空間の統計学(4)：空間集積性

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

今回紹介する手法は、空間疫学などの分野で用いられる「空間集積性」に関する方法です。この方法は、①空間データが特定の地域周辺に集積している（あるいはランダムに分布している）かどうかを仮説検定する方法と、②空間データが集積している場所を検出する方法とに大別されます。空間集積性の検証には、基本的に、空間データの座標情報、属性の観測値及び期待値が用いられます。空間属性の期待値は、前回紹介した確率地図などを用いて計算されます。今回、これらの方法を適用するために、RのDClusterという空間疫学の分析用に開発されたパッケージを使います。

前回、経験ベイズ法を用いた相対リスクの計算方法を紹介しました。DClusterパッケージには、前回紹介しなかったポアソン・ガンマモデルと対数正規モデルを用いて相対リスクを計算する関数が含まれています。そこで、今回はまず、前回の続きとして前半で、経験ベイズ法に関する2つの手法を紹介し、後半では、空間集積性に関する検定手法と空間クラスターの場所を検出する手法を紹介します。

2. 経験ベイズ法（前回の続き）

演習には、前回の演習でも用いた、2006年の都道府県別人口数と心疾患死亡者数に関するデータを、総務省統計局の「社会生活統計指標－都道府県の指標－」[1]からダウンロードして用います。

```
library(spdep)
library(DCluster)
jpn_pref <-
readShapePoly("jpn_pref.shp",IDvar="PREF_CODE")
pref_pnt <-
readShapePoints("pref_gov.shp")
hd06 <- read.table("76dat.csv", sep=",",
header=T)
ID.match <- match(jpn_pref$PREF_CODE,
hd06$PREF_CODE)
jpn_hd06 <- hd06[ID.match,]
jpn_pref_hd06 <- spCbind(jpn_pref, jpn_hd06)
```

⑴ ポアソン・ガンマモデル

ポアソン・ガンマモデルを用いて観測値と期待値からなる相対リスクを平滑化する方法です。地区 i ($i=1, \ldots, n$) の観測値 O_i、期待値 E_i 及びガンマ関数のパラメータ α と ν を用いて、平滑化相対リスク $\frac{O_i+\nu}{E_i+\alpha}$ を得ます。

ポアソンモデルとガンマモデルの2段階の階層ベイズモデルから、事前情報を得ます。

$$O_i \mid \theta_i \sim Po(\theta_i, E_i) \quad (1)$$

$$\theta_i \sim Ga(\nu, \alpha) \quad (2)$$

観測値 O_i の条件付き事後分布は、生起回数 ν、確率 $\alpha/E_i+\alpha$ の負の二項分布に従います。このとき、相対リスクの推計値は

$$\hat{\theta}_i = E[\theta_i \mid O_i, E_i] = \frac{O_i+\nu}{E_i+\alpha} \quad (3)$$

となり、パラメータ α と ν の推計値 $\hat{\alpha}$ と $\hat{\nu}$ は、次の2式を繰り返し計算し、その平均値を求めることで得られます。

$$\frac{\hat{\nu}}{\hat{\alpha}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i \quad (4)$$

$$\frac{\hat{\nu}}{\hat{\alpha}^2} = \frac{1}{n+1}\sum_{i=1}^{n}\left(1+\frac{\hat{\alpha}}{E_i}\right)\left(\hat{\theta}_i - \frac{\hat{\nu}}{\hat{\alpha}}\right)^2 \quad (5)$$

Rでは、DClusterパッケージの **empbaysmooth( )**関数を使って、ポアソン・ガンマモデルを用いた平滑化経験ベイズ法による相対リスクを計算することができます。

```
jpn_hd06_pm <-
probmap(jpn_pref_hd06$HD06,
jpn_pref_hd06$POPJ06/100)
summary(jpn_hd06_pm)
jpn_hd06_sm <-
empbaysmooth(jpn_hd06$HD06,
jpn_hd06_pm$expCount)
jpn_hd06_sm
```

⑵ 対数正規モデル

対数正規モデルを用いて観測値と期待値からなる相対リスクを平滑化する方法です。相対リスクの対数値 $\beta_i=\log(\theta_i)$ は、平均 $\phi=\mu_i$、分散 σ^2 の多変量正規分布に従います。観測値がゼロでも対数相対リスクが存在するように、

$$\theta_i = \frac{O_i+1/2}{E_i} \quad (6)$$

とすると、

$$\log\left(\frac{O_i+1/2}{E_i}\right) \sim N(\phi, \sigma^2) \quad (7)$$

となります。

EMアルゴリズムにより繰り返し計算することで、平滑化経験ベイズ推定値 $\hat{\beta}_i$ を得ます。

$$\hat{\beta}_i = \frac{\phi + (O_i+1/2)\hat{\sigma}^2\log[((O_i+1/2)E_i)] - \hat{\sigma}^2/2}{1+(O_i+1/2)\hat{\sigma}^2} \quad (8)$$

ここで、平均と分散の推定値は、

$$\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n}\beta_i \quad (9)$$

$$\hat{\sigma}^2 = \frac{1}{n}\left\{\hat{\sigma}^2\sum_{i=1}^{n}[1+\hat{\sigma}^2(O_i+1/2)]^{-1} + \sum_{i=1}^{n}(\beta_i-\hat{\phi})^2\right\} \quad (10)$$

となります。

Rでは、**lognormal**( )関数を使って、対数正規モデルを用いた平滑化経験ベイズ法による相対リスクを計算することができます。

```
jpn_hd06_ln <-
lognormalEB(jpn_hd06$HD06,
jpn_hd06_pm$expCount)
jpn_hd06_ln
```

3. 空間集積性に関する検定

空間的な事象がランダムに発生する場合、観測値O_iの発生確率はポアソン分布や二項分布、負の二項分布といった確率分布に従うと考えることができます。空間属性が集積しているかどうかは、空間データの観測値がランダムな確率分布に従って生起しているかどうかを基準として判断することができるといえます。そこで、空間属性の集積性に関する検定方法として、「空間データがランダムに分布している（＝特定地域周辺に空間属性が集積していない)」という帰無仮説H_0と、「空間データがランダムに分布していない（＝特定地域周辺に空間属性が集積している)」という対立仮説H_1をもとに、仮説検定する方法が提案されています。Moran's Iを用いた空間的自己相関の検出も、空間集積性に関する検定の一種といえます。

DClusterパッケージでは、データの過分散を検証するために、ポアソン分布または負の二項分布を用いたブートストラップ法を適用することができます。

ここで、観測値O_i及び期待値E_iについて、地域全体の合計を、それぞれ

$$O_+ = \sum_{i=1}^{n} O_i \qquad (11)$$

$$E_+ = \sum_{i=1}^{n} E_i \qquad (12)$$

と書くことにします。

演習には、東京都[2]、神奈川県[3]、千葉県[4]、埼玉県[5]のHPから入手可能な平成18年度の人口動態に関する統計をダウンロードして用いることにします。

(1) ピアソンのχ^2検定

空間属性がランダムに分布しているかどうかを、ピアソンのχ^2検定を用いて検証する方法です。

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - \theta E_i)^2}{\theta E_i} \qquad (13)$$

ここで、相対リスク$\theta = O_+/E_+$です。

内部標準化を行うと、$O_+ = E_+$であることから、帰無仮説H_0と対立仮説H_1は、

$$H_0 : \theta = 1$$

$$H_1 : \theta \neq 1$$

となります。このとき、自由度$n-1$のχ^2検定を行います。

Rでは、**achisq.test**( )関数を用いて、χ^2検定を適用できます。

```
achisq.stat(data76_OE, lambda=1)
achisq.test(Observed~offset(log(Expected)),
data76_OE, model="poisson", R=100)
data76_achb_pb <- boot(data76_OE,
statistic=achisq.pboot, sim="parametric",
ran.gen=poisson.sim, R=100)
plot(data76_achb_pb)
```

(2) Potthoff-Whittinghillの検定

全ての地区について、相対リスクが互いに

類似しているかどうかを検証する方法です。帰無仮説H_0と対立仮説H_1を、それぞれ以下のようにおきます。

$$H_0 : \theta_1 = \cdots = \theta_n = \lambda$$

$$H_1 : \theta_1 \sim Ga(\lambda^2/\sigma^2, \lambda/\sigma^2)$$

このとき、次式で表される統計量を用いて、仮説検定を行います。

$$E_+ \sum_{i=1}^{n} \frac{O_i(O_i - 1)}{E_i} \quad (14)$$

Rでは、**pottwhitt.test**( )関数を用いて、Potthoff-Whittinghillの検定を適用できます。

```
pottwhitt.stat(data76_OE)
pottwhitt.test(Observed~offset(log(Expected)),
data76_OE, model="poisson", R=100)
data76_pw_pb <- boot(data76_OE,
statistic=pottwhitt.pboot, sim="parametric",
ran.gen=poisson.sim, R=100)
plot(data76_pw_pb)
```

(3) Tangoの検定

空間オブジェクト同士の隣接性（隣接行列や距離行列）を用いて、空間属性の集積性を検証しようとする方法です。ここで、

$$r = [O_1/O_+, \ldots, O_n/O_+]^T \quad (15)$$

$$p = [E_1/E_+, \ldots, E_n/E_+]^T \quad (16)$$

$$a_{ij} = \exp(-d_{ij}/\lambda) \quad (17)$$

とし、行列Aが要素a_{ij}からなる行列とします。ただし、d_{ij}は地点ij間の距離です。次式で表されるTango統計量を用いて集積性の有無をχ^2検定することができます。

$$T = (r - p)^T A (r - p) \quad (18)$$

tango.test( )関数を用いてTangoの検定を適用してみましょう。

```
data76_OE <- cbind(data76_OE,
x=data76_spdf$Easting,
y=data76_spdf$Northing)
coords <-
as.matrix(data76_OE[,c("x","y")])
dlist <- dnearneigh(coords, 0, Inf)
dlist <- include.self(dlist) dlist.d
<- nbdists(dlist, coords)
col.W.tango <- nb2listw(dlist,
glist=lapply(dlist.d,
function(x){exp(-x)}), style="C")
tango.stat(data76_OE, col.W.tango,
zero.policy=TRUE)
tango.test(Observed~offset(log(Expected)),
data76_OE, model="poisson",
R=100, list=col.W.tango,
zero.policy=TRUE)
data76_tn_pb <- boot(data76_OE,
statistic=tango.pboot, sim="parametric",
ran.gen=poisson.sim, R=100,
listw=col.W.tango, zero.policy=TRUE)
plot(data76_tn_pb)
```

(4) Whittermoreの検定

Tangoの検定と同様に、空間オブジェクト間の距離d_{ij}を要素とする距離行列Dを用いて空間属性の集積性を検証する方法ですが、期待値を使わない点で異なります。

$$W = r^T D r \quad (19)$$

whittermore.test( )関数を用いてWhittermoreの検定を適用してみましょう。

```
col.W.whitt <- col.W.tango
whittermore.stat(data76_OE,
col.W.whitt, zero.policy=TRUE)
whittermore.test(Observed~offset
(log(Expected)), data76_OE, model="poisson",
```

```
R=100, list=col.W.whitt, zero.policy=TRUE)
data76_wt_pb <- boot(data76_OE,
statistic=whittermore.pboot,
sim="parametric", ran.gen=poisson.sim,
R=100, listw=col.W.whitt, zero.policy=TRUE)
plot(data76_wt_pb)
```

4. 空間クラスター位置の検出

この方法は、「スキャン統計量」という方法を用いて、クラスターを検出する「ウィンドウ」の中に含まれる観測値O_iが、ウィンドウ内の空間属性の期待値E_iに等しくなるかどうかを仮説検定することで、空間データが集積している場所を検出する方法です。

⑴ Geographical Analysis Machine（GAM）

分析対象地域上にグリッドを描き、正方グリッド上の交点k（$k=1, \ldots, p$）を中心に様々な半径の円を描き、これらをウィンドウとします。

ウィンドウkに含まれる地区（地点）の観測値の合計をO_{k+}、期待値の合計をE_{k+}としましょう。このとき、O_{k+}がE_{k+}に対して有意に大きい値をとるとき、交点kはクラスターの候補となり、地図上にマークされます。Rでは、**opgam**( )関数を用いてGAMを計算することができます。

```
thegrid <- as(data76_OE,
"data.frame")[,c("x","y")]
data76_opg <-
opgam(data=as(data76_OE, "data.frame"),
thegrid=thegrid,radius=20000,
step=1000, alpha=0.05)
data76_opg
plot(data76_OE$x,data76_OE$y,
xlab="Easting", ylab="Northing")
points(data76_opg$x, data76_opg$y,
col="red", pch="*")
```

この方法は、クラスター候補を決める際に、ウィンドウの半径と位置を変動させるたびに統計的有意性を検定しないという点や重なるクラスターが多く生成される点などが批判されていますが、空間クラスターを検出する基本的な考え方になっているといえます。

⑵ Besag and Newellの統計量

空間データの観測値O_iが非常に小さい場合に用いられる方法です。クラスター候補となる地区の中心点（セントロイド）をラベル0とし、当該地区をA_0とします。そして、A_0から地区中心点が近い順に、他の地区をA_j（$j=1, 2, \ldots, i-1$）とラベル付けします。地域iの人口数をP_i、対象地域全体の人口総数をP_+と表すとき、次のような統計量D_i及びu_iを用いて、クラスター候補の有意水準を求めます。

$$D_i = \sum_{j=0}^{i} O_j - 1 \qquad (20)$$

$$u_i = \sum_{j=0}^{i} P_j - 1 \qquad (21)$$

期待されるクラスターのサイズkに対して、

$$M = \min\{i : D \geq k\} \qquad (22)$$

とすると、最近隣のM地区に対して最も近いk個のクラスターが形成されます。Mの観測値がmであるとすると、クラスターとなりうる場合の有意水準は次式のようになります。

$$Pr = (M \leq m) = 1 - \sum_{s=0}^{k-1} \exp(-u_m(O_+/P_+))(u_m(O_+/P_+))^s/s! \qquad (23)$$

式㉓を用いて、例えば、有意水準５％以下の場合のクラスターをマップ上に描くことができます。

```
data76_bn_perboot <- boot(data76_OE,
statistic=besagnewell.boot, R=100, k=20)
plot(data76_bn_perboot)
```

⑶ Kulldorffの統計量

地区 i の中心点を中心とする同心円で表されるウィンドウを Z_i とし、Z_i の要素を z とします。このとき、次式で表される最大尤度比をとるウィンドウをクラスターの候補と考えます。

$$\max_{z \in Z_i} \left(\frac{O_z}{E_z} \right)^{O_z} \left(\frac{O_+ - O_z}{E_+ - E_z} \right)^{O_+ - O_z} \quad ㉔$$

また、最大尤度比を得るために、モンテカルロシミュレーションで求めたｐ値を用いて有意性を判断します。

```
data76_kn_pboot <- boot(data76_OE,
statistic=kullnagar.pboot,
sim="parametric", ran.gen=poisson.sim,
R=100, fractpop=0.5)
plot(data76_kn_pboot)
```

⑷ Stoneの検定

この方法では、各地区の相対リスクが等しいという帰無仮説 H_0 に対して、特定の地域からの距離が増える毎に相対リスクが減少するという対立仮説 H_1 をおき、仮説検定を行います。すなわち、帰無仮説 H_0 と対立仮説 H_1 は次のようになります。

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_n = \lambda$$

$$H_1 : \theta_1 \geq \theta_2 \geq \cdots \geq \theta_n$$

推定されるリスク源から順にデータが並び替えられているとすると、次式の検定統計量を用いて、その有意性を評価します。

$$\max \frac{\sum_{j=1}^{i} O_j}{\sum_{j=1}^{i} E_j} \quad ㉕$$

Ｒでは、**stone**( )関数を用いてStoneの検定を行うことができます。

```
stone.test(Observed~offset(log(Expected)),
data76_OE, model="poisson", R=100,
region=which(row.names(data76_OE)=="20"),
lambda=1)
data76_st_pb <- boot(data76_OE,
statistic=stone.pboot,
sim="parametric",
ran.gen=poisson.sim, R=100,
region=which(row.names(data76_OE)=="20"))
plot(data76_st_pb)
```

＊参考URL

[1] 総務省統計局：社会生活統計指標－都道府県の指標－2009（http://www.stat.go.jp/data/ssds/5.htm）.

[2] 東京都福祉保健局：人口動態統計（http://www.fukushihoken.metro.tokyo.jp/kiban/chosa_tokei/eisei/jinkou/）.

[3] 神奈川県：平成18年神奈川県衛生統計年報統計表（http://www.pref.kanagawa.jp/osirase/tiikihoken/joho/nenpo/H18/jinkodoutaitop.html）.

[4] 千葉県：平成18年千葉県衛生統計年報（http://www.pref.chiba.lg.jp/syozoku/c_syafuku/joho/h18eiseitoukei/h18eiseitoukei.html#jinko）.

[5] 埼玉県：平成18年埼玉県の人口動態概況（http://www.pref.saitama.lg.jp/A04/BA00/jindo%2815-%29/jindo18/jindo18.htm）.