INLA による時空間の従属性を考慮した頻度モデル

佐野誠一郎*

要旨

例えば料率算定において GLM による頻度モデルを用いる場合，データに地域的な傾向が見られたとしても，地域を説明変数として用いるのはその多面的な構造を表現することが難しいため，単純ではない．

この課題に対応する手法として，隣接関係にある地域間の従属性を反映する条件付き自己相関（Conditional AutoRegressive：CAR）モデルがある．GLM に変量効果として CAR モデルを組み込むことで，従来の頻度モデルに地域間の従属性を反映できる．

本稿では，この変量効果を更に拡張し時系列のトレンドを反映した時空間（Spatio-Temporal）モデルを用いて，GLM に地域間および時系列の従属性を反映したモデルを提案する．また計算手法として，一般的にベイズ推定に用いられる MCMC よりも計算負荷が少なく，時空間モデルなどの複雑なモデルにも柔軟に対応できる Integrated Nested Laplace Approximation（INLA）を用いる．

GLM に時空間モデルを INLA により組み込むことで，従来の頻度モデルに地域間および時点間の相関関係を上手く反映し，これにより説明力と予測精度にバランスをもたらすモデルとなり実務的に有用であることを，一般統計を用いた事例により示す．

キーワード

GLM，CAR モデル，時空間モデル，INLA

* 共栄火災海上保険株式会社, se.sano@kyoeikasai.co.jp

1　はじめに

保険数理の実務である料率算定やリスク分析などで，一般統計を使う機会は少なくない．一般統計は統計値に関連する様々な区分により分類されているが，その区分の中には都道府県などの地域が含まれていることがある．さらに地域ごとの結果を見ると，実際に用いることはないとしても，何らかの地域間の傾向が見られることがある．例えば疾病であれば，空間疫学として Yamamoto (2003) などで研究されているように，その種類によっては発生頻度に地理的な影響が存在すると考えられている．

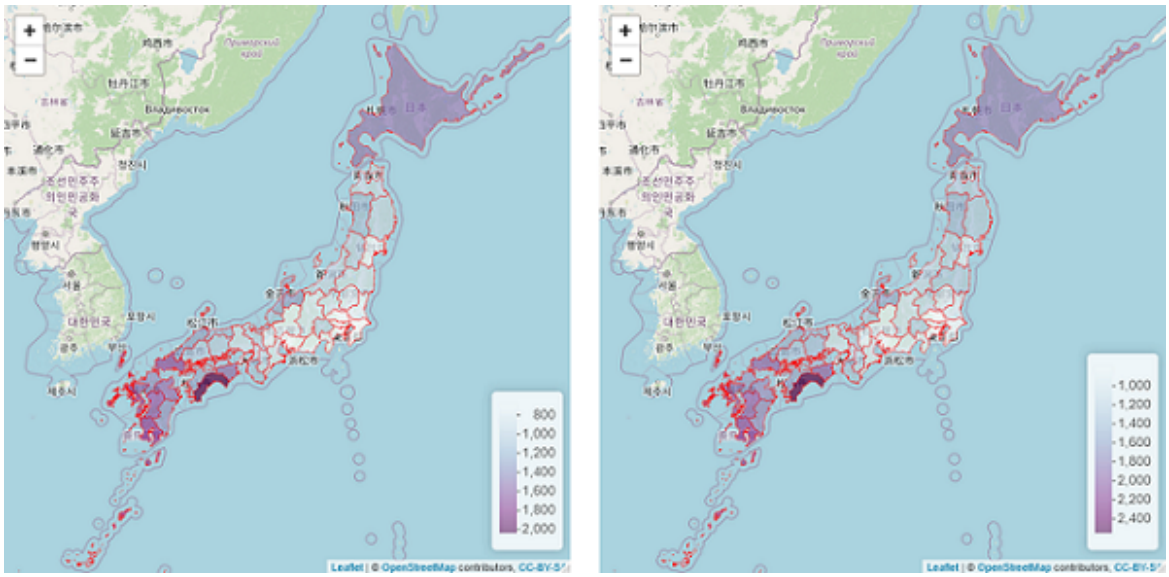

図1　(左) 厚生労働省患者調査平成 29 年第 17 表より都道府県別の入院受療率 (病院)，(右) 厚生労働省医療施設調査平成 29 年第 13 表より都道府県別の人口 10 万人当たり病床数 (病院)

図 1 の左側は，都道府県別の入院受療率（人口 10 万人あたりの入院患者数）を図示している．入院受療率は地域的な傾向が見られるため，より精度の高いモデルを作るのであれば，この傾向を考慮する必要がある．さらに図 1 の右側は，都道府県別の人口 10 万人あたり病床数を図示している．これらを比較すると，入院受療率は病床数に強い影響を受けており，その結果として地域別の格差が発生している．この場合は都道府県を，病床数の分布を表す二次的な指標として，区分に用いることが考えられる．

一方で，このようなデータを用いた保険数理の実務に関する手法として，一般化線形モデル (GLM) によるポアソン分布などを仮定した頻度モデルがある．GLM は多様な変数を組み込むことができるため，従来の料率区分として用いている変数に地域を加えることで，地域間の格差をモデルに反映できる．しかし，頻度モデルとして GLM を用いる場合，例えば疾病に対して既存の料率区分で

ある性別や年齢を変数に入れることはあっても，地域を変数として追加するのは単純ではない．まず，地域を入れるとモデルが冗長になり安定性が低下し，さらにデータが地域間で分断されることにより信頼性が低下するおそれがある．また，モデルが冗長になることを防ぐ手段として，地域に対する変数選択やグループ化もあるが，地域を変数に組み込むのは各地域を独立に捉えることとなるため，たとえグループ化したとしても隣接地域間の格差についての客観的・定性的な評価の課題は解消しない．さらに，モデルに組み込みたいのは地域間の格差そのものではなく，地域間の特性（近い地域であるほど似た傾向を持ち，離れるほど傾向が弱まるなど）であることが多い．地域間の特性を上手く取り込むことができれば，説明力を保持したままモデルの精度を良くする可能性が高まるためである．

この課題に対応する手法として，空間（Spatial）モデルのひとつである条件付き自己相関（Conditional AutoRegressive：CAR）モデルがある．空間モデルにはいくつか種類があるが，ここではデータが持つ地理情報に基づいて各データを一定の区間に分割した上で，その区間の従属性を考慮したモデルを考える．特に CAR モデルは，空間モデルの中でも，隣接関係にある地域間の従属性を考慮したモデルである．CAR モデルは経済学 (Kavanagh, Lee, and Pryce 2016)，疫学 (Jin, Carlin, and Banerjee 2005)，生物学 (Pettitt, Weir, and Hart 2002) など様々な分野の研究に用いられている．また，保険実務への応用については Brechmann and Czado (2014) で紹介されており，他にも Shi and Shi (2017) の例がある．説明力の観点からすると，地域間の境界に大きな障害がある例外を除き，近い地域の相対リスクが類似するという概念をモデルに反映することは，納得感がある手法と考えられる．

本稿では，GLM に空間モデルとして CAR モデルを組み込んだものを拡張して，更に時系列のトレンドを考慮したモデルを用いる．

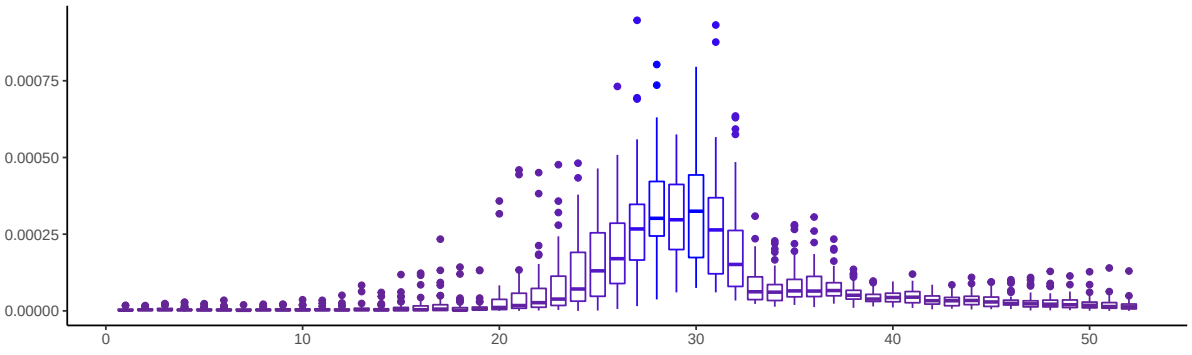

図2　厚生労働省感染症発生動向調査事業年報 2019 年度第 10-1 表より手足口病の発生率の箱ひげ図による推移（週ごと）

例えば，図 2 の箱ひげ図は週ごとの手足口病の発生率を図示し，図 3 はそのうち発生率が高い週を抜粋して，都道府県別の推移を図示している．図 2 によると，発生率は徐々に上昇し，夏ごろにピークを迎えてまた下降している．さらに図 3 によると，発生率はピークを迎える前後で徐々に北

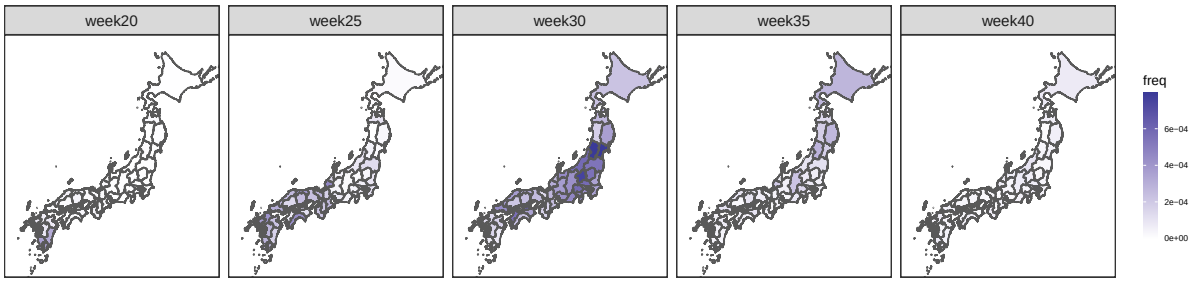

図3　手足口病の発生率の都道府県別推移（週ごと・抜粋）

上していくという地域的な傾向が見られる．したがって，発生率は時系列の変動性と地域的な従属性の両方を特徴として併せ持つため，モデル化するときはそれぞれを考慮する必要がある．

このような空間モデルに時系列を取り入れたモデルを，時空間（Spatio-Temporal）モデルという．時空間モデルも CAR モデルと同様に疫学の分野などで研究されており（例えば Waller et al. (1997) など），保険でも Wang, Schifano, and Yan (2017) の例がある．

CAR モデルを使用した時空間モデルはベイズモデルのひとつであり，ベイズ推定のうち解析的に解くことが難しい場合には，推定手法として MCMC を使うことが最も一般的である．ただし，GLM を拡張した一般化線形混合モデル（GLMM）として，既存の変数を保持したまま地域間の従属性に加えて時系列を反映すると，モデルは複雑になり，サンプリングを繰り返す MCMC では計算負荷が大きく結果が上手く算出できないという難点がある．そのため本稿では，MCMC に代わる手法として INLA（Integrated Nested Laplace Approximation）(Rue, Martino, and Chopin 2009) を用いる．INLA とは，潜在ガウスモデルと言われる階層モデルにラプラス近似を用いる，ベイズ推定手法のひとつである．潜在ガウスモデルは GLM の特性を含むだけでなく，更に変量効果として空間モデルや時空間モデルを組み込むことができるため，効率よく柔軟に様々なモデルに対応できる．

本稿では，疾病に対して一般的に影響があると考えられる性別・年齢に加えて地域別の区分を持つ一般統計のデータとして，人口動態調査に基づくがんによる死亡率を分析し，時空間モデルによりモデル化する．さらに，GLM やそれに対して地域を変数として加えた場合と空間モデル・時空間モデルを比較することにより，直近の実績に対するモデルの精度を測る．各モデルを比較した結果を確認することで，従来の GLM に地域相関と時系列のトレンドを加味することが，モデルに納得感のある事前情報を反映させるだけでなく，予測精度も向上させることを示す．

以降の構成は次の通りである．第 2 節では使用するデータの内容を説明し，地域的な相関の有無を含めた分析を行う．第 3 節では CAR モデルを，第 4 節は時空間モデルおよびそれを算出するために用いる INLA の計算方法を説明する．それから第 5 節で使用するモデル，第 6 節で結果を説明し，第 7 節でその結果を考察する．

2　使用するデータ

本稿の分析には，国立がん研究センターがん情報サービス「がん登録・統計」（人口動態統計）にある「都道府県別がん死亡データ」を用いる．このデータは，がんによる死亡者数および死亡率が，1995 年度から毎年度，性別・年齢区分別（5 歳刻み）・都道府県別に把握できる．

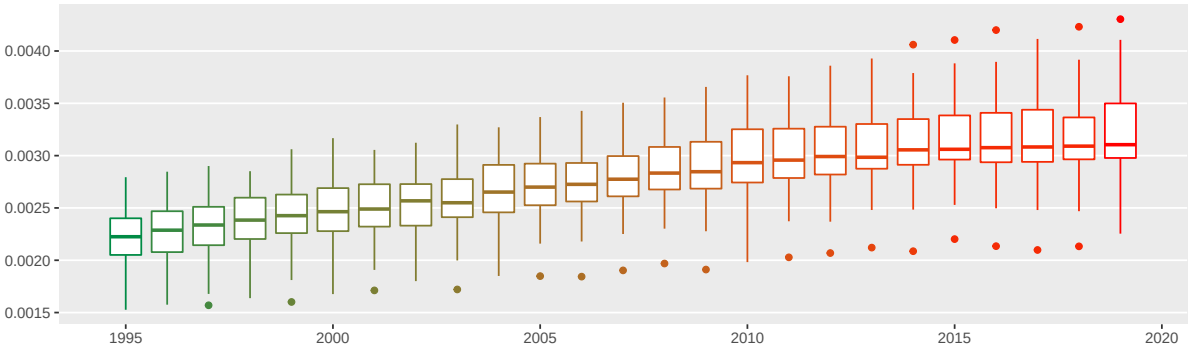

図4　1995 年度から 2019 年度までのがんによる死亡率の箱ひげ図による推移（年度ごと）

図 4 は，年度ごとのがんによる死亡率の箱ひげ図を図示している．図からは，年度ごとの死亡率が徐々に増加する傾向が見られる．これだけを見ると，モデルには時系列として増加トレンドを加味する必要があると判断される．しかし実態として，死亡率の増加トレンドは個人の死亡率そのものが増加しているのではなく，人口における年齢構成の変化が主な要因である．

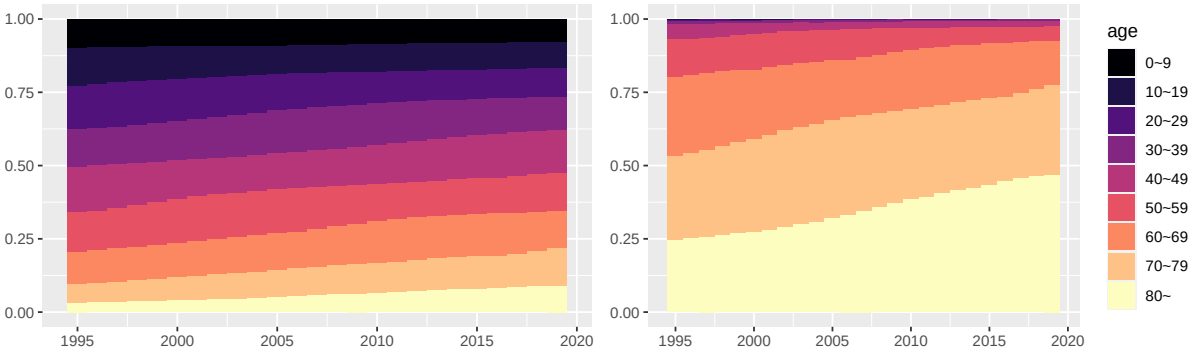

図5　1995 年度から 2019 年度までの年齢区分別の占有率（左が人口，右ががんによる死亡者数）

図 5 は，人口およびがんによる死亡者数の年齢構成に対する年度ごとの推移を図示している．人口は高齢者層が徐々に大きな割合を占めてきており，同時に，死亡者数は高齢者がほぼ全てを占めている．したがって図 4 の情報だけでは本質的な時系列の傾向を把握するのは難しいため，年齢構成の変化による影響を除去して，各年齢の死亡率状況を見る必要がある．

図6　1995 年度から 2019 年度までの性別・年齢区分別のがんによる死亡率推移（年度ごと）

図 6 は，性別・年齢区分別に見た年度ごとのがんによる死亡率を図示している．図 4 で見た死亡率は増加傾向を示していたが，性別・年齢区分別で見ると，各区分とも減少傾向にある．この図は，性別・年齢を変数として用いること，およびこれらを変数として用いた上で更に時系列のトレンドを考慮することの必要性を示している．

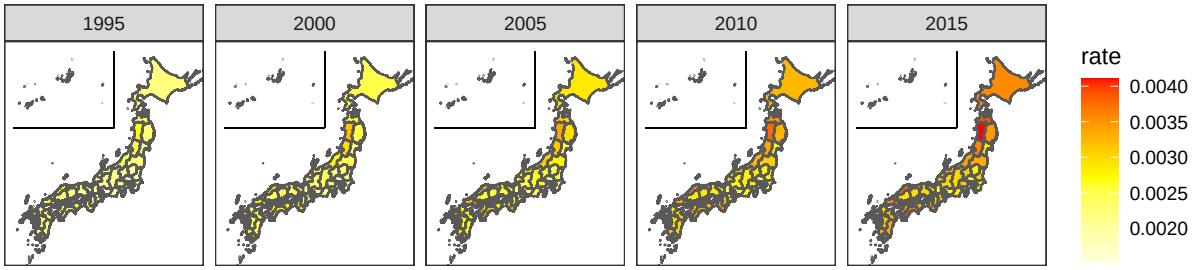

図7　1995 年度から 2015 年度までの 5 年ごと都道府県別のがんによる死亡率推移

図 7 は，5 年単位で見た都道府県別のがんによる死亡率を図示している．年度ごとに地域格差が発生し大きくなり，2010 年以降は，地域間に何らかの傾向または相関が見られる．

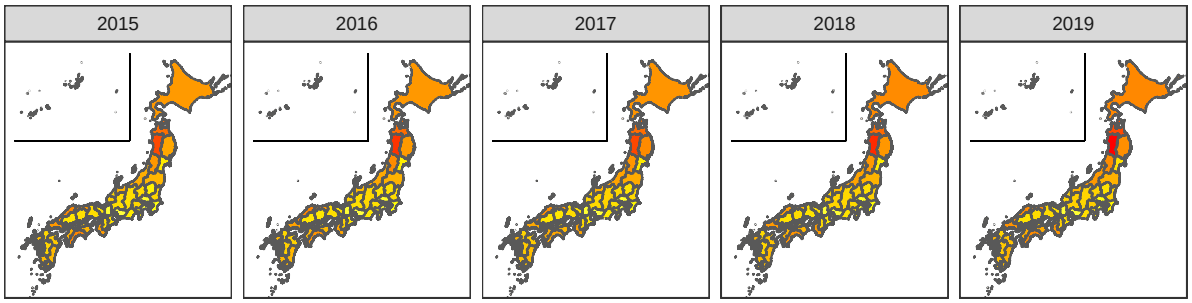

図8　2015 年度から 2019 年度までの直近 5 年間における都道府県別のがんによる死亡率推移

図 8 は，直近 5 年間における都道府県別の死亡率を図示している．図 7 で徐々に強くなっている格差と傾向が，直近年度は明確に確認できる．また，各年度の間で大きな変動は見られない．

2.1　Moran's I 統計量

地域間の影響をモデルに反映する動機として，空間自己相関の程度を測る．空間自己相関とは，ある地域の観測値が周辺地域の影響を互いに受ける度合いを言う．データは都道府県別に分類されているため，都道府県間のがんによる死亡率に対する空間的な影響の強さを測る．初めに，互いに影響を受ける周辺地域として，地域間の隣接関係を設定する必要がある．

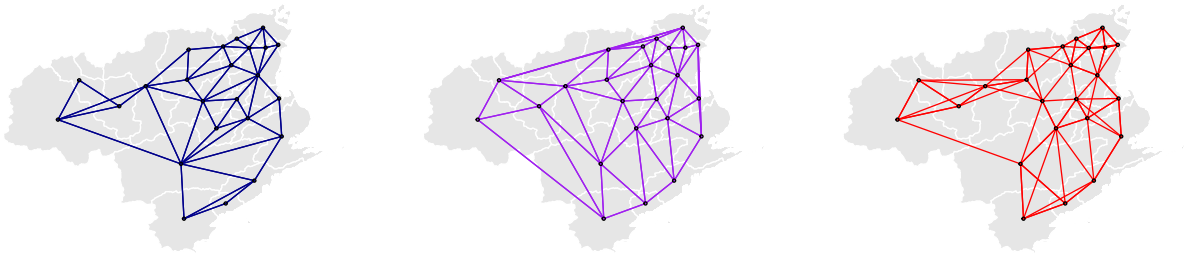

図9　隣接関係の種類（左からルーク型，ドロネー三角網，近隣 4 ゾーン）

図 9 は，R のパッケージ `spdep`(Bivand and Wong 2018) に基づく隣接関係を例示している．地域間のリンクが隣接関係を表している．本稿では，図 9 のルーク型（地域間に共有する境界がある場合を隣接とする）を基本として，隣接関係を設定する．

図10　都道府県間の隣接関係

図 10 は，本稿のモデルに用いる都道府県間の隣接関係を図示している．隣接関係の設定には，地理情報システム（Geographic Information System：GIS）のデータが必要である．上記の隣接関係は，「国土数値情報（行政区域データ）」（国土交通省）から取得した都道府県別のシェイプファイルを加工して設定している．

次に，空間自己相関を表す統計量として，Moran's I 統計量 (Moran 1950) を用いる．Moran's I 統計量は，地域 i $(i = 1, 2, ..., n)$ の標本 y_i に対して，下式で定義される．

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \tag{1}$$

ここで w_{ij} は，地域 i と地域 j が隣接する場合は 1，しない場合は 0 となる．Moran's I 統計量は Pearson 相関係数を空間的に見たもので，データによるが概ね -1 から 1 の間の値を取り，小さいほど負の相関，大きいほど正の相関となる．空間自己相関の場合，正の相関は周辺地域が似た傾向を持つこと，負の相関は周辺地域が異なる傾向を持つことを意味する．

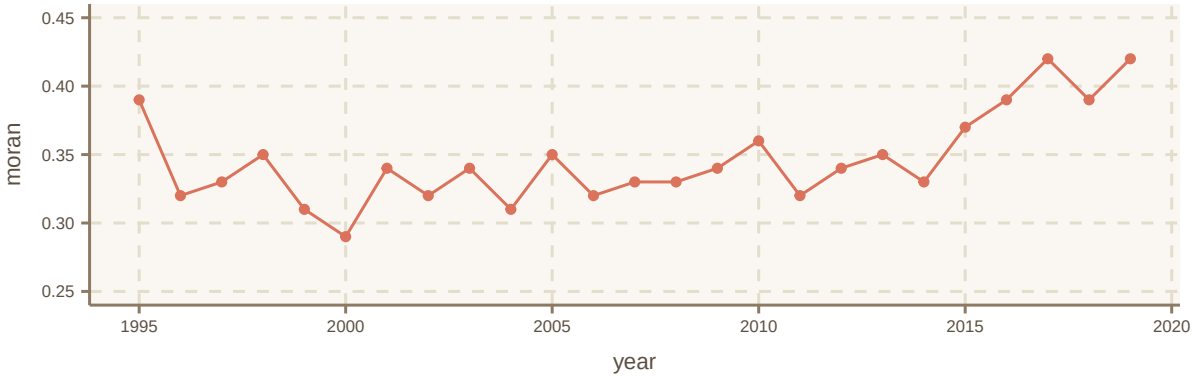

図11　年度ごとの Moran's I 統計量の推移

図 11 は，都道府県別のがんによる死亡率に対する Moran's I 統計量を，年度ごとに図示している．図からは，都道府県間の死亡率にある程度の正の相関が存在し，直近年度は特に強い傾向があることが分かる．上記の空間的な正の相関を踏まえて，これを反映するモデルを考える．

3　CAR モデル

本稿では空間的な相関を反映したモデルとして，CAR モデル (Besag 1974) を用いる．CAR モデルは，一定の地域（州や都道府県など）別に分類されたデータに対して，その地域間の従属性を考慮したモデルである．

$\boldsymbol{u} = (u_1, u_2, ..., u_K)^{\mathrm{T}}$ を，地域ごとの確率変数から成るベクトルとする．このとき，条件付き独立として下式が成り立つと仮定する．

$$p(u_k|\boldsymbol{u}_{-k}) = p(u_k|u_j,\ j \in \mathcal{N}(k)) \tag{2}$$

ここで，$\boldsymbol{u}_{-k}$ $(k = 1, ..., K)$ は u_k 以外の $\boldsymbol{u}$ の成分，$\mathcal{N}(k)$ は地域 k と近接関係にある地域の集合を表す．このような $\boldsymbol{u}$ をマルコフ確率場（Markov Random Field: MRF）(Banerjee, Carlin, and Gelfand 2014) という．CAR モデルとは，MRF を仮定した確率変数 $u_k \in \boldsymbol{u}$ が次の完全条件付き（full conditional）分布に従うものをいう．

$$u_k|\boldsymbol{u}_{-k} \sim N\left(\mu_k + \sum_{j \in \mathcal{N}(k)} c_{kj}(u_j - \mu_j),\ \sigma_k^2\right) \tag{3}$$

ここで，μ_k は u_k の平均，σ_k^2 は u_k の条件付き分散を表す．c_{kj} は $c_{kk} = 0$ を満たす定数で，地域 k と地域 j の従属性を表す．上式の解釈として，CAR モデルは事前分布として，地域間の関係（各

地域の値 u_k は周辺地域の値 u_j $(j \in \mathcal{N}(k))$ とその従属性 c_{kj} により決まる）を入れている．さらに，Brook's lemma により $\boldsymbol{u}$ の同時密度関数が一意に存在する．

$$\boldsymbol{u} \sim N\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix},\ (I-C)^{-1}M\right) \tag{4}$$

ここで，I は単位行列，C は各成分が c_{kj} の $K \times K$ 行列，M は対角成分 (k,k) が σ_k^2 の対角行列を表す．このような多変量正規分布に従う MRF を，ガウスマルコフ確率場（Gaussian Markov Random Field: GMRF）(Rue and Held 2005) という．CAR モデルは GMRF のひとつとして，条件付き独立の仮定が持つマルコフ性により，精度行列（分散共分散行列の逆行列）がスパースとなることで計算効率が良いモデルである．c_{kj}・σ_k^2 の設定により，いくつかのモデルが存在する．

例として，Intrinsic CAR（ICAR）モデル (Besag, York, and Mollié 1991) を挙げる．ICAR は，上記の u_k に対して $\mu_k = 0$ を仮定した上で，下式として表される．

$$u_k|\boldsymbol{u}_{-k} \sim N\left(\frac{\sum_j w_{kj}u_j}{\mathcal{N}_k},\ \frac{\sigma^2}{\mathcal{N}_k}\right) \tag{5}$$

ここで，w_{kj} は地域 k に対して地域 j が隣接する場合は 1，しない場合は 0 となる．$\mathcal{N}_k$ は，地域 k と隣接する地域の数を表す．ICAR は，事前分布として「各地域の値 u_k は，所与の隣接する地域の値の下でその値の加重平均が条件付き平均となり，さらに隣接する地域が多いほど条件付き分散が小さくなる」と考えるモデルである．

他にも，ICAR に無構造の変量効果 $u_k^{(2)}$ を加えて周辺地域に影響されない独立の誤差を同時に反映した Besag-York-Mollie（BYM）モデル (Besag, York, and Mollié 1991) や，ICAR の分散共分散行列を正則にするため行列 C にパラメータ $\phi(<1)$ を乗じた Proper モデル (Banerjee, Carlin, and Gelfand 2014) などがあるが，本稿では，事前分布としての CAR モデルに Leroux モデル (Leroux, Lei, and Breslow 2000) を用いる．Leroux モデルは，上記の u_k に対して下式として表される．

$$u_k|\boldsymbol{u}_{-k} \sim N\left(\frac{\rho\sum_j w_{kj}u_j}{\rho\mathcal{N}_k + 1 - \rho},\ \frac{\sigma^2}{\rho\mathcal{N}_k + 1 - \rho}\right) \tag{6}$$

パラメータ ρ は $0 \leq \rho \leq 1$ の値をとり，これで地域間の従属性の強さをコントロールしている．$\rho = 1$ のときは ICAR モデルとなり，$\rho = 0$ のときは各 u_k が独立となる．

4　時空間モデル

CAR モデルは空間的な相関を表現したモデルだが，これだけでは時系列のトレンドを反映できない．そこで時空間モデルを考える．時空間モデルとは，空間モデルと時系列モデルを組み合わせたもので，本稿では CAR モデルと時系列モデルを組み合わせたものをいう．CAR モデルは近い地域間で影響し合うと仮定するが，同様に，時系列の近い期間で影響し合うと仮定する．例えば自己相関（AR）モデルは GMRF のひとつであり，AR(1) モデル（$x_t = \phi x_{t-1} + \epsilon_t$，$\epsilon_t \sim N(0, \sigma^2)$）を条件付き分布として表すと，下式となる．

$$x_t | x_1, x_2, ..., x_{t-1} \sim N(\phi x_{t-1}, \sigma^2) \tag{7}$$

このように，時系列モデルを空間モデルと同様の枠組みで考慮して，時空間モデルを計算する．

時空間モデルの計算方法はベイズ推定として，R のパッケージ `CARBayesST`(Lee, Rushworth, and Napier 2018) のように MCMC によるものが一般的である．ただし MCMC による時空間モデルの推定は，地域および時系列で集約された観測値を対象として想定している．この場合，他の変数は，地域および時系列の代表値としてその平均などを用いることとなる．一方で本稿のモデルが対象とする観測値は，地域および時系列に加えて他の変数を含むため，より細かく分類されている．このように観測値が細分化されて複雑なモデルの場合，サンプリングを繰り返す MCMC は計算負荷が大きく適していない．そこで MCMC に代わる手法として，INLA を用いる．

4.1　INLA

INLA は，潜在ガウスモデルに対してラプラス近似を用いる，ベイズ推定の決定論的手法である．INLA は MCMC より計算負荷が少なく，複雑なモデルに対応できるという利点がある．これを踏まえて，本稿における時空間モデルの推定に用いる．

観測値としての確率変数 y_i（$\in \boldsymbol{y} = (y_1, y_2, ..., y_n)$）に対して，その平均 μ_i と線形予測子 η_i がリンク関数 $g(\cdot)$ を通じてリンクしている（つまり $g(\mu_i) = \eta_i$）とする．η_i は下式とする．

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(z_{li}) \tag{8}$$

ここで β_0 は切片，$\boldsymbol{\beta} = \{\beta_1, \beta_2, ..., \beta_M\}$ は変数 $\boldsymbol{x} = (x_1, x_2, ..., x_M)$ に関する固定効果のパラメータを表す．また，$\boldsymbol{f} = \{f_1(\cdot), f_2(\cdot), ..., f_L(\cdot)\}$ は変数 $\boldsymbol{z} = (z_1, z_2, ..., z_L)$ に関する関数の集合で，時空間に関する変量効果などを表す．これら潜在パラメータの集合を $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$ とする．さらに，

$\boldsymbol{y}$ の分散や $\boldsymbol{f}$ の分散パラメータなどの集合を，ハイパーパラメータベクトル $\boldsymbol{\psi} = \{\psi_1, \psi_2, ..., \psi_K\}$ とする．

$\boldsymbol{y}$ に条件付き独立を仮定する．つまり $p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta}, \boldsymbol{\psi})$ とする．また，潜在パラメータベクトル $\boldsymbol{\theta}$ に条件付き独立を仮定し，さらに事前分布として正規分布を設定する．これらにより $\boldsymbol{\theta}$ は GMRF に属する．この階層モデルを，潜在ガウスモデル (Rue, Martino, and Chopin 2009) という．潜在ガウスモデルは空間モデルや時空間モデルだけでなく，GLM の要素（指数型分布族，線形予測子，リンク関数）を含んでいる．そのため，潜在ガウスモデルの変量効果に CAR モデルなどを組み込むことで，GLM を拡張して空間従属性などを反映できる．

パラメータ推定で重要なのは，$\boldsymbol{\theta}$ の同時事後分布ではなく，y_i に関連する $\boldsymbol{\theta}$ のパラメータ θ_i の周辺事後分布であることが多い．INLA は各パラメータの周辺事後分布をラプラス近似により求めて，その積分式を周辺尤度の和として近似する．$\boldsymbol{\theta}$ および $\boldsymbol{\psi}$ における各成分の周辺事後分布は，下式となる．

$$
\begin{aligned}
p(\theta_i|\boldsymbol{y}) &= \int p(\theta_i, \boldsymbol{\psi}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\psi} = \int p(\theta_i|\boldsymbol{\psi}, \boldsymbol{y}) p(\boldsymbol{\psi}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\psi} \\
p(\psi_k|\boldsymbol{y}) &= \int p(\boldsymbol{\psi}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\psi}_{-k}
\end{aligned}
\tag{9}
$$

初めに，ハイパーパラメータの事後分布 $p(\boldsymbol{\psi} \mid \boldsymbol{y})$ を下式により算出する．各 ψ_k の周辺事後分布 $p(\psi_k \mid \boldsymbol{y})$ は，$p(\boldsymbol{\psi} \mid \boldsymbol{y})$ から算出する．

$$
p(\boldsymbol{\psi} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \boldsymbol{y})}{p(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \boldsymbol{y})} \propto \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}|\boldsymbol{\psi}) p(\boldsymbol{\psi})}{p(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \boldsymbol{y})} \tag{10}
$$

ここで，分母 $p(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \boldsymbol{y})$ は閉形式ではないことが多いため，分子とは異なり計算が容易ではない．そのため Luke and Joseph (1986) にある周辺事後分布のラプラス近似と同様の手法により，$\boldsymbol{\theta}$ の完全条件付き分布 $p(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \boldsymbol{y})$ を正規分布 $\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \boldsymbol{y})$ に近似する．また，上式の左辺 $p(\boldsymbol{\psi} \mid \boldsymbol{y})$ は $\boldsymbol{\theta}$ によらないことを踏まえ，簡略化のため $\boldsymbol{\psi}$ に対する $\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \boldsymbol{y})$ のモード $\boldsymbol{\theta}^*(\boldsymbol{\psi})$ で計算する (Blangiardo and Cameletti 2015)．

$$
p(\boldsymbol{\psi}|\boldsymbol{y}) \approx \left. \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}|\boldsymbol{\psi}) p(\boldsymbol{\psi})}{\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \boldsymbol{y})} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\psi})} =: \tilde{p}(\boldsymbol{\psi}|\boldsymbol{y}) \tag{11}
$$

さらに，$\tilde{p}(\boldsymbol{\psi}|\boldsymbol{y})$ を算出したのちに，$\tilde{p}(\boldsymbol{\psi}|\boldsymbol{y})$ の密度を表す代表的な点の集合 $\{\boldsymbol{\psi}^{(j)}\}$ を取得する．

次に，$p(\theta_i \mid \boldsymbol{\psi}, \boldsymbol{y})$ を下式により近似する（$\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \boldsymbol{y})$ から直接算出する手法もあるが，精度が高くないため，他の手法が提案されている）．

$$
\begin{aligned}
p(\theta_i|\boldsymbol{\psi},\boldsymbol{y}) = \frac{p(\{\theta_i,\boldsymbol{\theta}_{-i}\}|\boldsymbol{\psi},\boldsymbol{y})}{p(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})} &= \frac{p(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{y})}{p(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})} \\
&\propto \frac{p(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{\psi})p(\boldsymbol{\theta}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{p(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})} \\
&\approx \left.\frac{p(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{\psi})p(\boldsymbol{\theta}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{\tilde{p}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})}\right|_{\boldsymbol{\theta}_{-i}=\boldsymbol{\theta}^*_{-i}(\theta_i,\boldsymbol{\psi})} =: \tilde{p}(\theta_i|\boldsymbol{\psi},\boldsymbol{y})
\end{aligned}
\tag{12}
$$

ここで，$\tilde{p}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})$ は $p(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})$ を前述と同様のラプラス近似により正規分布に近似したもの，$\boldsymbol{\theta}^*_{-i}(\theta_i,\boldsymbol{\psi})$ は $\tilde{p}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})$ のモードを表す．

ただし計算負荷が大きいため，上式の分母・分子を更に 3 次までテイラー展開する simplified laplace approximation という手法を用いて近似する (Rue, Martino, and Chopin 2009)．

これら $p(\boldsymbol{\psi}\,|\,\boldsymbol{y})$ と $p(\theta_i|\boldsymbol{\psi},\boldsymbol{y})$ を合わせて，潜在パラメータの周辺事後分布 $p(\theta_i\,|\,\boldsymbol{y})$ を算出する．$p(\theta_i\,|\,\boldsymbol{y})$ は前述の積分式を $\tilde{p}(\theta_i\,|\,\boldsymbol{\psi},\boldsymbol{y})$ および $\tilde{p}(\boldsymbol{\psi}|\boldsymbol{y})$ に置き換えて計算するが，前述の $\{\boldsymbol{\psi}^{(j)}\}$ を使い，積分式を下式として近似する．

$$
p(\theta_i\,|\,\boldsymbol{y}) \approx \int \tilde{p}(\theta_i|\boldsymbol{\psi},\boldsymbol{y})\tilde{p}(\boldsymbol{\psi}|\boldsymbol{y})\mathrm{d}\boldsymbol{\psi} \approx \sum_j \tilde{p}(\theta_i|\boldsymbol{\psi}=\boldsymbol{\psi}^{(j)},\boldsymbol{y})\tilde{p}(\boldsymbol{\psi}=\boldsymbol{\psi}^{(j)}|\boldsymbol{y})\Delta_j \tag{13}
$$

ここで Δ_j は，$\boldsymbol{\psi}^{(j)}$ に対応するウェイト（積分式を離散化したことによる調整）を表す．

5　モデル

前述のとおり，潜在ガウスモデルは GLM の要素を含むため，INLA を用いる場合は GLM を基礎としてモデル化できる．これを踏まえて，GLM に変量効果として時空間モデルを組み込んだものとして，下式を計算する．

$$\begin{aligned} y_i &\sim \mathcal{P}o(\lambda_i) \\ g(\lambda_i) &= \log \lambda_i = \eta_i \\ \eta_i &= \log E_i + \beta_0 + \beta_{\text{gender}} x_{1i} + \beta_{\text{age}} x_{2i} + u_{k(i)} + \gamma_{t(i)} \end{aligned} \tag{14}$$

ここで，y_i $(i = 1, \cdots, n)$ はがんによる死亡者数を表す．分布はポアソン分布を仮定し，リンク関数は対数リンク，E_i はオフセット項として人口とする．β_0 は切片であり，それとは別に固定効果として，性別（β_{gender}）・年齢区分（β_{age}）を変数に入れている．また，各 β_i には漠然（vague）事前分布として $N(0, 10^3)$ を設定する．

$u_{k(i)}$ は空間変量効果として，CAR モデルを組み込む．CAR モデルには，前述のとおり Leroux モデルを用いる．

$$u_{k(i)} | \boldsymbol{u}_{-k(i)} \sim N\left(\frac{\rho \sum_j w_{k(i)j} u_j}{\rho \mathcal{N}_{k(i)} + 1 - \rho},\ \frac{\tau_u^{-1}}{\rho \mathcal{N}_{k(i)} + 1 - \rho} \right) \tag{15}$$

$\gamma_{t(i)}$ は時系列に関する変量効果として，二次増分が正規分布に従うランダムウォーク (Rue and Held 2005) を組み込む．このとき，$\gamma_{t(i)}$ は確率変数のベクトル $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_T)^{\mathrm{T}}$ に対して下式を満たす．

$$\Delta^2 \gamma_t = (\gamma_t - \gamma_{t+1}) - (\gamma_{t+1} - \gamma_{t+2}) \sim N(0, \tau_\gamma^{-1}) \tag{16}$$

これにより，γ_t の条件付き分布および $\boldsymbol{\gamma}$ の同時密度関数は下式となる．

$$\gamma_t | \gamma_{t-1}, \gamma_{t-2} \sim N\left(2\gamma_{t-1} - \gamma_{t-2}\ ,\ \tau_\gamma^{-1}\right) \tag{17}$$

$$\begin{aligned} p(\boldsymbol{\gamma} | \tau_\gamma) &\propto \tau_\gamma^{\frac{T-2}{2}} \exp\left(-\frac{\tau_\gamma}{2} \sum_{t=1}^{T-2} (\Delta^2 \gamma_t)^2 \right) \\ &= \tau_\gamma^{\frac{T-2}{2}} \exp\left(-\frac{\tau_\gamma}{2} \boldsymbol{\gamma}^T \boldsymbol{R} \boldsymbol{\gamma} \right) \end{aligned} \tag{18}$$

ここで $\boldsymbol{R}$ は，$\boldsymbol{\gamma}$ の時系列の隣接関係を表す構造行列である．

潜在ガウスモデルにおけるハイパーパラメータは，このモデルの場合は変量効果である $u_{k(i)}$（Leroux モデル）の τ_u，ρ および $\gamma_{t(i)}$（ランダムウォーク）の τ_γ である．これらは，事前分布としてそれぞれ $\log\tau_u$ および $\log\tau_\gamma$ に $\mathrm{LogGamma}(1, 0.0005)$，$\log\rho/(1-\rho)$ に $N(0, 0.45^{-1})$ を設定する．

6　解析例と結果

地域間の相関が直近年度で明確に見られることを考慮して，2014 年度〜2019 年度の 6 年分のデータを用いてモデルを計算する．そのうち 2014 年度〜2018 年度の 5 年分のデータを訓練データ，2019 年度をテストデータとして，モデルの精度を評価する．また比較のために，次のモデルを同時に計算する．

- GLM(1)：上記モデルを固定効果のみとしたモデル

$$\eta_i = \log E_i + \beta_0 + \beta_{\mathrm{gender}} x_{1i} + \beta_{\mathrm{age}} x_{2i} \tag{19}$$

- GLM(2)：上記モデルに固定効果として地域を変数に入れたモデル

$$\eta_i = \log E_i + \beta_0 + \beta_{\mathrm{gender}} x_{1i} + \beta_{\mathrm{age}} x_{2i} + \beta_{\mathrm{region}} x_{3i} \tag{20}$$

- 空間モデル：上記モデルの変量効果を CAR モデルのみとしたモデル

$$\eta_i = \log E_i + \beta_0 + \beta_{\mathrm{gender}} x_{1i} + \beta_{\mathrm{age}} x_{2i} + u_{k(i)} \tag{21}$$

GLM(1) は標準的なモデルとして，他とは異なり，地域に関する変数を組み込んでいない．このモデルを他と比較することで，地域を考慮する必要性を測る．GLM(2) は，地域に関する変数を固定効果（β_{region}）として組み込んでいる．このモデルは地域ごとのがんによる死亡者数を独立に扱っているため，説明が難しく実務で用いることはあまりないが，比較のために作成している．この GLM(2) と空間モデル・時空間モデルを比較することで，地域を単純に変数として考慮するだけでなく地域相関を考慮する必要性を測る．さらに空間モデルと時空間モデルを比較することで，地域相関だけでなく時系列のトレンドも考慮する必要性を測る．

各モデルは R を用いて算出する．特に INLA については，R のパッケージ `R-INLA`（www.r-inla.org）を使用する．また，当てはまりの良さを測るために，ベイズ推定に対するモデル比較の一般的な基準である逸脱度情報量規準（Deviance Information Criteria: DIC）(Spiegelhalter et al. 2002)，および広く使える情報量規準（Widely Applicable Information Criterion: WAIC）(Watanabe

and Opper 2010) を用いる．さらに，予測精度を測るために，ベイズ推定に対する leave-one-out クロスバリデーションに基づく指標である CPO（Conditional Predictive Ordinate），ならびにテストデータに対する，人口を重みとした加重平方平均二乗誤差（Weighted Root Mean Squared Error: WRMSE）を併せて用いる．

6.1 結果

表1　各モデルにおける固定効果パラメータの事後分布の平均と標準偏差（抜粋）

	GLM(1)		GLM(2)		空間モデル		時空間モデル	
	mean	sd	mean	sd	mean	sd	mean	sd
Intercept	-10.7207	0.0491	-10.6085	0.0492	-10.7249	0.0527	-10.7350	0.0521
age 40~44	2.8411	0.0498	2.8404	0.0498	2.8404	0.0498	2.8403	0.0498
age 45~49	3.4314	0.0495	3.4306	0.0495	3.4306	0.0495	3.4323	0.0495
age 50~54	4.0423	0.0494	4.0410	0.0494	4.0410	0.0494	4.0421	0.0494
age 55~59	4.6177	0.0493	4.6158	0.0493	4.6158	0.0493	4.6164	0.0493
gender female	-0.6994	0.0015	-0.7006	0.0015	-0.7006	0.0015	-0.7009	0.0015

表 1 は，固定効果に関するパラメータの一部について，事後分布の平均と標準偏差を抜粋して示している．各値ともに共通して，モデル間で大きな差異はない．

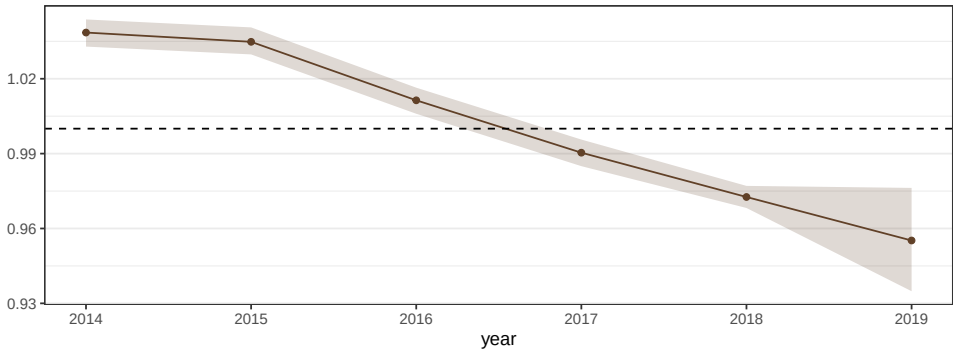

図12　時空間モデルの時系列に関する変量効果の事後分布

図 12 は，時空間モデルの時系列に関する変量効果（$\gamma_{t(i)}$）の事後分布の平均と 95%信頼区間を抜き出して図示している．図 6 で見た年齢区分別のがんによる死亡率と同様に，年度ごとの減少傾向が確認できる．

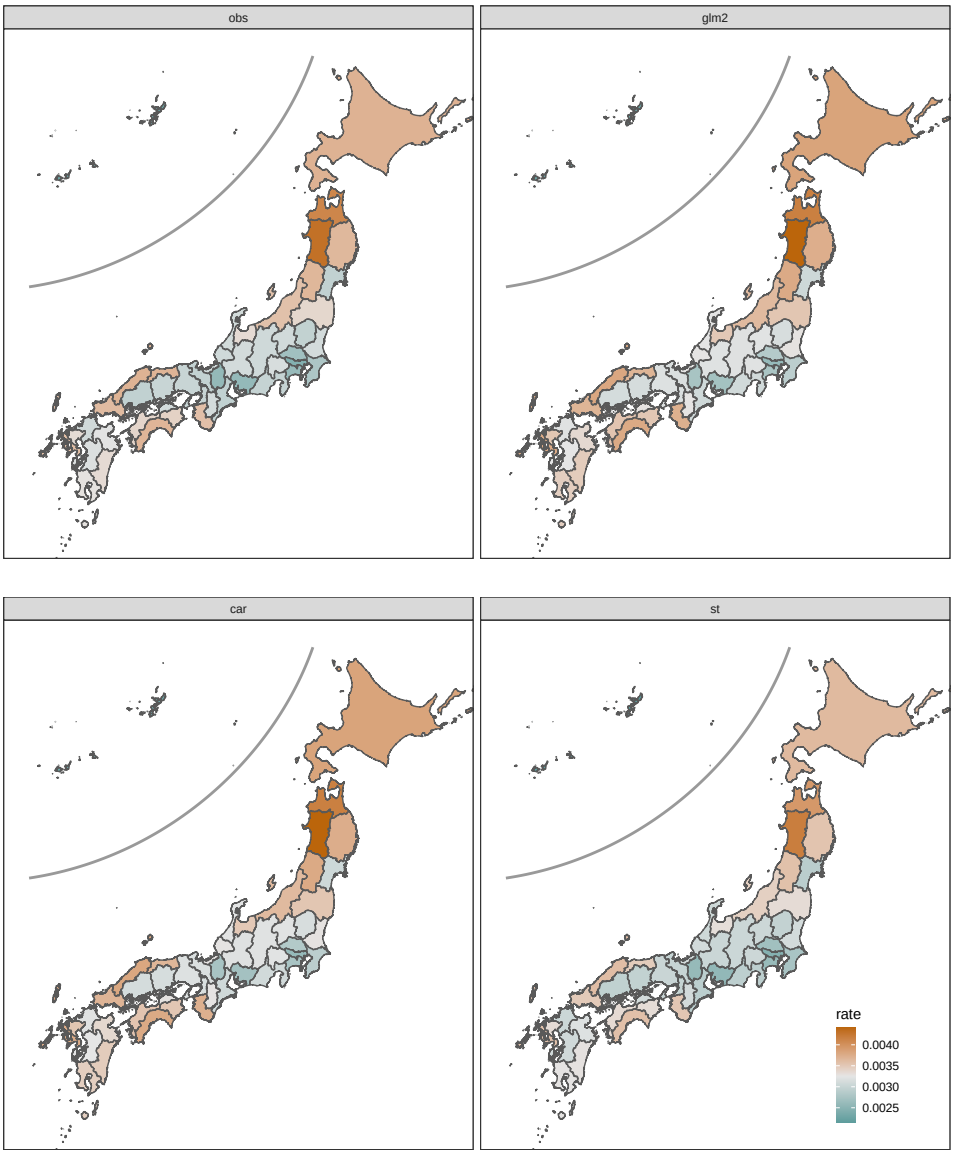

図13　2019 年度における都道府県別のがんによる死亡率の実績と予測値（左上から実績，GLM(2)，空間モデル，時空間モデル）

図 13 は，2019 年度のテストデータに対するがんによる死亡率の実績と各モデルの予測値を，都道府県別に図示している．GLM(1) は，都道府県間の格差がないため省略している．実績に対して GLM(2) の予測値は全体的に高い水準となっているが，GLM(2) と空間モデルの間に明確な差異はない．一方で時空間モデルは，図 12 にある時系列のトレンドを反映したことで，GLM(2) および空間モデルと比較して，より実績に近い水準となっている．

表2　各モデルの統計量

	DIC	WAIC	CPO	WRMSE
GLM(1)	76635	76877	38438	96.22
GLM(2)	72278	72616	36309	86.25
空間モデル	72277	72613	36307	86.13
時空間モデル	**71122**	**71442**	**35722**	**65.89**

表 2 は，モデルごとの DIC，WAIC，CPO および WRMSE を示している．各統計量は，値が小さいほど当てはまりまたは予測精度が良いことを示す．いずれの指標も，モデルが複雑になるとともに，当てはまりおよび予測精度が良くなっている．

7　考察

統計的な指標によるモデルの比較結果は，いずれも時空間モデルが最良であることを示していた．GLM(1) から GLM(2) への指標の改善は，データに対して地域別の格差を考慮する必要があることを意味し，空間モデルへの改善は更に地域間の従属性を考慮する必要があること，時空間モデルへの改善は時系列のトレンドを併せて考慮する必要があることを意味している．

一方で固定効果である性別・年齢区分の事後分布は，モデルが複雑になってもその影響をあまり受けず，モデル間で大きな差異はなかった．したがって，性別・年齢区分はデータの中で強い影響を持っていると考えられる．これは，空間モデルまたは時空間モデルを使う場合であっても，従来の GLM で考慮している性別・年齢区分がなおも必要であり，一般的に時空間モデルが対象とする地域および時系列で集約された観測値では精度が向上しないことを示唆している．

ただし性別・年齢区分の強い影響により，図 13 のとおり，GLM(2) と空間モデルの差異は明確には確認できておらず，空間モデルが CAR モデルを組み込むことでどのようにモデルの精度が改善したかは不透明である．そのため次の 2 点を別の角度から分析することで，空間モデルが GLM に与える影響として，CAR モデルの効果を確認する．

7.1　CAR モデルの効果（地域間格差の平滑化）

CAR モデルには，地域間の従属性を反映することで，予測値に対して隣接地域間の格差を平滑化する効果がある．そのため，例えば料率算出において地域間の区分を決定する際に，データ値の定量的な格差に加えて定性的な判断を組み込む動機がある場合には，CAR モデルによる平滑化が合理的な手法の選択肢となる．これを踏まえて，CAR モデルを変量効果に組み込むことにより地域

間の従属性が反映されることの影響を確認する．ただし Wall (2004) にあるように，CAR モデルは合理性と計算効率のバランスを考慮して条件付き独立を仮定することにより，分散共分散行列ではなく精度行列を指定しているため，分散構造に関するパラメータの評価からモデルを解釈するのが難しい．また前述のとおり，本稿のモデルは既に性別・年齢による変数の影響が強いデータを用いているため，地域間の従属性を考慮することは単純に変数として加えるより客観性は高いが，その結果が大きく変化しているわけではない．

そのため，年齢区分（β_{age}）を固定してその影響を抑えた上での結果について，モデルを比較する．具体的には，年齢区分 50〜54 歳のデータを抽出し，それに対して GLM(2) と空間モデルを計算する．各モデルのうち GLM(2) の固定効果 β_{region}（以下，GLM）と空間モデルの変量効果 $u_{k(i)}$（以下，CAR モデル）が地域間格差を直接反映したパラメータであるため，これらの事後分布の平均を比較することで，CAR モデルの効果を確認する．

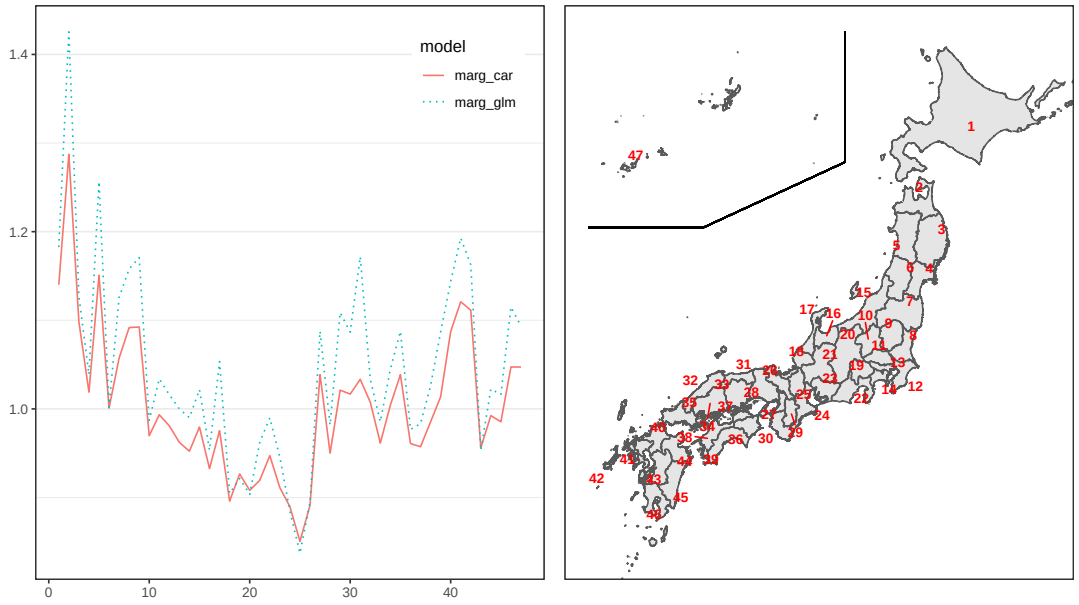

図14　(左) 地域格差に関するパラメータの事後分布の平均（点線が GLM，実線が CAR モデル），(右) 左図に対応する都道府県のインデックス

図 14 は，GLM と CAR モデルそれぞれの事後分布の平均による相対リスクを，横に並べて図示している．各地域を独立に見る GLM に対して，CAR モデルは格差が縮小し，全体的にぶれが小さくなっている．

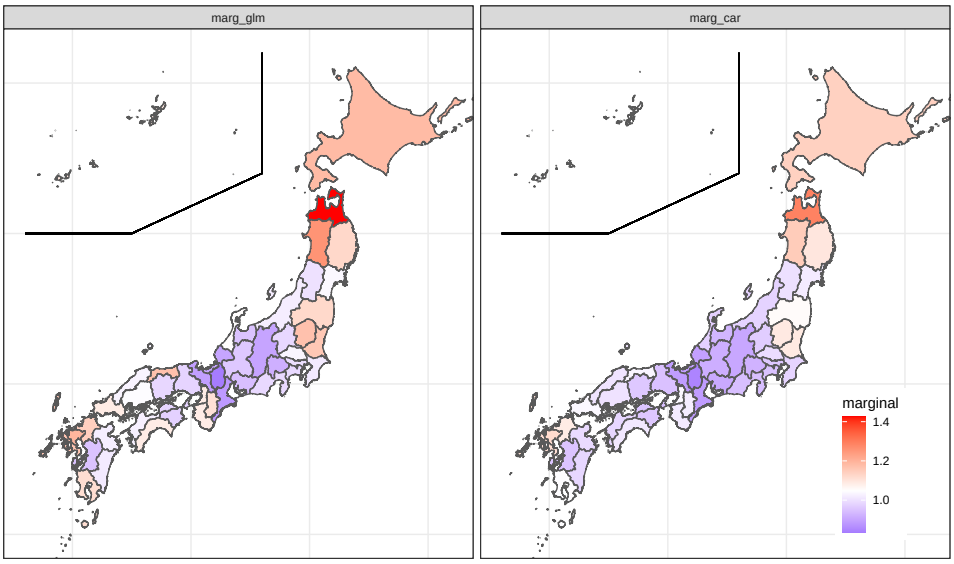

図15　地域格差に関するパラメータの事後分布の平均を地図で見た場合（左から GLM，CAR モデル）

図 15 は，図 14 の相対リスクを地域的に比較できるよう図示している．CAR モデルによる地域間の従属性を反映した結果として，GLM に比べて地域間の格差が減少し，全体的に滑らかになっている．

7.2　CAR モデルの効果（未知の地域の補間）

ここでは，観測されなかった地域がある場合の，その地域の予測を考える．実務では，これまで保険商品の販売実績がなかった地域に対して，新たに販売網を展開する際のリスク分析などのケースが挙げられる．

CAR モデルは前述のとおり，各地域が周辺地域の値により決まるという事前分布を設定している．これにより，Ver Hoef et al. (2017) にあるように，CAR モデルは未知の地域に対して，ベイズ予測としての予測の際に周辺地域の情報に基づき予測値を補間するという効果がある．そのため，データがない地域に対して CAR モデルを適用することで，より精度の高い分析が期待できる．この効果を確認するために，前述のデータを用いて，各都道府県を欠測とした場合の予測についてモデルを比較する．具体的には，訓練データの各都道府県をそれぞれ欠測としてモデル化し，テストデータに対するその欠測とした都道府県の予測値により，各モデルの精度を確認する．

表3　欠測とした都道府県の予測値に対する各モデルの WRMSE

	WRMSE
GLM(1)	96.22
空間モデル	86.13
時空間モデル	**65.89**

表 3 は，予測精度を示す統計量として，モデルごとの WRMSE を示している．GLM(2) は，地域に関する変数を固定効果（β_{region}）として各地域を独立に見ており，欠測している地域は予測できないことから省略している．それ以外の WRMSE は，6.1 節の結果と同様に，空間モデルが GLM(1) より，時空間モデルが他の 2 モデルより予測精度が良くなっている．

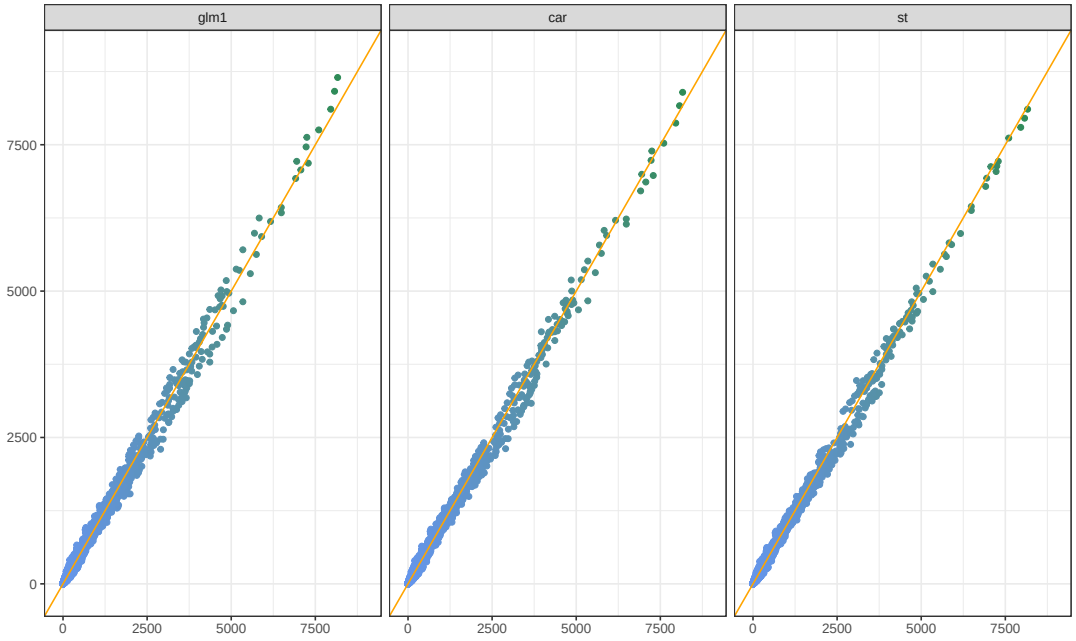

図16　モデルごとの q-q プロット（左から GLM(1)，空間モデル，時空間モデル）

図 16 は，モデルごとの WRMSE の視覚化を目的として q-q プロットを図示している．x 軸が実績，y 軸がモデルによる予測値を表している．図からは，特に裾の値について，GLM(2) より空間モデル，空間モデルより時空間モデルの当てはまりが良くなっているのが確認できる．

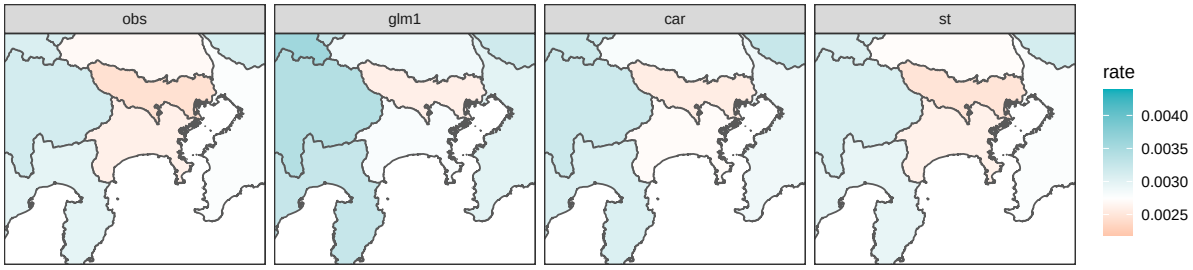

図17　2019 年度における神奈川県に対する実績と予測値（左から実績，GLM(1)，空間モデル，時空間モデル）

図 17 は例として，神奈川県（各図内の中央）を欠測とした場合のテストデータに対する実績と各モデルの予測値を図示している．GLM(1) の予測値は都道府県全体の平均が適用されるため，実績と比較して高い水準となっている．一方で，空間モデルは周辺情報からの補間により GLM(1) より低くなり，時空間モデルは時系列トレンドの反映により更に低い値を取ることで，より実績に近い水準となっている．

8　まとめ

本稿では，通常の GLM では難しい地域間および時系列の従属性という多面的な構造を反映したモデルとして時空間モデルを，さらに，複雑なモデルのベイズ推定に柔軟かつ効率的に対応する手法として INLA を提案した．今回のケースでは，地域相関および時系列のトレンドを持つがんによる死亡者数のデータに対して，時空間モデルは従来考えられるモデルよりも高い予測精度を示した．時空間モデルは，GLM に変量効果として CAR モデルや時系列モデルを組み込むことにより，地域間または時系列で近いものは似た傾向をもつという事前情報を加えることができるため，地域や時系列を単純に変数として加えるよりモデルの前提に納得感を持たせることができる．アクチュアリーの業務は，特に料率算定などでモデルを使用する場合，納得感のある前提の説明が求められることが多い．そのため時空間モデルは，一定の条件に対して主体性はあるが無理のない事前情報を持つ特性とその精度から，使い勝手のよい手法である．

なお，本稿は従来の実務で用いられることの多い GLM を拡張するというアプローチで時空間モデルを提案するために，既存の考慮すべき変数（性別・年齢）が一定の強さを持つ性質のデータに対して，潜在変数として地域間および時系列の従属性を反映するモデルを用いている．そのため，CAR モデルや時系列モデル自体の種類ごとの比較（例えば CAR モデルにおける前述の ICAR・BYM・Proper・Leroux の比較など）は，結果に特段の差異がなかったため記載していない．一方で，前述の手足口病のように，伝播経路として時点間および空間の影響が強い感染症の発生率をモデル化する場合は，時空間モデルの内容をさらに深める動機が強く，さらに対象とするデータは地域および時系列で集約されたものが多い．したがって，このような事例に対して地域および時系列

で集約されたデータを用いた手法を紹介することがあれば，本稿のモデルだけでなく，地域と時系列の交互作用を組み込むことで時間の経過とともに地域間の従属性が変化するモデルなど，時空間モデルを更に発展させる可能性を持っている．

謝辞

本稿の執筆にあたり，大同火災海上保険株式会社の川上良一氏と Guy Carpenter Japan, Inc. の藤田卓氏には非常に多くの助言をいただきました．この場をお借りし御礼申し上げます．

参考文献

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand. 2014. *Hierarchical Modeling and Analysis for Spatial Data*. CRC press.

Besag, Julian. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society: Series B (Methodological)* 36 (2): 192–225.

Besag, Julian, Jeremy York, and Annie Mollié. 1991. "Bayesian Image Restoration, with Two Applications in Spatial Statistics." *Annals of the Institute of Statistical Mathematics* 43 (1): 1–20.

Bivand, Roger, and David W. S. Wong. 2018. "Comparing Implementations of Global and Local Indicators of Spatial Association." *TEST* 27 (3): 716–48.

Blangiardo, M, and M Cameletti. 2015. *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. John Wiley & Sons, Ltd.

Brechmann, Eike, and Claudia Czado. 2014. *Spatial Modeling*. Edited by Edward W. Frees, Richard A. Derrig, and GlennEditors Meyers. Vol. 1. Predictive Modeling Techniques. Cambridge University Press.

Jin, Xiaoping, Bradley P Carlin, and Sudipto Banerjee. 2005. "Generalized Hierarchical Multivariate CAR Models for Areal Data." *Biometrics* 61 (4): 950–61.

Kavanagh, Leo, Duncan Lee, and Gwilym Pryce. 2016. "Is Poverty Decentralizing? Quantifying Uncertainty in the Decentralization of Urban Poverty." *Annals of the American Association of Geographers* 106 (6): 1286–98.

Lee, Duncan, Alastair Rushworth, and Gary Napier. 2018. "Spatio-Temporal Areal Unit Modelling in r with Conditional Autoregressive Priors Using the CARBayesST Package." *Journal of Statistical Software* 84 (9).

Leroux, Brian G, Xingye Lei, and Norman Breslow. 2000. "Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence." In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, 179–91. Springer.

Luke, Tierney, and B. Kadane Joseph. 1986. "Accurate Approximations for Posterior Moments and Marginal Densities." *Journal of the American Statistical Association* 81 (393): 82–86.

Moran, Patrick A P. 1950. "Notes on Continuous Stochastic Phenomena." *Biometrika* 37 (1/2): 17–23.

Pettitt, Anthony, Iain Weir, and Andrew Hart. 2002. "A Conditional Autoregressive Gaussian Process for Irregularly Spaced Multivariate Data with Application to Modelling Large Sets of Binary Data." *Statistics and Computing* 12: 353–67.

Rue, Håvard, and Leonhard Held. 2005. *Gaussian Markov Random Fields: Theory and Applications*. CRC press.

Rue, Håvard, Sara Martino, and Nicolas Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society: Series b (Statistical Methodology)* 71 (2): 319–92.

Shi, Peng, and Kun Shi. 2017. "Territorial Risk Classification Using Spatially Dependent Frequency-Severity Models." *ASTIN Bulletin: The Journal of the IAA* 47 (2): 437–65.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. 2002. "Bayesian Measures of Model Complexity and Fit (with Discussion)." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (4): 583–616.

Ver Hoef, Jay, Erin Peterson, Mevin Hooten, Ephraim Hanks, and Marie-Josee Fortin. 2017. "Spatial Autoregressive Models for Statistical Inference from Ecological Data." *Ecological Monographs* 88.

Wall, Melanie M. 2004. "A Close Look at the Spatial Structure Implied by the CAR and SAR Models." *Journal of Statistical Planning and Inference* 121 (2): 311–24.

Waller, Lance A, Bradley P Carlin, Hong Xia, and Alan E Gelfand. 1997. "Hierarchical Spatio-Temporal Mapping of Disease Rates." *Journal of the American Statistical Association* 92 (438): 607–17.

Wang, Chun, Elizabeth D Schifano, and Jun Yan. 2017. "Geographical Ratings with Spatial Random Effects in a Two-Part Model." *Variance* 13 (1): 20.

Watanabe, Sumio, and Manfred Opper. 2010. "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory." *Journal of Machine Learning Research* 11 (12).

Yamamoto, Masaharu. 2003. "Epidemiological Studies on the Distribution and Determinants of Biliary Tract Cancer." *Environmental Health and Preventive Medicine* 7 (6): 223–29.

A frequency model with Spatio-Temporal dependence using INLA

Seiichiro Sano

Abstract

When we use a frequency model through GLM for pricing, even if there is a regional trend in the data, it is not enough just adding the region as an explanatory variable as it is difficult to express the multifaceted structure of the data.

The Conditional AutoRegressive (CAR) model, which reflects the dependence between adjacent regions, is one of a variety of methods that addresses this issue. By incorporating CAR into GLM as a random effect, it is possible to reflect the dependence among regions in contrast to the conventional frequency models.

In this paper, we further extend the random effect by using a Spatio-Temporal model which incorporates the trend within time series, and thus is a model that reflects not only the dependence among regions but also the dependence within time series in GLM. As a computational method, we use Integrated Nested Laplace Approximation (INLA), which is less computationally demanding than MCMC. INLA is generally used for bayesian inference, and it is flexible enough to handle complex models such as the Spatio-Temporal model.

By incorporating a Spatio-Temporal model into GLM using INLA, we show that the model can successfully reflect the effects of regional and time-series influences on the conventional frequency models, and thus bring balance between explanatory power and predicting accuracy, which is of practical use.

Keywords

GLM, CAR, Spatio-Temporal model, INLA