近畿各府県における フラット35利用者の特性

住宅金融支援機構 調査部 専任研究員 外山（とやま） 信夫（のぶお）

1982 年早稲田大学政治経済学部卒業。住宅金融公庫入庫。日本経済研究センター経済分析部等を経て、2014 年４月より現職。日本統計学会会員
著書等「The R Book ―データ解析環境 R の活用事例集―」（共著、九天社、2004 年）、「R による GAM 入門」（共著、行動計量学第 34 巻1号、2007 年）、「R と SVM」（共著、大阪電気通信大学情報科学センター、2008 年）等

❶ はじめに

住宅金融支援機構では、年２回（年度上半期、年度全体）フラット 35 利用者のデータについて集計・分析を行い、「フラット 35 利用者調査」として公表している。

同調査の利用者データには、利用者の地域に関するデータも含まれており、それを使用することにより地域分析が可能となる。ここでは、地域分析のケース・スタディの一例として、近畿各府県におけるフラット 35 利用者の特性について、2013 年度全体に係る調査に基づいてその概要を分析した例を紹介する。

分析に先立ち、全国に占める近畿圏のフラット 35 利用者の概数について概観する。

近畿圏のフラット 35 利用者数は、合計で 10,174 件、全国のフラット 35 利用者数に占めるシェアは 16.3%となっている【図表１】。住宅の種類別に見ると、近畿圏のシェアはマンションと中古戸建で 20%を超え、比較的高くなっている。これに対し、注文住宅における近畿圏のシェアは 10%強と比較的少ない。

❷ 主要指標の１次元的な分布について

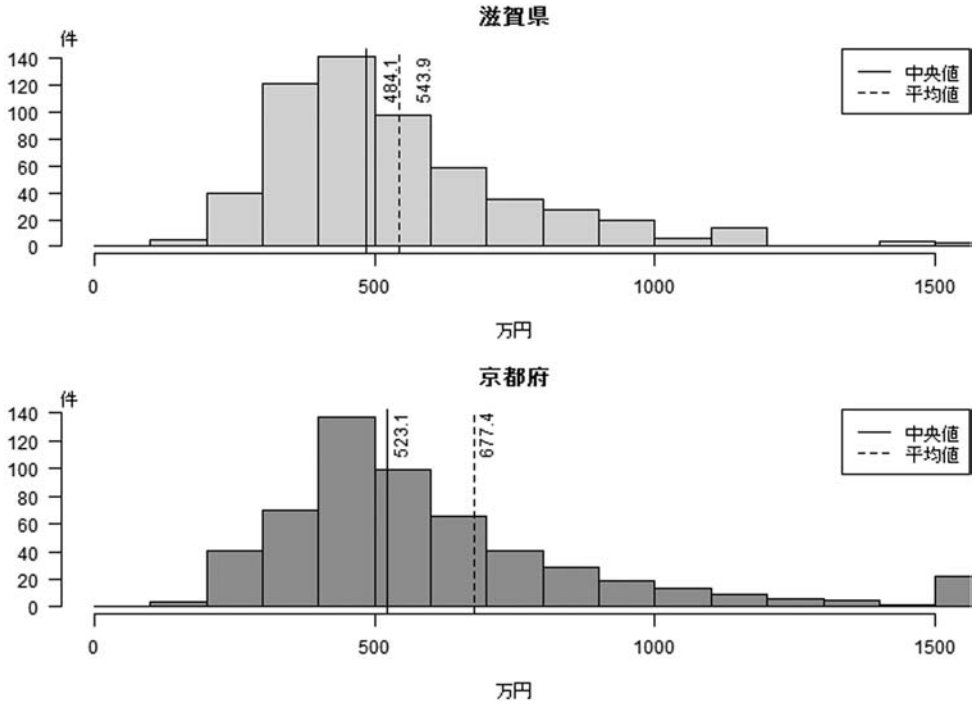
フラット 35 利用者の特性を表す主要指標の代表として、近畿圏の各府県における利用者の世帯年収を見てみる。【図表２】は、滋賀県と京都府における利用者の世帯

図表１　近畿圏のフラット35利用者の全国に占めるシェア

	全国	近畿圏	同シェア
注文住宅	14,507	1,629	11.2%
土地付注文住宅	18,887	2,912	15.4%
建売住宅	9,965	1,562	15.7%
マンション	10,326	2,346	22.7%
中古戸建	3,605	781	21.7%
中古マンション	5,065	944	18.6%
合計	62,355	10,174	16.3%

（資料）住宅金融支援機構「2013 年度フラット 35 利用者調査報告」（以下、同じ）

図表２　滋賀県と京都府の世帯年収の分布



図表３　京都府における世帯年収ベスト10

20,373.9	4,550.0	3,763.6	3,694.0	3,555.8
3,322.0	3,150.0	2,872.9	2,700.0	2,340.0

年収の分布とその平均値及び中央値（小さい順から並べた場合、その中央に位置する値）を示したものである。どちらにおいても、分布は左右対称ではなく、右裾が長い形状となっている。すなわち、高い世帯年収の分布が比較的厚く、長い。この結果、滋賀県においてはそれほど顕著ではないが、京都府においては平均値が677.4万円と中央値の523.1万円を大幅に上回っている。

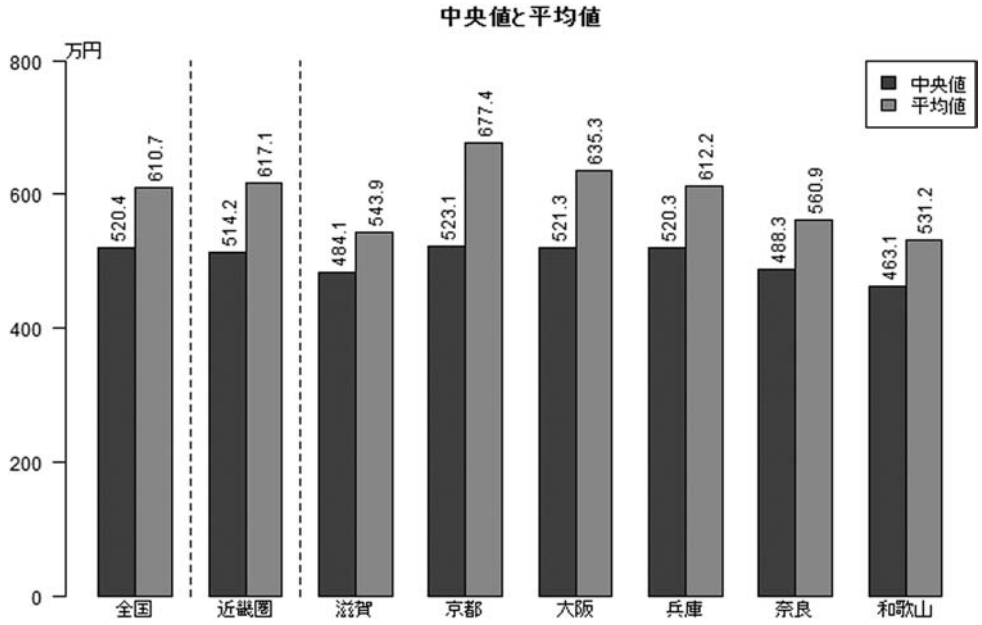
京都府の世帯年収の平均値は、どうしてこんなに高いのだろうか。その理由は、京都府における利用者の世帯年収の上位10件を見ると明らかになる【図表３】。

世帯年収のデータには、世帯年収が２億374万円と他からかけ離れた、明らかな外れ値が存在している。この１件を除いて世帯年収の平均値を再計算すると、677.4万円から642.4万円へ低下する。京都府の利用者数564件の平均値が、わずか１件の外れ値を除くだけで約35万円も低下していることになる。さらに、やや恣意的であるが上位10件を除いた場合の平均値は、598.8万円と600万円を下回ることになる。

このように平均値の値が不安定なのは、平均値という指標のブレークダウン・ポイント（破綻点）が極めて小さいからである。ブレークダウン・ポイントとは、観測値のうちどれだけを $\pm\infty$ の彼方へと近づけていくと、その指標は影響を受けるかを表す概念である。平均値は、ただ１件の観測値が $\pm\infty$ へ近付くだけでブレークダウンしてしまう。平均値のブレークダウン・ポイントは０である。これに対し、中央値は観測値の約50％弱が $\pm\infty$ へ発散しても、その値は不変である。中央値のブレークダウン・ポイントは0.5である。このように、平均値は極めて脆弱な指標であるのに対し、中央値はロバスト（頑健）な指標である。ある変数を代表する指標としては、中央値を採用することが望ましい。

【図表４】は、世帯年収の平均値と中央値について、近畿各府県の比較を行ったものだが、これも上述した点を踏まえて見られるべきものである。すなわち、平均値に

図表４　世帯年収の中央値と平均値



着目すると、各府県間で大きな差があるように見られ、京都府は大阪府より高く、大阪府は兵庫県より高いように見える。しかし、中央値で比較すると、この３府県では大きな差はないことがわかる。

❸ 主要指標の２次元的な分布について

１次元的な分布では、主要指標相互の関係がわからない。そこで、次に２つの指標について、散布図を作成することにより指標間の関連性を調べてみる。また、作成された散布図から近畿各府県の全国における位置を把握することとする。

その一例として、マンション比率（各都道府県の利用件数全体に占めるマンションの比率）と各都道府県の住宅面積の中央値について作成された散布図が【図表５】である。なお、マンション比率については、都道府県間の格差が大きいため常用対数により対数変換している。

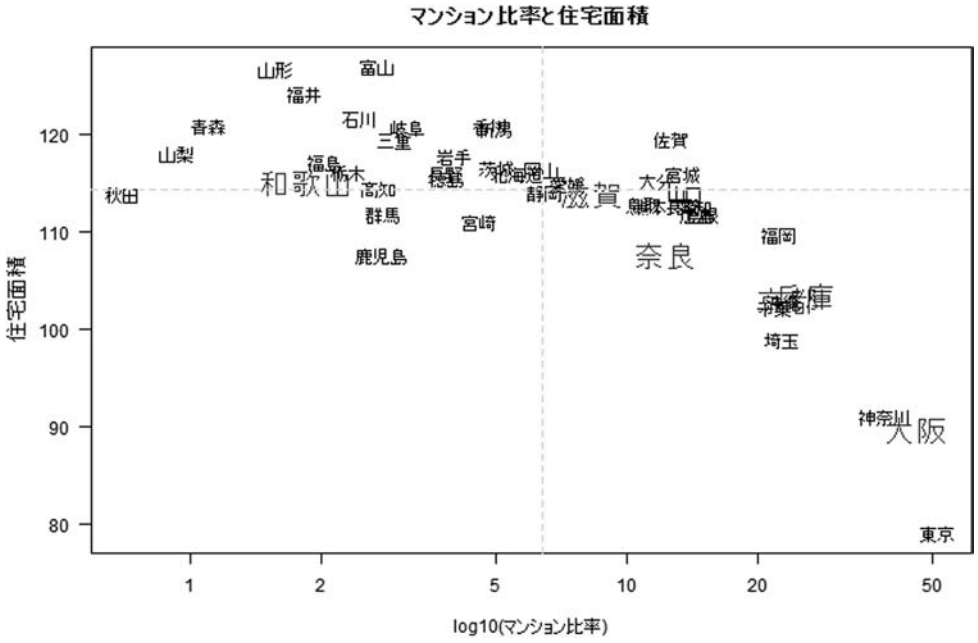
予想されるとおり、マンション比率が高い都道府県ほど、住宅面積の中央値が小さい（これを仮に都市化の程度とする。）。両者の間には、緩やかな負の相関が存在する。都市化が最も進んでいるのは東京都で、次に大阪府が神奈川県とほぼ同じ位置で並んでいる。次いで、京都府と兵庫県が千葉県とほぼ同じ位置で並んでいる。一方、奈良県、滋賀県、和歌山県と移るにつれ、都市化の程度が低下している。

このような２次元的な分布を示す散布図は、指標の対ごとに作成することができるが、指標の数が増えるにつれ、組合せの数が大幅に増加してしまう。例えば、代表的な指標として 10 の指標を採用すると、組合せの数は (10×9)/2＝45 と、45 通りの散布図を描かなければならなくなる。それらを個別に評価することは、著しく困難となっていく。

さらに、指標の対ごとにおける観測値間（ここでは都道府県間）の距離はわかっても、それらを統合した結果は必ずしも明らかでない。

このため、多次元空間における何らかの統合距離を算出し、その統合距離に基づいて各観測値をグルーピングする手法が考えられる。グルーピングする手法としては、自己組織化マップなど様々な手法が考えられるが、ここでは代表的な手法としてクラスター分析を採用する。

図表５　マンション比率と住宅面積



❹ 多次元データの要約（１）
―クラスター分析―

統合距離としては、ここでは次式で表されるユークリッド距離を採用する（これ以外によく使用される距離としては、マンハッタン距離などがある。）。

$$dist_{ij}=\sqrt{\sum_{k=1}^{p}(x_{ki}-x_{kj})^2}$$

ここで、$dist_{ij}$ は都道府県 i と都道府県 j の距離、$k=1,\ldots,p$ は使用される指標 x のインデックスである。すべての都道府県間の距離を計算するためには $(47\times46)/2=1081$ 通りの距離を計算することとなる。

都道府県の特性を表す指標として、ここでは当初金利、申込人年齢、住宅面積、返済負担率、log 戸数、log 世帯年収、log マンション比率、log 所要資金合計、log フラット 35 借入額及び log 現住宅面積の 10 の指標（$p=10$）を採用する。このうち log 戸数と log マンション比率以外は、各都道府県における当該指標の中央値を使用する。

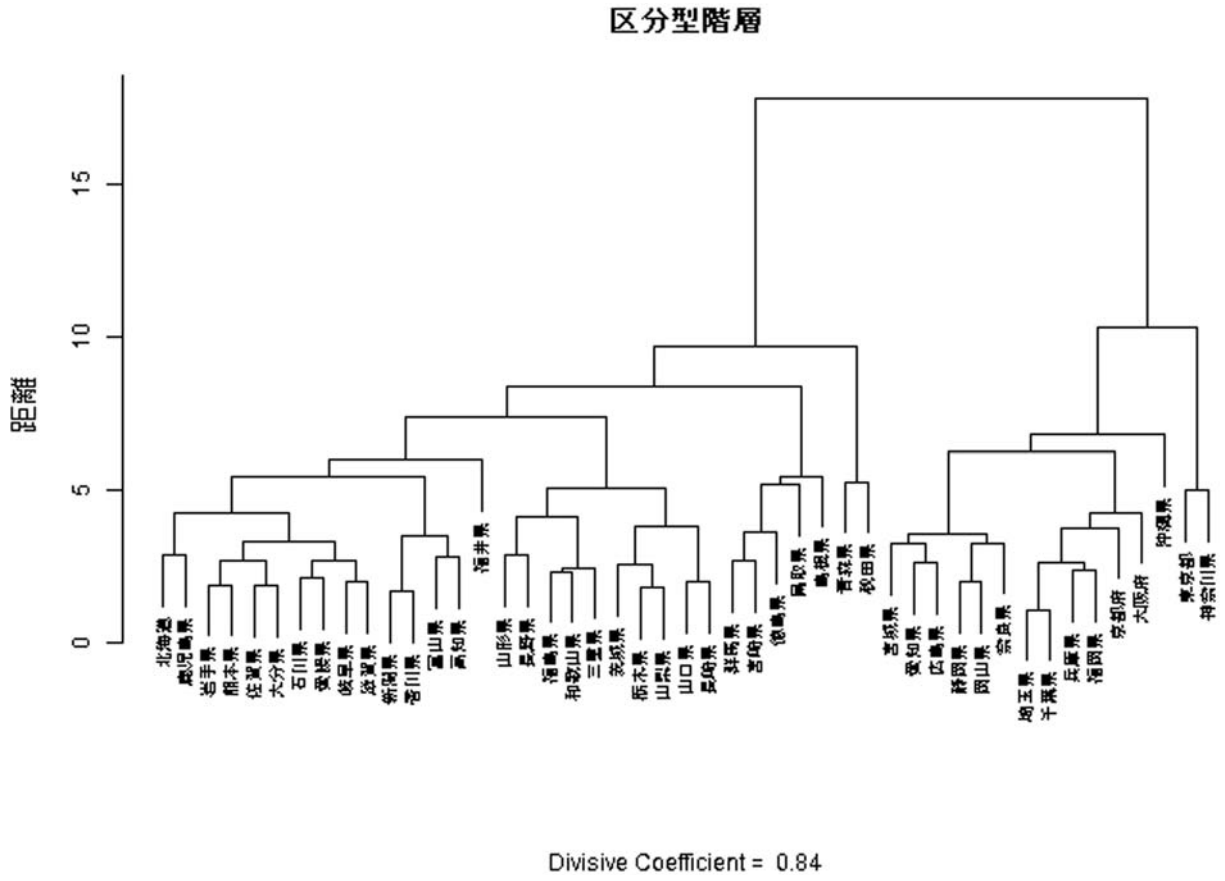
また、統合距離の計算に際しては、各指標の測定単位が異なるため、あらかじめ標準化を行っている（標準化に当たっては、少しでもロバストな値となるよう標準偏差ではなく、平均絶対偏差を使用している。）。

このようにして算出した距離に基づいて、都道府県をグルーピングし、全体を１つの階層として構成するわけであるが、階層の構成法としては、大別して凝集型階層による方法と区分型階層による方法とがある。凝集型階層による方法では、個々の隣接する都道府県に着目し、最も距離が近いものから順次結合していく。いわば下（各論）から上（総論）へ積み上げていく方法である。これに対し、区分型階層による方法では、全体のバラツキを俯瞰し、最も分離の良いグループから段階的に区分していく。いわば上（総論）から下（各論）へ区分していく方法である。

ここでは、紙幅の関係上、区分型階層による方法の結果のみを【図表６】として示す。区分型階層の方が分離が良く、全体の特徴をより良く表現しているからである。結果を概説すると、次のとおりである。

主として政令指定都市がある都道府県が右側のグループとしてまとめられる。兵庫県は、千葉県、埼玉県、福岡県といった県に近い。京都府と大阪府は、徐々に東京都、神奈川県に近付いていく。奈良県は、少しだけこれらの都府県から離れている。和歌山県は、三重県と同じグループを構成している。滋賀県は、特性が他の近畿圏の府県とは大きく異なる。

図表６　区分型階層による都道府県のクラスター分析



Divisive Coefficient = 0.84

このようにクラスター分析では、グルーピング結果は理解しやすい。しかし、グループ化に当たって、どの指標がどの程度、あるいはどの方向で影響しているかがわからないという欠点がある。

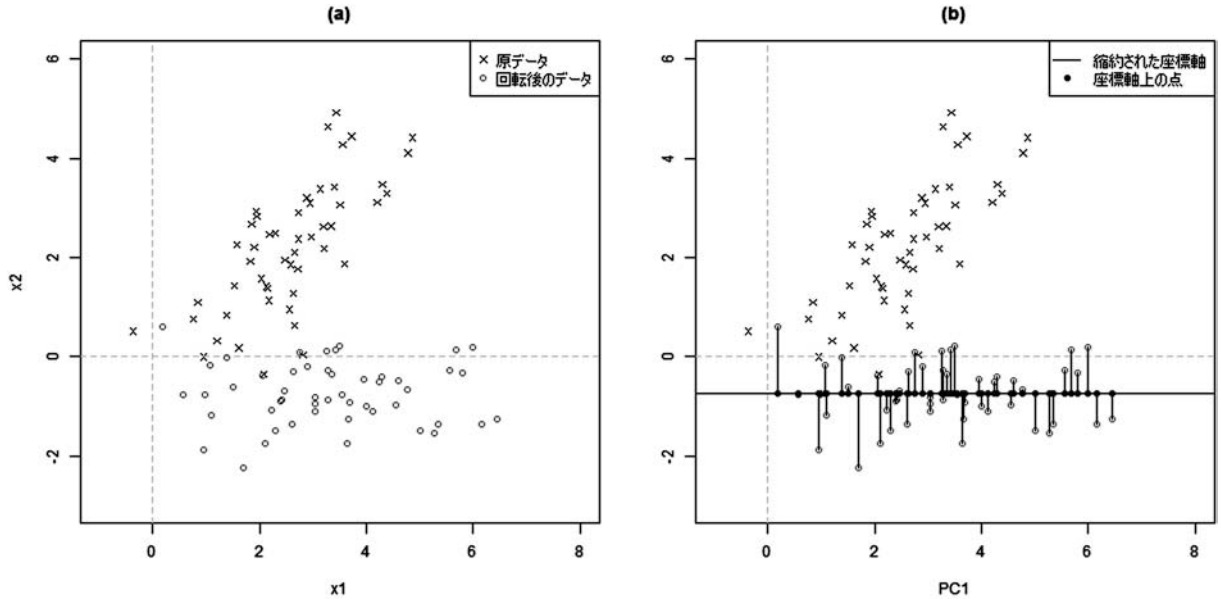
❺ 多次元データの要約（２）―主成分分析と因子分析―

この欠点を克服するため、主成分分析という手法を用いる。主成分分析は、できるだけ少数の合成指標で各都道府県の位置を特定しようとするものである。合成指標は、各指標にプラス又はマイナスの方向の一定の重みを乗じることにより、すべての指標の重み付き合計得点として作成される。その指標ごとの重みで、各指標の影響度を評価することができる。

「できるだけ少数の」ということを、簡単な概念図で説明する。【図表７】(a) の×印で表された各点は、点線で描かれた x_1 と x_2 の２つの座標軸からなる２次平面におけるデータ点である。目的は、情報の損失をできるだけ少なくしつつ、データに適切な回転を加えて１次元直線上の値として表現することにある。○印で表された各点は、回転後の点である。【図表７】(b) の実線で描かれた水平な軸 PC1 は回転後の１次元の軸である。情報の損失をできるだけ少なくするには、○印の各点からこの軸に下ろされた垂線の距離を最小にすることである。別の手法として、垂線と PC1 軸との交点（これが回転後の軸における各データ点となる。）の分散を最大化する方法がある。実は、分散最大化は情報損失最小化と等価である。

都道府県別データは 10 変数あるので、10 次元空間内のデータ点をより少数の座標軸で表現することが、ここでの目的である。主成分分析により、【図表８】のような結果が得られる。同表は、求められた座標軸（それによって説明される分散の、データ全体の分散に占める割合が大きい順に並べられ、それぞれ第１主成分、第２主成

図表７　回転による次元圧縮の概念図



図表８　都道府県別データの主成分分析の結果

	主成分１	主成分２	主成分３	主成分４	主成分５
分散の割合	0.4847	0.1578	0.1106	0.0915	0.0603
同累計	0.4847	0.6425	0.7531	0.8446	0.9049
	主成分６	主成分７	主成分８	主成分９	主成分 10
分散の割合	0.0431	0.0274	0.0173	0.0046	0.0026
同累計	0.9480	0.9755	0.9928	0.9974	1.0000

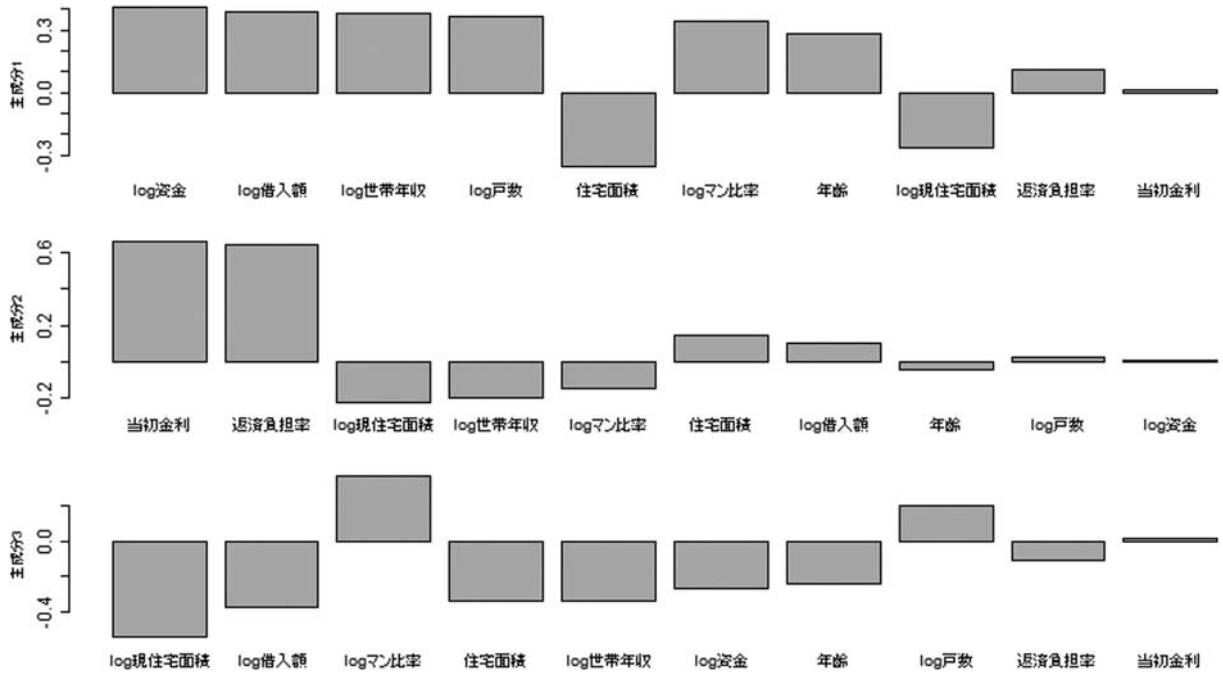
分等と呼ばれる。）によって説明される分散の割合とその累計値を示したものである。第１主成分から第３主成分までの３つの主成分だけで 10 変数全体の分散の 75.3%が説明できることになる。

主成分は、各変数の重み付き合計によって構成される。その重みを、主成分負荷量という。主成分負荷量を第３主成分まで示したものが【図表９】である。棒グラフの棒が、重みの大きさとその正負を示している。各都道府県の各主成分上の位置は、その都道府県の各変数にこの重みを乗じて、合計することにより得られる合成得点によって計算される。これを、主成分得点という。主成分分析は、各都道府県の主成分得点を主要な主成分ごとに吟味して、評価することが分析の目的となる。しかし、その評価の基準となる各主成分のそれぞれの重みを総合した場合の、その意義付け・解釈が、必ずしも明らかでないときがある。そのようなとき、解釈が分析者の主観に依存してしまうことになる。

これに対し、観測値によって表されている結果の元となっている要因は、本来、比較的単純な構造をしているはずであり、その単純構造を求めるように回転の方向を定めるという手法がある。そのような発想から発生した分析手法が、因子分析である。因子分析では、説明がしやすいように回転が行われることから、主成分分析のように解釈の問題が比較的生じにくい。回転方法としては、各因子の変数ごとの重み（因子負荷量という。）が特定の変数に集中するように回転させる、バリマックス回転がよく使用される。

都道府県別データにバリマックス回転による因子分析を行った結果が、【図表 10】である。第４因子までで、データ分散の 73.6%が説明されている。また、因子ごと

図表９　主成分分析による主成分負荷量



図表 10　バリマックス回転による因子分析の結果

	因子１	因子２	因子３	因子４	因子５
分散の割合	0.2865	0.2224	0.1378	0.0897	0.0743
同累計	0.2865	0.5089	0.6467	0.7364	0.8106

の因子負荷量を示したものが【図表 11】である。

【図表 11】における各因子の解釈は、比較的容易である。第１因子では、log 所要資金合計、log フラット 35 借入額及び log 世帯年収が大きな正の値をとっており、富裕度を表す指標であると解釈できる。第２因子では、住宅面積が大きな負の値となる一方で、log マンション比率及び log 戸数が大きな正の値をとっており、都市性を表す指標であると解釈できる。第３因子は返済負担率、第４因子は log 現住宅面積を表す指標であろう。

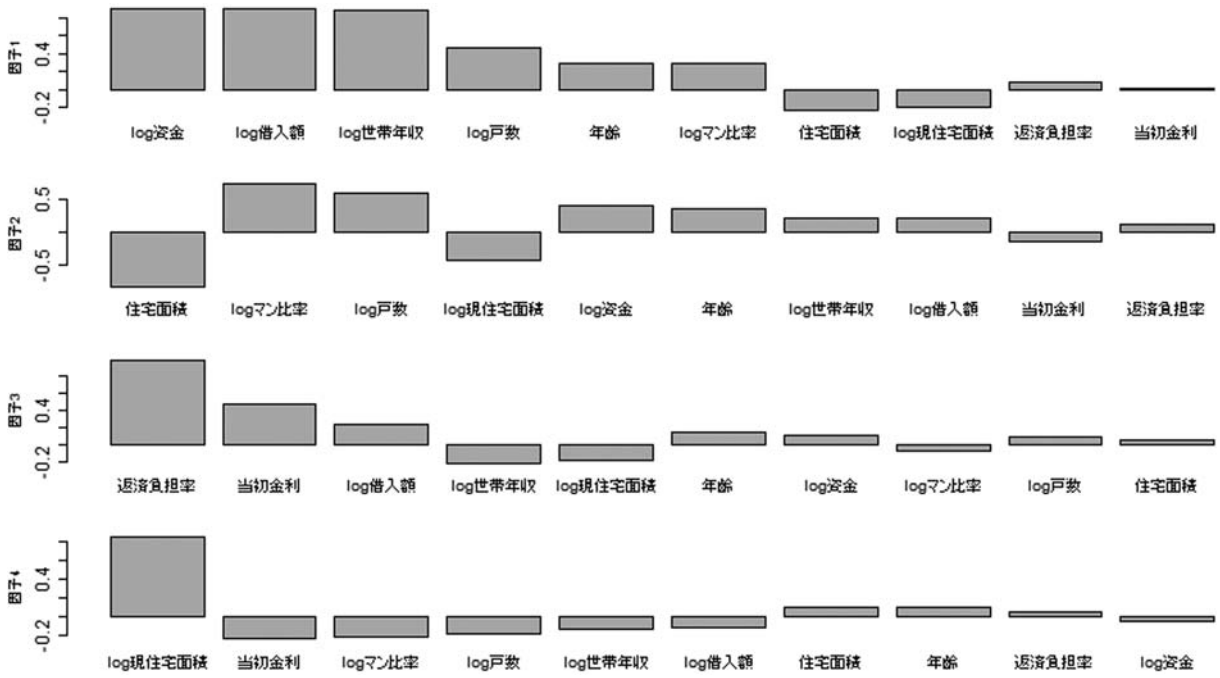
【図表 12】は、第１因子と第２因子を軸とした場合、各因子負荷量が表すベクトルの方向と都道府県別の因子得点を併せて表示した図で、バイプロットと呼ばれるものである。第１因子である富裕度においては、東京都が他を引き離し極めて大きな値となっているのに対し、近畿圏の各府県においては富裕度はそれほど高くない（文字が重なって見えにくいが、和歌山県はむしろ若干低いようである。）。第２因子である都市性においては、大阪府はわずかではあるが東京都を上回っている。また、兵庫県、奈良県及び京都府の都市性は、神奈川県及び埼玉県を下回り、千葉県と同程度であることがわかる。

（第３因子と第４因子のバイプロットについては、紙幅の関係上、割愛させていただいた。）

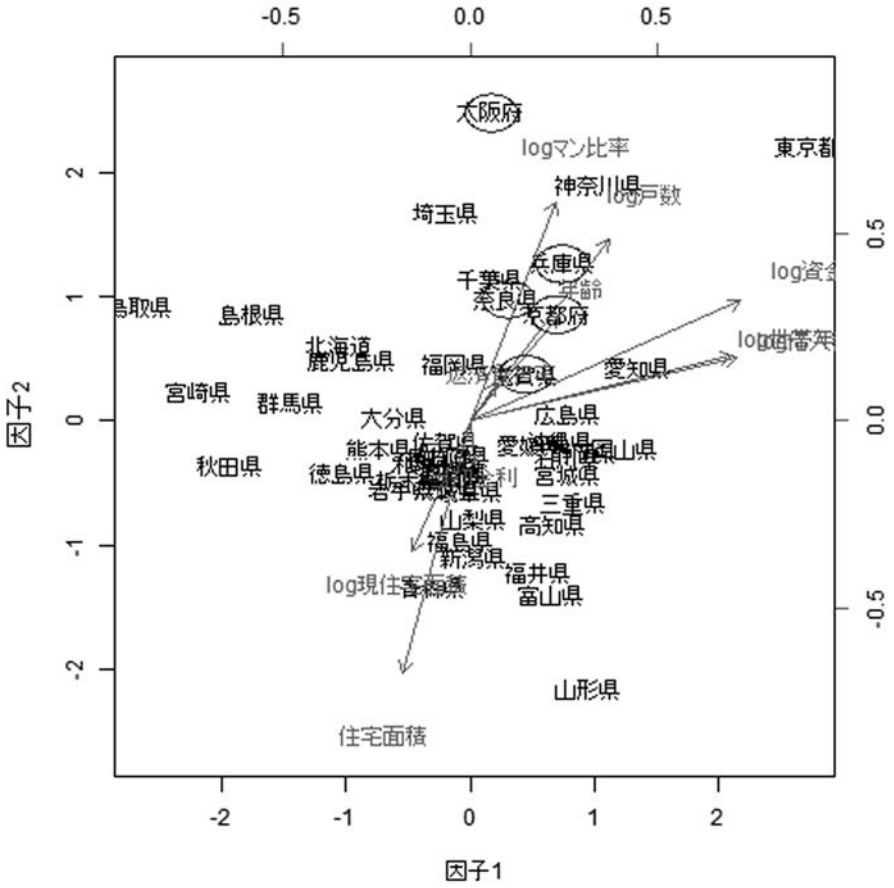
❻ ベイジアン・ネットワークによる因果関係探索の試み

主成分分析や因子分析では、各指標の重要度を要約することができる。しかし、各指標相互はどのような関連性があるのだろうか？　特に、各指標のどちらが原因で、どちらが結果なのだろうか？　そのような疑問に答え、変数相互の因果関係を、有向非循環グラフ（Directed Acyclic Graph：DAG）として構築することを目指す試みが、ベイジアン・ネットワークである。

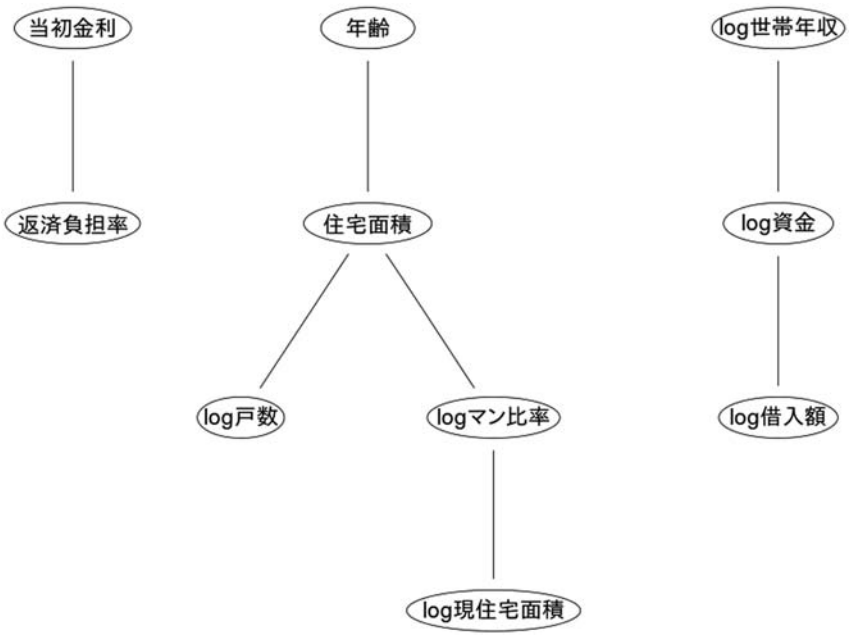
図表 11　因子分析による因子負荷量



図表 12　因子分析による因子負荷量



図表 13　Grow-Shrink アルゴリズムによるベイジアン・ネットワーク



データからグラフの構造を推定することを、構造学習という。構造学習アルゴリズムとしては、いくつかの手法が提案されているが、まず最初に制約基準アルゴリズムによるアルゴリズムのグループを取り上げる。制約基準アルゴリズムとは、ある変数間の条件付独立性を検定することにより、条件付独立であるという帰無仮説が棄却されたときに、両変数の間には何らかの関連性があるものとする。仮説検定には、ピアソンの線形相関検定、フィッシャーの Z 検定、相互情報検定などが使用されるが、ここではピアソンの検定を採用する。さらに、検定を実行する具体的な手法も複数存在するが、ここでは Grow-Shrink アルゴリズムを採用する。

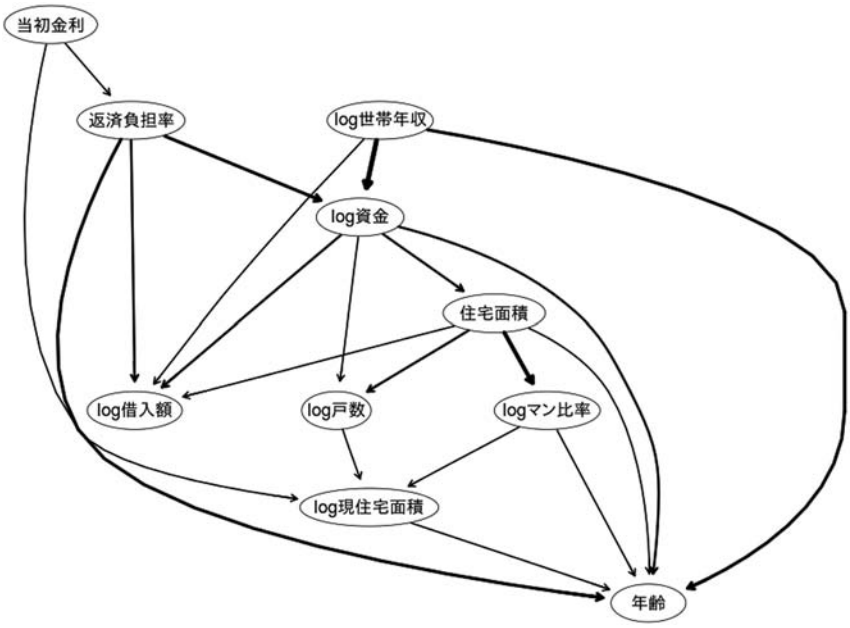
Grow-Shrink アルゴリズムによる構造学習の結果を、【図表 13】に示した。その図の特徴は、第１にすべての関連性が、無向アーク（矢印のない線分）によって表されていることである。すなわち関連性があることは確かだが、どちらが原因でどちらが結果なのかが不明である。第２は、全体が３つのグラフに分離されており、すべての変数を包括したひとつのネットワークが構築されていないという点である。しかし、個々のグラフを吟味すると、比較的納得のいく結果が得られているのではないだろうか。

構造学習の第２のグループは、スコア基準アルゴリズムによるものである。条件付独立性検定によるアルゴリズムは、個々の変数間の独立性に焦点を当てるが、スコア基準アルゴリズムでは、ネットワーク全体がどれだけ良好にデータの構造を表しているかを、何らかのスコアによって評価する。スコアとしては、いくつもの尺度が提案されているが、よく使用されているものは BIC（ベイジアン情報量基準）であり、ここでもそれを採用する。スコアを使用して最適なネットワークを探索する手法にも複数の手法が存在するが、ここでは比較的単純な Hill-Climbing による探索を採用する。

Hill-Climbing では、アークが何もない DAG から始めて、一度にひとつの有向アークを追加、除去、反転させ、スコアが最大になるネットワークを採用する。その結果、得られた最終的な DAG は、【図表 14】に示すとおりである（太い線で描かれたアークは、関連性が強いことを表している。）。

ここでは、当初金利と世帯年収がすべての原因となっている。当初金利→返済負担率→借入額などの因果関係は資金制約を表した関連性を表していると解釈される。また、住宅面積→マンション比率→現住宅面積は、都市性を表した関連性を表していると解釈できる。そして、注目すべきは、すべての変数の最終的な結果が、申込人の年齢とフラット 35 借入額となっていることである。

図表 14　Hill-Climbing アルゴリズムによるベイジアン・ネットワーク



すなわち、ここで表された因果関係の階層を通じて、最終的には住宅取得のタイミングとその借入額が決定されるということが、このネットワークからいうことができそうである。

❼ 今後の課題

ベイジアン・ネットワークを使用すれば、階層的な因果関係を探ることができる。従来、このような統計分析には回帰モデルが多く使われてきた。回帰モデルは、予測変数（説明変数）が横一線に並び、そこから応答変数（被説明変数）に向かってすべての矢印が直接結合するようなモデルであるといえる（交互作用を勘案したモデルは、多少異なるが）。しかしベイジアン・ネットワークの階層的な因果関係を活用すれば、従来の回帰モデルを大幅に改善する可能性を秘めているということができる。

しかし、ベイジアン・ネットワークにおける構造学習アルゴリズムの構築は、現在も多様な手法が提案され続けている状況であり、手法ごとに得られる DAG の構造も様々に異なる。そのどれが最適な手法なのか。あるいは状況によって最適な手法が異なるなら、どのような状況にどのような手法が適しているのだろうか。これらの点を探るのが、今後の課題として残されている。

〈参考文献〉

Hastie, T.J., Tibshirani, R.J. and Friedman, J.（2009）. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer.

Maronna, R., Martin, D. and Yohai, V.（2006）. *Robust Statistics. Theory and Methods*. Weily.

R Core Team.（2014）. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Scutari, M. and Denis J.B., S.（2014）. *Bayesian Networks with Example in R*. CRC Press.

水野　欽司（1996）。多変量データ解析講義。朝倉書店