104. 複数の点分布を比較・分類する手法

Comparison and Classification of Multiple Distributions of Points

貞広 幸雄*
Yukio Sadahiro*

Comparison and classification of point distributions are a useful first step of analyzing the relationship between the distributions. A new method is proposed for comparing and classifying multiple distributions of points. Point distributions are converted into surfaces to be robust against error and uncertainty in spatial data. Distance functions are defined based on the surface that consider explicitly the spatial scale of analysis. Similarity between a pair of point distributions is decomposed into elements of different spatial scales. A single measure is also proposed that represents the overall similarity between distributions, which is independent of spatial scale. The measure provides a means of classifying the distributions in the form of distance matrix. An empirical study is performed to test the validity of the proposed method, which reveals spatial patterns that are not detectable by visual analysis or other existing methods.

Keywords: point distributions, comparison, classification
点分布, 比較, 分類

1. はじめに

複数の点分布間の空間的關係は、都市解析における重要な分析対象の一つである。様々な種類の商業施設、多様な民族の混合などを点分布群として表現し、相互の關係性を決定づける背後の要因を推測、確信的分析においてその検証を行う。

複数の点分布に対する基礎的分析の一つに、分布同士の比較と分類がある。2つの点分布の比較には、区画法 (Upton and Fingleton (1985)) によって点分布を区画データ化し、 χ^2 検定を用いて一致度を測る方法が最も一般的である。あるいは、Lee (1979)の提案する最近隣空間相関指標 R^* や、 R^* を拡張した条件付き最近隣空間相関指標 (Okabe and Miki (1984)) などの距離指標、Ripley (1976)による相互 K 関数法も利用可能である。一方、点分布の分類には、2つの分布間の類似度を単一の指標で定量化し、距離行列を利用して Ward 法などのクラスター分析手法を適用するという方法がある。前述の類似度指標のうち、単一指標である χ^2 値や R^* は分類にも有効である。

点分布の比較と分類では、空間尺度が媒介として重要な役割を果たす。類似性とは、分析の空間尺度に依存する概念である。例えば、2つの点分布が広域的には類似しているが局所的な差異を含んでいる、部分的な類似性は認められるが全体的には様相が異なるなど、類似性の評価は比較に用いる空間尺度の関数と見なすことができる。従って点分布の比較には、空間尺度を明示的に取り込んだ評価が必要である。この観点から見ると、 R^* は空間尺度概念を明示的に考慮していないという問題点を有する。また χ^2 値では、区画の大きさが空間尺度のパラメータの役割を担っているが、空間尺度と分布間の類似性との関係は明確ではない。他方相互 K 関数法は、空間尺度と類似性との関係を直接表

現する手法であるが、単一指標ではないため点分布の分類に直接用いることはできない。さらに χ^2 値や相互 K 関数法は、階段状の関数に基づいて定義されるために、比較や分類の結果が位置情報の誤差によって大きく変動する。データの誤差に対する頑健性の欠如は、誤差の存在が一般的な空間情報の分析手法としては不適切な性質である。

以上を鑑み本論文では、複数の点分布を比較、分類するための新たな手法を提案する。空間尺度を明示的に取り込み、尺度に応じた類似性を評価できると同時に、分布の分類にも適用可能な枠組みを提示する。以下、2節で手法の概略を述べると共に、簡易な数値計算例を示す。3節では手法を実データに適用する。4節では結論をまとめた上で、今後の課題を論ずる。

2. 手法と適用

いま、2次元空間 R 上に存在する N 種類の点分布群を $\Gamma = \{\Gamma_1, \Gamma_2, \dots, \Gamma_N\}$ と表記する。点分布 Γ_i は n_i 個の点で構成され、 j 番目の点とその位置をそれぞれ P_{ij} , $\mathbf{z}_{ij}$ とする。

位置誤差に対する頑健性を担保するために、本論文では点分布を連続分布に変換する。具体的にはウィンドウ幅 h のカーネル関数 $\rho(\mathbf{x}, \mathbf{z}_{ij}, h)$ を利用し、点分布 Γ_i を

$$F_i(\mathbf{x}, h) = \frac{1}{n_i} \sum_j \rho(\mathbf{x}, \mathbf{z}_{ij}, h) \quad (1)$$

に置き換える。カーネル関数としては、本論文では最も一般的なガウス関数を用いる。

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

$$\rho(\mathbf{x}, \mathbf{z}_{ij}, h) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{|\mathbf{x}-\mathbf{z}_{ij}|^2}{2h^2}} \quad (2)$$

$\rho(\mathbf{x}, \mathbf{z}_{ij}, h)$ は、 $h=0$ のときには $\mathbf{z}_{ij}$ だけ平行移動したデルタ関数

$$\int_{\mathbf{x} \in R} g(\mathbf{x}) \rho(\mathbf{x}, \mathbf{z}_{ij}, 0) d\mathbf{x} = g(\mathbf{z}_{ij}) \quad (3)$$

であり、 $h \rightarrow \infty$ のときには $\mathbf{x}$ に関わらず一定値となる。ウィンドウ幅 h は、分析の空間尺度を表すパラメータの役割を果たす。後で見るように、小さな h は局所的な、大きな h は広域的な分析視点を与える。

位置 $\mathbf{x}$ における関数 $F(\mathbf{x}, h)$ の値は、 h が小さな場合には主として $\mathbf{x}$ の近傍の点分布によって定義されるが、 h が大きいときには広範な点分布全てに影響される。従って後者は前者と比べて、点分布の持つより多くの情報を取り込んでいると言える。この観点から見ると、 $F(\mathbf{x}, h_0)$ は点分布 Γ_i に含まれる様々な尺度の空間パターンのうち、 h_0 よりも小さな空間尺度で計測されたものの集合と捉えることができる。これを数式で表すと、

$$F_i(\mathbf{x}, h_0) = \int_0^{h_0} f_i(\mathbf{x}, h) dh \quad (4)$$

である。 $F(\mathbf{x}, h_0)$ を h で微分すると

$$\frac{d}{dh_0} F_i(\mathbf{x}, h_0) = f_i(\mathbf{x}, h_0) \quad (5)$$

となる。先ほどの解釈に倣えば、関数 $f(\mathbf{x}, h_0)$ は空間尺度 $h=h_0$ における点分布 Γ_i の空間パターンを表すと言える。

次に点分布 Γ_i と Γ_k を比較するために、関数 $F(\mathbf{x}, h)$ を用いて距離指標 $D_{ik}(h)$ を定義する。

$$D_{ik}(h) = \frac{1}{2} \int_{\mathbf{x} \in R} |F_i(\mathbf{x}, h) - F_k(\mathbf{x}, h)| d\mathbf{x} \\ = \frac{1}{2} \int_{\mathbf{x} \in R} \left| \frac{1}{n_i} \sum_j \rho(\mathbf{x}, \mathbf{z}_{ij}, h) - \frac{1}{n_k} \sum_j \rho(\mathbf{x}, \mathbf{z}_{kj}, h) \right| d\mathbf{x} \quad (6)$$

$h \rightarrow \infty$ のときに $\rho(\mathbf{x}, \mathbf{z}_{ij}, h)$ は全て一定値となり、 $D_{ik}(h)$ は最小値 0 をとる。一方 $h=0$ のとき、 $\rho(\mathbf{x}, \mathbf{z}_{ij}, h)$ と $\rho(\mathbf{x}, \mathbf{z}_{kj}, h)$ の差は、点分布 Γ_i と点分布 Γ_k の双方の点が位置する地点（例えば $\mathbf{z}_{ij}=\mathbf{z}_{kj}$ となる地点。この集合を R_{ik} とおく）においてのみ 0 となり、他の全ての地点ではデルタ関数となる。この場合に $D_{ik}(h)$ は以下の最大値をとる。

$$D_{ik}(0) = \frac{1}{2} \int_{\mathbf{x} \in R} \left| \frac{1}{n_i} \sum_j \rho(\mathbf{x}, \mathbf{z}_{ij}, h) - \frac{1}{n_k} \sum_j \rho(\mathbf{x}, \mathbf{z}_{kj}, h) \right| d\mathbf{x} \\ = \frac{1}{2} \int_{\mathbf{x} \in R_{ik}} \left| \frac{1}{n_i} \sum_j \rho(\mathbf{x}, \mathbf{z}_{ij}, 0) - \frac{1}{n_k} \sum_j \rho(\mathbf{x}, \mathbf{z}_{kj}, 0) \right| d\mathbf{x} \\ = \frac{1}{2} \left\{ \frac{1}{n_i} \int_{\mathbf{x} \in R_{ik}} \sum_j \rho(\mathbf{x}, \mathbf{z}_{ij}, 0) d\mathbf{x} + \frac{1}{n_k} \int_{\mathbf{x} \in R_{ik}} \sum_j \rho(\mathbf{x}, \mathbf{z}_{kj}, 0) d\mathbf{x} \right\} \\ = \frac{1}{2} \left(\frac{n_i - m_{ik}}{n_i} + \frac{n_k - m_{ik}}{n_k} \right) \\ = 1 - \frac{n_i + n_k}{2n_i n_k} m_{ik} \quad (7)$$

Γ_i と Γ_k の全ての点が異なる位置にあれば $m_{ik}=0$ であり、 $D_{ik}(h)$ の変域は 0~1 となる。

図-1 は、1次元空間における9つの点分布 $\Gamma_1 \sim \Gamma_9$ を表している。このうち、 $\Gamma_2 \sim \Gamma_9$ を Γ_1 と比較したときの距離指標 $D_{ik}(h)$ が図-2a である。 $D_{ik}(h)$ は、 h が小さいと局所的な、大きいと広域的な分布間の差異を表す。 h が増加すると微細な差異が隠されるため、 $D_{ik}(h)$ は h の単調減少関数となる。

点分布 $\Gamma_1 \sim \Gamma_6$ はいずれも広域的には均一分布であり、 $\Gamma_2 \sim \Gamma_6$ と Γ_1 の差異は局所的である。そのため図-2a では、 $D_{12} \sim D_{16}$ はいずれも h が 0 から増加してある値を超えると急速に減少して 0 に近づく。 $D_{ik}(h)$ が減少を開始するときの h は、 Γ_k の Γ_1 に対する位置のずれの程度を表している。一方点分布 $\Gamma_7 \sim \Gamma_9$ では点の局所的な集中が見られ、 Γ_1 とは広域的に見ても大きく異なる。そのため、 $D_{17} \sim D_{19}$ の減少は $D_{12} \sim D_{16}$ と比べて緩やかである。なおここで、 $D_{17} \sim D_{19}$ の減少が開始するときの h が $D_{13} \sim D_{16}$ のそれよりも小さいことに注意を要する。これは $\Gamma_7 \sim \Gamma_9$ の中に、 $\Gamma_2 \sim \Gamma_6$ の点よりも Γ_1 の点により近いものが含まれていることによる。即ち、2つの点分布が互いに位置の非常に近い点同士を有している場合、その近接性が局所的な類似性として評価され、その結果として $D_{ik}(h)$ が減少する。小さな h によって局所的に分布を比較すると、このような点分布間の部分的類似性が検出できる。

図-1 1次元空間における点分布群

図-2 図-1の点分布における(a) $D_{ik}(h)$, (b) $S_{ik}(h)$, (c) $s_{ik}(h)$.

次に距離指標 $D_{ik}(h)$ を用いて、類似度指標 $S_{ik}(h)$ を定義する。

$$S_{ik}(h) = 1 - D_{ik}(h).$$

(8)

図-2bに図-2aの各関数の $S_{ik}(h)$ を示す。定義からも明らかとおり $S_{ik}(h)$ は $D_{ik}(h)$ の鏡像であり、従って h の単調増加関数である。 h が大きくなると局所的視点よりも広域的視点が卓越するため、全ての分布が類似しているように見える。 $S_{ik}(h)$ の増加は、 Γ_1 と広域的に類似している $\Gamma_2 \sim \Gamma_6$ では急であり、 $\Gamma_7 \sim \Gamma_9$ では緩やかである。

前述の通り、 $F(\mathbf{x}, h_0)$ は h_0 よりも小さな空間尺度で計測された空間パターンの集合であり、その微分である $f(\mathbf{x}, h_0)$ は $h=h_0$ における空間パターンであると解釈できる。この議論を敷衍すると、 $S_{ik}(h)$ の h による微分 $s_{ik}(h)$ は、 $h=h_0$ における類似度を示す指標となる。

$$\begin{aligned} s_{ik}(h) &= \frac{d}{dh} S_{ik}(h) \\ &= -\frac{1}{2} \frac{d}{dh} \int_{\mathbf{x} \in R} |F_i(\mathbf{x}, h) - F_k(\mathbf{x}, h)| d\mathbf{x} \end{aligned} \quad (9)$$

関数 $s_{ik}(h)$ は、点分布 Γ_i と Γ_k の全体的な類似性のうち、空間尺度 h における類似性の占める相対的な割合を表す。図-2cは図-2bの各関数の $s_{ik}(h)$ である。広域的に類似している分布同士の場合、 h が比較的小さいうちに $S_{ik}(h)$ が1.0に近づくため、通常 $s_{12}(h)$ のように $s_{ik}(h)$ は小さな h で最大値を取る。これは、 Γ_i と Γ_k の局所的類似性が全体的類似性の中で大きな割合を占めていることによる。即ち、広域的に類似した分布同士は、広域的に異なるものと比べて、敢えて大きな空間尺度の視点を取り込まなくとも十分に類似して見えるということである。 $s_{13}(h) \sim s_{16}(h)$ では、最大値に達するときの h が $s_{12}(h)$ と比べて徐々に大きくなるが、これは $\Gamma_3 \sim \Gamma_6$ と Γ_1 の分布のずれの程度を反映している(図-1)。またこの h の差異は、全体的類似性のうちで局所的類似性の占める割合が $s_{13}(h) \sim s_{16}(h)$ では $s_{12}(h)$ と比べて相対的に小さいことを意味する。一方 $s_{17}(h) \sim s_{19}(h)$ は比較的小さい h で最大値を取っており、そのときの h は $s_{13}(h) \sim s_{16}(h)$ が最大値となる h よりも小さい。これは前述の通り、 $\Gamma_7 \sim \Gamma_9$ は局所的に見て、 $\Gamma_3 \sim \Gamma_6$ よりもさらに Γ_1 と類似している箇所があることに依る。 h がさらに増加すると $s_{17}(h) \sim s_{19}(h)$ の減少は緩やかになるが、これは $\Gamma_7 \sim \Gamma_9$ が広域的には Γ_1 と類似していないことを表している。

以上見たとおり、関数 $s_{ik}(h)$ の解釈では、絶対的な類似度ではなく相対的な類似度の割合であるという点に注意を要する。例えば $h=0.37$ のときの $s_{17}(h)$ の最大値 s_{17} は s_{13} よりも大きい。これは $h=0.37$ という空間尺度で見たときに、 Γ_7 の方が Γ_3 よりも Γ_1 に類似していることを示すわけではない。 Γ_1 と Γ_7 の類似性は $h=0.37$ という空間尺度で卓越しているが、 Γ_3 の場合はそれよりもやや広域的な空間尺度の割合が大きいと解釈すべきである。

関数 $s_{ik}(h)$ は相互 K 関数法と同様、空間尺度の概念を明示的に考慮した指標である。従って分布間の比較には有効であるが、単一指標ではないために分布の分類に直接適用するのは難しい。分類には、分布間の類似性を総合的に捉える単一の指標が必要である。そこで次に、空間尺度 h に依存しない類似度指標を提案する。

類似度指標 $S_{ik}(h)$ は、 h と共に0から1に単調に増加する。広域的にも局所的にも類似した分布同士の場合には増加が早く、広域的に類似していない場合には遅い。従って、 $S_{ik}(h)$ の h に対する増加の速度を見れば、 Γ_i と Γ_k の総合的な類似性を評価できる。そこでここでは、類似度指標 $S_{ik}(h)$ が一定値 α に達するときの $h=\eta_{ik}(\alpha)$ によって単一の類似度指標とする。 $\eta_{ik}(\alpha)$ の定義は以下のように間接的に与えられる。

$$\alpha = \int_{h=0}^{\eta_{ik}(\alpha)} s_{ik}(h) dh.$$

(10)

右辺は $h=0$ から $h=\eta_k(\alpha)$ までの $s_k(h)$ の累積を表し、従って $\eta_k(\alpha)$ は累積が α に達するときの h である。 Γ_i と Γ_k が類似していれば $s_k(h)$ の増加は速く、 $\eta_k(\alpha)$ は小さな値を取る。類似していない場合には、 $\eta_k(\alpha)$ は大きな値となる。このように $\eta_k(\alpha)$ は、 Γ_i と Γ_k の総合的な差異を表す指標であり、空間尺度 h には依存しない。

パラメータ α は分析者が任意に定めるが、3 つ以上の点分布を同時に比較する場合には、 α を一定値に保つ必要がある。また、 α が小さいと局所的類似性だけを評価することになり、総合的な類似性評価という目的には合致しない。他方 $\alpha=1.0$ では常に $\eta_k(\alpha)=\infty$ となり、分布の比較には不適切である。本論文では $\alpha=0.9$ あるいは $\alpha=0.95$ など、比較的大きな値の α を推奨する。

図-2b は $\alpha=0.9$ のときの $\eta_k(\alpha)$ である。この場合、 $\eta_k(\alpha)$ の最も小さい Γ_2 が Γ_1 に最も類似しているのに対し、 Γ_8 は最も異なっていると判断する。また、分布 $\Gamma_2\sim\Gamma_6$ は $\Gamma_7\sim\Gamma_9$ よりも Γ_1 に類似しているということになる。

$\eta_k(\alpha)$ は空間尺度に依存しない単一の指標であり、 $\eta_k(\alpha)$ に基づく距離行列を用いたクラスター分析手法によって点分布を分類することも可能である。既存の階層的分類手法としては、例えば最短距離法や Ward 法などがある。非階層的分類手法では K-means 法が最もよく用いられるが、K-means 法には距離行列だけでなく各要素の属性値も必要であるため、 $\eta_k(\alpha)$ に基づく分類には K-medoids 法などが適している (Everitt et al. (2011); Hennig et al. (2015))。

3. 実証分析

本節では、前節で提案した手法の有効性を検証するための実証分析を行う。オーストラリアのブリズベン都市圏(図-3)における go card (バスを中心とする公共交通機関用のスマートカード) の利用者、特に乗降頻度の高い利用者に焦点を当て、移動行動の特徴抽出を目指す。

分析に用いるデータは、go card を 2015 年 3 月 9~15 日の 1 週間に 1 回以上利用した 3026 人の移動履歴である。このデータには各トリップの出発と到着の時刻、利用バス停、バス系統番号が記録されており、バス停の位置データが別途、提供されている。

分析では特に、利用者の自宅と訪問地点の分布に着目する。一日の移動履歴を見ると、その大半は起点と終点が一致している。これは、自宅から出発して再び自宅に戻る移動を表しているものと考えられることから、起終点の位置を自宅の位置と見なす。複数トリップ間の節点は、トリップ間の時間間隔が 30 分未満であれば乗り換え、30 分以上であれば何らかの用務を行う訪問と見なす。そして 1 週間のうち、1 日に 4 回以上の訪問記録があるカード利用者を高頻度利用者 (以下、利用者と略す) と定義し、本論文の分析対象とする。この定義に従うと、3026 人のうち 257 人が該当する。

図-3 分析対象地域。

利用者の自宅と訪問地点の分布パターンを把握するために、各利用者の 1 週間の全訪問地点をそれぞれ点分布とみなして分類を行った。まず、各利用者対間の訪問地点の差異を $\eta_k(\alpha=0.9)$ によって定量化し、 257×257 の距離行列を作成した。次に利用者をランダムに分類し、K-medoids 法による収束計算を実施、分類が安定した時点で計算を停止した。分類数として 3~7 の 5 通りを試みたところ、分類結果が安定していたため、ここでは最も詳細な 7 分類の結果を示す。

表 1 は分類結果の概要である。 γ は各群の medoid に相当する利用者から他の利用者までの $\eta_k(\alpha)$ の平均値、 ρ は各群内利用者間平均距離であり、いずれも各群内における利用者の訪問地点のばらつきを表す。群間行列は各群の medoid 間の距離であり、群同士の間での訪問地点の差異を示す。

各群の利用者数を見ると、G01~G03 が全体の 3/4 を占めていることが分かる。このうち G02 は、利用者の数に比べて γ や ρ が小さく、利用者の訪問地点が互いに類似していることが分かる。反対に G06 では、利用者は少ないが γ や ρ は大きく、訪問地点のばらつきが大きくなっている。

群間の距離行列を見ると、G01~G04 の相互間で距離が小さく、全体として大きな 1 群を構成している。G05~G07 はいずれも各群との距離が大きく、訪問地点が他とは大きく異なっている。

表 1 各群内の平均距離と群間距離行列。γは各群の medoid から他の利用者までの平均距離、ρは各群内の利用者相互間の平均値、群間距離行列は各群の medoid 間の距離をそれぞれ表す。

利用者数	γ	ρ	G01	G02	G03	G04	G05	G06	G07
G01	37	45.59	77.76						
G02	110	10.35	19.07	250.0					
G03	48	27.31	44.36	359.4	62.5				
G04	19	30.84	52.65	484.4	187.5	80.1			
G05	14	18.67	35.90	812.5	515.6	406.3	281.3		
G06	19	46.84	87.10	523.4	828.1	937.5	1062.5	1000.0	
G07	10	6.24	12.91	968.8	679.7	578.1	453.1	1000.0	128.9

図4 は各群に含まれる利用者の自宅と各訪問地点における全利用者の延べ訪問回数である。自宅と訪問地点の間には全体的に強い近接関係が見られ、利用者は主として自宅周辺を訪れていることが分かる。G01～G04 の訪問地点はいずれも Brisbane 及びその周辺であり、表1 で示唆された群間の類似性が確認できる。一方、G05～G07 の訪問地点はいずれも Brisbane 以外の地域に広がっており、G05 は Sunshine Coast, G06 と G07 はそれぞれ Gold Coast の中部と南部などに集まっている。G05 と G06 及び G07 の間の差異が顕著であるが、これは表1にある群間距離の大きさとよく符合している。訪問地点の分布傾向を個別に見ると、G01 と G06 では広範に広がっているのに対し、G02 と G07 では非常に集中している。前者は主として通勤と仕事上の移動、後者は自宅近辺での勤務のための移動と推定される。

以上の結果より、利用者は大きく2つの群、即ち Brisbane 及びその周辺に居住する G01～G04 と、その他の地域に居住する G05～G07 に分類できる。このうち後者の特徴は図4にも比較的明瞭に見えるが、前者については、群間の差異がそれほど明確ではない。そこで次に、G01～G04 に含まれる214名の利用者を7つの群に再分類した。ここでもα=0.9としてK-medoids法を適用、新たに7つの群G11～G17を得た。

分類結果を表2及び図5に示す。表2を見ると、訪問地点のばらつきはG14とG17で小さく、反対にG13とG16で大きいことが分かる。訪問地点はここでも居住地周辺に多く、特にG12、G13、G15でこの傾向が顕著である。一方G11、G14、G17は、居住地に加えて隣接地域もよく訪れている点が特徴的である。G11、G14の利用者はそれぞれBrisbaneの南と北に居住しており、居住地域とBrisbaneの中心域の双方を訪れている。Brisbaneの中心域に居住するG17の利用者は、Brisbaneの南北に隣接する地域も訪問している。Brisbaneの中心とその周辺地域は、総体的な一つのまとまりを構成しており、移動と訪問が相互間で活発に行われていることが分かる。

表 2 利用者間及び群間距離。γは各群の medoid から利用者までの距離の平均値、ρは各群内の利用者相互間の平均値、右側は群間の距離行列。

利用者数	γ	ρ	G11	G12	G13	G14	G15	G16	G17
G11	42	12.08	21.26						
G12	11	15.21	26.76	281.3					
G13	11	22.42	42.86	437.5	103.5				
G14	37	7.10	13.10	101.6	121.1	281.3			
G15	12	11.28	19.07	91.8	351.6	500.0	179.7		
G16	16	30.84	52.65	82.0	406.3	562.5	234.4	117.2	
G17	85	4.87	9.11	44.9	187.5	343.8	23.4	128.9	168.0

次に、各群における訪問地点分布を詳しく分析する。図-6は各群内の利用者相互間 $s_{ik}(h)$ の平均である。G11、G14、G15、G17はいずれも比較的小さな h のときに明確な最大値をとっており、利用者の訪問地点が局所的にも広域的にも互いに類似していることが分かる。この結果は、これらの群でγとρがいずれも小さいという表2の結果と適合的である。他方、γとρが大きいG13とG16を見ると、前者では $s_{ik}(h)$ が最大となる h が比較的大きく、広域的に見て訪問地点がやや異なることが示唆される。これは実際、各利用者の訪問地点を示す図-7aで確認できる。一方G16は、極大値を2つ有している点が特徴的である。各利用者の訪問地点を図-7bで見ると、大半の利用者は点線で囲まれた地域を訪れており、それに加えて各利用者はそれぞれ異なる地域も訪れている。つまり、訪問地点の分布が部分的に類似しているが、広域的には異なっている。このような局所的な類似性は、分布の視覚的分析だけでは検出が困難であり、 $s_{ik}(h)$ の効用を示す結果であると言える。

以上のように本節では、前節で提案した手法の有効性を検証するために、ブリズベンにおけるスマートカード利用者を分類、各利用者の訪問地点分布の地理的特性について論じた。分類結果の厳密な正当性検証は困難であるが、表1、2の数値的指標と図4、5の訪問地点分布とはよく合致しており、結果は概ね妥当であると判断できよう。

本手法の利点の一つは、空間尺度を明示的に考慮しており、類似度評価の空間尺度を制御可能なことである。これまで点分布の分類には、 χ^2 値や R^* で構成される距離行列に基づいたクラスター分析という方法が一般的であった。この方法でも本節と同様の分類は可能であるが、結果の依拠する空間尺度を判断できないという問題が残る。それに対して本節の分類結果は、α=0.9という、広域的な空間尺度までを含む視点に基づくことが明示されており、図4や図-5からも、広域的な評価が行われていることが確認できる。

また図-6のような類似度の空間尺度ごとの評価を行うと、分布間の局所的類似性が検出できる。これは視覚的分析や既存の分布比較手法では不可能であり、今後、さらに空間データが増大したときには、その有効性はいっそう顕著となる。

図4 利用者の自宅（左上）と、各訪問地点における全利用者の延べ訪問回数。括弧内は各群の利用者数を示す。

図-5 利用者の自宅（左上）と、各訪問地点における全利用者の延べ訪問回数. 括弧内は各群の利用者数を示す.

図6 各群における関数 $s_k(h)$ の平均値.

図7 G13 及び G16 の各利用者の訪問地点. 色の差異は利用者の差異を表す.

4. 終わりに

本論文では、複数の点分布を比較、分類するための新たな手法を提案した。この手法は空間尺度を明示的に考慮しており、広域的、局所的類似性を評価できると同時に、点分布の分類にも適用可能である。また、位置データの誤差に対しても既存手法より頑健であるという利点を有する。

以下、今後の課題と拡張について述べる。

まず第1に、類似度の統計的な有意性検定がある。 $s_k(h)$ 及び $\eta_k(\alpha)$ は点分布 Γ_i と Γ_k の類似度を定量的に示す指標として有効であるが、それらの統計的な有意性についてはここでは論じていない。検定の方法としては例えば、 Γ_i と Γ_k と同数の点がランダムに分布する場合を帰無仮説としたモンテカルロシミュレーションを行い、 $s_k(h)$ と $\eta_k(\alpha)$ の確率分布を算出して比較するという方法が有り得るが、検定の度にシミュレーションを行うという手間が生ずる。解析的に $s_k(h)$ と $\eta_k(\alpha)$ の確率分布を導出するか、別の統計量を用いた検定手法の開発が必要である。

第2に、 α の決定に関する理論的裏付けを検討する必要がある。本論文では $\alpha=0.9$ あるいは $\alpha=0.95$ という値を提案しているが、これらの厳密な根拠に関する検討は十分とは言えない。通常の統計的枠組みでも、5%や1%などの有意水準に関する厳密な理論的検討は十分ではなく、 α についても同様に困難とも思われるが、 α の値と分析結果の関係

を詳しく見ることで、妥当な決定方法を探りたいと考えている。

第3に、空間尺度ごとの絶対的類似度指標の開発がある。関数 $s_k(h)$ は点分布 Γ_i と Γ_k の全体的な類似度に占める、空間尺度 h ごとの類似度の相対的な割合を表す値であり、その積分値は常に1である。そのため、異なる点分布対の間での直接的な比較には必ずしも適していない。 $\eta_k(\alpha)$ と併用すれば空間尺度ごとの絶対的類似度を評価できるが、単一指標の方がより簡便である。関数 $s_k(h)$ の定義を再検討するか、新たな指標の開発を行うことが望ましい。

【参考文献】

- Everitt BS, Landau S, Leese M, Stahl D (2011) *Cluster analysis, 5th edition*. Wiley,
- Hennig C, Meila M, Murtagh F, Rocci R (2015) *Handbook of cluster analysis*. Chapman and Hall/CRC,
- Lee Y (1979) A nearest-neighbor spatial-association measure for the analysis of firm interdependence. *Environment and Planning A* 11: 169-176
- Okabe A, Miki F (1984) A conditional nearest-neighbor spatial-association measure for the analysis of conditional locational interdependence. *Environment and Planning A* 16: 163-171
- Ripley BD (1976) The second-order analysis of stationary point processes. *Journal of applied probability*: 255-266
- Upton G, Fingleton B (1985) *Spatial data analysis by example. Volume 1: Point pattern and quantitative data*. John Wiley & Sons Ltd.,