

2023年度 統計データ分析コンペティション

統計数理賞 [高校生の部]

地価に関する最適モデルの構築と手法提案

柏原 昊隼・田原 睦己・大西 裕貴

(雲雀丘学園高等学校)

論文の概要

都市部への人口集中や地方の過疎化の問題に着目し、地価データを用いて因子分析及び重回帰分析を行い、地価の高い地域に共通する因子の特徴を推定することで、地方の人口増加に貢献する要因を見出した。

論文審査会コメント

先行研究などの丁寧な分析から課題を抽出している点が評価される。因子の抽出と交互作用項の重回帰分析への適用など分析技術的には新規性すらあり、レベルがとても高い。ただし、データの吟味や因子分析結果等の表示には、もう少し工夫が必要である。地価についても住宅地や商業地の区分が必要であろう。

地価に関する最適モデルの構築と手法提案

柏原 晃隼 田原 睦己 大西 裕貴
雲雀丘学園高等学校

第1章 序論

1-1 研究目的と研究背景

近年、都市部への人口集中や地方での過疎化が問題となっており、まちおこしや地方公共団体の権限強化などがそれらの対策として実行されている。しかし、状況は依然として好転しない。そこで我々は、地方の魅力を上げるために、その土地の魅力を示す尺度として、地価に関するデータの変化によって地方都市の人口が増加するための要因を見つけることにした。黒木(1999)[1]は、地価を「その土地の立地するポイントの持つ自然的・社会的・経済的及び行政的なポテンシャルを価格という尺度で表現したもの」と論じている。この定義を踏まえると、地価はその土地の魅力を総合的に算出する値と考えることができる。よって、地価を高めることはその土地の魅力を高めることにつながると考えた。

したがって、私たちの最終的な目標は、地価の高い地域に共通する因子の特徴を推定することで、地方の人口増加に貢献する要因を見いだすことである。そのため、本研究の目的は、地価に関するデータをもとに、総務省の統計データを使い、地方の魅力に関するよりよい説明モデルを構築することである。持続的かつ効果的な発展のために、どの年にも比較的共通するような因子の特定や、時代ごとに変化する因子を見いだすことにした。

1-2 問題の検証と先行研究

まず、地方部の過疎化について図1で示した。図1は転入超過数の年ごとの推移を示している。図1からわかるように年々都市部と地方部の転入超過数の差が開いていることから、地方部の過疎化が進行していることがわかる。

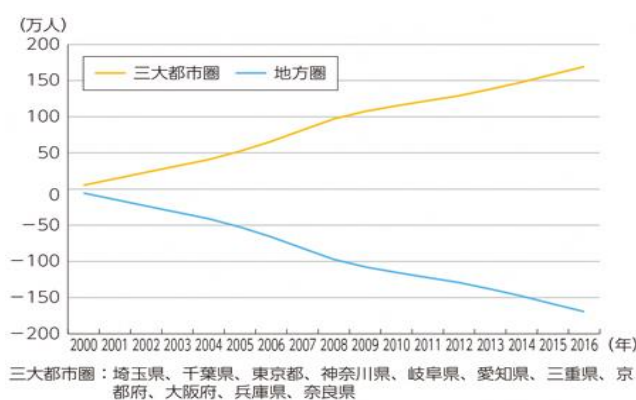


図1 転入超過数の推移

次に、地価によってその土地の魅力を総合的に算出する裏付けとして、黒木の定義と大東建託の地価と住み心地に関する調査[3]では、散布図から両者に強い正の相関（相関係数0.708）が示されている。このことから、本研究では、地価はその土地の魅力を総合的に算出された値であると見なして用いることとした。

地方創生に関する先行研究は数多くあり、政策面における研究が散見される。しかしながら、経済面からのアプローチを行う研究は、政策面の研究に比べ、多く見受けられない。

このことから、本研究では、経済面から土地の価値について考察し、新たな事実を発見することを目指している。土地の魅力という主観的なものに対して、アンケート等の方法を用いず、客観的なデータを用いて統計的なアプローチを用いることにした。

本稿の構成は、第2章は使用するデータの説明、第3章は研究方法、第4章は研究結果、第5章が考察、第6章が結論と今後の課題とした。

第2章 使用するデータ

2-1 用意したデータ

本研究において使用するデータには、欠損値等の理由により後から加えたもの（後述）を除いて、総務省統計局が提供する「統計でみる市区町村のすがた」と総務省が提供する、「地方公共団体の主要財政指標一覧」、各都道府県が調査・公表する「都道府県地価調査」、総務省統計局が提供する「国勢調査」を使用した。菅野・塩濱ら(2021)[4]は研究の際に使用する地価として、他の地価よりも厳密に測定された公示地価を使用している。菅野・塩濱らは公示地価に対して肯定的だが、否定的な意見もある。例えば、井筒・中川・笠間ら(2022)[5]は公示地価は標準点の地価を示しているため、標準値以外の任意の地点の地価を正確に予測することは難しいと述べている。しかし、両者ともに公示地価が厳密に測定されている点は共通しており、本研究では公示地価を使用する。また、東京について、人口などの様々な観点において他の県庁所在地との差が大きいため、本研究では対象から外すことにした。

本研究では、「統計でみる市区町村のすがた」と「地方公共団体の主要財政指標一覧」のデータを使用するため、使用データ一覧を示す(表1-1,表1-2)。ただし、2015年のデータについて、別途示すデータは、SSDSEからのデータを用いた(表1-3)が、同項目の他年度のデータについては、出典元が同じであるため、連続性は保証されていると考えられる。また、総人口と総面積については「統計でみる市区町村のすがた」やSSDSEを用いて集めたが、後述の研究において、直接的には説明変数として用いていない。ラスパイレス指数とは、国家公務員の給与に対する地方公務員の給与の比を表すものである。

表1-1 統計でみる市区町村のすがたのデータ

幼稚園数【園】，小学校数【校】，小学校教員数【人】，中学校数【校】，中学校教員数【人】，高等学校数【校】，出生数【人】，死亡数【人】，婚姻件数【組】，離婚件数【組】，一般診療所数【施設】，病院数【施設】，歯科診療所数【施設】，介護老人福祉施設数(詳細票)【所】，保育所等数(詳細票)【所】，自市区町村で従業・通学している人口【人】，15歳未満人口【人】，15～64歳人口【人】，65歳以上人口【人】，外国人人口【人】，課税対象所得(納税義務者1人当たり)【千円】，流出口(他県で従業・通学している人口)【人】，流入人口(他県に常住している人口)【人】，高齢単身世帯数(65歳以上の者1人)【世帯】，昼間人口【人】，昼夜間人口比率【%】，核家族世帯数【世帯】，単身世帯数【世帯】，高齢夫婦世帯数(高齢夫婦のみ)【世帯】，母子世帯数【世帯】，父子世帯数【世帯】
--

表1-2 地方公共団体の主要財政指標一覧のデータ

財政力指数，経常収支比率，実質公債費比率，ラスパイレス指数

表1-3 2015年においてのみSSDSEから用いたデータ

総人口，15歳未満人口，15～64歳人口，65歳以上人口，外国人人口，単身世帯数，高齢夫婦のみの世帯数，高齢単身世帯数(65歳以上の者1人)
--

2-2 データの修正

2-1のデータのうち、欠損のあるものがあつたため、他の統計を用いて補完することとした。加えて、欠損値を補完する統計を探す際に、2-1よりも詳しいデータを見つけたため、2-1のデータを置換したものについても示す。また、統計年度により書き方が異なるデータもあるが、判別可能と思われるので特記しない。初めに用意したデータのうち欠損を埋める手段がなかったものやほかのデータで十分同じ意味合いを持つと考えられるデータについては2-1には示しておらず、研究にも使用していない。補完するために使用したデータの他項目と「統計でみる市区町村のすがた」のデータが一致していることの確認を行うか、残りの年度分のデータも置き換え先のデータの統計から用いることで、データの連続性を担保している。

表2 他と統計によって置換した項目

他統計によって置換した項目（置換元）	置換先	使用した統計
自市区町村で従業・通学している人口【人】	・自宅で従業している人口【人】 ・自宅外の自市区町村で従業・通学している人口【人】 ・従業も通学もしていない人口	国勢調査
流出人口（他県で従業・通学している人口）【人】	他県で従業・通学【人】	国勢調査
流入人口（他県に常住している人口）【人】	うち他県に常住【人】	国勢調査

表3 他統計により補完した項目

他統計により補完した項目	年度	使用した統計
幼稚園数	平成17年度	学校基本調査
昼夜人口比	平成17年度、平成22年度	国勢調査
流出人口（県内他市区町村で従業・通学している人口）【人】	平成17年度、平成22年度	国勢調査
流入人口（県内他市区町村で従業・通学している人口）【人】	平成17年度、平成22年度	国勢調査

2-3 元データの加工

本研究では、元データの偏りを軽減する試みとして、以下の加工を行った。なお、加工に際してはExcelを使用した。

(1) 人口、面積による調整

使用した元データの多くには、人口や面積といったデータが過大な影響を与えているのではないかという懸念から、人口・面積のデータと元データ¹の相関係数を確認した。その結果を以下に示す。

- ・人口 相関係数の最大値はどの年でも0.99を超えており、最小値もどの年でも0.65を超えていることから、ほとんどのデータと人口との間に相関があることを示唆していると思われる。
- ・総面積 相関係数の最大値がどの年でも0.10程度、最小値がどの年でも-0.30程度という結果から、相関はほとんどないと考えられる。

¹ 2-3でいう「元データ」は、明らかに人口あたりとすることが不自然である「地方公共団体の主要財政指標一覧」から引用したデータや、もともと人口あたりデータとなっている昼夜間人口比率は除いた。

結果から、元データに加えて、脚注1に示した例外以外のデータの人口あたりのデータも使用した。例えば、人口あたりのデータとは、ある市の人口が10万人、中学生数が3千人であったとき、中学生数の3,000(人)を人口の100,000(人)で割った0.03(人/人)を新たなデータとして用いた。なお、データがあまりに小さくなり見づらくなることを防ぐため、本研究では人口1人あたりではなく、人口10万人あたりとなるよう加工した。

(2) 各データの相関係数の確認²

前項において、人口が各データに与える影響が大きいことを懸念して人口あたりのデータを加えたため、本項においては、人口との相関が高い元データの排除を行う。昼夜間人口比率、財政力指数、経常収支比率、実質公債費比率、ラスパイレス指数 については、割合や指数であり、人口あたりデータとするのが不適切である。課税対象所得(納税義務者1人当たり)【千円】については、既に人口あたりデータであることから、人口、面積のどちらの相関分析においても考慮していない。

さらに、同じデータを用いていなければ後に行う年度同士の比較が成立しないため、一年で人口との相関が高かったデータは他の年度においても排除する。具体的には、相関係数が0.97程度と高いものが多いため、人口との相関係数が、全体の第二四分位数を上回っているものを削除した。その結果、排除したデータを表4に示す。また、排除されなかったデータの一覧は省略するが、排除されなかったデータの項目数は、人口あたりのデータ、元データ合わせて55項目である。

表4 相関係数を基準に排除したデータ

幼稚園数【園】，小学校数【校】，小学校教員数【人】，中学校数【校】，中学校教員数【人】 出生数【人】，死亡数【人】，婚姻件数【組】，離婚件数【組】，15歳未満人口【人】，15～64歳人口【人】 65歳以上人口【人】，昼間人口【人】，核家族世帯数【世帯】，高齢夫婦世帯数(高齢夫婦のみ)【世帯】 父子世帯数【世帯】，従業も通学もしていない人口【人】
--

2-4 平均地価の算出

平均地価の算出方法について説明する。平均地価は、都道府県地価調査の各都道府県庁所在地のデータのうち、2005、2010、2015年のそれぞれでの基準地の地価の平均値を用いた。

第3章 研究方法

本研究では重回帰分析を用いてモデルを作成し、地価に影響を与える説明変数を見だし、時代ごとの要因や変遷について考察する。そのためには、どのような変数を説明変数にするのか評価・検証を行いながら、モデルを構築することにした。次の研究方法に従って進めつつ、適宜、評価・検証を行って研究を進めていく。

方法1 因子分析を用いて因子を得る。

方法2 因子を加工せずに重回帰分析を用いてモデルを構築する。

方法3 因子を加工して重回帰分析を用いてモデルを構築する。

方法4 2つのモデルを評価し、比較する。

方法2と方法3で2つのモデルを構築するが、この違いについて説明する。方法2は、方法1で得られた因子を加工せずに説明変数として用いて重回帰分析を行い、モデル構築した(以下、一次モデル)。方法3は、方法1で得られた因子を加工して説明変数として用いて重回帰分析を行いモデル構築した(以下、二次モデル)。方法4では、この2つのモデルを比較した。

² 因子分析の際に、人口あたりのデータと元データが異なる因子として分類されることもあることから、本研究においては人口あたりのデータと元データの相関係数については考慮しない。

モデル構築と評価のデータは、46道府県のうち、40道府県を学習データ、6道府県をテストデータとした。具体的には、各道府県庁所在地に乱数を振り、乱数値の上位40道府県を学習データを用いて一次モデル、二次モデルを構築し、残りの6都市をテストデータとして評価に用いた。テストデータは和歌山市、盛岡市、長野市、山口市、佐賀市、金沢市である。因子分析はExploratory、重回帰分析はHAD[6]を使用し、その他はExcelを使用した。

第4章 研究結果

最適なモデルを構築するために、研究方法を繰り返しながら実施した。

4.1 因子に対する評価・検証

方法1 直交回転による因子分析

因子分析について、固有値が1.00以上で累積寄与率が75%以上となった因子を抽出し、バリマックス法により軸の直交回転を行った。因子分析の結果、6つの因子に統合された。

方法2 重回帰分析を用いたモデル構築

方法1の因子分析により得た6つの因子を説明変数、道府県庁所在地の平均地価を目的変数として重回帰分析を行った。その際、p値によって説明変数を選別した。選別の閾値は、一般的に用いられている0.05とした。また、選別はP値が最も高い説明変数一つ削るごとに、再度重回帰分析を行い、p値を確認する流れをすべての説明変数のp値が0.05を下回るまで行うことにした。その結果が表5である。

表5-1 (2005年)			表5-2 (2010年)			表5-3 (2015年)		
変数名	係数	p値	変数名	係数	p値	変数名	係数	p値
切片	134353.457	.000	切片	126982.457	.000	切片	123503.174	.000
Factor_1	3528.503	.000	Factor_1	4143.421	.000	Factor_1	4911.138	.000

表5より、各年の地価に影響を与える変数がどれもFactor1の1つだけに定まった。この結果は、年度ごとの共通点を見つけることには適しているが、年度ごとの違いや変遷を見いだすためには不本意な結果である。そのため、方法1の因子分析で直交回転を用いたが、因子間相関が0.80等、一般的に強い相関があると判断できる変数の関係が見つかった。このことから、次は斜交回転を用いて因子分析を行うこととした。

4.2 改善した因子分析によるモデル構築

方法1 斜交回転を用いた因子分析と名付け

因子分析は、固有値が1.00以上で、累積寄与率が75%以上となった因子までを抽出し、プロマックス法によって各因子を軸の斜交回転を行った。因子分析の結果をもとに、それぞれの因子に名づけを行った。人口因子をF1、非労働人口因子をF2、離婚因子をF3、県外人口流入因子をF4、県内人口流入因子をF5、若年人口因子をF6とする。

表6 因子の名付け

因子	Factor_1	Factor_2	Factor_3	Factor_4	Factor_5	Factor_6
名付け	人口因子	非労働人口因子	離婚因子	県外人口流入因子	県内人口流入因子	若年人口因子

方法2 重回帰分析を用いたモデル構築（一次モデル）

学習用データである40道府県における因子分析で得られた因子を説明変数、道府県庁所在地の平均地価を目的変数として各年において重回帰分析を行った。その際、4.2方法2と同様の手順によって、すべての説明変数のp値が0.05を下回るまで行った。モデルは以下に示す。

- 2005年の最適モデル $\square = 3285 \times (\square 1) - 1257 \times (\square 2) + 134765$
- 2010年の最適モデル $\square = 4343 \times (\square 1) + 2807 \times (\square 5) + 127881$
- 2015年の最適モデル $\square = 5219 \times (\square 1) + 2459 \times (\square 5) + 4308 \times (\square 6) + 123656$

方法3 因子を加工して重回帰分析を用いたモデル構築（二次モデル）

方法2で構築したモデルの回帰係数より人口因子が最も地価に影響を与えていることは明らかであるため、人口因子との影響が強い説明変数を作成することにした。本研究では、人口因子（F1）とその他の5つの因子（F2～F6）の積を新たな因子として加工した。つまり、F1×F2、F1×F3等の5つの変数を作成した。40道府県のF1～F6と加工した5つの変数を説明変数として重回帰分析を用いてモデルを構築した。モデルは以下に示す。

- 2005年の最適モデル $\square = 4802 \times (\square 1) + 194 \times (\square 1 \times \square 2) - 196 \times (\square 1 \times \square 3) + 148452$
- 2010年の最適モデル $\square = 5849 \times (\square 1) + 3023 \times (\square 6) + 154 \times (\square 1 \times \square 2) + 141341$
- 2015年の最適モデル $\square = 5772 \times (\square 1) + 2885 \times (\square 5) + 2127 \times (\square 6) + 106 \times (\square 1 \times \square 2) + 201 \times (\square 1 \times \square 6) + 132374$

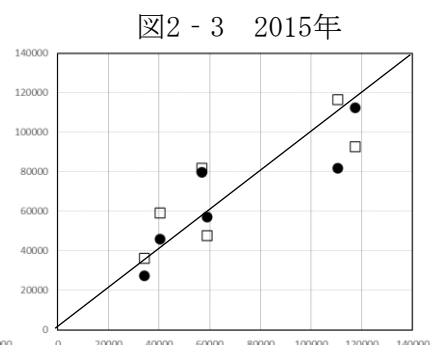
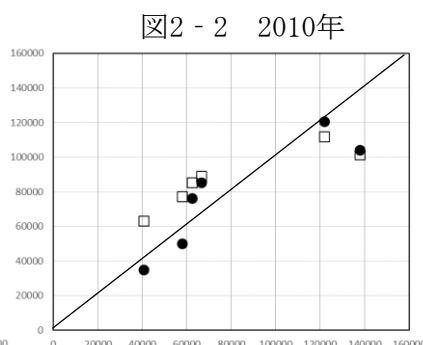
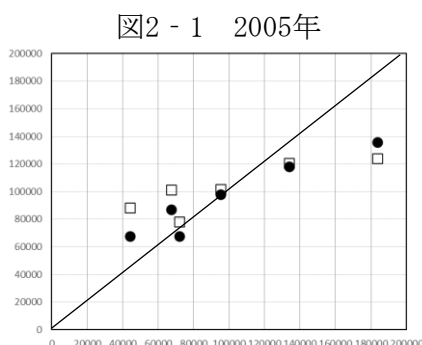
方法4 一次モデルと二次モデルの比較による評価

2005年、2010年、2015年の一次モデル、二次モデルにテストデータの和歌山市、盛岡市、長野市、山口市、佐賀市、金沢市のデータを代入し、予測値を求める。実際の地価と予測値の相関係数を用いてモデルの評価と比較を行い、その結果を表7に示す。また、実際の地価と予測値の散布図を図2に示す。実際の地価は公表されているその都市の地価の平均値を用いた。

表7、図2より、2010年、2015年は一次モデルと二次モデルの差はそれほど大きい差ではないが、2005年は二次モデルの方が最適であると判断できる。

表7 地価と予測値の相関係数

相関係数	2005	2010	2015
一次モデル	0.86	0.90	0.85
二次モデル	0.96	0.88	0.88



□一次モデル・●二次モデル、横軸が実際の地価、縦軸が予測値、単位は円

4.3 モデルによるシミュレーション

二次モデルの説明変数から2つの因子を選択し、それらに数値を代入してシミュレーションを行った。掛け合わせた説明変数ではなく、掛け合わせる前の因子の値を変動させた。選択しない因子は定数として扱い、結果が実際の状況に近づくよう、分散の大きいものを変数として選択した。モデルで定数とする因子の数値は、その年の因子得点の中央値とした。各年のモデルにおける因子は2005年が人口因子と非労働因子、2010年が人口因子と県内人口流入因子、2015年が人口因子と非労働因子である。結果が図3である。高さが地価、平面軸が選択した二つの因子である。

図3-1 2005年

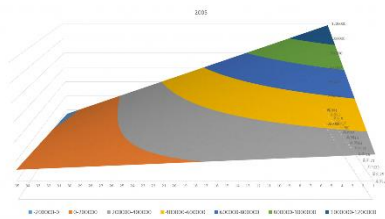


図3-2 2010年

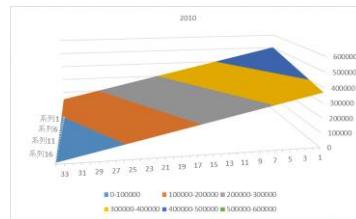
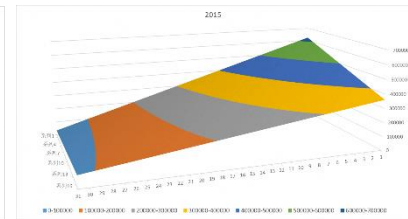


図3-3 2015年



第5章 考察

本研究の目的は、地価に関するデータをもとに、総務省の統計データを使い、地方の魅力に関するよりよい説明モデルを構築することである。そのために、使用するデータ、因子分析の活用、重回帰分析によるモデル構築を行った。その過程において、地価に影響を与えるよりよい説明変数を見いだすために、一次モデルと二次モデルを構築した。結果、2005年のみであるが、一次モデルに比べ、二次モデルの方が最適なモデルであることを示した。このことを踏まえ、各変数から地価に影響を与える要素について考察する。

第一に、二次モデルによって最適なモデルが構築できたことから、単に線形的に因子を用いた説明変数よりも、影響力の大きい因子と掛け合わせて変数を加工することで、モデルに影響を及ぼす説明変数が現れる。つまり、人口因子によって見えにくい要因が、人口因子と掛け合わせた変数とすることで目的変数に影響を与える要因分析につながる要素を見いだせる可能性が示唆できた。

第二に、2005年のモデルに離婚因子（回帰係数は負）が説明変数となったことは特徴的である。これは、共働き世帯数の増加が関連すると考えられる。共働きでない夫婦が離婚すると、専業主婦であった夫もしくは妻が地元に戻るというケースが一般的であり、離婚するとその都市での土地の需要が低下していた。しかし、共働きの夫婦は離婚しても職場は変わらず、別居してもその都市に住み続けることが多くなったからではないかと考えた。

第三に、2015年のモデルでは若年人口因子が地価に影響を与える説明変数となったことがその他の年との特徴である。これは出生率の低下が要因であると考えられる。2009年以降死亡数が出生数を上回り、人口増減率が負であることから、若年者がどこにでもいるという状況ではなくなり、若年者が多い地域は魅力ある地域であると判断できる可能性がある。ゆえに、若い人の希少性が高まり、若年人口因子が説明変数に加わったと考えた。

第6章 結論と今後の課題

本研究では、地価に関するデータをもとに、総務省の統計データを使い、地方の魅力を地価と代替し、地価に関する説明モデルが構築できた。単に重回帰分析を用いるのではなく、説明変数となり得る値を検討し、評価・検証を行いながら構築した。結果、説明変数において、因子分析の結果をそのまま変数とする場合と、加工して変数を作成する場合を比較することで、加工した変数を用いた方が最適なモデルをつくることができる可能性を示唆できた。

加えて、因子分析の結果を用いて重回帰分析を行う場合の変数の設定方法についての手法を示すことができた。さらに、本研究の変数の加工によって、影響の大きい変数によって見えなかった要因分析につながる要素が見いだせる可能性を見いだしたことも1つの成果である。

因子分析と重回帰分析を活用した研究を行うことで、年ごとの地価に着目し、持続的に地価、すなわち、その土地の魅力の向上に資する要因の探索を実施した。結果、人口因子は先行研究からも明らかであったが、2005年の離婚因子、2015年の若年人口因子が地価に影響を与えることについて、その時代背景に応じた要因を見いだす要素と判断できる可能性がある。

しかしながら、それらと地価との因果関係についてはより詳細な研究が必要となることには注意が必要である。使用したデータが限定的で、都市の規模についても小規模なものは検証できていないため、そのような点で課題が残る。

今後の展望として、2008年のリーマンショックや2011年の東日本大震災など、経済的に多くの影響を及ぼしうる年の分析と関連させる。これらの年は、例年と比べ外れ値となりうると考えられるが、特殊な年の検証は、普遍的に地価に影響を及ぼす要因と対比する研究は有用であると考えている。また、対象とする年の範囲の長さについても改善の余地があり、高度経済成長期から今に至るまでの地価の決定要因を考察できれば、大局的に地価の昂騰、急落、維持を観測できるため理想的である。

参考文献

- [1] 黒木勇人(1999). GISを用いた大分市における地価形成要因分析, 大分大学卒業論文.
- [2] 総務省(2017). 平成29年版情報通信白書人口減少社会の課題と将来推計.
- [3] 大東建託(2022). 街の住み心地の良さと公示地価に相関はある?. <https://dime.jp/genre/1590527/> (2023年9月4日参照)
- [4] 菅野雄太, 塩濱敬之 (2021). 地理的加重回帰分析モデルによる東京都公示地価の分析 その推定値の経年変化について, 統計研究彙報, Vol178, pp55-74.
- [5] 井筒雄介, 中川歩, 笠間俊夫 (2022). 機械学習を用いた土地価格の予測, 人工知能学会第二種研究資料, 金融情報学研究会, SIG-FIN-029-11.
- [6] 清水裕士 (2016). フリーの統計分析ソフトHAD: 機能の紹介と統計学習・教育, 研究実践における利用方法の提案 メディア・情報・コミュニケーション研究, 1, 59-73.