Bayesian Model Averagingを用いたモデル選択の実証研究
-市街化区域外を対象としてGISを活用した地価推計モデルの構築-

筑波大学　○堤　盛人
筑波大学大学院　宮下将尚
株式会社パスコ　瀬谷　創
寿精版株式会社　佐藤尚秀
朝日航洋株式会社　篠田順弘，今村政夫

1. はじめに

GIS の普及によって，様々な計量・分析モデルのための変数のデータが入手可能となり，多くの場面でモデルの推定精度の向上が期待できる反面，候補となる変数が多くなり，どの変数を採用するのかというモデルの特定化(model specification)が問題となる．従来，回帰モデルの特定化に際しては，変数増減法(ステップワイズ法)が用いられてきた．しかし，変数増減法によるモデル特定化は，最も良いと考える一つのモデルを求めるものであり，特定されたモデルに対する不確実性は考慮されない．これに関して，モデルの不確実性を明示的に考慮した Bayesian Model Averaging(BMA)と呼ばれる手法が考案され，理論的・実証的な研究が積み重ねられている．

本研究では，従来，地価分析の研究対象にほとんどならなかった市街化調整区域，非線引き区域を「市街化区域外」と総称し，その標準地を対象として，GIS によってモデルで用いる説明変数の数が増えた際のモデル特定化に関して，BMA を手法として活用することにより解決を図る．

2 分析の概要と予備的分析

2.1 対象地域

本研究では，茨城県を対象として，平成19年度の地価公示による住宅地の標準地のデータを用いて地価モデルを構築する．ここで，住宅地とは，地価公示において用途が分類されている内の，住宅地，宅地見込地，準工業地，市外化調整区域内の現況宅地とする．該当する標準地は161地点である．

表1　基本等計量

	サンプル数	平均値(円/m^2)
全域	656	40,606
市街化区域	495	47,443
市街化調整区域	104	16,878
非線引き区域	57	24,213

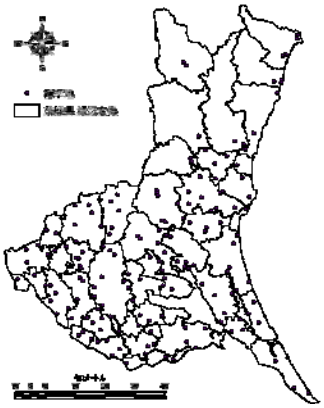

図1　茨城県の市街化区域外の標準地(平成19年度)

本研究では，都市計画区域における市街化調整区域及び，非線引き区域を「市外化区域外」と呼称する．従来，市街化区域外は地価分析の対象にはほとんどされてこなかったが，実際に推定を行ってみると地価公示から得られる情報だけでは充分な推定精度を得られず，GISを活用して説明変数の候補を増やす必要がある．その際問題となるモデル特定化に関して，変数増減法及び，BMAによってモデル構築を行い，その結果の比較を行う．

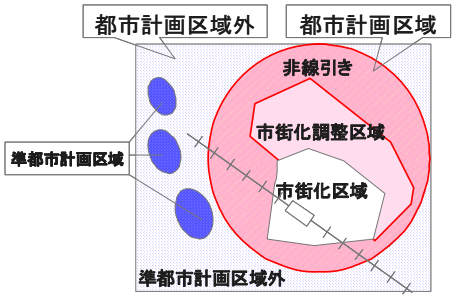

図 2　市街化区域外の概要

2.2 分析に用いるモデル

地価公示から得られたデータを次の線形回帰モデルの式に当てはめ，地価関数を推定する．

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (1)$$

ここで，nをデータ数，$(p-1)$を説明変数の数として，$\boldsymbol{y}$は公示地価 (円/m^2) の$n\times1$ベクトル，$\boldsymbol{X}$は$n\times p$の説明

変数行列，$\boldsymbol{\beta}$は$p\times1$のパラメータベクトル，$\boldsymbol{\varepsilon}$は平均0，分散σ^2の*i.i.d*誤差の$n\times1$ベクトルである．

2.3 予備的な分析

本研究では，まず，地価公示から入手可能な標準地の地価に関する情報の内，定量的な項目を説明変数としてモデルの構築を行った．都市供給施設については，各標準地において下水道か都市ガスのいずれかが整備されていれば 1，両方とも整備されていなければ 0 としている．また，標準地の前面道路の幅員に関する情報もダミー変数とした．具体的には，幅員が 0m～4m，4m～6m，6m～8m，8m より広いときの 4 つに分類し，幅員(0-4)ダミー，幅員(4-6)ダミー，幅員(6-8)ダミーにおいて該当する場合を 1 とした．

表 2 地価公示から取得した変数とパラメータの推定結果

地価公示から取得した変数	係数	標準誤差	t値
切片	33408	4541	7.36
最寄駅距離(km)	−789	132	−5.96
容積の割合(※)	−8033	1916	−4.19
幅員(0-4)ダミー	3115	2440	1.28
幅員(4-6)ダミー	4250	2182	1.95
幅員(6-8)ダミー	4435	2380	1.86
都市供給施設ダミー	4111	1272	3.23
自由度調整済み決定係数	0.329		

※容積率は通常(%)で表示されるが，ここでは 100 で除した数値を用いている．

2.4 地価公示から取得可能な説明変数の追加

公示地価のデータには，標準地周辺の様子を文章で記述した「周辺地利用現況」と呼ばれる項目が存在する．この項目から，重複している単語などをキーワードとして抽出し，ダミー変数とした．表3に示す15個のダミー変数を説明変数に加えてモデルの推定を行ったが，これら，地価公示から入手可能なデータだけでは十分な推定精度が得られなかったため，地価の決定要因に影響を及ぼしていると考えられるデータを独自に取得することとした．

表 3 周辺地利用現況から取得した変数

周辺地利用現況から取得	
住宅ダミー	周辺に住宅がある;1，ない;0
農家ダミー	周辺に農家がある;1，ない;0
工場ダミー	周辺に工場がある;1，ない;0
公共施設ダミー	周辺に公共施設がある;1，ない;0
空地ダミー	周辺に空地がある;1，ない;0
店舗ダミー	周辺に店舗がある;1，ない;0
分譲住宅地域ダミー	標準地周辺が分譲住宅地域である;1，ない;0
既成住宅地域ダミー	標準地周辺が既成住宅地域である;1，ない;0
住宅地域ダミー	標準地周辺が住宅地域である;1，ない;0
商業地域ダミー	標準地周辺が商業地域である;1，ない;0
農家集落地域ダミー	標準地周辺が農家集落地域である;1，ない;0
大規模開発ダミー	標準地周辺が大規模開発地である;1，ない;0
中規模開発ダミー	標準地周辺が中規模開発地である;1，ない;0
道沿いダミー	標準地が国道・県道沿いである;1，ない;0
区画化ダミー	標準地周辺が区画整理等されている;1，ない;0

3 GIS を用いた変数の追加について

地価公示から得られる情報のみでは，モデルの推定精度が十分では無い．GISを活用することで説明変数を増やし，推定精度を上昇させることを試みた．以下，3つの観点からGISを利用して10種類の説明変数を取得した．(表4)

3.1 人口密度について

人口密度の高さは，地価と密接な関係があると考えられる．平成 17 年度国勢調査から，町丁目ごとに人口密度を算出し，GIS によってデータに結合した．

3.2 標準地からの距離に着目した変数

県庁所在地にある水戸駅，関東圏の中心地である東京駅までの距離が近い標準地ほど，地価は高く算出されると予想される．同様に標準地から至近の市街化区域までの距離や，インターチェンジまでの距離を GIS によって測定し，説明変数とした．

また，迷惑施設との距離が近い土地の地価は低くなることが予想される．迷惑施設として焼却場，火葬場，原子力発電所を想定し，距離を測定し説明変数とした．

3.3 地理的な要因に着目した変数

茨城県内にある霞ヶ浦は，地価に対して何らかの影響を与えている可能性が考えられる．同様に，茨城県の標高にも着目し，GIS によって説明変数とした．

表 4 GIS によって取得した変数

GISを用いて取得した変数
人口密度(千人/km2)
水戸駅までの距離(km)
東京駅までの距離(km)
市街化区域までの距離(km)
インターチェンジまでの距離(km)
焼却場までの距離(km)
火葬場までの距離(km)
原子力発電所までの距離(km)
霞ヶ浦までの距離(km)
地点の標高(m)

4 モデル特定化の問題に対する対策

4.1 説明変数の増加に伴うモデルの特定化の問題

説明変数を増やすことによって推定精度の向上が期待される反面，候補となる説明変数が多くなると，どの変数を採用するのかというモデルの特定化(model specification) が問題となる．説明変数の候補が多い場合には，通常，*F* 値等を用いて変数選択を行う変数増減法(ステップワイズ法)がモデル選択の手法の一つとして用いられる．変数増減法は多くの市販の統計ソフトウェアにも機能として備わっており，その手順に従えば，誰がやっても同じ一つのモデルが得られるという利点を有する．しかしながら変数増減法は，最も良いとされる一つのモデルを求めるものであり，特定されたモデルの不確実性は考慮されない．これに関して

90 年代中頃，モデルの不確実性を明示的に考慮した Bayesian Model Averaging (BMA) と呼ばれる方法が考案され，理論的・実証的な研究が積み重ねられている（例えば，Rafety and Madigan(1998)；Hoeting *et al.* (1999)）．

4.2 Bayesian Model Averaging

BMA とは一言でいえば，あるモデル下での推定値を，そのモデルの事後確率で重みづけし，それを考慮するすべてのモデルについて平均したものである．以下，BMA について簡単に説明する．

$M=[M_1,M_2,...,M_q]$ を考えられるすべてのモデルの集合としよう．ただし，$q=2^p$ である．今，各モデルの事前確率を $p(M_k)$ としたとき，ベイズの定理によりモデルの事後確率が次式により与えられる．

$$p(M_k \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid M_k)p(M_k)}{\sum_{l=1}^{q} p(\boldsymbol{y} \mid M_l)p(M_l)} \tag{2}$$

ここで，

$$p(\boldsymbol{y} \mid M_k) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k \mid M_k)d\boldsymbol{\theta}_k \tag{3}$$

は，モデル M_k に対する周辺尤度であり，$\boldsymbol{\theta}_k$ はモデル M_k に対するパラメータベクトルである（$\boldsymbol{\theta}=(\boldsymbol{\beta},\sigma^2)$）．$\boldsymbol{\beta}_k$ に対する事後分布は，次式により得られる．

$$p(\boldsymbol{\beta}_k \mid \boldsymbol{y}) = \sum_{k=1}^{q} p(M_k \mid \boldsymbol{y})p(\boldsymbol{\beta}_k \mid M_k, \boldsymbol{y}) \tag{4}$$

ただし，モデルに変数が含まれていない，すなわち $\boldsymbol{\beta}_k = 0$ となる場合は平均計算から除く（Montgomery and Nyhan (2008)）．

モデル M_1 と M_2 のどちらが支持されるかは，ベイズファクターと呼ばれる，周辺尤度の比を用いて判断することが多い（例えば，砂田 (2006)）．ベイズファクターは次式により与えられる．

$$B_{12} = \frac{p(\boldsymbol{y} \mid M_1)}{p(\boldsymbol{y} \mid M_2)} \tag{5}$$

実際には，ベイズファクターの常用対数値 $\log_{10}B_{12}$ を用いた Jeffreys の基準などによって判断する．例えば，$B_{12}>1$ であれば，M_1 は M_2 に比べて強く支持されよう．

ここで，ベイズファクターは，計算負荷の問題から解析的にこれを解くのが困難である．本研究では現在最も広範に用いられる BIC(Bayesian Information Criterion)を用いたベイズファクターの近似を用いているが，近年では，MC^3 (Markov chain Monte Carlo model composition) を用いた，シミュレーションで周辺尤度を直接求める方法が提案され，有用性が確認されている．これらの比較等については，別の機会に報告したい．

また，M_k の事前確率 $p(M_k)$ は $1/q$，すなわち一様とすることが多く，本研究もこの方法を用いているが，このような方法は，説明変数同士の強い相関，すなわち多重共線性が存在するときに問題があるとの指摘がある（例えば，Hoeting *et al.* (1999)）．

BMA は，政策科学，経済学（格差分析），生態学など幅広い分野で用いられているが，ヘドニック・アプローチを用いた地価の推定モデルについては，Cotteleer *et al.* (2007) 等，これまでのところ限られた例しかない．

変数増減法とBMAの推定結果の比較に関してWang, Zhang and Bakhai (2004)では10回のシミュレーション(試行)を行っており，そこでは，BMAは10回のうち8~9回は最適なモデルを選ぶが，変数増減法ではすべての試行(occasion)において標準誤差を過小評価した変数を有意な変数として選出するという結果を得ている．

4.3 変数増減法とBMAの推定結果の比較

以下，変数増減法による推定結果と，BMA による推定結果の比較を行う．なお，前者には市販の統計ソフトである SPSS を用いている．本研究では，変数の投入時の有意水準を 5%, 除却時の有意水準を 10%として変数増減法を用いて推定を行った．BMA には R の BMA パッケージを用いている．

表 5 は, BMA（事後確率が大きいものから順に 5 つ），及び変数増減法によるモデルの推定結果を示している．自由度調整済み決定係数を比較すると，BMA の *model* 1 では 0.618 であるが, 変数増減法は 0.560 である. *model* 1 だけでなく *model*5 まで，BMA による推定結果は，変数増減法による推定結果よりも高い決定係数となることが分かる．また，GIS の使用により入手した説明変数は，変数増減法では，ほとんど選択されていないことが分かる．

表 5 BMA と変数増減法による推定結果

説明変数	model1	model2	model3	model4	model5	変数増減法
(定数)	43289	46502	47669	27115	44683	24543
最寄駅_距離(km)	−512	−512	−489	−505	−493	−559
容積ダミー	−5266	−6079	−6632	−5647	−5830	−4209
幅員(0-4m)ダミー	.	.	.	.	.	
幅員(4-6m)ダミー	.	.	.	.	.	
幅員(6-8m)ダミー	.	.	.	.	.	
都市供給施設ダミー	2382	.	.	2380	2057	2442
住宅ダミー	2611	2652	2565	2302	2545	2976
農家ダミー	.	.	.	.	.	
工場ダミー	.	.	.	.	.	
公共施設ダミー	.	.	.	.	.	
空地ダミー	.	.	.	.	.	
店舗ダミー	.	.	.	.	.	
分譲住宅地域ダミー	.	.	.	.	.	
既成住宅地域ダミー	.	.	.	.	.	
住宅地域ダミー	.	.	.	.	.	
農家集落地域ダミー	.	.	.	.	.	
商業地域ダミー	.	.	.	.	.	
大規模ダミー	.	.	.	.	.	
中規模ダミー	.	.	.	.	.	
道沿いダミー	.	.	.	.	.	
区画化ダミー	.	.	5154	.	4223	
人口密度(1000人/km2)	3649	3821	3419	3521	3343	3698
水戸_距離(km)	−215	−219	−222	−518	−217	
東京_距離(km)	−158	−169	−169	.	−160	
市街化区域_距離(km)	455	437	417	291	436	
インターチェンジ_距離(kr	.	.	.	.	.	
焼却場_距離(km)	.	.	.	.	.	
火葬場_距離(km)	.	.	.	−357	.	
原子力発電所_距離(km)	.	.	.	397	.	
霞ヶ浦_距離(km)	211	221	220	194	211	
標高(m)	−112	−112	−108	−107	−108	
nVar	10	9	10	11	11	5
R2(adjusted)	0.618	0.606	0.615	0.624	0.623	0.560
BIC	−114.6	−113.5	−113.3	−113	−112.8	−
post prob	0.079	0.045	0.042	0.036	0.032	−

表 6 は, BMA によって算出された事後平均(Posterior mean)と $p(\theta \neq 0 \mid y)$，及び変数増減法による推定結果である．数値が高いほど，地価の決定要因に影響を与えている変数であると言える．変数増減法で選択された変数以外にも高い事後確率を算出している変数は多く，地価の決定要因に影響を与えているであろうことが推

測される．例えば，変数増減法で選択された都市供給施設ダミーの BMA における事後確率 $p(\theta \neq 0 | y)$ は，57.5 となっているが，これに対して変数増減法では選択されていない変数の内，東京_距離(km)が 89.3，市街化区域_距離(km)が 69.2，霞ヶ浦_距離(km)，標高(m)が 89.7 となっており，都市供給施設ダミーよりも高い事後確率を算出している．従って，実際に任意地点の土地価格を推定する際には，これらの GIS によって新たに追加された変数が地価の決定要因に影響を与えている可能性も考慮に入れるべきである．

表6 事後平均と変数増減法による推定結果

	BMA		変数増減法
説明変数	p(θ≠0|y)	事後平均	推定値
(定数)	100.0	36101	24543
最寄駅_距離(km)	100.0	−517	−559
容積	99.3	−5856	−4209
幅員(0-4m)ダミー	0.0	0	
幅員(4-6m)ダミー	0.0	0	
幅員(6-8m)ダミー	0.0	0	
都市供給施設ダミー	57.5	1346	2442
住宅ダミー	76.2	1936	2976
農家ダミー	0.0	0	
工場ダミー	0.0	0	
公共施設ダミー	0.0	0	
空地ダミー	2.1	51	
店舗ダミー	0.0	0	
分譲住宅地域ダミー	3.5	151	
既成住宅地域ダミー	1.1	18	
住宅地域ダミー	0.0	0	
農家集落地域ダミー	0.0	0	
商業地域ダミー	3.4	286	
大規模ダミー	0.0	0	
中規模ダミー	1.2	70	
道沿いダミー	3.4	−81	
区画化ダミー	31.1	1579	
人口密度(1000人/km2)	100.0	3560	3698
東京_距離(km)	89.3	−76	
水戸_距離(km)	47.1	−315	
市街化区域_距離(km)	69.2	268	
インターチェンジ_距離(km)	1.5	2	
焼却場_距離(km)	0.0	0	
火葬場_距離(km)	39.5	−135	
原子力発電所_距離(km)	43.6	165	
霞ヶ浦_距離(km)	89.7	189	
標高(m)	89.7	−94	
自由度調整済み決定係数		0.604	0.560

BMAによって構築された地価の推計モデルを実際の土地価格の推定に用いる場合は，事後平均を用いる．これは，あるモデル下での推定値を，そのモデルの事後確率で重みづけし，それを考慮するすべてのモデルについて平均したものである．BMAによる事後平均と，変数増減法によって構築されたモデルの比較を行う．

2つの手法によって構築されたモデルによって任意地点の地価推定をする際の推定精度を見るために，4-fold Cross Validation を行い二乗平均平方根差(root mean square error; 以後RMSE）を算出し比較を行う．

Cross Validation(交差検定法)とは，データをランダムにn個のデータ郡に分割する．ある1つのデータ郡をテストデータ，残りのデータ郡をトレーニングデータとする．トレーニングデータによって構築したモデルを用いてテストデータの推定を行う．これをn回繰り返し，平均を取った値を推定精度とする．本研究では，テストデータ算出した残差を基に，RMSEを算出した．

表7はBMAと変数増減法の結果をCross Validationによって比較した結果である．誤差の絶対値及び，誤差を平均価格で割った誤差率を算出して，手法ごとに比較を行った．BMAのほうが，変数増減法よりも誤差が少なく算出されていることが分かる．

RMSEでは単位面積当たり約250円，誤差率では1.2%ほど誤差が少ない結果となった．

表7 4-fold Cross Validation の結果比較

RMSE	事後平均	変数増減法
4-fold_平均(円/m^2)	6967	7205

誤差率	事後平均	変数増減法
4-fold_平均(%)	35.6	36.8

5 おわりに

BMA及び，変数増減法によって構築されたモデルを比較すると，BMAによって選択されたモデルの方が，自由度調整済み決定係数が高くなる結果となった．

モデルによる推定の結果算出された誤差を見ると，BMAの方が確かに変数増減法より推定誤差は少ない結果となったが，その差はあまり大きなものでないことが分かる．

変数増減法で選択されない説明変数の内，GISを用いて追加した変数である東京_距離(km)，市街化区域_距離(km)，霞ヶ浦_距離(km)，標高(m)は，BMAでは事後確率が高く算出されており，地価に対して影響を与えていると考えられる．すなわち変数増減法で見逃されたこれらの変数が，地価に対して影響を与えていることが，BMAを用いることにより確認することができた．

参考文献

砂田洋志 (2006). 周辺尤度の推定理論とその応用－ギブス・サンプリング法，及び M=H 法で推定した場合－，「山形大学紀要（社会科学）」, 36 (2), 2345.

Cotteleer, G., Stobbe, T., and van Kooten, G.C., (2007). Bayesian model averaging in the context of spatial hedonic pricing: An application to farmland values, No 2007-07, Working Papers from University of Victoria, Department of Economics.

Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial, *Statistical Science*, 14 (4), 382417.

Montgomery, J., and Nyhan, B. (2008). Bayesian model averaging: Theoretical developments and practical applications, http://www-personal.umich.edu/~bnyhan/montgomery-nyhan-bma.pdf

Rafety, A.E., and Madigan, D. (1998). Bayesian model averaging for linear regression models, *Journal of the American Statistical Association*, 97,179191.

Wang, D., Zyang, W., and Bakhai, A. (2004). Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression, *Stasistics in Medicine*, 23, 3451-3467