町丁字別犯罪データの空間分析

大下　祐樹

2007年11月15日

1　研究動機

卒業論文、修士に進学し研究してきた岡山市交番管轄ごとの空き巣データであった。空間的集積性、空間自己回帰モデル、地理的加重回帰法を主な解析手法として研究を進めてきたが、交番管轄の面積は岡山市中心部と郊外との差が大きく、よって空間的な従属性の差にもつながっておりモデリングの際に苦労してきた。対処法として、岡山市中心部のみのデータを用いて解析するなどしてきたがデータ数の関係もあり良い結果は得られなかった。そこで、埼玉県川口市のホームページに町丁字別窃盗データが公開されているのを知り、解析を行った。本論文では空き巣データの空間的な集積性を分析した上で、空間自己回帰モデル、地理的加重回帰モデルを推定しその結果を示す。

2　空間的な集積性

2.1　Moran 散布図

Moran 散布図とは横軸に地区 i の犯罪認知数 y_i , 縦軸に当該地区 i に隣接する地区群の平均犯罪認知数 ${y^*}_i$ としたものである。${y^*}_i$ を説明する。ここで、空間的位置関係を示す記号として１次の隣接性指標を定義する。

$$c_{ij} = \begin{cases} 1 & \text{地区 } i \text{ と } j \text{ が隣接} \\ 0 & \text{そうでないとき} \end{cases} \tag{1}$$

ここで c_{ij} を１次の連結性行列と呼ぶ。地区 i ごとに（行ごとに）総和が１となるように基準化した

$$c'_{ij} = \frac{c_{ij}}{\sum_{j=1}^{n} c_{ij}} \tag{2}$$

n:地区数
つまり、

$$\sum_{j=1}^{n} c'_{ij} = 1$$

となる。この基準化した c_{ij} を用いて、地区 i に隣接する地区群での平均犯罪認知数 ${y^*}_i$ は

$${y^*}_i = \sum_{j=1}^{n} c'_{ij} y_j \tag{3}$$

となる。

2.2 Moran 'I 統計量

Moran$I$ 統計量とは

$$I = \frac{1}{W} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{{\sigma_y}^2} \tag{4}$$

でる。なお

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{5}$$

$${\sigma_y}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{6}$$

w_{ij} は空間的近接性を示す重みであり、$c_{ij}, {c'}_{ij}$, 地区間距離 d_{ij} の逆数 $\frac{1}{d_{ij}}$ などがよく用いられる。W はこの重みの総和である。つまり

$$W = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \tag{7}$$

Moran 散布図は空間的はずれ値の検出に利用できる。回垂直方向に大きくは外れた地区は、周辺地区の水準から特異には外れた属性値を持つ。Moran 散布図は、対象地域全体に見られる空間的変動の傾向の中で個別地域の位置づけを果たす。

2.3 Local Moran 'I 統計量

Moran 'I 統計量は、全体的な空間的変動の規則性を簡潔に要約している。しかし、犯罪などの空間的集積に関心のある場合、全体的な傾向ばかりでなくホットスポットと呼ばれる大きな属性値の集積する地域など、特定の地域に着目した特性を検討したい。ここで、先ほどの Moran の統計量を各区域ごとの成分の L_i の和として書き直す。

$$I = \frac{\sum_{i=1}^{n} L_i}{W} \tag{8}$$

$$L_i = \frac{(y_i - \bar{y}) \sum_{j=1}^{n} w_{ij}(y_j - \bar{y})}{{\sigma_y}^2} \tag{9}$$

Local Moran 'I 統計量は指標の局所的類似性の度合いを示している。ここで、Moran 散布図の縦軸と横軸をそれぞれ平均値 $\bar{y}$ で区切り、4 つの象限に分けて考える。
すると、L_i が正になるのは第 、 象限に位置する地区であり、この値が有意に大きい場合、属性値の特異な空間的集積性が、当該地区 i を中心に存在する。ただし、第 象限内の点では、

平均よりも大きな属性値の集積を、第　象限では平均よりも小さな属性値の集積を示す。
L_i が負になるのは第　、　象限に位置する地区であり、この値の絶対値が有意に大きい場合、属性値の特異な空間的変化が、当該地区 i とそれを取り囲む地区間に存在する。ただし、第　象限内の点では、当該地区 i で平均よりも小さな属性、近傍の地区群では平均よりも大きな属性、第　象限ではその逆の空間的変化を示す。

2.4　モンテカルロ検定

乱数を利用して各地区の値を並び換え、適当な組の仮想データを作り、そのおのおのの組で I 統計量を算出する。そして元のデータで観測された I 計量の値よりも大きな I 統計量が得られる割合を上側並び替え p 値とする。両側 5% 検定を行うときは、$p < 0.025$ なら正の空間的自己相関が、$p > 0.975$ なら負の空間的自己相関が、それぞれ有意と判定する。ローカルモラン I 統計量の検定の検定の場合は、ローカルモラン I 統計量の絶対値について検定する。

2.5　回帰モデルと残差の空間的自己相関

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{10}$$
$$\varepsilon_i \sim N(0, {\sigma_\varepsilon}^2) \tag{11}$$

すると地区 i の残差は

$$e_i = y_i - \hat{y}_i \tag{12}$$
$$= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \tag{13}$$

残差分散は

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n} {e^2}_i \tag{14}$$

平均平方誤差 RMSE は

$$\mathrm{RMSE} = \sqrt{\hat{\sigma}_\varepsilon^2} \tag{15}$$

決定係数 R^2 は

$$R^2 = 1 - \frac{\hat{\sigma}_\varepsilon^2}{{\sigma_y}^2} \tag{16}$$

モデルの適合度が優れていれば、誤差の大きさを表す平均平方誤差 (RMSE ≥ 0) は小さく、決定係数は大きくなる。
通常の回帰モデルでは、誤差は空間的に独立であり、互いに無相関であることを前提としてい

る。その仮定が満たされているかどうかの確認に空間的自己相関を利用する。次のように残差の Moran ’I 統計量を定義できる。

$$I_e = \frac{1}{W} \cdot \frac{\sum_i^n \sum_j^n w_{ij} e_i e_j}{\hat{\sigma}_\varepsilon^2} \tag{17}$$

3　埼玉県川口市の犯罪データの解析

本研究で用いるデータは埼玉県川口市のホームページで公開されている町丁字別犯罪（窃盗５手口）発生件数である。

平成１８年度の合計データと平成１９年度の各月の合計データが公開されており, 本研究では一年間の合計数である平成１８年度合計データを用いることにした。

犯罪データの一部を以下に示す。

地区ナンバー	町丁字	自転車盗	オートバイ盗	ひったくり	車上荒らし	空き巣
1	東川口５丁目	14	1	0	6	6
2	東川口４丁目	10	4	2	11	7
3	東川口１丁目	19	10	1	0	3
4	東川口３丁目	21	4	1	12	5
5	東川口６丁目	9	3	0	6	8

それぞれの犯罪の発生率（1,000 世帯あたり）の要約数を示す。

	統計量	自転車盗	オートバイ盗	ひったくり	車上荒らし	空き巣
1	Min.	0	0	0	0	0
2	1st Qu.	5.305	1.340	0	2.947	1.340
3	Median	8.016	3.058	0	5.014	36
4	Mean	13.073	4.464	1.069	6.909	4.073
5	3rd Qu.	11.850	5.728	1.566	8.071	5.488
6	Max.	321.472	66.667	9.804	47.619	34.884

どの犯罪にも外れ値が存在していることがわかる。
では、それぞれの犯罪がどの地区に集積しているのか分析をする。

3.1　Moran'I

	Moran'I
自転車盗	0.11*
オートバイ盗	0.07*
ひったくり	0.11**
車上荒らし	0.01
空き巣	0.23**

ひったくり以外には集積性が有意水準 5% で示唆された。

では、この中でも特に集積性が高い空き巣に焦点をおいて解析を進めることにする。

3.2 空き巣発生率と地域指標の関係

空き巣発生率と地域指標データとの関係を探りたい。そこで、平成１２年度国勢調査より町丁字別住宅の建て方のデータ（単位は世帯）と、平成１８年度川口市住民基本台帳のデータより町丁字年齢別人口数と世帯数のデータを手に入れた。地域的指標の一部を以下に示す。

	人口	６５歳以上人口数	老人化率	世帯数
東川口５丁目	2512	197	0.08	917
東川４丁目	2620	223	0.09	1075
東川口１丁目	1563	135	0.09	638
東川口３丁目	1042	66	0.06	463
東川口６丁目	2067	131	0.06	783

	戸建	共同住宅 1.2 階建	共同住宅 3-5 階建	共同住宅 6-10 階建	共同住宅 11 階建以上
東川口５丁目	194	105	85	387	0
東川口４丁目	316	208	322	32	0
東川口１丁目	108	80	128	112	45
東川口３丁目	65	74	200	21	0
東川口６丁目	198	106	230	97	0

上記の地域指標と空き巣発生率との散布図行列を以下に示す。なお、低層は 1.2 階建て、中層は 3-10 階建て、高層は 11 階建て以上とし、全世帯数で除することにより割合をデータとして用いる。

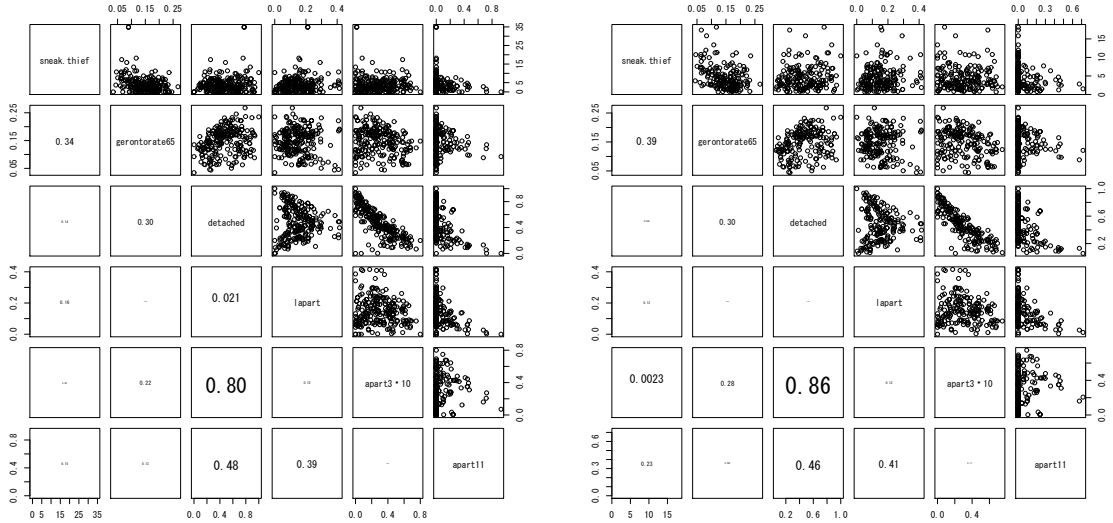

図 1: 散布図行列

この散布図行列から、空き巣発生率と関係がありそうな地域指標は老人化率である。そこで空き巣発生率を被説明変数、老人化率を説明変数として以下のモデルを構築しそれぞれの評価をする。

1 単回帰モデル

2 空間自己回帰モデル（SAR）

3 地理的加重回帰モデル（GWR）

各モデルの評価の基準として

1 残差二乗和

2 PSS

3 残差の Moran'I 統計量

各モデルの解析結果を次章で示す。

3.3 単回帰モデル

最小二乗法により、単回帰モデルの最小二乗推定量を求めた。

$$\hat{y}_i = 9.3 - 35.1x_i \tag{18}$$

残差 e_i の二乗和は

$$\sum_{i=1}^{n} e_i^2 = 4838.4 \tag{19}$$

残差 e_i の Moran'I は 0.13**。従って、単回帰モデルの誤差に独立性を仮定できない。そこで空間回帰モデルとして SAR、GWR モデルの推定を行った。

3.4 SAR モデル

$$\hat{y}_i = 6.7 - 26.71x_i + 0.35\sum_{j=1}^{n} w_{ij}y_j \tag{20}$$

残差 e_i の二乗和は

$$\sum_{i=1}^{n} e_i^2 = 4420.517 \tag{21}$$

残差 e_i の Moran'I は -0.03。従って、SAR モデルの誤差に独立性を仮定できる。

3.5 GWR モデル

$$\hat{y}_i = \hat{\beta_{0i}} + \hat{\beta_{1i}}x_i \tag{22}$$

残差 e_i の二乗和は

	min	mean	max
$\hat{\beta_{0i}}$	-0.23	6.04	26.14
$\hat{\beta_{1i}}$	-128.43	-13.90	19.87

$$\sum_{i=1}^{n} e_i^2 = 3197.92 \tag{23}$$

残差 e_i の Moran'I は -0.03。
従って、GWR モデルの誤差に独立性を仮定できる。

3.6　モデルの考察

誤差に集積性が示唆された単回帰モデルよりも、SAR モデルと GWR モデルがより良いモデルであるように思える。さらに言えば、SAR モデルよりも GWR モデルの方がより良いモデルである。単回帰モデルの老人化率の係数 $\hat{\beta}_1$ は負であり、卒業論文で用いた岡山市交番管轄ごとの犯罪データ解析と同じく、老人化率を空き巣の抑制力となっている。
興味深いのは、GWR の $\hat{\beta}_{1i}$ が正になる地区（下図の赤い地区）が見受けられたことである。この正となっている地区は老人化率の抑制の効果があるとはいえない。それは、高齢者の方々が監視役としての効果が薄いということである。このような、地区によって老人化率の抑制力が違うという空間的な変動を GWR モデルはとらえていることがわかる。
以下に GWR のパラメータの空間分布の図を示す。

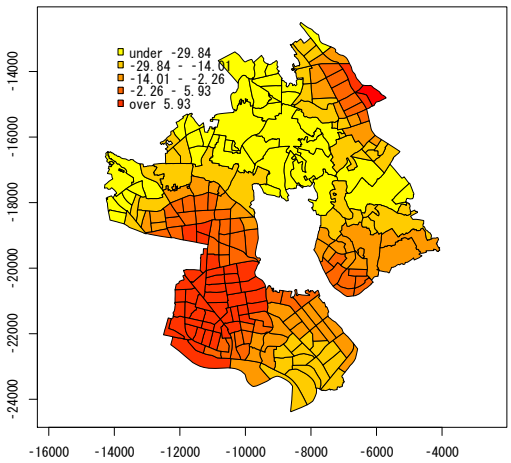

図 2: $\hat{\beta}_{1i}$

正となっている地区は南西部の横曽根地区、北東部の戸塚地区である。説明変数と被説明変数の空間的な変動が赤⇒黄へのグラデーションで表現されている。