総合都市研究　第65号　1998

集計単位問題についてのノート

青　木　義　次*

要　　約

本論は、集計単位のとり方によって分析結果が異なってきてしまうという集計単位問題についてのひとつの考察である。

最初に数値例を用いて集計単位問題の特質を見、その結果として、一般に言われているように集計単位を細かくすることによってはこの問題は本質的に解決しないこと、従来の相関分析や回帰分析では、集計単位を変化させることで異なる仮定のモデル設定となってしまっていること、空間情報を失ってしまう分析となっていることを明らかにした。

次に、空間での影響範囲を明確にするため、空間相関関数が利用できること、さらに、空間影響関数モデルにより、回帰分析と同様の分析が可能となることを示した。空間影響関数モデルの立場から再び従来の回帰分析を捉えると、回帰分析がまったく空間的情報を無視したモデル設定になっていることをあらためて示した。

最後に、集計単位の大きさがどのような意味を持つかを、フーリェスペクトル空間で考察すると、集計操作が高周波数成分をカットする一種のフィルターとなっていることから、集計単位を大きくすると空間的変数の細かな変動の情報を失うことを示し、集計単位を細かくすることが集計単位問題を本質的に解決することはないが細かな空間変動を捉えるという意味では有効であることを示した。

1．はじめに

コンピュータの普及によって都市データを用いた分析が頻繁に行われるようになってきている。さらに近年のGISのように地理的データを自由に活用することが容易になってきていることが、この傾向を加速している。こうした傾向の中で危惧されるのは、分析対象データが膨大になり分析方法が複雑になるに伴い、その過程で生じる誤り

*東京工業大学工学部建築学科

に鈍感になってしまうこと、また、その誤りが見つけづらくなってきていることである。

集計単位の違いによって分析結果が異なることは、古くから知られており集計単位問題とかエコロジカル・ファラシーとして指摘され、また、空間的な広がりを持つデータを分析する際に生じやすい誤りについては、空間データ分析パラドックスとして都市計画学会のワークショップで指摘されたこともあった。また、Openshaw (1984) およびArbia (1989) によりModifiable Areal Unit Problemとして組織的に研究がなされており、距離データについての集計問題については田頭 (1990) およびOkabe-Tagashira (1996) に綿密な検討が加えられている。しかし、近年の分析の中では、この問題に対しての配慮が全くされていないものも目立つ。

ここでは、こうした現状認識のもとで、この空間データの集計単位問題の一端を明らかにし、誤りを生じる基本的原因がどこにあるのか、また、それを回避するために必要な条件は何かを議論することとしたい。

2．集計単位の違いと分析結果の違いの例

本報告では、空間的に広がった変量間の相関分析、あるいはある変量を他の変量で説明する回帰分析で発生する問題にしぼって議論したい。

問題点を明確にするために、具体的な数値例を用いて検討することから始めたい。都市内の各地点で計測可能な物理量Xと心理的評価値Yとの関係を分析する例を用いることにしたい。

都市データは一般的に二次元的に広がった空間での値として表現されるが、本報告の範囲では、一次元としても問題の本質は変わらないので、以降では空間は一次元であるとして議論することとしたい。具体的な数値例として、6つの地域が線状に並んでいる都市を想定し、それぞれの地域iで計測された物理的環境要因x (i)，i＝1～6と環境の心理的評価値y (i)，i＝1～6のデータが実数で表1のように得られているとする。分析の目的は、心理的評価値Yは物理的環境要因Xと関係しているのかどうか、あるいは、心理的評価値Yは物理的環境要因Xで説明可能かということであるとする。前者は、心理的評価値Yと物理的環境要因Xとの相関分析、後者は、心理的評価値Yを被説明変量、物理的環境要因Xを説明変量とする以下の回帰モデルを前提とする回帰分析に持ち込まれる。

$$y(i) = a\ x(i) + b + \varepsilon \qquad (1)$$

ここで、a、bは推定される回帰係数、εは両変数に独立な正規誤差とされる。

以下では、本質的な問題は両分析方法で大きな差がないので、相関分析を行う想定で議論を進めたい。また、説明変量が複数になったとしても、以下の議論は容易に一般化できるので、本報告の中では、心理的評価値Yと物理的環境要因Xだけを想定する。

表1　数値例（6地域分割）

地域	物理的要因X	心理的評価値Y
1	-1.000	0.000
2	-0.500	-0.866
3	-0.500	0.866
4	0.500	-0.866
5	0.500	0.866
6	1.000	0.000

このとき、第1のケースとして、表1のデータをもとに、変量x (i)，i＝1～6とy (i)，i＝1～6の相関関係を以下のように分析する分析者を想定しよう。すなわち、両変数間の相関係数を求めてみると、

$$C(x, y) = 0.0 \qquad (2)$$

となり、有意性を検定したとしても、両者は独立と判断される。事実、変量x (i)，i＝1～6とy (i)，i＝1～6のデータをプロットした図1を見るかぎり完全にばらついており、両者に関係が無いように見える。

次に、第2のケースとして、隣接する2つの地域をひとつの地域としてまとめて、対象地域を3ゾーンに分割して集計しなおして分析する分析者を想定しよう。つまり、第1地区と第2地区のそれぞれの変量の平均として再定義され、第1ゾー

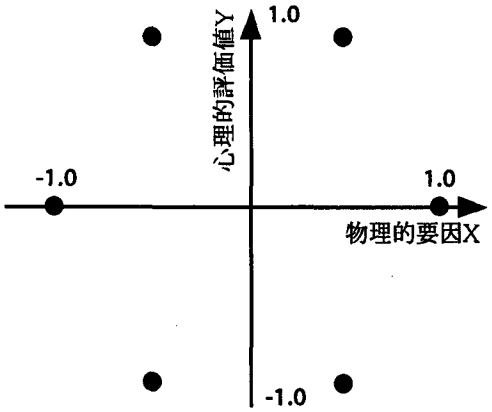


図1　表1データの散布図

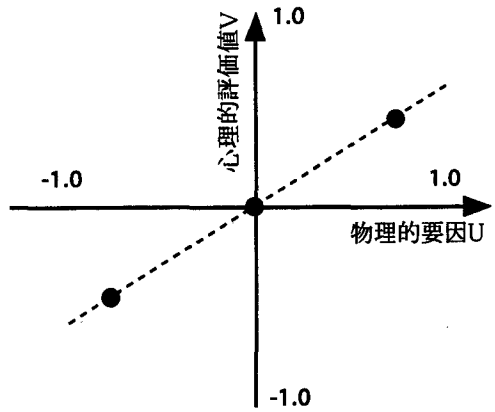


図2　表2データの散布図

ンのそれぞれの値となる。同様に3つのゾーンのそれぞれの変量の値が得られる。この関係は、

$u(j)=(x(2j-1)+x(2j))/2$、$j=1\sim3$　(3-1)

$v(j)=(y(2j-1)+y(2j))/2$、$j=1\sim3$　(3-2)

であり、表1のデータを用いると表2のようになる。物理的環境要因と心理的評価値との関係を調べると、データをプロットした図2を見ると完全に直線の上に乗っており、さらに、変量$u(i)$，$i=1\sim3$と$v(i)$，$i=1\sim3$の相関係数は、

$C(u,v)=1.0$　(4)

となり、第2の分析者は、物理的環境要因と心理的評価値は完全な相関関係にあり、心理的評価値は物理的環境要因で説明可能であると判断することになる。

第1の分析者と第2の分析者の結論は全く異なったものとなっている。一般にこのような状況では、第2の分析者は「荒い」データを用いて分析したために誤った結論に至ったのであり、より「細かい」精度のデータで分析する方が真実に近いと判断されやすい。しかし、そうした判断に根拠があるわけではない。その根拠がないことを数値例で示すことができる。実は、表1のデータ自体も、より細かな地域をまとめて6つの区域としたにすぎず、より細かな12小地域のデータから得られたと考えることができ、事実、表3に示すデータをもとに、

$x(i)=(s(2i-1)+s(2i))/2$、$i=1\sim6$　(5-1)

$y(i)=(t(2i-1)+t(2i))/2$、$i=1\sim6$　(5-2)

として得られたものが表1のデータなのである。従って、より細かいデータの方が良いとする考え方にそうならば、表3のデータより得られる結果が正しいということになるが、この場合のデータをプロットしたものが図3であり、変量$s(i)$，$i=1\sim12$と$t(i)$，$i=1\sim12$の相関係数は、

$C(s,t)=0.97$　(6)

となり、表3のデータを分析した分析者の結論は、物理的環境要因と心理的評価値は極めて高い相関関係にあり、心理的評価値は物理的環境要因で説明可能であると判断することになる。このことは、表1のデータの分析結果を否定することになる。従って、より細かい分析をしたから第1の分析の方が荒い分析をした第2の分析より真実に近いということは、表1よりも細かい表3に基づいて分析をした結果が表2よりも真実に近いはずの表1の分析結果を覆してしまうという点で矛盾していることになる。つまり、この場合、データが細かいということだけで真実に近いと判断することは、なんら根拠を持っていないと言わざるを得ない。ちなみに、表1のデータが得られている段階

表2　数値例（3地域分割）

地域	物理的要因U	心理的評価値V
1	-0.750	-0.433
2	0.000	0.000
3	0.750	0.433

表3　数値例（12地域分割）

地域	物理的要因S	心理的評価値T
1	-4.464	-2.000
2	2.464	2.000
3	-3.964	-2.866
4	2.964	1.134
5	-9.160	-4.134
6	8.160	5.866
7	-8.160	-5.866
8	9.160	4.134
9	-2.964	-1.134
10	3.964	2.866
11	-2.464	-2.000
12	4.464	2.000

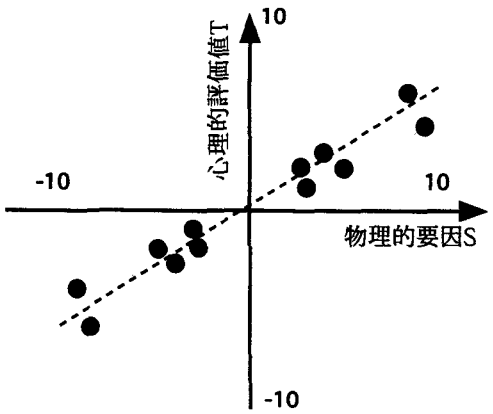


図3　表3データの散布図

では、より細かな集計をしなおしたとき、表4のようなデータになっているときもありうる。事実、表4のデータから2つの小地域をまとめることで表1が得られる。つまり、

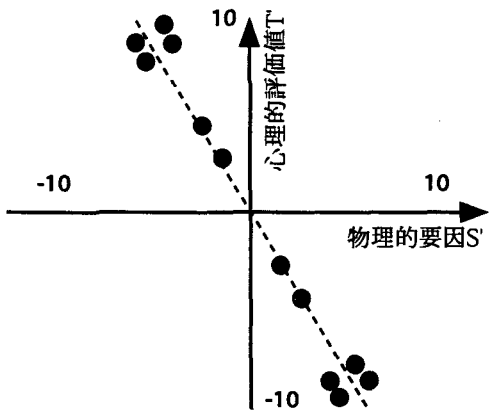


図4　表4データの散布図

$$x(i) = (s'(2i-1) + s'(2i))/2,\ i=1\sim6 \quad (7\text{-}1)$$

$$y(i) = (t'(2i-1) + t'(2i))/2,\ i=1\sim6 \quad (7\text{-}2)$$

となっており、変量s'(i)，i=1～12とt'(i)，i=1～12をプロットしたものが図4で相関係数は、

$$C(s',t') = -0.98 \quad (8)$$

となっており、この場合、物理的環境要因と心理的評価値は極めて高い逆相関関係となっている。いずれにしても、集計単位の違いによる分析結果の違いは、単に、集計単位の細かさだけで判断できるものではない。

表4　数値例（12地域分割）

地域	物理的要因S'	心理的評価値T'
1	-6.000	8.660
2	4.000	-8.660
3	-5.500	7.794
4	4.500	-9.526
5	-2.500	4.330
6	1.500	-2.598
7	-1.500	2.598
8	2.500	-4.330
9	-4.500	9.526
10	5.500	-7.794
11	-4.000	8.660
12	6.000	-8.660

3．空間情報の損失

表1のデータの分析について、問題点を調べることから検討を始めたい。確かに、6つの地域が地区1から地区6に線状に連結していたはずであるが、ここで、この地区の並びをランダムに取り替えてみると、表5のデータが得られる。これは、表1のデータと比べ、各地区の場所だけが入れ代わったために、空間的な地区の配置の情報は、表1と全く異なってしまっている。しかし、表5に基づいて、物理的環境要因と心理的評価値をプロットしてみると、図1と全く同じになり、相関係数も0となって同じになる。このことは、第1の分析者の分析手続きでは、地区の空間配置情報を全く利用していないということを示している。

この点をより正確に述べれば、6つの地区のそれぞれについて、各地区内の物理的環境要因は、

表5　数値例（表1の地域の並びかえ）

地域	物理的要因X	心理的評価値Y
1	0.500	0.866
2	1.000	0.000
3	-1.000	0.000
4	0.500	-0.866
5	-0.500	0.866
6	-0.500	-0.866

その地区の心理的評価値にしか影響しないという仮定が暗黙のうちになされていたのが、前述の分析の基本的な特徴なのである。また、3ゾーンにまとめた分析でも、各ゾーンの物理的環境要因は各ゾーンの心理的評価値にしか影響しないという仮定が暗黙のうちになされていたことになる。

つまり、物理的環境要因が影響する影響範囲について表1と表2の分析では異なる仮定が暗黙のうちになされて分析されたことにより、異なった結果となったと言える。

以上の検討から、前述の分析結果の混乱を回避するためには、基本的に、各地区の位置情報を損失しないようなモデル設定と、説明変量が被説明変量へ影響を与える範囲を明示したモデル設定になっている必要があることがわかる。

4．空間相関関数と影響範囲

影響の範囲を事前に特定することは難しい。しかし、次のような相関係数を拡張した概念を導入することで、ある変量Xとある変量Yの相対位置情報を含めた相関関係を把握することができる。すなわち、

$$C_{XY}(\tau)=\frac{\Sigma(x(i-\tau)-\bar{x})(y(i)-\bar{y})}{\sqrt{\Sigma(x(i)-\bar{x})^2\Sigma(y(i)-\bar{y})^2}} \qquad (9)$$

これは、変数Yとτだけずれた場所の変数Xの相関係数となっており、空間的なずれτの関数となっているので、空間相関関数と呼ばれる（青木(1986a)参照）。もちろん、空間を2次元とした場合、τは2次元ベクトルになる。実際の都市データについて、この空間相関関数を求めてみると図5のようになる。同じ変量でも、近接している（小さいτ）ときは相関があっても、次第に相関関係は減少してゆく傾向がある。場合によっては、遠く離れた場所で相関関係が高くなることもある（青木(1986b)参照）。

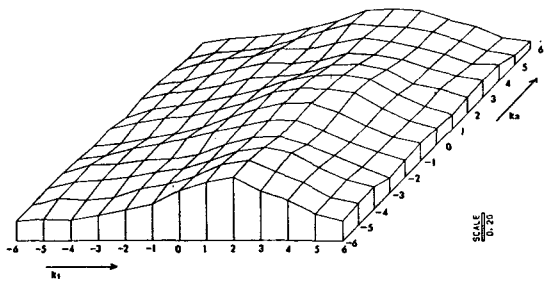
こうした関係を前提とすれば、同じ場所での相関関係を前提とするのではなく、近接した場所からの影響関係を仮定することの方が妥当である。そこで、空間が連続的に広がっているものとし、距離τだけずれた場所からの影響の度合いをτの関数として考える。先の物理的環境要因Xと心理的評価値Yの関係は、

$$y(s)=\int h(\tau)x(s-\tau)\,d\,\tau \qquad (10)$$

と記述することができる。空間が表1のように離散的に表されている場合は、

$$y(i)=\sum_k h(k)x(i-k) \qquad (11)$$

と表される。上記の関係は、畳み込み積分と呼ばれるものとなっている。ここでは、上記のモデルを空間影響関数モデルと呼ぶ（青木(1987)参照）。



X：商業業務用地面積
Y：道路面積
τ：（Ra_1、Ra_2）

図5　空間相関関数の例

5．空間影響関数モデルからみた集計単位問題

集計単位による混乱の原因を空間影響関数モデルをもとに考察することにしよう。

ある集計単位のもとで相関分析や回帰分析を行うということは、上記の離散型空間影響関数モデルで、

$$y(i)=h(0)x(i-0)=hx(i) \quad (12)$$

という隣接した場所からの影響が全くないとしたモデルになっている。すなわち、同一の集計単位の中だけしか両変量は関係しないというモデル設定をしていることになる。この仮定のもとで、集計単位を例えば2倍に変化させた場合には、影響しあう空間範囲が2倍であると仮定することを意味してしまい、結局、こうしたモデル設定のもとでは、集計単位を変化させるたびに、空間的な影響範囲の仮定を変えてしまっていることを意味する。すなわち、モデル自体が変化しているのである。こうした状況では、前述のように分析結果が異なるのは当然と言えよう。

以上の議論を整理すると、集計単位を考慮しない相関分析、回帰分析では、

①空間的配置情報を失っている。

②集計単位を変えるたびに異なるモデルを仮定している。

ということになり、これが、集計単位問題の本質的な点であると言える。

6．集計単位のフィルター効果

上記の検討から集計単位の中だけで影響しあうことを想定したモデルでは、地域の空間的配置情報を失い、集計単位を変えるたびに異なる仮定をおいたモデルとなり比較可能でなくなってしまうことがわかるが、空間影響関数モデルのように設定した場合には、集計単位の効果が完全に除去できるのだろうか。この点を考察するために、空間のスペクトル概念を導入する。

通常、集計は、ある一定の範囲ごとに集計されたり、あるいは平均化される。そこで、集計単位の範囲（1次元の空間を考えているので、区間となる）の長さを2Tとし、各変量Xを各集計単位で平均化していると考えると、集計された量は、次のようになる。

$$x_T(s)=\int_{-T}^{T} x(s-t)\,dt/2T \quad (13)$$

具体的には、上記のように集計されるが、必ずしも集計単位内を一様と考える必要もないので、一般的には、この集計操作自体は、

$$x_T(s)=\int_{-\infty}^{\infty} f(t)x(s-t)\,dt \quad (14)$$

と表現できる。

このとき、各変量$x(s)$,$x_T(s)$,$f(s)$のフーリェ変換をそれぞれ$X(w)$,$X_T(w)$,$F(w)$とすると、上記の関係式は、次のスペクトル空間での関係に変換される。

$$X_T(w)=F(w)X(w) \quad (15)$$

この関係式は、スペクトル空間で一定の変換$F(w)$を施すことを意味している。すなわち、もとの変量Xに対して、そのスペクトルを歪めていることを意味している。また、集計操作が、その地域の近傍での集計であることは、一般的に変換$F(w)$はスペクトルの高い周波数成分を減少させる操作に対応していることになる。すなわち、変換$F(w)$は高周波数成分をカットする一種のフィルターであることになる。この事実より、集計データを用いた場合、当然のことながら、空間的に細かに変動する影響関係がはっきりと推定できないことを意味する（青木他(1986)参照）。

以上の結果から、集計単位をできるたけ細かくとることで空間的に細かく変動する情報を抽出できることがわかる。すなわち、集計単位問題について、集計単位を細かくすることは本質的な解決とはなっていないものの、細かな空間変動の情報を失わないという点で集計単位を細かくすることには一定の意味があるのである。

7．まとめ

集計単位のとり方によって分析結果が異なってきてしまうという集計単位問題が指摘されているにもかかわらず、この問題に配慮しないまま分析がなされている現状が続いている。本論は、こうした問題についてのひとつの考察である。

最初に数値例を用いて集計単位問題の特質を見、その結果として、一般に言われているように集計単位を細かくすることによってはこの問題は本質的に解決しないこと、従来の相関分析や回帰分析では、集計単位を変化させることで異なる仮

定のモデル設定となってしまっていること、空間情報を失ってしまう分析となっていることを明らかにした。その上で、空間での影響範囲を明確にするため、空間相関関数が利用できること、さらに、空間影響関数モデルにより、回帰分析と同様の分析が可能となることを示した。空間影響関数モデルの立場から再び従来の回帰分析を捉えると、回帰分析がまったく空間的情報を無視したモデル設定になっていることをあらためて示した。最後に、集計単位の大きさがどのような意味を持つかを、フーリェスペクトル空間で考察すると、集計操作が高周波数成分をカットする一種のフィルターとなっていることから、集計単位を大きくすると空間的変数の細かな変動の情報を失うことを示し、集計単位を細かくすることが集計単位問題を本質的に解決することはないが細かな空間変動を捉えるという意味では有効であることを示した。

上記の考察から、都市データを分析する場合には、空間的な影響の範囲に注意をはらった分析が必要であり、空間影響関数モデルはそうした方法のひとつであると言える。

また、本論で考察された集計単位問題は、集計単位問題のひとつの側面であり、今後、この問題に対する体系だった研究が必要であるといえよう。

参考文献

Openshaw,S., "Ecological fallacies and the analysis of areal cencus data", *Environment and Planning*, 16, pp.17-31, 1984.

Arbia, G., *Spatial data configuration in statistical analysis of reagional economic and related problems*, Kluwer Academic Publishers, 1989.

田頭直人「空間集計問題－データを空間的に集計することによるモデル推定への影響－」,『都市計画論文集』25, p.361-366, 1990.

Okabe,A. and Tagashira,N., "Spatial aggregation bias in a regression model containing a distance variable", *Geographical Systems*, 2, pp.83-101, 1996.

青木義次「メッシュデータ解析の一方法としての空間相関分析法の提案　その1．メッシュデータ解析の問題点と空間相関分析法の理論」,『日本建築学会計画系論文報告集』364, p.94-101, 1986a.

青木義次「メッシュデータ解析の一方法としての空間相関分析法の提案　その2．土地利用の連担性・共存性・排斥性の計量化への応用」,『日本建築学会計画系論文報告集』368, p.119-125, 1986b.

青木義次「メッシュデータ解析の一方法としての空間相関分析法の提案　その3．空間影響関数モデルの有効性と問題点」,『日本建築学会計画系論文報告集』377, p.29-35, 1987.

青木義次他「都市メッシュデータ解析におけるメッシュサイズの効果」,『都市計画論文集』21, p.247-252, 1986.

Key Words（キー・ワード）

Ecological Fallacy（生態学的誤謬）, Spatial Aggregation Problem（空間集計問題）, Modifiable Areal Unit Problem（集計単位問題）, Convolution Integral（畳込み積分）, Fourier Spectrum（フーリェスペクトル）

A Note on Spatial Aggregation Problem

Yoshitsugu Aoki*

*Department of Architecture, Tokyo Institute of Technology
Comprehensive Urban Studies, No.65, 1998, pp.17-24

This paper discusses a problem when we meet in analyses of relationships between apatial distributed variables by use of aggregated data. More exactly, the calculated value of correlation coefficient between two spatial distributed variables changes if we change the size of aggregating units. This type of problem was pointed as "Ecological Fallacy" and discussed as "Spatial Aggregation Problem" or "Modifiable Areal Unit Problem".

At first, we show an artificial numeric example of two spatial distributed variables and try ordinal regression analyses in different size of aggregation units. The results show the fact that we can not conclude the smaller size of aggregation units produces the more true relationship.

The second, we show the fact the results do not change when we change the locations of units. This means one of fundamental cause of the fallacy is loss of the information of spatial distribution. Then we propose a model using a convolution integral by spatial parameters instead of ordinal regression models.

Finally, we show that aggregation procedure can be formulated as a kind of spatial filter in the Fourier spectrum. This means we lose the information of microscopic variation of spatial distribution if we use large size aggregation units.