特集：疾病地図の推定，疾病の集積性の検出に関する最近の進展

疾病地図と疾病集積性
—疾病指標の正しい解釈をめざして—

Disease mapping and spatial disease clustering
—Toward an appropriate interpretation and use of disease indices—

Toshiro TANGO

Mapping incidence and mortality from diseases such as cancer usually display relative rates in each region, as measured by a standardized mortality ratio, or age-sex adjusted mortality rate. Neither of these approaches is fully satisfactory in that they are heavily influenced by the heterogeneity of population size and also it cannot be used for the detection of disease clusters.

This paper reviews recent developments of statistical methods and models for disease mapping and the detection of spatial disease clustering and illustrate some methods proposed by the author with several real examples.

Key words : Empirical Bayes, Bayesian hierarchical model, Global test for clustering, Focused test for clustering, Monte Carlo Markov Chain

1．はじめに

公衆衛生の分野では，市区町村別の健康状況，疾病状況を比較検討するためにある疾患の年齢調整死亡率（有病率），標準化死亡比などを数区分に色分けして視覚的に表示した疾病地図がよく利用されている．また，ある疾患の年齢調整死亡率を被説明変数，市区町村毎の社会経済的指標，環境変数などを説明変数とした回帰分析などもよく行われている．しかし，日本では，これらの「日常的な行為」が実は統計学的に適切でないことはほとんど知られていない．その結果として，誤った（という自覚はない）解析・解釈が横行している．一方，欧米ではすでにその問題点が指摘され，問題解決への様々な方法論が提案されて実用化されている．

本論文では，これらの問題点を改めて紹介し，その解決に向けた方法論をEmpirical Bayes，MCMCを利用したFull Bayes，疾病の集積性の検出法，の三つに分けて最新の方法論とともに解説したい．

Division of Theoretical Epidemiology
Department of Epidemiology, The Institute of Public Health
4-6-1 Shirokanedai, Minato-ku, Tokyo 108, JAPAN
Tel : 03-3441-7111, ext.258
Fax : 03-3446-7164
E-mail : Tango@iph.go.jp

2．何が問題か？

まず，問題の本質を理解していただくために，図1(a)を見ていただきたい．Tsutakawa et al.(1985)，丹後(1988)が疾病地図の問題点を指摘するのに使用したMissouri州（男性，45－64歳，1972－1981年）の胃癌死亡率の市別データである．この図は，あたかも，死亡率が人口に反比例し，人口が減るにつれて死亡率が増加することを示している．「そんなばかな！」である．図1に示した「直線A」は

$$y=\frac{0}{\text{人口}}=0$$

であり，図1の「曲線B」は関数

$$y=\frac{1}{x}=\frac{1}{\log_{10}(\text{人口})}$$

をx軸を対数目盛りで描いたものである．つまり，単純な，誰でも計算できる死亡率

$$r=\frac{d}{n}\times 100{,}000 \quad (d：\text{死亡数}，\ n：\text{人口})$$

をそのまま使用している点が実は大きな落とし穴で，

> 各地域の人口の変動が大きいと，対象としているk個の地域毎に計算した率（r_1，r_2，…，r_k）が，地域間の死亡率の大きさを比較するのに適切な指標とならない

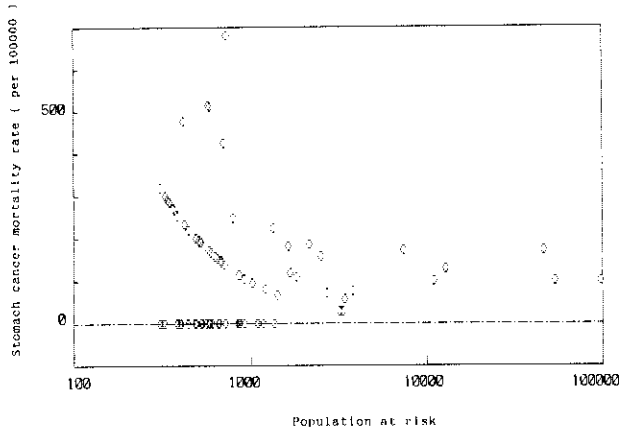

図1(a). Missouri州の市別の人口と胃がん死亡率(男性, 45-64歳, 1972-1981).

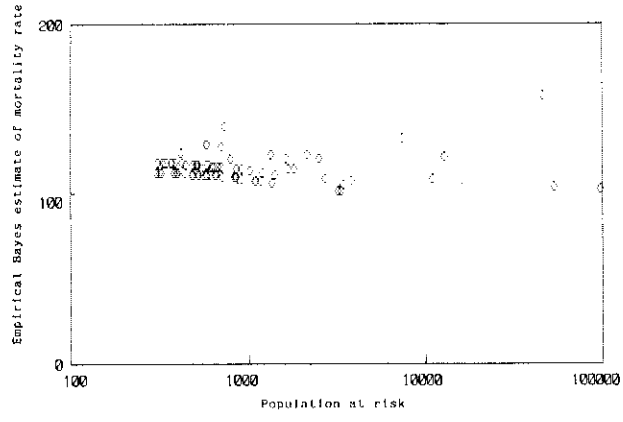

図1(b). Missouri州の市別の人口と胃がん死亡率のEmpirical Bayes推定値(丹後, 1988)

のである．何故だろうか？その理由を考えるために，次の4地域の観測死亡率を比較してみよう．

A地域：$r_A=1/5,000=20$（人口10万対，以下同様）
B地域：$r_B=10/50,000=20$
C地域：$r_C=100/500,000=20$
D地域：$r_D=1,000/5,000,000=20$

この4地域はいずれも，人口10万人当りで，20人と計算され，この値で比較する限り死亡状況は同程度であると判断される．A地域は人口5,000とD地域に比べると1,000分の1の少なさである．さて，A地域では，死亡が1人だけ観察されたため，死亡率が人口10万対で20と計算されたが，もし1人減って，0人となっていれば死亡率は

$$r_A=(1-1)/5,000=0\text{人}$$

となり，もし1人増えて死亡数が2人となれば死亡率は

$$r_A=(1+1)/5,000=40\text{人}$$

となる．すなわち，死亡数のわずかな数の増減で死亡率r_Aが大きくばらつき，不安定な，信頼性の乏しい指標となってしまう．これに対して，D地域では，死亡数が1人減っても，死亡率は

$$r_D=(1,000-1)/500\text{万人}=19.98\text{人}$$

一人増えても，死亡率は

$$r_D=(1,000+1)/500\text{万人}=20.02\text{人}$$

となり，いずれも，死亡率20人とほとんど同じであり，安定で,信頼性の高い指標であることが理解できるであろう．つまり，死亡率r_iは，人口の少ない地域では，わずかな死亡数の増減の影響が大きく反映され，不安定な指標となってしまうのである．このように，人口の大きさに起因する精度を有する死亡率で地図を作成する「行為」は，「k種類の精度の異なる物差しの測定結果を同じレベルで比較すること」と等価であり，サイエンスの世界では到底考えられない事である．日本の公衆衛生の分野ではそれがまかり通っている．

3．死亡率の精度

しかし，死亡率の精度と言うと，次のような反論がでるかもしれない．

> 疾病地図で問題にしている死亡率は，通常，各地域毎の全数調査（人口動態統計）で「計算」されたものであり，標本調査（random sampling）により「推定」された死亡率ではない．したがって，当該地域を母集団とした標本抽出によるサンプリング誤差は考えられない．つまり，計算された地域毎の死亡率$r_i=\frac{d_i}{n_i}$は，その地域の真の死亡率（母数）と考えられる．

さて，この反論に対しては次の様に解答することが可能である．

1．ある期間のある地域における死亡率がpであるとは，この地域の一人一人がこの期間で死亡する平均的確率がpであると考えられる．
2．一人，一人の死亡は互いに独立な確率現象と考えると，この期間での死亡数は確率的に変動する変量となり，観測死亡数はその実現値である．

具体的には，人口n人の地域で，この期間にd人死亡する確率は次の二項分布

$$P(d|p)=\binom{n}{d}p^d(1-p)^{n-d} \tag{1}$$

となる．死亡率pは一般に10万人当りで表示されるように，1より極めて小さいので，二項分布は次のポアッソン分布に近似される．

$$f(d|n,p)=\frac{(np)^d\exp(-np)}{d!} \tag{2}$$

このとき，$r=d/n$と計算される死亡率rの期待値と標準偏差は

$$E(r)=p \tag{3}$$

表1．高知県の市町村別人口，男性の結腸・直腸がんの死亡数，期待死亡数，SMR，Empirical Bayes SMR（今井，1998）．

地域	総人口（1987～1996）		結腸がん・直腸がん（男）			
					死亡指標	
	男	女	総死亡数	期待死亡数	SMR	EBSMR
高知県	3,868,896	4,345,571	632	632.0	100.0	100.0
高知市	1,476,788	1,692,957	225	189.6	118.7	114.7
室戸市	107,442	119,874	31	18.2	170.8	119.6
安芸市	110,621	124,533	16	18.8	85.0	90.8
南国市	226,985	247,962	28	35.1	79.8	86.8
土佐市	150,732	162,630	27	25.6	105.3	98.6
須崎市	145,739	152,767	18	23.6	76.3	86.8
中村市	166,894	187,313	29	27.0	107.3	99.6
宿毛市	121,362	135,319	20	19.6	102.3	96.8
土佐清水市	95,792	111,445	19	18.0	105.8	97.8
東洋町	20,101	22,415	6	4.1	147.6	99.2
奈半利町	20,674	23,955	8	4.4	183.0	103.5
田野町	16,880	19,307	5	3.4	149.0	98.5
安田町	19,098	20,911	2	3.7	53.8	90.0
北川村	7,914	9,022	1	2.0	49.0	91.3
馬路村	6,475	6,469	0	1.3	0.0	90.4
芸西村	20,745	24,657	0	4.0	0.0	84.4
赤岡町	17,412	19,966	7	2.8	246.5	105.0
香我美町	30,092	31,318	7	5.7	123.5	97.8
土佐山田町	105,541	120,294	12	18.5	64.9	83.9
野市町	67,610	75,795	6	10.5	57.3	85.5
夜須町	21,803	24,733	5	4.4	112.8	95.8
香北町	26,984	31,137	9	7.6	119.1	98.2
吉川村	10,017	10,990	1	1.7	58.8	92.2
物部村	17,005	19,234	4	4.7	84.9	92.7
本山町	24,777	26,905	3	5.7	53.0	88.2
大豊町	36,092	39,774	5	10.0	50.0	84.2
土佐山村	8,235	8,936	1	1.9	53.3	91.7
土佐町	6,762	6,769	0	1.5	0.0	90.1
大川村	25,999	28,584	3	6.1	49.1	87.3
本川村	3,440	3,714	0	0.9	0.0	91.4
伊野町	5,967	4,742	0	1.3	0.0	90.5
池川町	114,861	124,344	19	18.1	105.1	97.6
春野町	12,335	14,253	2	4.0	50.0	89.4
吾川村	70,556	80,093	11	12.6	87.3	92.1
吾北村	16,583	18,727	5	4.4	114.0	96.0
中土佐町	19,265	20,755	7	4.9	142.5	99.6
佐川町	37,011	41,955	9	7.6	118.5	98.1
越知町	73,603	82,151	14	14.1	99.5	95.4
窪川町	37,995	43,305	8	8.3	95.9	94.2
梼原町	75,698	86,043	9	16.0	56.4	82.2
大野見村	24,006	25,689	2	5.5	36.7	86.2
東津野村	8,489	9,478	1	2.3	43.9	90.8
葉山村	14,477	16,195	2	3.4	58.9	90.7
仁淀村	23,025	24,980	4	5.0	80.1	92.1
日高村	14,195	15,940	3	3.6	84.0	92.9
佐賀町	29,451	32,395	8	5.3	150.5	101.1
大正町	22,566	24,828	3	4.2	71.9	91.5
大方町	17,568	18,902	1	3.5	29.0	88.1
大月町	49,848	56,561	12	10.0	120.1	99.5
十和村	36,560	41,405	6	7.5	80.5	91.5
西土佐村	19,505	20,853	2	4.1	48.9	89.2
三原村	19,764	21,963	2	4.3	46.7	88.7
	9,557	10,329	4	2.2	184.6	98.9

$$S(r)=\sqrt{\frac{\tilde{p}}{n}} \tag{4}$$

となり，不偏推定量であるものの，そのバラツキは人口サイズの平方根に逆比例することがわかる．すなわち，人口の小さいところでは指標のバラツキが大きく，人口の大きいところではバラツキが小さくなるという「当り前」のことがわかる．バラツキが大きいということは，上述したように，本当は全国平均と比べて差がないのに，確率現象としての変動が大きいので，ある期間では高度に死亡率が大きくなったり（危険地域，赤で表示されることが多い），別の期間では極めて死亡率が低くなる（安全地域，青で表示）という見かけ上の変動で悩まされることになる．現実の疾病地図をみるとこのような現象は少くない．

4．年齢調整でも不十分

もちろん，地域間比較においては，単純な「率」ではなく，年齢・性などの分布の違いを調整した指標がよく利用される．代表的な指標として，直接法として知られる年齢調整死亡率 DAR（Directly age-Adjusted Rate）

$$\mathrm{DAR}_k=\sum_{j=1}^{J}\frac{N_j}{N_+}\frac{d_{kj}}{n_{kj}},\ (k=1,\ldots,K;\ j=1,\ldots,J) \tag{5}$$

ここで，

d_{kj}：k 地域，j 年齢階級の観察死亡数

n_{kj}：k 地域，j 年齢階級の人口（正確には人年）

Nj：標準人口の j 年齢階級の人口

$N_+=N_1+\cdots N_K$

がある．この指標は直接に観測死亡率 d_{kj}/n_{kj} を利用しているので，すでに述べた理由に加えて年齢階級の人口の分布の影響もあり，「地域比較の指標としては不適当な指標」である．その異常な性質の具体的例については福富（1980），丹後（1988）を参照されたい．Mosteller and Tukey (1977) はこのような現象を ill-determined rate と呼んでいる．これに対して，間接法と呼ばれる標準化死亡比 SMR（Standardized Mortality Ratio）

$$\mathrm{SMR}_k=\hat{\theta}_k=\frac{d_{k+}}{\sum_{j=1}^{J}n_{kj}P_{0j}}=\frac{d_{k+}}{e_k},\ (k=1,\ldots,K) \tag{6}$$

P_{0j}：標準人口における第 j 年齢階級の死亡率

e_k：k 地域の期待死亡数

は年齢調整死亡率ほどは人口の変動の影響は受けにくい（福富，1984）が，それでも

$$\mathrm{SMR}_k=\frac{1}{\sum_{j=1}^{J}q_{kj}P_{0j}}\frac{d_{k+}}{n_{k+}},\ (k=1,\ldots,K) \tag{7}$$

$$q_{kj}=\frac{n_{kj}}{n_{k+}},\ n_{k+}=n_{k1}+\cdots+n_{kJ}$$

と変形すればわかるように，地域全体の人口が相対的に小さければやはり粗死亡率（crude mortality rate）d_{k+}/n_{k+} の関数であるからやはり人口の影響は大きい．その例として図2(a)に高知県の53の市町村別男性の結腸・直腸がんの SMR（1987－1996）を利用した疾病地図を示す（今井，1998）．図3(a)には，人口を x 軸（常用対数）に SMR を y 軸にしてプロットした．人口の少ない市町村で SMR が高低に激しく変動していることがわかるだろう．人口の最大は高知市の1,476,788人，最小は大川村の3,440人であり，その比はほぼ430：1である．さて，SMR の最大値は赤岡町の250（死亡者数7人），最小値は死亡者0の5町村であった．これらのデータは表1に示した．このような図を見ると，論文等でよくみかける次のような回帰分析が如何に馬鹿げているか理解できるはずである．

$$\mathrm{SMR}_k=\beta_0+\beta_1 x_{1,k}+\cdots+\beta_m x_{m,k}+\text{誤差}$$

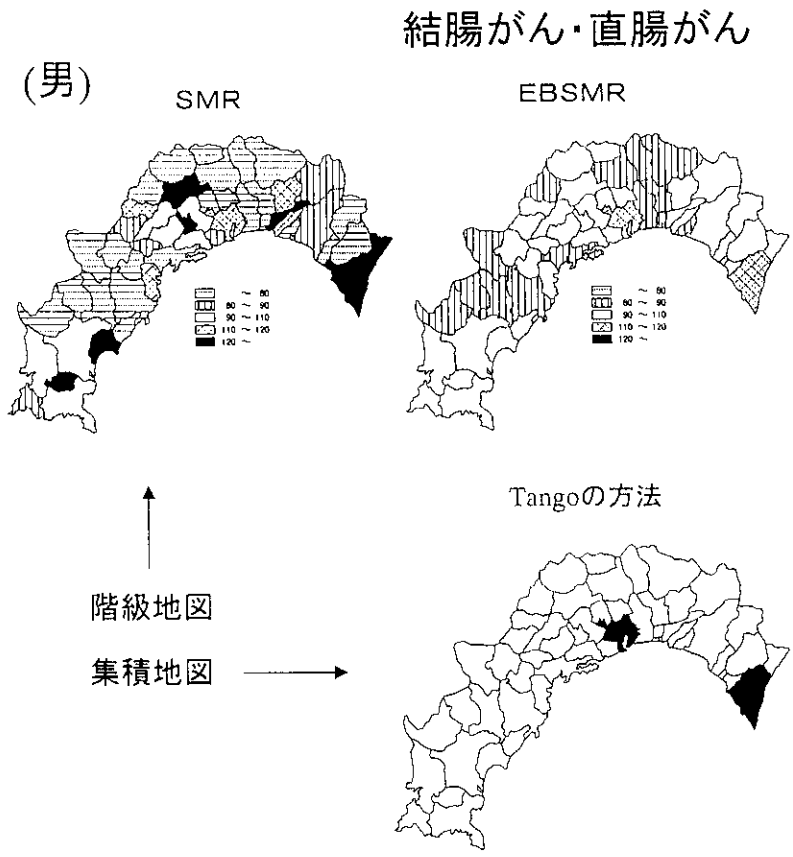

図2．1987－1996年の高知県の市町村別男性の結腸・直腸がんの疾病地図(a) SMR, (a) Empirical Bayes SMR, (c) Tango の集積性の検定で検出された市町村（今井, 1998）.

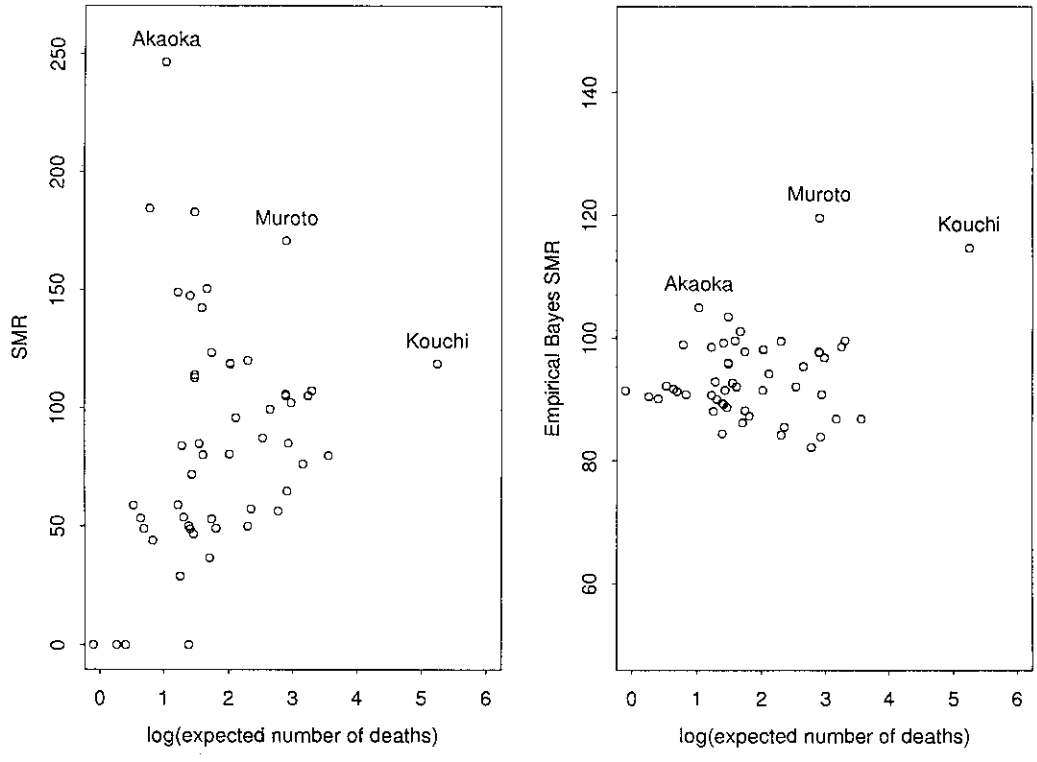

図3．期待死亡数と SMR の関連．(a) x 軸：log（期待死亡数），y 軸：SMR, (a) x 軸：log（期待死亡数），y 軸：Empirical Bayes SMR

このように，地域の比較を行うためには，「人口の大きさを調整」しなければならない。一つの簡単な方法は重み付回帰分析

$$\log SMR_k=\beta_0+\beta_1 x_{1,k}+\cdots+\beta_m x_{m,k}+\text{誤差},\quad Var(\log SMR_k)=1/d_{k+} \tag{8}$$

を実施することである（より正確な方法は5.2節参照）。もっとも，現在の行政区域を無視してでも，人口の変動を調整する一番簡単な方法は，各地域の人口がほぼ等しくなるように地域の再編成をしてから疾病地図を描くことであろう。例えば，日本全国での比較においては，二次医療圏の疾病地図であれば人口の変動はすくないので人口の影響は小さい．Selvin *et al*.（1988），Schulman *et al*.（1988）らは人工的に等しい人口（密度）をもつようなzoneを新しく作成する方法論を提案したが実用的ではなかった．

5．Bayesian approach

統計学的推測を大きく分類すると頻度論者流推測（Frequentist inference）とBayesian流推測（Bayesian inference）に分類される．伝統的な統計学的方法と言えば前者を指し，後者は統計学的推測に自分の信念（事前情報，事前分布）を付加する．ある地域の標準化死亡比θ_kの推定問題でも同様で，Frequentistであればθ_kは定数と考えるが，Bayesianではある分布に従う確率変数と考える．これをBayesの事前確率（分布）という．Bayesianでは，データをとる前の信念（事前分布）をデータをとることによって更新し，更新された信念（事後分布）の平均的な値（期待値，中央値など）で推測しようというものである．

ところで，死亡率には地域差があり，全体としてあるなめらかな連続分布にしたがうということは，決して不自然な考え方ではないだろう．したがって，地域毎の母標準化死亡比（$\theta_1,\ldots,\theta_K$）も，なめらかな連続分布（事前分布）にしたがうと考えられる．さて，ここで，「連続分布」を事前分布として仮定するということは，「推定される標準化死亡比$\hat{\theta}_k$が，極端に高いまたは低い値を持たないようにバラツキの大きさを制御する」ことを意味する．さて，事前分布を$g(\theta|\eta)$としよう．ここにηは分布を規定するパラメータである．観測死亡数d_{k+}は式(2)のように期待死亡数e_kをもつポアッソン分布．

$$f(d_{k+}|\theta,e_k)=\frac{(\theta e_k)^{d_{k+}}\exp(-\theta e_k)}{d_{k+}!} \tag{9}$$

に近似できるから，θ_kの事後分布はBayesの定理より

$$h(\theta_k|e_k,d_{k+},\eta)=\frac{g(\theta_k|\eta)f(d_{k+}|\theta_k,e_k)}{\int_0^\infty g(\theta|\eta)f(d_{k+}|\theta,e_k)d\theta} \tag{10}$$

と計算できる．したがって，SMRθの推測は，事後分布からの期待値

$$\hat{\theta}_k \Leftarrow E(\theta_k|e_k,d_{k+},\eta)=\int_0^\infty \theta h(\theta|e_k,d_{k+},\eta)d\theta \tag{11}$$

$$=\frac{\int_0^\infty \theta g(\theta|\eta)f(d_{k+}|\theta,e_k)d\theta}{\int_0^\infty g(\theta|\eta)f(d_{k+}|\theta,e_k)d\theta} \tag{12}$$

で行う．実は，この推定値$(\hat{\theta}_1,\ldots,\hat{\theta}_K)$は次の平均2乗誤差

$$E\left(\frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_k-\theta_k)^2\right) \tag{13}$$

を最小にするという意味でFrequentistの最尤推定量より優れている（例えば，Efron and Morris（1973）参照）．

5.1 Empirical Bayes

さて，Bayesian inferenceの問題は事前分布のパラメータηの設定である．一つのアプローチは，死亡数d_{k+}の周辺尤度

$$\prod_{k=1}^{K} Pr\{d_{k+}|e_k,\eta\}=\prod_{k=1}^{K}\int_0^\infty g(\theta|\eta)f(d_{k+}|\theta,e_k)d\theta \tag{14}$$

に基づく最尤推定法で推定する方法でEmpirical Bayes推定と呼ばれる．中でも，最も簡単で，かつ，解釈も容易な方法はガンマ分布

$$g(\theta|\alpha,\beta)=\frac{\alpha(\alpha\theta)^{\beta-1}\exp(-\alpha\theta)}{\Gamma(\beta)} \tag{15}$$

を仮定することである．ここで，$\eta=(\alpha,\beta)$（α：尺度パラメータ，β：形状パラメータ）

$$\text{平均値：}E(\theta)=\frac{\beta}{\alpha} \tag{16}$$

$$\text{分散：}\mathrm{Var}(\theta)=\frac{\beta}{\alpha^2} \tag{17}$$

である．なぜなら，この場合，式(10)より

$$h(\theta_k|e_k,d_{k+},\alpha,\beta)=\frac{(\alpha+e_k)[(\alpha+e_k)\theta_k]^{(\beta+d_{k+})-1}\exp\{-(\alpha+e_k)\theta\}}{\Gamma(\beta+d_{k+})}=g(\theta_k|\alpha+e_k,\beta+d_{k+}]) \tag{18}$$

と事後分布も尺度パラメータ$\alpha+e_k$，形状パラメータ$\beta+d_{k+}$のガンマ分布にしたがうからである．このように事後分布が事前分布と同じである場合，共役な事前分布（conjugate prior）と言う．Bayes推定値は

$$\hat{\theta}_{EB,k}=\frac{\hat{\beta}+d_{k+}}{\hat{\alpha}+e_k}=\frac{e_k}{\hat{\alpha}+e_k}\frac{d_{k+}}{e_k}+\frac{\hat{\alpha}}{\hat{\alpha}+e_k}\frac{\hat{\beta}}{\hat{\alpha}} \tag{19}$$

ここで，$\hat{\alpha}$，$\hat{\beta}$は推定値である．この式の形から$\hat{\theta}_{EB,k}$は

1．人口が大きい場合には（e_k→大），通常の標準化死亡比$\hat{\theta}_k=d_{k+}/e_k$に近づき，
2．人口が少ない場合には（e_k→小），地域全体の平均値$\hat{\beta}/\hat{\alpha}$に近づく．

という性質を持つ．さて，死亡数 d_{k+} の周辺尤度は負の二項分布（negative binomial distribution）

$$\Pr\{d_{k+}|e_k, \alpha, \beta\}=\frac{\Gamma(\beta+d_{k+})}{\Gamma(\beta)d_{k+}!}\left(\frac{\alpha}{\alpha+e_k}\right)^{\beta}\left(\frac{e_k}{\alpha+e_k}\right)^{d_{k+}} \quad (20)$$

となるので，(α, β) の最尤推定値は，次式を満たす．

$$\frac{\hat{\beta}}{\hat{\alpha}}=\frac{1}{K}\sum_{k=1}^{K}\frac{\hat{\beta}+d_{k+}}{\hat{\alpha}+e_k} \quad (21)$$

$$\sum_{k=1}^{K}\sum_{s=0}^{d_{k+}-1}\frac{1}{\beta+s}=\sum_{k=1}^{K}\log\left(1+\frac{e_k}{\hat{\alpha}}\right) \quad (22)$$

実際にはモーメント推定値を初期値とした Newton-Raphson 法で計算する．Appendix には Fortran program を参考のため掲載した．まず，図1(a)の Missouri 州のデータに適用したのが図1(b)である．この場合は SMR ではなく死亡率 $r_k=d_{k+}/n_{k+}$ であるから，上記の計算を

$$n_{k+} \leftarrow e_k$$

と置き換えて計算すれば良い．人口の少ないところはほとんど一定であることがわかる．高知県のデータに適用した結果が図2(b)，図3(b)である．最高の「120－」の階級に入る市町村が SMR では10もあったのに対し，Empirical Bayes 推定ではその様な地域は一つもなくなっている．また最低の「－80」の階級に属する市町村の数も「25→0」と激減している．Empirical Bayes 推定では，室戸市（EBSMR＝119.6），高知市（EBSMR＝114.7）の2つの市が高いが他は一塊で特に差は見られない．

5.2 Bayesian Hierarchical model

前節の Empirical Bayes 推定では人口の調整だけを考慮に入れたが，疾病指標に基づいた実際の解析では，地域毎の共変量を説明変数とした回帰分析，また，近接地域は類似の死亡率（有病率）であると仮定できる場合にはそれを考慮にいれた空間平滑化（spatial smoothing）のモデルを導入したり，といろいろな解析が必要となることがある．このような場合には，前節のアプローチでは計算上多くの困難が伴い実用的でなく，以下に説明する Bayesian 階層的ポアッソン回帰モデル（Bayesian hierarchical Poisson regression model）で議論するのが便利である．例えば，共変量 $(x_1, \ldots, x_m)$ による説明と，近接地域の類似性を考慮に入れたモデルの一つとして条件付き自己回帰モデル（Conditional autoregressive model）

$$\log E(d_{k+}) \stackrel{\text{def}}{=} \log\mu_k = \log e_k+\beta_1 x_{1,k}+\cdots+\beta_m x_{m,k}+\eta_k+\phi_k \quad (23)$$

d_{k+} ～Poisson 分布（期待値：μ）

$\eta_k \sim N(0, \sigma^2)$（：標準化死亡比の地域差）

$\phi_k|\phi_{h\neq k} \sim N\left(\bar{\phi}_k, \frac{1}{n_{h\sim k}}\tau^2\right)$：空間 smoothing

$n_{h\sim k}$＝地域 k の近接地域の数

$\bar{\phi}_k=\frac{1}{n_{h\sim k}}\sum_{h\sim k}\phi_h$

が考えられる．このモデルでは SMR が

$$\hat{SMR}_k \stackrel{\text{def}}{=} \hat{\theta}_k=\frac{\hat{\mu}_k}{e_k} \quad (24)$$

と推定される．この種の Bayes モデルの統計解析には Gibbs sampling に基づく MCMC（Markov Chain Monte Carlo）法（Gilks *et al.*, 1996）を利用した統計ソフト BUGS（Spiegelhalter *et al.*, 1995）を利用すると簡単である．疾病地図への具体例は，Mollie(1996)，Lawson *et al.*(1999)等を参照されたい．

6．疾病の集積性

前節までは，疾病地図の適切な解釈には人口のサイズ，他の共変量を調整する重要性とその方法論としての Bayesian approach を議論してきた．ところで，どんな推定値であれ，小さい順に並べれば必ず最低と最高が存在する．したがって，保健対策上重要なのは本当に健康状況が思わしくない地域はどの辺なのか？という疾病の地域集積性を検討する必要がある．なぜなら，限りある予算・資源を効率的に投入して地域ニーズに対応したきめ細かい保健対策の立案・実施を行うためには，対策が最も（本当に）必要とされる優先地域を選定する必要があるからである．そこで，本節では，

1．Global test
対象地域における疾病の地域集積性の有無を統計学的に検定し，有意な集積性が認められた場合にその地域はどこか？を教えてくれる方法

2．Focused test
ごみ焼却・危険物廃棄・原子力発電施設などの事前に定まっている地点の周辺に居住する地域住民に関連する疾病の集積性があるか否かを検討する方法

の二つについて解説する．疾病の集積性に関する研究は Marshall（1991）の総説にもあるように数多くの方法論が提案されてきた．Global test としては Turnbull et al.(1990), Cuzick and Edwards(1990), Besag and Newell (1991), Tango (1995), Kulldorff (1995)らの方法が代表的であり，Focused test としては Stone（1988），Besag and Newell（1991），Waller et al.（1992），(1995)，Lawson (1993)，Tango（1995）などがよく利用されている．ここではどちらも筆者の方法を紹介する．

6.1 Global test

検定仮説は

帰無仮説 H_0：調査地域に集積性はない

対立仮説 H_1：どこかに集積している

である．

ここでは，交絡因子として年齢だけを考えるが，複数の交絡因子の調整も同様である．集積性がないという帰無仮説の下では，j 年齢階級での観察死亡数

$$(d_{1j}, d_{2j}, \ldots, d_{Kj})$$

は周辺度数 d_{+j} が一定という条件の下では多項分布

$$\mathbf{p}_j^t=(p_{1j}, \ldots, p_{Kj}),\ p_{kj}=\frac{n_{kj}}{\sum_{k=1}^{K} n_{kj}}$$

$$=\frac{n_{kj}}{n_{+j}},\ \text{for } k=1, \ldots, K\ ;\ j=1, \ldots, J. \tag{25}$$

にしたがうサンプルサイズ d_{+j} の無作為標本と考えられる．とすると，k 地域の全年齢での期待死亡数 e_k は

$$e_k=\sum_{j=1}^{J} d_{+j}p_{kj}=\sum_{j=1}^{J} n_{kj}\frac{d_{+j}}{n_{+j}}=\sum_{j=1}^{J} n_{kj}\bar{P}_j$$

($\bar{P}_j$=調査対象地域全体での死亡率) (26)

となる．標準人口を利用して計算された期待死亡数(式(6))とは異なることに注意．ここで，

$$\mathbf{p}=\sum_{j=1}^{J}\frac{d_{+j}}{d_{++}}\mathbf{p}_j=(e_1, e_2, \ldots, e_K)^t/d_{++}=\mathbf{e}/d_{++} \tag{27}$$

とおいておく．このような情報が入手でき，疾病の頻度がrareである場合の集積性の検定法として，Tango (1995)は次の検定統計量を提案した．

$$C_\lambda=(\mathbf{r}-\mathbf{p})^t\mathrm{A}_\lambda(\mathbf{r}-\mathbf{p})$$

$$=\sum_{k=1}^{K}\left\{\sum_{h=1}^{K} a_{ij}(\lambda)(d_{k+}-e_k)(d_{h+}-e_h)\right\}/d_{++}^2 \tag{28}$$

$$=\sum_{k=1}^{K} U_k(\lambda) \tag{29}$$

ここに，

$$\mathbf{r}=(d_{1+}, d_{2+}, \ldots, d_{K+})^t/d_{++} \tag{30}$$

$\mathrm{A}_\lambda=(a_{kh}(\lambda))$：2地域 $(k,\ h)$ 間の近さの尺度の $K\times K$ 行列 (31)

$$a_{kh}(\lambda)=\exp\left\{-4\left(\frac{d_{kh}}{\lambda}\right)^2\right\} \tag{32}$$

d_{kh}=2地域 $(k,\ h)$ 間の距離：人口中心点の緯度，経度から計算 (33)

である．ここで，パラメータλは，クラスター(集積が見られる地域群)の大きさ（ほぼ最大距離）の尺度であり，それ以上の距離にある任意の二つの地域はクラスターとは考えないモデルである．したがって，λを小さく設定すれば大きなクラスターは検出力が低く，反対にλを大きく設定すれば小さなクラスターは検出力が低くなる．実際，事前に存在するクラスター大きさを予想できるわけがなく（データを見た後でクラスターの大きさを見積もって検定を適用することは事前の選択バイアスによる検定の誤用である），したがって，λの値をいく通りかに変えて適用することになるが，ここに検定の多重性が問題となる．この問題はTangoの方法ばかりでなく，ほとんどすべての提案された方法の問題点である．この問題を回避するためにTango (1998, 1999a)はλを連続的に動かして，λの関数としてのプロファイルp値の曲線を計算しその最小値$\mathrm{P}_{\min}$を検定統計量とすることを提案した．

$$P_{min}=\min_{\lambda}\Pr\{C_\lambda>c_\lambda|H_0,\ \lambda\}$$

$$=\Pr\{\mathrm{C}_\lambda>\mathrm{c}_\lambda|H_0,\ \lambda=\lambda^*\} \tag{34}$$

ここにc_λはあるλに対する式(28)の統計量の実現値であり，λ^*が最小値を達成する値である．実際の計算にはλ小刻みに変化させて最小値を探す一次元探索法で簡単に計算できる．P_{min}の帰無仮説の下での分布はMonte Carloシミュレーションにより計算する．なお，λは

$$0<\lambda\le\frac{d_{max}}{4}\ (d_{max}=\text{調査地域間の最大距離}) \tag{35}$$

の範囲で変化させれば十分であろう．

1．帰無仮説の下でのP_{min}の分布の推定

以下の手順をM回繰り返す（例，M=9999）．

Step 0．rep ←---0（rep 繰り返し数）

Step 1．rep ←— $rep+1$

Step 2．i ←—0

Step 3．i ←— $i+1$

Step 4．Computer乱数を利用して，年齢階級毎に帰無仮説の下で式(25)で定義されている多項分布にしたがう死亡数を発生させ，式(26)より期待死亡数を計算する．

Step 5．λ←—$(i-1)w$（wは刻みの最小単位（距離））

式(28)を計算する．その値が$C_\lambda=c_\lambda$であるとき，p値を次の近似式で計算でする（Tango, 1990）．

$$p(\lambda)=\Pr\{C_\lambda>c_\lambda|H_0, \lambda\}$$

$$\approx\Pr\left\{\chi_\nu^2>\nu+\sqrt{2\nu}\left(\frac{c_\lambda-E(C_\lambda)}{\sqrt{Var(C_\lambda)}}\right)|\lambda\right\} \tag{36}$$

ここに，χ_ν^2は自由度νのχ^2分布にしたがう確率変数であり，

$$E(d_{++}C_\lambda)=tr(\mathrm{A}_\lambda\mathrm{V}) \tag{37}$$

$$Var(d_{++}C_\lambda)=2tr(\mathrm{A}_\lambda\mathrm{V})^2 \tag{38}$$

$$\nu=8/\{\sqrt{\beta_1(C_\lambda)}\}^2 \tag{39}$$

$$\sqrt{\beta_1(C_\lambda)}=2\sqrt{2}\ tr(A_\lambda\mathrm{V})^3/\{tr(A_\lambda\mathrm{V})^2\}^{1.5} \tag{40}$$

$$\mathrm{V}=\sum_{j=1}^{J}\frac{d_{+j}}{d_{++}}\{diag(\mathbf{p})-\mathbf{p}\mathbf{p}^t\} \tag{41}$$

である．ここに$diag(\mathrm{p})$はベクトルpに基づく対角行列である．

Step 6．If$\lambda=d_{max}/4$ then goto Step 7, otherwise, goto Step 3.

Step 7．Q_{rep}←— $\min_\lambda p(\lambda)$

Step 8．If rep=M then goto Step 9, otherwise, goto Step 1.

Step 9．$\{Q_{rep},\ rep=1, 2, \ldots, M\}$を小さい順に並べた値を$\{Q_{(i)}, i=1, 2, \ldots, M\}$とする．これが帰無仮説での$P_{min}$のシミュレートされた分布である．

2．集積性の検定のp値の計算

Step 10．実際のデータで上記の手続きのStep 2, 3,

5, 6, 7を計算する．計算されたp値の最小値をここではQ^*としよう．その時のλの値をλ^*とする．

Step 11．集積性の検定のλの選択による多重性を調整したp値を

$$P_{min}=\frac{\#\{i: Q_{(i)}\geq Q^*\}+1}{M+1} \tag{42}$$

で計算する．

もし，有意な集積性が認められた場合には，クラスターの中心として（最も）疑われる地域は

$$U_k(\lambda^*)=\sum_{h=1}^{K} a_{kh}(\lambda^*)(d_{k+}-e_k)(d_{h+}-e_h)\}/d_{++}^2$$

または，

$$\frac{U_k(\lambda^*)}{C_{\lambda^*}}\times 100(\%)：k\text{ 地域の寄与率} \tag{43}$$

の値が他に比べて，大きく飛び離れていることが期待される．

さて，高知県の表1のデータに適用してみよう．Windows上で統計ソフトS-PLUSを利用して解析した結果を図4に示す．画面の右側はコマンドの操作画面と計算結果の表示画面であるが，画面の左側に二つの図が示されている．左の図はx軸をλにしたプロファイルp値であり，λの値が最小のときにp値が最小値をとっている．つまり，クラスターが複数の隣接地域で発生しているのではなく，散発的に発生している可能性を示唆している．図上にも記載されているが，式(42)による調整されたp値は$P_{min}=0.006$であり，高度に有意な集積性が見られた．右の図は，式(43)の各地域（region IDがx軸）の寄与率（%）を表示する図であるが，region ID＝1（高知市）が断然トップで，次にregion ID＝2（室戸市）もやはり飛び離れている．この結果は図2(c)に示す通りであり，図3(b)のEmpirical Bayes推定の結果と一致している．

6.2 Focused test

今，ある地域k^*が問題にしている危険施設のある地域であるとしよう．すると，その地域の周辺に疾病集積性があるか否かを検討する検定統計量が

$$C_{F:\lambda}=\mathbf{w}'\mathrm{A}_\lambda(\mathbf{r}-\mathbf{p}) \tag{44}$$

で定義できる（Tango, 1995）．ここに

$$w_{k^*}=1；w_{k\neq k^*}=0(k, h=1, 2, \ldots, K)$$

である．もし，危険施設が数地域に点在している場合にはそれぞれの地域で$w_k=1$とすればよい．帰無仮説の下で

$$Z=\frac{C_{F:\lambda}}{\sqrt{\mathrm{Var}(C_{F:\lambda})}}=\frac{C_{F:\lambda}}{\sqrt{\mathbf{w}'\mathrm{A}_\lambda\mathrm{V}\mathrm{A}_\lambda\mathbf{w}}}\sim N(0, 1) \tag{45}$$

となる．したがって，式(34), (35)と同様の手続きで集積性の検定のp値が次式で計算できる（Tango, 1999b）．

$$P_{F:min}=\min_{\lambda}\Pr\{C_{F:\lambda}>c_{F:\lambda}\,|\,H_0, \lambda\} \tag{46}$$

ここでは，方法の紹介だけにとどめるが，Focused testの

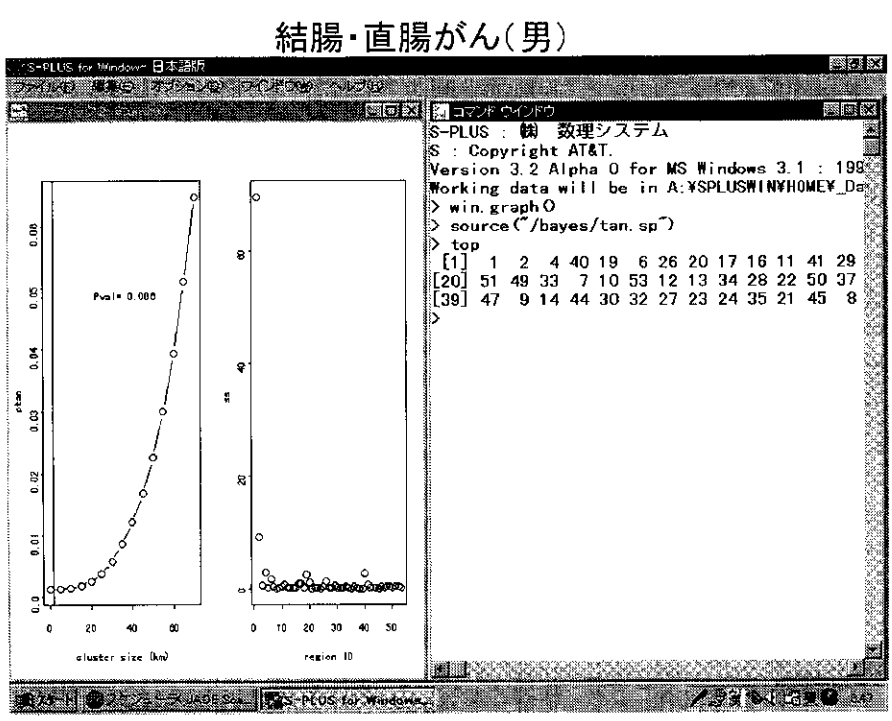

図4．Tangoの集積性の検定．Windows上で統計ソフトS-PLUSを利用して解析している画面．画面の右側はコマンドの操作画面と計算結果の表示画面．画面の左側に二つの図が示されている．左の図はx軸をλにしたプロファイルp値であり，右の図は，式(43)の各地域（region IDがx軸）の寄与率（%）を表示する図である，region ID＝1（高知市）が断然トップで，次にregion ID＝2（室戸市）もやはり飛び離れている．

Appendix. Empirical Bayes SMR の Fortran プログラム

```
C              Fortran Subroutine Program of
C                Empirical Bayes Estimation
C    Maximum Likelihood Estimator for Poisson-Gamma Model
C      < Input >
C          NK: Number of regions
C          P : Expected number of deaths
C          D : Observed number of deaths
C        ALP: initial value of alpha (ex. moment estimate)
C        BET: initial value of beta (ex. moment estimate)
C      < output >
C        ALP: estimated value of alpha
C        BET: estimated value of beta
C        ES : Empirical Bayes estimates of SMR or O/E ratio.
C        LM : =0 (converged), =1 ( Not converged)
C             In the case of non-convergence, initial values of
C             ALP and BET are returned.
C
      SUBROUTINE COMP(NK,P,D,ALP,BET,LM)
      DIMENSION P(200),D(200)
      LM=0
      ALP0=ALP
      BET0=BET
      EPS=0.0001
      LMAX=30
      II=0
  303 II=II+1
      IF(II .GT. LMAX)THEN
        LM=1
        ALP=ALP0
        BET=BET0
        RETURN
      ENDIF
      F1=0
      F2=0
      F11=0
      F12=0
      F22=0
      DO 300 I=1,NK
      NN=D(I)-1
    S=0
    T=0
      DO 310 M=0,NN
      XM=M
      S=S+1/(BET+XM)
      T=T+1/(BET+XM)**2
  310 CONTINUE
      F1=F1+BET/ALP - (BET+D(I))/(ALP+P(I))
      F2=F2+ALOG(ALP/(ALP+P(I))) +  S
      F11=F11-BET/ALP/ALP + (BET+D(I))/(ALP+P(I))**2
      F12=F12+1/ALP-1/(ALP+P(I))
      F22=F22-T
  300 CONTINUE
      DEL=F11*F22-F12*F12
      ALP2=ALP-(F1*F22-F2*F12)/DEL
      BET2=BET-(F2*F11-F1*F12)/DEL
      DA=ABS((ALP2-ALP)/ALP2)
      DB=ABS((BET2-BET)/BET2)
      IF((DA .LE. EPS) .AND. (DB .LE. EPS))GOTO 330
      ALP=ALP2
      BET=BET2
      GOTO 303
  330 ALP=ALP2
      BET=BET2
      RETURN
      END
```

具体的適用事例については，Lawson(1993)，Waller *et al.* (1992, 1995) などを参照されたい．

7．考　　察

本論文では，疾病地図を描く場合には人口の調整をする Bayesian approach が重要であることを示し，それとともに，疾病の集積性を検討する方法論を合わせて適用することにより疾病対策上有用な情報が得られることを強調した．ここでは紹介しなかったが，Global test の有効な方法の一つに Kulldorff (1995, 1997) の方法がある，この方法は，最も集積度が大きい思われる cluster (most likely cluster) を推定する方法である．Tango の方法との比較では（Tango and Kulldorff, 1998）真の cluster の個数が一つで，かつ，人口の少ない地域 (rural area) で発生している場合に検出力が高い．Windows 95上で稼動するソフト (Kulldorff, 1996) が彼から入手可能である．Tango の方法 (1999b) については，統計ソフト S-Plus 上で稼動するプログラムが利用できる．

本論文で述べた疾病の集積性の検討方法は市区町村毎の頻度データを利用する場合であるが，患者の住所情報が入手可能であれば，より細かい精度の高い集積性の検討が可能である．

最近は，computer の格段の進歩により，欧米では，computer をフルに利用した方法，5.2節で紹介した MCMC 法に基づく Bayesian hierachical model などの適用が疾病地図に限らずいろいろな医学，公衆衛生学分野の統計解析において盛んになってきている．日本の医学，公衆衛生分野においてもそうなる時代はそう遅くはないだろう．ただ，日本の医学系の研究者に，本論文で紹介した程度の数理を理解できる研究者が育たない限りこの種の情報化の分野で日本は後進国を余儀なくされるかもしれない．

参考文献

(1) Anderson, N.H. and Titterington, D.M. 'Some methods for investigating spatial clustering, with epidemiological applications', *Journal of the Royal Statistical Society, Series A*, 160, 87-105 (1997).

(2) Besag, J. and Newell, J. "The detection of clusters in rare diseases", *Journal of Royal Statistical Society, Series A*, 154, 143-155 (1991).

(3) Clayton, E. and Kaldor, J. 'Empirical Bayes estimates of age-standardized relative risks for use in disease mapping', *Biometrics*, 43, 671-681 (1987).

(4) Cuzick, J. and Edwards, R. 'Spatial clustering for inhomogeneous populations (with discussion)'. *Journal of the Royal Statistical Society, Series B*, 52, 73-104 (1990).

(5) Efron, B. and Morris, C. Stein's estimation rule and its competitors- an empirical Bayes approach. *J. Amer. Statist. Assoc.*, 68, 117-130 (1973).

(6) Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., *Markov Chain Monte Carlo In Practice*, Chapman & Hall, London, (1996).

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

(8) Kulldorff, M. 'A Spatial scan statistic', *Communications in Statistical Theory and Methods*, 26, 1481-1496 (1997).

(9) Kulldorff M., Rand, K., and Williams, G. SaTScan, version 1.0, program for the space and time scan statistic. Bethesda, MD: National Cancer Institute, 1996.

(10) Lawson, A.B. 'On the analysis of mortality events associated with a prespecified fixed point'. *Journal of the Royal Statistical Society, Series A*, 156, 363-377 (1993).
(11) Lawson, A. *et al. Advanced Methods of Disease Mapping and Risk Assessment for Public Health Decision Making*, Wiley, In Press (1999)
(12) Marshal R.C. "A review of the statistical analysis of spatial patterns of disease", *Journal of Royal Statistical Society, Series A*, 154, 421-441 (1991).
(13) Mollie, A. 'Bayesian mapping of disease', in *Markov Chain Monte Calro in Practice* (Eds. by Gilks, W., Richardson, S. and Spiegelhalter, D.), 359-379, Chapman and Hall, (1996).
(14) Mosteller F. and Tukey J.W. *Data Analysis and Regression*, Addison-Wesley, (1977).
(15) Selvin, S., Merrill, D.,Schulman, J. *et al.* 'Transformed map to investigate clusters of diseases', *Social Science and Medicine*, 26, 215-221 (1988).
(16) Schulman, J., Selvin, S. and Merrill, D. 'Density equalized map projections: a method for analyzing clusters around a fixed point', *Statistics in Medicine*, 7, 491-505 (1988).
(17) Spiegelhalter, D.J., Thomas, A., Best, N. and Gilks, W. R. Bugs: Bayesian Inference using Gibbs sampling, version 0.50. Technical Report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University, (1995).
(18) Stone, R.A. 'Investigation of excess environmental risks around putative sources: statistical problems and proposed test', *Statistics in Medicine*, 7, 649-660 (1988).
(19) Tango, T. 'The detection of disease clustering in time', *Biometrics*, 40, 15-26 (1984).
(20) Tango, T. 'Asymptotic distribution of an index for diesease clustering', *Biometrics*, 46, 351-357 (1990).
(21) Tango, T. 'A class of tests for detecting 'general' and ' forcused' clustering of rare diseases', *Statistics in Medicine*, 14, 2323-2334 (1995).
(22) Tango, T. and Kulldorff, M. 'Detection of spatial clusters of disease: power comparison of general tests', *Nineteenth International Society for Clinical Biostatistics Meeting*, Invited paper, August, Dundee (1998).
(23) Tango, T. 'Comparison of general tests for disease clustering', in *Advanced Methods of Disease Mapping and Risk Assessment for Public Health Decision Making*, (Lawson et al. eds), Wiley, In Press (1999a).
(24) Tango, T. 'A test for spatial disease clustering adjusted for multiple testing'. *Statistics in Medicin*, 18, In press (1999b).
(25) Turnbull, B.W., Iwano, E.J., Burnett, W.S., Howe, H.L. and Clark, L.C. 'Monitoring for clusters of disease: application to leukemia incidence in upstate New York', *American Journal of Epidemiology*, 132, S136-143 (1990).
(26) Tsutakawa, R.K., Shoop, G.L. and Marienfeld, C.J. ' Empirical Bayes estimation of cancer mortality rates', *Statistics in Medicine*, 4, 201-212 (1985).
(27) Wallenstein, S., Naus, J. and Glaz, J. 'Powers of the scan statistic for detection of clustering', *Statistics in Medicine*, 12, 1829-1843 (1993)
(28) Waller, L.A., Turnbull, B.W., Clark, L.C. and Nasca, P. 'Chronic disease surveillance and testing of clustering of disease and exposure: application to leukemia incidence and TCE-contaminated dumpsites in upstate New York', *Environmetrics*, 3, 281-300 (1992).
(29) Waller, L.A., Turnbull, B.W., Gustavsson, G. *et al.* ' Detection and assessment of clusters of disease: an application to nuclear power plant facilities and childhood leukemia in Sweden', *Statistics in Medicine*, 14, 3-16 (1995).
(30) 今井淳. 高知県における疾病の地域集積性について－死亡指標の評価と疾病地図への応用－, 平成10年度国立公衆衛生院特別課程疫学統計コース・調査研究報告書, 57-(1998).
(31) 丹後俊郎. 死亡指標の経験的ベイズ推定量について－疾病地図への適用－. 応用統計学, 17, 81-96 (1988).
(32) 福富和夫. 昭和50年職業別訂正死亡率にみられた異常な値について. 日本公衛誌, 27, 229 (1980).
(33) 福富和夫. 死亡指標の意味と性格. 日本公衛誌, 31, 289-295 (1984).