〈研究ノート〉

地理情報を利用した探索的データマッチングの試み

経済学科　小　川　　浩

要旨

近年，個票を用いた分析がかなり一般的なものとなってきている。しかし，自分で行った調査以外では個票データに含まれている情報だけでは分析目的には不足することが多い。その場合，何らかのキー情報を用いて他の調査とのデータ結合を行うことになる。再集計のために個票を用いるのであれば，集計する範囲での特性値がある程度の誤差範囲に収まれば問題ないが，個人の行動分析のために個票データ自体を分析対象とする際には可能な限り正確な推定値を与えるデータ結合を行う必要がある。

本稿では「全国消費実態調査」の資産データから土地評価額を計算するために地価情報（地価公示・地価調査）と組み合わせた例を用いて，マッチング過程で地理情報を援用することにより，推定が改善される可能性があることを示した。

1．マイクロデータ分析における複数データソース結合時の問題点

近年，SSJ データアーカイブ[1]や政府統計ミクロデータの試行的提供[2]，消費生活に関するパネル調査[3]などにより個票データが利用可能になってきたことをうけ，集計データではなく個票を用いた分析がわが国でも一般化しつつある。しかし，各調査はそれぞれの目的のために設計されているため，再集計あるいは再分析を別の目的で行う際には分析に必要な情報の一部が個票に含まれていないことが多い。この場合，個票自体には含まれていない情報を他の調査から得て，何らかのキーを用いて結合した上で処理することになる。個票データの使用目的が再集計であれば，再集計範囲での特性値（平均や標準偏差など）が必要な誤差範囲に収まれば問題は生じないが，個票データ自体を個人の行動分析に用いる場合には集計による平均化が期待できないだけに慎重な処理が必要となる。

このことを簡単な例で示そう。表１は仮想的な調査１から得られたデータ A, B と調査２から得られたデータ C を何らかの方法で結合して作成したデータセットの一部を示している。このデータセットを地域ごとの再集計に用いた場合，結果は表２のようになり地域「い」と地域「ろ」はデータ A, B, C からみてかなり似た地域と判断できる。しかし，データ A とデータ C を個別にプロットした図１を見ると，二つの地域間ではデータ A とデータ C の関係が逆になっている。データ A とデータ C は別の調査で得られたデータを結合しているため，図１に示したような傾向の違いが本当にデータの性質によるものか，結合時の手法の問題なのかを判別することは困難である。

実際にはこのように極端な差が出ることは稀であると期待されるが，この例は，データ結合時には得られる情報を可能な限り利用することが望ましいことを示している。

本稿では，全国消費実態調査の実物資産額推計に用いられている地価公示・地価調査データとのマッチ

1)　東京大学社会科学研究所附属　日本社会研究情報センター

2)　一橋大学経済研究所附属　社会科学統計情報研究センター

3)　財団法人　家計経済研究所

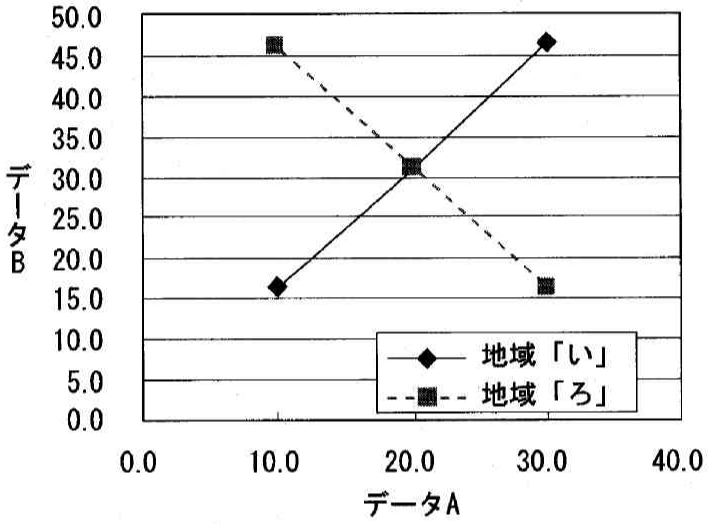
表１　２つの調査を結合したデータセット例

レコード番号	地域	調査１		調査２
		データ A	データ B	データ C
1	い	10.0	5.0	16.5
2	い	20.0	10.3	31.2
3	い	30.0	15.9	46.7
4	ろ	10.0	5.7	46.1
5	ろ	20.0	10.6	31.0
6	ろ	30.0	15.7	16.2

表２　地域別集計結果

変数名	地域	平　均	標準偏差
データ A	い	20.0	10.0
	ろ	20.0	10.0
データ B	い	10.4	5.5
	ろ	10.7	5.0
データ C	い	31.5	15.1
	ろ	31.1	15.0

図１　データ A, C の地域別プロット



ングを例に取り，総務省統計局が集計用に採用しているマッチング方法と（高山他 1989）で用いられた方法，さらに筆者が「全国消費実態調査」の個票から土地資産額を推計する際に用いた地理情報を援用したマッチング手法を比較し，マイクロデータ分析における地理情報の重要性を確認する。

2.「全国消費実態調査」地価推定の方法

「全国消費実態調査」は総務省統計局が５年に１回行っている調査であり，国民生活の実態について家計収支，貯蓄・負債，耐久消費財，住宅・宅地などの家計資産を総合的に明らかにすることを目的としている。しかしながら，住宅や宅地については，所在地（調査区あるいは市区町村）および面積の情報は調査されているものの，価格は調査対象となっていない。そのため，土地資産額の推定を行うためには何らかの地価情報とマッチングを行い，単位面積あたりの価格を求める必要がある。

わが国で土地の取引価格の指標として調査・公開されているデータとしては国土交通省が毎年公表している公示価格と，都道府県が公開している地価調査が一般的に利用される。他の価格指標としては路線価や固定資産評価額などもあるが，土地資産評価としては転売時にどの程度の価格になるかが重要であることを考慮すると，取引時の参考として公開されている地価公示・地価調査のデータを用いるケースが多い。

そこで問題となるのが全国消費実態調査と地価公示・地価調査のデータのマッチング方法である。地価公示・地価調査はそれぞれ標準地，基準地というポイントでの価格を公表しているものであり，たとえ全国消費実態調査の調査区に地価公示や地価調査の調査地点が入っていたとしても調査区内の全ての土地の単価が調査地点とあまり差がないかどうかは定かでない。また，そもそも調査区の中に調査地点が存在す

るか否かも全く保証されない。

以下では，総務省統計局が「全国消費実態調査」の集計表を作成する際に利用している方法，（高山等1989）で個別世帯の資産推計のために用いられた推定方法，及び筆者が個別世帯の資産推計のために用いた方法を概観する。

2．1　総務省統計局の方法

総務省統計局は集計データを作成するために土地資産を評価している。そのため各世帯データへのマッピングについては（1）現居住地宅地，（2）現居住地以外の宅地，の2つに区分して比較的単純な方法で行っている。

(1)　現居住地宅地

宅地単価は，各調査単位区（国勢調査での近接する2つの調査区をまとめたもの）に最も近い地価公示の標準地あるいは地価調査の基準地を選び，その1平方メートルあたり評価額を用いる。標準地あるいは基準地が十分な密度で存在していればこの方法はかなりよい近似を与えることが期待できるが，残念ながら町村レベルでは標準地と基準地を合わせても高々数件というケースが少なくない。この場合「最も近い」といっても，場合によってはkm単位で離れている可能性もあり，果たしてどの程度適切な近似となっているかは疑問である。

ただし，この推計はあくまで集計用の推計であるから，集計範囲が広ければあまり問題は生じない可能性が高い。

(2)　現居住地以外の宅地

所在地情報を市区町村単位でしか調査していないため，地価公示および地価調査の調査地点のうち，当該市区町村に含まれる「住宅地」および「市街化調整区域内現況宅地」を抽出し，中位数を評価額とする。この方法から分かるように，同一の市区町村に含まれる現居住地以外の宅地は全て同じ価格として扱われている。しかしながら，市街化調整区域内の現況宅地については転売後に住宅建設が可能である保証がないことを考えれば過大推定となっている可能性は否めない。

2．2「日本の家計資産と貯蓄率」法

（高山他1989）では土地面積を調査していない年度の全国消費実態調査データを用いて計算しているためまず住宅敷地面積の推計から行い，推計した面積に調査区が存在する市・郡別の用途区分住宅地の公示価格分布から求めた中位数を価格として乗じている。この際，公示価格の件数が少ない郡部では，近隣市部の地価の第1・十分位または第1・四分位を郡部の地価として用いたとしている。つまり，同一市郡に含まれる調査区については同一の地価を用い，なおかつ郡部では公示価格の件数が少なければ近隣市部のデータで近似する[4]という方法を用いていることになる。

2．3　地理情報を用いた推計方法

属性区分

全国消費実態調査の調査票[5]では現居住地宅地とそれ以外の宅地では調査項目の細かさにかなりの差がある。現居住地宅地については，建物設備として水洗式便所，風呂・シャワー，都市ガス，プロパンガスの有無や業務用部分の面積，耕地面積などの追加情報があるが，それ以外の土地については単に所在地（市区町村レベル）と面積のみの調査となっている。一方，地価公示あるいは地価調査では調査地点の住所，単位面積あたり価格，面積，土地の形状，利用区分，ガス・水道・下水の有無，最寄り駅までの距離，建坪率・容積率，区域区分，利用現況，周辺地利用現況などの情報が公開されている。

---

4)　（高山他1989）には近隣市部の第1・十分位と第1・四分位の使い分け基準は書かれていない。

5)　以下，特にコメントしない場合は平成11年調査の調査票での調査項目を前提とする。当然，年次によってはここで利用した属性を調査していないこともある。

これらの情報のうち，住所情報以外で実際にマッチングに使えるものを選ぶと表3の3属性になる。これらの属性は，(1) 属性ごとにある程度の地価情報が存在する，(2) 属性によって地価が大きく変動する，(3) 全国消費実態調査のデータから推定可能である，の3つの条件を満たしている。3属性がそれぞれ2つの値を取り得るので組合せ数は8通りとなるが，実際には地価データ数があまり多くない組合せもある。サンプル数不足から市区町村ごとに安定した推定値が得られない可能性もあるため，実際には表4に示した10種類の組合せおよび市区町村全体による中央値および平均値を地価データとして計算した。表4の件数は，平成11年の地価公示データに含まれる当該属性を満たす標準地数である。現居住地以外の家計資産推定には，属性を付けない中央値を用いる。

サンプル数の確保と地理情報

属性を細かく分けたことにより，市区町村単位での集計では十分な件数の地価データが得られず，安定した推定値を計算することが困難になるという問題が生じる。地価データサンプル数の不足に対し，(高山他 1989) では郡部については郡単位での中央値を計算する，あるいは近隣の市部データの一部を利用するといった処理をしていた。本稿では，郡のようにあまり意味のない行政単位よりも実際に人々が移動する距離に着目して中央値を計算する範囲を選ぶことを試みる。具体的には，着目している市区町村の役所・役場から隣接する市区町村の役所・役場までの距離を計算し，近い市区町村から順次中央値を計算する範囲に加えていくという方法を採用した。範囲拡大を打ち切る基準は，(1) 平均の90% 信頼区間の幅が平均値の10% 以下にまで縮小した，(2) 計算に用いた地価の数が50カ所を越えた，(3) 計算に用いた市区町村の数が9を越えた，のいずれかの条件が成立した場合である。この際，各市区町村役所・役場の座標は国土交通省が公表している国土数値情報の経緯度データから得た。

この場合であっても郡部の地価データに近隣市区の地価データをそのまま入れた場合には都市中心部の極端に高い地価が入り問題が生じる可能性があるため，郡部のデータに市のデータを加える場合は当該市内での中央値以下のデータのみを，区のデータを加える場合には当該区内での第1・四分位以下のデータのみを用いている。

さて，ここまでの考察がどの程度妥当であるか，実際の地価公示データを地図上にプロットしたデータで観察してみる[6]。図2は静岡県東部の1999年の地価公示データを地図上にプロットしたものである。○で表示されているポイントは市街化区域内の標準地，△で表示されているポイントは調整区域内の標準地を表しており，各ポイントの色はこの範囲での地価の四分位数になるように塗り分けられている。色が濃い方が安く，色が薄い方が高い地点を表している。

まず，調整区域と市街化区域を比較すると，調整区域内の標準地は一カ所の例外を除き全て第1・四分位以下の地価となっている。このことは，表4での大きな属性区分として市街化区域とそれ以外を用いたことが妥当であることを示している。

次に，地価データの件数が少なすぎる際に，郡単位で集計する方法と距離を重視して集計する市区町村を決定する方法を比較してみる。図3の影をつけた部分は，この地域の郡部（田方郡）に含まれる町村[7]を示している。いま着目している町が田方郡北端で三島市，沼津市と接している函南町であると仮定しよう。函南町の地価公示標準地は7カ所[8]あるが，住宅地のデータは南西部に集中しており，比較的地価が高い。しかし，函南町の地価を求めるために田方郡全体の標準地データを用いて計算した場合は，南の町に多くある低地価の標準値の影響を受けて過少推定となると予想する。

---

6)　©ZENRIN，許諾番号 Z06B-第2092号の地図利用

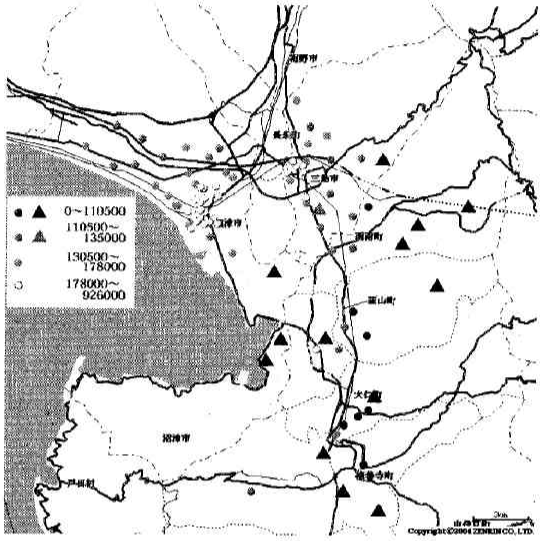
7)　平成16年，17年に田方郡の多くの町村が合併したため，平成18年1月現在の田方郡に属する町は函南町だけである。ここでの図示は平成11年当時の田方郡を用いている。

8)　図ではデータ点が近すぎて重なって見えるため6カ所しか見えない。

表３　全国消費実態調査と地価公示のマッチング属性

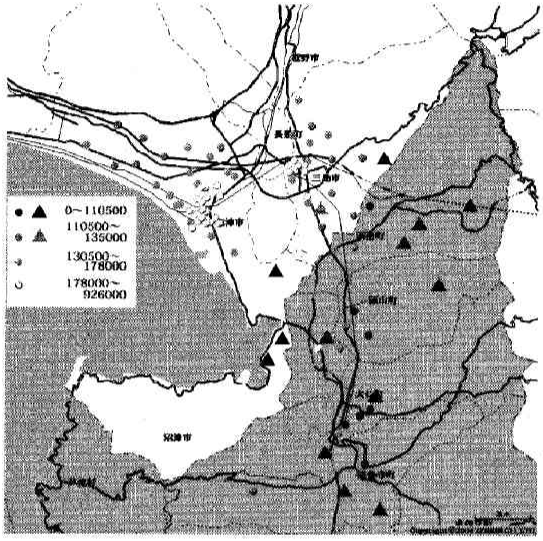
全国消費実態調査	地価公示
都市ガスの有無	ガス供給の有無
耕地の有無	市街化地域／調整地域
業務用面積の有無	事務所併設住宅／住宅のみ

図２　静岡県東部の標準地分布と地価



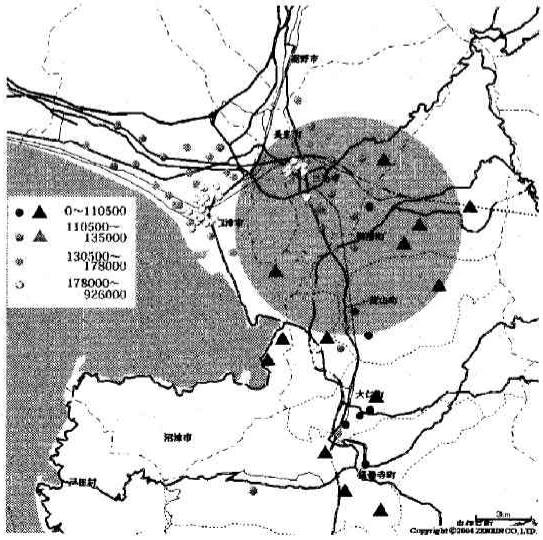
資料出所：国土交通省「地価公示」平成11年より筆者作成

図３　郡単位での集計



資料出所：国土交通省「地価公示」平成11年より筆者作成

図４　距離を基準とした集計（イメージ）



資料出所：国土交通省「地価公示」平成11年より筆者作成

表４　地価公示における属性ごとの標準地数

属性の組合せ	件数
非市街化区域，ガスあり	116
非市街化区域，ガスなし	1499
非市街化区域全域	1615
市街化区域，ガスあり，事業併設	1842
市街化区域，ガス有り，事業なし	12227
市街化区域，ガスあり	14069
市街化区域，ガスなし，事業併設	1191
市街化区域，ガス無し，事業なし	8127
市街化区域，ガスなし	9318
市街化区域全域	23387

資料出所：国土交通省「地価公示」平成 11 年より筆者集計

一方，図４は着目している町からの距離を基準に集計範囲を作成するイメージ[9]を表している。函南町を中心に集計範囲を広げた場合には，北隣にある三島市と南隣にある韮山町がまず集計対象になるが，三島市のデータについては「町村の地価を推計している際には市のデータは中央値以下のみ利用する」というルールにより市の中心部にある地価の高い部分は推計に使われない。一方，韮山町については全データが推計用データとして用いられることになり，市部の中心データによるバイアスを排除しつつ，データの件数を増やすことができる。

3. 推定方法の比較

ここでは，上記の方法で家計資産推定時の地価を割りあてた場合にどの程度の割当が可能かを静岡県田方郡函南町のデータを例に計算してみる。

3. 1　総務省統計局方式

総務省統計局が用いている方法は，調査区に最も近い標準地あるいは基準地の地価を調査区の地価として採用するという方式である。図２に示すように，地価情報は町内で均一な分布で採られているわけではないため，調査区をランダムに選択して町内の平均を考えた場合，地価情報が疎に分布している部分のウェイトが高くなる。実際の分布では地価情報が疎な部分は相対的に地価が低い地域に相当しており，全体としては過少評価になる可能性がある。ただし，全国消費実態調査の調査区は国勢調査調査区を流用しているため，同じ大きさのメッシュとして分布しているわけではないため，実際にどの程度の差があるかの評価は難しい。

3. 2　「日本の家計資産と貯蓄率」方式

（高山他　1989）での推計方法は，郡部については郡内での中央値を採用するというものであった。つまり，郡内の地価はどこも同じであると仮定していることになる。田方郡について中央値を計算した結果は表５に示すように 104,000 円／平米であり，実際の地価とはかなりの乖離がある。郡単位より大きな範囲での集計用に用いるならば問題は少ないかもしれないが，個別の家計資産として評価する際には注意が必要である。

9)　実際の処理では，役所・役場間の距離を用いて市区町村単位で集計範囲を拡大していくため，単純な円では図示できない。

表５　地理マッチングと郡名を用いた中央値の比較

(円／m^2)

標準地名	市街化？	ガス？	事業併設？	地価公示	地理マッチング	田方郡中央値
函南-1	Yes	No	No	137000	119500	104000
函南-2	Yes	Yes	No	105000	136000	104000
函南-3	Yes	No	No	125000	119500	104000
函南-4	Yes	No	No	103000	119500	104000
函南 10-1	No	No	No	81200	87500	104000
函南 10-2	No	No	No	94500	87500	104000
函南 10-3	No	No	No	33300	87500	104000

資料出所：国土交通省「地価公示」平成 11 年より筆者推計

3. 3　地理マッチングを用いた割当方式

本稿で提案している方式での割当の結果を表５の地理マッチング欄に示す。郡名によって得た中央値より公示価格に近いデータが得られていることが分かる。例えば函南-1 の地価は，函南町，韮山町，三島市，清水町，伊豆長岡町の５市町，10 カ所のデータから得られた中央値となっている。これらの市町のうち，函南町と同じ田方郡に属するものは韮山町，伊豆長岡町だけであり，三島市と清水町のデータは郡名ベースでの集計では用いられなかったはずのデータである。地理的条件を用いたデータマッチングによって改善された部分と考えていいだろう。

4. まとめと考察

既存の２つの方法は，いずれも地価に付随する土地の情報，例えば上下水道の有無や都市ガスの有無，土地が市街化地域にあるか調整区域にあるかなどの条件を一切無視して単純に距離や同一行政単位に存在するという部分だけを利用している。この理由を推測すると，おそらく地価公示や地価調査の調査地点数があまり多くないため，属性によって区分すると行政単位内で特性値を計算できない程度の数になってしまうからであろう。つまり，データの制限のため，本来ならば区分したい属性を無視していると予想する。

しかしながら，土地の資産価値を考える際にこれらの属性を無視することによりかなりの推定バイアスが生じる。例えば市街化調整区域は原則的に住居の建設が制限されており，現況で住宅となっているからといって転売後もそこに住宅が建設可能である保証はない。この場合，同じ行政単位内に存在していても，市街化地域と調整区域の単位面積あたりの資産価値が同一であると考えることは個別世帯の資産推定という観点からは問題が多い。

本稿で提案した地理情報を用いたマッチング方法は，地価情報のサンプル地点が少ないために属性別の集計が困難となっている問題を解決する一つの方法である。確かに計算量は増えるものの，本稿で行った程度の探索的マッチングであればパーソナルコンピュータの処理速度でも問題なく実行可能であり，個票データを扱う場合は単純なコードマッチングではなく，データの内容に応じて地理的な情報をも加味した探索的マッチングを行うことが望ましい。

参考文献

高山憲之他，「日本の家計資産と貯蓄率」経済分析，経済企画庁経済研究所，第 116 号，1989 年９月