは別に、ハイパーパラメータと呼ばれるパラメータが多く存在する。ハイパーパラメータの例としては学習の基本となる学習率や学習回数を意味する epoch、各層のニューロン数やバッチサイズが挙げられる。また過学習と呼ばれる問題に対処するために設定する weightdecay や Dropout の値もハイパーパラメータである。ハイパーパラメータは一般的に試行錯誤を繰り返し人手で事前に設定することが多い。

ニューラルネットワークの学習では、損失関数の値を可能な限り小さくするようなパラメータを見つけることを目的としている。ニューラルネットワークでは最適なパラメータを見つけるために、パラメータの勾配を使って、勾配方向にパラメータを更新するというステップを何度も繰り返し行われる。これはパラメータ空間が非常に複雑な問題の大域的最適解を探索することであるが、パラメータによっては局所的最適解に囚われてしまい、大域的最適解にたどり着くことができない場合がある。すなわち適切なハイパーパラメータを設定しなければ性能の悪いモデルが出来上がってしまうため、大域的最適解にたどり着く、性能が良いモデルを作るためにはハイパーパラメータの調整が必要不可欠である。

ニューラルネットワークにおける学習率とは、訓練データの学習時に、1 回の学習でどれだけ学習するべきか、言い換えると 1 回の学習でどれだけパラメータの更新を行うかに対応している。学習率は一般的に 0.01 や 0.001 といった値を前もって設定する必要がある。一般的に学習率を大きく設定すると学習回数が少なくて済むが、学習結果が不安定になる傾向がある。また、学習率を小さく設定すると学習回数が多くなり、学習結果が局所的最適解にとらわれて、大域的最適解に達しない可能性が高くなる。従って、適切な学習率の設定が学習を正しく行うために不可欠である。

ハイパーパラメータの最適化を行う上で重要なポイントは、良い性能ができるハイパーパラメータの範囲を徐々に絞り込んでいくことである。最初は 0.01 や 0.001 といった対数スケールでおおまかに範囲を設

定し、その範囲の中から性能が良くなるようなパラメータをピックアップし、その値で性能の評価を行うことが良いとされている。この徐々に範囲を狭めていく作業を繰り返し行なうことが適切なハイパーパラメータの設定の基本である。

epoch とはニューラルネットワークの学習において訓練データをすべて使いきった場合の単位を表すパラメータである。ミニバッチ学習を行っている場合、訓練データから確率的に取ってきた訓練データで学習を繰り返し、訓練データの総数分学習した場合を 1epoch とする。

ニューラルネットワークの学習においては、学習に使ったデータセットだけに過度に適応したパラメータが学習されてしまい、テストデータに対して性能が出ない過学習と呼ばれる状態に陥ることがある。過学習を抑えるために Weightdecay と呼ばれる正則化の手法が存在する。過学習は、そのコーパスの特徴だけに特化した重み付けがされてしまい、汎化性能が低下してしまう問題である。そこで Weightdecay では、学習の過程において、大きな重みを持つことに対してペナルティを課すことで、過学習の抑制を図る。具体的には、ニューラルネットワークの学習において損失関数に重みの 2乗ノルムを加算することで重みが大きくなることを抑える。重みを W とした場合、2乗ノルムを用いた Weightdecay は $1/2\lambda W^2$ となり、この $1/2\lambda W^2$ を損失関数に加算する。ここで、 λ は正則化の強さをコントロールするハイパーパラメータである。 λ を大きく設定すると、大きな重みに対して重いペナルティを課すことに対応する。

過学習を抑制するもう一つの手法に Dropout と呼ばれる手法が提案されている。Dropout は、ニューロンをランダムに消去しながら学習する手法である。訓練データの学習時に、あらかじめ設定された割合に基づいて、隠れ層のニューロンをランダムに選択し、その選択したニューロンを消去する。

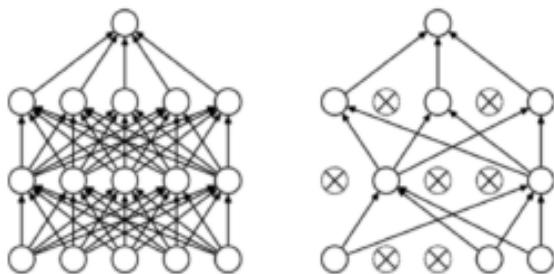


図 5.5 Dropout の概念図

図 5.5 は Dropout の概念図である。図 5.5 に示すように、消去されたニューロンは信号の伝達が行われなくなる。なお、テスト時には、すべてのニューロンが信号を伝達するが、各ニューロンの出力に対して、訓練時に消去した割合を乗算して出力する。機械学習の分野では、アンサンブル学習と呼ばれる手法が用いられる。アンサンブル学習とは、複数のモデルを個別に学習させ、推論時には、その複数の出力を平均する手法である。ニューラルネットワークでも複数の同じ構造のネットワークを用いて、それぞれ学習を行い、テストのときに出力の平均を答えとするように用いられる。アンサンブル学習を行うとニューラルネットワークの認識性能が数 % 向上することが実験的に分かっている。Dropout はこのアンサンブル学習を一つのニューラルネットワーク内で擬似的に再現する手法である。Dropout を用いると、学習時にニューロンをランダムに削除する。これは毎回異なるモデルを学習していると考えられる。また、推

論時には、ニューロンの出力に対して削除した割合を乗算することで、モデルの平均を取っていると考えられるからである。

5.3.6 その他のニューラルネットワーク

現在、研究の成果として様々な特徴を持ったニューラルネットワークが考案されている。本節では各タスクにおいてベースとされている 3 つのネットワークについて概説する。

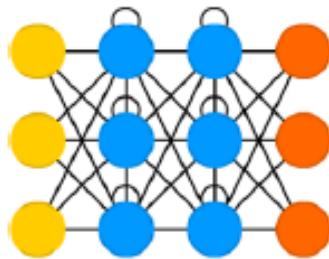


図 5.6 RNN のネットワーク構造

まず 1 つ目にリカレントニューラルネットワーク（以下 RNN）について説明する。RNN（図 5.6）は現在の隠れ層の入力に 1 つ前の隠れ層の値を入れることで出力層を得る、定式化すると、式(5.9)に 1 つ前の入力を足して以下のようになる。

$$h_t = g(Uh_{t-1} + W_1x_1 + b_1) \quad (5.9)$$

RNN は自然言語処理の様々なタスクで成功を収めているが、誤差伝搬時に系列の初めに行くに連れて勾配が伝わらなくなるという問題を抱えている。

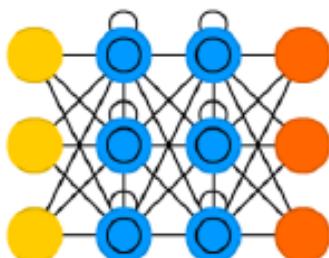


図 5.7 LSTM のネットワーク構造

Long Short Term Memory（以下、LSTM）について説明する。LSTM（図 5.7）は上で述べた RNN の問題を解決するための拡張である。LSTM の中心となるのは、各ステップにおける入力に対してのメモリセル c である。このセル c は入力ゲート、忘却ゲート、出力ゲートの 3 つのゲートにより制御され

ている。以下に LSTM の定式化を示す。

$$h_{t+1}, s_{t+1} = LSTM(w_t, s_t; \theta_{lstm}) \quad (5.10)$$

$$P_{Dec}(w_{t+1}) = softmax(W_{out}h_{t+1}) \quad (5.11)$$

CNN

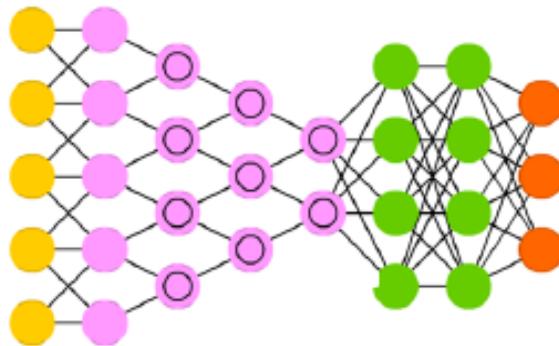


図 5.8 CNN のネットワーク構造

畳み込みニューラルネットワークは画像認識や、音声認識など、至るところで使われている。特に画像認識の領域では、ディープラーニングによる手法の殆どが CNN をベースとしている。CNN では、図 5.8 のように隠れ層は「畳み込み層」と「プーリング層」で構成されます。畳み込み層は、前の層で近くにあるノードにフィルタ処理して「特徴マップ」を得る。プーリング層は、畳み込み層から出力された特徴マップを、さらに縮小して新たな特徴マップとする。この際に着目する領域のどの値を用いるかであるが、最大値を得ることで、画像の多少のズレも吸収される。したがって、この処理により画像の位置移動に対する普遍性を獲得したことになる。畳み込み層は画像の局所的な特徴を抽出し、プーリング層は局所的な特徴をまとめあげる処理をしている。つまり、これらの処理の意味するところは、入力画像の特徴を維持しながら画像を縮小処理することになる。

従来の画像縮小処理と異なるところは、画像の特徴を維持しながら画像の持つ情報量を大幅に圧縮できるところである。これを言い換えると、画像の「抽象化」とも言え、ネットワークに記憶された、この抽象化された画像イメージを用いて、入力される画像を認識、つまり画像の分類をすることが可能となる。

5.4 不動産価格の予測モデル

5.4.1 データの確認

本分析では、株式会社リクルート住まいカンパニーから提供を受けた、東京 23 区において 2000 年～2010 年に成約したと考えられる戸建物件のデータを用いる。同データの要約統計量を確認する（表 5.2）。成約価格では最小値は 2,050 万円、最大値は 2 億円と大きな価格の開きがある。また、建築後年数に着目すると、最小値は築後 2 年であるのに対して、最大値は 49 年と築浅の物件からきわめて古い物件まで含まれていることがわかる。

表 5.2 戸建住宅データの要約統計量

	平均値	標準偏差	最小値	最大値
成約価格 (万円)	6,235.40	2,955.90	2,050.00	20,000.00
部屋数	3.95	1.04	2.00	8.00
最寄り駅までの徒歩時間 (分)	9.95	4.50	2.00	29.00
東京駅までの時間 (分)	31.68	7.53	8.00	48.00
ターミナル駅までの時間 (分)	14.91	6.05	1.00	33.00
建ぺい率	57.07	6.98	50.00	90.00
容積率	164.19	63.25	100.00	360.00
延床面積 (m ²)	109.65	36.27	50.12	247.89
土地面積 (m ²)	102.83	42.57	50.01	249.77
建築後年数	14.70	8.92	2.01	49.72
サンプル数 : 5565				

このようなデータを線形モデルで分析する場合、市場が分割されており、市場構造が異なれば、市場を層別化をしたうえでモデルを構築する必要がある。しかし、以降で行う機械学習を用いた非線形モデルでそのような市場の異なりをアルゴリズムが認識できるかを確認する意味でも、今回はすべてのデータを対象として分析を行うこととした。

まず、このデータで OLS を用いて線形モデルとして推計した結果が表 5.3 となる。回帰係数の符号を確認すると、実際の感覚と異なる部分は現れていない。また、自由度調整済み決定係数は 0.857 と比較的説明力の高いモデルとして推計されていることがわかる。

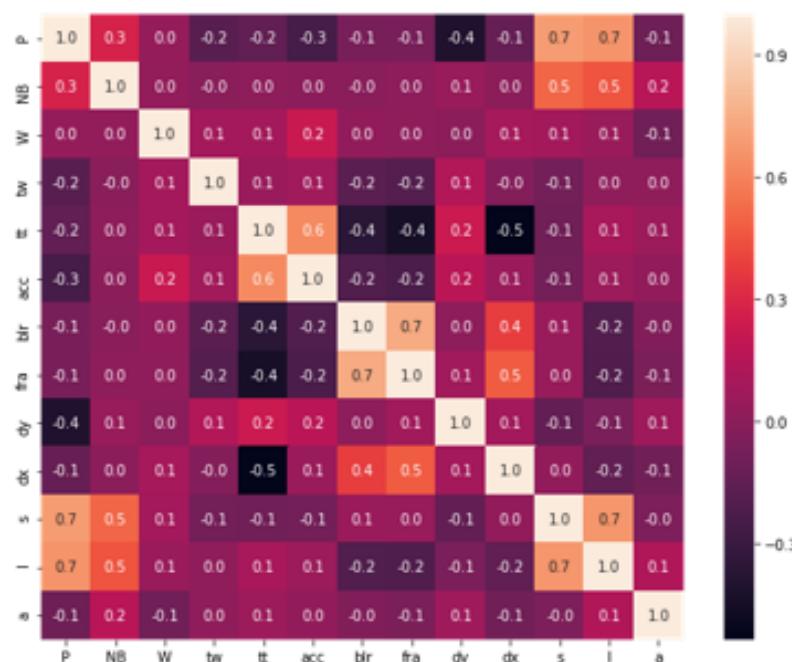


図 5.9 相関マトリックス

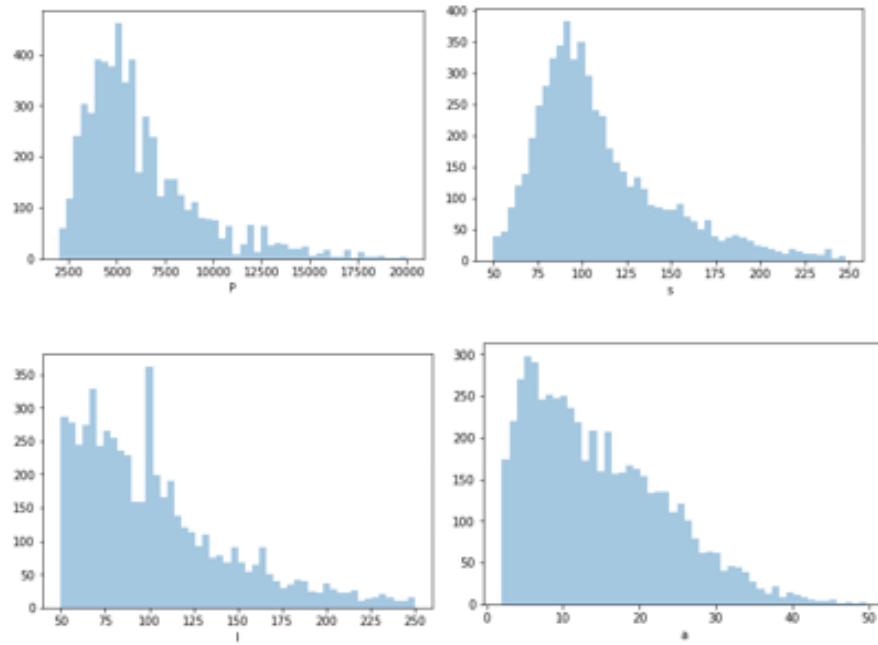


図 5.10

表 5.3 OLS による住宅価格関数の推計結果

	回帰係数	標準誤差	P値
定数項	8.897	0.049	0.000
延床面積	0.003	0.000	0.000
土地面積	0.005	0.000	0.000
建築後年数	-0.007	0.000	0.000
最寄り駅までの徒歩時間	-0.010	0.001	0.000
ターミナル駅までの時間	-0.006	0.000	0.000
前面道路幅員	0.002	0.000	0.000
建ぺい率	-0.003	0.001	0.000
木造ダミー	-0.009	0.007	0.177
都市計画用途：商業	0.029	0.016	0.069
都市計画用途：工業	-0.050	0.050	0.000
市区町村ダミー	Yes		
成約時点ダミー	Yes		
サンプル数：5565			
自由度調整済み決定係数：0.857			

5.4.2 手法間の比較

この節では、線形モデルである OLS をベースラインとし、機械学習手法がどのような特性を持っているのか確認する。OLS を用いた線形回帰モデルによる分析に必要な基礎的な統計知識として、清水千弘 (2016)[14] ではデータの持つ誤差の話から、記述統計の説明が丁寧に解説されており有用である。比較対

象として、前節で紹介した、機械学習の手法である決定木をもとにしたランダムフォレストと XGBoost を選択した。また、ニューラルネットワークによる手法として多層パーセプトロンを選択した。以下で各手法についての概要を説明する。OLS は定数項（切片）と説明変数の係数によって値を予測する線形モデルであり、最小二乗法によって係数と切片を決定する。ランダムフォレストは、決定木とバギングを組み合わせた手法であり、決定木を大量に生成し、各決定木の結果を集計して予測を行う。各決定木は独立して異なる特性を持つように学習する。XGBoost は、決定木とブースティングを組み合わせた手法であり、決定木を逐次的に増やしていく、生成済みの決定木の誤差を補正するように、新たな決定木を生成し学習を進めていく。多層パーセプトロンは、入力、中間、出力の 3 層からなるニューラルネットワークの手法であり、バックプロパゲーションを用いた学習を行う。

この節における機械学習手法では、グリッドサーチによるパラメータチューニングを行い、過学習を抑制するためにパラメータを定めてからモデルの構築を行った。

5.4.3 評価

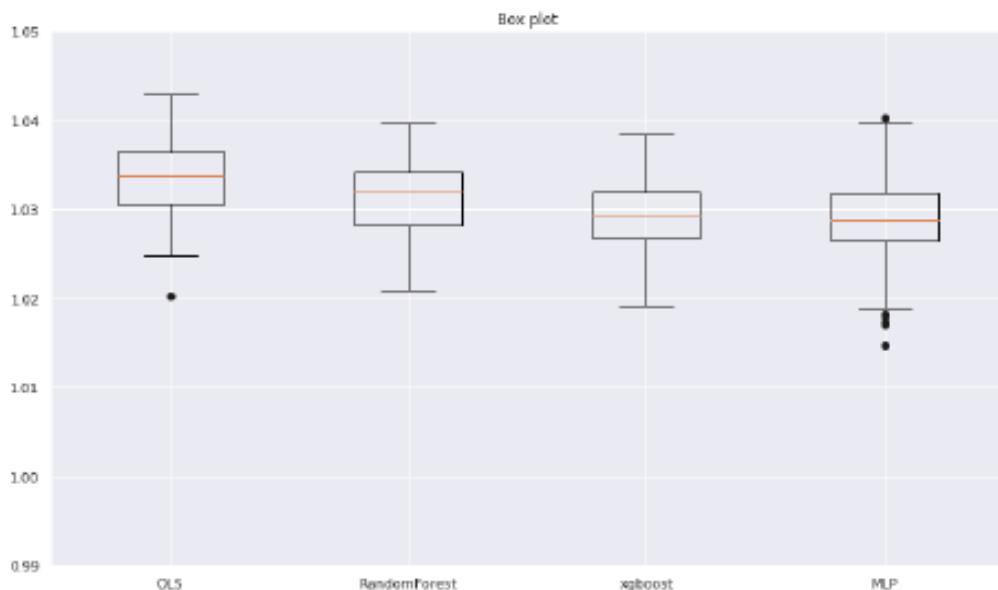


図 5.11 各手法の誤差率の平均値の分布

まず、手法間の誤差率分布を比較した結果を図 5.11 に示す。データは学習データとテストデータを 8 対 2 でランダムに分割し、各手法において同じ説明変数を投入したモデルで一つ一つの物件に対して取引価格の予測を行い、予測価格/実際の取引価格で定義される誤差率を算出した。その実験を 200 回繰り返し行い、誤差率の平均値の分布を作成した。結果から、OLS や決定木をベースにした手法は予測結果が大きくはずれてしまう可能性が少ないと見て、ニューラルネットワークをベースにした多層パーセプトロンでは誤差率の平均値の分布に外れ値が見られた。このことから多層パーセプトロンでは推計結果に大外れが生じてしまうことが確認できた。

5.4.4 適切なパラメータ設定

機械学習モデルを構築する際，アルゴリズムが認識して設定するパラメータと，人手によって設定しなければいけないパラメータが存在することを意識する必要がある。5.3節のハイパーパラメータについて説明している部分でも繰り返し述べている通り，アルゴリズムに対して人間からどういった指標を元に，どのように学習すればよいかということをパラメータとして設定する必要がある。

表 5.4 パラメータチューニングの有無による RSS の比較

手法	チューニング有	チューニング無
OLS	1,603,106	-
RandomForest	1,681,227	1,689,219
XGBoost	1,174,908	1,489,999

表 5.4 で示す結果は，今回用いた決定木による手法でのパラメータ設定の有無の比較である。ここでは，実験に用いた python ライブラリのデフォルト値をパラメータとして用いた場合と，グリットサーチによりパラメータを探索して定めた場合を比較している。木の深さや葉の部分に含まれるサンプル数といったパラメータしか設定しない RandomForest では大きな違いは見られないものの，学習率の考え方に入ってくる，構築した木の正誤を次の木に反映させるようにモデルを構築する XGBoost ではパラメータのチューニングの有無で結果に大きな違いが見られた。

5.5 不動産市場における介入効果の測定

不動産市場において介入効果を計測する際には，第2章で指摘されているように，消費者の効用関数における顯示選好を市場価格関数の中で測定するために，ヘドニック・アプローチは，有力な手法となる。しかし，ヘドニック・アプローチの適用において最も注意すべき問題として，内生性(endogeneity)の問題が挙げられる(詳細は，清水・唐渡(2007)[9]を参照されたい)。内生的(endogenous)とは，説明

変数と誤差項との間に存在する相関性である。この場合、ヘドニック・アプローチにより得られた推定量は一致性を持たず，バイアスを持つため，介入の効果を正しく評価することができない。内生性が発生するメカニズムとしては，観測誤差，除外変数，同時性，セレクション・バイアスなどが挙げられる。これらはいずれも対処すべき重要な課題であるが，既存のデータセットを活用・流用する場合に遭遇することが多いセレクション・バイアスに注目する。セレクション・バイアス(sample selection bias)は，データ収集の際の制約や問題により，得られたサンプルの処置群と対照群で属性が異なる場合に生じるバイアスである。セレクション・バイアスへの対処法としては，Heckman(1974)[4]が提唱した2段階推定法や，Besley and Case(2000)[1]で知られる DID(difference in difference: 差の差分法)，そして傾向スコア(propensity Score)が挙げられる。

本章では，Deng et al(2018)[2]に基づき，傾向スコアを用いてサンプルを調整することで，セレクション・バイアスに対処したヘドニック・アプローチを紹介する。傾向スコア(propensity score)を用いた

解析は，Rosenbaum and Rubin(1983)[12]により提唱された。複数の共変量を傾向スコアという一つの数値に集約することで，交絡因子の影響を除去する方法として用いられる（大林, 2016）[11]。傾向スコアを用いることで，いくつか前提のもとで偏りのある標本をより無作為抽出に近い標本になるべく近くなるように調整できることから，前節で挙げたセレクション・バイアスに対処するための技法として適用できる可能性があり，2000年代初頭以降注目を集めた。適用範囲は広く，医学，経済学，心理学で多くの研究が報告され，Web調査の調整にも利用される事例もある（星野, 2004）[6]。

傾向スコアを用いた分析では，まず，介入が行われる群（処理群）にサンプルが割り当てられるか否かを共変量によって説明するモデルを設定し，そのモデルのパラメータの推定を行う。推定されたパラメータの推定値を用いて，サンプルごとに処理群に割り当てられる予測確率を計算し，これを傾向スコアの推定値とする。次に，推定した傾向スコアをもとにして，似通った2つの群を作り出し，その2群間での大きな違いは割り当て変数が1か0かのみという状況を作り出す。その上で，効果を推定するという手順で分析が進められる。

傾向スコアを用いてデータ処理すると，共変量の平均値は2群間で一定程度のバランスを取ることができ，内生性の問題に対応することができる。また，多次元の共変量を1次元に集約するため，層別化を用いた詳細な分析にも耐えることができる。

今，以下のように定義する。

Y_1 ：介入をうけた場合 ($D = 1$) の従属変数

Y_0 ：介入をうけない場合 ($D = 0$) の従属変数

D ：割り当て変数 (1: 介入をうけた場合, 0: 介入をうけない場合)

x ：共変量として観測可能な属性。

ただし， x は (x_1, \dots, x_i) を要素とするベクトルとする

$P(x) = \Pr(D = 1 | x)$: x の属性を持つサンプルが $D = 1$ に割り当てられる予測確率

介入が与える効果を， $D = 1$ に割り当てられたサンプルと， $D = 0$ に割り当てられたサンプルの従属変数の期待値の差とした場合，その効果は次の式で表される。

$$E(Y_1 | D = 1) - E(Y_0 | D = 0) \quad (5.12)$$

このとき，共変量が $D = 1$ への割り当てられやすさに影響を与えていた場合，推定結果にはバイアスが生じる可能性がある。そこで，介入による効果を次の式のように考える。

$$\Delta_{D=1}(x) = E(Y_1 - Y_0 | P(x), D = 1) \quad (5.13)$$

この式は，実際に介入が行われたサンプルについて，介入が行われた場合と仮に介入が行われなかった場合の従属変数の差の期待値を表している。このように，実際に介入が行われたサンプルにおける，仮に介入が行われなかった場合の従属変数がわかれば，純粋な介入の効果が取り出せる。しかし，通常はそのようなデータは観察できないため，これらを代理する値を $P(x)$ でウェイト付けした，実際に介入が行われていないサンプルの従属変数から推定する。この値と，実際に介入が行われたサンプルをマッチングさ

せると、介入の効果の推定値は以下のようなになる。

$$\hat{\Delta}_{D=1}(\mathbf{x}) = \frac{1}{n_1} \sum_{\substack{i=1 \\ \{D_i=1\}}}^{n_1} \{Y_{1i}(\mathbf{x}_i) - \hat{E}(Y_{0i} | P(\mathbf{x}_i), D_i = 0)\} \quad (5.14)$$

ここで、 n_1 は実際に介入が行われたサンプルの数を示す。ここで、代理として用いられる介入が行われなかったサンプルの従属変数の期待値は次の式で表される。

$$\hat{E}(Y_{0i} | P(\mathbf{x}_i), D_i = 0) = \sum_{\substack{j=1 \\ \{D_j=0\}}}^{n_0} W_i(P(\mathbf{x}_i)) Y_{0j} \quad (5.15)$$

n_0 は介入が行われなかったサンプルの数を、 W は $P(x)$ が与えられた場合のウェイトを示す。比較するときのウェイト付けとしては最近傍マッチング (Nearest Neighbors Matching) が用いられる。最近傍マッチングでは、介入が行われる確率が最も近い、介入が行われたサンプルと介入が行われなかったサンプルをマッチングさせ、実際に介入が行われたサンプルが仮に介入が行われなかった場合の従属変数の代理として、マッチングされた介入が行われていないサンプルの従属変数を用いる。

なお、属性 x を持つサンプルにおいて、介入が行われる群に割り当てられる確率を推計するためには、ロジスティック回帰や以下に挙げるプロビット回帰が用いられることが多い。ここで、 β は $(\beta_1, \dots, \beta_i)'$ を要素とするベクトルとする。

$$P(\mathbf{x}) = \Pr(D = 1 | \mathbf{x}) = \Phi(\mathbf{x}\beta) = \int_{-\infty}^{\mathbf{x}\beta} \frac{1}{2\pi} \exp\left(-\frac{z^2}{2}\right) dz \quad (5.16)$$

5.6 傾向スコアを用いた実証分析の事例

5.6.1 傾向スコアマッチングによるサンプル調整と課題

Deng et al(2018)[2] では、東京のオフィス市場を対象として、環境認証を持つオフィスビルと持たないビルにおいて、家賃に乖離が存在するのかどうか、存在するのであればどの程度の乖離が存在するのかということを実証的に分析している。環境認証とは、環境配慮型社会を実現するために、世界的な規模で不動産が発生する環境への負荷を縮減していくこうとする取り組みの中で、環境への負荷が小さい不動産に対して認証を与える制度である。ここでの仮説は、このような環境性能の高いオフィスには、テナントが高い家賃を支払っても借りたいと考えているのではないか、その結果として、環境認証を持つ不動産はプレミアムを持つのではないか、ということである。

傾向スコアを用いた分析では、共変量が似通っているサンプルを形成することで、説明変数と被説明変数が共変量から受ける影響を取り除くアプローチをとる。

分析に用いた変数の記述統計を表 5.5 の (1) に示す。表 5.5 の (2) および (3) は、それぞれ (1) のうち環境認証を付与されたサンプル、付与されていないサンプルの記述統計である。

この分析では、オフィスビルの新規賃料に影響を与える環境認証の有無とその他の説明変数を特定化することでオフィス新規賃料関数を推定する。オフィスビルの新規賃料は、そのオフィスビルの属性情報お

より環境認証取得状況をもとに、一般的には以下のようなヘドニック価格関数として表される。

$$\ln R_i = \alpha + green'_i \cdot \beta + x'_i \cdot \gamma + \varepsilon_i \quad (5.17)$$

ここで、左辺に示される $\ln R_i$ は成約事例 i の新規賃料の対数をとったものである。また、 $green_i = (0, 1)$ は環境認証ダミーであり、 $x'_i = (x_{1i}, x_{2i}, \dots, x_{ni})$ は成約事例 i における n 種類の属性を表すベクトルである。

(5.17) 式のベーシックなヘドニック・モデルで OLS 推定した結果を表 5.6 の (1) に示す。推計結果は、 $+0.0439(0.0115)$ とプラスで有意となった（括弧内は標準誤差）。これは、環境認証が付与されている物件はそうでない物件に比べ、新規賃料が約 4% 高いことを示している。

この結果をもって、環境認証の取得は、より高い収益を得ることが期待できる経済合理性にかなった行動であることを示唆していると言い切れるだろうか。以下のような留意すべき点があると考える。

まず、環境認証が付与されているのは、資金力に富み、環境認証を取得するために必要なコストを支払う余裕がある大手デベロッパーや REIT が開発、所有する大規模で築年数が若い物件が多い。そのため、このモデルの中での説明変数には、デベロッパーの違いや REIT が取得したものかどうかなどといった変数が含まれていないために、まったくそのような相違は、モデルの中ではコントロールができない。また、築年数や規模などはコントロールされていたとしても、十分に機能していない可能性もある。

分析用データを環境認証の有無別に集計した表 5.5 の (2) と (3) を比較すると、環境認証が付与されているサンプルの延床面積および築年数の平均値はそれぞれ $61,717m^2$, 8.78 年なのに対し、付与されていないサンプルではそれぞれ $16,513m^2$, 24.57 年と小規模、築古である傾向がある。環境認証ダミーの係数として推定された $+4.39\%$ は、本当に環境認証の有無によるものなのか、規模、新しさ、立地そして成約時期の影響が混ざっている可能性はないのか、慎重に分析する必要がある。

そこで、傾向スコアを用いて、分析対象をコントロールした上で、環境認証の効果を識別することが有効になる。傾向スコアの推定に際し、環境認証を付与されているか否かを被説明変数としたプロビット

表 5.5 分析用データの変数と記述統計量

	(1) Full Sample				(2) Green-Label		(3) Non-Green-Label	
	Number of Observations = 6,758				Number of Observations		Number of Observations	
	mean	standard deviation	minimum	maximum	mean	standard deviation	mean	standard deviation
新規賃料	5.169.73	1.862.02	1.845.25	16,649.60	7.638.37	2.154.42	5.030.42	1,743.04
環境認証ダミー	0.05	0.22	0.00	1.00	1.00	0.00	0.00	0.00
延床面積	18,928.33	37,444.15	992.56	379,447.90	61,717.76	80,973.86	16,513.61	31,669.82
築年数	23.73	11.83	0.00	59.91	8.78	9.77	24.57	11.36
地上階数	11.69	7.70	3.00	60.00	20.17	12.50	11.21	7.04
基準階面積	780.24	795.11	99.87	9,834.71	1,533.13	1,217.62	737.76	742.05
都心 5 区ダミー	0.77	0.42	0.00	1.00	0.86	0.35	0.76	0.42
徒歩分数	3.36	2.31	0.00	15.00	2.98	1.80	3.38	2.34
OAフロアダミー	0.68	0.46	0.00	1.00	0.96	0.21	0.67	0.47
個別空調ダミー	0.80	0.40	0.00	1.00	0.81	0.39	0.80	0.40
機械警備ダミー	0.83	0.37	0.00	1.00	0.87	0.33	0.83	0.38
リニューアルダミー	0.13	0.34	0.00	1.00	0.12	0.33	0.13	0.34

回帰モデル（(5.16)式参照）を用いた。説明変数は、新規賃料を被説明変数としたヘドニック回帰モデルと同じものを用いる。具体的には、延床面積（坪）、基準階面積（坪）、地上階数（階）、築年数（年）、エリアダミー、最寄駅からの徒歩分数（分）、OA フロアダミー、個別空調ダミー、機械警備ダミー、四半期ダミーを採用する。推定されたプロビット回帰モデルの対数尤度は -831.1474、AIC は 1814.2950 であった。統計的に有意な結果が得られた変数としては、延床面積で +0.4199（標準誤差 0.1331）、築年数で -0.0727（標準誤差 0.0146）、リニューアルダミーで +1.0639（標準誤差 0.1257）が得られた。この結果から、築年が若い、もしくはリニューアルが行われた高品質で設備が整った規模が大きいオフィスビルほど、環境認証を取得する確率が高いことが示され、前項で指摘した内生性バイアスが実際に生じていることが示唆された。

推定された傾向スコアを用いて最近傍マッチングを行うと、環境認証が付与されているオフィスビル（Green building）のサンプルが 361 件、付与されていないオフィスビル（Non-Green building）のサンプルが同数の 361 件、合計 722 件のサンプルが得られた。記述統計量を環境認証の有無別に比較すると、延床面積では Green building が平均 61,717m²（標準偏差 80,973m²）、Non-Green building が平均 48,572m²（56,058m²）であった。また、築年数では Green Building が平均 8.78 年（標準偏差 9.77 年）、Non-Green building が平均 8.36 年（標準偏差 8.03 年）であった。

表 5.5 のサンプル調整前と比較すると、傾向スコアマッチングにより、2 つの群の間の延床面積と築年数の平均値と差が縮まり、似通ったサンプルが作り出されていることが見て取れる。

しかし、マッチングにより抽出された 722 サンプルにおけるヘドニック・アプローチによる推定では、環境認証ダミーの係数推定値は -1.667(1.8874) と有意な結果は得られなかった（表 5.6 の (2)）。この結果は、傾向スコアマッチングを行う前の表 5.6 の (1) での分析とは異なる結果を示している。

表 5.6 全サンプルおよび傾向スコアマッチングによるヘドニック回帰の結果

	(1)	(2)
	Baseline	PS マッチング
環境認証ダミー	0.0439*** (0.0115)	-1.6673 (1.8874)
築年数	-0.0089*** (0.0003)	-0.0628 (0.0554)
切片	Yes	Yes
成約時点ダミー	Yes	Yes
エリアダミー	Yes	Yes
他の共変量	Yes	Yes
サンプル数	6,758	722
環境認証ありの割合	5.34%	50.00%
自由度調整済み決定係数	67.70%	92.13%
モデル	OLS	OLS

ここで注意を要するのは、最近傍の傾向スコアのサンプルを1対1でマッチングするという手法の都合上、抽出されたサンプルのみがヘドニック回帰分析の対象となる点である。本分析の傾向スコアマッチングでは、傾向スコアが高い大規模・新しいオフィスビルが多くマッチングされた一方で、傾向スコアが低い中小規模・古いオフィスビルのサンプルではマッチングされず、ヘドニック回帰分析の対象から除外されやすい。そうすると、本当に環境認証は、経済的なプレミアムを持たないのか、傾向スコアの作り方に問題があったのかといった新しい問題に直面してしまうのである。

このように、傾向スコアマッチングによりサンプルを調整することで、内生性に対処した上で介入効果を確認することが可能になる。しかし、その一方で、マッチングで除外されたサンプルにより構成される市場での介入効果が確認されなくなるという問題が生じる可能性もあるために、十分に吟味しながら結論を道祖びいていかなければならないことがわかる。

5.6.2 傾向スコアによる層別化

前節での問題点を解決するため、傾向スコアの大きさによりサンプルを5層に層別化する。分割の境界としては5分位点を用いた。5.6.1項のプロビット回帰結果で示したように、傾向スコアが小さい層ほど小規模で築年数が古く、傾向スコアが大きい層ほど大規模で築年数が新しい傾向にある。各層においてヘドニック回帰を行うことで、傾向スコアマッチングでは除外されやすい市場においても介入効果の観察が期待できる。

各層においてヘドニック回帰した結果を表5.7の(2)~(3)に示す。なお、傾向スコアが比較的小さい第1層から第3層においてはGreen buildingの割合が少なく(第1層0%，第2層0.22%，第3層0.3%)、十分な結果が得られなかつたため、第4層および第5層の結果のみ示す。

表5.7 傾向スコアによる層別化を用いたヘドニック回帰の結果

	(1)	(2)	(3)
	Baseline	PS層別化 第4層	PS層別化 第5層
環境認証ダミー	0.0439*** (0.0115)	0.1378*** (0.0328)	-0.0058 (0.0105)
築年数	-0.0089*** (0.0003)	-0.0111*** (0.0015)	-0.0130*** (0.0008)
切片	Yes	Yes	Yes
成約時点ダミー	Yes	Yes	Yes
エリアダミー	Yes	Yes	Yes
他の共変量	Yes	Yes	Yes
サンプル数	6,758	1,351	1,352
環境認証ありの割合	5.34%	2.22%	23.96%
自由度調整済み決定係数	0.6770	0.6218	0.7684
モデル	OLS	OLS	OLS

傾向スコアが最も高い第5層(表5.7の(3))では、環境認証ダミーの偏回帰係数は $-0.0058(0.0105)$ とほぼゼロに近いマイナスとなった。第5層のサンプルの平均値を見ると、延床面積でGreen buildingが $66,651m^2$ 、Non-Green buildingが $44,826m^2$ 、築年数でGreen buildingが7.3年、Non-Green buildingが10.7年と、大規模で築年数が新しいサンプルが多く、層別化によりサンプルの調整もされている。この結果は傾向スコアマッチングを用いた表5.6の(2)の結果と整合的である。

第4層(表5.7の(2))では $+0.1378(0.0328)$ と有意にプラスの効果があることが示されている。第4層のサンプルの平均値は、延床面積でGreen buildingが $19,894m^2$ 、Non-Green buildingが $14,674m^2$ 、築年数でGreen buildingが19.3年、Non-Green buildingが22.2年であった。第5層と比較して中規模の古いオフィスビルが抽出されている。このことから、傾向スコアによる層別化を用いることで、傾向スコアマッチングでは除外されやすかった市場においても、サンプル調整を行いながら介入効果を測定できることがわかる。

なお、第4層内のサンプルを観察すると、環境認証が付与されたサンプルはそうでないものに比べ傾向スコアが高い傾向にあった。層別化により比較的近い傾向スコアのサンプルに限定したとはいえ、各層内共変量からの影響を取り除く余地がまだ存在する可能性がある。そこで、以下の2つの方法により頑健性の確認を行った。

まず、層別化されたサンプル集団の中で傾向スコアマッチングを追加して行い、より調整されたサンプルを抽出した上でヘドニック関数を推定する。第4層について追加的に傾向スコアマッチングを行った。環境認証ダミーの偏回帰係数は $+0.1297(0.0370)$ であり、有意な推定結果が得られた。

5.7 不動産市場分析における機械学習の応用と課題

近年における機械学習に代表される技術進歩は、各固有分野において、統計的な解析を容易にするだけでなく、従来の手法以上に高い予測力や識別を可能としてきていることは否定するものではない。しかし、データが自動的に集まり、機械学習の各種手法を適用すれば、簡単に予測や識別ができるというものではなく、依然として高い専門性が要求されている。

不動産価格の予測に機械学習を適用しようとした場合には、推計手法の選択以上に、データ資源をどのように構築していくのかといったことが重要であることが理解できるであろう。また、不動産価格の形成要因を特定し、特定できないとしても必要と考えられるデータを収集・生産しなければならない。

さらに、不動産価格を形成する各要因が、価格に対してどの程度影響を持っているかを解釈したいときには、次数が1の線形回帰モデルが好ましい。なぜなら築年数が大きくなると、不動産の価格はいくら下がるといった解釈が可能であるからである。一方で予測力に注目すると、4章で言及されていたように、予測のバイアスとバリアンスはトレードオフの関係であることから、柔軟性を高めて予測のバイスを抑制しようとすると、複雑なモデルとなり、解釈が困難になってくる。分析時にモデルを利用して回帰係数を推定して解釈を行いたいのか、大量のパラメータから予測を行いたいのかによって手法を使い分ける必要がある。また、モデルの解釈を可能にしつつ、予測力も向上させていきたいというモチベーションから、統計的なアプローチと機械学習によるアプローチの併用も考えていく必要がある。

また、不動産市場分析において、環境認証の有無などの介入効果を計測しようとした場合には、通常の回帰分析などでは、十分に品質差を制御することができなかったり、内生性が存在するために、正しい推

計値を得ることができなかったりすることが理解できたであろう。このような問題には、ニューラルネットワークなどの機械学習の各種手法は、解釈性が弱いものが多いために有効ではない。このような場合には、傾向スコアを用いたマッチング法などといった手法が有効となる。

ビッグデータが容易に入手できるようになり、機械学習の手法が発達する中で、容易に不動産市場分析ができるようになってきている。しかし、このような分析例からも明らかなように、データの量に任せて、容易に利用できるようになった簡易の機械学習のプログラムを適用したとしても、精度の高い予測もできなければ再現性も担保できないことの方が多い。

サービスの利用者は、データの量と分析手法の容易さに騙されないようにしていかなければならない。また、不動産市場を対象として分析を行うデータサイエンティストは、やはり不動産市場に深く精通していかなければならないのである。

参考文献

- [1] Besley, T., & Case, A. (2000). Unnatural experiments? Estimating the incidence of endogenous policies. *The Economic Journal*, 110(467), 672-694.
- [2] Deng, Y., Onishi, J., Shimizu, C., & Zheng, S. (2018). The Economic Value of Green Office Buildings in Tokyo. CSIS Discussion Paper No.155.
- [3] Fuerst, F., & Shimizu, C. (2016). Green luxury goods? The economics of eco-labels in the Japanese housing market. *Journal of the Japanese and International Economies*, 39, 108-122.
- [4] Heckman, J.J.(1974) “Shadow Prices, Market Wages, and Labor Supply”, *Econometrica*, 42(4), pp.679-694.
- [5] Hill , R , R , M. Scholz , C. Shimizu and M. Steurer (2018) , “An Evaluation of the Hedonic Methods Used by European Countries to Compute their Official House Price Indices” ,*Economie et Statistique* n 500-501-502, 221–238.
- [6] 星野崇宏, & 繁樹算男. (2004). 傾向スコア解析法による因果効果の推定と調査データの調整について. 行動計量学, 31(1), 43-61.
- [7] Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1), 34–55.
- [8] Shimizu , C and K.G.Nishimura (2006) , “Biases in appraisal land price information: the case of Japan ,” *Journal of Property Investment & Finance* , 24(2) , 150- 175.
- [9] 清水千弘・唐渡広志 (2007), シリーズ：応用ファイナンス講座 4 『不動産市場の計量経済分析』，朝倉書店.
- [10] Harrison , D and D.L.Rubinfeld (1978) , “Hedonic Housing Prices and the Demand for Clean Air ,” *Journal of Environmental Economics and Management* , 5 , 81-102.
- [11] 大林準. (2016). ロジスティック回帰分析と傾向スコア (propensity score) 解析. 天理医学紀要, 19(2), 71-79.
- [12] Rosenbaum,P.R., & Rubin, D.B.(1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.
- [13] 斎藤康毅 (2016), ゼロから作る Deep Learning Python で学ぶディープラーニングの理論と実装, オライリー・ジャパン.
- [14] 清水千弘 (2016) ,「市場分析のための統計学入門」朝倉書店.
- [15] 清水千弘. (2016). ヘドニック・アプローチを利用した環境価値の計測. 都市住宅学, 2016(92), 12-16.

-
- [16] Shimizu , C. , K. G. Nishimura and K. Karato (2014) , “Nonlinearity of Housing Price Structure -Secondhand Condominium Market in Tokyo Metropolitan Area- ,” *International Journal of Housing Markets and Analysis* , 7(3) , 459-488.
 - [17] Shimizu , C , K.G.Nishimura and T.Watanabe(2016) , “House Prices at Different Stages of Buying/Selling Process ,” *Regional Science and Urban Economics* , 59 , 37-53.
 - [18] 高橋祐介・清水千弘 (2020) ,「不動産市場データを用いた機械学習の評価」 mimeo.

第6章

不動産市場分析における GIS の活用

6.1 不動産市場分析と GIS

不動産市場は、空間的に連担する対象をから、その規則性を定量的に時間軸との変化とあわせて分析を行うことが多い。具体的には、クロスセクショナルな意味での地価の空間格差や時系列的な意味での価格変動要因または空間的時間的な変動構造とその見通しを行うものであるが、いずれの場合においても、空間特性をどのように考慮するのかといった問題はきわめて重要である。二次元または三次元にまたがる空間特性への考慮を、どのようにするのかといったことが、他の経済市場分析などと大きく異なる点である。

そのような中で、地理情報システム (GIS : Geographic Information System) の活用は、極めて有効である。地理情報システムは、コンピューター技術および電子地図の発展によって、急速に普及し、今では、不動産市場分析を行うためには必要不可欠な技術となっている。

具体的には、空間的なデータ整備に利用されたり、また空間構造から何らかの規則性を発見しようとするデータマイニングツールとして活用されたり、さらには座標データとリンクさせることで空間統計学の手法を適用するといったことが行われる。とりわけデータ整備には、従来できなかった多くの困難性を解決していった。空間計測または空間検索・交差検索などのデータ構築である (清水 (2004) 第6章)。そのようなことが実現する過程で、空間統計、空間計量経済学が発達していくこととなった。

そこで、本章では、不動産市場分析における GIS の活用と空間統計学的手法との連動可能性について整理する。さらに、5章からの発展として、不動産価格の予測モデルの構築における空間特性の配慮の事例を示す。

6.2 GIS の活用

6.2.1 データ利用

空間座標が付与されたデータは不動産市場を分析する際ににおいては、データの取得において極めて有効である。例えば、最寄り駅までの距離を測定したい場合、駅と物件までの緯度経度座標があれば計算可能であり、徒歩 10 分圏内のスーパーの店舗数なども瞬時に把握できる。

GIS は空間に関わるストック、フローなどを地物として表示し、分析を行うシステム全般を指す。地物とは空間上に存在する街路樹、住宅、鉄道などの点、線、面的オブジェクトであり、さらに GIS では実際

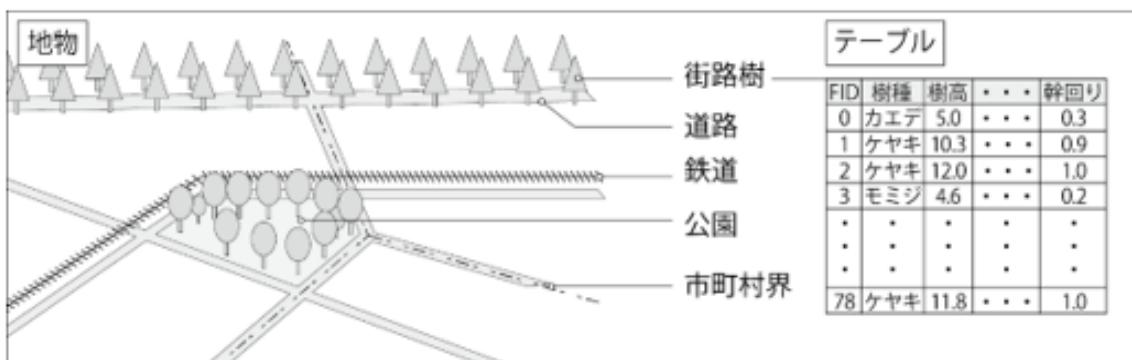


図 6.1 地物とテーブルのイメージ

には見えない境界線、用途地域なども地物として捉えられる。このような情報は表形式で格納されたテーブルに格納され、行はレコード、列はフィールドと呼ばれる。各フィールドに対して数値、文字列などのデータを格納することができ、街路樹の場合、樹種や幹回りなどが該当する（図 6.1）。このように、GIS ではこのような地物の重ね合わせによって、空間の見通しを良くするシステムといえる。GIS で利用するデータは一般的な統計で利用されているデータとは異なり、座標が付与されている。GIS は世界共通で利用できる反面、そのデータ構造や投影法などを正確に理解し、定義する必要がある。

6.2.2 データ形式とそのモデル化

データ形式には大きく分けてラスター（raster）形式とベクター（vector）形式が存在する（図 6.2）。前者はデジタル写真や気温図など、データがセル内に格納されており、行と列に整理されている。ラスターデータはその構造がシンプルであるため、降雨量など連続的な変化量の可視化やメッシュベースの空間解析に優れているが、地点間の距離測定などは難しい。後者は幾何学的解析に優れている一方で、空間的な連続量を可視化するためには空間補間（spatial interpolation）を行い、連続量の推計を行う必要がある。なお、不動産市場の分析では衛星画像などのラスターデータを用いるよりもベクターデータの方が主であるため、以降ベクターデータを対象として説明する。

6.2.3 ジオコーディングと座標系

得られた情報がテキストベースの住所である場合、GIS 上にマッピングするには緯度経度座標を付加する必要がある。このような操作をジオコーディング（geocoding）という（図 6.3）。これは不動産取引データをポイントベクターとして扱う際に必要となるため、重要な前処理である。2019 年 10 月現在、東京大学空間情報科学研究センターが提供する CSV アドレスマッチングサービス（<http://newspat.csis.u-tokyo.ac.jp/geocode/>）などを用いることにより、住所の表データから緯度経度座標を自動的に生成することができる。

ジオコーディングしたデータを GIS 上に投影する際にもいくつかの注意がある。まず、測地系とよばれる特定の空間座標を示すために必要な測量方法及び基準の体系である。わが国では日本測地系と世界測

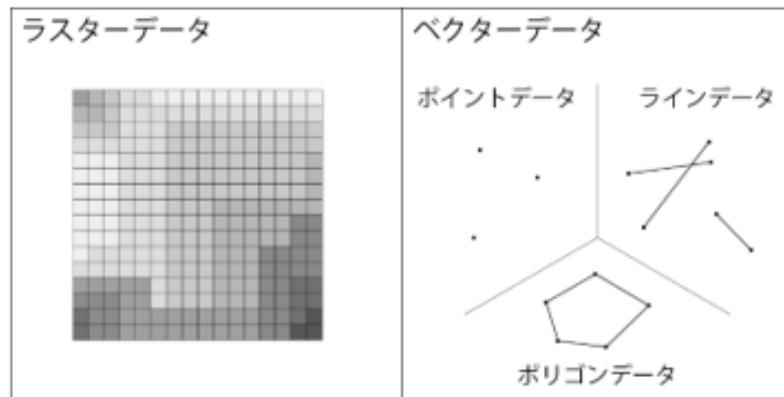


図 6.2 ラスター形式とベクター形式の差異

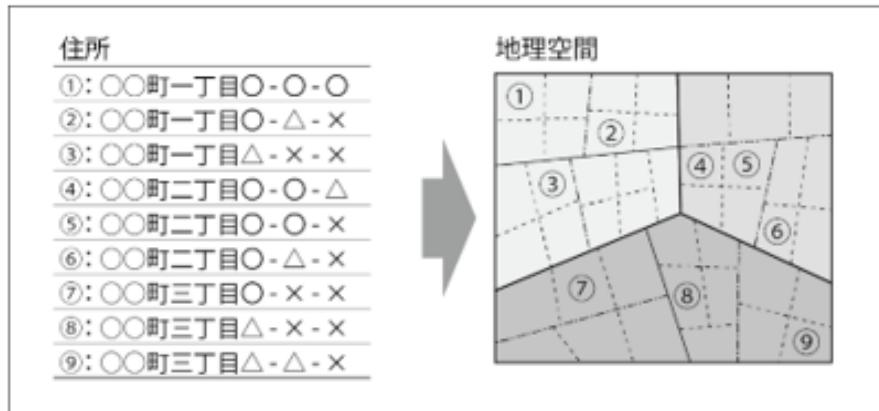


図 6.3 ジオコーディングのイメージ

地系が用いられている。日本測地系はベッセル橙円体に基づく測地系であり、旧国立天文台跡地を経緯度原点として定めている。一方で、世界測地系は人工衛星の計測に基づく、地球形状を高い精度で測量した測地系である。日本測地系は独自の測地系であるため、航空機の GPS による把握などの際、世界測地系と相対誤差が生じる。そのため、2002 年の測量法改正以降、国土地理院はわが国の測量基準を世界測地系としている。世界測地系の例としては、WGS84 や JGD2000 などがある。

座標の表現についても、いくつかの方法がある。代表的なものは緯度経度を度数で表す地理座標系である。これは、設定した測地系の橙円形上にある座標を表現しており、正確な位置が捕捉される。しかしながら、実務では位置・方向・距離等を平面上に投影して測量計算するほうが簡単に処理できる。これは平面直角座標とよばれ、現在でも公共測量で用いられている。投影法はガウス・クリューゲル図法を採用し、わが国では 19 の座標系が存在している。例えば、東京都は第 9 系に位置し、座標原点から東西約 130km が適用範囲である。

6.2.4 利用可能な不動産関連データの例

現在、わが国の基盤となる地図情報の多くは GIS データとして利用可能であり、分析者の用途に応じてデータを取得することができる。

国土地理院は主にわが国での基本的な地図情報を公開しており、道路、建物、行政界等を提供する基盤地図情報や、土地利用、ハザードマップ、都市施設などを幅広くカバーする国土数値情報がある。前者は全国で整備されており、都市計画区域内では縮尺 1/2,500 で、都市計画区域外では縮尺 1/25,000 でそれぞれ提供されている。有償の住宅地図が利用出来ない際、基盤地図情報をベースマップとして利用するのに優れている。後者は幅広いデータを提供しているため、分析者の関心に合わせてダウンロードするのが良い。不動産分析に関係するものとして、例えば地価公示ポイント、学校などの公共施設ポイント、用途地域ポリゴン、浸水想定区域ポリゴンなどがある。加えて、将来人口推計の 500m 及び 1,000m メッシュも提供されているため、将来時点でのシミュレーションを行う際に有用である。

総務省は小地域、メッシュ単位で境界データを提供しており、国勢調査や経済センサスなどと対応させることができる。国勢調査は人口構成や建物の建て方別世帯数などを公開しており、信頼性の高い結果を空間上で得ることができる。経済センサスでは産業別・従業者規模別全事業所数などがわかるため、例えば不動産価格と商業集積との関係を分析する際に利用を検討される。

近年、民間企業が構築しているデータにも注目が集まっている。2019 年 10 月現在、東京大学空間情報科学研究センターは JoRAS (Joint Research Application System) という共同研究利用システムを運営しており、その枠組みのなかで必要な民間企業提供データの利用が可能となっている。さらに、国立情報学研究所情報学研究データリポジトリでも共同研究を前提として民間企業データの提供が行われている。以下、不動産分析に関連するデータをいくつか挙げる。

ゼンリン住宅地図は道路、建物ポリゴンなどを現地調査から作成しており、基盤地図情報よりも精度の高いデータを得られる。さらに、建物ポリゴン内に用途、階数などの情報が格納されており、それらはヘドニック価格推定において重要な変数となる。

アットホーム株式会社は不動産取引データを JoRAS 経由で提供しており、住宅種別及び分譲・賃貸別にデータが整備されている。

国立情報学研究所経由で提供される LIFULL HOME'S データセットは、賃料、面積、築年数などの住宅特性だけでなく、高解像度の間取り図画像データも提供している。従って、画像データを利用して住宅特性を補間することが可能であり、画像認識分野との共同分析など今後の発展が見込まれる。

住宅・土地統計調査などの公的統計は GIS データとして直接利用可能ではないが、市区町村別で集計されている場合市区町村コードを GIS のポリゴンと対応させることで分析可能である。例えば、市町村単位での空き家率を地理空間上に投影し、分布の傾向や空間集積などを考察することができる。

6.3 空間集計における基本操作

6.3.1 基本量の測定

不動産分析を行う際、例えば対象物件から最寄り駅までの直線距離や建築面積を知りたい場合がある。この場合、GIS を用いることで瞬時に計算可能である。また、ある用途地域にかかる敷地の面積や、対象とする地域だけ取り出して分析したい場合など、空間的な重ね合わせが役に立つ。加えて、GIS ではある施設までの距離が最近隣となる領域などを求めることも可能である。本節では空間集計の操作について基本的な事項を述べる。

ベクターデータにはポイントデータに緯度経度座標が格納されており、2 点のポイントデータ $(x_i, y_i), (x_{i+1}, y_{i+1})$ が存在する場合、その間のユークリッド距離 (Euclidean distance) は

$$d_i = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \quad (6.1)$$

で定まる。複数の点が存在する場合、各二点間で距離測定を行えば良い。続いて、線分に囲まれた多角形の面積について求める。左記の二点に加えて $(x_{i+2}, y_{i+2}), (x_{i+3}, y_{i+3})$ も存在し、図 6.4 の線分の内側の面積 S を求めたいとする。この時、 S_1 の台形の面積は

$$S_i = \frac{1}{2}(x_{i+1} - x_i)(y_{i+1} + y_i) \quad (6.2)$$

と表せる。よって、台形 S の面積は

$$S = \frac{1}{2}\{(x_{i+1} - x_i)(y_{i+1} + y_i) + (x_{i+2} - x_{i+1})(y_{i+2} + y_{i+1}) - (x_{i+2} - x_{i+3})(y_{i+2} + y_{i+3}) - (x_{i+3} - x_i)(y_{i+3} + y_i)\} \quad (6.3)$$

によって求積可能である。

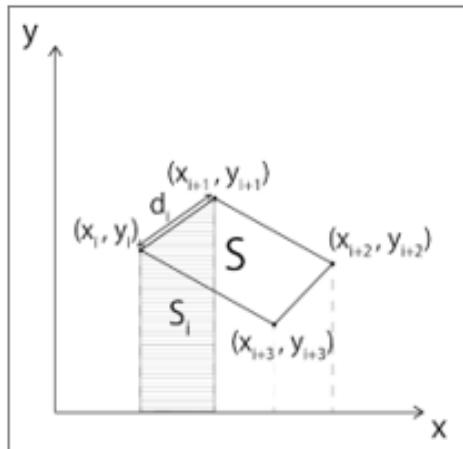


図 6.4 線分と面積の測定方法

このような方法により、対象物件から最寄り商店までの距離や建築面積などを計算できる。

6.3.2 ジオプロセシング

ジオプロセシング (geoprocessing) とは、GIS に関連するデータ処理を行い、新しいデータを出力する操作全般のことをいう。GIS では空間の領域生成や重ね合わせにより、分析者のニーズに応じた操作を行うことができる。領域生成には主にバッファー (buffer) 生成によるものやボロノイ分割 (Voronoi division) によるものなどがある。空間的重ね合わせには複数オブジェクトの足し引きで新たな領域を生成し、インターセクト (intersect), クリップ (clip), ユニオン (union) が主に利用される操作である。さらに、各ポリゴンのフィールドに基づく空間生成にもいくつか種類が存在し、マージ (merge), ディゾルブ (dissolve) が代表的な操作である。

任意のオブジェクトに対して、その近傍の領域を生成する操作をバッファーといい、ポイント、ライン、ポリゴンのそれぞれで生成可能である。例えば対象物件から 10 分圏内の領域を生成することができ、その領域内のコンビニエンスストア数の集計などに利用される。

ある点において近傍の領域を求める際に用いられるのが、ボロノイ図 (Voronoi diagram) である。ある母点 $i \in (1, 2, \dots, I)$ について、距離空間内の有限部分集合 $P = (p_1, p_2, \dots, p_I)$ は、

$$V(p_i) = \{p \mid d(p, p_i) \leq d(p, p_j), j \neq i\} \quad (6.4)$$

と表現できる。ただし、 $d(\cdot)$ は距離関数である。各ボロノイ領域 $V(p_i)$ の集合 $\{V(p_1), V(p_2), \dots, V(p_I)\}$ をボロノイ図と呼び、各母点からの近接する領域を知ることができる。さらに、各母点同士を結ぶことで描かれる三角形をドローネ図 (Delaunay diagram) と呼び、ボロノイ図と双対の関係にあることが知られている。このような図を作成することで、スーパー・マーケットの商圈や各駅の駅勢圏などを明らかにすることが出来る（図 6.5）。

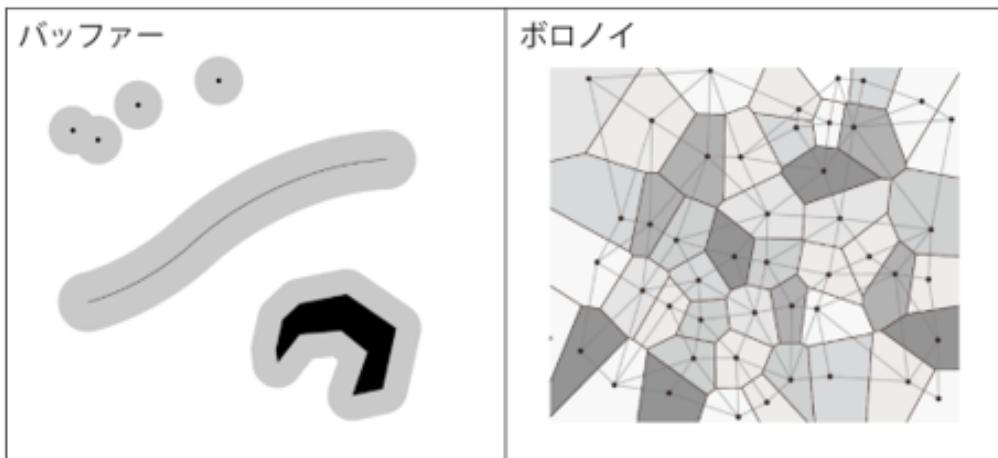


図 6.5 バッファーとボロノイによる空間的重ね合わせ

次に、空間的重ね合わせについて述べる。インターセクトはあるポリゴン A と B の積集合 ($A \cap B$) を求める操作である。生成されたポリゴンは A と B のどちらのフィールドも受け継がれる。一方、クリップはあるポリゴン A と B の積集合 ($A \cap B$) を求める操作であるが、インターセクトと異なり、

ある空間領域 B に入るポリゴン A を取り出す操作である。そのため、ポリゴン B の属性は生成されたポリゴンに含まれない。ユニオンはあるポリゴン A と B の和集合 ($A \cup B$) を求める操作である。生成されたポリゴンは空間領域で分割され、 A と B のどちらのフィールドも受け継がれる。例えば、複数市町村についてそれぞれ駅勢圏を分割した時にはユニオンを用いることで全ての情報が保たれて分割される。あるいは、 A 市のみに対して駅勢圏を抜き出したい場合には駅勢圏を A 市境界のポリゴンでクリップすると良い。

続いて、フィールドに基づく空間生成について述べる。マージはユニオンと同様、 A と B の和集合 ($A \cup B$) が出力されるが、同じフィールドを持つ空間データにより統合される。従って、生成されたポリゴンはマージに用いたフィールドに対してユニークに分割される。ディゾルブはフィールドに基づく空間データの集約であり、単体の空間データでも操作可能である。上記の演算イメージとテーブルの変化は図 6.6 のようにまとめられる。

	インターフェクト	クリップ	ユニオン																																				
空間演算 イメージ																																							
テーブル	<table border="1"> <thead> <tr> <th>FID</th><th>市町村名</th><th>用途地域</th><th>面積</th></tr> </thead> <tbody> <tr> <td>0</td><td>○○市</td><td>住居地域</td><td>10</td></tr> </tbody> </table>	FID	市町村名	用途地域	面積	0	○○市	住居地域	10	<table border="1"> <thead> <tr> <th>FID</th><th>市町村名</th><th>学区</th><th>面積</th></tr> </thead> <tbody> <tr> <td>0</td><td>○○市</td><td>A 学区</td><td>10</td></tr> </tbody> </table>	FID	市町村名	学区	面積	0	○○市	A 学区	10	<table border="1"> <thead> <tr> <th>FID</th><th>市町村名</th><th>用途地域</th><th>学区</th><th>面積</th></tr> </thead> <tbody> <tr> <td>0</td><td>○○市</td><td>住居地域</td><td>Null</td><td>8</td></tr> <tr> <td>1</td><td>○○市</td><td>Null</td><td>A 学区</td><td>8</td></tr> <tr> <td>2</td><td>○○市</td><td>住居地域</td><td>A 学区</td><td>2</td></tr> </tbody> </table>	FID	市町村名	用途地域	学区	面積	0	○○市	住居地域	Null	8	1	○○市	Null	A 学区	8	2	○○市	住居地域	A 学区	2
FID	市町村名	用途地域	面積																																				
0	○○市	住居地域	10																																				
FID	市町村名	学区	面積																																				
0	○○市	A 学区	10																																				
FID	市町村名	用途地域	学区	面積																																			
0	○○市	住居地域	Null	8																																			
1	○○市	Null	A 学区	8																																			
2	○○市	住居地域	A 学区	2																																			
	マージ		ディゾルブ																																				
空間演算 イメージ																																							
テーブル	<table border="1"> <thead> <tr> <th>FID</th><th>市町村名</th><th>用途地域</th><th>面積</th></tr> </thead> <tbody> <tr> <td>0</td><td>○○市</td><td>住居地域</td><td>10</td></tr> </tbody> </table>	FID	市町村名	用途地域	面積	0	○○市	住居地域	10	<table border="1"> <thead> <tr> <th>FID</th><th>市町村名</th><th>学区</th><th>面積</th></tr> </thead> <tbody> <tr> <td>0</td><td>○○市</td><td>A 学区</td><td>10</td></tr> </tbody> </table>	FID	市町村名	学区	面積	0	○○市	A 学区	10	<table border="1"> <thead> <tr> <th>FID</th><th>市町村名</th><th>用途地域</th><th>面積</th></tr> </thead> <tbody> <tr> <td>0</td><td>○○市</td><td>住居地域</td><td>5</td></tr> <tr> <td>1</td><td>○○市</td><td>住居地域</td><td>2</td></tr> <tr> <td>2</td><td>○○市</td><td>住居地域</td><td>3</td></tr> </tbody> </table>	FID	市町村名	用途地域	面積	0	○○市	住居地域	5	1	○○市	住居地域	2	2	○○市	住居地域	3				
FID	市町村名	用途地域	面積																																				
0	○○市	住居地域	10																																				
FID	市町村名	学区	面積																																				
0	○○市	A 学区	10																																				
FID	市町村名	用途地域	面積																																				
0	○○市	住居地域	5																																				
1	○○市	住居地域	2																																				
2	○○市	住居地域	3																																				
	<table border="1"> <thead> <tr> <th>FID</th><th>市町村名</th><th>用途地域</th><th>学区</th><th>面積</th></tr> </thead> <tbody> <tr> <td>0</td><td>○○市</td><td>住居地域</td><td>Null</td><td>10</td></tr> <tr> <td>1</td><td>○○市</td><td>Null</td><td>A 学区</td><td>10</td></tr> </tbody> </table>		FID	市町村名	用途地域	学区	面積	0	○○市	住居地域	Null	10	1	○○市	Null	A 学区	10	<table border="1"> <thead> <tr> <th>FID</th><th>市町村名</th><th>用途地域</th><th>面積</th></tr> </thead> <tbody> <tr> <td>0</td><td>○○市</td><td>住居地域</td><td>10</td></tr> </tbody> </table>	FID	市町村名	用途地域	面積	0	○○市	住居地域	10													
FID	市町村名	用途地域	学区	面積																																			
0	○○市	住居地域	Null	10																																			
1	○○市	Null	A 学区	10																																			
FID	市町村名	用途地域	面積																																				
0	○○市	住居地域	10																																				

図 6.6 ジオプロセシングの空間演算と出力テーブル

6.4 空間データの相関と補間

6.4.1 空間的自己相関

本節では、不動産分析に関連する空間解析手法について、空間解析特有の概念である空間的自己相関と空間補間にについて説明する。下記のような分析により、不動産市場における空間的構造の理解を深めるとともに、出力結果を用いてデータベースの充実を図ることも可能である。

ある空間領域について、興味のあるフィールドと相関があるかは分析を行ううえで重要である。例えば、市場滞留期間の長い不動産物件に集積があるかについて、客観的な判断が出来れば将来の投資判断につながると考えられる。以下、多くの既往論文で用いられている Moran's I を用いて概念について述べる。

一般に相関関係を求める際、二変数でよく用いられるのはピアソンの相関関数であり、以下の式で定義される。

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (6.5)$$

ただし、 \bar{x} \bar{y} はそれぞれ x と y の平均値を表す。

では、このような一般的な相関と空間相関との差異とは何であろうか。空間データは特定の座標上に観測値を持つため、空間上のどの位置を参照するかによって対応関係が異なる。特に、同一の観測値に関して空間的位置に起因する相関を空間的自己相関 (spatial autocorrelation) とよぶ。

空間的自己相関について、代表的に用いられている指標に Moran's I が存在し、以下の式で表される。

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6.6)$$

ただし、 $S_0 \equiv \sum_{i=1}^n \sum_{j=1}^n w_{ij}$ 、 w_{ij} は地域 i と j の距離に関する重み付け値であり、 $d(i, j)$ を i と j 間の距離とすると、 $d(i, j) > d(i, k)$ のとき、 $w_{ij} < w_{ik}$ と定義する。これは、観測値同士が空間的に近いものほど似通っているというトブラーの地理学の第一法則 (Tobler, 1970[15]) を反映したものといえる。なお、 $w_{ii} = 0$ とする。 w_{ij} の決め方としては、隣接している観測値を 1、そうでないものを 0 とおく場合や、 $d^{-\alpha}$ や $\exp(-\alpha d)$ などの距離減衰関数を設定して数値化する場合がある。上記の統計量は、大域的な空間的自己相関を判定するため、グローバルモラン統計量 (global Moran's I) とよばれている。グローバルモラン統計量が大きいときに正の自己相関であり、一方で小さいときには負の自己相関を示す (図 6.7)。

地域 i に着目してモラン統計量を算出する方法も存在する。さきほどの I について地域 i を固定すると、

$$I_i = \frac{n(y_i - \bar{y}) \sum_{j=1}^n w_{ij}(y_j - \bar{y})}{\sum_{j=1}^n (y_j - \bar{y})^2} \quad (6.7)$$

と書き下せる。これをローカルモラン統計量 (local Moran's I) とよび、近隣地域との類似性について表す (Anselin, 1995[1])。 I_i が正であるとき、地域 i は近隣と類似している一方で、負であるときには類似していない。なお、 $I = \sum_{i=1}^n I_i$ である。

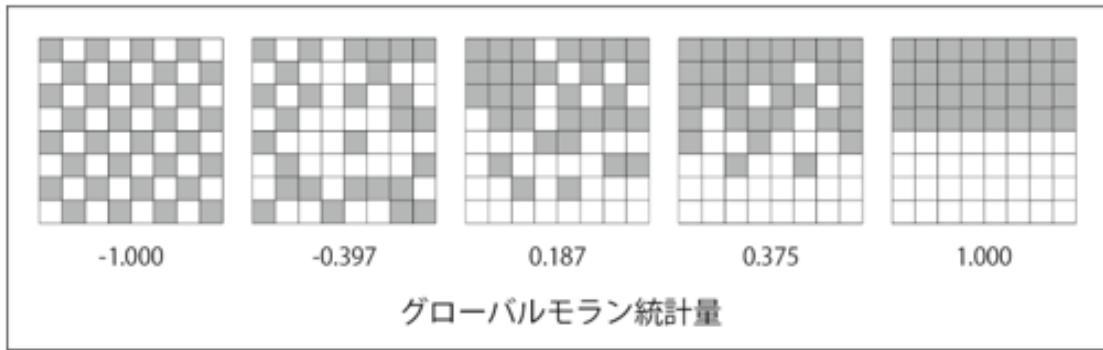


図 6.7 グローバルモラン統計量による空間的自己相関の図化

このような統計量を用いることで、空間的自己相関に関する分析を行うことができる。具体的には、各地域のローカルモラン統計量を計算し、対象地域と周辺地域との関係性について類型化し、それぞれの関係性について相関の程度を High と Low で表す。これにより作成される図をモラン散布図 (Moran scatterplot) といい、図 6.8 のように 4 分類して分析することが一般的である (Anselin, 1996[2])。例えば、図 6.8 の第一象限にある地域は High-High を示しており、自地域と周辺地域が共に高いローカルモラン統計量をもつような状態である。これはホットスポットと呼ばれ、例えば産業集積の立地や度合いを分析する際に利用される。一方で第三象限にある Low-Low に当たる地域はクールスポットとよばれる。不動産市場において、高経年マンションの立地傾向の分析や、高層マンションの集積度合いを把握する際に役立つ。空間的自己相関の扱いについて、その影響を考慮した空間統計学 (spatial statistics) という分野が発展しているが、紙面の都合上割愛する。詳細は瀬谷・堤 (2014) [11] を参照されたい。

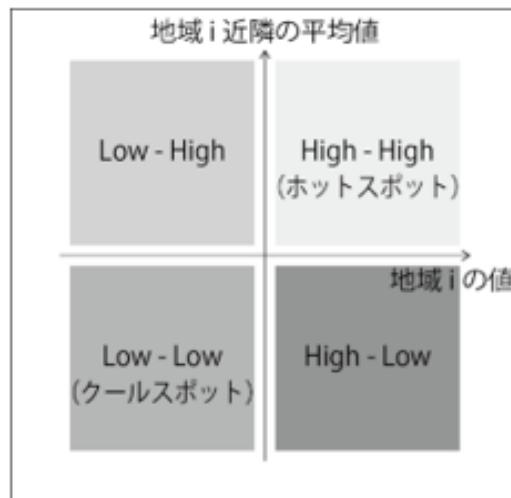


図 6.8 モラン散布図

6.4.2 空間補間

ここまで空間的な相関関係についてみてきたが、そもそも近隣の観測値が不明である場合、それを補間する方法がある。例えば、ある地点の地価を知りたい時に近隣の地価からある程度の推定が可能である。空間補間には大きく補間点近傍の観測値を用いる場合と、大域的な観測値を求める場合の2通り存在する。本節では、比較的容易に補間可能な前者について述べる。なお、空間補間に関する全般的な説明は Lam (1983)[7]において詳述されている。

空間補間のなかで最も簡便なものが最近隣法 (nearest neighbor method) である。この方法では、任意のデータに対して、最近隣の観測点を当てはめる手法である。従って、観測点を母点としたボロノイ分割によって、任意の点に対する最近隣のデータ点を求めることができる。最近隣法は推定が比較的簡単である一方で、各ボロノイ領域は離散的な値をとるため、ボロノイ境界付近での値に誤差が生じる場合がある。すなわち、最近隣法による空間補間は観測点が十分密である場合には有効であるが、観測点間に大きなギャップが存在する場合や観測値の変動が大きい場合、良好な精度を保証できるとは限らない。

続いて、不整形三角網 (TIN: Triangulated Irregular Network) から近傍を求める手法について述べる。これは、何らかの方法で不整形三角網を構築し、各三角形の頂点の観測値から内部にある任意のデータを補間するものである。この時、一般的に利用されるのが、ボロノイ図を描画する際に登場したドローネ三角網である。ドローネ三角網は各三角形の最小角を最大化するように作成するため、より正三角形に近い三角網を構築可能である。補間推定値を求める際には、 xy 座標と観測値 z について、 $A(x_1, y_1, z_1), B(x_2, y_2, z_2), C(x_3, y_3, z_3)$ の三つの頂点からなる平面を考える。平面上の任意の点を $P(x, y, z^*)$ とすると、ベクトル \vec{AP} が \vec{AB} と \vec{AC} で書き表せる。すなわち、 $\vec{AP} \vec{AB} \vec{AC}$ が一次従属であることと同値である。従って、行列式を用いて、

$$\begin{vmatrix} x - x_1 & x_2 - x_1 & x_3 - x_1 \\ y - y_1 & y_2 - y_1 & y_3 - y_1 \\ z^* - z_1 & z_2 - z_1 & z_3 - z_1 \end{vmatrix} = 0 \quad (6.8)$$

を解くことで補間推定値 z^* を求めることができる。これは地形表現などでよく利用され、可能な限りコンパクトな観測値の組から補間値を得られる一方で、最近隣法と同様に表面が平滑化されない。

最後に、一定の仮定をおいて連続的な表面を作成する方法について述べる。代表的なものが逆距離加重法 (IDW: Inverse Distance Weighted) である。これは、空間的自己相関と同様に地理学の第一法則を踏まえたものである。いま、観測点数 n 、補間点 z^* と観測点 z_i との距離を d_i 、距離に応じた重み付け関数を $w(\cdot)$ とする。このとき、補間推定値は

$$z^* = \frac{\sum_{i=1}^n z_i w(d_i)}{\sum_{i=1}^n w(d_i)} \quad (6.9)$$

で表される。なお、 $w(\cdot)$ は通常距離で減衰し、空間的自己相関のように距離に関して $d^{-\alpha}$ などの仮定をおき、数値化する。この式の分母は距離による重みの和を基準化するための項であり、分子で距離的な重みに基づき観測値を足し合わせている。逆距離加重法の利点は直観的に自然な連続表面を作成することであり、補間推定値は連続値として求められる。一方で、真の空間分布が連続的に変化しない場合や、設定する距離関数が適切でない場合、真値とうまくフィットしない。この方法では距離関数やパラメータの精

度を確認する必要があり、主に交差検証が用いられる。これは、観測値の一部を評価用にしてパラメータ推定に利用せず、推定補間値が評価用観測値とどの程度乖離しているか検証するものである。評価には、観測値と補間値の差分二乗値を平均化する二乗平均誤差（MSE: Mean Square Error）や、その平方根をとる二乗平均平方根誤差（RMSE: Root Mean Square Error）などが用いられる。以下、図 6.9 に空間補間のイメージをまとめた。

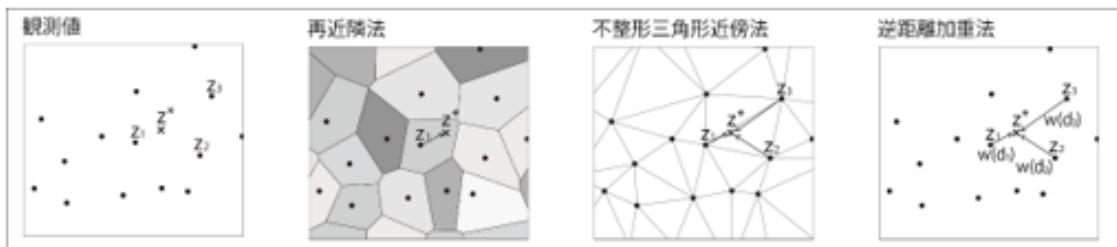


図 6.9 空間補間のイメージ

6.5 空間特性に配慮した不動産価格構造の推定^{*1}

空間的な相違が不動産価格の影響を与えることは容易に連想できる。しかし、より厳密に考えれば、不動産は土地と建物から構成されていることを考えれば、空間的な相違は、建物価格に影響を及ぼすことはないが、土地価格に対して影響を与えるということを考えないといけない。このように考えれば、最寄り駅までの距離や都心までの接近性といった特性も、建物価格には影響を及ぼさないが、土地価格には強く影響を与える。一方、建物の建築後年数の増加は、建物価格に影響を及ぼすが、土地価格には影響を及ぼすことではない。このように考えると、不動産価格の予測においては、土地と建物を区分して推計していくかなければならないことがわかる。

通常、不動産価格情報を用いたヘドニック関数を推計する場合には、5章で紹介したように土地と建物が一体となった不動産価格を対象として分析されることが多い。一般的には、不動産価格は、土地価格と建物価格によって構成されるものであるし、2章で紹介したヘドニック・アプローチに基づき、関数を推計していく。一方、土地と建物を区分して価格構造を考えようとした場合には、生産関数から不動産価格を考えることになる。生産者、つまりビルダーは、土地と建物がそれぞれ投入し、不動産を生産する。そのように生産された不動産から派生するサービスの対価として価格が形成される。このような生産関数から導かれる不動産価格の測定方法が、ビルダーズモデルと呼ばれる手法である。

ビルダーズモデルでは、土地の価格と建物の価格を足したものが、住宅価格として形成されることを仮定する。この手法は、推計の複雑性という問題を持つものの、理論的な優位性を持つ。今、不動産価格について考えてみる。建物が完成した後の価格は、建物の延べ床面積 (S) × 単位面積当たりの建築費 (β_t) と土地の面積 (L) × 単位面積当たりのコスト (α_t) に等しい。ここでは β_t, α_t は時間ごとに変化する係数であるとする。今、取引期 t において、延べ床面積 S_{tn} 、土地面積 L_{tn} で価格が V_{tn} であるような不動産を考える（ただし、 $t = 1, \dots, T$, $n = 1, \dots, N(t)$ であり、 $N(t)$ は時点 t におけるサンプル数を表している）。この時、これらの価格が土地と建物価格の総和に誤差項 (ε_{tn}) を加えたものに等しいとする

(ただし, ε_{it} は互いに独立な正規分布に従う) すると, 取引期 t におけるパラメータ α_t と β_t は次式のようなヘドニック回帰モデルとして表現できる。

$$V_{tn} = \alpha_t L_{tn} + \beta_t S_{tn} + \varepsilon_{tn} \quad (6.10)$$

(6.10) は取引期 t , 物件 n における土地面積 L_{tn} と建物の延べ床面積 S_{tn} という測定量と, 時点 t における土地の平米単価 α_t と建築コストの平米単価 β_t という一定品質の価格から成り立っている。そのため, (6.10) によって定義されるヘドニック価格モデルは, 新築の場合に相当している。

一般的に古い物件の場合には, 経年減価によって新築物件よりも価格が安くなる。そこで, 物件 n の取引期 t における建築後年数 $A(t, n)$ が分かっており, 幾何学的な減価モデルを仮定することで, より現実的なヘドニック回帰モデルとして, (6.11) のようにビルダーズモデルを考えることができる。

$$V_{tn} = \alpha_t L_{tn} + \beta_t (1 - \delta)^{A(t, n)} S_{tn} + \epsilon_{tn} \quad (6.11)$$

ここで, (6.11) 中のパラメータ δ は正味の経年減価率を表している。築後年数による幾何学的な減価を取り入れたことにより, (6.10) では単純な線形回帰モデルであったが, (6.11) では非線形回帰モデルになっている。

また, (6.11) で定義されたヘドニック回帰モデルには, 多重共線性という重大な問題がある。(6.11) で定義されるようなヘドニック回帰モデルでは, 土地面積と建物面積の間に多重共線性があるため, 土地価格と建物価格を同時に正確に推計することが不可能であると経験的に知られている。そこでこの問題を避けるために, 新築建物の価格の初期値として国土交通省が公表している 1 平米あたりの建築コスト P_{St} に延床面積をかけたものを用いる。すると経年減価を取り入れた新しいビルダーズモデルは結果的に, (6.11) の β_t を建築コスト P_{St} で置き換えることで

$$V_{tn} = \alpha_t L_{tn} + P_{St} (1 - \delta)^{A(t, n)} S_{tn} + \epsilon_{tn} \quad (6.12)$$

と表せる。これをベーシックなビルダーズモデルとする。

実際には (6.12) に, 様々な物件属性を追加して最終的なモデルとする。特に従来では, 土地価格に対する空間的な効果として区ダミーをモデルに取り入れることが多かったが, 本章では双線型スプライン補間近似を用いる。この理由として, 区ダミーだと隣り合った区にかかわらず境界で大きな差が発生してしまうが, 現実的にそのようなことは考えにくい。本章のような補間近似を行うことで, 空間的な土地価格の変化をなだらかな変化として取り入れることが可能となる。詳細については次節以降で詳しく説明していく。

6.6 空間構造の取り扱い

6.6.1 単位正方形における双線型補間

この節では, ベーシックなビルダーズモデルの初項である土地価格部分に, より一般的な空間構造を取り入れるための手法である双線型補間にについての説明を行う。この補間方法は、単位平方上で定義される

2変数関数を近似する基本的な方法である。まず $f(x, y)$ を $x(0 \leq x \leq 1), y(0 \leq y \leq 1)$ の2変数からなる連続的な関数とし、 f は単位平方の四隅で γ_{ij} の値、すなわち

$$\gamma_{00} \equiv f(0, 0); \gamma_{10} \equiv f(1, 0); \gamma_{01} \equiv f(0, 1); \gamma_{11} \equiv f(1, 1) \quad (6.13)$$

をとるとする。このとき、関数の四隅の高さを推定できると仮定すると、単位正方形の四隅にある (6.13) を満たし、単位正方形の境界を構成する4つの線分に沿った線形関数であるような近似関数を探す。Colwell(1998)[4] ではそれらの条件を満たす x, y の2次関数 $g(x, y)$ として

$$g(x, y) \equiv \gamma_{00}(1-x)(1-y) + \gamma_{10}x(1-y) + \gamma_{01}(1-x)y + \gamma_{11}xy \quad (6.14)$$

が示されており、 $g(x, y)$ が $\gamma_{00}, \gamma_{10}, \gamma_{01}, \gamma_{11}$ の加重平均であることも示されている。そのことを踏まえて、(6.14) の特性をより理解するために (6.15) のように

$$g(x, y) = \gamma_{00} + (\gamma_{10} - \gamma_{00})x + (\gamma_{01} - \gamma_{00})y + [(\gamma_{00} + \gamma_{11}) - (\gamma_{01} + \gamma_{10})]xy \quad (6.15)$$

と書き換える。すると、もしも $\gamma_{00} + \gamma_{11} = \gamma_{01} + \gamma_{10}$ であれば $g(x, y)$ は線形関数であることがわかる。しかしながら、 $\gamma_{00} + \gamma_{11} \neq \gamma_{01} + \gamma_{10}$ であれば $g(x, y)$ はサドル関数、すなわち $\nabla^2 g(x, y) = -[(\gamma_{00} + \gamma_{11}) - (\gamma_{01} + \gamma_{10})]^2 < 0$ であるから、 $\nabla^2 g(x, y)$ は正の固有値と負の固有値を1つずつもつことがわかる。

本節では単位正方形で定義された関数 $g(x, y)$ を例に考えたが、次節ではより一般的な例として、グリッド上の連続関数を定義する方法について説明する。

6.6.2 グリッドにおける双線型スプライン補間

正方格子で Colwell(1998)[4] の手法がどのような働きをするか説明するために、 3×3 の格子の場合について考える。この方法は、 X, Y 空間の長方形領域上で定義される変数 X および Y に適用され、 X と Y は次式 (6.16) のように

$$X_{\min} \leq X \leq X_{\max}, Y_{\min} \leq Y \leq Y_{\max} \quad (6.16)$$

で定義されているとし、 X, Y を0から3の範囲に収まるように、スケーリングした x と y を次式 (6.17) のように定義する。

$$x \equiv 3(X - X_{\min}) / (X_{\max} - X_{\min}), y \equiv 3(Y - Y_{\min}) / (Y_{\max} - Y_{\min}) \quad (6.17)$$

ここで、 x に関する3つのダミー変数関数を次のように定義する。

$$\begin{cases} D_1(x) \equiv 1 & \text{if } 0 \leq x < 1; \\ D_2(x) \equiv 1 & \text{if } 1 \leq x < 2; \\ D_3(x) \equiv 1 & \text{if } 2 \leq x \leq 3; \end{cases} \quad \begin{cases} D_1(x) \equiv 0 & \text{if } x \geq 1 \\ D_2(x) \equiv 0 & \text{if } x < 1 \text{ or } x \geq 2 \\ D_3(x) \equiv 0 & \text{if } x < 1 \end{cases} \quad (6.18)$$

従って、 $0 \leq x \leq 3$ であれば、それぞれのダミー変数の和は1になることがわかる。また、 y に関する3つのダミー変数関数も (6.18) と同様に定義できる。そして、 $3 \times 3 = 9$ 個の x, y に関するダミー変数関数 $D_{ij}(x, y)$ を (6.19) のように

$$D_{ij}(x, y) \equiv D_i(x)D_j(y), \quad i = 1, 2, 3; j = 1, 2, 3 \quad (6.19)$$

と定義する。このとき、 $D_{ij}(x, y)$ の定義域は、長さ 3 の各辺を持つ 2 次元空間の正方形 $S_3 \equiv \{(x, y) : 0 \leq x \leq 3; 0 \leq y \leq 3\}$ である。そして、 S_3 に属する (x, y) は、 $\sum_{i=1}^3 \sum_{j=1}^3 D_{ij}(x, y) = 1$ であることから、 $D_{ij}(x, y)$ は、 S_3 を構成する 9 つの単位正方形セルのいずれかに $(x, y) \in S_3$ を割り当てる。 $D_{ij}(x, y) = 1$ となる x, y に対応する領域のセルを C_{ij} で示すと、例えば、 $0 \leq y < 1$ を満たす y 値に対応する 9 つのセルのグリッド内の 3 つのセルは、 C_{11}, C_{21}, C_{31} になる。同様に、 $1 \leq y < 2$ では、 C_{12}, C_{22}, C_{32} 、 $2 \leq y \leq 3$ では、 C_{13}, C_{23}, C_{33} である。

いま、 $f(x, y)$ を、近似したい S_3 で定義された関数だとする。そして次のように、単位面積セルのグリッドの 16 頂点で関数 $f(x, y)$ の高さ γ_{ij} を (6.20) のように定義する。

$$\gamma_{ij} \equiv f(i, j), \quad i = 0, 1, 2, 3; j = 0, 1, 2, 3 \quad (6.20)$$

このとき、任意の $(x, y) \in S_3$ において、Colwell(1998)[4] の双線形スプライン補間近似 $g_3(x, y)$ は $f(x, y)$ に対して (6.21) のように定義される。

$$\begin{aligned} g_3(x, y) \equiv & + D_{11}(x, y) [\phi_{00}(1-x)(1-y) + \phi_{10}(x-0)(1-y) + \phi_{01}(1-x)(y-0) + \phi_{11}xy] \\ & + D_{21}(x, y) [\phi_{10}(2-x)(1-y) + \phi_{20}(x-1)(1-y) + \phi_{11}(2-x)(y-0) + \phi_{21}xy] \\ & + D_{31}(x, y) [\phi_{20}(3-x)(1-y) + \phi_{30}(x-2)(1-y) + \phi_{21}(3-x)(y-0) + \phi_{31}xy] \\ & + D_{12}(x, y) [\phi_{01}(1-x)(2-y) + \phi_{11}(x-0)(2-y) + \phi_{02}(1-x)(y-1) + \phi_{12}xy] \\ & + D_{22}(x, y) [\phi_{11}(2-x)(2-y) + \phi_{21}(x-1)(2-y) + \phi_{12}(2-x)(y-1) + \phi_{22}xy] \\ & + D_{32}(x, y) [\phi_{21}(3-x)(2-y) + \phi_{31}(x-2)(2-y) + \phi_{22}(3-x)(y-1) + \phi_{32}xy] \\ & + D_{13}(x, y) [\phi_{02}(1-x)(3-y) + \phi_{12}(x-0)(3-y) + \phi_{03}(1-x)(y-2) + \phi_{13}xy] \\ & + D_{23}(x, y) [\phi_{12}(2-x)(3-y) + \phi_{22}(x-1)(3-y) + \phi_{13}(2-x)(y-2) + \phi_{23}xy] \\ & + D_{33}(x, y) [\phi_{22}(3-x)(3-y) + \phi_{32}(x-2)(3-y) + \phi_{23}(3-x)(y-2) + \phi_{33}xy] \end{aligned} \quad (6.21)$$

(x, y) がグリッドの各頂点である場合、 $g_3(x, y)$ は S_3 上の x と y の連続関数であり、 $g_3(x, y)$ は基になる関数 $f(x, y)$ と等しい。つまり、 S_3 の 16 の頂点に対して (6.22) のような等式

$$g_3(i, j) = \gamma_{ij} \equiv f(i, j), \quad i = 0, 1, 2, 3; j = 0, 1, 2, 3 \quad (6.22)$$

が成り立つ。つまり、グリッドの単位面積の各正方形について、 $g_3(x, y)$ は (6.14) で定義された双線形補間関数 $g(x, y)$ のように振る舞うことがわかる。

Colwell(1998)[4] に従って $y = j; j = 0, 1, 2, 3$ を仮定した場合、 x の関数 $g_3(x, j)$ は、 $0 \leq x \leq 3$ の線形スプライン関数になる。つまり、 $g_3(x, j)$ は、 $x = 1$ および $x = 2$ で勾配を変更できる 3 つの線形区間を持つ x の連続的な区分線形関数である。同様に、 $x = i; i = 0, 1, 2, 3$ を仮定すると、 y の関数 $g_3(i, y)$ は、 $0 \leq y \leq 3$ の y の線形スプライン関数でもある。したがって、 $g_3(x, y)$ は、 x および y 方向のこれらの線形スプライン関数を 2 つの変数の一貫した連続関数にマージする補間関数として見ることができる。

以上より、Poirier(1976)[9] および Colwell(1998)[4] に従って、(6.21) で定義された補間モデルから計量経済推定モデルに移行することができる。そのために 3 つの仮定、(1) N 個の観測に対して x と y を観測できる、(2) $n = 1, \dots, N$ に対して $f(x_n, y_n)$ も観測できる、(3) S_3 上で関数 $f(x, y)$ を $g_3(x, y)$

で近似できる、が成り立つとする。このとき、 $\gamma \equiv [\gamma_{00}, \gamma_{10}, \dots, \gamma_{33}]$ を (6.21) に現れる 16 個の γ_{ij} ベクトルとし、 $g_3(x, y)$ を $g_3(x, y, \gamma)$ に書き換え、 γ を次式の線形回帰モデルに現れるパラメーターのベクトルとして表示する。

$$z_n = g_3(x_n, y_n, \gamma) + \varepsilon_n, \quad n = 1, \dots, N \quad (6.23)$$

近似誤差 ε_n が平均 0、一定の分散で独立分布していると仮定する場合、(6.23) の未知のパラメーター γ_{ij} （頂点の「真の」関数 $f(x, y)$ の高さ）は、最小二乗回帰によって推定できる。境界のあるセットに 2 次元の表面をフィットさせるこの方法は、本質的にノンパラメトリックな方法であり、観測値の数 N が十分に大きく、観測値がグリッド上でほぼ均一に分布している場合、グリッドをより細かくして、真の基になる関数により近い近似を得ることができる。

表面の推定に対するこのノンパラメトリックアプローチが地理的領域の土地区画の販売においてどのように適用できるかを見るために、ある取引期において、 N 個の土地区画の販売価格に関する情報があると仮定し、不動産の面積が L_n 平方メートルであるような土地区画 n の販売価格が P_n であるとする。この時、 $n = 1, \dots, N$ の土地区画 n の緯度経度 X_n, Y_n のデータがあれば、定義式 (6.23) および (6.17) を使用して、これらの空間座標を変数 x_n, y_n に変換およびスケーリングすることができる。そして N は十分に大きく、観測値は 3×3 の地理的グリッドの 9 つのセルすべてに分散しているのであれば、次の線形回帰モデルを推定することにより、対象としている地理的領域の真の土地価格の近似値を得ることができる。

$$\frac{P_n}{L_n} = g_3(x_n, y_n, \gamma) + \varepsilon_n, \quad n = 1, \dots, N \quad (6.24)$$

(6.24)において $g_3(x_n, y_n, \gamma)$ は、観測サンプルの各 (x_n, y_n) に対して (6.21) によって定義される。したがって、(6.21) の 16 の未知の高さパラメーター γ_{ij} の推定値は、単純な最小二乗最小化問題を解くことによって得られる。

観測サンプルが豊富な場合、グリッドをより細かくすることができる。つまり、 3×3 グリッドを $k \times k$ に置き換えることができる。ここで、 k は任意の正の整数である。この場合、(6.17) は $x \equiv k(X - X_{\min}) / (X_{\max} - X_{\min})$ および $y \equiv k(Y - Y_{\min}) / (Y_{\max} - Y_{\min})$ に置き換えられる。

6.7 実データを用いた推計例

今回分析で使用するデータは、5 章で利用したデータと同様である。内訳としては、戸建住宅 5,580 件に加え、土地 8,493 件であり、モデル構築のための変数として V ：価格 (1000 万円)、 S ：延床面積 ($100m^2$)、 L ：土地面積 ($100m^2$)、 A ：築後年数、 NB ：部屋数、 W ：前面道路幅員 (0.1m)、 TW ：最寄駅徒歩分数、 TT ：最寄駅から東京駅までの所要時間、 X ：経度、 Y ：緯度、 $P_s : m^2$ あたりの建築コスト (10 万円) を用いる。また、取引期 t として 2000 年 1Q～2010 年 4Q までの 44 期あるため、 $t = 1, \dots, 44$ である。そしてこれら変数の記述統計量を表 6.1 にまとめる。

表 6.1 をみるとわかるように、平均販売価格が約 6,250 万円、平均延床面積が $43.5m^2$ 、平均土地面積が $103.9m^2$ 、平均築後年数は 5.8 年 (土地を除くと平均 14.7 年)、平均部屋数は 3.95、平均前面道路幅員は 4.7m、最寄駅までの平均徒歩時間は 9.4 分、最寄り駅から東京までの平均所要時間は 31.2 分である。

表 6.1 使用データの記述統計量

	N	MEAN	ST.DEV	VAR	MIN	MAX
V	14,045	6253.70	2902.30	8423500	1800.00	20000.00
S	14,045	0.43	0.58	0.34	0.00	2.48
L	14,045	1.04	0.40	0.16	0.50	2.50
A	14,045	5.82	9.12	83.17	0.00	49.72
NB	14,045	1.57	2.04	4.17	0.00	8.00
W	14,045	46.81	12.53	157.02	25.00	90.00
TW	14,045	9.37	4.30	18.53	1.00	29.00
TT	14,045	31.25	7.39	54.66	8.00	48.00
P _s	14,045	17.73	0.29	0.09	17.30	18.50

6.7.1 ビルダーズモデルの一般化

(6.12) で示したビルダーズモデル立地などを全く考慮していない、最もベーシックなものである。そこで、(6.23) で示した空間構造の補間関数や、その他の説明変数を組み込むことで、より一般的なモデルへ拡張する。今回は最終的なモデルとして

$$\begin{aligned}
 V_{tn} = & \alpha [D_{S,tn} + \phi D_{L,tn}] g_k(x_{tn}, y_{tn}, \gamma) f_L(L_{tn}, \lambda) [1 + \tau(TW_{tn} - 1)] \\
 & \times [1 + \rho(TT_{tn} - 8)] [1 + \sigma(W_{tn} - 25)] + P_{St} (1 - \delta)^{A(t,n)} \\
 & \times f_S(S_{tn}, \mu) \left[\sum_{i=2}^6 \kappa_i D_{NB,tn,i} \right] + \epsilon_{tn}
 \end{aligned} \tag{6.25}$$

を用いる。以下で (6.25) に新たに表れた項について説明する。

まず、 $D_{L,tn}$ は

$$D_{L,tn} \equiv \begin{cases} 1 & \text{時点 } t \text{ における観測 } n \text{ のレコードが土地のみである場合} \\ 0 & \text{それ以外} \end{cases} \tag{6.26}$$

で定義されている。(6.26) を見るとわかるように、時点 t 、物件 n が土地データのみのデータであるかどうかを分けるダミー変数関数である。そのため、(6.25) において $D_{S,tn} = 1 - D_{L,tn}$ である。

次に、 $f_L(tn, \lambda)$ は (6.27) のように定義される。

$$\begin{aligned}
 f_L(L_{tn}, \lambda) = & D_{L,tn,1} [\lambda_0 L_0 + \lambda_1 (L_{tn} - L_0)] \\
 & + D_{L,tn,2} [\lambda_0 L_0 + \lambda_1 (L_{tn} - L_0) + \lambda_2 (L_{tn} - L_1)] \\
 & + D_{L,tn,3} [\lambda_0 L_0 + \lambda_1 (L_{tn} - L_0) + \lambda_2 (L_{tn} - L_1) + \lambda_3 (L_{tn} - L_2)] \\
 & + D_{L,tn,4} [\lambda_0 L_0 + \lambda_1 (L_{tn} - L_0) + \lambda_2 (L_{tn} - L_1) + \lambda_3 (L_{tn} - L_2) \\
 & + \lambda_4 (L_{tn} - L_3)]
 \end{aligned} \tag{6.27}$$

この項は土地面積の価格に対する影響を単純に線形として取り入れるのではなく、4 つの土地面積帯別に線形区分関数としてモデルに組み込んでいる。面積区分は $L_0 = 0.5, L_1 = 1, L_2 = 1.5, L_3 = 2$ によっ

て分けられている。また、 $D_{L,tn,m}$ は取引期 t 、物件 n がどの土地面積帯に所属するかのダミー変数関数である。

次に、 $f_S(S_{tn}, \mu)$ は次式のように定義される。

$$\begin{aligned} f_S(S_{tn}, \mu) &\equiv D_{S,tn,1} [\mu_0 S_0 + \mu_1 (S_{tn} - S_0)] \\ &+ D_{S,tn,2} [\mu_0 S_0 + \mu_1 (S_1 - S_0) + \mu_2 (S_{tn} - S_1)] \\ &+ D_{S,tn,3} [\mu_0 S_0 + \mu_1 (S_1 - S_0) + \mu_2 (S_2 - S_1) + \mu_3 (S_{tn} - S_2)] \end{aligned} \quad (6.28)$$

(6.27) は土地面積に関する線形区分関数であったが、(6.28) は建物面積に関する線形区分関数を表す項であり、 $S_0 = 0.5, S_1 = 1, S_2 = 1.5$ によって 3 つの面積区分に分けられている。

最後に物件の部屋数に関するダミー変数関数 $D_{NB,tn,i}$ が (6.29) のように

$$D_{NB,tn,i} \equiv \begin{cases} 1 & \text{時点 } t \text{ における観測 } n \text{ の部屋数が } i \text{ である} \\ 0 & \text{それ以外} \end{cases} \quad (6.29)$$

として定義される。

6.8 推計結果

本節では、(6.25) のモデルを用いて推計したいくつかの結果を示す。今回の計算では、グリッド数としては $k = 7$ を用いる。この際、海だけのセルも存在するが、このようなセルでは当然ながら、データが存在していない。また、サンプル数が少なく、推計結果が安定しない頂点の値も 0 として計算を行った。

また、Diewert and Shimizu(2017)[5] では全時系列に対してのみ推計を行っていたが、本節では (I)2000-2002, (II)2004-2006, (III)2008-2010 の 3 時点それぞれについても推計を行った。その結果を表 6.2 に示す。

まず表 6.2 をみるとわかるように、決定係数はどの期間においても 0.8 を超えており、説明力が高いモデルであることがわかる。また、最寄駅までの徒歩分数の係数と最寄駅から東京駅までの所要時間の係数は負値になっている。これは最寄り駅から遠くなれば遠くなるほど、東京駅から離れれば離れるほど価格が安くなることを意味しており、一般的な感覚とも一致する。それ以外の係数は正值であり、階数が高くなればなるほど、部屋の数が多くなればなるほど価格が高くなるということを示している。

さらに、各時点の係数を比較するとわかるように、大きく変わらないものの若干異なっている。このことから、各係数は年々ずつと一緒ではなく、時点によって異なることがわかる。

次に空間効果の時点変化を図 6.10 に示す。図 6.10 上段は各期間における格子点の値を、下段は期間 I と II, II と III, I と III の比を示している。まず上段を見るとわかるように、土地価格は 23 区中心部ほど高く、離れるほど徐々に安くなる傾向がどの時点においても見て取れる。次に下段に注目すると、II と III の比較では、端の格子点で変化が大きいため特徴を捉えづらくなっているが、期間 I と III を比較した図をみるとわかるように、全体的に土地価格が高くなっていることがわかる。従来のビルダーズモデルでは、空間効果をモデルに取り入れるために、本書 5 章のように区ダミーを用いていた。そのため、区を境として大きなギャップが生じるという問題があった。しかしながら、本章のように空間効果を取り扱うことで、なだらかに空間構造の変化を捉えることが可能となる。

6.9 不動産市場分析の発展可能性

GIS に代表される空間解析技術や空間に紐づけられた新しいデータ資源は、不動産市場分析を大きく発展させる可能性がある。

不動産市場分析に有用な GIS 的処理は、データベース構築のための空間演算から高度な解析まで多岐にわたる。本章では GIS の基本的な処理及び分析について多くの時間を割いたが、それぞれの概念は相互に関連していることがわかる。例えば空間補間を行う際、再近隣法ではボロノイ領域の概念を利用し、逆距離加重法の重み付け関数は空間的自己相関のものと類似した考え方である。従って、一度空間解析の基礎を身に付ければ、様々な応用に派生可能といえる。

これまでに紹介した分析は、実際には GIS に関連するソフトウェアにより計算されるため、その習熟も重要になる。一般的に利用されるのは ESRI 社の ArcGIS やフリーソフトウェアである QGIS である。

表 6.2 ビルダーズモデルの推計結果

	2000-2010		2000-2002		2004-2006		2008-2010	
	coef.	std. err.						
ϕ	1.126	0.010	1.287	0.025	1.124	0.020	0.998	0.014
λ_0	1.508	0.037	1.301	0.068	1.344	0.056	1.866	0.077
λ_2	1.134	0.033	1.064	0.060	1.070	0.051	1.155	0.065
λ_3	1.212	0.036	1.094	0.064	1.210	0.056	1.256	0.088
λ_4	0.984	0.067	1.230	0.124	0.930	0.101	0.730	0.185
τ	-0.013	0.000	-0.013	0.001	-0.012	0.001	-0.012	0.001
ρ	-0.007	0.000	-0.007	0.001	-0.006	0.001	-0.007	0.001
σ	0.004	0.000	0.005	0.000	0.004	0.000	0.003	0.000
μ_0	0.923	0.132	1.675	0.207	0.946	0.261	-0.137	0.295
μ_1	1.344	0.168	1.137	0.245	0.786	0.288	2.034	0.388
μ_2	2.359	0.168	1.790	0.213	2.541	0.377	2.421	0.396
μ_3	1.767	0.175	1.138	0.199	1.153	0.314	3.957	0.505
δ	0.041	0.002	0.031	0.003	0.043	0.005	0.061	0.004
κ_2	1.101	0.057	1.092	0.079	1.227	0.145	1.168	0.110
κ_3	1.028	0.052	1.111	0.079	1.187	0.138	0.952	0.086
κ_4	0.867	0.047	1.005	0.076	0.847	0.104	0.872	0.085
κ_5	0.733	0.042	0.947	0.075	0.837	0.104	0.527	0.071
Spatial Structure	Yes							
Time Dummy	Yes							
R^2	0.849		0.844		0.865		0.836	
sample	14,045		4,646		4,429		3,558	

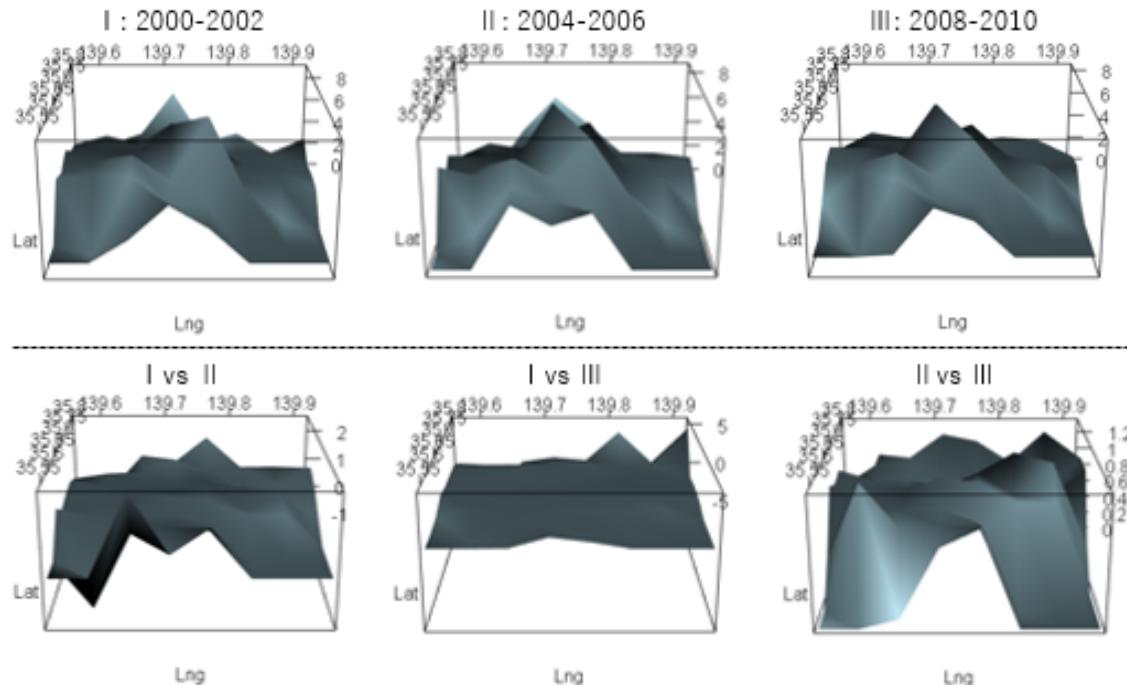


図 6.10 空間効果の時点変化

さらに、統計プログラミング言語である R でもいくらかの処理を行うことができ、例えば spdep パッケージを利用すればドローネ三角網の図化や空間的自己相関の確認などが可能である。

近年になり、不動産分析における地理的特性や対象物件周辺のアメニティが重要視されており（例えば Shimizu, 2014[13]），今後さらに GIS を用いた情報の取得が重要になってくると考えられる。このような潮流に乗るため、浅見ら（2015）[3] や貞広・山田・石井（2018）[10] なども参考にすると理解が深まる。なお、空間回帰分析や空間相互作用モデルなど、やや発展的な解析も不動産分析を行ううえで重要である。深く学びたい場合は Longley et al. (2005)[8]などを参照すると良い。

また、不動産市場分析でしばしば利用されている線形回帰モデルやニューラルネットワークなどの適用では、土地建物一体で価格の推計を行っていることが多い。しかし、前述のように、経年減価は建物価格にしか影響しないし、立地条件の多くは土地価格にしか影響を与えないというように、土地建物一体で分析するのは不適切である。また、本書の 5 章では推計を行う際の空間効果として区ダミーを用いていたため、区の境界で土地効果が急激に変動してしまうが、実際にはその変化は考えづらい。そこで本章では、土地建物を分離して推計する手法であるビルダーズモデルに、空間構造の補間効果を取り入れたモデルで推計を行った。これによって、土地建物それぞれのみに寄与する効果を分離することが可能となり、また土地価格の空間的な効果をなだらかに表現することが可能となった。

このような分析が可能となるだけでなく、モバイル空間統計に代表されるように、リアルタイム性を持った位置情報に紐づけられた新しいデータ資源が登場してきている。また、7 章で紹介するように、GIS を用いた新しいエリア指標が開発されるなど、技術・データ資源の両面において発展が著しい領域である。不動産市場分析を行うものは、GIS の技術の習得も併せて行うとともに、新しいデータ資源の登場

にも関心を持ち続けないといけないであろう。

参考文献

- [1] Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical analysis*, 27(2), 93-115.
- [2] Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess lo-cal instability in spatial association. *Spatial Analytical Perspectives on GIS*, 111-125.
- [3] 浅見泰司・矢野桂司・貞広幸雄・湯田ミノリ (2015) 地理情報科学 GIS スタンダード. 古今書院.
- [4] Colwell, P. F. (1998), “A Primer on Piecewise Parabolic Multiple Regression Analysis via Estimations of Chicago CBD Land Prices”, *Journal of Real Estate Finance and Economics* , 17:1, 87-97.
- [5] Diewert, W.E. and C. Shimizu (2017), “Alternative Approaches to Commercial Property Price Indexes for Tokyo”, *Review of Income and Wealth* , 63, 492-519.
- [6] Diewert, W.E. and C. Shimizu (2020), “Residential Property Price Indexes: Spatial Coordinates versus Neighbourhood dummy Variables”, *Discussion Paper, Vancouver School of Economics, The University of British Columbia* , 20-01.
- [7] Lam, N. S. N. (1983). Spatial interpolation methods: a review. *The American Cartographer*, 10(2), 129-150.
- [8] Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographic information systems and science*. John Wiley & Sons.
- [9] Poirier, D.J. (1976), “The Econometrics of Structural Change”, *Amsterdam: North-Holland Publishing Company*.
- [10] 貞広幸雄・山田育穂・石井儀光 (2018) 空間解析入門 都市を測る・都市がわかる. 朝倉書店..
- [11] 濱谷創・堤盛人 (2014) 空間統計学: 自然科学から人文・社会科学まで. 朝倉書店.
- [12] 清水千弘 (2004) ,『不動産市場分析』住宅新報社.
- [13] Shimizu, C. (2014). Estimation of Hedonic single-family house price function considering neighborhood effect variables. *Sustainability*, 6(5), 2946-2960.
- [14] Shimizu, C., H. Takatsuji, H. Ono, and K.G. Nishimura (2010), “Structural and Temporal Changes in the Housing Market and Hedonic Housing Price Indices: The Case of the Previously Owned Condominium Market in the Tokyo Metropolitan Area”,*International Journal of Housing Markets and Analysis* , 3(4), 351-368.
- [15] Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Eco-*

nomic geography, 46(sup1), 234-240.

第7章

GIS を用いたエリア指標の開発

7.1 エリア指標と不動産テック

消費者の住宅選択において、土地及び建物だけでなく、その住宅を取り巻く周辺の環境水準は、極めて重要になってきている。人々は、広義のエンターテイメントに基づく「アメニティ」によってもたらされる文化的・自然的消費の機会を重視するようになることで、住宅選び=エリア選びといった方向へと、多くの国において、エリアの価値が重視されるようになってきている。米国では、2007年から「Walk Score」と呼ばれる指標が開発され、そして、不動産情報サイトによって公開されるようになってきたことから、またそのようなスコアが不動産価格の予測モデルの中の特徴量として利用されていることから、不動産テック分野においても高く注目される研究領域になってきている。

このようにエリア指標が注目されるようになった理由には、かつての労働集約型の企業が大部分を占めていた経済構造から、情報と知識集約型産業が主となる形へとシフトし、人々の生活において余暇を楽しむ機会が増えたことが挙げられる (Fogel (2000)[2], Glaeser, Kolko and Saiz (2004)[3])。このような傾向は、「働き方改革」が推し進められる、今後の日本において一層強くなっていくものと考える。

都市の集積のメカニズムの変化は、都市の役割を「生産のための場」から「消費のための場」へとシフトさせた (Glaeser, Kolko, and Saiz (2004)[3])。いわゆる、経済学でいう生産関数から消費関数へ、企業から家計へと主役が変化し、都市の民主化が進み、「Consumer City Theory」が発展してきたのである。

そのような中で、GIS情報を用いて、暮らしやすさの観点から不動産の立地環境を表す指標「日本版WalkScore(仮称)」を開発した。具体的には、株式会社ゼンリン保有の各種施設ポイントデータ、ネットワークデータからデータベースを構築し、個々の場所とその場所から徒歩でアクセス可能な周辺アメニティとを紐づけ、周辺アメニティの充実度からその場所の暮らしやすさを評価したスコアを算出した。

住宅向け、オフィス向けといった「用途別スコア」に加え、家族世帯向け、単身世帯向け、高齢者向けなどの「タイプ別スコア」など、多様なニーズに応じたスコア提供へと発展させることが可能である。また、スコアは現状で50m メッシュごとに算出が可能であり、街区レベルに匹敵する詳細な立地環境の把握が可能である。

これにより、同一駅周辺の複数の物件についても周辺環境を一目で比較できるようになる。さらには、ヒートマップ上から立地環境の優れたエリアを見定めたり、レーダーチャートの形状から類似した立地環境の物件を提案したり、といった幅広いサービス展開を行うことができる。

本章では、エリア指標の開発におけるデータ資源と技術について解説したい。

7.2 不動産の価値評価

7.2.1 不動産価値評価の現状

不動産価値の評価手法は種々あるが、近年、Real Quality Rating (RQR) という評価指標が注目されている。現状、不動産の価格 (market price) については多くの情報にアクセスが可能であるが、それらの不動産の質 (market quality) についての情報を入手することは容易ではない。RQR は、不動産の立地環境情報や建物自体の情報、建物内についての情報から総合的に不動産を評価することでその品質を明らかにし、投資を行う上での判断基準となる指標として 2017 年に開発されたものである。

近年開発された不動産価値評価の指標としては、RQR 以外にも Well Building Standard (2012-)[9] や Fitwel (2017-)[1] などがあるが、いずれも建物自体の評価にとどまらず、建物を利用する人の快適性や健康、ウェルビーイングといった観点から評価項目が選定されているという特徴がある。我が国においても、日本政策投資銀行が創設した DBJ Green Building 認証 (2014-)[12]、CASBEE-ウェルネスオフィス (2019-)[11] で同様の視点が取り入れられており、利用者目線を考慮した不動産価値評価の考え方が世界的な潮流となりつつある。

表 7.1 近年開発された不動産価値評価の指標（例）

名称	Well Building Standard	DBJ GREEN BUILDING 認証制度	Real Quality Rating	Fitwel	CASBEE-ウェルネス オフィス
発表年／国	2012年／アメリカ	2014年／日本	2017年／フランス	2017年／アメリカ	2019年／日本
目的	建物使用者の健康、快適性、知能に影響を与える問題に対処すること	環境・社会への配慮がなされた不動産の評価を通じ、事業者と金融機関・投資家の架け橋となること	不動産の質 (market quality) を明らかにし、投資判断の基準となる指標となること	建物の設計・運用面の改善を通じて個人やコミュニティの健康を支えること	オフィスワーカーが知的生産性向上を健康な状態で実現すること
評価内容	<11カテゴリー> - Air - Water - Nourishment - Light - Movement - Thermal Comfort - Sound - Materials - Mind - Community - Innovation	<5カテゴリー> - Energy &Resources - Amenity - Resilience - Community &Diversity - Partnership	<3カテゴリー> - Location - Built structure - Workspace	<12カテゴリー> - Location - Building Access - Outdoor Spaces - Entrances + Ground Floor - Stairwells - Indoor Environment - Workspaces - Shared Spaces - Water Supply - Cafeterias + Prepared Food Retail - Vending Machines + Snack Bars - Emergency Procedures	<5カテゴリー> - 健康・快適性 - 利便性 - 安全・安心 - 準備・管理 - プログラム

7.2.2 立地評価

不動産価値評価指標について、特に立地の評価の考え方にも近年変化が見られる。米国では、2007年から「Walk Score」という指標が開発されサービス提供されている。Walk Scoreは徒歩での生活のしやすさを表す指標であり、任意の住所に対してその周辺に「Dining&Drinking(飲食店)」、「Shopping(買い物)」、「Parks(公園)」、「Schools(学校)」などの都市アメニティがどれだけ充実しているかを算出し100点満点のスコアを提供している。徒歩でアクセスできるアメニティが多いほど生活しやすい場所という評価がなされ、このスコアを見るだけでその物件の周辺環境の良し悪しを把握することができる。WalkScoreは、米国の大手不動産ポータルサイトに掲載されており、物件探しをしている人が気になる物件の詳細ページを開いた際に、家賃や間取りなどの情報に併記される形でスコアが表示されており、物件検索条件の一要素として定着しつつある。また、米国を中心に普及している不動産価値評価指標「Fitwel」においては、このスコアが立地評価における採点基準としても取り入れられている。

一方、我が国においてはこのような一般向けの不動産立地環境評価の指標は確立されていない。しかし、私たちが住む家を探す際には、コンビニやスーパーが近くに充実しているかどうかや、小さな子供がいる世帯であれば子供が遊べるような公園が近くにあるかどうかなどは当然のように関心を持っている。また、オフィスの立地についても、ランチで行くレストラン・弁当屋や夜の居酒屋などが充実しているか



図 7.1 米国 WalkScore ウェブサイトのトップページ (<https://www.walkscore.com>)

どうかはオフィスワーカーにとって有益な情報となりうる。オフィスの中でも、建物内に飲食店が併設されていることが多い大規模オフィスとは異なり、中小規模のオフィスの場合には、建物の周辺にどれだけ利便施設が充実しているかという情報は相対的にニーズが大きいと考えられる。投資家目線では、同じ中小規模のオフィスであっても、周辺施設がより充実しているものほど市場価値が高いため、そちらにより多くの投資を行うという判断を行うことができる。国全体で不動産ストックの老朽化が大きな課題となっている我が国において、優れた立地ポテンシャルを持ち市場価値の高いストックを峻別し再投資を行う上で、このような客観的な判断基準が整備されることには大きな意義がある。

7.3 「日本版 WalkScore (仮称)」の開発

7.3.1 データ資源とデータベース構築

従来、不動産価値評価では、主に家賃、最寄り駅までの距離、建物・設備スペックについての情報が用いられてきた一方で、不動産周辺環境については分かりやすい客観的な評価指標が整備されてこなかった。

本取組みは公共及び民間が保有する GIS 情報を用いて、上述したような我が国の社会課題解決に向けて、不動産立地環境に関する新たな評価指標「日本版 WalkScore (仮称)」の開発を行うものである。

「日本版 WalkScore (仮称)」は暮らしやすさの観点から、不動産の立地環境（周辺の都市アメニティ充実度）を表す指標である。全国の市街化区域を対象として、不動産とそこから徒歩でアクセス可能なアメニティ群（スーパー、コンビニ、公園、飲食店、カフェなど）のデータを紐づけ、アメニティ分類ごとの周辺立地数をもとにその充実度を 100 点満点でスコア化するものである。

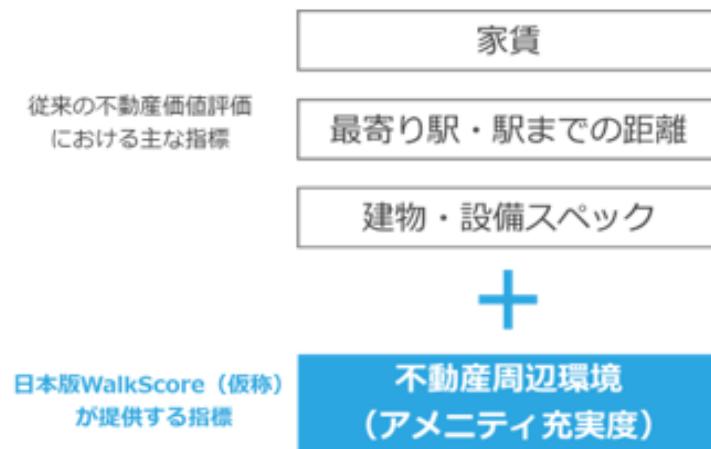


図 7.2 従来の不動産価値評価における主な指標と日本版 WalkScore (仮称) が提供する指標

(データ資源)

主に使用したデータは、様々な都市アメニティの位置情報を有したポイントデータと、徒歩での所要時間を算出するための経路情報であるネットワークデータとに分けられる。いずれも株式会社ゼンリンの

データを使用しており、前者はテレポイント Pack!データ、建物ポイントデータ及び POI データ、後者は主に歩行者ネットワークデータを用いている。

テレポイント Pack!データではアメニティ業種が 2000 超まで細分化されており、その詳細な業種ごとに個々のアメニティの位置情報が入手できる。コンビニ等の業種に関しては個別の企業ブランド名（例：セブンイレブン、ローソン、ファミリーマートなど）まで補足できるなど、非常に詳細なアメニティ分類となっている。

また、テレポイント Pack!データ、建物ポイントデータ、POI データそれぞれに、データの収集・作成の過程で、アメニティ業種ごとの捕捉率に差異が生じていることがある。しかし、業種ごとに各データを突き合わせて比較することで、より実態に即した適切なデータを採用している。

スコアを集計する際には、各種分析を通じて、それら膨大なアメニティ業種の中から、不動産価値評価においてより重要度が高いと思われるアメニティ項目を選定して使用している。当然ながら、様々なニーズに合わせ、スコア集計時に採用するアメニティ業種をカスタマイズしていくことも可能である。実際に、目的別スコア、タイプ別スコアはそれぞれの用途・タイプごとに選好されるアメニティ業種を考慮し、スコアごとに集計対象のアメニティ業種を変化させている。

(データベース構築)

日本全国の市街化区域を 50m メッシュで分割し、その 1 つ 1 つのメッシュに対してスコアを算出している。スコア算出に際し、まず各 50m メッシュを起点として徒歩で到達可能な範囲として「徒歩圏」を設定し、その「徒歩圏」内に立地するアメニティを特定する。その上で、その 50m メッシュと「徒歩圏」内アメニティとを紐づける。この操作を全ての 50m メッシュに対して行うことで、日本全国の任意の場所とそこから徒歩でアクセス可能なアメニティに関するデータベースを構築することができる。

50m という距離は、徒歩 1 分で 80m の距離を進むとした場合、徒歩で約 40 秒程度の距離である。50m メッシュという地理的に非常に細かい単位でスコア集計をしていることにより、より正確な周辺環境の評価が可能となる。例えば、同じ駅であっても、駅の西側は昔ながらの商店街が続く商業エリア、駅の東側は大学キャンパスが広がる文教エリア、などというように、駅の出口ごとに街の特徴が大きく異なる場合は少なくない。このような場合、立地環境の評価を駅という単位で行ってしまうと、当然ながら実態を上手く反映した評価ができないということは想像に難くない。駅の東西、南北でエリアの特徴が大きく異なるような場合でも、50m メッシュ単位で評価を行うことでスコアにその差異を十分に反映させることができる。

これにより、同じ東京駅周辺であっても、皇居側の丸の内エリアと、日本橋側の八重洲エリアの差異を正しく捕捉することができ、さらに言えば、同じ丸の内エリアの中でもより周辺アメニティが充実している場所とそうでない場所とを把握することができる。丸の内エリアという業務中心地にオフィスを構えたいたいと考える企業は多く存在するが、丸の内エリアの中でも周辺環境が異なり、従業員の満足度を左右し得るという点から見れば、50m メッシュでのスコア提供はオフィス立地選択における新たな価値提供にもつながると考えられる。

また、不動産から徒歩でアクセス可能なアメニティを集計する際に、単純なポイント間の直線距離ではなく、歩行者ネットワークデータを使用した徒歩経路距離を用いている点も重要である。例えば、横断で

きる地点が限られている幹線道路や大規模な施設が存在するエリアでは、直線距離と比べて徒歩経路距離がより長くなっていることが考えられ、直線距離を用いて周辺のアメニティを定義した場合には実態とのかい離が大きくなってしまう。しかし、今回は徒歩経路を用いることで、そのような実態との乖離をなくしている。また、大規模な公園内の歩道なども反映されており、実状を正確に反映した周辺アメニティの捕捉が可能になっている。

なお、各 50m メッシュを起点とした「徒歩圏」は移動距離に応じて複数設定しており、のちのスコア算出の際に距離減衰を考慮した重み付けを行っている。各アメニティ数は、「徒歩圏」ごとの距離に応じた重み付けを行う場合には、下記のように集計できる。

$$AM_i = \sum_j a_j \cdot AM_{ij} \quad (7.1)$$

AM_i : アメニティ数

SM : スーパーマーケット

CS : コンビニエンスストア

DS : 薬局

PF : 公共施設

⋮

a_j : 各「徒歩圏」の距離に応じた低減係数 ($j = 1, \dots, J$)

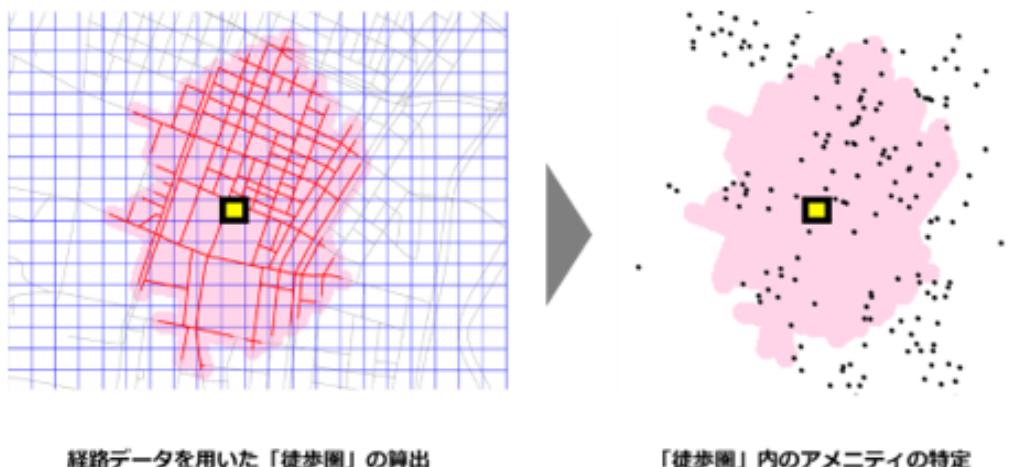


図 7.3 経路データを用いた「徒歩圏」の算出と「徒歩圏」内のアメニティの特定のイメージ

7.3.2 日本版 WalkScore の計算

以上のように構築されたデータベースを用いて、エリア別指標を計算していく。

(キーアメニティの抽出)

アメニティの業種数については、テレポイントデータの特性上、最多で2000超を区別することが可能である。しかし、この中には、不動産の立地環境の評価において必ずしも大きな影響を与えない業種も含まれているため、それらの業種を省き、より重要度の高い業種（キーアメニティ）のみを抽出してスコア算出に用いている。

キーアメニティの抽出には、ヘドニック理論と呼ばれる手法を活用する。

Rosen(1974)[6]によって提案されたヘドニックモデルは、差別化された生産物の市場均衡理論を発展させ、住宅のような財をどのように分析することができるのかを、経済理論と計量経済モデルの両面から示した。具体的には、商品供給者のオファー関数（offer function）、商品需要者の付け値関数（bid function）およびヘドニック価格関数の構造との間の関係を厳密に検討し、市場価格を消費者および生産者の行動から特徴づけている。そうすると、ヘドニック理論を援用することで、住宅の選択者がどのような要因に基づき、そして、どの程度のウェイトをもってその要因を重視して住宅選択をしているのかを定量的に把握することができる。

そこで、下記の3つの手続きによって、アメニティ業種の抽出作業を行っている。第1段階として、地価分析を行っている。商業・業務地と住宅地とを区別した上で、各アメニティの立地数と地価との関係性を分析し、地価に対して有意な影響を与えてアメニティを採用している。第2段階としてデモグラフィック分析を行っている。各アメニティ立地数と人口動態との関係性を分析し、様々な社会的属性を持つ人々が選好していると考えられるアメニティを採用している。第3段階として、既往の類似指標調査等を行っている。上記のような分析に加え、既存の類似指標や不動産ポータルサイト等で用いられているアメニティを参照し、重要度が高いと考えられるアメニティを採用している。

地価分析に際しては、ヘドニックアプローチを用いている。被説明変数として地価、説明変数として、用途地域、実行容積率等の都市計画条件およびスーパー、コンビニ、公園、飲食店等のアメニティ充実度を用いている。

一つの例を挙げれば、一般的なヘドニック理論に基づく地価関数は、下記のように推計されることが多い。

$$\log P = a_0 + \sum_h a_{1h} X_h + \sum_j a_{2j} Z_j + \sum_k a_{3k} \cdot LD_k + \sum_l a_{4l} \cdot RD_l + \sum_m a_{5m} \cdot TD_m + \varepsilon \quad (7.2)$$

DP/GA	: 戸建て住宅価格 (円/m ²)
X_h	: Main variables
GA	: 土地面積
FS	: 建物面積
RW	: 前面道路幅員
Age	: 建築後年数
TS	: 最寄り駅までの時間
TT	: 都心までの時間
Z_j	: Other variables
ZD	: 土地利用規制:容積率・建ぺい率・用途規制
BC	: その他の要因:南向きダミー等
LD_k	: Location(Ward) Dummy ($k = 0, \dots, K$)
RD_l	: Railway Dummy ($l = 0, \dots, L$)
TD_m	: Time Dummy ($m = 0, \dots, M$)

そのような関数に、アメニティや地域環境を追加していく。

$$\begin{aligned} \log P = & a_0 + \sum_h a_{1h} X_h + \sum_j a_{2j} Z_j + \sum_k a_{3k} \cdot LD_k + \sum_l a_{4l} \cdot RD_l + \sum_m a_{5m} \cdot TD_m \\ & + a_6 Dm_{(l \leq LA < m)} + a_7 Dm_{(m \leq LA)} + a_8 (LA)(Dm_{(l \leq LA < m)}) + a_9 (LA)(Dm_{(m \leq LA)}) \\ & + \sum_i a_{10i} \log V_i + \sum_{h,i} a_{11h,i} X_h \cdot V_i + a_{12u} u + a_{13v} v + \varepsilon \end{aligned} \quad (7.3)$$

V_i	: Neighborhood Effects
AM	: アメニティ
CS	: 世帯特性:国勢調査
u, v	: longitude, latitude

(スコアの算出と可視化)

データベースからキーアメニティの立地数を集計し、最終的に 50m メッシュごとに 100 点満点のスコアを算出している。なお、スコア集計に際し、同一アメニティについての効用遞減、歩行経路距離による効用遞減を考慮している。

現時点では、「用途別スコア」と「タイプ別スコア」の 2 種類を開発している。「用途別スコア」は、住宅向け (for Residence), オフィス向け (for Office) のスコアを想定し、用途に応じたスコア算出手法を用い、住宅用途、オフィス用途それぞれに対応している。住宅向けについては、さらに「タイプ別スコア」として、家族世帯向け (for Family), 単身世帯向け (for Family), 高齢者向け (for Elderly) を想定し、住む人に応じたスコア算出手法を用い、各タイプに対応している。

本スコアは、不動産を探す希望エリアの 1 次スクリーニングツールとして活用することが想定される。スコアをヒートマップとして表示することで、例えば東京 23 区全体を俯瞰して、どのエリアが特にスコ

アが高いかを直感的に把握することができる。鉄道駅周辺や主要路線沿線一体が赤く表示され、特にアメニティ充実度が高いことが見て取れる。逆に、大きな河川沿いや大規模な公園の付近では、川や公園によって「徒歩圏」内のアメニティ数が比較的少なくなってしまい、アメニティ充実度の観点からは比較的低いスコアとなっている。また、駅から離れたエリアでも、大規模な商業施設や商店街が近いエリアでは、アメニティが充実し、スコアが高い傾向になっている。

通常、不動産を探す場合には駅からの近さを重視する傾向があるが、駅から多少離れていても周辺環境が優れた不動産をきちんと評価、可視化することで、不動産の価値をより正確に伝えることが可能になる。

また、レーダーチャート表示で、アメニティ分野別のスコアを表現することで、その不動産の立地環境の特徴を一目で把握できるようになる。また、レーダーチャートの形状により、まちの類型化が可能になり、第一希望のエリアは家賃面で断念せざるを得ないという顧客に対し、それと類似した形状のレーダーチャートを持つエリアを代替候補として提案するなど、顧客ニーズに合ったエリアの提案を客観的なデータをもとに効果的に行えるようになる。

これにより、不動産の需要者と供給者の間の最適なマッチングを促進し、わが国の不動産市場の活性化を促進することも可能となる。

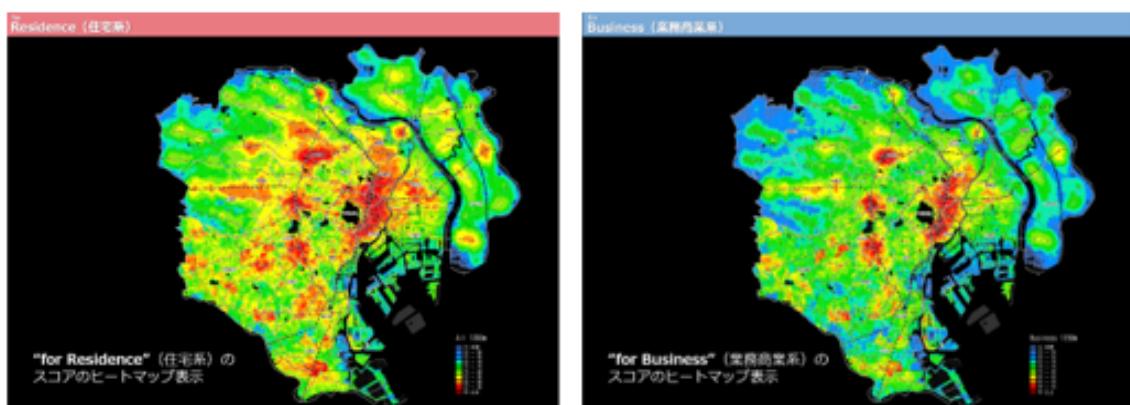


図 7.4 日本版 WalkScore (仮称) のスコア可視化イメージ (ヒートマップ表示の例)

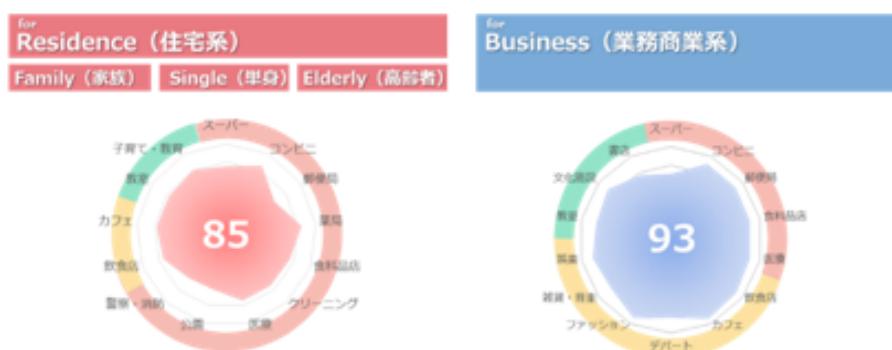


図 7.5 日本版 WalkScore (仮称) のスコア可視化イメージ (レーダーチャート表示の例)



図 7.6 日本版 WalkScore (仮称) のスコアを用いた街比較の例

7.4 日本版 WalkScore 研究の発展可能性

「日本版 WalkScore (仮称)」に代表される地域指標の開発は、今後、大きな研究分野として成長していくことが予想される。

今後も急速に進む高齢化もまた住宅を取り巻くアメニティとの関係に大きな変化をもたらす。働く時間、通勤する時間の縮小と働き方が大きく変化すれば、または労働から解放された人々が増加していく中では、住宅およびそれを取り巻く地域での過ごす時間の質と密度が大きく変化していく。休息をとるためだけの家であれば、職場と近接した場所に住まうことで通勤時間を節約し、住宅から受けるサービスの水準を上昇させることができた。そのため、従来の住宅選択では、「最寄り駅や都心までの距離」や「大きさ」、各種性能といった「住宅」の物理的機能だけにしか関心がなく、そのような機能だけによって価格差が生まれてきた。

しかし、住宅を中心とした歩行可能空間での消費活動が活発化することで、つまり家で過ごす時間が長くなることで、多様な消費ができる地域に人々が集まる傾向は強くなっていくであろう。

しかし、ここで重要なのが、「Scene:シーン」という概念である。Shimizu et al (2014)[7]では、アメニティの種類をシーンと呼んだ。これは、Clark 教授が主導した国際比較プロジェクトの中で出てきた概念である。そして、そのシーンは、それぞれの街を見る個人によって異なる評価が存在することを意味している。例えば、バー やカラオケが集積している街のシーンを見た時に、ある主体はワクワクとするようなエキサイティングをする場合もあれば、違う主体では嫌悪感を覚える場合もある。緑や公園を見て、心が穏やかになる人もいれば、寂しくなる人もいる。また、そのシーンも一日の中での時間や一年の中での季節によっても変化していく。

Walk Scoreなどの定量的な指標は、そのような意味では平均的な街の顔を写像しているだけであり、街の特性を十分に踏まえた万能な指標ではない。歩行可能な空間が魅力的な街になりうる場合もあれば、それが人によっては違う街の顔として映る場合もある。

街づくりや都市計画を進めるにあたり、誰の、どの視点から街を眺め、評価していくべきかという観点

か。都市が縮退し、高齢化が進展する中で、多数決原理では最適な解は見つけることは出来ず、多様な主体からの見え方を重視していかなければならない。将来を見据えた時には、子供や若者などの目から見えるシーンを大切にしなければならないはずである。色のついていないできる限り純粋な目をもって、街を眺めていくことが重要になってきていると考える。

それを実現していくためには、道路傾斜等も考慮した徒歩経路の設定や、時間帯を考慮したスコア集計（飲食店の営業時間等を考慮したナイトライフスコアなど）などを組み込んでいくことも考えられる。このように、日本版 WalkScore（仮称）には様々な発展可能性がある。現在、日本版 WalkScore（仮称）と同様、前述の RQR についても日本に適した形で開発を進めている。これらの指標開発及び実用化を通じて、不動産立地環境情報の見える化を実現し、不動産に関する情報の非対称性を解消し、不動産とユーザーとのマッチングを推進することが、我が国における不動産市場のさらなる活性化につながるものと考える。

今まで可視化できていない不動産にかかる情報を、テクノロジーの進化によって実現できるようになってきているのである。

参考文献

- [1] Fitwel ウェブサイト (最終閲覧日 : 2019 年 12 月 5 日)
<https://www.fitwel.org>
- [2] Fogel,R.W. (2000), "The Fourth Great Awakening the Future of Egalitarianism", University of Chicago Press.
- [3] Glaeser, E., Kolko, J.K., Saiz, A. (2004), "Consumers and Cities, The City as an Entertainment Machine", Research in Urban Policy 9, Elsevier, 177-184.
- [4] Navarro, C. J., Mateos, C. and Rodriguez, M.J. (2012) Cultural scenes, the creative class and development in Spanish municipalities, European Urban and Regional Studies, 21: 301-317
- [5] Real Quality Rating ウェブサイト (最終閲覧日 : 2019 年 12 月 5 日)
<http://rqr-global.com/>
- [6] Rosen, S.(1974), "Hedonic Prices and Implicit Markets, Product Differentiation in Pure Competition," *Journal of Political Economy*, Vol.82, pp34-55.
- [7] Shimizu, C., S. Yasumoto, Y. Asami and T. N. Clark (2014), "Do Urban Amenities drive Housing Rent? ", CSIS Discussion Paper: (The University of Tokyo), No.131.
- [8] Silver, D., T. N. Clark and C. J. Navarro. (2010) Scenes: Social Context in an Age of Contingency, Social Forces 88 (5): 2293-2324
- [9] The International WELL Building Institute ウェブサイト (最終閲覧日 : 2019 年 12 月 5 日)
<https://www.wellcertified.com>
- [10] Walk Score ウェブサイト (最終閲覧日 : 2019 年 12 月 5 日)
<https://www.walkscore.com>
- [11] 建築環境・省エネルギー機構ウェブサイト, "CASBEE ウェルネスオフィス評価認証" (最終閲覧日 : 2019 年 12 月 5 日)
http://www.ibec.or.jp/CASBEE/certification/WO_certification.html
- [12] 日本政策投資銀行ウェブサイト, "DBJ Green Building 認証" (最終閲覧日 : 2019 年 12 月 5 日)
https://www.dbj.jp/service/finance/g_building/

第8章

不動産間取り図の認識と応用^{*1}

8.1 市場探索行動における不動産間取り図

不動産市場で情報を探索するものにとって、間取り情報は極めて重要である。第1章で整理しているように、買い手となる消費者は、自分に最も適した情報を求めて不動産市場を探索する。探索する情報としては、7章で紹介したエリア情報と併せて、物件、そして部屋特有の情報を探索することとなる。そのなかで、部屋の特徴を示す間取りは、最も重要な情報の一つとして位置付けることができる。不動産間取り図とは、不動産物件においてその間取りを簡潔に表現した図である。間取り図は、人々が物件を評価する際に非常に有用な情報となる。しかし、間取り図は規格化がなされておらず、その作成者や作成方法が様々であるためイラストのスタイルやカラーリング、文字フォントに至るまで表記ゆれが激しくなっている。その様子を図8.1に示す。このような表記ゆれが、間取り図を画像として直接認識・処理することを困難にする一因である。

間取り画像を認識し、構造化することができれば様々な応用が期待される。例えば、不動産物件検索システムへの応用が挙げられる。既存の物件検索システムは、賃料や立地、築年数などに対しては詳細な検索ができる一方、間取りに対してはワンルームや2LDKといったいわゆる部屋のタイプでしか検索できない。これは、「リビングは何帖以上は欲しい」、「水回りが集中している利便性の高い配置がいい」といった間取りに対するあらゆる希望を持っているユーザーのニーズを満たしていない。このようなニーズに沿った検索を可能にするには、間取り図の内容を計算機に認識させ、それらに対して適切にマッチングを行う必要がある。

そこで本章では、間表記ゆれの大きい取り図画像を対象に、グラフという数学的に構造化された表現に変換することを考える。これによって、間取りの比較や評価が容易になり、新たな間取り検索システムをはじめ様々な応用に繋がることが期待できる。間取り図の解析には深層学習による画像認識が役立つと考えられ、ここではそれを用いて間取り図を自動でグラフ化する手法を紹介する[1],[2]。

8.2 関連研究

関連する研究として、間取り図の解析を行う研究、グラフ化された間取り情報を応用する研究を紹介する。また、間取り図のデータセットについても述べる。間取り図を解析する研究では、間取り図を解析す



図 8.1 間取り図の例 [3]

ることで間取りをどのようにして計算機に認識させるかが課題となっている。グラフ化された間取り情報を応用する研究は、間取りがグラフで表現されることによってどのようなことが可能になるのか、どういった恩恵が得られるのか、といった間取り図をグラフ化することによる応用例である。

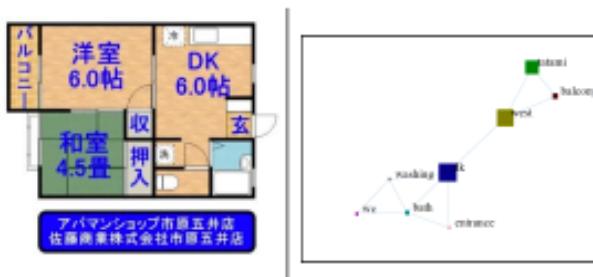
8.2.1 間取り図の解析

グラフ化を目的としていなくても、間取り図の解析自体を行っている研究は様々なものがある。例えば Ahmed[4] らは、間取り図の外内壁とテキストを解析することで、間取りに含まれる部屋の検出率を向上させることに成功している。また、Liu[5] らは、間取り図中の壁の交差点を検出することや、部屋の領域や浴槽等の物体を認識することなどにより、間取り図をラスターイメージからベクターイメージへ変換するという研究を行っている。Dodge[6] らも、深層学習を用いて間取り図をセグメンテーションする研究を行っており、その応用として間取りの 3D モデリングと家具配置の最適化を挙げている。

間取り図の解析だけではなくグラフ化を試みている例としては、我々の先行研究 [7] がある。そこでは、深層学習で間取り画像をセグメンテーションした後、領域の隣接性から図 8.2(a) のようにグラフを作成している。ただし、この先行研究ではグラフ作成時に単に距離の近い部屋同士を隣接と判定しているため、実際には行き来できない部屋の間にもエッジが存在してしまうという問題がある。つまり、ここで作成されているグラフは実際の間取り構造を正しく反映したものとは言えない。さらに作成されたグラフについての定量的評価も行われておらず、グラフ化の精度が不透明である。

8.2.2 グラフ化された間取り情報の応用

間取り図をグラフ化した先の応用としては、大原[8]らの間取りをクエリとする不動産検索システムが挙げられる。この研究では、ユーザーの間取りに対する細かい希望に応える検索システムとして、間取り図をグラフ構造として表し、そのグラフ間の類似度を用いて検索するという手法を提示している。図8.2(b)にその概要を示す。図の左側のグラフがシステムに入力するクエリであり、ユーザーが所望する間取りを表現したグラフである。それに対し右側の4つの間取り図はその検索結果であり、入力クエリのグラフに構造が似ている間取り図が検索されている。



(a) 間取り図のグラフ化 [7]



(b) 間取りをクエリとする検索システム [8]

図 8.2 関連研究の例

また、間取り検索の他にも、隣接グラフを用いて住宅の分類を行う研究を花里[9]らが、グラフ構造を組み込んだ回帰モデルによる賃料予測を行う研究を滝澤[10]らが行っている。花里らの研究では、グラフの型によって間取りを6つのタイプに分類できるとし、海外に比べ日本の住宅は「個室群化」傾向にあることを示している。滝澤らの研究では、部屋の隣接関係を表現したグラフ情報を用いることで賃料予測精度が向上し、室配置が賃料に大きく影響していることを示している。これらの研究から、間取り図をグラフ構造として扱うことで、間取り図をより有効活用することが可能であることがわかっている。

さらには、「現代風に感じる間取り」や「水回りが使いやすそうな間取り」といった物件に対する主観評価とグラフ構造を対応付けることで、間取り中のどの要素がそのような主観評価に影響を与えるかを明

らかにし、物件の設計の手助けとする試みもすでに始まっている [11]。

8.2.3 間取り図のデータセット

間取り図のデータセットとしては、[7] にて作成したものが挙げられ、それは間取り図画像とそのアノテーションからなる。[7] では、LIFULL HOME'S データセット [3] から間取り図画像をサンプルし、クラウドソーシングによりそれらの間取り図に対するアノテーションを付加している。LIFULL HOME'S データセットは、国立情報学研究所 [12] が株式会社 LIFULL [13] から提供を受けて研究者に提供しているデータセットである（今後に示す全ての間取り図画像はその出典を [3] とする）。アノテーションの付加は、クラウドソーシングプラットフォーム CrowdWorks [14] とオンラインアノテーションツール LabelMe [15] を利用して行われた。その結果、間取り図中の各領域のクラスラベルと輪郭座標の情報を保持したアノテーションが XML 形式で収集された。アノテーションでは 17 種類のクラスラベルが定義されており、その一覧を表 8.1 に示す。実際の学習には、間取り図画像 x とそのアノテーションを基に作成された正解画像 t のペア (x, t) を用いる。ここでいう正解画像とは、学習時に正解データ (Ground Truth) となるラベルマスク画像のことである。

表 8.1 クラスラベルの定義

ラベル	説明	ラベル	説明
wall	外壁、内壁	rouka	廊下
tatami	和室	stairs	階段
west	洋室	cl	クローゼット、押入れ、収納、下駄箱
dk	ダイニング、キッチン、リビング	fan	開き戸
wc	トイレ	slide	引き戸
bath	浴室	fold	折戸
washing	洗面所、脱衣所	window	窓
balcony	バルコニー、ベランダ、テラス	unknown	記載のない箇所、不明箇所
entrance	玄関		

8.3 間取り画像のグラフ化手法

本章で紹介する手法は、間取りの認識のみで終わることなく、間取り図画像からグラフ構造への変換までを行いうるものである。さらにそのグラフ構造には、ドアによる部屋同士の実際の接続関係を反映させる。これは、各部屋間の距離だけを判断材料としていたがゆえに実際の接続関係を捉えられていないかった既存手法の問題点を解決しようとするものである。手法は大きく次の 2 ステップからなり、その図解を図 8.3 に示す。

- Step1: 深層学習を用いて間取り図画像の semantic segmentation を行う
- Step2: 各部屋やドアの接続関係からルールベースでグラフ構造を作成する

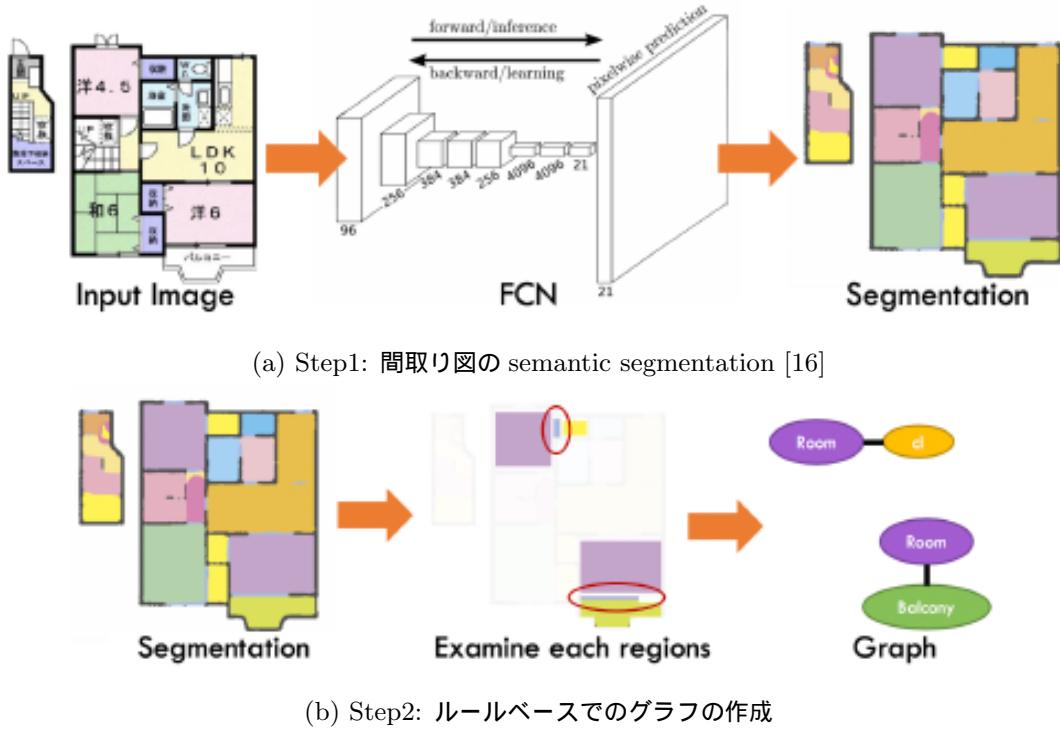


図 8.3 提案手法の流れ

8.3.1 間取り図の semantic segmentation

まず、深層学習を用いて、入力した間取り図画像の semantic segmentation を行う。semantic segmentation とは、画像の各ピクセルに対してラベル付けを行うことである。これにより、画像をピクセル単位で認識することができ、すなわちどこにどのような部屋が位置しているかという間取りの構造に関する情報を獲得することができる。その様子を図 8.3(a) に示す。

semantic segmentation を行うには、深層ニューラルネットワークモデルである Fully Convolutional Networks (FCN) [16] を用いる。FCN は、入力画像からピクセルごとのクラスラベルを End-to-End で学習できるモデルである。ここでは、FCN のアーキテクチャのうち、ベンチマークデータセットである PASCAL VOC 2012 [17] に対して最も高い精度を記録している FCN-8s を用いる [18]。そのアーキテクチャの概要を図 8.4 に示す。FCN-8s は多層な構造をとっており、層間では各種演算が繰り返される。各演算を f_{ks} (フィルタのカーネルサイズを k , ストライドを s) で表し、その入力の (i, j) 成分を x_{ij} とすると、出力 y_{ij} は、式 (8.1) のように求められる。

$$y_{ij} = f_{ks}(\{x_{si+\delta j, sj+\delta j}\}_{0 \leq \delta i \leq k, 0 \leq \delta j \leq k}) \quad (8.1)$$

FCN の最初の 5 層では、Convolutional Neural Networks (CNN) [19] で用いられているのと同様に、convolution と max pooling からなる演算によって入力画像から特徴マップが計算される。また、convolution の後には Relu 関数による activation が行われる。さらにその後には convolution が 3 層続き、ラベルの予測スコアのヒートマップが得られる。次に、deconvolution を行い、ヒートマップを入力

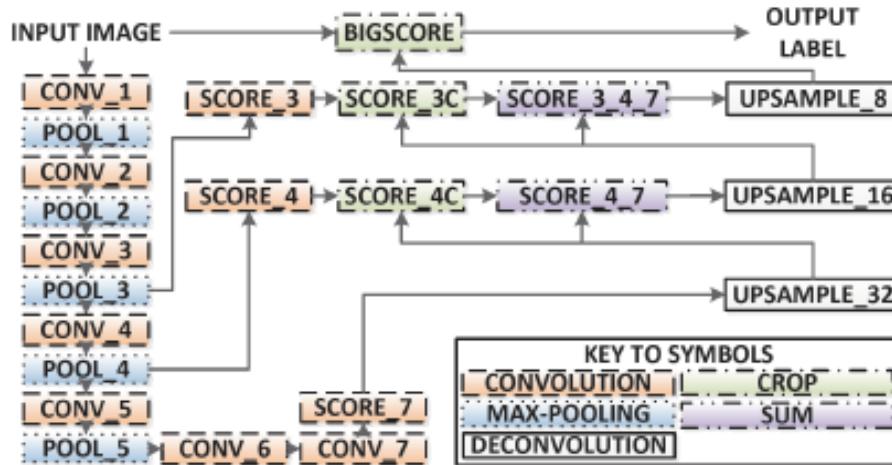


図 8.4 FCN-8s のアーキテクチャの概要 [20]

画像のオリジナルサイズまで upsample する。このとき，3 層目と 4 層目から求められる予測ヒートマップを図のように足し合わせて，最終的な出力を得る。

損失関数は正解画像に対するピクセルごとの cross entropy とし，ネットワークの各層における convolution または deconvolution フィルタの重みを更新することで学習を行う。

8.3.2 ルールベースによるグラフ化

次に，semantic segmentation を行った後の間取り図から各領域を抽出し，各部屋とドアの接続関係を調べてグラフを作成する。各部屋をノードとし，ドアで繋がっている部屋同士をエッジで接続する。これにより，実際の部屋の繋がりを反映したグラフを作成することができ，動線の最適化といった間取りの使いやすさを定量的に評価するような試みにも応用可能になると考えられる。我々のグラフ作成のアルゴリズムは以下の通りであり，その様子を図 8.3(b) に示す。

- ノードの作成
 - 同じラベルがひと続きになっている領域を抽出する。
 - その領域の面積が一定以上であれば，そのラベル名を冠したノードを作成する。
 - stairs ノードが複数存在する場合，それらをひとつのノードにマージする。(間取り図において階段は空間的に離れた位置に存在するのに対し，実際はそれらが繋がっていることを考慮する。)
- エッジの作成
 - 各ノードの領域が自分以外のノードの領域に隣接していれば，両ノード間にエッジを作成する。
 - 各ノードの領域が隣接しているドアを列挙する。
 - 同じドアに対して隣接しているノードの集合があれば，それらのノード間にエッジを作成する。

8.4 実験

本節では、提案手法の有効性を検証するために行った実験とその結果を示す。はじめに、実験に使用するデータセットについて、次に semantic segmentation の精度についてと最終的に変換したグラフの精度についてを順に議論する。そして最後に、提案手法によって変換したグラフを用いた応用例として、間取り類似物件検索を行う実験を示す。

8.4.1 データセットの改良

本研究では、[7] で作成したデータセットを改良したものを使用して以降の実験を行っていく。改良点は以下の 3 点である。次節で述べるが、これにより semantic segmentation の精度向上が確認できた。

- データ数を増加させたこと
- アノテーションの質を向上させたこと
- 正解画像の作成方法を修正したこと

ひとつめの改善点は、データを増強したことである。[7] で使用した間取り画像とその正解画像のペアは総計 2,635 組であったのに対し、本研究ではこれを 4,800 組に増加させた。これによってネットワークの学習に用いることのできるデータ数が増加するため、semantic segmentation の精度向上に寄与すると考えられる。

次の改善点は、アノテーションの質を向上させたことである。具体的には、unknown ラベルをできる限り減らし、正しいラベルを振り直すという作業を行った。unknown は記載のない不明箇所に付けられるラベルであり、例えば図 8.5(a) のような箇所である。一方で、間取り図には文字記載は無いものの明らかにそれとわかる領域も存在する。特に、廊下や下駄箱等の小さい収納スペースは文字記載が省略されている場合が多い。修正前のアノテーションでは、これらの箇所にも unknown が付けられており、例えば図 8.5(b) は明らかに廊下である領域に unknown ラベルが付けられてしまっている例である。ここでは、これらのラベリングを修正し、ラベル名が明らかな領域には正しいラベル名を振り直した。これを全 4,800 枚に対して行うことより、unknown ラベルの個数の割合は 8.4% から 4.7% へと減少し、より質の高いアノテーションデータとすることができた。

最後の改良点は、正解画像の作成方法を修正したことである。[7] の正解画像の一部には、図 8.6(b) のように、ドアが不必要的領域にまではみ出しているものが存在した。これは、アノテーションから正解画像を作成する方法に問題があったためである。これに対し、新たな正解画像の作成方法ではそのようなはみ出し部分をなくすように修正した。この改良も正解画像の質を高め、結果としてより精度の高い semantic segmentation が実現できると考えられる。

8.4.2 semantic segmentation の評価

前節で述べたデータセットを用いて、FCN による間取り図画像の semantic segmentation の精度評価実験を行う。実験は学習フェイズと評価フェイズからなる。学習フェイズでは、間取り図画像と正解



(a) 正しい unknown ラベル

(b) 修正すべき unknown ラベル

図 8.5 unknown ラベルの例

画像を繰り返し入力することでネットワークの学習を行う。ネットワークのアーキテクチャは FCN-8s を用い、エポック数や学習率などの実験条件は表 8.2 の通りとした。学習は max epoch 数だけ行い、Validation data に対して最も良い精度を示したときのモデルを保存する。その後の評価フェーズでは、保存したモデルを用いて Test data に対して accuracy と IoU を算出し評価を行う。

accuracy は最も単純な評価指標であり、正解ピクセル数を総ピクセル数で除した値である。クラス c についての accuracy は式 (8.2) のように計算される。ただし、 n_{ij} はクラス i に属しクラス j と予測されたピクセルの総数、 t_i はクラス i に属するピクセルの総数である。そして、全クラスの accuracy を平均した値を mean accuracy とすることで、全体としての認識精度を評価する。accuracy は単純でわかりやすいという利点がある一方、予測が外れた場合のペナルティを考慮していないという欠点がある。そこで、その点を考慮したより厳格な評価指標として Intersection over Union (IoU) がある。こちらは予測

表 8.2 semantic segmentation 性能評価実験の各種条件

データセットの分割			パラメータの設定	
Train data	データ数増加前	1,635	max epoch	500
	データ数増加後	3,800	learning rate	1e-5
Validation data	500		momentum	0.99
Test data	500		weight decay	5e-4

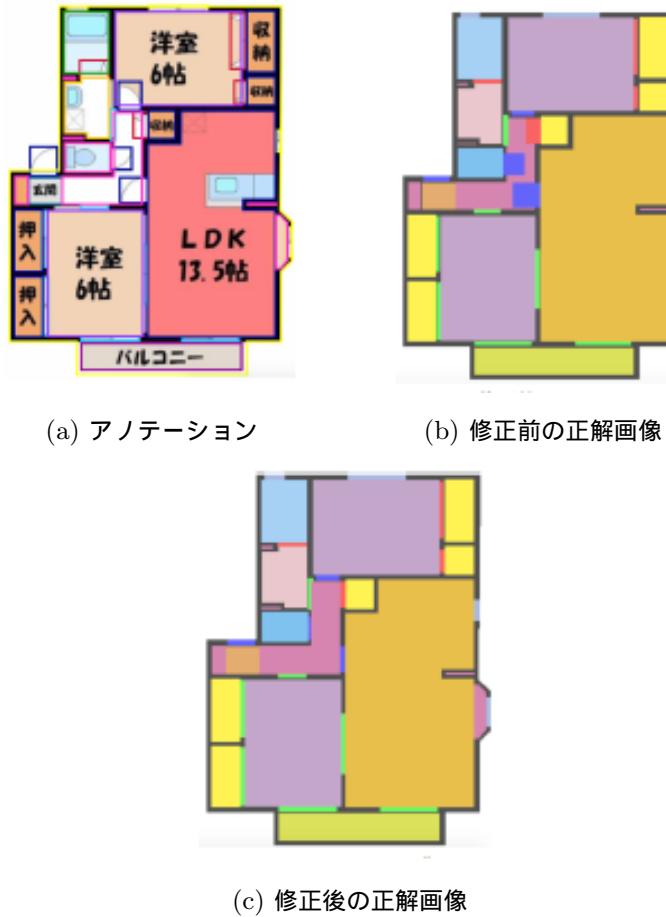


図 8.6 正解画像の作成方法を修正した例

したピクセル数をも分母に含めることで accuracy の欠点を解消しており、式 (8.3) のように計算される。結果として、IoU は正解領域と予測領域との重なり具合を表す尺度となる。そして同様に、全クラスについての IoU を平均した値を mean IoU とする。

$$\text{accuracy}_c = \frac{n_{cc}}{t_c} \quad (8.2)$$

$$\text{IoU}_c = \frac{n_{cc}}{t_c + \sum_i n_{ic} - n_{cc}} \quad (8.3)$$

クラス設定については、fan/slide/fold/window をひとつのクラス doors にマージし、合計 14 種類のクラス設定とする。アノテーションには元々 17 種類のラベルが付けられているが、ここでは後のグラフ化を目的とした場合の必要最小限なクラス設定に改変する。fan, slide, fold はそれぞれドアの種類を表しているが、グラフ化におけるドアの役割は各部屋の接続関係を表すことであるため種類別に分ける必要はない。したがって、これらの 3 種類のドアはひとつにマージする。また window は窓を表すが、必ず屋外領域と接していること、ドアと表記が酷似していることから特別に分類する必要はないと考え、ドアと同一クラスとする。

データセット改良による認識精度への影響

まず、データセットの改良がセグメンテーションの精度に与える影響を調べるために、改良後のデータセットだけではなく改良前のデータセットをも用いて学習を行い、その結果を比較した。両者の実験条件は同じであり、使用するデータセットだけを変えている。その結果算出された Test data に対する平均の mean accuracy と mean IoU を表 8.3 に示す。この結果から、データセットの改良によって mean accuracy, mean IoU ともにそれぞれ約 4%, 5% 向上したことがわかる。また、このときの semantic segmentation 結果の例を図 8.7 に示す。括弧内の数値はそれぞれ mean accuracy と mean IoU である。ここでは、認識精度が高い例、平均的な例、低い例をそれぞれ複数個ずつ示している。認識精度の高い例では、面積の大きい洋室や和室などの部屋だけでなく、洗面所や小さい収納スペースなども隅々まで認識できている。さらに、壁とドアについても正確に認識できていることがわかる。認識精度が平均的な例、すなわち mean accuracy と mean IoU がそれぞれおよそ 90%, 84% である例については、各領域の境界部分、特に廊下との境界部分にはみ出しが見られるなど細かい部分で誤認識が起こっているが、概ね問題なく認識できていることがわかる。

表 8.3 mean accuracy と mean IoU

データセット	mean accuracy	mean IoU
改善前	0.868	0.788
改善後	0.906	0.840

クラスごとの認識精度

次に、より詳細に認識精度を評価するためにクラスごとの accuracy と IoU を図 8.8 に示す。この結果から、FCN による semantic segmentation の認識精度は、各クラス領域の面積、特徴、文字表記の有無に関わっているということがわかった。

それぞれのクラスを順に見ていくと、まず和室、洋室、リビングダイニングキッチンについては非常に高い精度で認識できていることがわかる。これらの共通点は面積が広く、ほとんどの場合で文字表記が付されているということであり、それが高い認識精度の要因であると考えられる。次いで精度の良いトイレ、浴室、洗面所、バルコニーについては、必ずしも文字表記があるわけではないがそれぞれ特徴的な領域である（便器や浴槽、洗面台があったり、屋外に近い場所に位置していたりする）ことから、高い認識精度が達成できたと考えられる。逆に精度の低いクラスについては、unknown が特に低く、続いてドア、廊下、玄関、壁の精度が低いということがわかる。Unknown クラスは他のどのクラスにも属さない領域が対象であるため、共通する特徴が少なく、認識精度が低くなるのは納得できる結果である。廊下と玄関は文字による表記が無い場合が多く特徴の少ない領域であることから認識精度が低くなっていると考えられる。ドアと壁については、両者とも他クラスに比べ特徴的な領域ではあるものの、その面積が小さいことから高い精度を達成するのは難しいということが考えられる。特にドアは後のグラフ化で重要な役割を果たすため、さらなる認識精度の向上が望まれる。

以上のことから、面積が広く特徴的であり、かつ文字表記が存在する箇所の認識精度は高く、そうではない箇所の認識精度は低くなる傾向にあるといえる。



図 8.7 FCN によるセグメンテーション結果の例

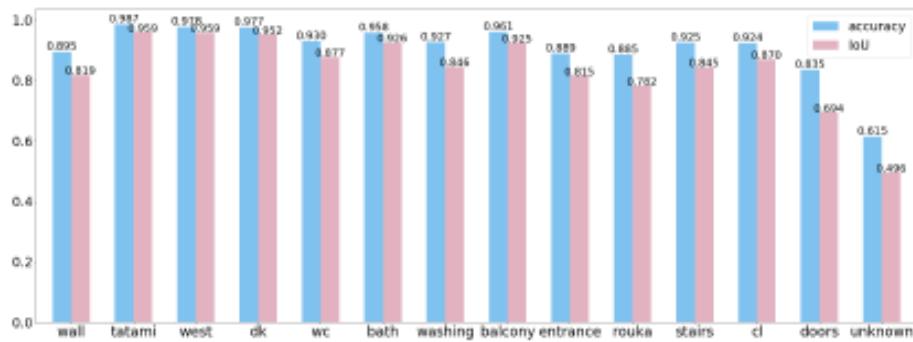


図 8.8 Test data に対するクラスごとの accuracy と IoU

間取り図中の文字の有無による認識精度への影響

さらに、間取り図中にある文字の有無が semantic segmentation の精度に与える影響を調べるための実験を行った。オリジナルの間取り図から文字を消去したものを用意し、それを保存したモデルに入力して semantic segmentation を行う。その結果を図 8.9 に示す。図 8.9 の上段が入力した間取り図、その下段がそれぞれの semantic segmentation の結果である。図 8.9(b) は全ての文字を消去した場合、図 8.9(c) は“LDK”という文字のみを残した場合、図 8.9(d) は“収納”という文字のみを消去した場合である。これによると、文字の有無によって semantic segmentation の結果に差があることがわかる。文字を全て消去した図 8.9(b) では、多くの場所でセグメンテーションに誤りが見られる。一方、特定の文字だけを残した図 8.9(c) と図 8.9(d) では、残した文字に対応する領域がそれぞれ正しく semantic segmentation が行われていることがわかる。以上のことから、FCN は文字を頼りに（とは言っても意味は理解しておらず、記号的な使い方をして）semantic segmentation を行っていることが示唆される。

8.4.3 グラフ化の評価

提案手法の Step2 におけるルールベースでのグラフ化をし、その評価を行う。そのためにはまず Test data に含まれる 500 枚の間取り図に対してその正解グラフを作成し、提案手法のグラフ化アルゴリズムを semantic segmentation 後に適用して作成したグラフがどれだけ正解グラフに類似しているかを調べる。その類似度を既存手法と提案手法それぞれのアルゴリズムで作成したグラフについて算出し比較を行う。

グラフ同士の類似性評価指標としては、一般的で扱いやすいことから、MCS による類似度を用いる。Maximum Common Subgraph (MCS) とは、2 つのグラフの共通サブグラフのうちエッジ数が最大のグラフである。グラフ G_1 のノード数とエッジ数の和を $|G_1|$ と表し、グラフ G_1 とグラフ G_2 の MCS を $MCS(G_1, G_2)$ とすると、グラフ G_1 とグラフ G_2 の類似度は式 (8.4) で計算される。MCS 類似度は、一般的なグラフの類似度指標として Labrijli ら [21] によって紹介されており、例えば創薬の領域で Cao ら [22] によって分子類似性を測る指標として用いられている。MCS 類似度は 2 つのグラフが全く異なる場



図 8.9 文字を消去したときの semantic segmentation 結果の例

合は 0 , 全く等しい場合は 1 となり , 0 ~ 1 に正規化されている。

$$\text{sim}(G_1, G_2) = \frac{|\text{MCS}(G_1, G_2)|}{\max(|G_1|, |G_2|)} \quad (8.4)$$

グラフ化アルゴリズムは第 3 章で示した通りであり , ここではノードを作成する際の面積のしきい値を , 各ラベルごとに Train data 中の最小値とする。[7] のアルゴリズムでは , 面積 A 以上の領域をノードとし , 各ノードの領域間の最小距離が D 以下であれば両ノード間にエッジを作成するというもので , ここでは論文の通り $A = 1000$, $D = 30$ とする。それぞれのグラフ化アルゴリズムで作成したグラフと正解グラフとの類似度を算出する。

以上の実験結果を表 8.4 に示す。これより , 提案手法は既存手法に比べて約 40% 高い値を記録してい

ることがわかる。既存手法は上述したように本来繋がっていない部屋の間にもエッジを作成してしまうため、その分だけ正解グラフとの類似度が低下していると考えられる。この実験結果から、本研究の提案手法による間取り図のグラフ化によって、平均して正解グラフとの類似度が 0.8 程度となるグラフを作成できるということが確認できた。

表 8.4 それぞれのグラフ化アルゴリズムを用いて作成したグラフと正解グラフとの類似度

グラフ化アルゴリズム	正解グラフとの類似度
既存手法 [7]	0.581
提案手法	0.810

また、図 8.10 と図 8.11 に実験結果の一例を示す。図 8.10 は一般的な 1K、図 8.11 は 2 階の 3LDK の間取りである。図 8.10B と図 8.10C、あるいは図 8.11B と図 8.11C を比較すると、提案手法のグラフ化アルゴリズムの方が実際の間取り構造である各部屋の接続関係を正しく反映していることがわかる。また図 8.11 を見ると、既存手法では接続関係が絶たれてしまう階段についても、提案手法ではひとつのノードとしてまとめることでノードが途切れることなく、正しい間取りを表現できていることがわかる。ただし、階段が複数種類存在する間取りの場合はグラフ化に失敗するため、この点は改善の余地がある。もちろん、図 8.10C のように、semantic segmentation の精度が原因でグラフ化に失敗するという例もある。図 8.10C のグラフは、玄関横に存在する下駄箱に相当する cl ノードがなく、この点は正解グラフと異なっている。

8.4.4 間取り類似物件の検索

以上の提案した手法による間取り図のグラフ化を応用して、間取り類似物件の検索を行う。ここでは、クエリとして任意の間取り図画像を入力し、その間取りとの類似度が高い間取りを持つ物件の間取り図画像を出力するというシステムを作成する。実験ではクエリとして Test data を用い、検索候補として LIFULL HOME'S データセットからさらに 25,000 件の間取り図を用いる。この 25,000 件にはこれまでの実験で用いた 4,800 枚は含まれていない。検索の手順は以下の通りである。

1. 提案手法を用いて検索候補となる間取り図 25,000 件をあらかじめグラフ化しておく。
2. システムにクエリとなる間取り図画像を入力する。
3. 提案手法を用いてクエリ画像をグラフ化する。
4. クエリのグラフと検索候補のグラフとの間の類似度を計算する。
5. 類似度が高い順に検索結果を表示する。

このようにして類似する間取り図を検索した結果の例を図 8.12 と図 8.13 に示す。図の左上がクエリとなる間取り図、その右から 2 段目にかけて類似度が高い順に 5 位までの検索結果を示している。図 8.12(a) のクエリとなっている 2DK の部屋は、玄関から直接アクセスできる DK から和室と洋室それぞれへ接続しているという間取り構造を持っている。和室と洋室には収納がひとつずつあり、両者からアクセスできるテラスを持つ。また水回りは全て DK に繋がっている。一方で、検索された間取り図を見



(a) 間取り図

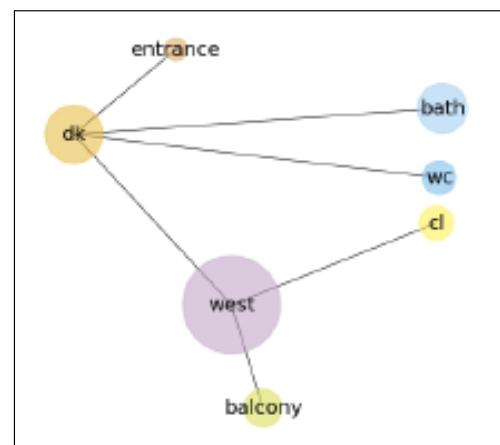
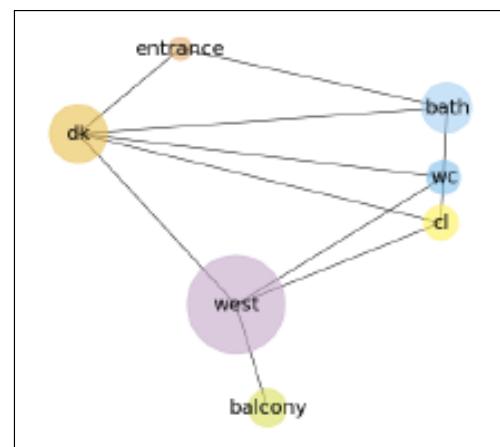
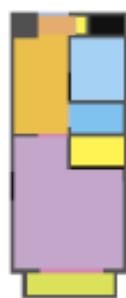
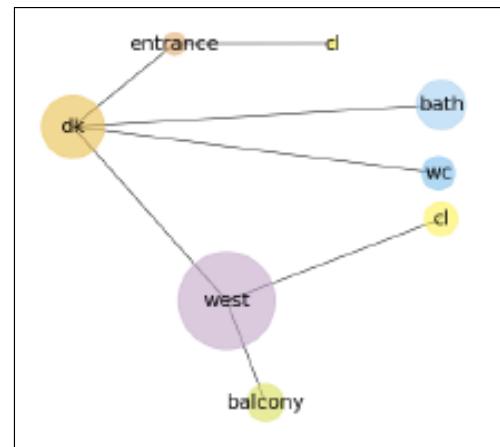


図 8.10 作成したグラフの例 1

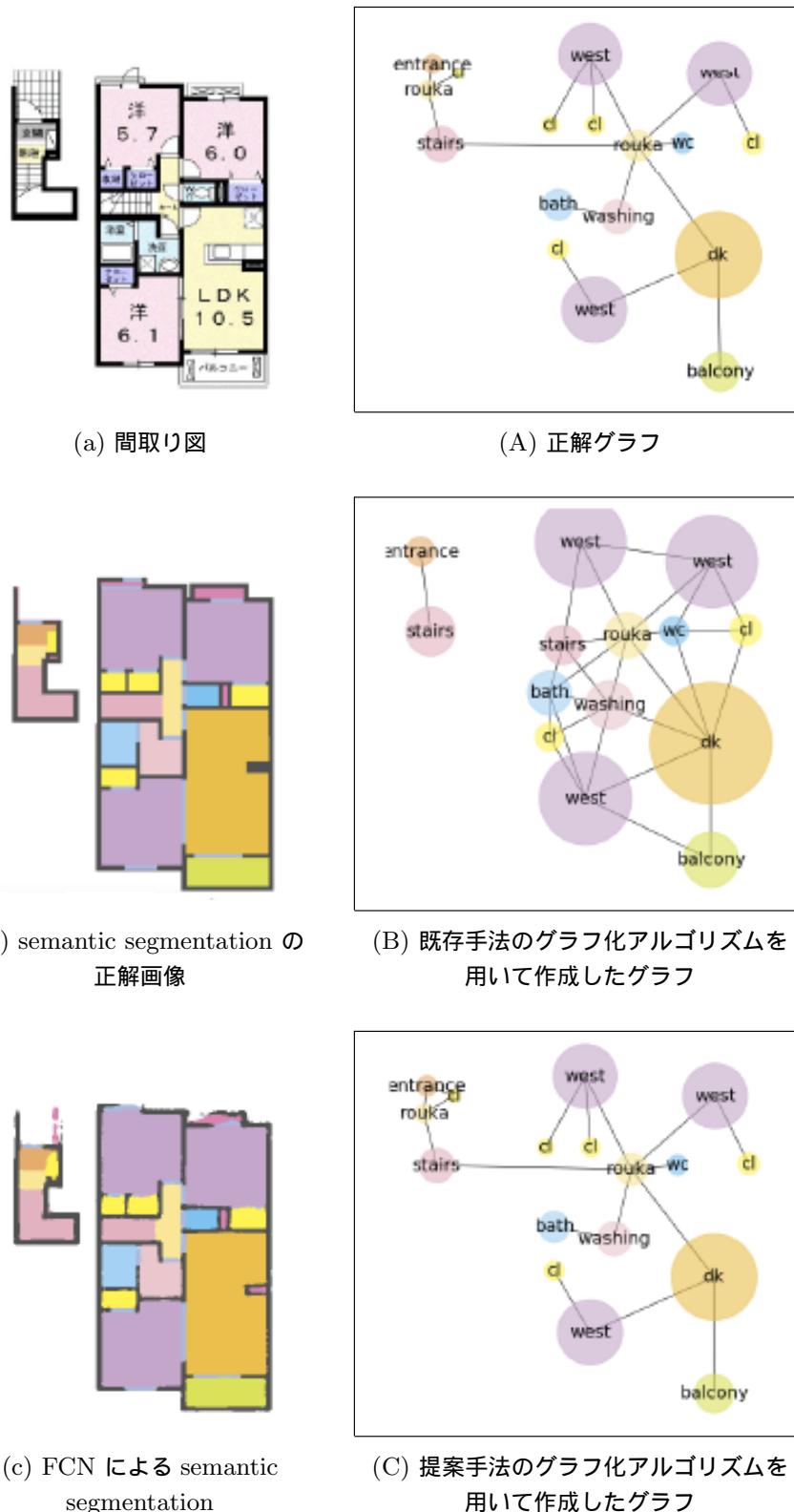


図 8.11 作成したグラフの例 2

てみると、これと同様の間取り構造を持っていることがわかる。次に図8.12(b)のクエリは、2階建ての2LDKで1階がリビング、2階が2つの洋室で構成されている。水回りは全て1階にあり、2階の洋室にはそれぞれクローゼットが、また一方の洋室にはバルコニーが付いている。これに対して、検索された間取り図も同様の構造を持っており、特に3位までの間取り図はクエリによく一致している。4位と5位は2階部分の部屋がひとつ多いが、その点以外は概ね一致しているといえる。

一方で、図8.13は類似の間取り図を検索できなかった例である。図8.13(a)のクエリとなる間取りは、玄関がLDKと洋室の両方に繋がっていて、中央位置する洗面所を介してLDKと洋室を行き来できるという特徴的な間取りである。しかし、検索結果にそのような間取りを含む物件は現れなかった。この場合、クエリとなっている間取り構造が極めてユニークであったため検索に失敗した可能性も考えられる。図8.13(b)については、グラフ化の精度が低いことが原因で検索に失敗したと考えられる例である。ここでのクエリは、図8.7(h)で示したsemantic segmentationの精度が低かった間取り図である。そのため、そこから変換したグラフの精度も低くなる。すると、クエリの間取り構造を正しくグラフで表現できていないまま検索を行うこととなり、類似の間取り図を発見することに失敗したと考えられる。

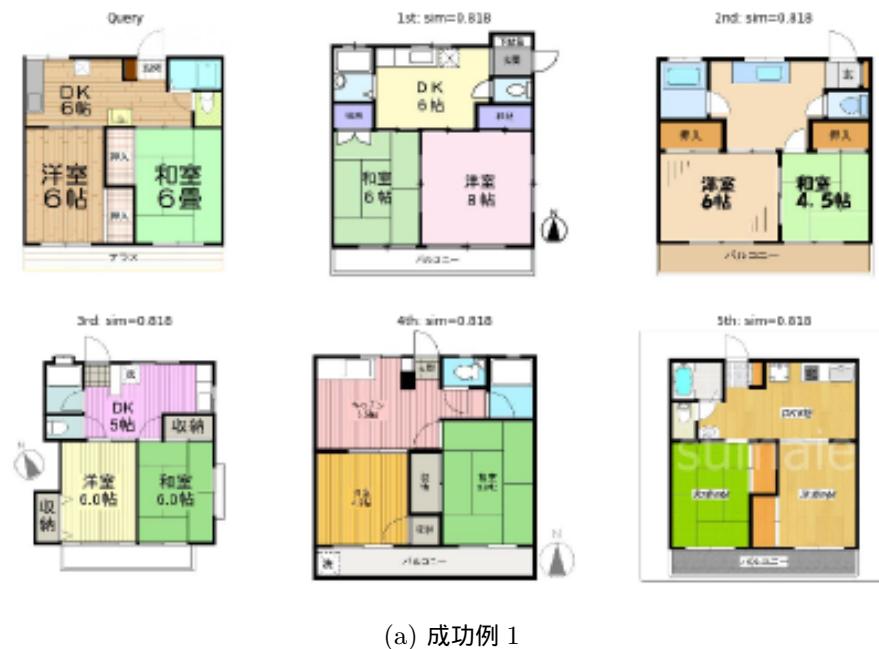
以上の結果から、間取りをグラフで表現しその類似度での検索を行うことで、間取り図のスタイルやカラーリングは異なっていたとしても間取りの構造が類似している物件を検索できることが確認できた。

8.5 まとめ

不動産間取り図の画像を自動でグラフ構造に変換する試みについて紹介した。これは、不動産業界でデジタル技術の導入が進む中、間取り図を機械的に扱いやすくしようという取り組みである。間取りをグラフで表現することができれば、例えば希望する間取りを細かく検索できる新たなシステムの開発に繋がることが期待できる。その他にも、間取り同士の比較や評価が容易になることで間取りの使いやすさを定量化できたり、賃料予測モデルに間取りを組み入れることで予測精度を向上させることができたりと、様々な応用に繋がると考えられる。

本研究で提案した手法により、間取り図画像をおよそ8割の精度でグラフに変換できることを示した。これは、深層学習を用いたsemantic segmentationによる画像認識精度の向上と、実際の部屋の接続関係を考慮したグラフ化アルゴリズムによって可能になったといえる。さらに、我々の手法を用いることによって任意の間取り図に対してそれと構造的に類似している間取りを検索する仕組みを実現できることを確認した。

一方で、本手法によるグラフ化は完全ではなく、その先の応用のためにもさらなる精度向上が望まれる。そのためには、間取り図、特にドアについての認識精度を向上させる、あるいはグラフ化アルゴリズムを改善する、といった改善が必要になる。また、類似物件検索への応用についてはより多くの課題が残されており、例えば検索された間取り図がユーザーの所望している間取り図に如何ほど合致しているのか、検索に用いる類似度指標としてMCS類似度は適切であるかといった点はさらなる検証が必要であると考えられる。また、ユーザーの間取りに対する要望をどのような形で計算機に伝えるか、そのインターフェースは実用上極めて重要であると考えられ、その考案は今後の課題であるといえる。



(a) 成功例 1



(b) 成功例 2

図 8.12 類似間取り図検索の成功例



(a) 失敗例 1



(b) 失敗例 2

図 8.13 類似間取り図検索の失敗例

参考文献

- [1] 山田万太郎, 汪雪, 山崎俊彦, 相澤清晴. 深層学習を用いた不動産間取り図のグラフ化と物件検索への応用. 2019 年度人工知能学会全国大会 (第 33 回), pp. 3N4-J-10-03, 2019.
- [2] Mantaro Yamada, Xuetong Wang, Toshihiko Yamasaki, and Kiyoharu Aizawa. 深層学習を用いた不動産間取り図のグラフ化と物件検索システムへの応用. 第 22 回画像の認識・理解シンポジウム (MIRU2019), pp.OS2B-7, 2019.
- [3] 国立情報学研究所 IDR 事務局. LIFULL HOME'S データセット, 2010-2019.
<https://www.nii.ac.jp/dsc/idr/lifull/homes.html>.
- [4] Sheraz Ahmed, Markus Weber, Marcus Liwicki, Christoph Langenhan, Andreas Dengel, and Frank Petzold. Automatic Analysis and Sketch-Based Retrieval of Architectural Floor Plans. Pattern Recognition Letters, Vol. 35, pp. 91 - 100, 2014. Frontiers in Handwriting Processing.
- [5] Chen Liu, Jiajun Wu, Pushmeet Kohli, and Yasutaka Furukawa. Raster-to-Vector: Revisiting Floorplan Transformation. In IEEE International Conference on Computer Vision, pp. 2214-2222, 2017.
- [6] Samuel Dodge, Jin Xu, and Björn Stenger. Parsing Floor Plan Images. In IAPR International Conference on Machine Vision Applications, pp. 358-361, 2017.
- [7] Toshihiko Yamasaki, Jin Zhang, and Yuki Takada. Apartment Structure Estimation Using Fully Convolutional Networks and Graph Model. In ACM Workshop on Multimedia for Real Estate Tech, RETech'18, pp.1-6, 2018.
- [8] 大原康平, 山俊彦, 相澤清晴. 間取りや広さをクエリとする直感的な不動産検索システム. 情報処理学会全国大会講演論文集, Vol. 78, No. 4, pp.4.311-4.312, 2016.
- [9] 花里俊廣, 平野雄介, 佐々木誠. 首都圏で供給される民間分譲マンション $100m^2$ 超住戸の隣接グラフによる分析. 日本建築学会計画系論文集, Vol. 70, No. 591, pp. 9-16, 2005.
- [10] 瀧澤重志, 吉田一馬, 加藤直樹. グラフマイニングを用いた室配置を考慮した賃料分析: 京都市郊外の 3LDK を中心とした賃貸マンションを対象として. 日本建築学会環境系論文集, Vol. 623, pp. 139-146, 2008.
- [11] Taro Narahara and Toshihiko Yamasaki. A preliminary study on attractiveness analysis of real estate floor plans. In IEEE 8th GlobalConference on Consumer Electronics (GCCE), 2019.
- [12] 国立情報学研究所. <https://www.nii.ac.jp/>.
- [13] 株式会社 LIFULL. <https://lifull.com/>.

- [14] 株式会社クラウドワークス. <https://crowdworks.jp/>.
- [15] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. LabelMe: A Database and Web-Based Tool for Image Annotation. International Journal of Computer Vision, Vol. 77, No. 1-3, pp. 157-173, 2008.
- [16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, 2015.
- [17] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision, Vol. 111, No. 1, pp. 98-136, 2015.
- [18] Leaderboards for the Evaluations on PASCAL VOC Data, Segmentation Results: VOC2012. <http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=6>.
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In International Conference on Neural Information Processing Systems, Vol. 1, pp. 1097-1105, USA, 2012.
- [20] Luis García-Peraza-Herrera, Wenqi Li, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, and Sébastien Ourselin. Real-Time Segmentation of Non-Rigid Surgical Tools based on Deep Learning and Tracking. In Computer-Assisted and Robotic Endoscopy, pp. 84–95, 2017.
- [21] Amine Labriji, Salma Charkaoui, Issam Abdelbaki, Abdelouhaed Namir, and El Houssine Labriji. Similarity Measure of Graphs. International Journal of Recent Contributions from Engineering, Science & IT, Vol. 5, No. 2, 2017.
- [22] Cao Yiqun, Jiang Tao, and Thomas Girke. A Maximum Common Substructure-based Algorithm for Searching and Predicting Drug-like Compounds. Bioinformatics, Vol. 24, pp. i366-i374, 2008.

第9章

不動産物件情報の流通と活用を支える データベース・情報アクセス技術

9.1 データベース・情報アクセス技術の発展

不動産市場において人工知能 (AI) の活用を図るにあたって絶対に欠かせないのが、市場に流通している（あるいは過去に流通していた、または将来流通する可能性のある）不動産物件に関する網羅性が高く正確な不動産物件データベースである。現在では、米 Zillow 社が公表している Zestimate とよばれる物件参考価格など、不動産物件データベースを利用して AI によって算出された情報が、不動産物件取引においても広く活用されつつある。また、不動産物件を探している利用者への物件情報の推薦、アパート・マンション開発業者による土地の仕入れ業務の効率化、不動産広告における情報の適正化（広告規約違反の是正、掲載画像の品質チェック）など、不動産物件データベースの用途はきわめて幅広い。

不動産物件データベースは、コンピュータの実用化、情報検索、インターネット、ビッグデータ処理など、データベース技術および情報アクセス技術の発達と軌を一にして発展してきた。本章では、データベース・情報アクセス技術の発展および不動産物件情報への応用の歴史を概観した上で、データベース・情報アクセスの基礎技術について解説するとともに、近年発達が著しい深層学習を、不動産物件画像に適用する取り組みの事例についても紹介する。

データベースとは、蓄積および検索が容易にできるように構築された情報の集合体を指す。コンピュータの出現以前においても、紙のカードなどを用いて大量の図書をタイトル、著者名などの索引から素早く検索する図書目録などの仕組みが 18 世紀には実用化されていた。

第二次世界大戦の時期にコンピュータが実用化されたことにより、コンピュータをデータベースに応用するための研究開発も盛んに行われてきた。コンピュータにおいてデータベースを実現する方法を論じた Young と Kent による先駆的な研究 [1] が 1950 年代に行われたのち、1960 年代には階層型データベースやネットワーク型データベースのシステムが実用化された [2]。現代におけるデータベースの主流であるリレーションナルデータベース管理システム (RDBMS) は、IBM の Codd によって提案された関係モデル [3] の考え方をもとにしている。また、現代の RDBMS の大部分は、SQL というデータベース言語によって、ある程度統一された方法によってデータの操作や定義などが可能である。関係モデルによって扱える情報は、構造化データ、すなわち一定の規則によって定型化された情報である。

一方で、非構造化データ、すなわち一定の規則による定型化が困難な情報をコンピュータで扱う方法についても、図書や論文、特許などの自然言語による情報を主な対象として、古くから行われてきた。1950年代の先駆的な取り組みとしては、IBMの研究所に在籍していた Luhn による文献検索の研究 [4] が広く知られている。その後、1957年のソ連による初の人工衛星打ち上げのニュースが米国を中心とする西側諸国に大きな衝撃を与えたといわれる「スプートニクショック」が、科学技術に対する政府からの支援を増やし、論文などの科学技術情報へのアクセス技術の研究を推し進めることにつながった。1960年代から1980年代にかけて、医学文献の検索システム MEDLARS [5] などが実用化されるとともに、Saltonを中心とするコーネル大学の研究グループによる SMART システム [6] など、文書検索システムの研究が推し進められた。

データベースおよび情報アクセス技術は、1980年代にはいったんの完成をみて、組織（政府、企業、研究機関など）の内部における業務システムのデータベース構築や、図書館の蔵書検索、論文検索、特許検索などの各種サービスに応用されていった。しかし、この時点では、データベースや各種の検索サービスの利用者は、組織内の専門職やライブラリアン、各分野の研究者などの一部の層に限られていた。

このような状況は、1990年代後半からのインターネットおよび World Wide Web（以下、Web と略す）の爆発的普及にともない大きく変化した。多くの人々が、Web を通じて RDBMS や文書検索システムによって提供される情報サービスを日常的に利用することが当たり前となるとともに、Web 上の情報量の激増にともない、新たなデータベースおよび情報アクセス技術の開発へのニーズが高まった。RDBMS は、構造化データを厳密に管理するための各種機能（たとえばデータ更新の一貫性を保つためのトランザクション機能）を有するが、Web 規模の情報量（テラバイト、ペタバイト単位以上）に対しては適用が難しい。また、図書や新聞記事に対して有効であった文書検索システムの技術は、激増する Web ページ数に適用しても良好な検索結果を得られず、利用者が必要とする情報が大量のスパムページに埋もれてしまうという問題に直面していた [7]。

こうした問題に対して、Web 規模の情報量に対応できるデータベース技術および情報アクセス技術の開発が 2000 年前後から進められてきた。RDBMS のような固定的な規則に縛られず、テラバイト、ペタバイト単位の情報量に対応可能な NoSQL とよばれるデータベース技術群が登場し、広く利用されるようになった [8]。また、利用者の閲覧・購買履歴などのビッグデータを活用する協調フィルタリング [10] などの情報推薦アルゴリズムや、Google の共同創業者である Larry Page と Sergey Brin が米スタンフォード大学の大学院生として考案した PageRank [9] など、Web 情報の特性を生かした順位付けの技術が開発された。これらの技術は、Google、Amazon、Facebook などに代表される現代の巨大プラットフォームビジネスを支える基盤となっている。

2010年前後からのスマートフォンの爆発的普及、深層学習（ディープラーニング）などの人工知能技術の成熟は、情報アクセス技術にも新たなブレークスルーをもたらしている。スマートフォンなどによって Web やソーシャルメディアにアップロードされた画像・動画が激増するとともに、利用者が見たい画像・動画を素早く探せる技術へのニーズが高まっていたところに、深層学習によって画像・動画の内容を理解し、映っている人物や物体の認識が高精度でできるようになったことで、Web に流通する情報は、テキスト中心から画像・動画中心にシフトしつつある。

9.2 不動産物件情報へのデータベース・情報アクセス技術の応用

不動産業の発展の歴史を振り返ると、現代のような形で大量の不動産情報が市場に流通する仕組みが整えられたのは、早くとも20世紀に入ってからである。19世紀以前の不動産物件の売買や賃貸は、いわゆる口コミに頼る部分が大きかった。全米リアルター協会(NAR)のWebサイト[11]によれば、不動産取引業者どうしが物件情報をやりとりするMultiple listing service(MLS)の仕組みは、18世紀後半に入つて不動産取引業者が定期的に各地域の協会の事務所に集まって売り物件の情報のやりとりをするようになったのが起りとされている。

1950年代から開発が進められてきたコンピュータによるデータベース技術は、不動産業界においても順次応用が進められてきた。とくに、関係モデルおよびRDBMSは、不動産物件情報の根幹をなす物件所在地(住所、緯度・経度)、物件種別(一戸建て・アパート・マンションなど)、賃料・価格、専有面積、築年数、建物構造などの各種情報がいずれも定型化が容易で親和性が高いため、MLSなどにおいて広く用いられている。日本においても、建設省(現国土交通省)所管の財団法人であった不動産流通近代化センターが中心となって不動産物件情報の交換を行うためのコンピュータネットワークシステムの開発が進められ、1990年よりREINS(レインズ、Real Estate Information Network Systemの略称)としてサービスが行われ、標準化された不動産情報の流通の仕組みが整えられた。

MLSなどの不動産物件情報データベースは、基本的に不動産仲介のライセンスをもつ会員事業者にのみアクセスが許されている。不動産物件を探している一般消費者には、大量の不動産情報にアクセスする手段は基本的に存在せず、新聞や情報雑誌などのマスメディアに不動産会社が掲載した広告などが、不動産情報を得る主な手段であった。1990年代後半からのインターネットおよびWebの爆発的普及は、こうした状況を一変させた。米国のZillow、英国のRightmoveなどに代表される不動産情報のポータルサイトが登場するとともに、Google AdWordsに代表される検索連動型広告が普及することによって、一般消費者がWebのキーワード検索で不動産物件情報を探すといった行動が当たり前となった。日本国内でも、SUUMO、LIFULL HOME'S、at homeなどの不動産情報サイトが、一般消費者向けに使いやすい物件検索サービスを提供している。多くの不動産会社も、集客のための広告にWebを積極的に活用するようになることで、Web上に大量の不動産物件情報が流通するようになった。また、ソーシャルメディア隆盛の流れは不動産の世界にも波及している。不動産会社や不動産物件への評価を口コミとして共有する情報サービスがいくつも立ち上がり、物件探しにおいて活用されるようになっている。

近年のスマートフォンの爆発的普及は、不動産物件情報の在り方にも大きな影響を与えている。スマートフォンで不動産物件情報を探すという行動が当たり前になるとともに、スマートフォンの小さな画面でも物件情報の直感的な取捨選択が可能な写真や間取り図、パノラマ写真、動画などの画像情報の充実度が利用者の行動に大きな影響を与えるようになりつつある。不動産会社や不動産情報サイトの運営者にとっても、高品質な画像情報を利用者に提供することがビジネスに直結するようになってきた。さらに、バーチャルリアリティ(VR)対応のデバイスの普及により、不動産物件のVRコンテンツなども充実しつつある。深層学習を中心に飛躍的な発展を遂げているAI技術は、不動産物件情報の品質向上や、新たな付加価値の創出にも活用されるようになってきた。深層学習を利用したいいくつかの事例について、9.4節にて紹介したい。

9.3 RDBMS の仕組み

現代において流通している不動産情報の大半は、RDBMS によって管理・運用されている。本節では、RDBMS の基本的な概念と、RDBMS における素早い情報アクセスを可能としているインデックスの仕組みについて簡単に紹介する。本節では不動産情報が RDBMS においてどのようにして扱われているかのイメージを伝えることに重点をおく。詳細な技術的解説についてはデータベース工学の専門書を参照されたい。

RDBMS は、あらゆる情報をテーブルの集合として扱う。不動産賃貸物件情報を表現するテーブルの例を図 9.1 に示す。テーブルの各々の行のことを組 (tuple) またはレコード (record) といい、各々の列のことを属性 (attribute) またはフィールド (field) という。図 9.1 の例では、組は各々の不動産賃貸物件に対応し、属性は物件情報を表現する各種のメタデータに対応する。各々の属性には、属性名 (例: 物件 ID, 物件名, 所在市区町村, 賃料, 間取り, 部屋面積) と、データ型 (例: 整数, 実数, 文字列) が定義されている。また、各々の属性にさまざまな制約 (例: 重複の禁止, 空データの禁止など) を定義することもできる。対象とする情報の性質に合わせてデータ型, 制約を定義することで、情報の不整合 (例: 同一の物件 ID をもつ物件情報の重複, 賃料や部屋面積への数値以外のデータの登録など) を防止し、データベース全体での情報の一貫性を保ちやすくなる。

						属性 (attribute)			
物件 ID	物件名	所在市区町村	賃料	間取り	部屋面積	属性名 (attribute name)	データ型 (data type)		
(整数)	(文字列)	(文字列)	(整数)	(文字列)	(実数)				
1 辰巳レジデンス405号室	東京都江東区	185000	1LDK	62.50					
2 申酉アパート201号室	横浜市中区	70000	1K	22.35	組 (tuple)				
3 子丑ヒルズ302号室	千葉市美浜区	100000	3LDL	95.20					

↑
関係
(relation)

図 9.1 RDBMS におけるテーブルの例

RDBMS が備える重要な機能の一つとして、任意の条件によるレコードの絞り込み検索がある。1万件の不動産賃貸物件が登録されているテーブルを例に考えてみよう。たとえば、賃料が 60,000 円以上、80,000 円未満の物件だけを検索したいとする。もっとも単純なアルゴリズムは、1万件のレコードのすべてについて、賃料のフィールドの値が「60,000 以上, 80,000 未満」であるかどうかを判定し、条件に合致するレコードだけを検索結果として返す処理を 1 万回繰り返すというものであろう。しかし、このアルゴリズムでは、検索にかかる所要時間は、テーブルに登録されているレコード数に比例して増えてしまう。

そこで、RDBMS においては、指定したフィールドに対してインデックスをあらかじめ作成しておき、検索処理にかかる時間を短縮することが一般に行われている。よく使われるインデックス作成アルゴリズ

ムとしては、B木などが知られている。図9.2に、B木によって作成されたインデックスの例を示す。たとえば、B木によるインデックスを賃料フィールドに対してあらかじめ作成しておくと、賃料のフィールドの値が「60,000以上、80,000未満」であるレコードを、一瞬にして絞り込むことが可能となる。

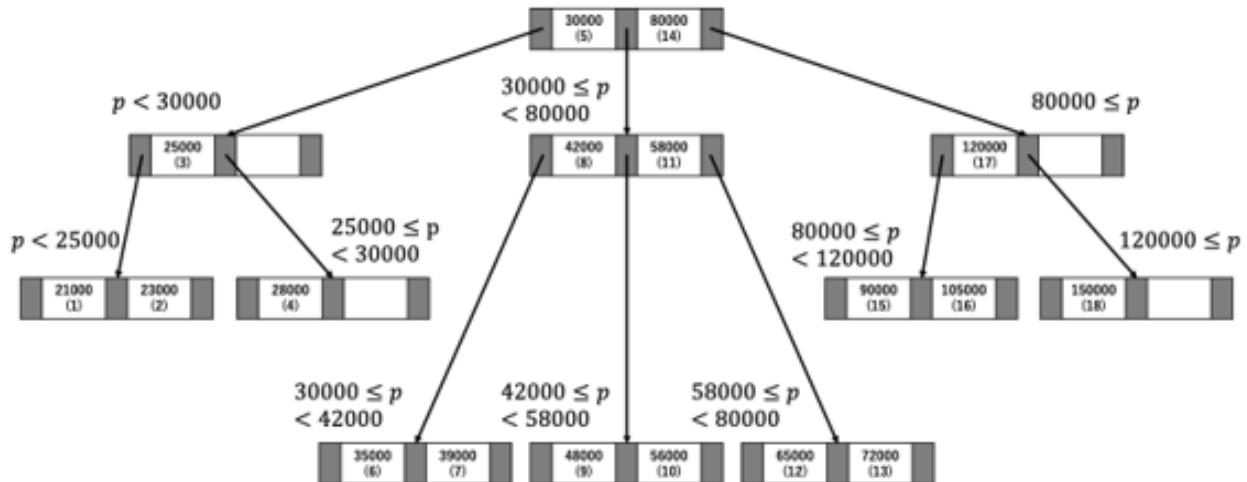


図9.2 B木によるインデックスの作成

9.4 不動産物件画像への深層学習の適用

第5章でも触れている深層学習は、さまざまなタスクにおいて活用が進められているが、深層学習の威力が広く知られるようになったきっかけが、2012年に開催された画像認識研究のコンペティションILSVRC^{*1}である。深層学習を採用したトロント大学のシステムSuperVisionが、他のチームを圧倒的に上回る精度を達成したことが、画像処理および人工知能の研究コミュニティに大きな衝撃を与えた。深層学習は、音声認識や機械翻訳、ロボットの制御や自動運転などあらゆる分野への応用が進みつつあるが、最も応用と手法の洗練が進んでいるのは、現在でも画像処理の分野である。今後も当分の間は深層学習の発展は画像処理分野を中心に進むことが予想される。

深層学習の有効性が広く知られるようになり、利用しやすいオープンソースのソフトウェアライブラリなども整備されてきたことから、不動産物件写真に深層学習を適用する研究や応用事例も増えつつある。本節では、最近のいくつかの事例を紹介する。

*1 ImageNet Large Scale Visual Recognition Challengeの略。画像に写っているのがどんな物体（ヨット、犬、猫、花など）なのかをコンピューターが当てるタスクが課される。ImageNetは、画像物体認識の研究促進を目的として整備されている画像データベースで、英語の概念辞書であるWordNetの同義語セット（synset）2万件以上に対応づけられた1,400万点以上の画像データから構成されている。

9.4.1 写真情報品質の向上の取り組み

前述の通り、不動産物件を探しているユーザーから非常に重視されている物件写真において、品質のばらつきは大きな課題となっている。なかには、「不動産の表示に関する公正競争規約」などに抵触する写真が掲載されるケースも見られる。不動産情報サイトを運営する各社では、人手によるチェックなどで情報品質の向上に努めているが、数百万点を超える写真データが毎日のように入稿される状況では、人手では限界があることから、深層学習などの最先端の画像処理技術を利用する取り組みが進められている。

菊田ら [13] は、不動産情報サイト SUUMO において、規定に抵触する異常写真を検出するタスクへの深層学習適用の事例を報告している。「人が映り込んでいる写真」を検出するタスクでは、画像処理向きの深層学習手法である畳み込みニューラルネットワーク (Convolutional Neural Network, CNN) の一種を利用し、異常写真を見逃してしまう確率を 5% 未満に抑えることができたと報告している。

石田ら [14] は、LIFULL HOME'S データセット [16] を利用し、13 種類の写真種別^{*2}の深層学習による自動判別の精度を評価している。同じく CNN を利用し、13 万点 (各種別ごとに 1 万点を無作為抽出) の写真データから学習を行うことで、誤り率 14.3% を達成したと報告している。図 9.3 左に示すように、「リビング」など人間による分類でも判断が揺れがちな種別では精度が低いものの、「キッチン」や「風呂」などではきわめて高い精度を達成している。誤り例 (図 9.3 右) を見ても、右上の例ではユニットバスの洗面台 (正解は「洗面」) を「風呂」と分類しているなど、複数の種別に分類されるとも考えられる例が少なくなかった。

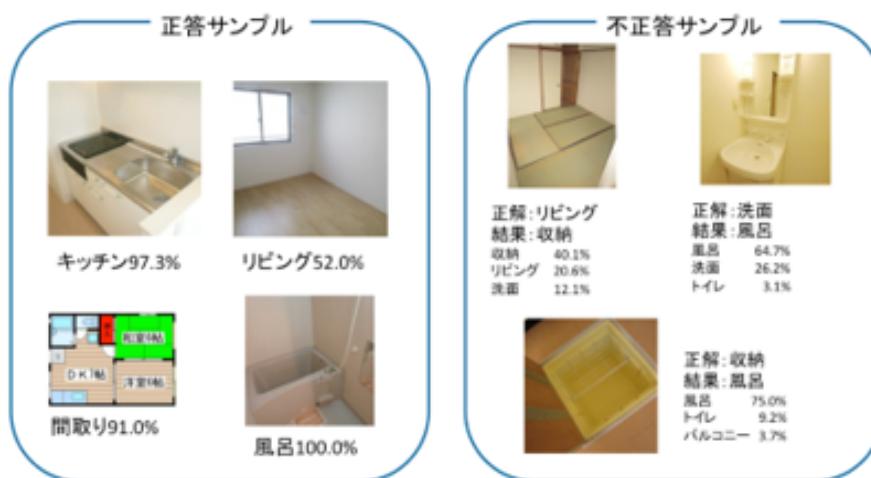


図 9.3 物件写真種別の深層学習による判別例

上記に述べた通り、深層学習による不動産物件写真の分類は現時点でも人間並みかそれ以上に高い精度を達成していることから、ビジネス現場での応用事例も報告されつつある。LIFULL HOME'S では、不動産会社から入稿される不動産物件写真のカテゴリ不整合を検出する仕組みを 2016 年 12 月より運用し

^{*2} 間取り、地図、玄関、居間、キッチン、風呂、トイレ、洗面、収納、設備、バルコニー、エントランス、駐車場の 13 種類。

ている [15]。LIFULL HOME'S は、ユーザーにとってより有益な情報を提供するという観点から、室内写真がより多く登録されている物件を検索結果において優先的に表示する仕組みをとっているが、図 9.4 の一番下の写真のように、なかには室内写真以外の写真が室内の種別で登録されるなどの不整合が起きていることが課題となっている。そこで、深層学習によって図 9.4 に示すような整合率の自動算出を行い、登録種別と不整合となっている写真については、登録元の不動産会社に是正を促すようにしている。

入稿画像	登録項目	整合率	判定結果
	キッチン	キッチン 70.3% 玄関 23.7% 設備 2.804%	○
	キッチン	キッチン 97.3%	○
	キッチン	キッチン 0% 玄関 0%	✗

図 9.4 深層学習による物件写真のカテゴリ不整合検出

9.4.2 写真解析による不動産物件情報の付加価値向上の試み

不動産物件を探すユーザーのニーズの多様化に合わせて、不動産情報サイトにおいても、「カウンターキッチン」「プロードバンド接続」「コンビニが近くにある」など、多種多様な検索条件を追加するなどの対応が行われている。しかし、物件の住みやすさに関する要素はあまりに多岐にわたることから、データベースの整備はニーズの多様化に追いついていないのが現状である。

そこで、物件写真から住みやすさに関連する指標を抽出することで、不動産物件情報の付加価値を向上させようという試みが行われている。石田ら [14] は、住みやすさに大きく影響する「キッチンの使いやすさ」に着目し、「キッチンの種類」と「ワークスペースの広さ」の 2 種類の指標を深層学習によって判別する実験を行っている。前者については、「システムキッチン」「簡易型システムキッチン」「非システムキッチン」「キッチン部位」「その他」の 5 種類に分類したデータセット（各種類 1000 点、計 5000 点の写真で構成）を作成し、CNN で学習することによって、誤り率 11.6% という高い精度を達成している。後者については、「とても狭い」～「とても広い」に「その他」を加えた 6 種類にカテゴリ分類したデータセット（図 9.5 上、計 5500 点の写真で構成）を作成し、同じく CNN で学習を行っている。カテゴリ判別の誤り率は 36.2% とそれほど良くないものの、混合行列（図 9.5 左下）でみるとある程度広さを識別できていることがわかる。各々のカテゴリに広さスコアを割り当てて相関係数を算出すると 0.717（図 9.5 右下）であり、データセットの拡充によって実用的な精度になることが期待できる結果となっている。

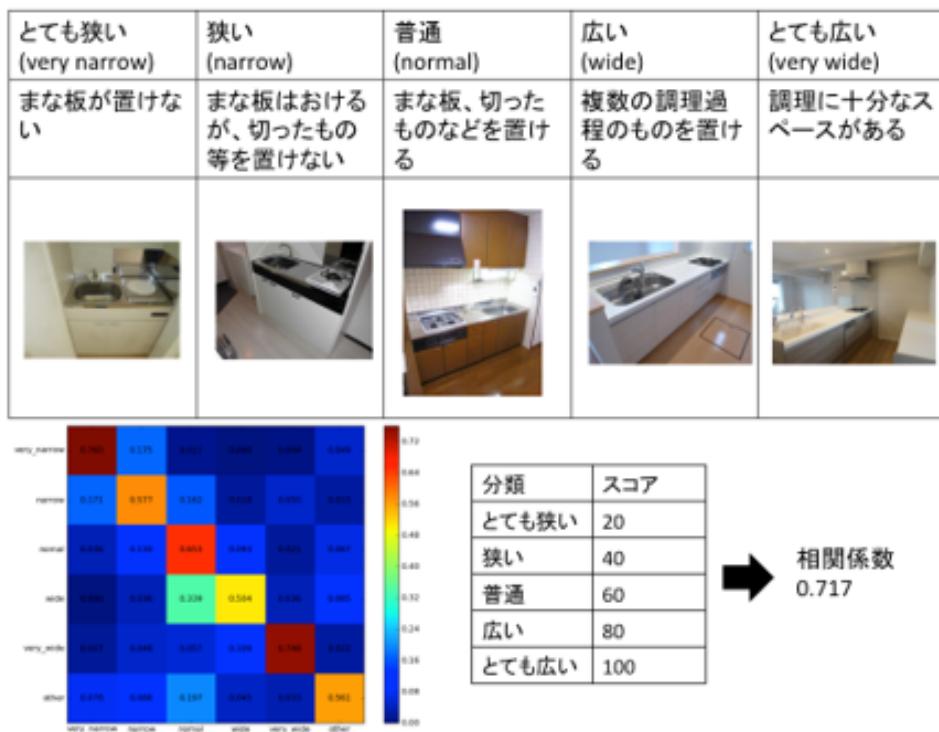


図 9.5 キッチンのワークスペースの広さ判別

9.5 質の高い不動産情報データベースの構築

本章では、不動産市場における AI ビジネスの創出の基盤となる不動産情報データベースの発達の経緯を、コンピュータの実用化以来のデータベース技術および情報アクセス技術の発展の歴史と関連付けて紹介するとともに、その果実としての大量の不動産画像情報や深層学習を、新たなビジネス価値創出に生かそうとする取り組みにも触れた。

データベース技術や情報アクセス技術などのコンピュータ科学の知見、および MLS などの標準化された不動産データベースの整備の嘗みが、不動産ビジネスを大きく発展させる土台となった経緯は、今後の不動産市場における新たなビジネス創出を考える上で大きな示唆を含んでいるように思う。産学官の枠組みを超えて、より質の高い不動産情報データベースを構築する努力を続けられるかどうかが、今後の不動産市場における AI ビジネスの成否を大きく左右するのではないだろうか。

参考文献

- [1] John W. Young, Jr. and Henry K. Kent. An abstract formulation of data processing problems. In Preprints of papers presented at the 13th national meeting of the Association for Computing Machinery (ACM '58). ACM, New York, NY, USA, pp. 1-4. 1958. DOI=<http://dx.doi.org/10.1145/610937.610967>
- [2] Cornelius T. Leondes. Database and Data Communication Network Systems: Techniques and Applications. p. 7. 2002.
- [3] E. F. Codd. A relational model of data for large shared data banks. Communications of the ACM, 13(6):377-387, 1970
- [4] H. P. Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, (2):159–165, 1958.
- [5] Ruth Atwood. Grass-Roots Look at MEDLARS. Bull Med Libr Assoc., 52(4):645–651, 1964.
- [6] G. Salton, editor. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, 1971.
- [7] Ira S. Nathenson. Internet Infoglut and Invisible Ink: Spamdexing Search Engines with Meta Tags. Harvard Journal of Law & Technology, 12(1), 1998.
- [8] Jing Han, Haihong E, Guan Le, and Jian Du. Survey on NoSQL Database. in Proceedings of the 6th International Conference on Pervasive Computing and Application, 2011.
- [9] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab, 1998. available from <<http://ilpubs.stanford.edu:8090/422/>>, (accessed 2019-08-20).
- [10] Jonathan L. Herlocker, Joseph A Konstan, Al Borchers, and John Riedl. An algorithmic framework for performing collaborative filtering. in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 230-237, 1999.
- [11] National Association of Realtors. “Multiple Listing Service (MLS): What Is It”. <https://www.nar.realtor/nar-doj-settlement/multiple-listing-service-mls-what-is-it> (accessed 2019-11-11)
- [12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in Proceedings of Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012, pp. 1097–1105.

- [13] 菊田 遥平, 野村眞平, 李 石映雪, 小林秀, 神津 友武. “Deep Learning 技術をベースとした異常画像検出,” in 第 30 回人工知能学会全国大会 (*JSAI 2016*) 予稿集, 2016, p. 1A4-OS-27b-1
- [14] 石田 陽太, 清田陽司. “住居選択支援を目的とした不動産物件画像からの深層学習による情報抽出の試み,” in *ARG Web インテリジェンスとインタラクション研究会第 8 回研究会 予稿集 (ARG WI2 No. 8)*, 2016, pp. 29–30.
- [15] 株式会社 LIFULL, “AI による物件の不整合画像検出を開始,” 2016. [Online]. Available: <http://lifull.com/news/7529/> [Accessed: 11-Apr-2017].
- [16] 国立情報学研究所 IDR 事務局, “情報学研究データリポジトリ LIFULL HOME'S データセット,” 2015. [Online]. Available: <http://www.nii.ac.jp/dsc/idr/lifull/homes.html> [Accessed: 13-Apr-2017].

第 10 章

官民ビッグデータを用いた空き家分布把握手法の開発

10.1 わが国における空き家の増加とその問題背景

不動産テックは、民間の不動産市場だけでなく、公的部門が直面する社会課題にも活用される事例が出てきている。ここでは、官民データを用いた空き家の空間的な分布の捕捉手法について紹介する。

近年、日本では人口減少や高齢化、都市部への人口移動などを背景に、全国で空き家が増加している。総務省統計局「住宅・土地統計調査」によると、平成 30 年の日本全国の空き家数は約 846 万戸、空き家率は 13.6% に達しており、空き家数・空き家率ともに過去の調査から比較しても、一貫して増加が続いている状況にある。なかでも、「その他の住宅」(別荘などの一時的に利用実態がある住宅や、賃貸用・売却用の住宅以外の住宅)の増加は著しく、平成 20 年調査から平成 30 年調査までの 10 年間に約 268 万戸から約 347 万戸へと約 1.3 倍に増加している。一部の管理不十分なその他の住宅は、腐朽・破損による倒壊危険性を有するだけでなく、地域の防犯性の低下や景観の悪化にもつながる。このような空き家は「特定空き家」とよばれ、近隣住民や地域全体に深刻な影響を及ぼす可能性が高いことから、特定空き家を含む空き家の実態把握はわが国にとって急務となりつつある [1, 2]。

こうした背景を受け、平成 27 年 5 月から「空家等対策の推進に関する特別措置法(空家等対策特措法)」が全面施行された。同法の施行により、自治体は空き家の所有者への適切な管理の指導や、空き家跡地の活用促進、特定空き家に対する助言・指導・勧告・命令、さらには罰金・行政代執行も可能となり、空き家の活用・除却といった行動を法的根拠に基づいて実施することが可能になった [3]。しかしこれらの行動を起こすためには、まずは既存自治体のどこにどの程度空き家が分布しているのか、という情報を把握する必要がある。しかし空き家の空間的分布を把握する手法は、現状では一棟一棟を個別に訪問し外観を見て判断する戸別目視が中心となっている。また、現地調査を実施する前に、空き家が数多く分布すると考えられる地域を予め把握する手法も確立されていないため、広域の空き家分布を継続的に把握し続けるには多大な労力と時間、そして費用を要する。これが自治体において空き家対策の取り組みを進めていく上で大きな障壁となっている。すなわちこれらの調査を「迅速」、「安価」かつ「継続的」に実施可能な手法の確立が期待されている。

10.2 既存の空き家分布把握の手法

空き家の空間的分布の把握を試みた研究はこれまでに数多く見られる。最も一般的な手法は、先述した現地調査により建物1棟1棟を個別に目視判定する手法[4, 5, 6]である。また現地住民や自治会などへの聞き取り調査によって空き家の分布を把握する手法[6, 7]もみられる。これらの手法は現地の状況を直接目視した結果や、地域の現状を熟知している住民等からの情報に基づいているため、建物ごとの空き家判定を高い信頼性を持って行うことができる。しかしこれらの手法を広域に亘って適用するには、相当な時間、労力、費用を要してしまうという問題がある。

一方、自治体データなどを活用した、広域を対象とした空き家分布把握の試みも見られる。例えば水道の閉栓情報を用いた自治体全体の空き家分布把握の例がある[8, 9]。ただし、これらの手法では水道が「閉栓」、「休止中」の物件全てを暫定的に空き家と定義しており、その根拠が明らかではない。また最終的に空き家か否かの特定を行うために大規模な現地調査を行っているため、時間及び費用面での問題を解決しているとはいえない。

以上を含め、先行研究では空き家の分布調査が様々な方法で実施されている[10]。また自治体においてもその取り組みは見られるものの多くの場合、水道の使用量や閉栓状況のみを用いて空き家の候補地を探すに留まっている[11, 12]。以上のように、自治体全域という広域の空き家の分布状況を、迅速かつ安価に、そして継続的に把握・推定し続ける手法はこれまでのところ確立には至っていないと言える。

10.3 空き家分布把握に有用なデータ

以上の課題解決を図るためにはどうすればよいだろうか。この課題に対して、筆者らは産官が保有する様々なデータ、特に建物ごとの情報把握に繋がる空間情報を統合して活用・分析することで、迅速かつ安価な空き家分布の把握・推定手法を確立し、空き家分布の継続的な把握につなげたいと考えている。また自治体における空き家の分布調査の実施およびその支援を実現するためには、可能な限り自治体が保有しているデータを用いて実現できる手法を確立することが望ましい。本章では10.4節「鹿児島県鹿児島市の事例」、10.5節「群馬県前橋市の事例」において具体的な取り組みを紹介するが、本節ではこれらの取り組みで活用した空き家分布把握に有用なデータを紹介する。

① 住民基本台帳（「官」保有）

居住者が住民登録を行う際に作成される住民票を編成したものであり、各市町村が必ず保有している情報である。住民基本台帳には氏名、生年月日、性別、続柄、住所などの情報が格納されている。ただし本章で紹介する事例では、個人を直接特定できないように居住者の氏名や生年月日、続柄などは秘匿されており、居住者の年齢、世帯人員数などのみを使用している。

② 水道使用量情報（「官」保有）

各契約世帯に対し、検針月ごとに収集された水道使用量に関する情報である。自治体によっては開栓・閉栓情報も有している。長期に渡り水道使用量が少ない場合や閉栓日から年月が経過している場合には、対象建物が空き家であることの参考となる。しかし、井戸水の利用が一般的である地域

では居住世帯の上水道利用が少なくなるか、あるいは上水道の契約をしていない場合もあるため注意が必要である。

③ 建物登記情報（「官」保有）

ここでいう「建物登記」とは不動産登記のうち、建物に関する情報のことである。不動産登記とは、不動産（土地・建物）の物理的な現況と権利関係を公示するため登記簿に登記することであり、土地と建物それぞれ独立した登記簿に登記される。建物に関する不動産登記（建物登記）からは、建物ごとの登録年月日、建物所在地の情報（住所）、家屋番号、建物の種類（用途）、構造、床面積などを知ることができる。

④ 家屋固定資産税台帳（「官」保有）

各市町村が固定資産の状況及び固定資産税の課税標準である固定資産の評価を明らかにするために備えるべき台帳であり、土地と家屋に分かれて存在する。家屋固定資産税台帳は住所、氏名、建物用途、建物構造、建築年などが登録されている。なお住所は地番表示であることに注意が必要である。

⑤ 地番図（「官」保有）

住所は住居表示と地番表示の二通りが存在し、後者は固定資産税台帳の管理のために自治体が保有している場合がある。地番図では正確な地番区画、町名などの情報が格納されており、地番表示の自治体データと対応させることが可能である。

⑥ 住宅地図（「民」保有）

住宅地図を提供する民間企業はいくつか存在するが、例えばゼンリン住宅地図は行政界、建物、道路縁、鉄道などのデータが全国規模で詳細に整備されている。さらに、建物データには建物ごとの建物名（表札・看板等）、建物種別、階数の情報などが格納されている。

上記の自治体・民間データは空き家の分布を推定する際に有用な情報を提供してくれるが、各自治体の特性を理解し、分析手法を確立させなければその効果は十分に発揮されない。次節以降、上記データを活用した空き家の空間分布把握の事例について紹介する。

10.4 鹿児島県鹿児島市の事例

本節では、自治体からの協力を得ることができた鹿児島県鹿児島市（以下、鹿児島市）を対象に、主に自治体データを利用して空き家の空間分布の把握を試みた秋山ほか（2019）[13] および Akiyama et al. (2020)[14] の事例を紹介する。具体的には鹿児島市が保有する複数の自治体データと民間データ、また一部地域で実施した現地調査の結果に基づいて、建物ごとの空き家率を推定するモデルを構築し、その結果を任意の集計単位（丁目、メッシュなど）で集計化することで、市域全体を対象に空き家数や空き家率を推計し、その結果を可視化することが可能となった。なお本事例で対象とする空き家は戸建住宅の空き家とし、共同住宅の空き部屋は対象外とする。また地方自治体で活用することを前提としているため、出来るだけ簡便な手法で広域を対象に迅速に処理可能な手法とすることも重視している。

なお鹿児島市は人口約 60 万人を有する比較的大規模な地方中核市であり、大都市的な様相を呈する都市部から、住宅地、工業地域、郊外型の住宅団地、中山間地域といった多様な特性を持つ地域を含んでい

る。また最新の国勢調査による高齢化率は24.8%（日本全国平均27.7%），最新の住宅・土地統計調査による空き家率は13.9%（日本全国平均13.6%）と，日本全国の平均的な動態と近くわが国の一般的な都市の特質を有すると期待される。そのため分析対象として相応しいと考えられる。

10.4.1 現地調査の実施と利用データ

まず空き家となる戸建住宅の特徴を把握するために，鹿児島市の一地域を対象に現地調査を実施するとともに，空き家分布把握手法を開発するためのデータベースの開発（以下，空き家データベース）を行った。

まず図10.1に示す鹿児島市の一地域を対象に現地調査を行い，戸建住宅ごとに空き家か否かの判定を行った。現地調査時における空き家か否かの判定方法は，国土交通省が作成した空き家判定の指針を元に作成された上田ほか（2016）[15]の手法に基づいて実施した。その結果，鹿児島市の現地調査地域における戸建住宅7,350棟のうち，空き家は353棟（約4.8%）であった。

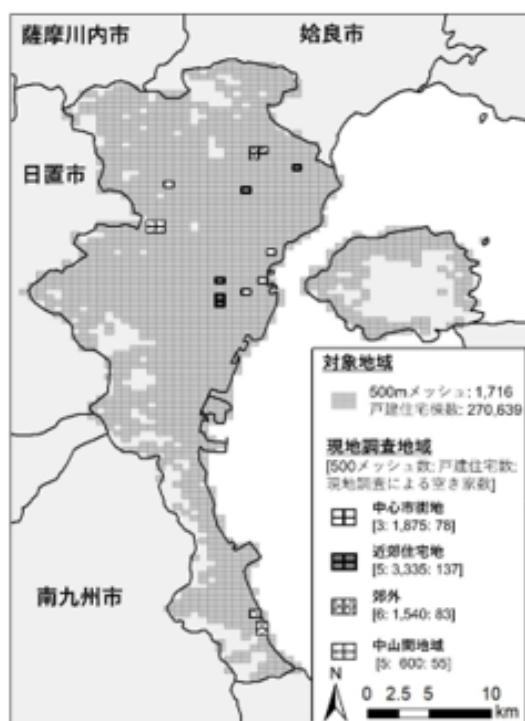


図10.1 分析対象範囲と現地調査地域

続いて鹿児島市全域を対象に，「空き家データベース」を作成した。空き家データベースは自治体保有のデータである住民基本台帳（2016年6月），過去水道使用量（2016年6月から過去5年分の月別情報），建物登記情報（2016年1月）と現地調査結果（2016年8月実施分）を，民間保有の住宅地図（2016年版）に収録された建物データに結合することで作成した。なお自治体データが持つ位置情報の多くは，住宅地図に直接結合できる経度緯度といった定量的な位置情報ではなく住所であるため，自治体データが持つ住所を全てアドレスマッチング（住所を経度緯度に変換する処理）することで，住宅地図との結合を実

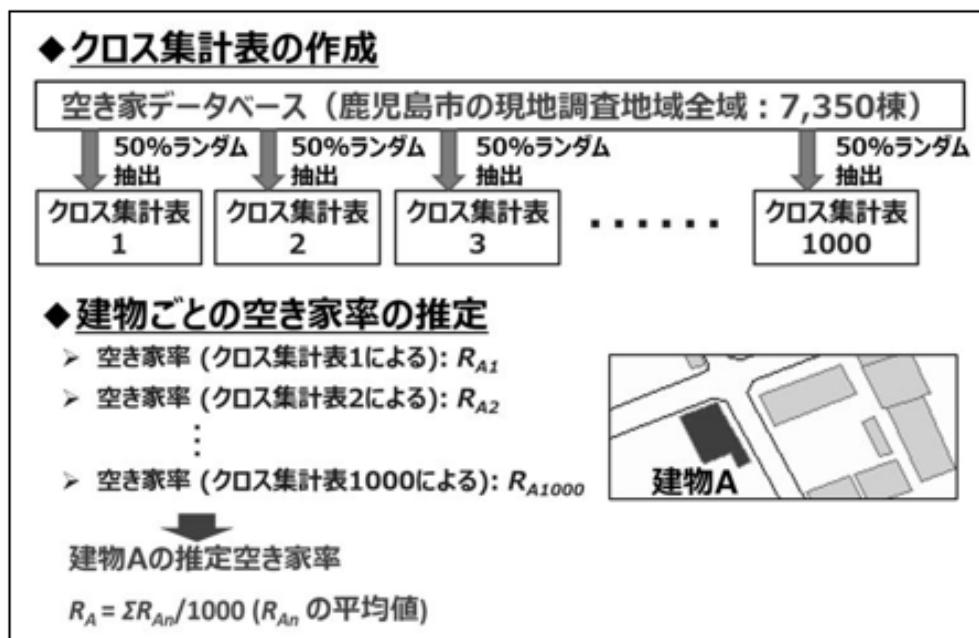
現した。

10.4.2 空き家分布把握のためのモデル構築

続いて鹿児島市の現地調査対象地域の結果を用いてクロス集計表を開発し、同クロス集計表を用いて空き家分布把握のためのモデルを構築するとともに、推定結果と真値を比較することで、同手法の信頼性を検証した。

まず各説明変数のセル分割を行う。なおここでいう「セル」とは、クロス集計表の1つ1つの組み合わせのことを言う。質的変数である建物用途や建物構造などはセル分割を行う必要は無いが、それ以外の量的変数はセル分割を行う必要がある。そこで全ての量的変数において様々な区切り方でセル分割を行い、サンプル数が少なかったり、空き家率が類似した結果になったりするセル同士を結合することで計算量を圧縮し、最適なセル分割方法を決定した。また鹿児島市におけるクロス集計表では最大で15種類の変数を用いることが出来るため、15種類の説明変数を組み合わせてクロス集計表を作成し、最も真値と一致するクロス集計表の選択を行った。その結果、15種類全ての変数を利用したクロス集計表を用いる場合が最も信頼性が高くなることが明らかになった。

続いて現地調査地域の全ての戸建住宅(7,350棟)を用いて建物ごとの空き家率を推定するモデルを構築した。ここでは現地調査地域の全ての戸建住宅からランダムに50%を抽出してクロス集計表を作成し、同様の処理を1000回繰り返して1000種類のクロス集計表を作成した。そしてこれらを用いて建物ごとに1000種類の空き家率を算出し、それらの平均値を最終的な空き家率として採用した(図10.2)。これは過学習を防ぐための措置である。



最後に建物ごとに与えた空き家率の推定値を図10.3の方法で集計化することで、任意の空間単位の空

き家棟数と空き家率を明らかにすることが出来る。

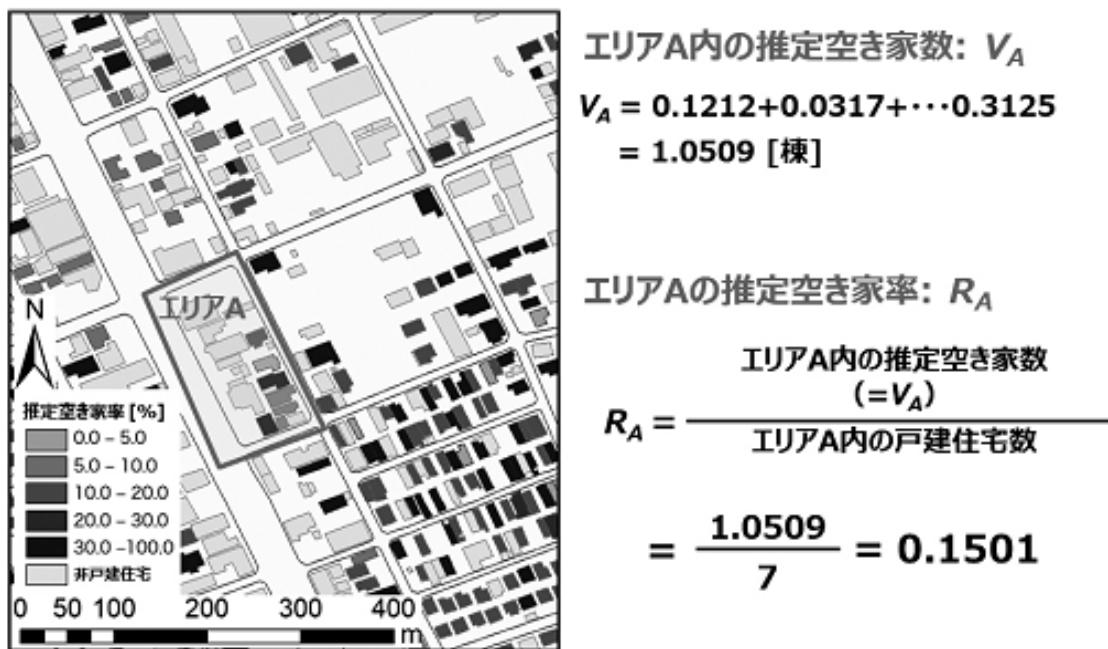


図 10.3 任意の空間単位における空き家棟数と空き家率の推定方法

表 10.1 に本研究の手法で推定した現地調査地域における 250m メッシュ , 500m メッシュおよび番地ごとに空き家数と , 現地調査で得られた空き家数の相関分析の結果を示す。250m メッシュや番地単位といったかなり細かい空間単位で集計して比較しても充分に高い相関が見られることが分かった。また平均絶対誤差も小さいことから , 本研究の手法で推定した空き家数と現地調査で得られた真値との間の平均的な誤差も小さいことが分かった。

表 10.1 集計単位ごとの空き家数の真値と推定値の比較結果（相関分析）

	集計単位		
	250m メッシュ	500m メッシュ	番地
決定係数 (R ²)	0.8520	0.9121	0.8229
自由度調整済決定係数 (R ^{2f})	0.7259	0.8599	0.7009
平均絶対誤差 (MAE)	1.7876	1.1201	1.8864
p 値	2.23E-26	4.56E-58	5.20E-24

10.4.3 鹿児島市全域の空き家分布推定結果

図 10.4 に 10.4.2 節で開発した 1000 種類のクロス集計表を用いて推定した , 鹿児島市全域における 4 次メッシュ (500m メッシュ) 単位の推定空き家棟数と推定空き家率を示す。鹿児島市では中心市街地 (図 10.4 の A) 周辺の住宅街や , 吉野町 (図 10.4 の B 周辺), 喜入中名町 (図 10.4 の C 周辺), 喜入生

生町（図 10.4 の D 周辺）、桜島地区の錦江湾沿岸地域（図 10.4 の E 周辺）などで特に推定空き家数が多くなっていることが分かった。また空き家率は市北部や西部の山間部で特に高い値となっていることが分かった。さらに喜入生生町では空き家数、空き家率ともに高い値であった。

以上のように本研究の手法で鹿児島市全域の空き家数と空き家率の空間的分布を推定することが可能となった。ただし本研究の結果は推定値であるため、現実の空き家数・空き家率を市域全体で必ずしも正確に示しているものではない可能性がある点には留意されたい。

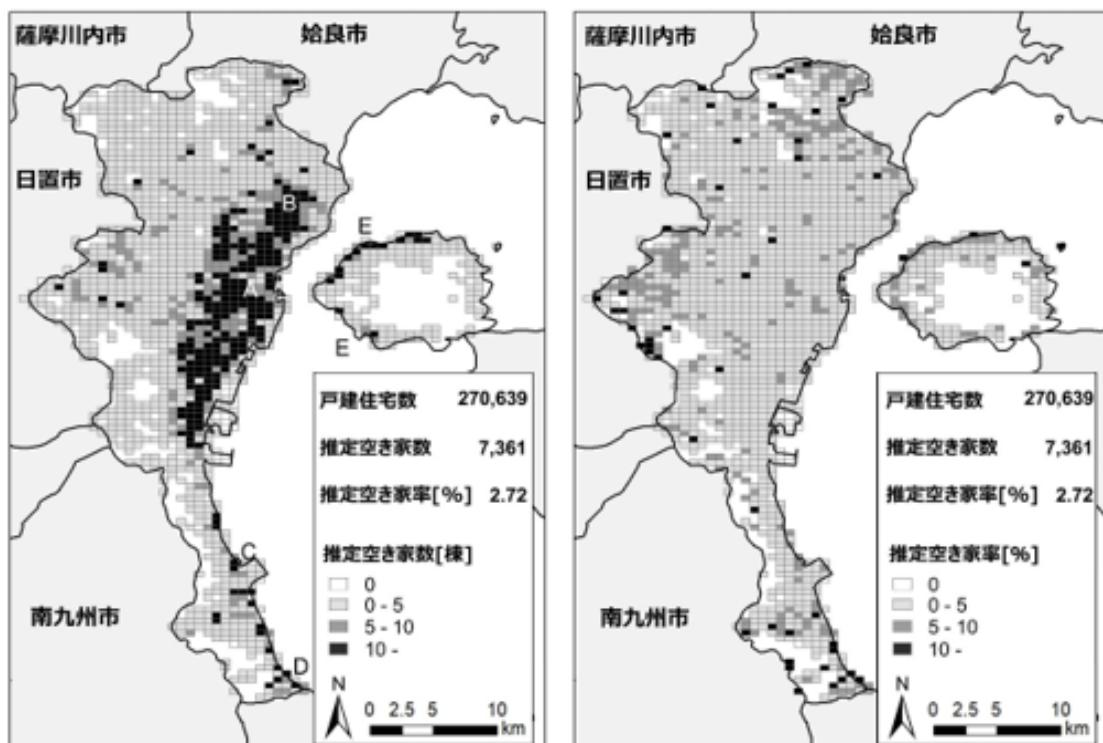


図 10.4 鹿児島市全域の空き家数（左）と空き家率（右）の推定結果（500m メッシュ単位 2016 年の場合）

10.4.4 小括

本事例では自治体における空き家分布調査の実施およびその支援を目的として、自治体データを活用することで空き家の分布把握を出来るだけ簡便な手法で実施するための手法開発の取り組みを紹介した。まず鹿児島市の一部地域の現地調査結果と複数の自治体データを組み合わせて空き家データベースを構築した。続いて空き家データベースから複数種類のクロス集計表を作成し、それらを用いて各建物の空き家率を推定した。さらにその結果をメッシュ等で集計化することで、鹿児島市全域を対象に任意の空間単位で空き数や空き家率を推定することが可能になった。

今後は本手法が他の自治体においても適用可能かどうかの検討を進める必要がある。また空き家に関する現地調査データの収集をさらに進めることで、推定精度の向上が期待される。そのためには現地調査を効率的に行うためのスキームやアプリケーションの開発も必要になるものと考えられる。今後も以上の課

題を考慮しながら、様々な自治体と連携しつつ、引き続き本研究の手法の改良を図るとともに、現地調査データのさらなる充実と蓄積を進めていくことで、本事例で紹介した手法を更に洗練させていきたいと考えている。

10.5 群馬県前橋市の事例

続いて本節では、群馬県前橋市（以下、前橋市）における自治体データを利用して空き家の将来分布を予測する試みについて馬場ほか（2019）[16] を参考にして述べる。具体的には、過年度の自治体データから空き家推定モデルを構築し、将来時点での空き家予測確率の地図を作成というものである。そして最終的には、自治体における空き家現地調査のためのツールとして実装することを狙いとしている。

対象地域は自治体からの協力を得ることが出来た前橋市とする。前橋市は東京都心から約 100km 圏内に位置し、東京への人口流出が顕著な中核市である（図 10.5）。市はこのような状況を鑑み、中心市街地活性化基本計画において中心市街地における空き家への対応を重要な項目の 1 つとしている。このように、前橋市は大都市圏に隣接する中核市として一般的な特質を有すると考えられ、分析対象として相応しいと考えられる。本節では前橋市の中でも中心市街地活性化基本計画で空き家への対応を重点的に行うとされている市の中心部を研究対象地域とする。



図 10.5 分析対象範囲の位置図

10.5.1 利用データと手法

前橋市の協力のもと、自治体データとして住民基本台帳、水道使用量情報、家屋固定資産税台帳を収集した。住民基本台帳は 2013 年 11 月 1 日現在、2017 年 3 月 31 日現在の 2 時点のデータを収集した。水道使用量情報は月ごとの使用量情報から 2014 年度及び 2018 年度をそれぞれ集計して利用した。固定資産税台帳（家屋）は 2013 年 1 月 1 日現在、2018 年 1 月 1 日現在の 2 時点を収集しており、いずれかの時

点で存在する家屋を全て抽出した。また住民基本台帳、水道使用量情報には住居表示で住所が併記されているため、東京大学空間情報科学センターが提供する CSV アドレスマッチングサービスを利用して緯度経度座標を与えることができた。ところが、家屋固定資産税台帳は地番での住所が掲載されており、上記アドレスマッチングが必ずしも有効であるとはいえないかった。そこで、前橋市から地番図を提供頂き、地番図住所と完全一致したものを利用した。なお、データの集計は建物単位で行っており、2016 年のゼンリン住宅地図の建物ポリゴンに自治体データをリンクさせた。

教師データとなる空き家の空間分布に関するデータは、前橋工科大学が 2015 年 7~9 月及び 2016 年 2 月にかけて実施した中心市街地の現地調査結果を用いた。同調査では空き家評価を損傷の程度や流通されているか等で分類しており、どの評価を受けた空き家を空き家データとして利用するべきか、という点で議論の余地がある。なお本研究では前橋市から市場に流通するような空き家の分布についても把握したいとの要望があり、流通の有無に関わらず空き家と判定された物件全てを教師データとして利用した。空き家の現地調査結果は緯度経度座標を直接得ており、522 件の空き家が建物ポリゴンと対応した。最終的に、いずれかの自治体データと結合した建物ポリゴンは 2,111 件であり、本節ではこれらを分析対象とした（表 10.2）。

表 10.2 自治体データのアドレスマッチング結果と建物との対応

	自治体データ			
	住民 基本台帳	固定資産 税台帳	水道 使用量	空家 現地調査
サンプルサイズ	10,572	9,089	9,997	534
建物重複削除後サンプルサイズ		5,137	2,551	
アドレスマッチング成功サンプル	9,375	4,044	2,418	
補足率 [%]	88.7	78.7	94.8	
建物と結合出来たサンプル	9,093	4,044	2,293	522
全数からみた採用率 [%]	86.0	78.7	89.9	97.8
対象範囲内の住宅建物数	2,587			
いずれかの自治体データと結合した住宅建物数	2,111			
建物と対応した自治体データ数	1,161	1,937	1,227	328

将来空き家分布予測の概念図を図 10.6 に示す。図 10.6 は t 時点の特徴量と $t+1$ 時点の教師データを用いてモデルを構築し、構築したモデルに k 年後の特徴量を投入するイメージを表現している。なおこのモデルは k 年間で空き家の形成傾向が変化しないという仮定のもとで成立するものであり、3~5 年程度の短期間に空き家の形成傾向に大きな変化はない、という前提を設けているといえる。

推定に際して、機械学習的分類手法のひとつである XGBoost (eXtreme Gradient Boosting) [17] を利用して空き家の分類を行った (Chen and Guestrin, 2016)。これはアンサンブル学習のひとつで、ランダムフォレストを基本として弱学習器である決定木間の重み付けに勾配ブースティングという手法を合わせたものである。当該手法は推定精度が高く、欠損値が混入していても処理が可能である。しかしながら、推定精度向上のためにはパラメータチューニングを行う必要があり、複雑なデータで汎化性能が下が

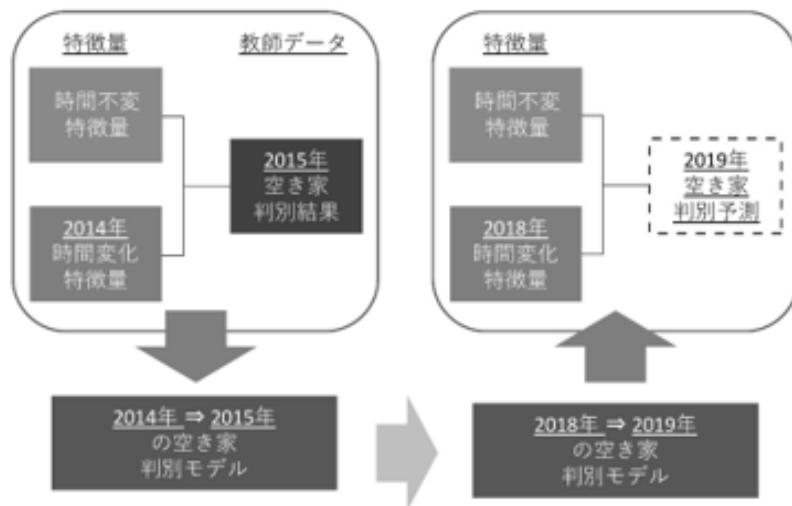


図 10.6 将来空き家分布推定モデルの概念図

るという欠点がある。

学習の際、時間不变特徴量である家屋固定資産税台帳は 2013 年、2018 年から重複するものを除いて全て利用し、時間変化特徴量について 2013 年住民基本台帳、2014 年度水道使用量情報を用いてモデルを構築した。続いて、2015 年空き家現地調査を教師データとしてパラメータチューニングを行った。その後、構築されたモデルに 2017 年住民基本台帳、2018 年度水道使用量情報を投入して 2019 年空き家率予測確率を算出した。具体的な特徴量は、建物人員内最高年齢、最少年齢、建物人員数、住基有無ダミー、建物構造ダミー、建物用途ダミー、建物面積ダミー、水道使用量である。特徴量の二乗項及び交差項もモデルに含め計 144 の特徴量を用いた。

10.5.2 将来空き家率分布予測の構築

はじめに、グリッドサーチによるパラメータチューニングを行ってモデルを構築した後、2015 年現地調査結果と検証データによる将来空き家の判別結果をクロス集計し、モデルの妥当性を交差検証した。ここでは全サンプルの 7 割を訓練データとしてモデルを構築し、残り 3 割を検証データとして分類した。検証データ 634 件のうち、将来空き家の予測が的中したのは 530/634 (83.6%) であり、一方で空き家でないと予測されて現地調査で空き家である割合は 53/566 (9.4%) であった。後者は、予測結果に基づき現地調査対象を定める際には最小化したい誤差であるが、データ数が十分でないことや今回の空き家の定義が広範であることが原因であり、十分な精度を出せていない可能性がある。なお、これは 2015 年時点でのモデルの予測結果を他地域に外挿した結果をみており、将来空き家分布予測に対する妥当性を検証している訳ではない。

続いて、構築されたモデルを用いて 2019 年時点の空き家予測確率の推定を行った。図 10.7 は 2015 年現地調査結果と将来空き家確率の地理的分布を示している。空き家予測確率の高い建物は地理的に分散しており、必ずしも空き家の空間的分布が地理的法則性をもつとは言えない。これは、住民基本台帳データから抽出される年齢等の情報が有効に作用する一方で、地理的な規則を持って分布しないためであると考えられる。

えられる。図 10.7 の分布を比較すると、中央前橋駅周辺では 2015 年時点の空き家分布を反映しているようみえるが、JR 前橋駅北西部では必ずしも空き家予測の分布が対応しているとは言えない。これは、対象地南東部で建物人員内最高年齢などの特徴量に変化があったためであると考えられる。

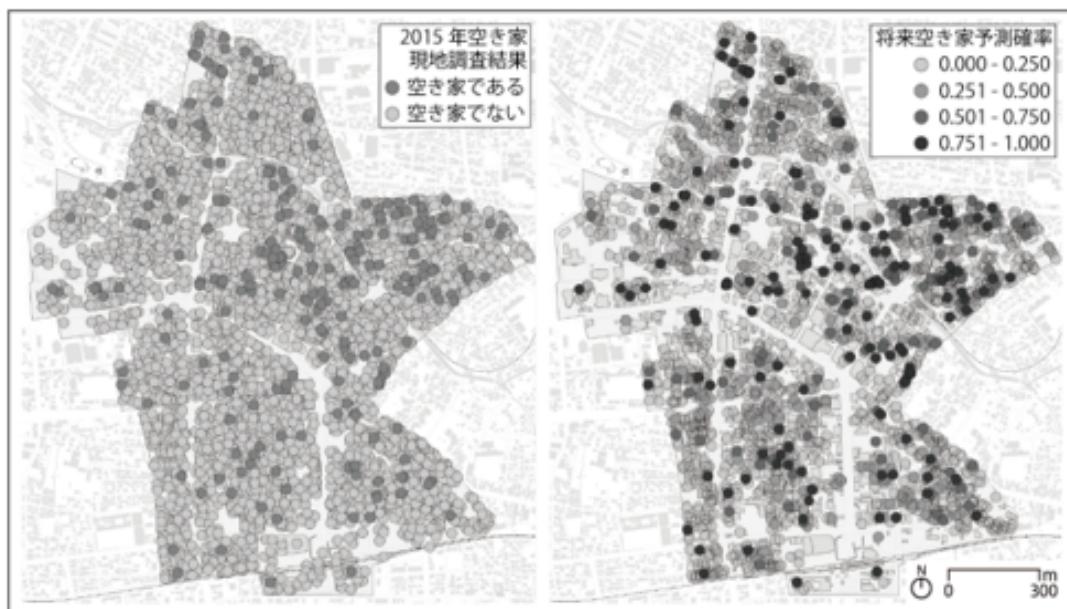


図 10.7 2015 年現地調査結果と将来空き家確率の地理的分布

10.5.3 小括

本節では自治体データから将来空き家分布を予測するため、XGBoost を用いて 2015 年の空き家分布を教師データとして 2019 年の空き家分布を推定した。このような手法を用いることで、欠損値が多く発生する自治体データでも容易に分析を行うことができ、また R の”xgboost”パッケージなどで既に実装されたものを利用可能である。一方、本節で紹介した前橋市での取組みはまだ発展段階であり、今後は現地調査を行い、本節の分析から得られた将来空き家分布の信頼性について再度検証する予定である。さらに、本節の研究成果の活用方法について前橋市と議論を行っており、老朽空き家や活用可能な店舗の特定など、今後の活用が期待されるところである。

10.6 空き家は予測できるのか?

本章では様々な空間情報を統合的に活用することで、空き家の空間的分布を把握するための最新の取り組みを紹介した。鹿児島市、前橋市何れの事例においても、主に自治体が保有する様々な空間情報を組み合わせてモデルを構築することにより、高い精度で空き家の分布状況を推定することが可能になりつつある。しかも空き家の教師データは一部地域の一時点の現地調査結果のみながらも、自治体全体の空き家の分布状況や、将来の空き家分布状況をデータからごく短時間で推定できるようになりつつあり、これは先

述した今自治体が求めている「迅速」、「安価」かつ「継続的」な空き家分布の把握手法につながるものとなる可能性を秘めていると言えよう。

ただし本章で紹介した多様な自治体データを活用するためには、基本的には自治体内における個人情報保護審査会（自治体によっては審議会）において、これらのデータ利用についての審査を通過する必要がある。そのため自治体によってはこれら的一部、あるいは全てを本章で紹介したように使用することが困難な場合も考えられる。また財政力が小さい自治体の場合、本章で使用した民間データを継続的に入手し続けることが難しい可能性もある。そのため今後は利用可能なデータの組み合わせに応じた空き家分布推定手法の開発を進めていく必要があると言えよう。

また公共データ、民間データともに十分な質・量を確保できない場合に備えて、他の手法も準備しておくべきであろう。この課題に対して筆者らはドローンを用いて上空から建物の熱分布（生活由来の排熱の分布を把握）と夜間光の分布（生活由来の照明の使用を把握）の情報を収集することで、空き家の分布状況を広域的に推定する手法の開発にも着手している[18]。また近年では高解像度のカメラを搭載した人工衛星が撮影した夜間光画像（例えば米国の Suomi NPP による VIIRS 画像など）も利用可能になりつつあり、今後更に高精細な夜間光画像の利用が可能になれば、衛星画像を用いた広域を対象とした空き家分布推定が実現する可能性もあるだろう。さらに電力会社が保有するスマートメータを活用する手法にも期待が高まっている。スマートメータは電力使用量をリアルタイムで収集しているため、月単位などの短期間でも対象地域の空き家率を可視化することができる可能性がある。さらに、水道使用量などの自治体データと組み合わせることで、より高精度な空き家分布の予測が可能になると考えられる。

以上の様に、自治体における「迅速」、「安価」かつ「継続的」に空き家の分布を把握する手法に関する研究はかなり進展しつつあるものの、まだ課題も多い。今後は様々な自治体・民間データと現地調査データを組み合わせた、また利用可能なデータによって様々なデータの組み合わせに応じた空き家の空間分布の把握手法を開発・発展させていきたい。またその過程で教師データとなる空き家の情報を蓄積し続けることで、様々な地理的条件に適用可能な手法・モデルを開発し、継続的に改善し続けていきたいと考えている。

参考文献

- [1] 浅見泰司：都市の空閑地・空き家を考える，プログレス，2014
- [2] 日本弁護士連合会法律サービス展開本部自治体等連携センター：深刻化する「空き家」問題 全国実態調査からみた現状と対策，明石書店，2018
- [3] 宮路和明・西村明宏・山下貴司：空家等対策特別措置法の解説，大成出版社，2015
- [4] 遊佐敏彦，後藤春彦，鞍打大輔，村上佳代：中山間地域における空き家およびその管理の実態に関する研究 山梨県早川町を事例として，日本建築学会計画系論文集，第 71 卷，第 601 号，pp. 111-118，2006.3
- [5] 片山直紀，海道清信，村上心，前田幸栄：空き地・空き家実態からみた郊外住宅団地の持続可能性についての考察 名古屋都市圏・可児市と多治見市における事例調査より，都市住宅学，第 2006 卷，第 55 号，pp. 70-75，2006
- [6] 久保倫子，益田理広：岐阜市中心部における空き家増加の実態，日本地理学会発表要旨集 2015 年度日本地理学会春季学術大会，pp. 247，2015.3
- [7] 矢吹剣一，西村幸夫，窪田亜矢：歴史的市街地における空き家再生活動に関する研究 長野市善光寺門前町地区を対象として，都市計画論文集，第 49 卷，第 1 号，pp. 47-52，2014.4
- [8] 山下伸，森本章倫：地方中核都市における空き家の発生パターンに関する研究，都市計画論文集，第 50 卷，第 3 号，pp. 932-937，2015.11
- [9] 西山弘泰：宇都宮市における空き家の空間的特徴，日本地理学会発表要旨集，2014s(0), 100310, 2014
- [10] 益田理広・秋山祐樹：昨今の空き家研究の盛衰と手法について 日本における研究状況に着目して，地理空間，XX，2020
- [11] 牛久市空き家対策課：平成 29 年度牛久市空き家等実態調査 結果報告書，2018. URL: http://www.city.ushiku.lg.jp/data/doc/1528847877_doc_233_0.pdf
- [12] 呉市都市部住宅政策課：「吳市空き家実態調査」及び「住宅等の状況把握に関するアンケート調査」報告書，2016.URL: <https://www.city.kure.lg.jp/uploaded/attachment/20807.pdf>
- [13] 秋山祐樹・上田章紘・大内健太・伊藤夏樹・大野佳哉・高岡英生・久富宏大，公共データを活用した空き家の分布把握手法の高度化 自治体の公共データを活用した空き家の分布把握手法に関する研究（その 2），日本建築学会計画系論文集,764, 2165-2174, 2019.
- [14] Akiyama, Y., Ueda, A., Ouchi, K., Ito, N., Ono, Y., Takaoka, H. and Hisadomi, K., Estimating the Spatial Distribution of Vacant Houses using Public Municipal Data, Geospatial Technologies for Local and Regional Development, 165-183, 2020.

- [15] 上田章紘・秋山祐樹・大野佳哉, 空き家発生・分布メカニズムの解明に関する調査研究(その1),
PRI review, 61, 24-35, 2016.
- [16] 馬場弘樹・秋山祐樹・谷内田修 (2019) 群馬県前橋市における公共データを活用した空き家分布推定
手法の検討. 地理情報システム学会講演論文集, vol.28, CD-ROM (F-2-4).
- [17] Chen, T. and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. Proceedings of the
22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794.
- [18] 秋山祐樹・飯塚浩太郎・谷内田修・杉田暁, ドローンにより収集した熱赤外画像と可視光画像を用い
た空き家分布推定手法の基礎的研究, 第 28 回地理情報システム学会講演論文集, CD-ROM(F-2-3),
2019.

第 11 章

不動産金融市場における不動産テック^{*1}

11.1 不動産投資信託（REIT）市場におけるデータ資源

不動産テックは、不動産流通での活用がしばしば紹介されるが、不動産金融市場においても、多くのサービスが開発されてきている。ヘドニック・アプローチを応用して不動産投資の利回りを予測したり、不動産物件を保有する所有者が、各不動産を売却する確率を提供したりするサービスが出現してきているが、そのようなサービスを開発していくうえで、最も利用されているデータ資源として、不動産投資信託（REIT）市場で生成され、そして開示されている物件データがある。

不動産投資信託、または REIT とは、多くの投資家から集めた資金で、オフィスビルや商業施設、マンション、物流施設、ホテル等複数の不動産を購入し、その賃貸収入や売買益を投資家に分配する枠組みである（図 11.1）。不動産の購入には膨大な資金が必要であり、個別の投資家・企業が単独で行うのにはリスクも大きい。また、金融市場と比べて不動産市場は、取引に係る制度が複雑であり、物件の個別性が強く毎期の価格・賃料の情報が不足している等の不透明性がある。REIT の枠組みは、不透明な不動産市場の透明性を高め金融市場に近づける役割を果たしているといえよう。金融市場のうち、株式はキャピタルゲインを狙い短期間で売買を繰り返すのに対し、債券は長期保有によりインカムゲインを得ることが基本となる。REIT が保有する不動産は、基本的には賃料収入の安定性を重視し債券寄りの位置づけといえるが、時機を得た売却を通して株式のようにキャピタルゲインを狙うこと也可能である。



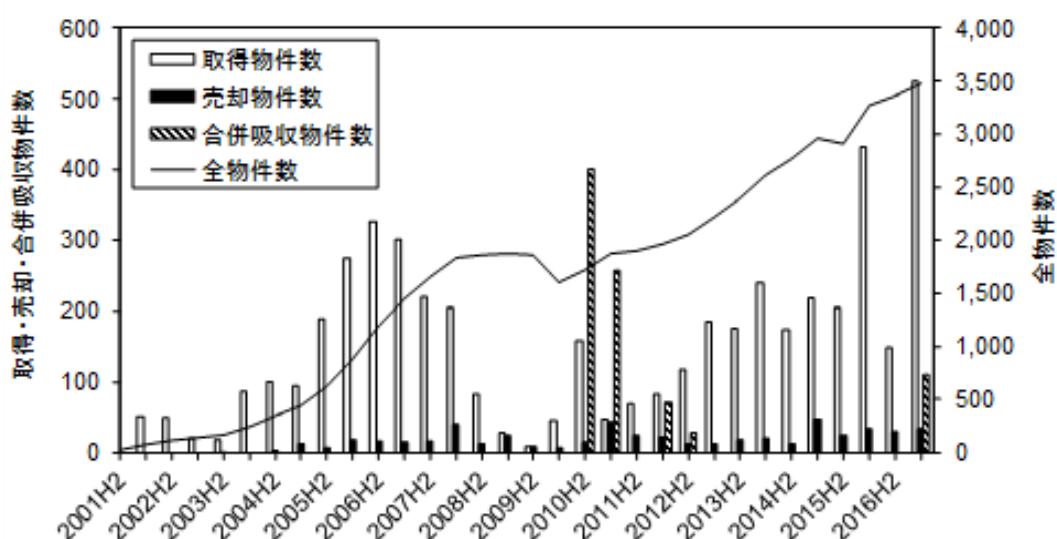
一般社団法人投資信託協会「そもそも J-REIT とは？」
(<https://www.toushin.or.jp/reit/about/what/>) を参考に作成した。

図 11.1 REIT 市場の仕組み

^{*1} 本章は、Suzuki, M., Ong, S. E., Asami, Y., & Shimizu, C. (2019). Long-Run Renewal of REIT Property Portfolio Through Strategic Divestment. *Journal of Real Estate Finance and Economics*, forthcoming を要約したものである。

REIT の成長戦略として、外部成長と内部成長がある。前者は、不動産等を追加取得することで資産規模を拡大すること、後者は、既に保有している不動産等の収益性を高めることである。物件取得後、マネジメントの改善や建物の改装等を経て利回りを高め、投資家への還元を目指す。最終的には不動産の売却益を新しい不動産の取得費用に回し、さらなる成長を目指すことになる。

日本では、国内の不動産への投資を中心とした J-REIT が 2001 年に発足した。図 11.2 は J-REIT 市場の動向を示しており、発足以来物件取得を拡大させ、2006 年前後が物件取得数の一度目のピークとなった。一転して 2008 年前後の金融危機では物件取得が困難となり、その後 2010～2012 年には REIT 間での合併に伴う物件の移転が生じた。2013 年以降は、再び物件取得数が増加し、近年では保有物件数が過去最高に達している。一方で、J-REIT の発足後数年が経過して以降、毎期一定数の物件が売却に至っている。近年では、設立後 15 年以上経過した REIT も存在することから、当初から運用してきた不動産が減価している可能性や、ポートフォリオが拡大する中で当該物件の位置づけが変化してきた可能性があり、資産入替の重要性が高まっていると考えられる。



縦軸左は取得・売却・合併吸収の物件数を、縦軸右は REIT が保有している全物件数を示す。

図 11.2 REIT 市場の動向

本稿では、第 11.2 節で、日本における REIT 市場のデータと国際的な REIT 研究の動向を概観する。第 11.3 節では、Suzuki et al. (2019)[7] に基づき長期的な資産入替行動の分析例を紹介する。第 11.4 節では、REIT 情報を用いた不動産市場分析の方向性を整理する。

11.2 REIT 市場データと REIT 研究の動向

11.2.1 銘柄情報

J-REIT については、各投資法人の決算資料等をもとに、銘柄レベルや物件レベルの各種情報が蓄積されているデータベースが存在する（例えば、東急不動産 TOREIT データベース、Japan REIT データ

ベース等がある）

J-REIT 市場では、投資家の判断に資する銘柄レベルの各種財務情報が開示されている。投資口に関する情報として、株価、時価総額、分配金・利回り等が、貸借対照表（BS）に関する情報として、資産総額、借入金等が、損益計算書（PL）に関する情報として、各期の収益・費用等が公開されている。また、格付け機関による格付け情報も含まれる。

当該 REIT がどのような物件を保有しているのかという集計情報も公開されており、オフィス・住宅等の物件タイプ、東京・首都圏等の物件所在地域の構成比率等の集計指標がある。

11.2.2 物件情報

銘柄レベルの情報に加えて、J-REIT では、個別物件の情報が開示されている。物件がポートフォリオに組み込まれている期間（取得してから売却に至るまでの期間）内は、半年（半期）毎に、鑑定評価額、収益（NOI, NCF）、費用（OPEX, CAPEX）等が記録されている。物件の取得時には取得価格が、売却時には売却価格が記録される。

なお、鑑定評価額とは、当該不動産の取引が生じなくても、不動産鑑定士が周辺の取引情報を参照したり当該物件の収益性を見積もったりしながら、毎期評価した価格である。NOI（Net Operating Income）とは、収入（主に賃料）から維持管理費、固定資産税等の日常的に発生した経費である OPEX（Operating Expenditure）を控除した純収益である。また、NCF（Net Cash Flow）は、NOI から、長期的観点での修繕費等の資本的支出である CAPEX（Capital Expenditure）を控除した金額であり、不動産鑑定評価に当たって収益還元法を用いるときには、この NCF をベースにすることが多い。

一般の不動産市場では、持ち家の場合は売買が生じるときにしか価格が観察されないが、REIT 市場では、毎期の鑑定価格・賃料収入を同時に観察できるという特徴がある。

11.2.3 REIT 研究の動向

日本以外の国では、REIT レベルでの保有物件の構成比は公開されるが、個別物件のパフォーマンスまでは公開されない。こうした背景があり、基本的には REIT レベルのデータを用いて、その他の金融市場データと組み合わせた分析が進んできた。不動産分野の国際学術誌をレビューした児島（2015）によれば、REIT 研究は、「REIT 価格・投資リターン」、「REIT 価格と実物不動産および他の証券との関連研究」、「エージェンシー問題・企業統治体制・経営者報酬」、「ディスクロージャー、会計数値と CF 数値、利益調整」、「IPO, SEO, 自社株買い、株式分割、非上場化および M&A」、「金融政策・経済状況の影響」といったトピックに分類できる。一方で、REIT に限らない投資不動産や商業不動産の動態については、国際的にも物件レベルの研究蓄積が限られており、日本の REIT 市場における充実した物件レベルのデータを通じた貢献が期待される。

11.3 長期的な資産入替の分析

11.3.1 長期的な資産入替

REIT は不動産を保有し、賃料収入から得た収益を安定的に投資家へ分配することが求められる。しかしながら、不動産には、建物部分が経年的に減価していくという特性がある。すなわち、空室が増え賃料収入は減少していく一方、維持管理に必要なコストが増加していく。そこで長期的には、こうした陳腐化した物件を売却することで資本を回収し新しい物件の取得に回していく。

そこで、Suzuki et al. (2019)[7] は、物件レベルパネルデータ（東急不動産 TOREIT データベース、2001 年下期～2017 年上期）を用いて、REIT 市場における資産入替のメカニズムや、売却直前のマネジメント戦略を検証している。日本以外の国では物件レベルのデータが存在しないことから、既往研究では、物件取得・売却が REIT のパフォーマンス（利回り等）に及ぼす影響の検証や、マクロ経済環境や目標レバレッジ水準等の REIT レベルの意思決定論理をもとにした物件取得・売却行動の分析にとどまってきた（Ooi et al., 2010[5]）。基本的に、REIT には、投資家への利回りを安定的に高めるインセンティブがある。売却によって次の投資物件への資本を回収し、ポートフォリオの効率性を高めることで、利回りを確保し銘柄の格付けを高めることにもつながる（Li et al., 2018[4]）。実際に売却される物件が、REIT レベルの論理と整合的かどうか、物件レベルでの検証が求められている。

11.3.2 保有資産の効率性

具体的には、REIT はどういった指標をもとに、保有資産の効率性を判断しているのであろうか。ここでは、3 つの観点から考える。まず、「経済的陳腐化」の観点があり、次の 2 つの指標から判断する。1 つは、運用費率（Operating Expense Ratio）であり、OPEX をグロス収入（Gross Rental Income）で割った値である（物件 i , t 期における定義を示す）：

$$OER_{i,t} \equiv \frac{OE_{i,t}}{GRI_{i,t}}$$

この指標が大きいほど、グロス収入に対し運用費の占める割合が大きく、その分 NOI が差し引かれてしまうことを示す。一般に、物件の経年減価が進むと運用費が増加していくことから、運用費率の大きい、経済的陳腐化の程度が大きい物件が売却されやすいと考えられる。

もう 1 つは、賃料利回り（Rental Yield）であり、NOI を鑑定価格（Appraised Value）で割った値である（物件 i , t 期における定義を示す）。

$$RY_{i,t} \equiv \frac{NOI_{i,t}}{AV_{i,t}}$$

この指標が大きいほど、NOI に対して鑑定価格が過少に見積もられている、すなわち、何らかのリスクにより割り引かれていることを示す。REIT は、利回りを「安定的に」一定水準に保つことが重要であり、リスクの大きい資産を手放す必要がある。よって、賃料利回りの大きい、経済的陳腐化の程度が大きい物件が売却されやすいと考えられる。

次に、「地理的な集中」の観点がある。REIT は、特定のエリアに投資を集中させることがあり、こうした投資方針は変更されることもある。投資方針に適さない物件は、売却される可能性が高まると考えられる。J-REIT の投資対象物件は東京中心区が多いことから、ここでは、東京中心区に投資を集中させている REIT において、それ以外の地域の物件が売却されやすいかどうかを検証する。具体的には、「東京中心区に投資を集中させている REIT（鑑定価格ベースで、東京中心 5 区の構成比を算出して定義）」ダミーと、地域ダミーの交差項の係数に着目する。

さらに、「キャピタルゲイン・ロス」の観点がある。当初の物件取得価格に対し、売却時における資産価格の上昇・下落を判断する必要があると考えられる。ここでは、物件を保有している間、毎期観察することができる価格として鑑定価格 ($AV_{i,t}$) を用いることで、当初の取得価格 (AP_i) に対する変化率を算出する（物件 i , t 期における定義を示す）。

$$\Delta AV_{i,t} \equiv \frac{AV_{i,t} - AP_i}{AP_i}$$

以上の 3 つの資産入替の判断指標（ポートフォリオの効率性）に関する仮説が正しいかどうか、まずは、基本統計量を確認してみたい。表 11.1 は、運用中、取得直後、売却直後のサブサンプルについて、各変数の平均値を示したものである。①売却直前の物件は運用中の物件と比べ、運用費率・賃料利回りが高く、経済的陳腐化が進んでいる物件が売却に至りやすい可能性を示している。また、②「東京中心区に投資を集中させている REIT」ほど、東京中心区の物件を売却しにくい（交差項をみると、「東京中心区に投資を集中させている REIT」において、売却直前の物件の 41.6% が東京中心区の物件であるのに対し、運用中の物件の 53.8% が東京中心区の物件である）、すなわち、それ以外の地域の物件を相対的に売却しやすい可能性がある。さらに、③運用中の物件は平均 0.4% というわずかなキャピタルゲインが見込まれるのに対して、売却直前の物件は平均 6.5% のキャピタルロスが見込まれることから、キャピタルロスの大きい物件が売却に至りやすい可能性がある（なお、取得価格は鑑定価格を下回ることが通例のため、取得直後はキャピタルゲインが見込まれる）。

また、売却に至るまでの平均期間は、7.6 半期（3~4 年）程度である。売却物件は、合併吸収により取得された物件である割合が相対的に高い（運用中の物件の平均が 19.1% であるのに対し、売却物件の平均は 27.7% である）。売却される物件は古く（運用中の物件の平均が築 12.6 年であるのに対し、売却物件の平均は築 16.8 年である）、規模の小さい物件である（運用中の物件の平均が $13.4 \times 10^3 m^2$ であるのに対し、売却物件の平均は $8.5 \times 10^3 m^2$ である）。運用中の物件は、東京中心区の構成割合が高く（棟数ベースで 31.1%），オフィス・住宅が中心となる（棟数ベースでそれぞれ 30.7%，50.0%）。

11.3.3 物件売却のモデル

3 つの資産入替の判断指標（ポートフォリオの効率性）に関する仮説を検証するため、物件売却を説明するプロビットモデルを構築する。REIT r が保有する物件 i について、 t 期の状態を考える。

$$DIV_{i,t} = \alpha + \sum_k \beta_k \Psi_{k,i,t} + \sum_l \gamma_l Z_{l,i,t} + REIT_r + T_t + \epsilon_{i,t} \quad (11.1)$$

ここで、 $DIV_{i,t}$ は、物件 i について t 期に売却されたとき 1 を、ポートフォリオ内にとどまる場合に 0 をとる 2 値変数である。 α は定数項、 $\Psi_{k,i,t}$ は効率性指標群（ β_k はそれらの係数）、 $Z_{l,i,t}$ は物件レベ

ルのコントロール変数群 (γ_i はそれらの係数), $REIT_r$ は銘柄ダミー, T_t は半期ダミー, $\epsilon_{i,t}$ は誤差項である。物件・銘柄・時期による基本的な要因をコントロールした上で、効率性指標群の係数に着目する。なお、一般に、連続変数ではなく、「あり・なし」という 2 値データの説明にあたってはロジットモデルやプロビットモデルが用いられるが、ここでは、売却という事象が生じる場合が極端に少ない（表 11.1）ため、プロビットモデルを用いた。

表 11.1 基本統計量

サブサンプル:	運用中	取得直後	売却直前
経済的陳腐化			
運用費率	0.269	0.225	0.300
賃料利回り(%)	2.697	2.748	2.804
地理的集中			
東京中心区 REIT	0.285	0.228	0.299
× 東京中心区	0.538	0.518	0.416
× 東京区部(中心区を除く)	0.183	0.214	0.174
× 東京圏(東京区部を除く)	0.115	0.103	0.143
× その他都市圏	0.135	0.134	0.193
× その他地域	0.028	0.031	0.075
キャピタルゲイン・ロス			
取得時からの資産価格の変化(%)	0.467	4.115	-6.500
パフォーマンス			
CAPEX(百万円)	19.5	21.1	8.6
NCF(百万円)	108.9	105.4	71.3
NOI(百万円)	113.1	107.8	72.9
鑑定価格(百万円)	4,569	4,096	2,762
取得			
REIT設立からの期間(半期)	13.914	7.105	12.885
運用期間(半期)	7.213	0.000	7.641
合併吸収による取得	0.191	0.000	0.277
建物			
築年数(年)	12.60	9.77	16.84
床面積(1,000 m ²)	13.44	12.59	8.54
立地			
東京中心区	0.311	0.273	0.338
東京区部(中心区を除く)	0.238	0.238	0.188
東京圏(東京区部を除く)	0.169	0.185	0.164
その他都市圏	0.212	0.216	0.214
その他地域	0.073	0.088	0.097
種別			
オフィス	0.307	0.249	0.366
商業	0.098	0.096	0.084
住宅	0.500	0.498	0.491
ホテル	0.029	0.046	0.020
物流	0.040	0.061	0.017
その他	0.026	0.050	0.022
サンプル数	46,420	4,908	538

運用中、取得直後(1 半期後)、売却直前(1 半期前)にサンプルを分け、平均値を示す。部分売却や、REIT 間の合併に伴い REIT 間で移転した物件は、売却事例に含めていない。

表 11.2 に、物件売却のプロビットモデルの推計結果を示す。基本的には、表 11.1 の基本統計量から示唆された傾向に沿った結果となっている。まず、①経済的陳腐化を表す運用費率・賃料利回りの係数は、ともに正で有意となっており、経済的陳腐化が進んでいる物件が売却に至りやすい。また、②地理的な集中については、東京中心区に投資を集中させている REIT ほど、東京中心区でない物件を売却しやすい。

具体的には、「東京中心区 REIT」ダミーと物件が位置する地域ダミーの交差項をみると、東京中心区を基準としたとき、その他の地域（東京区部（中心区を除く）、東京圏（東京区部を除く）、その他都市圏）の係数は正で有意となっている。投資方針に合致しない物件を手放す傾向にあるといえる。さらに、③キャピタルゲイン・ロスの程度については、鑑定価格ベースで 15% 以上の大きなキャピタルゲインが見込まれる場合や、鑑定価格ベースで 15% 以上の大きなキャピタルロスが見込まれる場合に係数が正で有意であり、取得時からの価格変動幅が小さい物件に比べ売却に至りやすい。

表 11.2 物件売却のプロビットモデルの推計結果

説明変数	係数	標準誤差
経済的陳腐化		
運用賃率	0.138	0.024 ***
賃料利回り(%)	0.081	0.025 ***
地理的集中		
東京中心区REIT	-0.261	0.123 **
× 東京中心区		(基準)
× 東京区部(中心区を除く)	0.314	0.125 **
× 東京圏(東京区部を除く)	0.302	0.134 **
× その他都市圏	0.288	0.128 **
× その他地域	0.197	0.212
キャピタルゲイン・ロス		
+15%以上	0.228	0.077 ***
[+5%, +15%)	0.152	0.069 **
[-5%, +5%)		(基準)
[-15%, -5%)	0.274	0.068 ***
-15%未満	0.343	0.069 ***
取得		
REIT設立からの期間(半期)	0.460	0.139 ***
運用期間(半期)	0.030	0.008 ***
合併吸収による取得	0.940	0.325 ***
建物		
築年数(s.d.)	0.159	0.021 ***
床面積(s.d.)	-0.149	0.033 ***
立地		
東京中心区		(基準)
東京区部(中心区を除く)	-0.244	0.072 ***
東京圏(東京区部を除く)	-0.248	0.085 ***
その他都市圏	-0.152	0.081 *
その他地域	0.015	0.114
種別		
オフィス	-1.551	0.245 ***
商業	-1.277	0.255 ***
住宅	-0.953	0.232 ***
ホテル	-1.157	0.341 ***
物流	0.322	0.656
その他		(基準)
定数項	-12.495	3.037 ***
REIT固定効果		
時間(半期)固定効果		
サンプル数	37,877	
Pseudo R ²	0.142	

経済的陳腐化、キャピタルゲイン・ロスの各変数については、1 期前の値を使用しタイムラグを持たせている（なお、2 期前、4 期前の値を使用した場合でも、同様の推計結果が得られた）。有意水準は、***10%、**5%、***1% である。

コントロール変数のうち、標準化した築年数（床面積）の係数は正（負）で有意である。これは、古く小規模な物件が売却されやすいことを示しており、たとえ賃料利回りの大きい物件を売却しても、REIT レベルでの投資家への利回りに対する影響を抑える効果がある（Suzuki et al. (2019)[7] では、当該物件を売却しても、REIT の利回りの変化は十分に小さいことを確認している）。物件取得からの期間の係数は正で有意であり、基本的には長期間保有した物件が売却される傾向にある。

また、その他の基本的な傾向として、REIT 設立からの期間は正で有意であり、設立から長期間が経過した REIT ほど、物件売却の傾向があることがうかがえる。また、合併吸収による取得物件は、ポートフォリオの整理に伴い売却されやすい。立地については、東京圏外の物件が売却されやすい傾向にある（ただし、東京中心区の物件は、売却されやすい）。また、オフィスは他の種別に比べ売却されにくい傾向にある。

なお、Suzuki et al. (2019)[7] では、運用費率・賃料利回りについて、売却前数年間・取得後数年間ににおける動態の比較を行っている。売却物件に比べ、運用費率が低く、利回りが低い（安全資産）、すなわち、効率性の高い物件を取得していることを確認している（ただし、賃料利回りについては、取得直後 1 年程度はプレミアムがみられる）。

また、REIT 市場が (i)REIT 設立当初の拡大期、(ii) 金融危機による縮小期、(iii)REIT 合併期、(iv) 近年の拡大期、という 4 つの期間に分かれる（図 11.2）ため、ここで検証した売却に至るメカニズムは期間により異なる可能性がある。そこで、Suzuki et al. (2019)[7] では期間毎のサブサンプル分析を行ったところ、(i)REIT 設立当初の拡大期、(iv) 近年の拡大期において、全サンプルでの分析と共通した売却メカニズムが観察された。これは、拡大期の資産入れ替えが、長期的な視点から非効率な物件を売却し、ポートフォリオの効率性を高めるためのものであることを示している。一方、(ii) 金融危機による縮小期には、既往研究で指摘されていたように、借入資金をやりくりする必要から、効率性の低い物件に限らずキャピタルゲインが狙える物件についても売却が進んでいたことが明らかとなった。また、(iii)REIT 合併期にも効率性の低い物件が売却されるとは限らない結果となり、合併に伴い立地・種別等の観点において、ポートフォリオ内で重複した物件が売却されるといった、他のメカニズムに基づいていた可能性が考えられる。

11.3.4 売却直前のマネジメント戦略

REIT は、実際に物件売却を行う半年前～1 年前から売却の検討を始めるが、その後の準備過程では何らかのマネジメント戦略が存在する可能性がある。一般に、会計基準が認めている範囲内で会計利益の数値を調整する行動は利益調整（Earnings Management）と呼ばれ、2 つの方法がある。1 つは、会計上の処理や見積もりを変更する方法であり、会計的利益調整（Accruals Earnings Management）と呼ばれる。もう 1 つがキャッシュ・フローそのものを変更する方法であり、実体的利益調整（Real Earnings Management）と呼ばれ、研究開発投資の抑制、売却時期の変更等が該当する。REIT 市場でもこうした利益調整がみられ、実体的利益調整については、物件売却の時期の変更等が指摘されてきた（Deng and Ong, 2018[1]）。

REIT の物件レベルで観察される実体的利益調整として、売却直前に、長期的観点からの修繕費である CAPEX を削減するインセンティブが生じると考えられる。売却直前に CAPEX を削減すると、NOI か

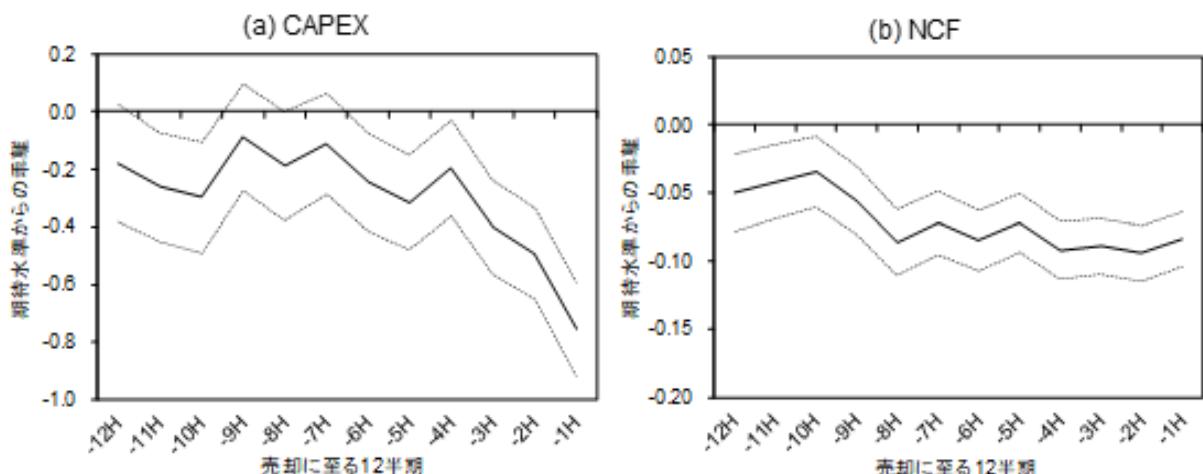
ら CAPEX を差し引いた値である NCF の低下が抑えられるが、これは、NCF は鑑定価格評価に用いられるために買い手にとって魅力的に映ると考えられる。

ここでは、次の推計式により、売却に至る 6 年前（12 半期前）からの CAPEX および NCF の動態を分析する。物件 i について、 t 期の状態を考える。

$$\ln Y_{i,t} = \alpha + \sum_{m=-12}^0 \beta_m^D D_{m,i,t} + \gamma Age_{i,t} + \theta_i + T_t + \epsilon_{i,t} \quad (11.2)$$

ここで、 $\ln Y_{i,t}$ は、CAPEX・NCF の対数値である。 α は定数項、 $D_{m,i,t}$ は売却直前の期間を表すダミー変数（ β_m^D はそれらの係数）、 $Age_{i,t}$ は時間変化のあるコントロール変数（築年数、 γ はその係数）、 θ_i は物件の固定効果、 T_t は半期ダミー、 $\epsilon_{i,t}$ は誤差項である。パネルデータでは、経年に変化しない主体固有の効果を「固定効果」として扱うことで、観察不可能な特性をコントロールし当該主体の時間的な変化に焦点を当てることができる。ここでは、物件の固定効果 θ_i を入れることで、当該物件の築年数・時期による要因をコントロールした上で、売却直前の期間中（売却の 12 半期前～売却期）の変動を捉えることが可能となる。ダミー変数 $D_{m,i,t}$ の係数 β_m^D は、売却直前の期間における、CAPEX・NCF の期待水準からの乖離を捉える。

図 11.3 に、売却直前の CAPEX および NCF の動態として、係数 β_m^D とその標準誤差（±1）を示す。縦軸が 0 の水準は、当該物件固有の効果や築年数を考慮したうえで想定される CAPEX/NCF の水準である。売却直前までは、CAPEX は期待水準で推移しているが、売却時期が近づくと、約 1 年前から CAPEX が大きく減少する（期待水準に比べ数十 % の低下となる）。一方、NCF については、売却に至る 4~5 年前から期待水準より低下しているが、売却直前にはそのままの水準で大きく変動しないことが分かる。これは、CAPEX を売却直前で削減したことで、NOI から CAPEX を差し引いた値である NCF の低下を防ぐ役割を果たしたといえ、物件レベルでの実体的利益調整を示していると考えられる。



係数を示す。点線は、±1 標準誤差を示す。

図 11.3 売却直前の CAPEX, NCF の動態

なお、Suzuki et al. (2019)[7] では、売却直前ににおける NOI・鑑定価格の変化を同様に分析している。CAPEX を売却直前に削減することで、NOI・鑑定価格は緩やかに低下することとなる。

11.3.5 分析結果のまとめ

Suzuki et al. (2019)[7] では、J-REIT 市場における物件レベルのパネルデータを用いて、長期的な観点からの資産入替メカニズムを明らかにした。REIT の基本方針は、効率性の高い物件を長く運用し (going concern)、安定的な賃料収益を得ることである。しかしながら、不動産は、資本的支出や維持管理の不足もあり、経年的に減価していく。そのため、経済的陳腐化が進んだ物件や地理的集中を高める方針から外れた物件を売却する結果、売却物件ではキャピタルロスの程度が大きく観察されることとなる。また、物件売却に至る直前には、当該物件への資本的支出 (CAPEX) を削減することで、鑑定評価に直結する NCF の低下を抑えるマネジメント戦略が明らかとなった。

ここでは、物件レベルの事情により、投資対象となる不動産の入替が生じるメカニズムを検証した。一方、既往研究で指摘されているように、REIT レベルの事情により売却に至ることもある。不動産価格形成の裏側には、こうした REIT の行動原理が関係しているといえよう。今後の課題として、物件レベルと REIT レベルの両方の論理から、資産入替行動を説明する枠組みの構築が求められる。

11.4 REIT 情報を用いた不動産市場分析の方向性

金融市場と資産市場を結びつけるという点では、銘柄レベルのデータから不動産市場の動向を捉える指標を構築する動きがある (清水ほか, 2019[9])。REIT は保有資産の殆どが不動産であることが義務付けられており、レバレッジの補正を行えば、株価の変動から不動産価格の変動を捉えることが可能となる。特に商業不動産等、取引頻度の少ない不動産については価格指数の構築が課題となってきたが、株価情報を用いることにより、高頻度で市場動向を掴むことができるようになる。

こうした研究動向の 1 つに、REIT 市場における物件セクター毎の利回り指標の構築がある (Geltner and Kluger, 1998[2]; Horrigan et al., 2009[3])。REIT 株式の利回りが、地域・種別に応じた物件セクターの利回りから構成されていることを利用する。すなわち、各 REIT の利回りは、各物件セクターの利回りを、当該 REIT のポートフォリオにおいて各セクターの物件が占める構成比で重み付けしたものと考えることができる。株式市場は日時で変動することから、不動産のセクター別の利回り水準やその指標を日時で算出することが可能となる。もう 1 つの研究動向に、REIT 株価を用いた不動産価格指標の構築がある (Shimizu et al., 2015[6])。REIT 株価が、運用中の物件の価格を合計したものとなることを利用する。すなわち、各物件の価格は、当該物件を保有する REIT の株価を、当該 REIT のポートフォリオにおける当該物件が占める構成比で分配したものと考えることができる。その上で地域・種別等の品質調整を行うことで、株式市場は日時で変動することから、不動産価格指標も日時で算出することが可能となる。この 2 つの研究動向は、利回り・価格という異なる指標を対象としアプローチも異なるが、基本的な考え方は共通していることから、統合した枠組みの構築が求められるところである。

日本の REIT 市場では、物件レベルの情報がデータベース化されている点で、海外に比べて充実しているといえるが、学術研究での活用例は限られる。こうした REIT 市場データを活用して、不動産市場のダイナミクスや金融市場との相互関係への理解が深まることが望まれる。

参考文献

- [1] Deng, X., & Ong, S. E. (2018). Real Earnings Management, Liquidity Risk and REITs SEO Dynamics. *Journal of Real Estate Finance and Economics*, 56(3), 410-442.
- [2] Geltner, D., & Kluger, B. (1998). REIT-Based Pure-Play Portfolios: The Case of Property Types. *Real Estate Economics*, 26(4), 581-612.
- [3] Horrigan, H., Case, B., Geltner, D., & Pollakowski, H. (2009). REIT-Based Property Return Indices: A New Way to Track and Trade Commercial Real Estate. *Journal of Portfolio Management*, 35(5), 80-91.
- [4] Li, Q., Ling, D. C., Mori, M., & Ong, S. E. (2018). The Wealth Effects of REIT Property Acquisitions and Dispositions: The Creditors' Perspective. *Journal of Real Estate Finance and Economics*, 1-30.
- [5] Ooi, J. T., Ong, S. E., & Li, L. (2010). An Analysis of the Financing Decisions of REITs: The Role of Market Timing and Target Leverage. *Journal of Real Estate Finance and Economics*, 40(2), 130-160.
- [6] Shimizu, C., Diewert, W. E., Nishimura, K. G., & Watanabe, T. (2015). Estimating Quality Adjusted Commercial Property Price Indexes Using Japanese REIT Data. *Journal of Property Research*, 32(3), 217-239.
- [7] Suzuki, M., Ong, S. E., Asami, Y., & Shimizu, C. (2019). Long-Run Renewal of REIT Property Portfolio Through Strategic Divestment. *Journal of Real Estate Finance and Economics*, forthcoming.
- [8] 児島幸治 (2015)「米国におけるREIT(不動産投資信託)研究の最近の動向」,『国際学研究』, 4(1), 65-81.
- [9] 清水千弘・鈴木雅智・大西順一郎 (2019)「不動産の価格決定構造と情報整備の課題」,『CSIS Discussion Paper (The University of Tokyo)』, 159.