標本に基づく空間データの利用が空間解析に及ぼす影響についての考察
- 空間的自己相関の解析を題材に -
Potential influence that use of sample data has on spatial analysis
- A case study with spatial autocorrelation analysis -

山田 育穂*
Ikuho Yamada*

This study investigated potential influence that use of spatial data that were obtained through a sample survey would have on spatial analysis based on statistical simulations. In the simulations, spatial patterns with clusters of high attribute values were stochastically generated and analyzed using the population data as well as sample data so as to assess if and how results of the analyses with the two types of data would differ from each other. The spatial analysis method used here was the local Moran statistic. Results suggested that occurrence of type I errors was not much affected by the difference of the data types. On the other hand, the statistical power of cluster detection was found to decrease as the sampling rate of the sample data decreased. In addition, this tendency was more critical near the border of the simulated clusters, implying that shapes of the clusters were difficult to accurately determine with the sample data.

Keywords: sample data, spatial analysis, spatial autocorrelation, local Moran statistic
標本に基づくデータ，空間解析，空間的自己相関，ローカル Moran 統計量

1. はじめに

統計的社会調査は人間の行動・意識や属性に関する量的なデータを収集するための調査手法である[1)]。国勢調査をはじめとした政府統計や研究者が自ら実施するアンケート調査など代表的な統計的社会調査は、地域や地域住民の特徴を把握・分析するため都市計画に係わる諸分野でも広く用いられている。国勢調査など限られた例外を除き、多くの統計的社会調査は標本調査であるため、結果には真の値からのある程度の乖離が避けられないが、調査が対象地域全体としての特徴を捉えるためのものであるならば、その乖離は一般の統計学における標本分布に関する理論により調整・把握することが可能で、分析結果や解釈に与える影響は最小限に抑えられる。例えば、住環境に対する不満足度や都市施設へのアクセシビリティ等、アンケート調査が対象とする事象の対象地域全体における傾向を把握することが調査目的であるような状況がこれに対応する。

しかしながら、対象地域がある程度の広さを持つときには特に、全体的な傾向のみならずその空間的な変動・分布（例えば、住環境に対する不満足度の高低が対象地域内でどのように変動しているか）が興味の対象となることも多い。こうした場合には、標本（すなわち調査の回答者である住民）は対象地域を分割した空間単位（町丁目、国勢調査区など）ごとに集計され、分析ではそれぞれが真の値からの乖離を持つ複数の値を扱うことになるため、個別の標本分布に関する理論だけでは、標本に基づくデータを用いることによる分析結果への影響を考慮しきれない可能性がある。これは例えば、空間統計学の手法を用いた空間パターンの分析では、個々の空間単位の属性値を統合して対象地域全体あるいは特定の空間単位の近傍に対する統計指標を算出することから、個々の乖離が蓄積されて全体として増幅したりあるいは逆に相殺されたりという状況が生じうるためである。しかしながら、一般の統計学とは異なり空間統計学の分野では、こうした標本に基づくデータを扱うことに関する議論は非常に限られている。

こうした状況を踏まえ本研究では、空間解析において統計的社会調査など標本調査により収集された空間データを用いることが及ぼす影響を明らかにすることを目的として、統計的シミュレーションに基づく分析を行う。特に、分析対象地域が複数の空間単位に分割され、それぞれの空間単位の特徴が属性値として与えられているいわゆる面データ（ポリゴンデータ）を想定し、その属性値が母集団全体を用いて算出されている場合と標本を用いて算出されている場合とで解析の結果を詳細に比較することによって、標本の使用により生じる影響を整理・検証していく。

空間統計学における標本に係わる議論をみると、その中心は標本を得る位置についてのものである。空間統計学のテキストの中には標本や標本抽出について言及のないものも見られるが、Haining[2)]や Cressie[3)]によるテキストは、地質や降水量のように連続的に変化する空間変数の分布を推定する際の調査地点の数や配置の適切な決め方について論じている。調査コストの大きい地質調査等では調査のできる地点数が限られるため、真の分布を効率的に推定するための調査地点の選び方に関する研究が近年でも行われている[4),5)]。統計的社会調査では、対象地域を構成する町丁目などの空間単位ごとに全住民あるいは全世帯を母集団と捉えて、そこから標本を抽出することが多く、このような連続平面からの任意の調査地点の抽出に関する知見を直接適用することは難しい。前述の Haining のテキスト[2)]では、このような標本抽出の特別な場合として、対象地域を構成する全空間単位の中から調査を行う空間単位を抽出する状況が言及されているが、これは本研究で着目している空間単位ごとに調査される属性値が母集団に基づくものか標本に基づく

* 正会員 東京大学空間情報科学研究センター（The University of Tokyo）

ものかという観点とは異なる議論である。

また、点分布の集塊・分散傾向を分析する基礎的手法として広く知られている最近隣距離法を提案した論文 [6]では、最近隣距離を母集団全体ではなくそこからランダムに抽出した標本に対して算出し分析する場合について言及している。最近隣距離法は、点分布を構成する全ての点からそれに最も近い点までの距離を測定し、その分布や平均値に基づき空間分布の傾向を評価する手法である。一般に分析対象とする点分布は母集団であると仮定されるが、この論文では、点分布の範囲を小さな空間単位に分割して、そこから調査対象とする空間単位をランダムに抽出し、抽出された空間単位の中にある点に対してのみ最近隣距離を算出するという方法も提案されている。これは、森林を構成する樹木の分布を分析したい場合など、母集団を網羅的に把握することが現実的でない状況を想定して提案されたものであるが、空間的に近い位置にある点同士の最近隣距離は独立ではなく、統計分析の基礎的な条件を満たしていないという問題への対処法とも捉えられる [7]。この最近隣距離法における標本抽出も、上述の Haining[2]の例と同様、標本を得る位置の抽出であり、本研究の観点とは異なるものである。

一方、統計的社会調査に係わる研究では、住所情報を地図上の位置に変換するジオコーディングの精度が空間解析に及ぼす影響を論じたもの [8),9)]が多いが、そうした研究では社会調査が標本調査であること自体の影響は論じられていない。埴淵・村中の著書 [1)]では、回収率や特定の設問に対する回答率の低下、その地域差など、調査票を配布して行われる従来型の社会調査が抱える問題点や、新たな調査手法として注目されるインターネット調査の課題や地域分析に用いる際の懸念点などについて、様々なデータや分析を示して詳細に議論されている。しかしながら、基本的には個別の地域に対する調査結果が議論の中心であり、それらの値の空間的な分布を統合的に分析した場合にさらなる問題が生じる可能性については考慮されていない。

以上のように、空間統計学の分野において標本に基づくデータを扱うこと、特に、標本に基づき空間単位ごとに算出されたデータを用いて空間解析を行うことについての検討はほとんどなされていない。一般の統計学の理論に基づいて十分な標本数を確保することで、空間単位ごとに集計された調査結果の真の値からの乖離はコントロールすることができるが、それら空間単位ごとの値を融合して扱う空間解析においては小さな乖離が蓄積し大きな乖離につながる可能性もある。社会調査によって得られる標本に基づく空間データが空間解析において広く利用されている現状を鑑みると、そうした可能性や生じうる問題を明らかにすることは重要な課題であると考えられる。そこで本研究では、標本に基づく空間データを用いた空間解析の結果が持つ特徴や問題点を、統計的なシミュレーションにより調査・分析して、空間解析における標本データの適切な利用を検討するための知見を得ることを目指す。

2. 方法

2.1. シミュレーションの概要

本研究では、正方形の地域を縦横に 10 分割した合計 100 個のセルからなる分析対象地域を想定し、その中で特定のセルの集合が他のセルより高い属性値を持ち、いわゆる「クラスター」を形成している空間パターンを確率的に発生させて、空間統計学的な手法でそのクラスターを検出するシミュレーションを行う。その際に、セルごとの属性値を母集団全体から算出した場合と、一定の抽出率で抽出した標本に基づき算出した場合とで結果を比較して、標本に基づく空間解析の特徴を明らかにしていく。具体的には、空間パターンが存在しないときに誤って検出してしまう第一種の過誤と、空間パターンが存在するときにそれを正しく検出する検出力に関して比較を行う。前述の住環境に対する不満足度の例で考えると、第一種の過誤は対象地域内における不満足度に本質的な空間的変動がないにもかかわらず、特定の地域を他に比べて不満足度が著しく高い、都市計画的対応の検討を要する地域と判断してしまう状況にあたる。検出力が低い解析は逆に、実際に不満足度が高く都市計画的対応が必要な地域を見逃してしまう状況である。統計解析の特性上、第一種の過誤の発生確率をゼロにしたり検出力を 100%にしたりすることはできないが、誤った結果の発生傾向を把握することは解析結果を適切に精査するための重要な知見となり得る。

分析の対象とするセルごとの属性値には、全人口に占める特定の特徴を持つ人の割合（例えば、住環境に不満足な人の割合や高齢化率、全人口を全建物に置き換えた場合の木造率など）を用いる。平均年齢や平均延べ床面積といった、個人や個々の建物が持つ属性値をセルごとに集計した代表値を分析することも考えられるが、政府統計や一般的なアンケート調査では、そうした属性値を数値で回答してもらう代わりに、属性値が取り得る値をいくつかに分類したカテゴリーの中から選択してもらう方法が頻繁に用いられることから、今回はより汎用性が高いと考えられる前者を扱うこととした。

以下では、発生させる空間パターンと、その検出に使用する空間統計学的な手法について、詳細に説明する。なお、シミュレーションには R ver. 3.5.3（https://www.r-project.org/）を用いた。

2.2. 分析対象とする空間パターン

図 1 に本研究のシミュレーションで使用した 12 通りの空間パターンを示す。これらは、10×10 の対象地域の中心に 4×4 のクラスターを配置したパターンを基準にして、クラスターの大きさ、位置、形状を変化させて作成したものである。下段の 6 つのパターンにおけるクラスターの形状はひとつの例外を除き、基準パターンのクラスターを縦横に分割して得られるものであるので、以下ではこれらを分割のパターンと呼ぶ。

今回のシミュレーションでは、各セルの人口は 2500 人[(1)]

に固定し、セルごとの人口変動は考慮しないこととした。特定の特徴 P を持つ人の割合pとクラスター内のセルにおけるpの上昇率 q は、それぞれ 0.2、0.5 に設定した。これは例えば、住環境のある要素に対して不満足な人が通常のセルでは平均して 500 人（=2500 人×0.2）、クラスター内のセルでは 750 人（=2500 人×0.2×(1+0.5)）存在する状況であり、空間統計学的手法を用いた空間パターンの分析は、この 250 人に相当する不満足割合の上昇が生じているセルの空間的なクラスターを検出することを目的として行う。一般に、上昇率qが大きいほどクラスターの検出は容易となる。本研究では上昇率qを 1.0 とした場合のシミュレーションも実施したが、非常に強い空間パターンであるために、母集団全体を用いた分析でも標本を用いた分析でも検出力は同様に高く両者の比較に適さなかったことから、ここでは上昇率qを 0.5 とした場合の結果についてのみ報告する。

次に、シミュレーションにおける空間パターンの作成と分析対象とする属性値の算出について説明する。まず、各セルi（$i = 1, \dots, n$; nは総セル数 100）の人口 2500 人に対して確率p（クラスター内のセルの場合は確率$p(1+q)$）で、特徴 P を持つかどうかをランダムに割り当てる。そして、2500 人のうち特徴 P を持つとされた人の割合x_iを算出する。これが母集団全体を用いて算出された属性値（真の値）に対応する。続いて、各セルの人口から抽出率 rで標本を抽出し、その標本内での特徴 P を持つ人の割合x'_iを算出する。これが標本に基づいて推定された属性値にあたる。なお、抽出率 rには、0.5、0.3、0.1、0.05、0.03、0.01 の 6 種類の値を用いた。これらはセルごとの標本数で見ると、それぞれ 1250 人、750 人、250 人、125 人、75 人、25 人に対応する。抽出率およびセルごとの標本数が小さくなるほど、推定された属性値の真の値からのばらつきは大きくなるため、空間解析の結果に与える影響も大きくなると予想される。シミュレーションでは、図 1 に示した 12 通りの空間パターンに対して、このプロセスをそれぞれ 1000 回ずつ繰り返した。

2.3. 使用する空間解析の手法

空間パターンの分析手法にはローカル Moran 統計量[10]を用いる。これは各セルの周辺における空間的自己相関の状態を示す指標で、対象地域全体における空間的自己相関の状態を示すグローバルな指標である Moran の I 統計量[11]をセルごとに分割した構造を持つ。空間的自己相関は、Tobler[12]が提唱した「地理学の第一法則」、空間的に近いものほどより密接に関連しているという概念を表すもので、類似した値が空間的に近く分布している状態は正の空間的自己相関、逆に大きく異なる値が近く分布している状況は負の空間的自己相関と呼ばれる。一般的な空間データは全体としては正の自己相関を持つことが多いが、局所的には正負の自己相関が混在することもあり、ローカル Moran 統計量を用いることによって対象地域内における空間パターンの変動を見ることができる。前節で説明した他より高い属性値を持つセルが集まってクラスターを形成している空間パターンの場合、クラスター内には類似した他より高い値が集中するため正の空間的自己相関を持つことになる。

Moran の I 統計量およびローカル Moran 統計量は必ずしもクラスター検出のための手法ではないが、もっともよく知られた空間解析手法のひとつであること、正の空間的自己相関としてクラスターを検出できること、GIS ソフトウェアや R のパッケージなどで提供されており空間統計学の専門家以外にも利用しやすいことなどを考慮し、本研究ではこれを用いることとした。

ローカル Moran 統計量は次式で定義され、それぞれのセルiに対して算出される。

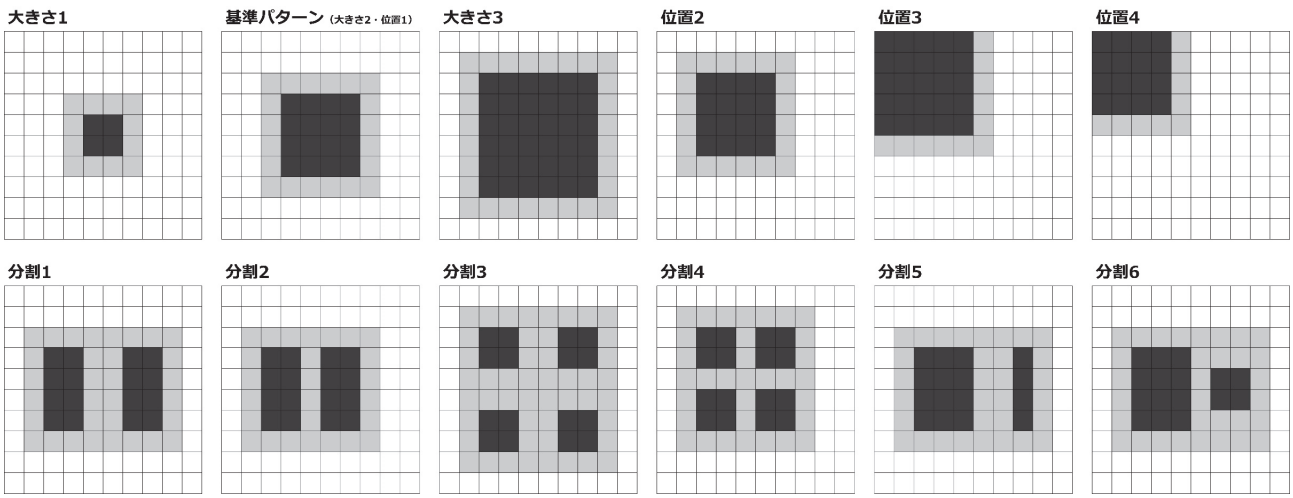

図 1： 分析対象とする 12 種類の空間パターン

※ 濃いグレーのセルがクラスターを示す。淡いグレーのセルは後述の Queen 形式の重みを用いたときにクラスターに近接すると見なされる範囲である。

$$I_i = \frac{n(x_i-\bar{x})\sum_{j=1}^{n} w_{ij}(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (1)$$

n：対象地域を構成するセルの数

x_i：セルiで観測された分析対象変数の値 $(i = 1, \cdots, n)$

$\bar{x}$： x_iの平均値

w_{ij}：セルiとjの空間的な近接性を示す重み

I_iは概ね-1から$+1$の間の値をとり、Pearson の積率相関係数と同様に、$+1$に近い値は強い正の相関、-1に近い値は強い負の相関、0付近の値は無相関と解釈される。I_iの期待値・分散については理論式が与えられており[10),13)]、観測されたI_iの値の有意性はそれらを用いて判断できる。さらに、I_iの値と、セルiと周辺セルの属性値の高低の組み合わせによって、ローカルな空間的自己相関の状態は4 つのタイプに分けられる（表1）。本研究のシミュレーションで扱う空間パターンの場合は、クラスターの内部にあるセルではHH、クラスターの外周を構成するセルではHL またはHH の空間的自己相関が検出されることが予想される。

表1： ローカルな空間的自己相関の分類

		セルiの周辺のセルjの値 x_j	
		高い（H）	低い（L）
セルiの値 x_i	高い（H）	HH（ホットスポット） 正の空間的自己相関	HL 負の空間的自己相関
	低い（L）	LH 負の空間的自己相関	LL（コールドスポット） 正の空間的自己相関

式(1)で用いられている近接性の重みw_{ij}は、一般にセルiとセルjが空間的に近いときに大きな値をとるように定義される。理論的には、検出したいクラスターの大きさなど分析対象とする空間パターンの特徴を考慮して決めることが望ましいが、実際の分析ではそうした特徴が事前に分からないことも多い。本研究ではそうした場合に分析の出発点としてよく使用されるQueen 形式の重みを用いることとした。Queen 形式の重みは、セルiとセルjが境界線か頂点を接している場合に1、それ以外の場合に0をとる構造で、通常、セルiに対する重みの総和（$\sum_{j=1}^{n} w_{ij}$）が 1 となるように隣接するセル数で除し基準化したかたちで用いられる。

3. 結果

3.1. クラスターが存在しない空間パターンの分析

統計的な仮説検定では有意水準を定めて観測された統計量の有意性を判断するため、クラスターを持たない空間パターンの分析であっても、ローカルな空間的自己相関が検出される第一種の過誤が相応数のセルで発生する。クラスターを持たない空間パターンの分析で発生する第一種の過誤は、特定の解析手法で生じる第一種の過誤の基本的傾向を表すものと考えられるため、図1 に示したクラスターを持つ空間パターンの分析の前に、クラスターを持たない、すなわち、特徴 P を持つ人の割合pが全てのセルで等しい空間パターンをローカルMoran 統計量により分析し、第一種の過誤の発生回数や空間分布の特徴をみた。

表2 に、1000 回のシミュレーションにおいて各セルでローカルな空間的自己相関が検出された回数をまとめた。クラスターのない空間パターンの分析であるので、これらの検出回数は第一種の過誤の発生回数に相当する。なお、検定の有意水準は0.05 とした。

表2： セルごとのローカルな空間的自己相関の検出回数とその分布の空間的自己相関

抽出率 r		1.0	0.50	0.30	0.10	0.05	0.03	0.01
HH	平均	16.96	15.23	15.24	15.97	16.55	15.84	17.35
	S.D.	4.05	3.98	4.06	4.46	4.14	3.54	3.79
HL	平均	17.85	17.32	17.38	17.53	17.8	17.41	17.66
	S.D.	3.84	3.94	4.04	4.11	4.06	4.51	4.21
LH	平均	16.85	16.68	17.15	16.32	16.38	16.73	15.97
	S.D.	4.12	4.08	4.25	3.84	4.58	3.62	4.38
LL	平均	14.96	15.29	15.56	15.01	14.83	14.4	13.48
	S.D.	4.54	4.12	4.02	4.30	3.59	3.50	3.84
合計	平均	66.62	64.52	65.33	64.83	65.56	64.38	64.46
	S.D.	7.82	7.17	7.18	8.03	7.86	7.60	7.18
第一種の過誤の空間分布								
	I統計量	0.089	0.134	0.117	0.046	0.074	0.187	0.012
	p値	0.065	0.007	0.016	0.291	0.118	0.000	0.682

※ 1000 回のシミュレーションのうち、各セルでローカルな空間的自己相関が検出された回数の平均値と標準偏差（S.D.）を示す。

ローカルな空間的自己相関の4 タイプを比較すると、標本の抽出率 rによらず、正の自己相関にあたる HH と LL に比べ、負の自己相関にあたる HL と LH の方がやや多く検出されている。また、4 つのタイプの合計検出回数はおよそ 65 回であり、抽出率 rによる違いは特にみられない。このシミュレーションでは、有意水準 0.05 で 1000 回の繰り返しを行っているので、各セルで空間的自己相関が検出される回数は二項分布 Bi(1000, 0.05)に従うと仮定でき、その 5%信頼区間は[36.49, 63.51]となる。今回のシミュレーションで観測された検出回数はいずれもこの上限を超えており、ローカルMoran 統計量による分析では第一種の過誤が発生しやすいことが示唆されている。この結果は、既存研究[10), 14), 15), 16)]で指摘されている、前述の理論式に基づくMoran 統計量の推定値の精度が分布の裾野部分で低下することによるものと考えられる。なお、空間的自己相関の 4 タイプごとに検出回数についても、抽出率rと第一種の過誤の発生回数の間に規則的な関係性は特にみられなかった。

さらに、表2 には、セルごとの検出回数の空間分布に特定の傾向があるかを調べるため、グローバルなMoran の I 統計量で分析した結果も記載してある。I 統計量の値はいずれも極弱い正の自己相関を示しているが、7 つの検定を同時に行っていることを考慮して Bonferroni 修正した有意水準0.007（=0.05/7）と比較すると、有意な結果は限定的で

あり、第一種の過誤の発生には明確な空間パターンはないと考えることができる。

これらの検証から、クラスターが存在しないところでそれを検出してしまう第一種の過誤に関しては、分析対象となるデータが母集団全体に基づくものか標本に基づくものかによる違いは小さいと言える。また、各セルにおける第一種の過誤の発生しやすさに空間的な傾向はほとんど検出されず、セルの位置による違い、例えば対象地域の端にあるセルでは第一種の過誤が起こりやすいといった傾向はないことが分かる。一方で、分析対象データの種類、セルの位置によらず、第一種の過誤の発生回数が理論値より大きくなることには注意が必要である。

3.2. セルごとに観測された属性値の特徴

次に、空間解析の対象となるセルごとに観測された属性値x_i（すなわちセルごとの特徴Pを持つ人の割合）の統計的な特徴についてみる。表 3 は、1000 回のシミュレーションで得られたx_iのセルごとの平均と標準偏差を、セルがクラスターの内か外かで分類して対象地域全体で集計した結果を示したものである。セルごとのx_iの値はクラスターの大きさや位置にはよらないため、ここでは基準パターンの結果を例として取り上げる。

2.2 節で述べたように、特徴Pを持つ人の割合pとクラスター内のセルにおけるpの上昇率qは、それぞれ 0.2、0.5 に設定されているため、クラスター外のセルでのx_iの期待値は 0.2、クラスター内のセルでは 0.3 となる。表 3 の結果をみると、x_iの値は平均的にはその期待値と一致していることが分かり、x_iを算出する際の標本の抽出率 rの影響はみられない。一方で、標準偏差に関しては抽出率 rの影響が顕著であり、抽出率 r=0.01 のときの標準偏差は 0.08～0.09 と、母集団全体を用いたとき（抽出率 r=1.0 に対応）の標準偏差の 10 倍にもなっている。この標準偏差の値は、本研究のシミュレーションで検出しようとしているクラスター内外の特徴Pを持つ人の割合の差に匹敵する大きさであり、標本抽出により発生するランダムな変動がかなり大きいことを示唆している。

さらに、空間統計学の手法を用いた空間解析では、式(1)の例が示すように個々のセルの観測値を統合して統計指標を算出することが一般的であるため、各セルを単独で分析する場合に比べ、こうした標準偏差の増大がより強く影響する可能性がある。特にローカルな解析では、各セルの近傍の少数のセルの値のみで統計量を算出することが多く、全てのセルの値を用いて対象地域全体の傾向を捉えるグローバルな解析に比べ、ランダムな変動同士が相殺されて全体としての影響が抑えられるといった効果はあまり期待できない。次節では、ローカルな解析の結果について詳細に検討する。

3.3. ローカル Moran 統計量による解析

表 4 に、ローカル Moran 統計量による空間的自己相関の検定結果（有意水準は両側 0.05）を示す。ここでは、クラスター内のセルで HH あるいは HL の空間的自己相関が検出されることを「(理論的に）正しい検出」、クラスター周辺のセル（図 1 の淡いグレーで示されたセル）で LH あるいは HH が検出されることを、そのセルがクラスター内にないという意味で望ましくはないが理論的には起こりうる状況と捉えて「許容できる検出」、その他のセルでタイプによらず空間的自己相関が検出されることを「誤った検出」と分類して、1000 回のシミュレーションで各セルにおいて当該の検出が発生した平均回数をまとめている。例えば、大きさ 1 の空間パターンの場合、抽出率 rが 0.3 以上であれば、クラスター内のセルでは 1000 回中 1000 回、つまり 100%正しく HH または HL の空間的自己相関が検出されたことを意味している。

表 3： 基準パターンにおけるセルごとの観測値x_iの特徴

抽出率 r	1.00	0.50	0.30	0.10	0.05	0.03	0.01
クラスター外（セル数 84）							
平均	0.20002	0.20001	0.20001	0.19998	0.19991	0.19990	0.20018
S.D.	0.00801	0.01136	0.01464	0.02539	0.03591	0.04654	0.08196
クラスター内（セル数 16）							
平均	0.29995	0.29996	0.29999	0.29994	0.30039	0.30079	0.29993
S.D.	0.00925	0.01304	0.01666	0.02928	0.04115	0.05324	0.09367

最初に表 4-1 の「正しい検出」をみると、抽出率 rが 0.3 以上の場合には、大きさ 3 と分割 5 を除く全ての空間パターンにおいて、概ね 95%以上の高い検出力となっていることが分かる。大きさ 3 と分割 5 の空間パターンで検出力が低い原因としては、前者ではクラスター内のセル、すなわち特徴Pを持つ人の割合が高いセルの数が他のパターンより多く、各セルの観測値x_iと比較する平均値$\bar{x}$の値が相対的に大きいこと、後者では右側の縦に 4 セルが並んだクラスターは Queen 形式の近接性を用いた場合には正の空間的自己相関とは捉えられないことが、それぞれ考えられる。

抽出率 rが 0.1 以下になると検出力は徐々に低下し、抽出率 r=0.03 では半分程度の空間パターンで、抽出率 r=0.01 では全ての空間パターンで 50%を下回り、クラスターか否かをランダムに割り振った場合よりも低い検出精度となっている。また、抽出率 rの減少につれた検出力の低下の度合いは空間パターンによって異なり、特に分割のパターンで顕著な低下がみられた。これは、クラスターが小さく分割されていると、Queen 形式の近接性の定義で各セルの近傍と見なされる範囲に、クラスター外のセルが含まれる状況が多く発生するためと推察される。表 4-1 の結果から、分割以外の空間パターンでは、80%程度の検出力を確保するために必要な標本抽出率 rは 0.1～0.05 程度と推定できる。

次に、表 4-2 をみると、クラスター周辺のセルにおける「許容できる検出」の発生回数は、抽出率 rの減少につれて増加するものの概ね有意水準と同程度に収まっている。表 4-3 に示したその他のセルにおける「誤った検出」も同様である。これは表 2 に示したクラスターが存在しない場

合の第一種の過誤の発生回数と比較すると、同程度かやや少ないという結果である。これらのことから、第一種の過誤という観点では、空間解析で標本に基づくデータを用いることによる影響はほぼないと見なすことができる。

ただし、抽出率 r が 1 の母集団全体を用いた解析であっても、「誤った検出」はほとんど発生しないものの、「許容できる検出」はいくつかの空間パターンで一定回数発生している。これはローカル Moran 統計量を用いたクラスターの検出では、その形状が多少曖昧になる可能性があることを示唆する結果である。

なお、大きさ 1 の空間パターンでは例外的に、クラスター周辺のセルにおける「許容できる検出」が多く生じている。この原因としては、前述の大きさ 3 の空間パターンについての議論とは逆に、クラスター内のセルの数が少なく平均値 $\bar{x}$ の値が相対的に小さいことが考えられる。しかしながら、クラスターが小さいことは、周辺セルの Queen 形式の近傍に含まれるクラスター内のセルが少ないことにつながり、LH や HH の空間的自己相関は生じにくいとも言える。この点についてはさらなる検証が必要と考えている。

最後に、セルごとの空間的自己相関の検出状況を、基準パターンを例に検証する。図 2 に、抽出率 r が 1、0.05、0.01 の場合について[2]、1000 回のシミュレーションのうち、セルごとに有意な空間的自己相関が検出された回数をまとめた。

抽出率 r=1 では、クラスター内のセルではほぼ 100%の確率で「正しい検出」がなされており、周辺セルにおける「許容できる検出」は合計 4 回、その他のセルにおける「誤った検出」は 1 回のみである。クラスターの検出力はほぼ 100%で、第一種の過誤もほとんどない、非常に精度の高い解析結果となった。

抽出率 r が 0.05 まで下がると、クラスター内のセルのうち中心の 4 セルでは 90%以上の確率で「正しい検出」がされたものの、残りの 12 セルの特に角に位置するセルではその確率は 50%以下に低下する。「許容できる検出」・「誤った検出」も増加し、特にクラスター周辺では、有意水準 0.05 で予想される第一種の過誤の回数 50 回を越えているセルも発生している。さらに抽出率 r が低下し 0.01 になると、クラスター内の中心のセルであっても「正しい検出」の確率は 50%程度に低下する。一方で、「許容できる検出」・「誤った検出」は抽出率 r=0.05 の場合と比較して大きくは変化していない。

以上の結果から、標本の抽出率 r が低いデータを用いて解析した場合、クラスターの中心部分しか捉えることができず、クラスターの発生状況を低く認識してしまう可能性が明らかとなった。一方で、クラスター周辺のセルでは有意水準 0.05 で予想される第一種の過誤より頻繁に「許容できる検出」・「誤った検出」が発生しており、クラスターの形状を正確に把握することが難しいことも示唆された。

表 4： ローカル Moran 統計量による空間的自己相関の検定結果

(1) クラスター内のセルの「正しい検出」の平均発生回数

抽出率 r	1.00	0.50	0.30	0.10	0.05	0.03	0.01
大きさ 1	1000.00	1000.00	1000.00	996.00	894.50	711.75	294.00
基準	999.94	997.31	986.56	899.13	774.63	637.44	321.94
大きさ 3	742.03	697.11	657.67	539.06	435.14	353.56	196.75
位置 2	999.88	997.06	986.63	901.25	770.81	634.69	333.38
位置 3	999.88	996.88	985.56	901.19	770.44	630.81	328.69
位置 4	1000.00	999.19	996.13	959.25	848.69	707.56	378.13
分割 1	999.81	994.06	971.75	820.00	655.50	506.56	239.50
分割 2	999.81	993.94	969.81	812.69	647.63	501.25	233.81
分割 3	999.69	985.38	937.00	666.69	454.06	323.94	157.63
分割 4	999.88	988.38	940.63	666.56	447.25	322.38	161.31
分割 5	783.69	783.25	774.06	696.63	585.44	474.88	242.56
分割 6	999.81	994.38	968.25	823.75	664.44	530.94	268.44

※ 「基準」は基準パターン（大きさ 2・位置 1）を指す。

(2) クラスター周辺のセルの「許容できる検出」の平均発生回数

抽出率 r	1.00	0.50	0.30	0.10	0.05	0.03	0.01
大きさ 1	93.67	141.08	159.33	158.75	138.92	112.42	69.67
基準	0.30	4.90	12.75	30.20	40.65	45.00	46.40
大きさ 3	0.00	0.00	0.00	0.61	3.00	5.57	13.96
位置 2	0.40	4.15	10.50	29.05	40.65	43.15	46.15
位置 3	17.10	31.30	43.40	64.45	73.30	74.65	64.60
位置 4	1.00	5.56	15.78	40.22	48.33	57.33	55.44
分割 1	0.28	2.81	7.13	21.06	31.03	35.91	41.75
分割 2	43.50	47.77	52.08	65.96	70.73	72.50	62.27
分割 3	0.00	0.02	0.21	5.40	13.69	20.63	30.35
分割 4	20.52	35.64	44.55	57.48	65.18	66.30	57.91
分割 5	0.28	3.09	7.81	23.94	33.34	37.59	44.69
分割 6	0.32	2.00	5.45	16.08	24.16	28.87	34.87

(3) その他のセルの「誤った検出」の平均発生回数

抽出率 r	1.00	0.50	0.30	0.10	0.05	0.03	0.01
大きさ 1	0.07	0.93	3.71	20.76	31.83	40.22	51.55
基準	0.01	0.34	2.03	12.60	22.04	28.78	40.54
大きさ 3	4.25	15.09	24.82	40.47	42.59	42.28	43.83
位置 2	0.01	0.42	2.13	13.50	21.22	28.58	40.30
位置 3	0.00	0.53	2.39	14.68	24.51	30.14	39.75
位置 4	0.00	0.62	3.06	17.02	28.82	36.03	47.20
分割 1	0.00	0.20	1.39	9.72	17.52	22.64	33.22
分割 2	0.01	0.28	2.05	12.24	19.81	25.79	35.41
分割 3	0.00	0.03	0.47	5.16	11.20	16.22	26.50
分割 4	0.00	0.26	1.35	9.12	17.29	22.28	30.98
分割 5	0.00	0.30	1.44	9.83	17.77	24.32	34.36
分割 6	0.01	0.35	1.69	9.82	19.06	24.52	35.08

4. おわりに

4.1. 本研究のまとめと考察

本研究では、空間解析において母集団ではなく標本に基づくデータを用いることが及ぼす影響、特に問題点を明らかにすることを目的として、分析対象地域内に他より高い値を持つクラスターが存在する多様な空間パターンを想定して、そのクラスターをローカル Moran 統計量により検出することを試みる統計的シミュレーションを行った。標本の抽出率 r は母集団全体に対応する 1 から 0.01 まで規則的に変化させ、その結果を比較することによって、標本に基

づくデータを用いることが及ぼす影響を、特に第一種の過誤とクラスターの検出力に着目して検証した。

まず、クラスターの存在しないところでそれを検出してしまう第一種の過誤に関する結果をまとめる。対象地域内にクラスターが全く存在しない空間パターンの場合、各セルで有意な空間的自己相関が検出される確率は、標本の抽出率 rによらず有意水準 0.05 で 6%程度となった。この 6%という値は有意水準に対応する 5%よりやや大きいが、これは 3.1 節でも述べたように、ローカル Moran 統計量の理論分布を用いた検定の問題点として既存研究でも指摘されていることである。一方で、クラスターの存在する空間パターンを分析した場合には、抽出率 rが大きいうちは第一種の過誤はほとんど発生しないが、抽出率 rが低下するにつれて徐々に増加し、今回のシミュレーションで用いた最小の値 $r = 0.01$では概ね 5%程度となった。また、いずれの場合にも、後者のクラスター周辺を除き、第一種の過誤の発生状況に空間的な分布傾向はみられなかった。以上のことから、第一種の過誤に関しては、標本に基づくデータを用いることによる解析結果への影響はほぼないものと結論づけることができる。

次に、実際にクラスターが存在する場合にそれを検出できる確率を示す検出力に関する結果をまとめる。今回検出を試みたクラスターは母集団全体を用いた解析ではほぼ 100%検出できる強いパターンであり、抽出率 rが 0.3 以上の場合の検出力は概ね 95%以上と高い水準となった。しかしながら、抽出率 rが小さくなると検出力は徐々に低下し、抽出率 r=0.03 では今回用いた 12 の空間パターンのほぼ半数で、抽出率 r=0.01 では全ての空間パターンで、50%を下回る結果となった。検出力 50%以下とは、ランダムにクラスターの有無を割り当てるよりも解析の精度が低いということであり、大きな問題である。なお、この検出力の低下の度合いは空間パターンにより異なるが、特にクラスターのサイズが小さい場合に顕著であった。シミュレーションの結果から、小さいクラスターの場合を除くと、検出力 80%程度を達成するために必要な標本抽出率 rは概ね 0.1～0.05 と推定できる。ただし、検出力はシミュレーションで用いたクラスターの強度とも深く関連しているため、この結果は絶対的な数値としてではなく相対的な傾向として理解すべきである。

さらに、空間的自己相関の検出状況の空間的な分布傾向をみると、抽出率 rが低い場合、特にクラスターの辺縁部で検出力が低下することが分かった。また、クラスターの周辺部では、正しくはないが理論的に起こり得る「許容できる検出」と「誤った検出」が混在し、第一種の過誤がやや増加する傾向もみられた。これらの傾向は、標本に基づく空間データにローカル Moran 統計量を適用してクラスターの検出を試みた場合、その形状を正確に捉えることは難しいことを示している。従って、空間解析の結果に基づいて都市計画的な対応等を検討する際には、検出されたクラスターのみでなくその周辺も含めて丁寧な検証を行うこと

抽出率 $r=1$　「正しい検出」・「許容できる検出」

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	999	1000	1000	1000	0	0	0
0	0	1	1000	1000	1000	1000	0	0	0
0	0	2	1000	1000	1000	1000	1	0	0
0	0	0	1000	1000	1000	1000	0	0	0
0	0	0	0	2	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

$r=0.05$　「正しい検出」・「許容できる検出」

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	2	14	83	85	13	3	0	0
0	0	21	431	857	846	457	15	0	0
0	0	74	854	944	932	858	78	0	0
0	0	88	857	949	943	837	80	0	0
0	0	22	451	831	870	477	21	0	0
0	0	3	22	76	95	18	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

$r=0.01$　「正しい検出」・「許容できる検出」

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	19	28	70	75	40	18	0	0
0	0	26	141	319	315	143	39	0	0
0	0	61	303	524	530	299	61	0	0
0	0	70	318	523	521	292	79	0	0
0	0	43	154	307	313	149	39	0	0
0	0	19	40	63	79	46	13	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

「誤った検出」

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

「誤った検出」

30	35	31	29	17	29	23	22	27	25
37	46	49	55	37	41	38	40	35	22
29	51	9	2	0	0	0	6	40	30
36	46	2	0	3	1	0	0	35	24
27	41	0	0	5	5	0	0	36	18
22	37	0	0	5	5	1	0	39	15
30	30	1	1	1	0	0	0	44	32
26	42	3	2	0	0	2	9	47	26
33	43	48	44	58	35	40	39	45	25
22	33	27	29	23	21	32	31	26	16

「誤った検出」

36	45	68	46	50	50	39	54	47	60
50	60	62	56	54	59	77	59	55	56
50	63	29	12	4	3	9	23	58	47
44	54	10	9	18	15	4	10	44	51
41	52	6	14	51	58	13	5	60	53
60	69	2	14	48	49	21	6	62	50
48	56	11	8	24	20	7	22	62	47
48	58	18	14	5	3	9	24	69	52
56	52	57	70	45	59	64	58	55	54
57	56	50	56	64	45	44	46	47	40

図２： 基準パターンにおけるローカルな空間的自己相関の検出状況

※　各セルの数値は 1000 回のシミュレーションのうち空間的自己相関が検出された回数を示す。

が必要と言える。可能であれば当該地域における標本数を増やした追加調査や現地調査などを行い、対象としている事象について実際に問題を抱えた地域（例えば、住環境に対する不満足度の高い町丁目の集合体）を把握した上で対応を進めることが望ましい。

4.2. 本研究の制約と今後の展望

前節で述べたように、本研究では、空間解析において標本に基づくデータを用いることにより生じうる問題を明らかにして、その対策につながるいくつかの知見も得ることができた。しかしながら、ここでは検証できなかったことがらも多く残されている。

まず、本研究では標本の大きさを抽出率 r で論じてきたが、空間単位ごとの標本数についての検討も不可欠である。各セルの人口を一定とした今回のシミュレーションとは異なり、実際の都市空間では空間単位ごとの人口は変動するため、一定の抽出率で標本を抽出しても、標本数に大きな差が生じる可能性もある。特に限られた資源で行うアンケート調査などでは標本数をどのように空間単位に割り振るかは重要な課題であり、空間単位ごとの抽出率rを一定にすべきか、標本数を一定にすべきかなどの検討は有用である。筆者がグローバルの Moran 統計量を用いて行ったシミュレーションでは、総標本数が同じ場合、後者の方がより高い検出力につながる傾向がみられており、ローカルな解析の場合についても分析を進めているところである。

また本研究では、分析対象とした変数、空間解析に用いた手法、クラスターの強度などが限られており、議論を一般化するためには、幅広い条件・手法での検証が必要である。特に、クラスターの検出にはMoran 統計量以外にも様々な手法があり、よりロバストな手法の調査も重要である。

本研究では、標本に基づくデータを用いて空間解析を行った場合、母集団の中に実際に存在する空間パターンを検出できない危険性があることを明らかにすることができた。標本調査であることの多い統計的社会調査は都市における事象を理解する上で重要な情報源のひとつであり、こうした危険性を軽減・回避するための方法論は、それを安全に利用するためにも必要である。今後はそうした方法論の構築に向けて、より包括的な分析を進めていきたい。

【謝辞】
本研究は JSPS 科研費（16H01830）の助成を受けた。研究チームの皆様、特に東京大学大学院工学系研究科都市工学専攻 浅見泰司教授、同大学院情報学環 貞広幸雄教授、東北大学大学院環境科学研究科先端環境創成学専攻 中谷友樹教授、東京大学大学院工学系研究科都市工学専攻 薄井宏行助教には、本研究を進めるにあたり多くのご助言をいただいた。記してお礼を申し上げる。

【補注】
(1) この値は、政府統計の総合窓口 e-Stat（https://www.e-stat.go.jp/）で公開されている「2015 年国勢調査」の小地域統計を用いて算出した東京都の町丁目ごとの平均人口に基づき設定した。統計的社会調査で得られる住所情報の最小単位が町丁目である場合が多いことから、この値を用いた。
(2) 紙面の制約により扱った全ての抽出率は掲載できないため、変化が分かりやすい 3 つを選んだ。

【参考文献】
1) 埴淵知哉・村中亮夫 編 (2018),『地域と統計ー＜調査困難時代＞のインターネット調査ー』. ナカニシヤ出版.
2) Haining, R. (1990), *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press.
3) Cressie, N.A.C. (1993), *Statistics for Spatial Data*, *Revised Edition*. John Wiley and Son.
4) Heim, A., L. Wehrli, W. Eugster, and M.W.I. Schmidt (2009), Effects of Sampling Design on the Probability to Detect Soil Carbon Stock Changes at the Swiss CarboEurope Site Lägeren. *Geoderma* 149, 347-354.
5) Ye, H., W. Huang, S. Huang, Y. Huang, S. Zhang, Y. Dong, and P. Chen (2017), Effects of Different Sampling Densities on Geographically Weighted Regression Kriging for Predicting Soil Organic Carbon. *Spatial Statistics* 20, 76-91.
6) Clark, P.J. and F.C. Evans (1954), Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecology* 35(4), 445-453.
7) Bailey, T.C. and A.C. Gatrell (1995). *Interactive Spatial Data Analysis*. Harlow: Longman.
8) Jacquez, G.M. (2012), A Research Agenda: Does Geocoding Positional Error Matter in Health GIS Studies? *Spatial and Spatio-temporal Epidemiology* 3, 7-16.
9) Zimmerman, D.L. and X. Fang (2012), Estimating Spatial Variation in Disease Risk from Locations Coarsened by Incomplete Geocoding. *Statistical Methodology* 9, 239-250.
10) Anselin, L. (1995), Local Indicators of Spatial Association-LISA. *Geographical Analysis* 27 (2): 93-115.
11) Moran, P.A.P. (1948), The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society Series B*, 10 (2), 243-251.
12) Tobler, W.R. (1970), A Computer Movie Simulating Urban Growth in The Detroit Region. *Economic Geography* 46, 234-40.
13) Rogerson, P.A. and I. Yamada (2009), *Statistical Detection and Surveillance of Geographic Clusters*. Boca Raton: CRC Press.
14) Cliff, A.D. and K. Ord (1971), Evaluating the Percentage Points of a Spatial Autocorrelation Coefficient. *Geographical Analysis* 3(1): 51-62.
15) Tiefelsdorf, M. and B. Boots. (1995), The Exact Distribution of Moran's *I*. *Environment and Planning A* 27: 985-999.
16) Yamada, I. and A. Okabe (2015), Examining Tail Distributions of Moran's *I* Statistic through Intensive Simulations. In Harvey, F. and L. Yee (eds.): *Advances in Spatial Data Handling and Analysis* (Select Papers from the 16th IGU Spatial Data Handling Symposium), 309-317.