数理解析研究所講究録
第1603巻 2008年 11-24

疾病地図と統計解析

国立保健医療科学院・技術評価部 高橋邦彦 (Kunihiko Takahashi)

Department of Technology Assessment and Biostatistics
National Institute of Public Health

1 はじめに

地域ごとの疾病状況を把握・検討するため空間疫学とよばれる研究が重要となってきている (Lawson(2006), Waller and Gotaway(2004), Elliott *et al.*(2000) など). 空間疫学では, 健康リスクを表わす症候・疾病・死亡の発生状況の地理的な格差・変動を記述するとともに, 人口統計学的要因, 環境要因, 行動要因, 社会経済学的要因, 遺伝的要因, 伝染性要因など疾病のリスクファクターの地理的変動を考慮に入れて, ランダムではない系統的な疾病の地理的変異を検出し, その要因の分析を行う比較的新しい学問である (丹後, 横山, 高橋 (2007)). 保健医療・公衆衛生分野などにおいて疾病に関する観察を行う場合, ひとつひとつの症例を個々に調べるだけではなく, 発生地点を空間的にとらえ, 地域全体としての状況把握も必要になる. このような空間データの統計解析を行うにあたって, まずはそのデータの分布の様子を把握することは, 最も基本的かつ重要なことである. そのための有用なツールとして疾病地図が用いられる.

疾病地図には大きくわけて "点データの地図" と "集計データの地図" の2つがある. とくに集計データの地図は日本における市区町村や二次医療圏, 都道府県単位など, また米国などでは州, 郡ごとに集計された疾病地図が広く利用されている.

ここではこの集計データの疾病地図に基づく統計解析について検討を行う.

2 死亡リスクの推定

2.1 標準化死亡比

一般に地域ごとの死亡数はポアソン分布に従うと仮定され, i 地域の基準集団に対する相対リスク (relative risk) を $\theta_i(>0)$ とする. いま対象としている地域が m 個の地域 (市区町村など) に分割されていると考え, i 地域の死亡数を d_i とおくと

$$d_i \sim \text{Poisson}(\theta_i e_i), \qquad (d_i \text{ と } d_j \ (i \neq j) \text{ は独立}) \tag{1}$$

とかける. ここで $e_i(>0)$ は i 地域の期待死亡数であり, i 地域の死亡リスク (危険度) が基準集団 (日本全国や解析対象地域全体など) と同じであるとしたときに, i 地域で観測が期待される死亡数である. 期待死亡数 e_i は通常, 住民の性・年齢構成などを考慮した人口に比例して定められる既知の定数である. もし i 地域が基準集団と同じ死亡リスクを持てば $\theta_i = 1$ であり, 基準集団より死亡リスクが大きければ $\theta_i > 1$ となる. このとき, θ の推定量として

$$\hat{\theta}_i = \frac{d_i}{e_i}, \qquad i = 1, 2, \cdots, m \tag{2}$$

が標準化死亡比 (standardized mortality ratio, SMR) として広く利用されている.

しかし, SMR はその推定誤差の大きさが期待死亡数, すなわちその地域の人口に影響され, 人口が多い地域と少ない地域でのリスクの推定値としての標準誤差が大きく異なってしまう. つまり人口サイズの異なる地域間でのリスクの比較には適していない指標となっている.

2.2 経験ベイズ推定量

SMR の問題を解決する方法として, 推定される $\hat{\theta}_i$ が極端に高いまたは低い値をもたないようにバラツキの大きさを制御する工夫の一つとして Bayes 推定が考えられる. ここでは, (1) の設定のもと, θ_i の事前分布として Gamma 分布を仮定する Poisson-Gamma モデルを考える. すなわち

$$\begin{aligned} d_i &\sim \text{Poisson}(\theta_i e_i) \\ \theta_i &\sim \text{Gamma}(\alpha, \beta) \end{aligned} \tag{3}$$

ただし $\text{Gamma}(\alpha, \beta)$ は平均 α/β, 分散 α/β^2 の Gamma 分布である. このとき θ_i の事後分布は Gamma 分布 $\text{Gamma}(\alpha + d_i, \beta + e_i)$ になる. このとき Gamma 分布のパラメータ α, β を d_i の周辺尤度から求めた最尤推定量 $\hat{\alpha}$, $\hat{\beta}$ を用いれば θ_i の事後分布の期待値から θ_i の経験ベイズ推定量は

$$\hat{\theta}_{i,\text{EB}} = \frac{\hat{\alpha} + d_i}{\hat{\beta} + e_i} \tag{4}$$

と求められる.

2.3 Poisson-Gamma モデルのフルベイズ推定量

Poisson-Gamma モデルの経験ベイズ推定では θ_i の従う Gamma 分布のパラメータ α, β を定数と考えている. 一方それらのハイパーパラメータも確率変数ととらえるフルベイズ法も提案されている.

$$\begin{aligned} d_i &\sim \text{Poisson}(\theta_i e_i) \\ \theta_i &\sim \text{Gamma}(\alpha, \beta) \end{aligned} \tag{5}$$

ここではハイパーパラメータの事前分布として無情報事前分布のひとつ, 期待値 20 をもつ指数分布

$$\alpha \sim \text{Exponential}(1/20)$$
$$\beta \sim \text{Exponential}(1/20)$$

を仮定する.

2.4 対数正規モデル

より柔軟なフルベイズモデルとして対数正規モデルが用いられる.

$$\begin{aligned} d_i &\sim \text{Poisson}(\theta_i e_i) \\ \log\theta_i &= \mu + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma_\varepsilon^2) \end{aligned} \tag{6}$$

ここではハイパーパラメータの事前分布として

$$\mu \sim N(0, 10^5)$$
$$1/\sigma_\varepsilon^2 \sim \text{Gamma}(0.5, 0.0005)$$

を仮定する. ただし ε_i は独立な (相関のない, 構造のない) 地域差を表わす変量効果を表わす. この計算には一般的には MCMC が用いられ, WinBUGS などを利用することができる (Lawson, Browne and Vidal Rodeiro(2003) など).

2.5 条件付自己回帰モデル

これまで紹介してきたモデルでは, 各地域の相対リスクは独立であるという仮定をおいて相対リスクの推定をしている. しかし, 地域間の距離が近ければ相対リスクは類似し, 遠ければ類似しないという相対リスクと地域間距離が負の相関を示すと考えることは自然であろう. この相関は空間相関, 空間依存性, 空間クラスタリングなどと呼ばれている. この空間相関を表現する変量効果を導入したモデルの 1 つとして Besag, York and Mollie (1991) の条件付自己回帰 (conditional autoregressive, CAR) モデルが有名である.

$$\begin{aligned} d_i &\sim \text{Poisson}(\theta_i e_i) \\ \log\theta_i &= \mu + \varepsilon + \phi_i \\ \varepsilon_i &\sim N(0, \sigma_\varepsilon^2) : \text{相関のない独立な地域差} \\ \phi_i|\phi_{j\neq i} &\sim N\left(\bar{\phi}_i, \frac{1}{m_i}\sigma_\phi^2\right) : \text{空間平滑化} \\ m_i &= i \text{ 地域の隣接地域の数} \\ \bar{\phi}_i &= i \text{ 地域に隣接する地域 } j \text{ での } \phi_j \text{ の平均} \end{aligned} \tag{7}$$

この ϕ_i の事前分布を CAR 事前分布という．ここではハイパーパラメータの事前分布として

$$\begin{aligned} \mu &\sim \text{一様分布 (improper prior)} \\ 1/\sigma_\varepsilon^2 &\sim \text{Gamma}(0.5, 0.0005) \\ 1/\sigma_\phi^2 &\sim \text{Gamma}(0.5, 0.0005) \end{aligned}$$

を仮定する．

2.6 Mixture モデル

さらに柔軟なモデルとして Mixture モデルが Lawson and Clark(2002) によって提案されている．

$$\begin{aligned} d_i &\sim \text{Poisson}(\theta_i e_i) \\ \log\theta_i &\sim \mu + \varepsilon_i + p_i\phi_i + (1-p_i)\psi_i \\ \varepsilon_i &\sim N(0, \sigma_\varepsilon^2) \\ \phi_i|\phi_{j\neq i} &\sim N\left(\bar{\phi}_i, \frac{1}{m_i}\sigma_\phi^2\right) \\ \pi(\psi_1, \cdots, \psi_m) &\propto \frac{1}{\sqrt{\lambda}} \exp\left(-\frac{1}{\lambda}\sum_{i\sim j}|\psi_i - \psi_j|\right) \\ p_i &\sim \text{Beta}(\alpha, \alpha) \end{aligned} \tag{8}$$

ここではハイパーパラメータの事前分布として

$$\begin{aligned} \mu &\sim \text{一様分布 (improper prior)} \\ 1/\sigma_\varepsilon^2 &\sim \text{Gamma}(0.5, 0.0005) \\ 1/\sigma_\phi^2 &\sim \text{Gamma}(0.5, 0.0005) \\ \alpha &= 0.5 \end{aligned}$$

を仮定する．

2.7 適用例

実際に上記のモデルを適用した疾病地図を比較する．日本における胆のうがんを含む胆道がんは，新潟県をトップとしてその周辺に高く発生しているといわれている (Yamamoto(2003))．そこで 1996〜2000 年の 5 年間における新潟県，福島県，山形県の市町村 ($m = 246$ 地域) ごとの男性の胆のうがんによる死亡について (i) SMR, (ii) Poisson-Gamma モデルの経験ベイズ推定, (iii) Poisson-Gamma モデルのフルベイズ推定, (iv) 対数正規モデル, (v) CAR モデル, (vi) Mixture モデルを計算し，その疾病地図を描いた (図 1〜6)．

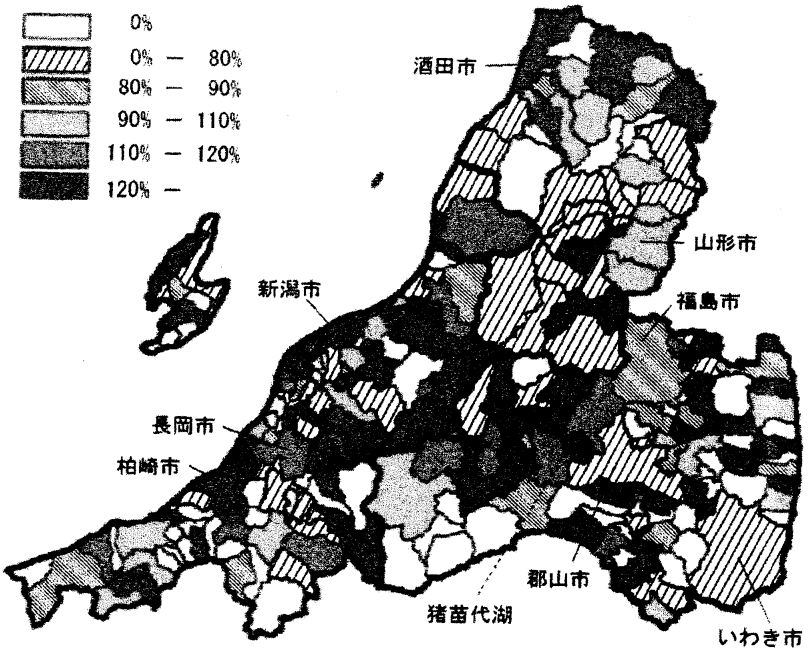


図 1. 1996〜2000 年新潟県, 福島県, 山形県の市町村ごとの男性の胆のうがんの SMR

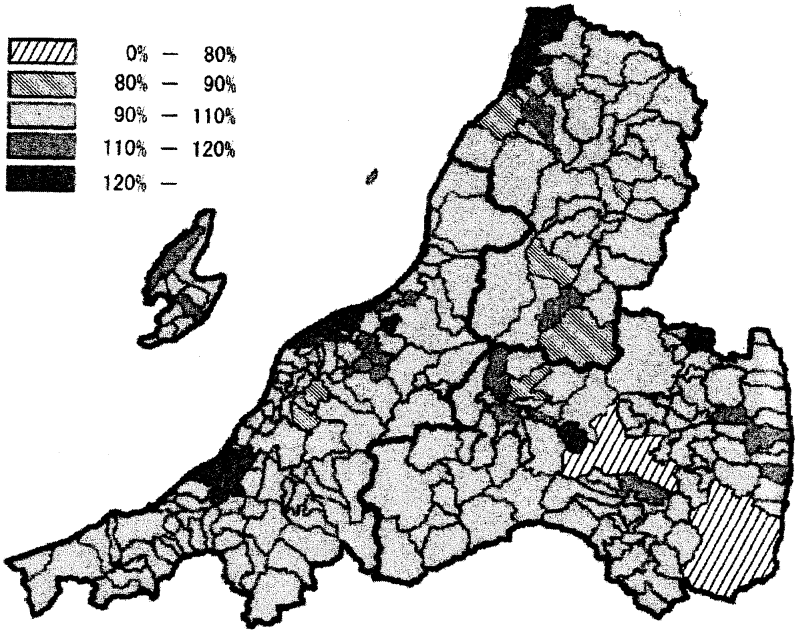


図 2. 1996〜2000 年新潟県, 福島県, 山形県の市町村ごとの男性の胆のうがんの Poisso-Gamma モデルの経験ベイズ推定値

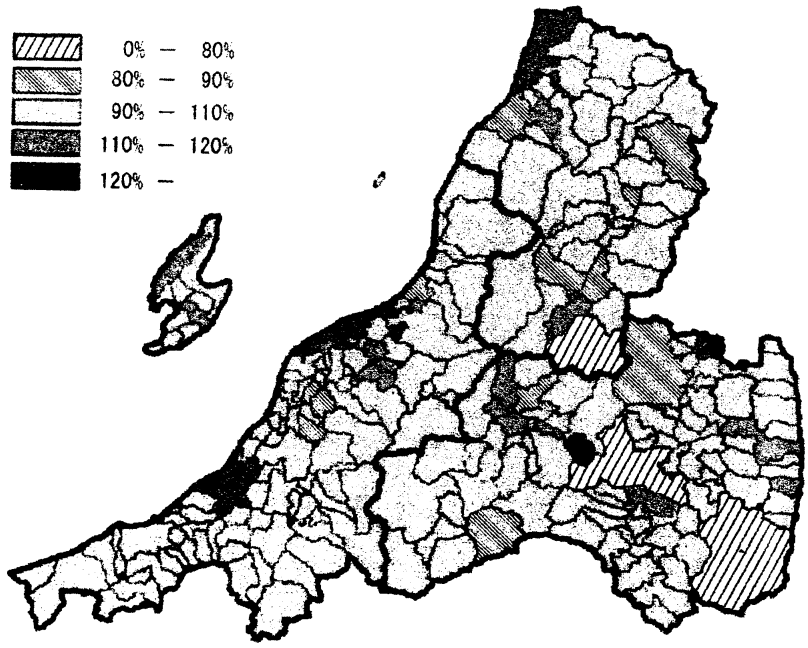


図 3. 1996～2000 年新潟県, 福島県, 山形県の市町村ごとの男性の胆のうがんの Poisso-Gamma モデルのフルベイズ推定値

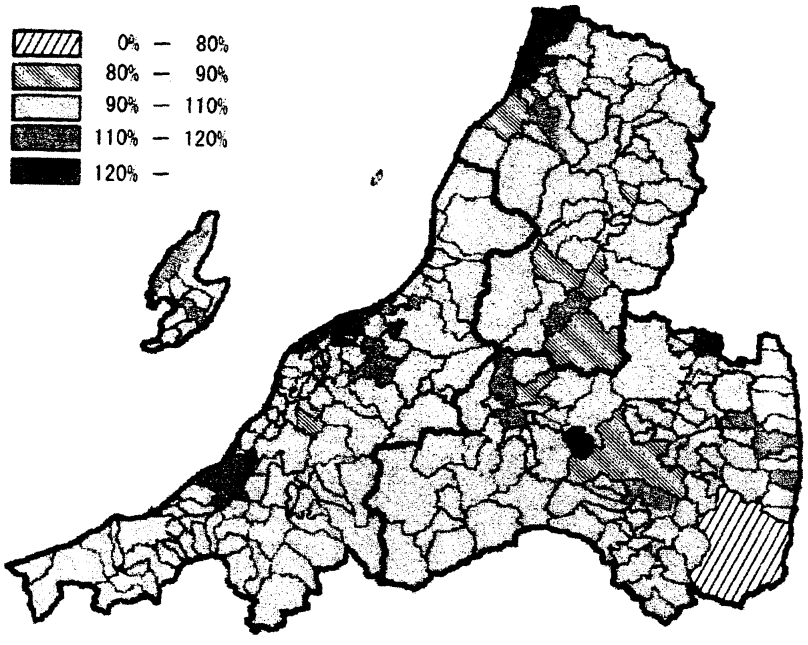


図 4. 1996～2000 年新潟県, 福島県, 山形県の市町村ごとの男性の胆のうがんの対数正規モデルのフルベイズ推定値

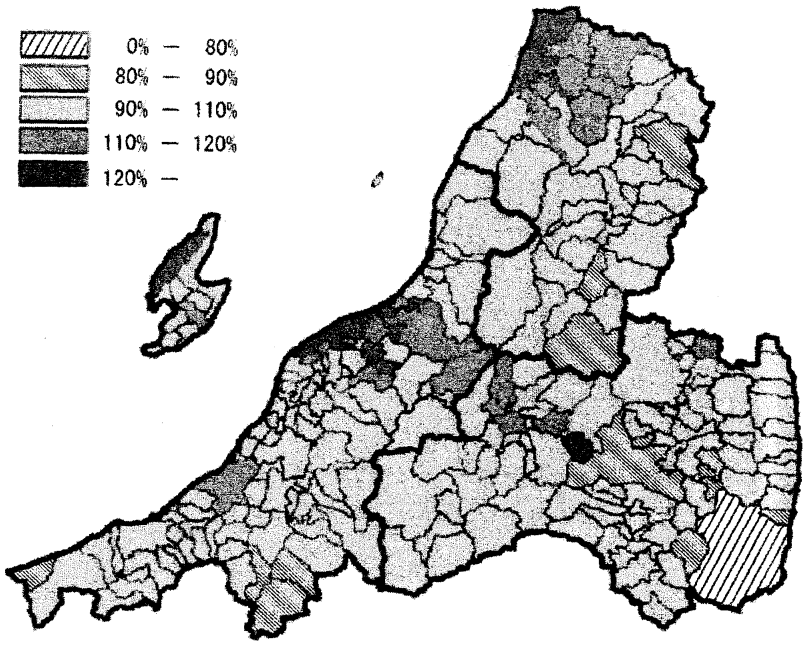


図 5. 1996〜2000 年新潟県, 福島県, 山形県の市町村ごとの男性の胆のうがんの CAR モデルのフルベイズ推定値

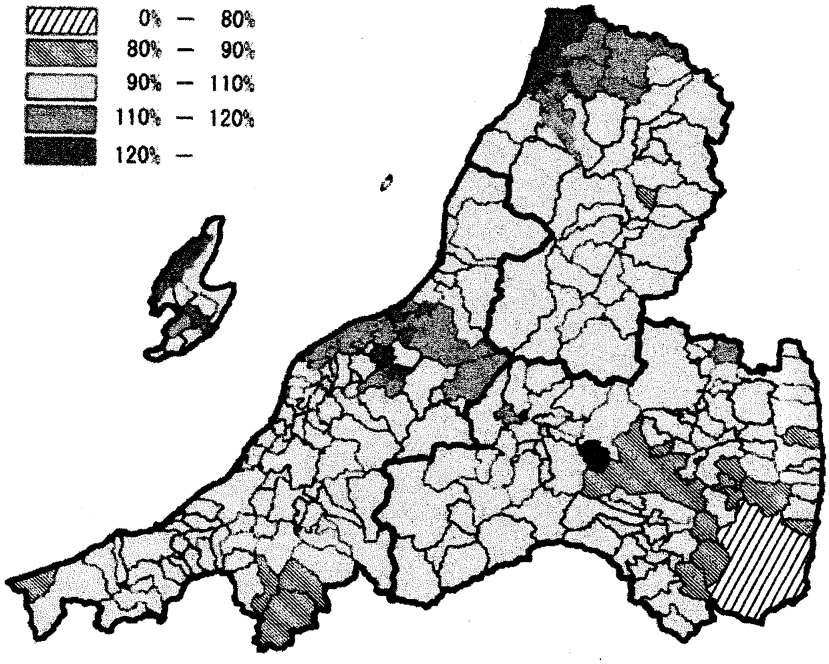


図 6. 1996〜2000 年新潟県, 福島県, 山形県の市町村ごとの男性の胆のうがんの Mixture モデルのフルベイズ推定値

3 スキャン統計量による疾病集積性の検討

これまでは，各市区町村の標準化死亡比（SMR）などの指標の値を色分けして視覚的に表した疾病地図について論じてきた．ところで，この疾病地図を観察すると，対象としている疾病のリスクの高い地域（もしくは低い地域）が，ある特定の地域に集中しているのではないかと思われることがある．もしこの疾病が集中して発生いるとすれば，その地域になんらかの原因があるかもしれないし，その疾病が流行性のものであるかもしれない．このように疾病の集積が観察された場合，集積地を中心に調査を行い，原因を特定したり対策を講じることが必要となるだろう．

しかし，疾病地図をみて，そこから集積地域を視覚的に見つけ出すだけでは説得力に欠けるであろう．これらの疾病地図だけでは、「どこかに集積しているか？それとも全体的にばらついているか？」の判断は難しい場合も少なくない．さらに集積しているとしても，どの範囲までかを客観的に判断することは難しいであろう．ここに，疾病集積性の有無を統計学的に客観的に決定する分析方法として集積性の検定が適用できる．さらに集積があると判定された場合,「集積地域はどこか？」を定める方法として Cluster Detection Test(CDT) が適用できる．この方法としていくつかの方法が提案されているが，それぞれ優れている点と同時に多少の弱点がある．

一般には, 死亡数, 患者数, 有病者数などさまざまなデータの解析を行うことができ, データに応じて Poisson model, Bernoulli model などを使い分けるが, ここではよく用いられる Poisson model に従って議論をすすめる．

いま, 解析を行う対象地域 G が m 個の region (市区町村, counties, zip codes など) に分割されているものとする．i 地区での case の数 (観測数)N_i が互いに独立に Poisson 分布

$$N_i \sim \text{Poisson}(\xi_i) \quad (i = 1, 2, \cdots, m)$$

に従うとし, その観測値を n_i とする．ただし, ξ_i は i 地区の人口に比例する値, または性・年齢などの共変量を調整した期待観測数とする．このとき集積地域 (cluster) の候補 window Z を考える．ただし Z は連結した region の集合であるとする．さらに window Z 内の case の数を確率変数 $N(Z)$, その観測値を $n(Z)$ であらわす．また window Z が集積地域でないという状況での $N(Z)$ の期待値を $\xi(Z)$ であらわし, さらに $N(G) = \xi(G)$ とする．ここで,

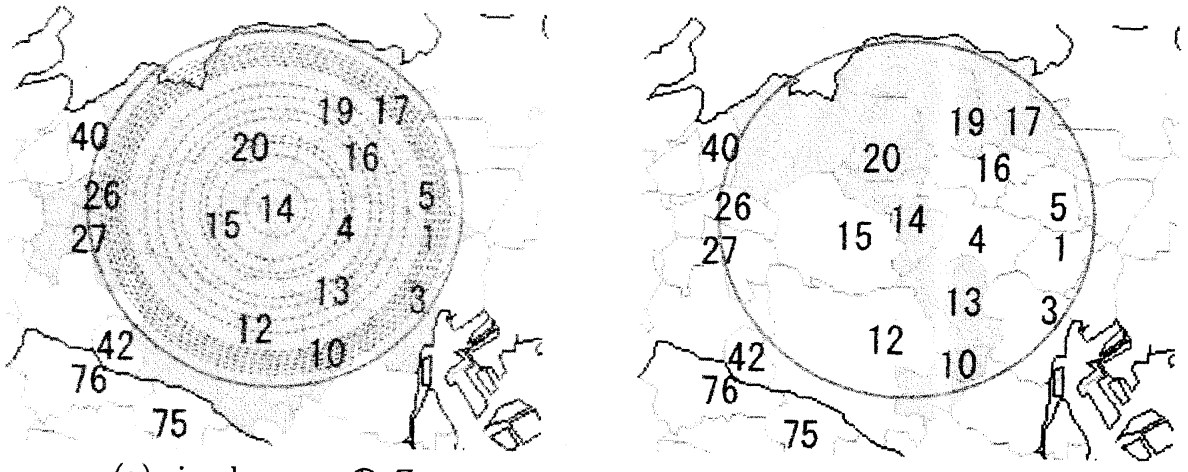
$$E(N(Z)) > \xi(Z) \tag{9}$$

となるような window Z を疾病の集積地域とする．つまり, 集積の有無は

$$\text{帰無仮説 } H_0 : E(N(Z)) = \xi(Z) \quad \text{for } {}^{\forall}Z \in \mathcal{Z}$$
$$\text{対立仮説 } H_1 : E(N(Z)) > \xi(Z) \quad \text{for } {}^{\exists}Z \in \mathcal{Z}$$

の仮説検定問題になる．そこで, このときの尤度比は

$$\lambda(Z) = \begin{cases} \left(\frac{n(Z)}{\xi(Z)}\right)^{n(Z)} \left(\frac{n(Z^c)}{\xi(Z^c)}\right)^{n(Z^c)} & (n(Z) > \xi(Z)) \\ 1 & (\text{その他}) \end{cases} \tag{10}$$



(a) circular scan の Z_{ik}　　(b) flexible scan の $Z_{ik(j)}$

図 7. $\mathcal{Z}$ の概念図. (a) $i = 14$ を中心に $k = 15$ の window Z_{ik} (b) $i = 14$ を中心に $K = 15$ としたときの $k = 6$ の 1 つの window $Z_{ik(j)}$

となり,

$$\lambda^* = \lambda(Z^*) = \max_{Z \in \mathcal{Z}} \lambda(Z)$$

のように最大尤度比 λ^* をとる window Z^* を most likely cluster(MLC) とし, これを cluster の候補と考える. ここで, この MLC が統計的に有意な集積性をもつかどうかの評価が必要となる. そのため帰無仮説のもとでの $\max_{Z \in \mathcal{Z}} \lambda(Z)$ の分布を使って有意性を見るが, 一般的には Monte Carlo 法を利用して数値的に求めた p 値によってその有意性が検討される.

ここで cluster を探し出す (scan する)window Z の全体集合 $\mathcal{Z}$ のとり方が重要であり, この違いによっていくつかの統計量が提案されている. Kulldorff(1995, 1997) は, 同心円状に, ある限界まで region を追加していく circular window の全体をとった circular scan statistic を提案した. $Z_{ik}(k = 1, 2, \cdots, K_i)$ を region i から近い順に, i 自身を含む k 個の region からなる集合とする. ただし各 i の座標はその region の代表点 1 点 (市区町村役場の所在地や人口重心など) であらわすものとする. このとき circular scan 法では, Z の全体集合として

$$\mathcal{Z}_1 = \{Z_{ik} \mid 1 \leq i \leq m,\ 1 \leq k \leq K_i\} \tag{11}$$

を考える. K_i としては cluster に含まれる最大距離や人口, 最大 region 数などが用いられる. この方法は簡便であるが, 明らかに円状の cluster しか同定できない (図 7(a)). そこで最近, 非円状の cluster も同定できるよう circular scan 法を拡張したいくつかの方法が提案されてきている. たとえば, Duczmal and Assunção(2004) の方法 (SA 法), Patil and Taillie(2004) の upper level set (ULS) 法の方法などが提案されてきている. これらの方法は非円状の window も同定できるようにしながら, また計算時間が大きくなりすぎないように工夫されている. しかし, これらの方法ではデータに応じて全体集合 $\mathcal{Z}$ が変化し, p

値を求めるための Monte Calro 計算の際にも毎回 scan する集合が変わってしまう．特に SA 法では同じデータを用いても結果の再現性が保証されない．また最大尤度比を求めるため，複雑な形状の大きな cluster を同定してしまう傾向がある．

そこで，Tango and Takahashi(2005) では，このような大きな cluster を防ぐよう制限された範囲内で非円状の cluster を同定する flexible scan statistic を提案した．まず region i を中心として i 自身を含み i から近い順に K 個の region からなる集合 Z_{iK} を定める．この Z_{iK} から，i を含み，連結している部分集合を考え，その全体 $\mathcal{Z}_2$ を考える．つまり Z_{iK} の中で i を含んで k 個の region からなる連結した window が j_{ik} 個あるとすると，Z の全体集合は

$$\mathcal{Z}_2 = \{Z_{ik(j)} \mid 1 \leq i \leq m,\ 1 \leq k \leq K,\ 1 \leq j \leq j_{ik}\} \tag{12}$$

とあらわされる (図 7(b))．

なお，Kulldorff の circular scan statistic はその解析を行うソフトウェア “SaTScan” が提供されており，また Tango & Takahashi の flexible scan statistic にはソフトウェア “FleXScan” が利用でき，それぞれ無料で公開されている．他の手法については現時点で一般ユーザーが利用できるソフトウェアは提供されていない．

そこで，前節で扱った 1996 年〜2000 年新潟県・福島県・山形県の市町村ごとの男性の胆のうがんの死亡について，この 3 県において，胆のうがんの死亡はどこかに集中（集積）して発生していると言えるだろうか？集積している場合，それはどの地区であろうか？例えば 3 県を基準とした SMR の疾病地図を見ると新潟市周辺から福島県西部の広い地域，あるいは山形県北部に SMR の高い地域が広がっている様子が観察できる．EBSMR, フルベイズ推定値を見ると高い地域が浮き彫りになり，CAR モデルのフルベイズ推定値では空間平滑化によってかなり滑らかな疾病地図ができあがり，新潟市周辺と酒田市周辺の 2 地域に高い地域が集積しているように観察される．しかしこれらの疾病地図だけでは，「どこかに集積しているか？それとも全体的にばらついているか？」の判断は難しい場合も少なくない．さらに集積しているとしても，どの範囲までかを客観的に判断することは難しいであろう．そこで circular scan と flexible scan を用いて 3 県の胆のうがんの集積性を検討してみる．

3.1 circular scan statisitc による解析

circular scan statistic によって同定された地域と検定結果を図 8 と表 1 に示す．表 1 の RR(relative risk) は SMR と同じ意味である．集積があると判定された地域は 2 箇所あった．もっとも集積していると判定された地域は山形県北部の酒田市周辺の 10 市町村であり，その SMR は 1.92, 集積の有意性は $p = 0.022$ であった．また 2 番目に高い集積性があると同定されたのは新潟市周辺の 16 市町村であり，その有意性は $p = 0.023$ であった．

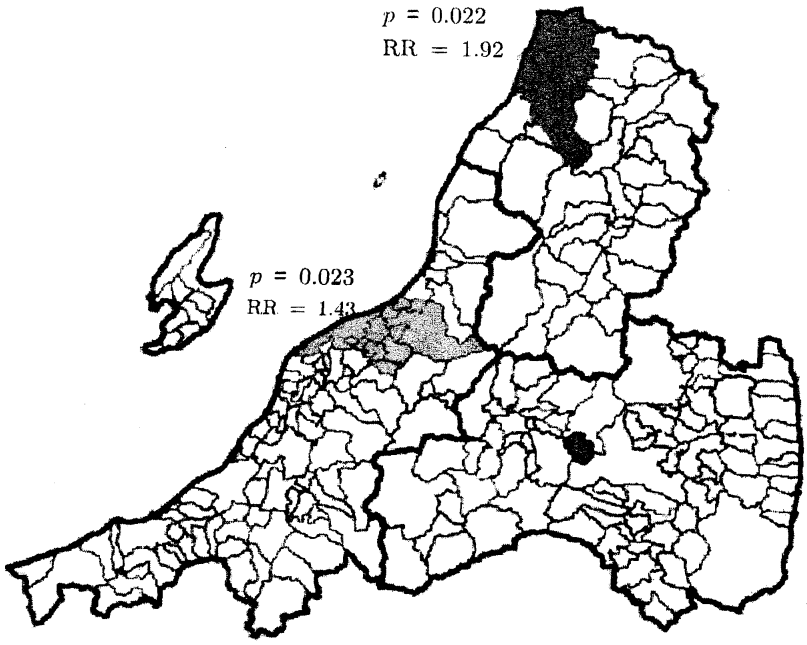


図 8. circular scan statistic によって同定された集積地域

表 1: 1996 年～2000 年新潟県・福島県・山形県の市町村ごとの男性の胆のうがんの死亡の集積性の検定結果

同定された地域	観測死亡数	期待死亡数	RR	p-value
circular scan statistic				
1 酒田市周辺の 10 市町村	46	23.97	1.92	0.022
2 新潟市周辺の 16 市町村	124	86.78	1.43	0.023
flexible scan statistic				
1 酒田市周辺の 8 市町村	46	21.05	2.19	0.022
2 新潟市周辺の 12 市町村	112	72.16	1.55	0.041

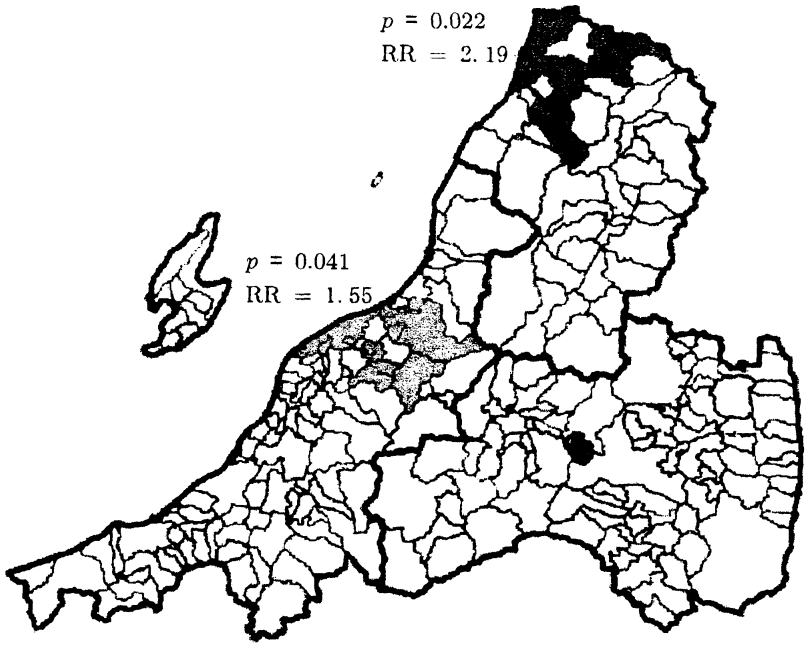


図 9. FleXScan によって同定された集積地域

3.2 flexible scan statistic による解析

同様に flexible scan statistic で解析を行った. 結果は図 9 と表 1 に示した. circular scan によって同定された地域とほぼ同様の地域が同定されたが, その中のいくつかの町村が集積地域から落ちていることが観察できる. 実際 circular scan よりも SMR が高い地域が同定されている.

4 おわりに

本論では, 死亡リスクをあらわす代表的指標である SMR, より複雑なモデルを用いた死亡リスク推定値の疾病地図を概観し, そこから疾病集積性の概念とその手法としてスキャン統計量に基づく 2 つの方法を論じた. これらの方法により, データの様子を視覚的に観察でき, さらに客観的に集積性の有意性の判定と, その集積地を同定することが出来る. しかし最初に述べたように, 疾病地図は空間データの様子を最初に観察するためのツールであり, また疾病集積性の検討にしても, その検定だけで強い疫学的な結論を出すことは難しいであろう. むしろ集積性が検出・同定されたことで「そこに何かあるのではないか？」「この疾病とこの地域に特有の環境要因等が関連しているのではないか？」というような次の研究へ続ける仮説を立てるための手段であり, その後の詳細調査や研究の必要性が示唆されると考えられる.

最近では平面および時間変化も考慮した空間での集積性の検定をもちいたサーベイランスの研究が注目されてきている．米国のいくつかのサーベイランスシステムなどでは実際に Kulldorff の方法を用いたサーベイランス解析が日々行われている．また Takahashi *et al.*(2008) によって flexible scan statistic を用いたサーベイランスのための集積性の検定法も提案されている．

一方で，集積性の検定においては，いかに精度良く集積地を同定できるかという問題も重要となっている．そのためには一般的な検出力では不十分であり，その評価指標として，Tango & Takahashi(2005) による bivariate power distribution や，Takahashi & Tango(2006) による extended power などが提案されている．これらの指標をもとに更に精度良く集積地を同定できる統計量の開発も今後の重要な課題となっている．

参考文献

Besag JE, York JC and Mollie A (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43**: 671–681.

Duczmal L and Assunção R (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* **45**:269–286.

Elliott P, Wakefield J, Best N and Briggs D(eds) (2000). *Spatial Epidemiology.* Oxford University Press.

Kulldorff M and Nagarwalla N (1995) Spatial disease clusters: detection and inference. *Statistics in Medicine* **14**: 799–810.

Kulldorff M (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods* **26**: 1481–1496.

Lawson AB (2006). *Statistical Methods in Spatial Epidemiology*(2nd ed.). Wiley.

Lawson AB, Browne WJ and Vidal Rodeiro CL (2003). *Disease Mapping with WinBUGS and MLwiN.* Wiley.

Lawson AB and Clark A (2002). Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine*, **21**: 359–370.

Patil GP and Taillie C (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* **11**:183–197.

Takahashi K, Kulldorff M, Tango T and Yih K (2008). A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics* **7**: 14.

Takahashi K and Tango T (2006). An extended power of cluster detection tests. *Statistics in Medicine* **25**: 841–852.

Tango T and Takahashi K (2005). A flexibly shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics* **4**: 11.

Waller LA and Gotway CA (2004). *Applied Spatial Statistics for Public Health Data.* Wiley.

丹後俊郎, 横山徹爾, 高橋邦彦 (2007). 空間疫学への招待. 朝倉書店.