

空間の統計学(5)：空間点過程

慶應義塾大学総合政策学部准教授

古谷 知之 (Furutani Tomoyuki)

■兵庫県生まれ。2001年東京大学大学院工学系研究科博士課程修了。博士（工学）。東京大学大学院助手、慶應義塾大学環境情報学部専任講師を経て、07年4月より現職。専門分野：空間統計学、都市交通計画、観光政策。



1. はじめに

今回は、「ポイントデータの分布がランダムかどうか」を分析するための手法を紹介します。この手法は、空間点過程やポイントパターンの分析として知られています。前回紹介した「空間集積性」に関する手法が面データや点データに対して適用されるのに対し、今回紹介する手法は点データに対してのみ適用されます。

空間点過程は、以下のように、広範な分野に応用されています。

- ① 空間疫学の分野：感染症（コレラ、インフルエンザなど）の発症者の空間分布解析など
- ② 生態学分野：森林などでの植生分布の均一性を検出する場合など
- ③ 地震工学：地震発生地点の空間分布解析など
- ④ 画像工学：プリンタインクの粒径分布の均質性を分析する場合など
- ⑤ 人間工学：アイマークレコーダでの注視点解析など

⑥ 都市工学：犯罪分布やGPSで得られた都市活動分布の解析など

⑦ 天体物理学：宇宙望遠鏡から得られた惑星や銀河の分布解析など

これらの分析で使われている点データには、発生数や種類などの属性が付与されており、マーク付き点過程と呼ばれています。今回の演習では、仮想的なデータとして、カラーブリンタのように4色（シアン、マゼンタ、イエロー、ブラック）のマーク付き点過程データを用いることにします。

2. カーネル密度関数

空間点過程の分析には、Rのspatstatというパッケージを用います。筆者のHP (<http://web.sfc.keio.ac.jp/~maunz/wiki>) からX600.txtというファイルをダウンロードして読み込んでください。このデータには、“c”・“m”・“y”・“b”というマークが付けられていますが、それぞれシアン、マゼンタ、イエロー、ブラックを意味します。spatstatパッケージ

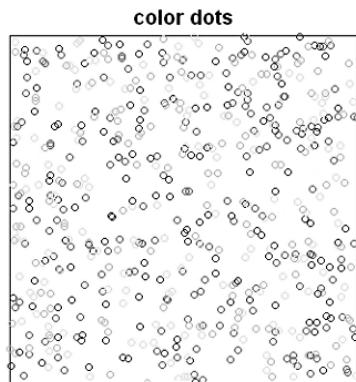
で2次元の空間点過程を分析するために、**ppp()**関数を使って座標値とマーク及び分析対象範囲（ウィンドウ）を定義します。

```
library(spatstat)
X600 <- read.table("X600.txt",
sep=",", header=T)
m <- sample(c("c","m","y","b"),
600, replace=TRUE)
m <- factor(m, levels=c("c","m","y","b"))
X <- ppp(X600$x, X600$y, c(0,1), c(0,1),
marks=m)
```

点過程全体の散布図を図1に示します。
ppp()関数で定義されたマーク付き空間点過程データをマーク別に表示・解析したい場合は、**split()**関数でマーク別データを抽出します。

```
plot(split(X)$b, main="color dots")
points(split(X)$m, col=14)
points(split(X)$y, col=7)
points(split(X)$c, col=5)
```

図1 点過程データ



density()関数を用いた場合、ガウス型カーネル関数を用いて計算される強度(intensity)が密度として表示されます。いま、 n 個の点データ $x_i (i = 1, \dots, n)$ が分析対

象地域 A に分布しているとします。地域 A の任意の地点 x に対するカーネルスマージング推計値 $\hat{\lambda}(x)$ は、次式から得られます。

$$\hat{\lambda}(x) = \frac{1}{h^2} \sum_{i=1}^n \frac{1}{h} k(x_i - x) / q(x) \quad (1)$$

ここで、 $k(x_i - x)$ はカーネル関数、 h は地点 x のバンド幅、 $q(x)$ は式(3)で示されるエッジ修正を意味します。カーネル関数には、ガウス関数、イパネクニコフ関数、四次関数 (biweight関数)、矩形関数、三角関数 (コサイン関数) などが適用されます。例えば、 $k(x_i - x)$ にガウス関数を適用する場合、カーネル関数は式(2)のように表されます。

$$k(x_i - x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|x_i - x\|^2}{2\sigma^2}\right) \quad (2)$$

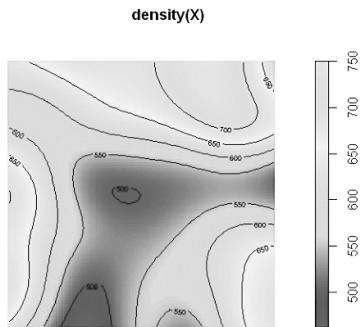
$$q(x) = \int_A k(v - x) \quad (3)$$

ここで、 σ^2 はカーネル関数の標準偏差、 v は対象地域の任意の地点を意味します。

図2では、バンド幅 h を0.1としたときの点過程全体のカーネル密度分布を示しています。
contour()関数を用いると、密度関数のコンター図を描くことができます。また、**persp()**関数を用いると、密度を3次元で表示することができます。

```
# 点過程全体の密度マップ
plot(density(X), 0.1)
contour(density(X), add=T)
# シアン色のみの密度マップ
plot(density(split(X)$c), 0.1)
contour(density(split(X)$c), add=T)
# 密度関数の3次元表示
D <- density(X)
persp(D)
```

図2 密度マップ



3. 点データのランダム性

点データのランダム性を検証するベンチマーク的な手法として、空間データの完全ランダム性（Complete Spatial Randomness：CSR）を示す方法が知られています。この方法は、点過程がランダムに分布する場合、均質なポアソン過程に従うと仮定し、「点過程が完全にランダムである」という帰無仮説に対して仮説検定を行い、CSRでないという証拠を見つけることが基本的な分析となります。

CSRを証明する代表的な手法として、① χ^2 検定やコルモゴロフ・スマイルノフ検定を用いる方法（便宜上、古典的仮説検定方法と呼びます）、②ポアソン過程に対して最尤法を適用しポアソンモデルを推定する方法、③距離に基づく統計量を用いる方法などがあります。このほか、点同士の相関や点と背景データとの相互作用を考慮したギブス点過程による方法などもありますが、誌面の都合上、割愛します。

(1) 古典的仮説検定方法

①コドラーを使った χ^2 検定

この方法は、分析対象地域を「コドラー

ト」と呼ばれる任意の方形サブ領域に分割し、各コドラーの点密度に対してピアソンの χ^2 検定を行う方法です。この方法では、 χ^2 検定によるp値が、例えば、非常に小さいとき帰無仮説を棄却し、点過程がランダムでないと判断します。Rでは、**quadrat.test()**関数を用いて計算することができます。図3は、 χ^2 検定の結果を各コドラー上に示したものですが、コドラー内の中左上が実測値（ポイント数）、右上が理論値、下が実測値と理論値との差を意味します。

```
quadrat.test(X, nx=5, ny=5)
plot(quadrat.test(X, nx=5, ny=5), col=2)
```

図3 コドラー法による検定結果

quadrat.test(X, nx = 5, ny = 5)

23 24 -0.2	27 24 0.61	26 24 0.41	34 24 -2	23 24 -0.2
23 24 -0.2	26 24 0.41	26 24 0.41	25 24 0.2	29 24 1
34 24 -2	17 24 -1.4	19 24 1	16 24 -1.6	17 24 -1.4
25 24 -0.2	22 24 -0.41	18 24 -1.2	26 24 0.41	26 24 0.41
25 24 0.2	16 24 -1.6	27 24 0.61	18 24 -1.2	32 24 1.6

②コルモゴロフ・スマイルノフ検定

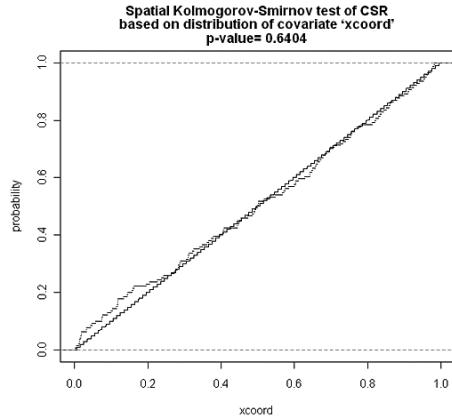
対象地域全体の観測点に対して、任意の関数から得られる期待地点を比較し、コルモゴロフ・スマイルノフ検定を行う方法があります。例えば、関数としてx座標を採用し、x座標の観測値と期待値との差の検定を行うことができます。Rでは、**kstest()**関数を用いてコルモゴロフ・スマイルノフ検定を適用することができます（図4）。

```

xcoord <- function(x,y) {x}
kstest(split(X)$c, xcoord)
plot(kstest(split(X)$c, xcoord))

```

図4 コルモゴロフ・スミルノフ法による検定結果



(2) ポアソンモデルを推定する方法

点過程 x_i と未知パラメータ θ からなる密度関数 $\lambda_\theta(x)$ を用いて、均一なポアソン過程の尤度関数は、次式のように書くことができます。

$$\log L(\lambda_\theta(x)) = \sum_{i=1}^n \log \lambda_\theta(x_i) - \int_A \lambda_\theta(x) dx \quad (4)$$

$$\log \lambda_\theta(x) = \sum_{j=1}^p \theta_j z_j(x) \quad (5)$$

ここで、 $z_j(x)$ は地点 x の共変量を意味します。密度関数 $\lambda_\theta(x)$ として、例えば、次のような関数が考えられます。

$$\lambda_\theta(x) = \theta_0 + \theta_1 x + \theta_2 y \quad (6)$$

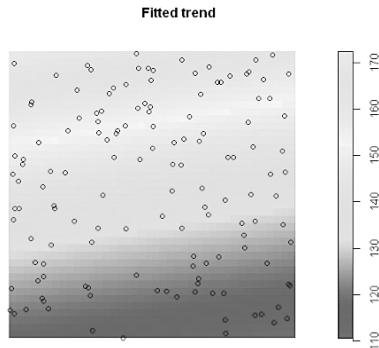
Rでは、**ppm()**関数を用いてポアソンモデルを推定することができます。モデルの当てはまり傾向を図5に示します。

```

modelc <- ppm(split(X)$c, ~x+y)
# モデルの当てはまり傾向
plot(modelc, how="image", se=FALSE)
# 予測トレンド
plot(predict(modelc, type="trend",
ngrid=512))

```

図5 ポアソンモデルの当てはまり傾向



(3) 距離に基づく統計量を用いる方法

距離に基づく統計量として、① F 統計量、② G 統計量、③ K 統計量 (L 統計量)、④ ペア相関、⑤ J 統計量などがあります。

① F 関数と F 統計量

任意の地点から他の最近隣地点までの距離について、 F 関数を使って得られる統計量です。 F 関数は任意の地点からの半径 r を用いて、次式のように表されます。

$$F(r) = 1 - \exp(-\lambda \pi r^2) \quad (7)$$

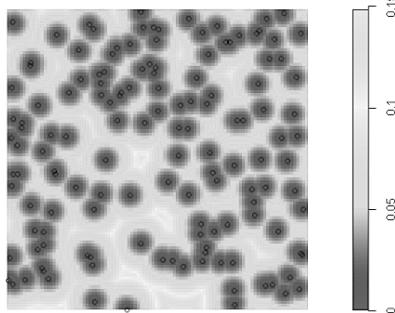
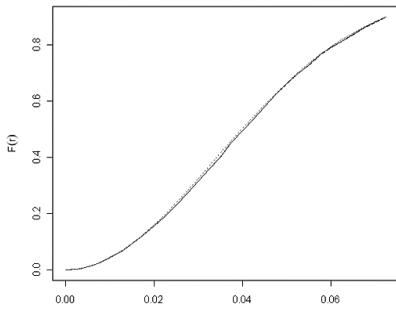
Rでは、**Fest()**関数を用いて F 統計量を計算することができます。各地点からの距離を図6に、 F 関数と F 統計量を図7に示します。

```

plot(distmap(split(X)$c))
plot(split(X)$c, add=T)
plot(Fest(split(X)$c))

```

図6 距離マップ

`distmap(split(X)$c)`図7 F 関数と F 統計量`Fest(split(X)$c)`

② G 関数と G 統計量

任意の地点からの最近隣距離の分布を計測することにより、 G 統計量が得られます。ある地点 i から他の地点 j ($\neq i$) との距離のうち、最小となる組み合わせ (= 最近隣距離) を d_{ij} とし、 $d_{ij} \leq r$ となる地点の数を $N(d_{ij}; d_{ij} \leq r, \forall i)$ と書くことにします。このとき、 G 統計量を得るための G 関数は次式のように書くことができます。

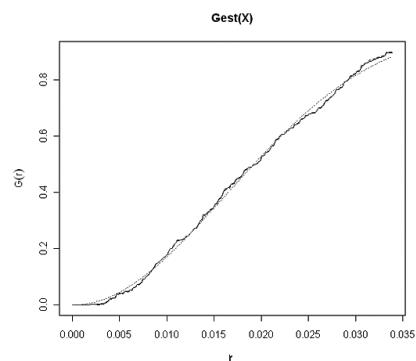
$$\hat{G}(r) = \frac{N(d_{ij}; d_{ij} \leq r, \forall i)}{n} \quad (8)$$

点過程が完全ランダムであるとき、 G 統計

量は次式のように表されます。

$$G(r) = 1 - \exp(-\lambda \pi r^2) \quad (9)$$

Rでは、**Gest()**関数を用いて G 統計量を計算することができます(図8)。

`plot(Gest(X))`図8 G 関数と G 統計量

③ K 統計量と L 統計量

地点 ij 間のペアワイズ距離 $s_{ij} = \|x_i - x_j\|$ を用いて、 $s_{ij} \leq r$ となる地点の数を $N(s_{ij}; s_{ij} \leq r, \forall i)$ と書くことにします。このとき、 K 統計量を得るための K 関数は次式のように書くことができます。

$$\hat{K}(r) = \frac{a}{(n-1)\pi} N(s_{ij}; s_{ij} \leq r, \forall i) q(x) \quad (10)$$

ここで、 a はウインドウの面積、 $q(x)$ は式(3)で示されるエッジ修正です。

点過程が完全ランダムであるとき、 K 統計量は次式のように表されます。

$$K(r) = \pi r^2 \quad (11)$$

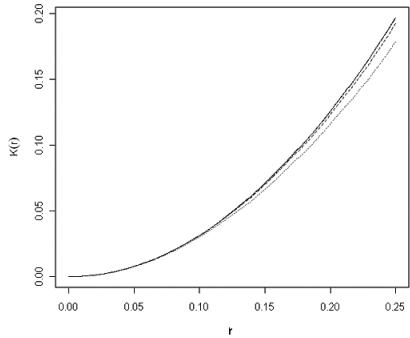
Rでは、**Kest()**関数を使って K 統計量を計算することができます(図9)。

```

K <- Kest(X)
plot(K)
Kc <- Kest(split(X)$c)
plot(Kc)

```

図9 K 関数と K 統計量



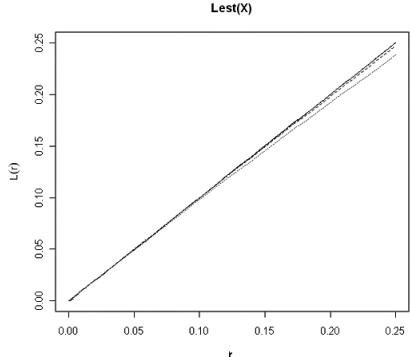
しばしば、 K 統計量を次式のように変形した L 統計量が一般的に用いられます。

$$L(r) = \sqrt{\frac{K(r)}{\pi}} \quad (12)$$

つまり、ポアソン過程では $L(r) = r$ となります。Rでは、**Lest()**関数を用いて L 統計量を計算することができます(図10)。

```
plot(Lest(X))
```

図10 L 関数と L 統計量



④ペア相関関数

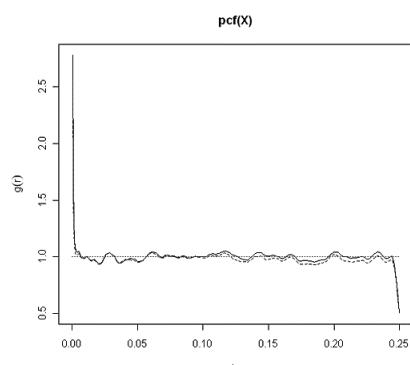
K 関数の導関数を用いて、ペア相関関数を定義することができます。

$$g(r) = \frac{K'(r)}{2\pi r} \quad (13)$$

Rでは、**pcf()**関数を使ってペア相関を計算することができます(図11)。

```
plot(pcf(X))
```

図11 ペア相関



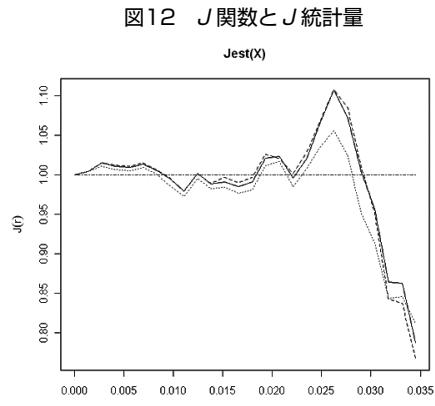
⑤J 統計量

G 統計量と F 統計量を組み合わせた統計量として、 J 統計量が知られています。

$$J(r) = \frac{1 - G(r)}{1 - F(r)} \quad (14)$$

Rでは、**Jest()**関数を用いて J 統計量を計算することができます。 J 関数と J 統計量を図12に示します。

```
plot(Jest(X))
```



4. マーク付き点過程の分析

前節で紹介した距離に基づく統計量を、マーク付き点過程に適用することにより、マーク間の点過程の独立性やランダム性を示すことができます。ただし、マークが量的変数か質的変数かによって分析方法が異なります。マーク付き点過程の分析方法は、①特定のタイプ（マーク） i と他のタイプ j との組み合わせを分析する方法と、②特定のタイプ（マーク） i と他のすべてのタイプの組み合わせを分析する方法とがあります。

前者は、 F 統計量、 G 統計量、 K 統計量（ L 統計量）、ペア相関、 J 統計量に対して適用することができます。Rを使ってマーク付き点過程の K 統計量を求める場合は、**Kcross()**関数を用いて計算することができます。

```
plot(Kcross(X, "c", "b"))
plot(alltypes(X, "K"))
```

また、後者は、 G 統計量、 K 統計量（ L 統計量）、 J 統計量に対して適用することができます。Rを使ってマーク付き点過程の K 統計量を求める場合は、**Kdot()**関数を用いて計算

することができます。

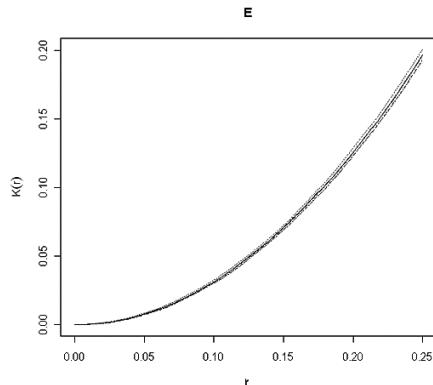
```
plot(alltypes(X, "Kdot"))
```

5. シミュレーションによる適合度分析

点過程のランダム性を示す統計量を得るために、 K 関数などの関数を用いてモンテカルロシミュレーションによる適合度分析を行うことができます。この方法を用いると、与えられた点過程がシミュレーションによる包絡線の最小値と最大値の間に含まれるかどうかを判断することができます。Rでは、**envelope()**関数を用いて適合度分析を行うことができます（図13）。

```
E <- envelope(X, Kest, nsim=99)
plot(E)
```

図13 シミュレーションによる適合度分析



*参考文献

A. Baddeley (2008) : Analysing spatial point pattern in R, Workshop Notes version 3, CSIRO (<http://www.csiro.au/files/files/pn0y.pdf>).