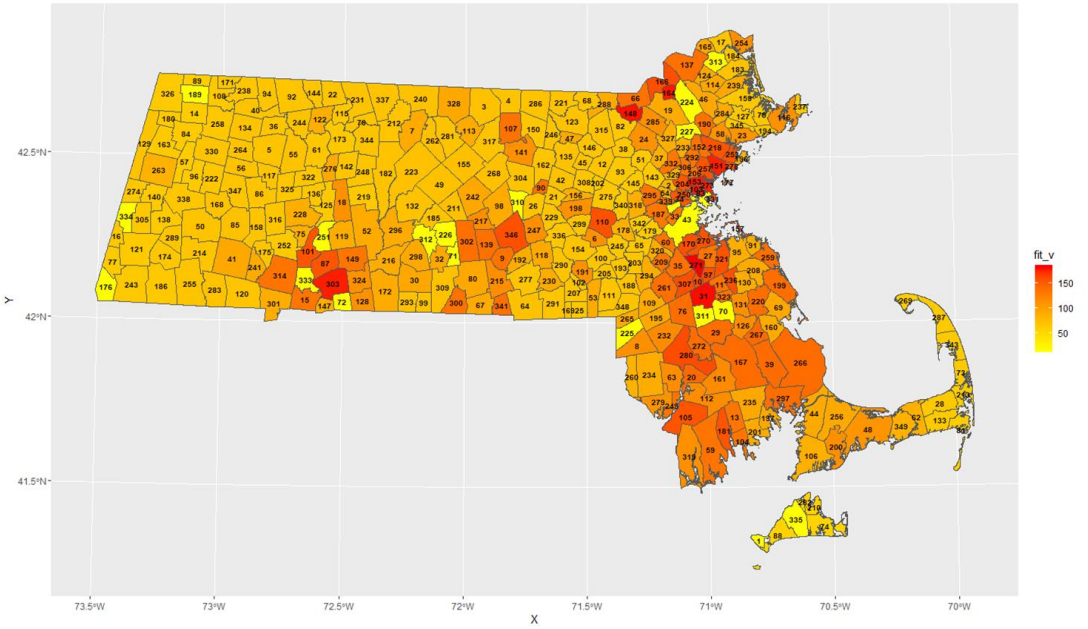
地域間の従属性を考慮した頻度モデル

2020年2月1日

川上 良一（大同火災海上保険）
佐野 誠一郎（共栄火災海上保険）

本発表研究の背景

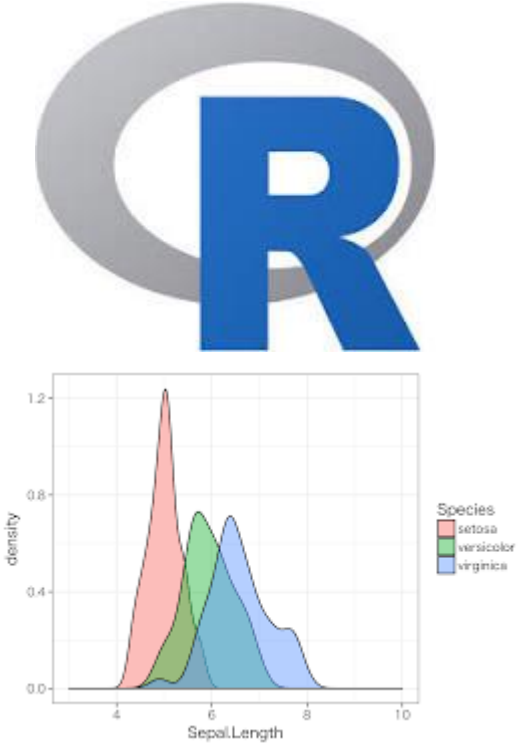
- ASTINの論文「TERRITORIAL RISK CLASSIFICATION USING SPATIALLY DEPENDENT FREQUENCY-SEVERITY MODELS」（Peng Shi and Kun Shi）について研究を行っていたことがスタート
- 損害保険分野では損害について頻度と損傷度に分けた分析が行われ、この中ではGLMが良く用いられる
- 上記論文においては、米国マサチューセッツ州の自動車事故についてGLMに地域間の従属性を織り込んだ頻度と損傷度の考察を行っている

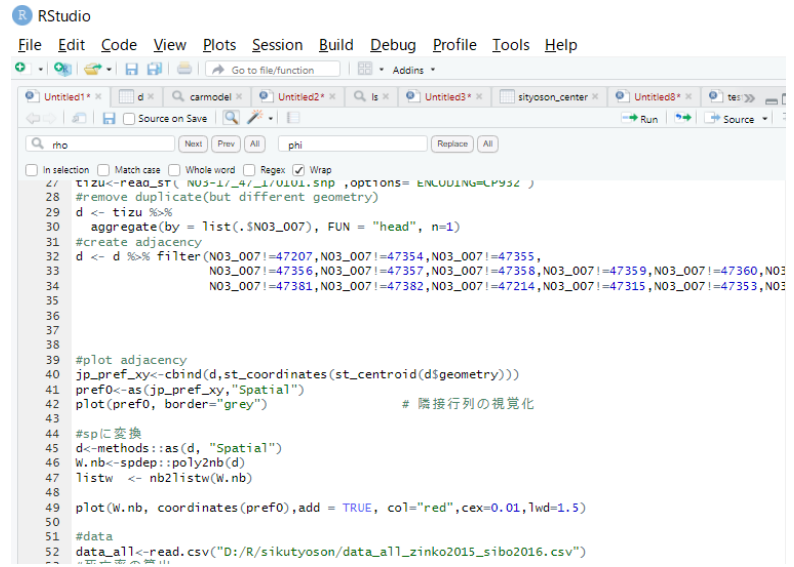


- データサイエンスでよく用いられるR言語にて日本の実データでも地域間の従属性を考慮した「空間モデリング」が行えないか？

R言語とは

- R言語はオープンソース・フリーソフトウェアの統計解析向けのプログラミング言語及びその開発実行環境
- データサイエンスに用いる言語であり、Pythonとともに人気の高い言語
- 統計解析に強みがあり、アクチュアリーにも人気が高い
- グラフィカルな図を描け、GISも行える
- 本発表ではRにてモデリング等を行っている





モデルとは

○モデルとは

- プラ「モデル」はプラスチック素材によるモデルであり、例えば飛行機のプラモデルは質量は実物の飛行機とは大きく異なり機能も無視しているが、形や色は精工にできている
- このようにモデルはあくまでも本物ではなく不必要な機能は落として、必要なエッセンスのみを取り上げたもの
- 統計を用いて何らかの事象を数式で表すものを「統計モデリング」という

○モデリングの効果

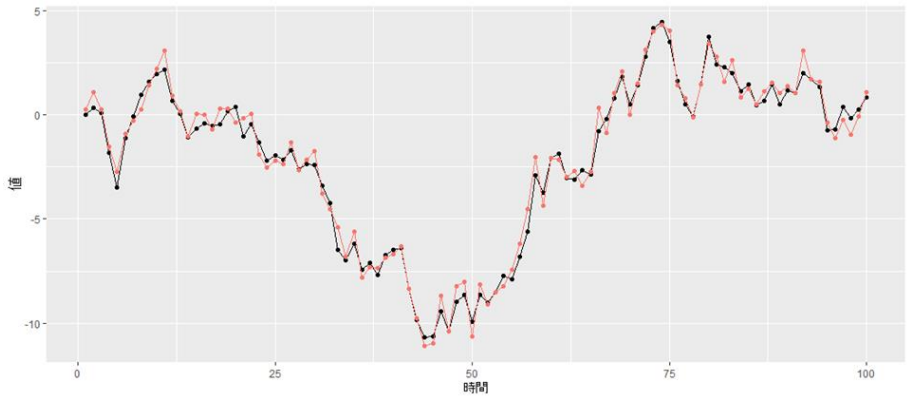
- 観測値の増減に対して、合理的に要因の変動要因が説明可能となる
- ある観察事象はたまたま得られた結果であると考えられ、全く同一の条件でも観測結果は変わると考えられる。モデリングにより、真の値に近い値を求められるかもしれない
- 例えばある地域の観測値がたまたま0であった場合に、実績値のみを用いると保険料の設定ができないが、モデリングで推計する場合は料率の設定が可能となる
- 空間モデリングでは不確実性を考慮して背後に潜む空間パターンを解析することができる

時系列モデルと空間自己相関モデル

○時系列モデル

- 時間的に変化する事象について、モデル化したもの
- 時刻tの現象は時刻t-1に依存するとする右図のモデルが事例としてある（$\rho = 1$ならランダムウォーク）
- 各事象は時間的に独立ではなく、時刻が近いほど同じような結果が得られやすい

似ている



$$y_t = \rho y_{t-1} + \varepsilon$$

時刻 tの状況はその一つまえの$t-1$にρを乗じて、更にランダム項としてεを追加したものとして表される

○空間自己相関モデル

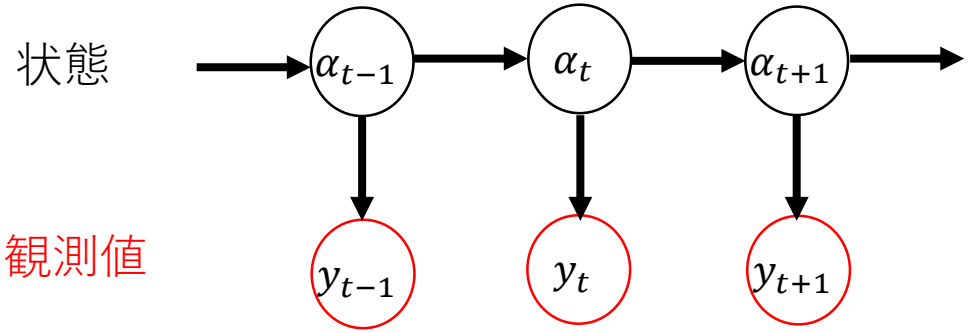
- 地理的に変化する事象について、モデル化したもの
- 各事象は地域について独立ではなく、地域相関がある
- 例えば、何らかの統計について、宜野湾市の観測値は浦添市や西原町の観測値に影響を受けている可能性が考えられるが、一方で国頭村の影響は受けていないかもしれない
- 本発表では空間モデルのうち、CARモデルを取り上げる

状態空間モデル（時系列モデル・空間モデル）

○状態空間モデル

- 状態空間モデルとは、実際の状態を表す変数と実際に観測できる変数が異なるような系を数式で表現したもの
- 状態は実際には観測されず、実際に観測されるのは観測値のみ

ローカルレベルモデルの例



観測モデル　$y_t = \alpha_t + \varepsilon_t\,, \epsilon_t \sim N(0, \sigma_\varepsilon^2)$

システムモデル　$\alpha_t = \alpha_{t-1} + \eta_t\,, \eta_t \sim N(0, \sigma_\eta^2)$

ε_tは観測誤差、 η_tはシステム誤差

○状態空間モデルのメリット

- 観測値は観測誤差を含んでおり、状態空間モデルでは真の値を推定することができる。
- 例えば、何らかの死亡率は実際に観察された死亡のみが観測値であり、観測されていない死亡がある。

空間隣接行列と空間重み行列

○空間隣接行列と空間重み行列

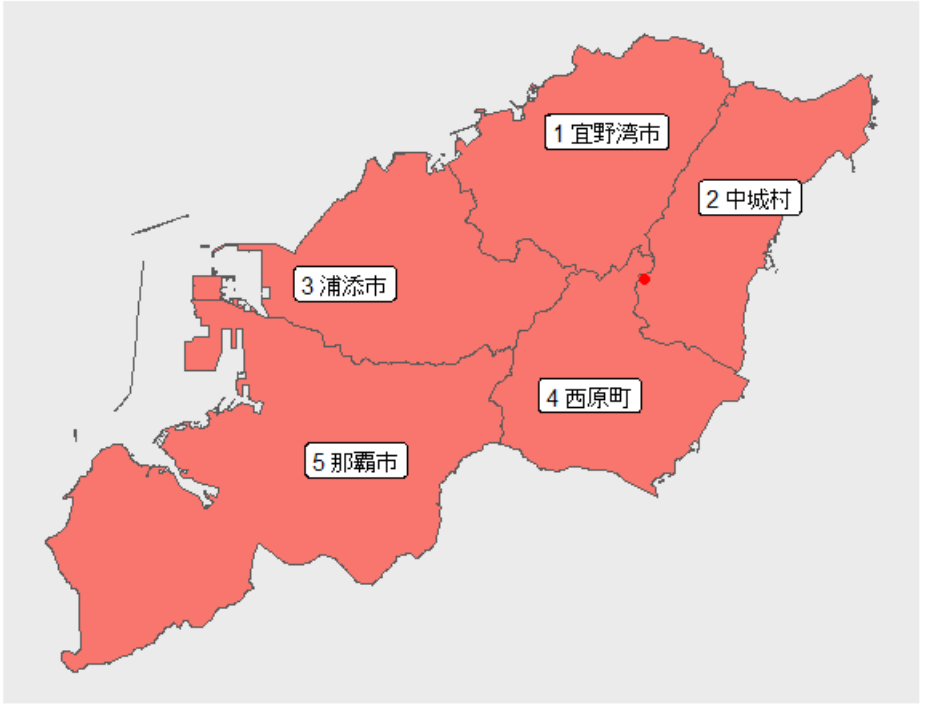
- 空間隣接行列Cは地域の隣接状況を行列形式で表したもの
- 空間重み行列Wは隣接状況より何らかの重みづけをしたもの

○５市町村による事例

- $y = \begin{pmatrix} y_{1\text{宜野湾}} \\ y_{2\text{中城}} \\ y_{3\text{浦添}} \\ y_{4\text{西原}} \\ y_{5\text{那覇}} \end{pmatrix}$のとき、空間隣接行列$C$と空間重み行列$W$の一例はつぎのとおり

- $C = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$：地域が隣接していると1、隣接していないときは０

- $W = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$：標準化（この例では行の合計を１とする）

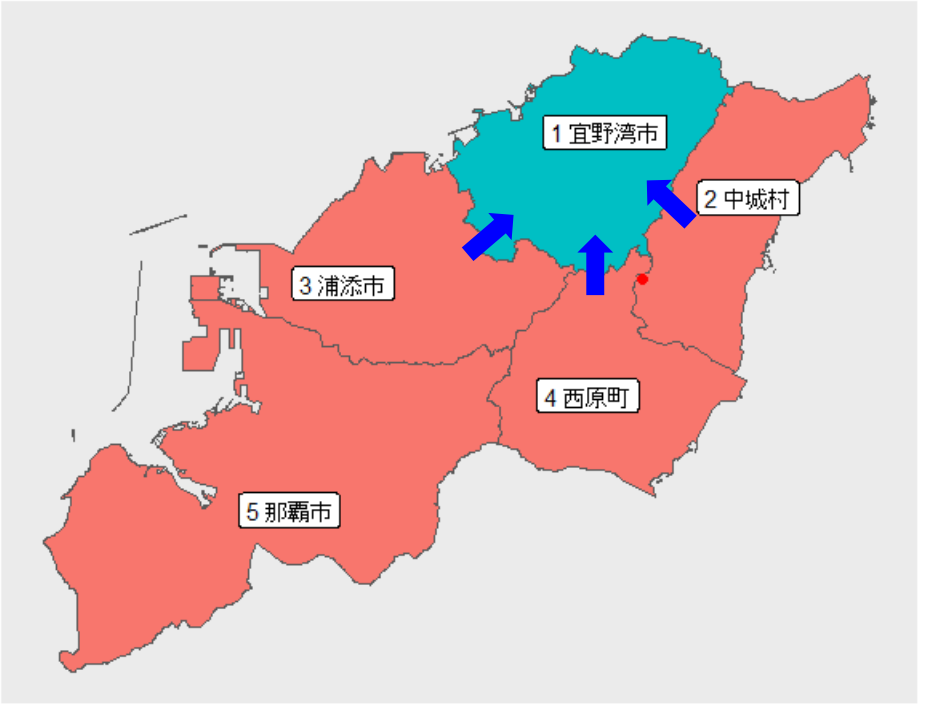


Cの1行目は宜野湾市を表しており、中城、浦添、西原に隣接し、那覇とは隣接していない

WはCのそれぞれの地域の隣接地域数で除すことにより求める

簡単な空間モデルの例

- $\boldsymbol{y} = \begin{pmatrix} y_{1\text{宜野湾}} \\ y_{2\text{中城}} \\ y_{3\text{浦添}} \\ y_{4\text{西原}} \\ y_{5\text{那覇}} \end{pmatrix}, W = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}, I = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$



$$\boldsymbol{y} = \rho W \boldsymbol{y} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$$

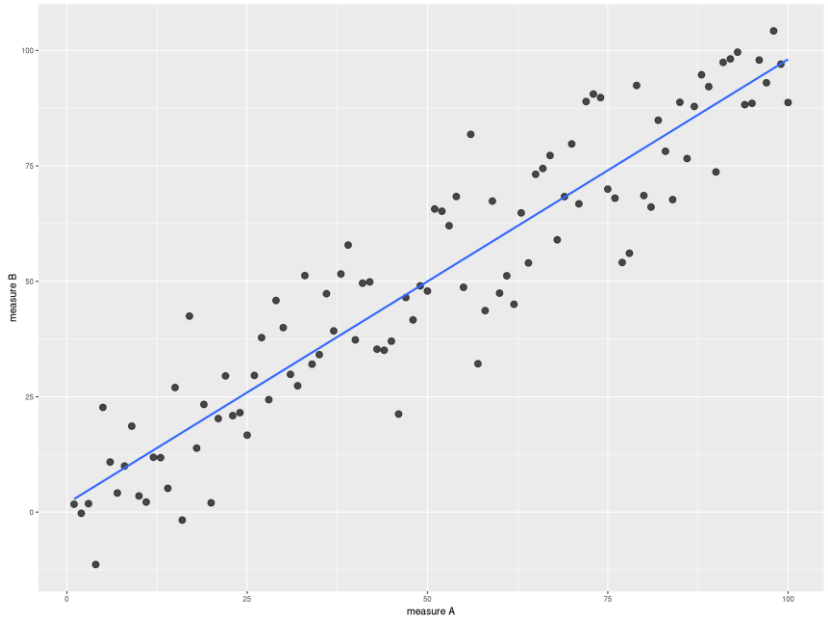
- $y_{1\text{宜野湾}} = \rho\left(\frac{1}{3} y_{2\text{中城}} + \frac{1}{3} y_{3\text{浦添}} + \frac{1}{3} y_{4\text{西原}}\right) + \varepsilon$

- 宜野湾の観測値は中城、浦添、西原のそれぞれ1/3の観測値の合計値をρ倍しランダム項εを加えたもの
- このモデルは簡単なモデルのため拡張が必要

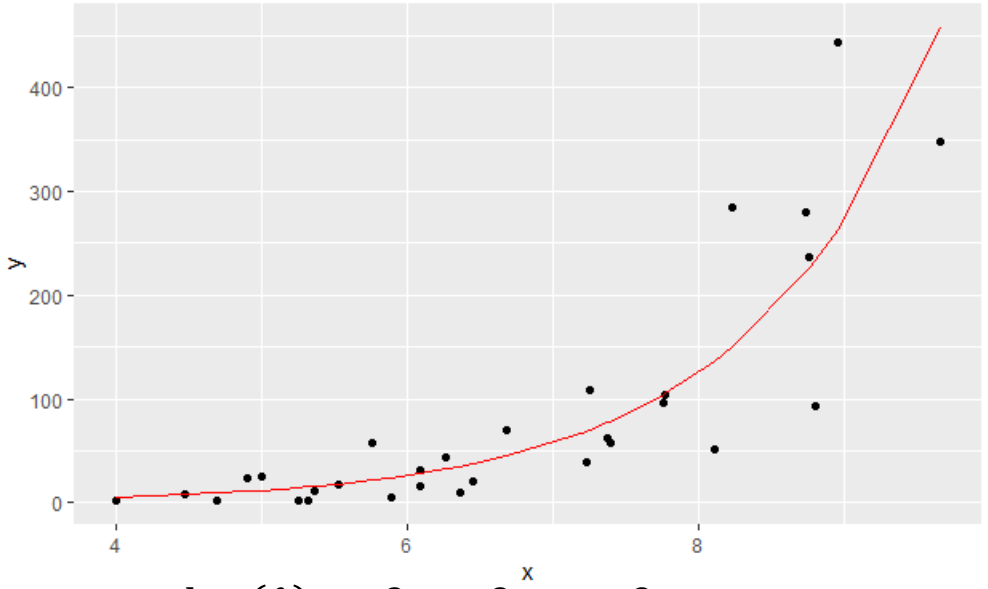
GLM（Generalized linear model:一般化線形モデル）とは

- 線形回帰（回帰直線）を一般化したもので、線形回帰の残差（実績値と予測値の差分）は正規分布に従っているが、GLMでは残差を任意の分布としたモデル
- 一般化線形モデルには線形回帰、ポアソン回帰、ロジスティック回帰などが含まれ、損害保険で基本的な予測モデリングの技法

線形回帰



GLM（ポアソン回帰）



$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

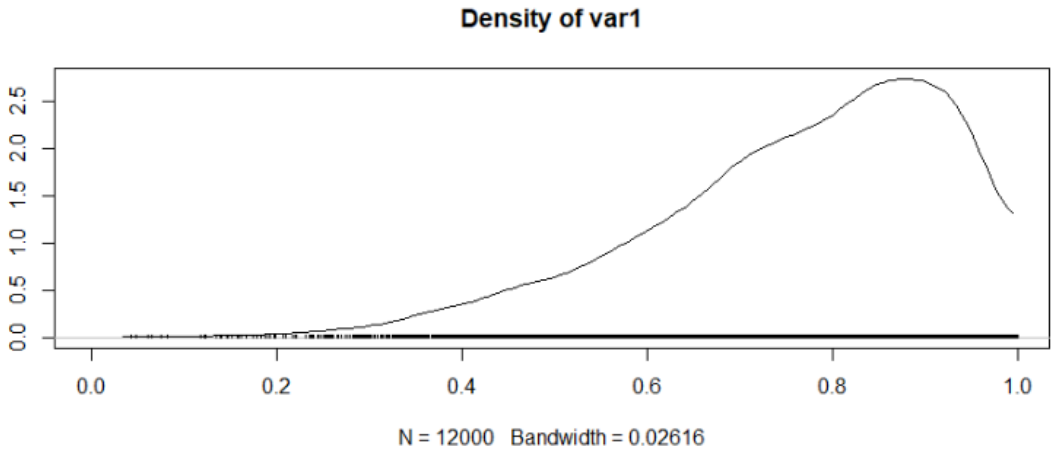
階層ベイズとは

- GLMを拡張した階層ベイズでは似たようなパラメータに共通の制約を与えることによってデータの個数よりもパラメータの多い場合でも統計モデルをあてはめることができる
- ベイズ推計では最尤法と異なり、事後分布の形を推定することができる
- 後述のCARモデルでは地域を変数とするが、階層ベイズを用いることにより、変数の数を抑えることができ、また、事前分布により条件を縛ることで滑らかさを表現する

＜ベイズの定理＞

$$\mathrm{P}(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \propto P(X|\theta)P(\theta)$$

事後分布＝尤度×事前分布



本論文で用いたCARBayesパッケージではMCMC（マルコフチェーン・モンテカルロ）により各パラメータを推定（上記はρ）

課題

○ 例えばGLMによる頻度モデルを考える

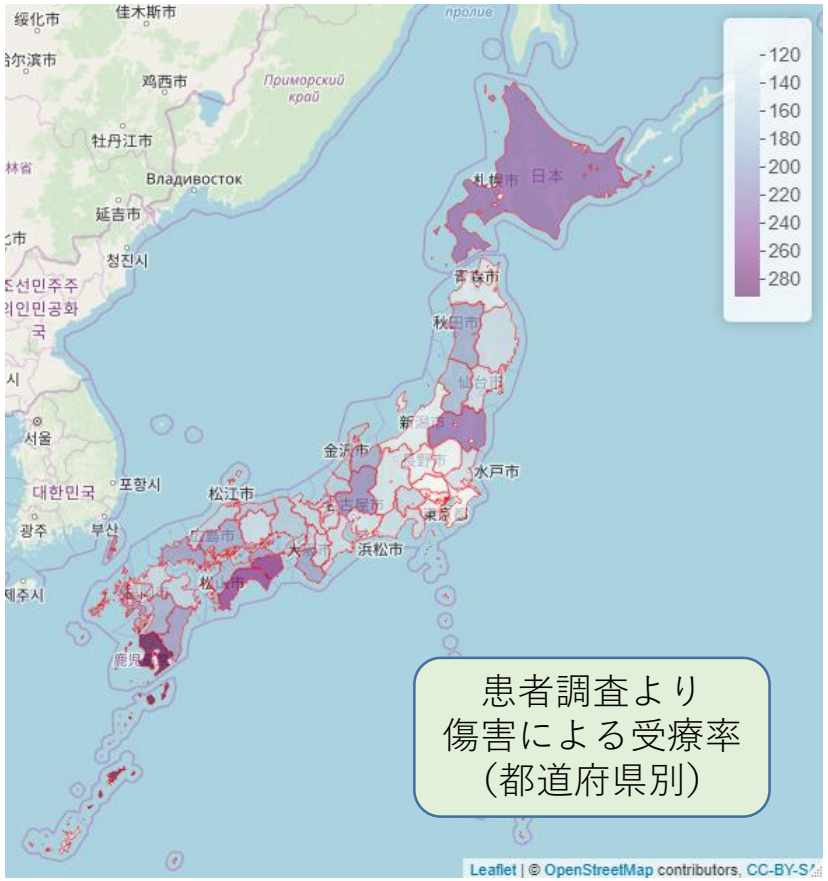
一般統計には地域を変数に持つものがよくある

- 患者調査（厚生労働省）
- 地域がん登録（国立がん研究センター）
- 人口動態調査（厚生労働省）

○ 地域間に何らかの影響が見られることもよくある

- 入院発生率（地域間のベッド数の差異）
- がん罹患率（生活習慣による地域差）

○ しかし、地域を変数に用いることはあまりない



患者調査より
傷害による受療率
（都道府県別）

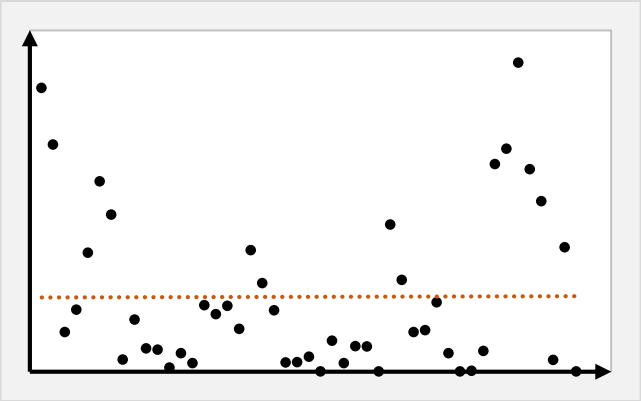
課題

○ なぜか

使わない場合

○ 地域間の差異が明確な場合には説明力が不十分

○ 精度が低い

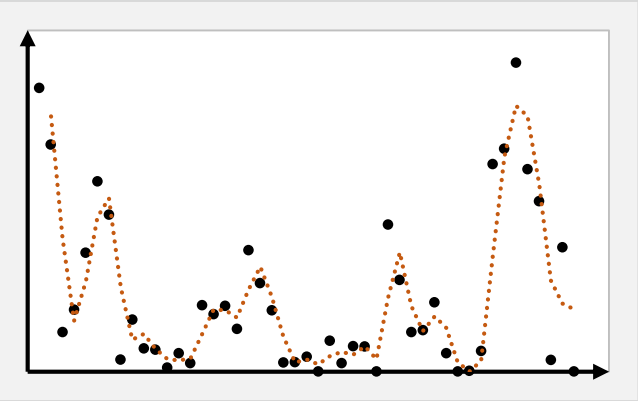


どっちもどっち

使う場合

○ 各地域を独立に見るため、地域間の影響を反映しない

○ データが地域ごとに分断されることで信頼性が低下



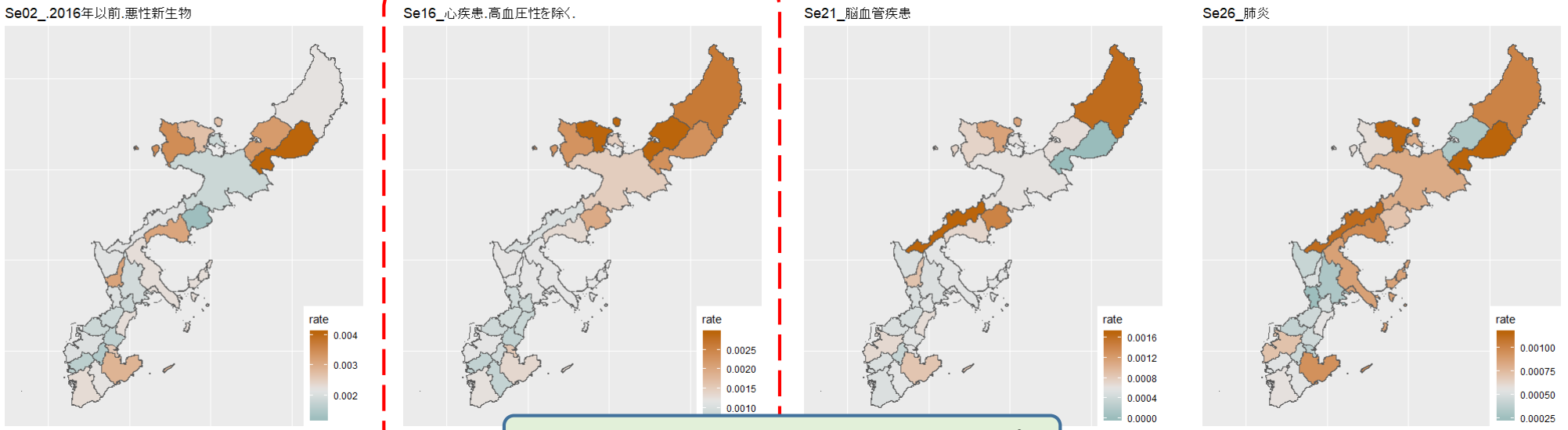
地域間の影響を考慮したモデリングがしたい

課題

○ 使用するデータ
- 人口動態調査（厚生労働省）より市町村・死因・性別の死亡者数
- 母数として、人口推計（厚生労働省）

○ 沖縄県（うち本島の26市町村）に限定して、市町村別に見る

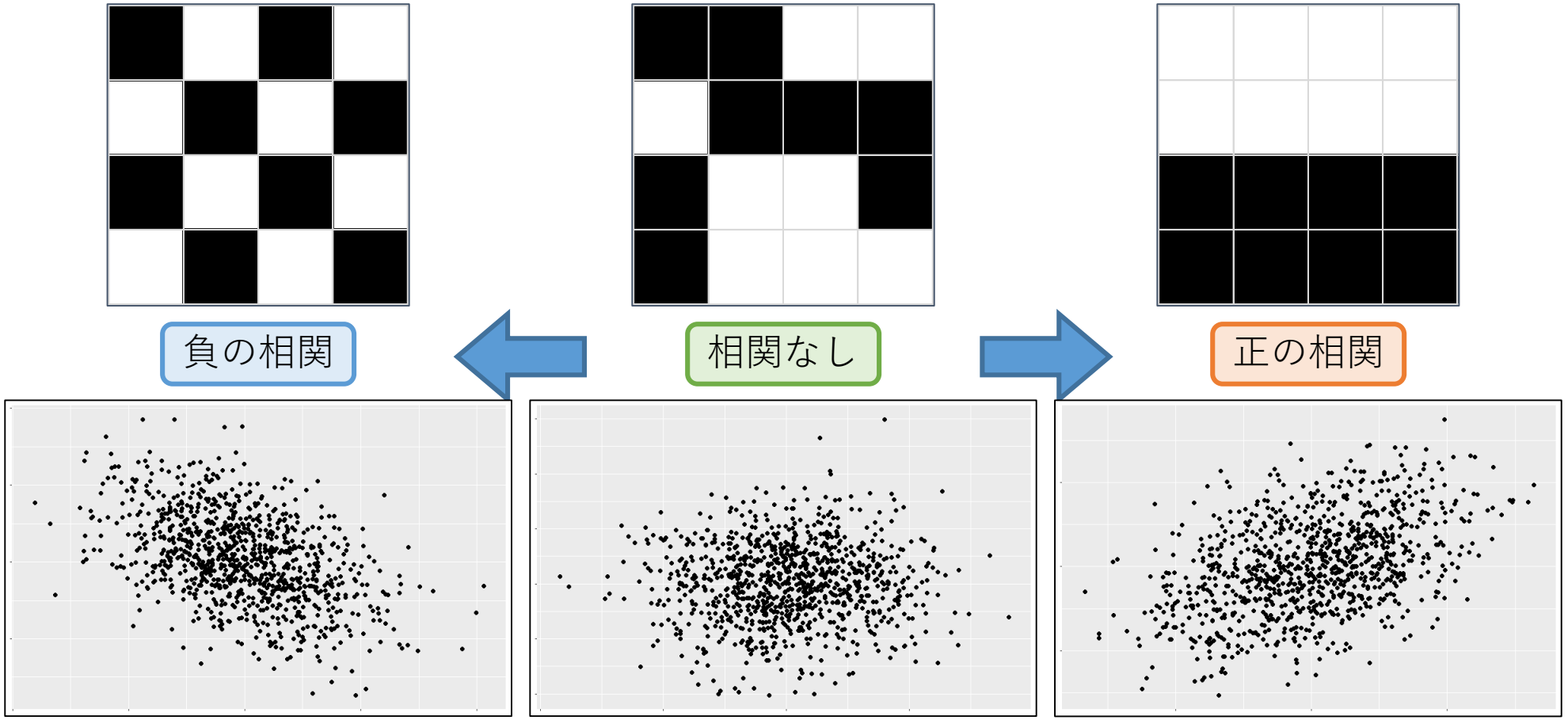
○ 死因のうち、地域間の影響がありそうな心疾患を分析



地域間の影響を測る

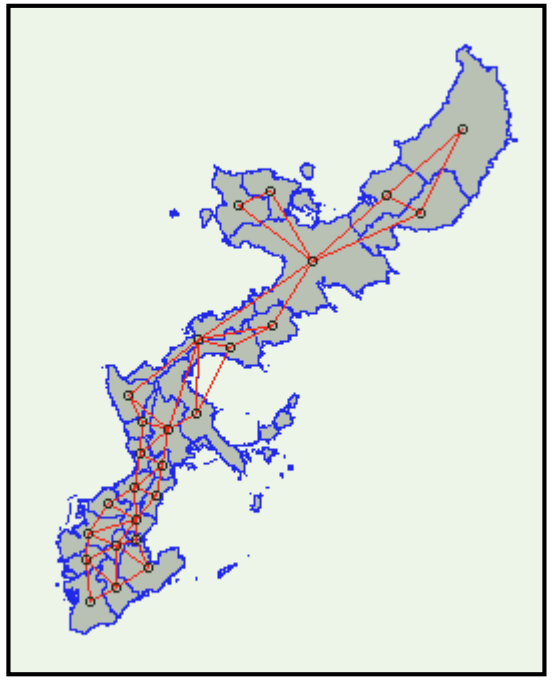
○ モデリングの前に、地域間の影響の有無を探索する

○ 空間自己相関・・・近くにある地域間の観測値が持つ相関

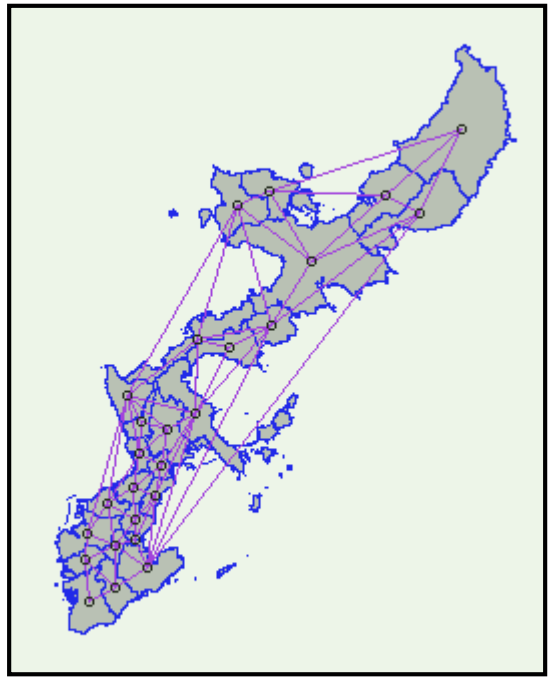


地域間の影響を測る

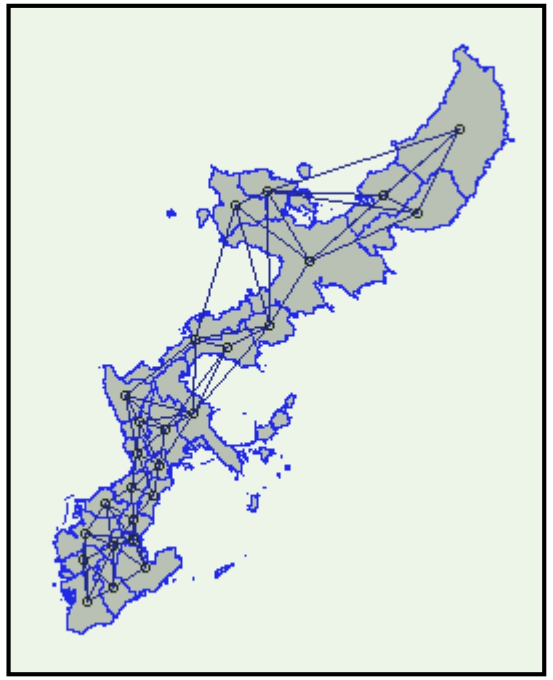
○ 隣接関係を定義する・・・色々な隣接の型



隣接（ルーク型）



ドロネー三角網



近隣4ゾーン

地域間の影響を測る

○ モランI統計量

$$\rho_I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(z_i - \bar{z})\left(z_j - \bar{z}\right)}{\sum_i (z_i - \bar{z})^2}$$

- z_i：地域iの観測値
- $\bar{z}$：観測値の平均
- w_{ij}：地域iと地域jが隣接する場合は1、そうでない場合は0
- n：地域の数

○ -1（負の相関）～1（正の相関）の値を取る

○ R package「spdep」のmoran.test関数を使う

地域間の影響を測る

○ モランI統計量（結果）

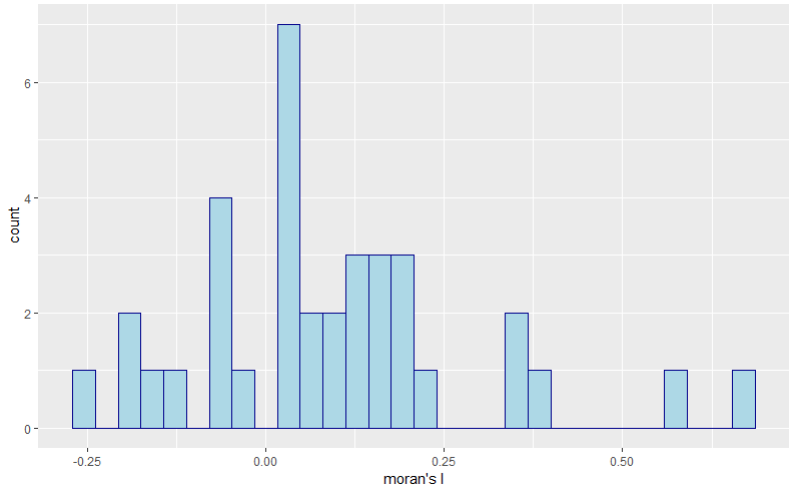
```
> moran

        Moran I test under randomisation

data:  tizu_data_genin$rate
weights: listw

Moran I statistic standard deviate = 5.1373, p-value = 1.393e-07
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.66021274       -0.04000000        0.01857753
```

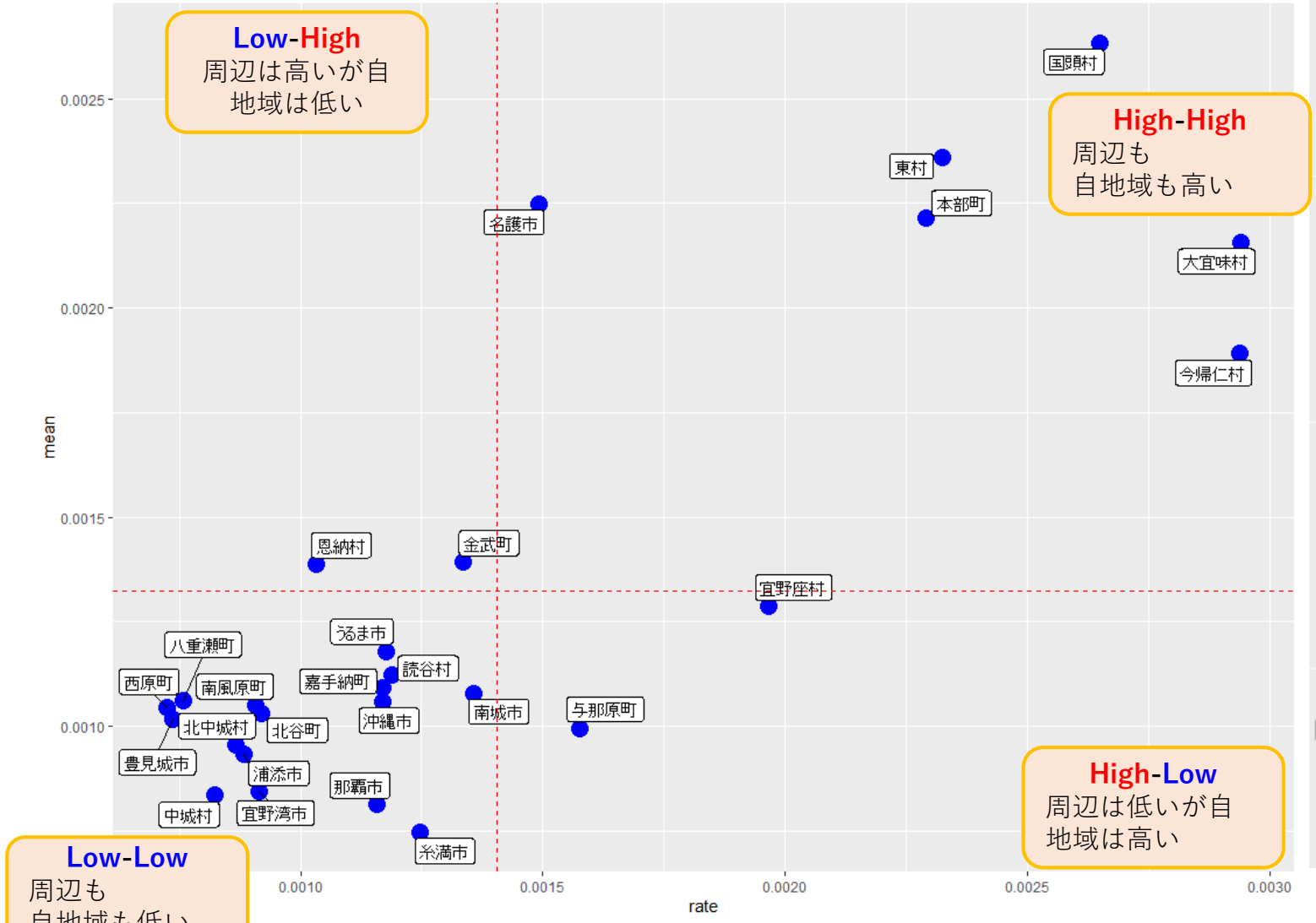
人口動態調査による
死因別死亡率の場合

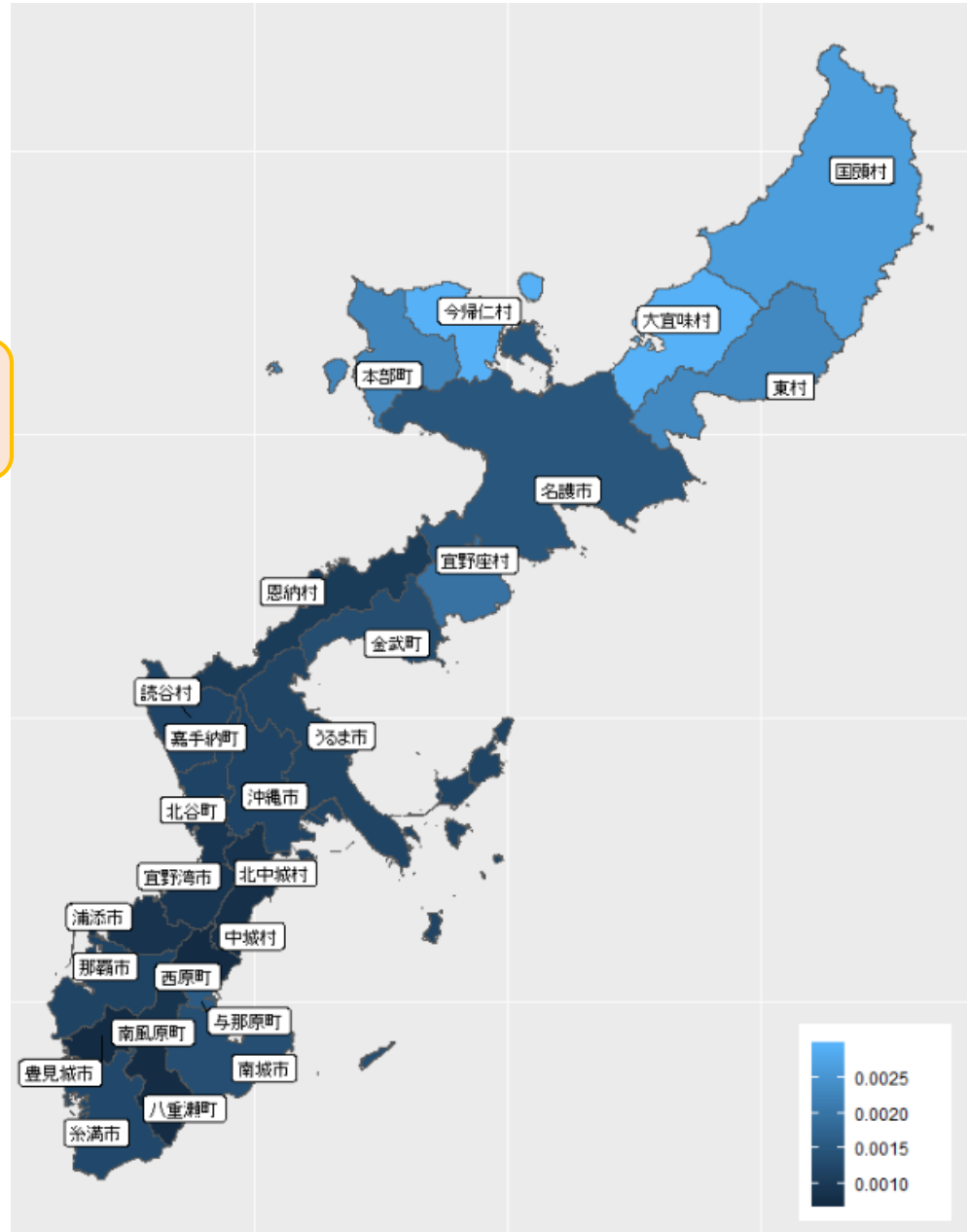


	youin	moranI
17	Se16_心疾患.高血圧性を除く.	0.660212741904978
1	Se00_総数	0.559095428558101
19	Se18_その他の虚血性心疾患	0.398444883674511
14	Se13_白血病	0.346600763211086
12	Se11_乳房の悪性新生物.腫瘍.	0.346107712104762
27	Se26_肺炎	0.215698556355638
30	Se29_肝疾患	0.204520918146924
29	Se28_喘息	0.192073103408867

地域間の影響を測る

○ （参考）心疾患のモラン散布図





地域間の影響を考慮したモデリング

○ 条件付自己相関モデル（Conditional AutoRegressive model）

・・・地域k（$k = 1, \cdots, K$）の値y_kが次式の条件付分布に従う

$$y_k | y_{-k} \sim N\left(\mu_k + \sum_{j=1}^{K} c_{kj}(y_j - \mu_j), {\sigma_k}^2\right)$$

- c_{kj}：定数、$c_{kk} = 0$・・・地域間の従属関係を表す
- y_{-k}：地域k以外の全ての地域

○「y_kは周辺地域の値y_jと従属性c_{kj}で決まる」という事前分布を与える

地域間の影響を考慮したモデリング

○ $\boldsymbol{y} = \{y_k\}$の同時密度関数が一意に存在する（Brook's lemma）

$$\boldsymbol{y}\left(= \begin{pmatrix} y_1 \\ \vdots \\ y_K \end{pmatrix}\right) \sim N_K\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix}, (I_K - C)^{-1} M_\sigma\right)$$

- $C = \{c_{kj}\}$：$K \times K$行列
- M_σ ：$(M_\sigma)_{kk} = {\sigma_k}^2$の$K \times K$対角行列
- ${c_{kj}\sigma_k}^2 = {c_{jk}\sigma_j}^2$を満たす

○ $\boldsymbol{y}$の分散構造が（特に精度行列が${M_\sigma}^{-1}(I_K - C)$として）決まっているため、計算効率がよい

○ この条件の中で色々なモデルが存在する・・・例えば

地域間の影響を考慮したモデリング

○ かんたんなCARモデルの例

・・・Intrinsic CARモデル（ICAR）＋混合モデル

$$y_k = x_k\beta + \phi_k \quad , k = 1, \cdots, K$$

固定効果　変量効果

$$\phi_k | \phi_{-k} \sim N\left(\frac{\sum_j w_{kj}\,\phi_j}{\sum_j w_{kj}}, \frac{\sigma^2}{\sum_j w_{kj}}\right)$$

- ϕ_{-k}：地域k以外の全ての地域に関する変量効果
- w_{kj}：地域kと地域jが隣接する場合は1、そうでない場合は0
- σ^2：分散パラメータ

⇒「各地域は隣接する地域の加重平均」と考える

地域間の影響を考慮したモデリング

○ 色々なCARモデル・・・R package「CARBayes」より

種類	条件付分布	特徴
BYM	$\phi_k = \phi_k^{(1)} + \phi_k^{(2)}$ $\phi_k^{(1)} \Big\vert \phi_{-k}^{(1)} \sim N\left(\frac{\sum_j w_{kj}\phi_j^{(1)}}{\sum_j w_{kj}}, \frac{\tau^2}{\sum_j w_{kj}}\right)$ $\phi_k^{(2)} \sim N(0, \sigma^2)$	ICARに無構造の分散$\phi_k^{(2)}$を加えることで、周辺地域に影響されない誤差を反映したモデル
Proper	$\phi_k \vert \phi_{-k} \sim N\left(\frac{\lambda \sum_j w_{kj}\,\phi_j}{\sum_j w_{kj}}, \frac{\tau^2}{\sum_j w_{kj}}\right)$	ICARに対してλを置くことにより分散共分散行列を正則にしたモデル
Leroux	$\phi_k \vert \phi_{-k} \sim N\left(\frac{\rho \sum_j w_{kj}\,\phi_j}{\rho \sum_j w_{kj} + 1 - \rho}, \frac{\tau^2}{\rho \sum_j w_{kj} + 1 - \rho}\right)$	ρにより変量効果の中で従属性の強さをコントロールできるようにしたモデル $\rho = 1$のときはICARに等しく、 $\rho = 0$のときは全て独立となる

地域間の影響を考慮したモデリング

○ 分析例（Lerouxモデル+GLM）

・・・死亡者数y_iを予測する

$$y_i \sim Poisson(\lambda_i)$$

$$\log \lambda_i = \log E_i + x_i \beta_{gender} + \phi_{k(i)}$$

- 地域の変量効果$\phi_{k(i)}$にはLerouxモデルを適用する
- E_i：人口（オフセット項）
- β_{gender}：性別

○ 比較としてGLM（地域変数（固定効果）の有無別）も見る

- $\log \lambda_i = \log E_i + x_i \beta_{gender}$（地域変数なし）
- $\log \lambda_i = \log E_i + x_{1i} \beta_{gender} + x_{2i} \beta_{region}$（地域変数あり）

Leroux+GLMの解説

○ $y_i \sim Poisson(\lambda_i)$

- 心疾患による死亡者数はポアソン分布に従う。ただし観察値の属性（性別や市町村）により、ポアソン分布のパラメータは異なる

○ $\log \lambda_i = \log E_i + x_i \beta_{gender} + \phi_{k(i)}$

- 式変形すると$\lambda_i = E_i \cdot \exp(x_i \beta_{gender}) \cdot \exp(\phi_{k(i)})$
- リンク関数として対数をとることで、線形予測子としてポアソン分布のパラメータλ_iを表すことができる

○ $\phi_{k(i)} | \phi_{-k(i)} \sim N\left(\frac{\rho \sum_j w_{k(i)j} \phi_j}{\rho \sum_j w_{k(i)j} + 1 - \rho}, \frac{\tau^2}{\rho \sum_j w_{k(i)j} + 1 - \rho} \right)$

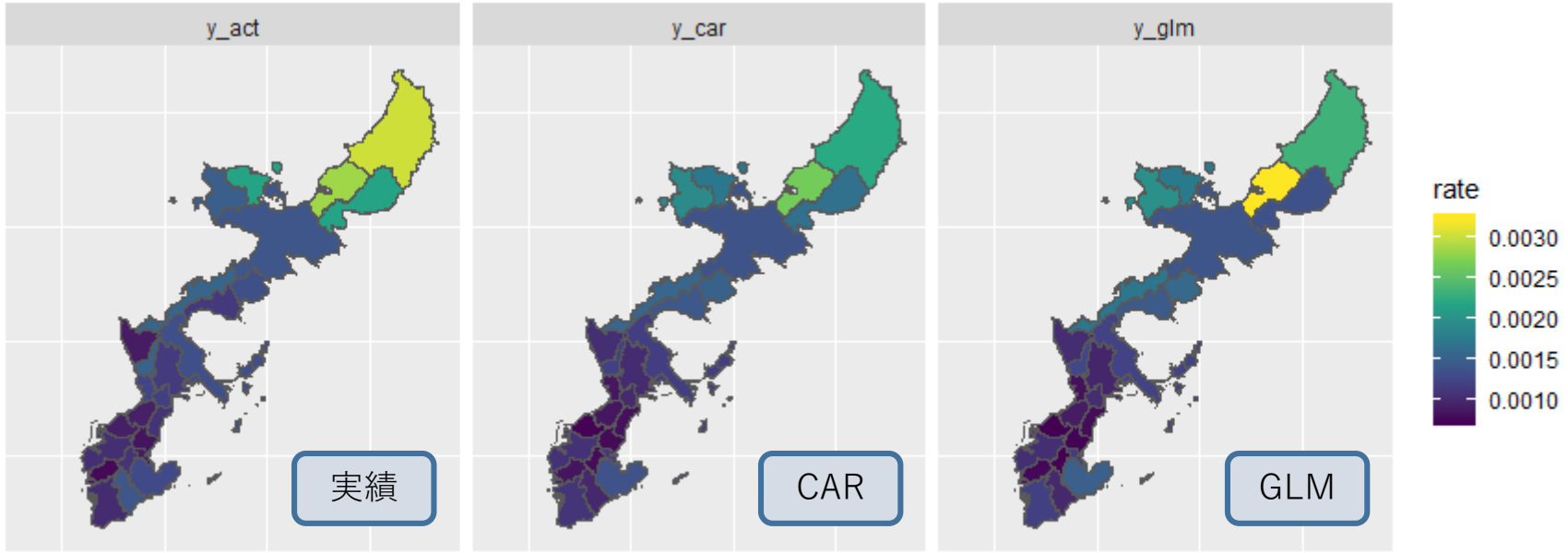
- 地域の変量効果を、次の階層構造を持つモデルでベイズ推計する
- $\tau^2 \sim \text{Inverse}-\text{Gamma}(1, 0.01)$、$\rho \sim \text{Uniform}(0,1)$

地域間の影響を考慮したモデリング

○ 分析例（Lerouxモデル+GLM）

- 2012年度～2016年度の5年分のデータを使用
- 2017年度を評価に用いる

○ 結果



地域間の影響を考慮したモデリング

○ 結果（地域変量効果の予測値（MAP推定量））

市町村	$\exp(\phi)$
那覇市	0.90
宜野湾市	0.72
浦添市	0.63
名護市	1.16
糸満市	0.92
沖縄市	0.85

市町村	$\exp(\phi)$
豊見城市	0.71
うるま市	0.99
南城市	1.15
国頭村	1.88
大宜味村	2.22
東村	1.38

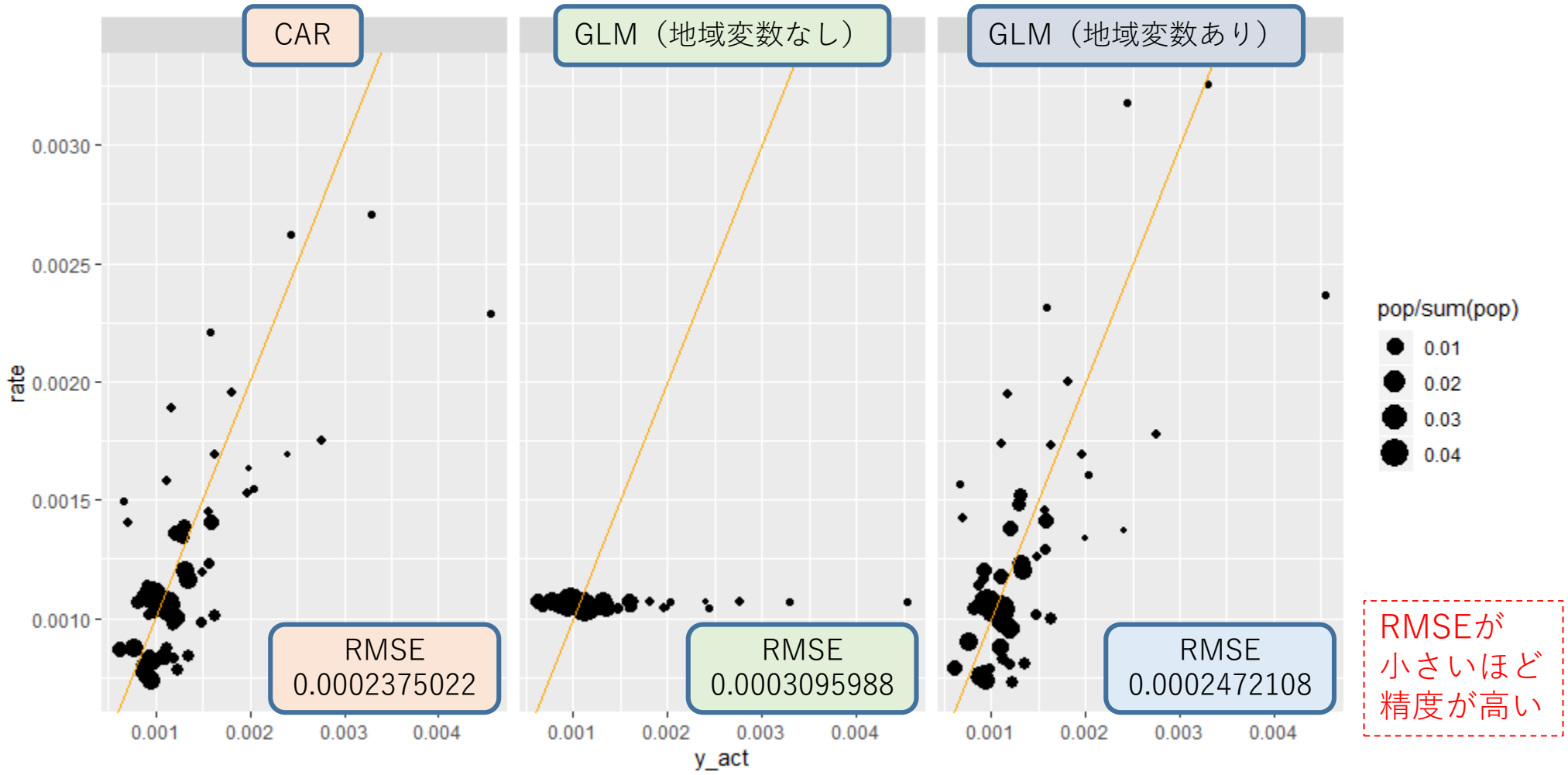
市町村	$\exp(\phi)$
今帰仁村	1.44
本部町	1.61
恩納村	1.30
宜野座村	1.27
金武町	1.19
読谷村	0.88
嘉手納町	1.01

市町村	$\exp(\phi)$
北谷町	0.72
北中城村	0.84
中城村	0.68
西原町	0.66
与那原町	0.94
南風原町	0.66
八重瀬町	0.83

- 数値は死亡率に対する相対的な倍率を表す。最大は大宜味村、次に国頭村、本部町と続き、北部地域が上位を占めている
- 最少は浦添市、その次に西原町、南風原町と中南部寄りの地域が多い
- 例えば地域の変量効果が死亡率に与える影響として、大宜味村と浦添市の格差は約３.5倍（$2.22/0.63$）ある

地域間の影響を考慮したモデリング

○ 結果（q-qプロット）



RMSEが
小さいほど
精度が高い

まとめ

○ 通常のGLMでは難しい、地域間の従属性という多面的な構造を反映したモデルとして、CARモデルを提案した

○ 今回のケースでは、CARモデルは、地域相関を持つデータに対して通常のGLMよりも（少し）高い予測精度を示した

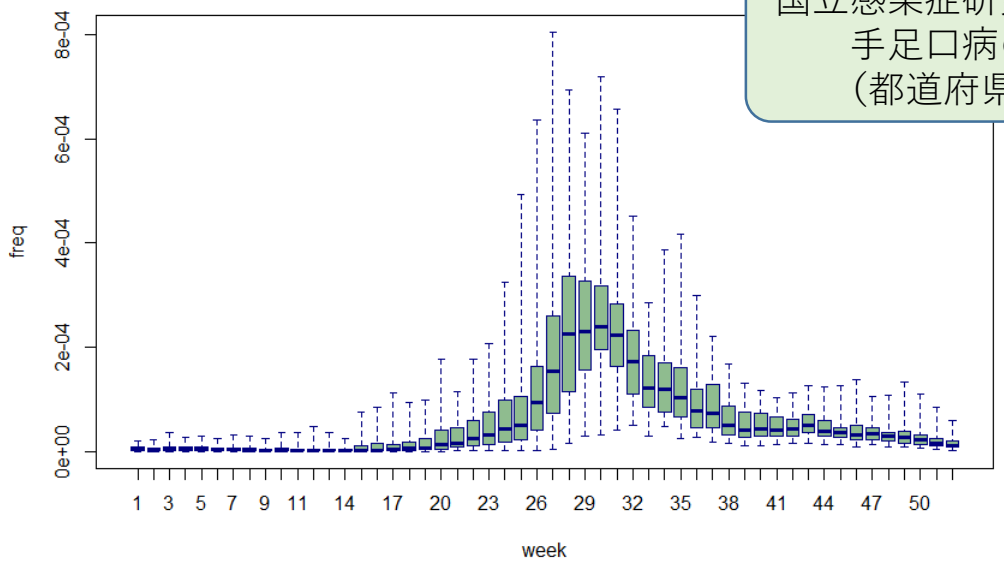
○ アクチュアリー業務のうち、結果より説明力が求められるケースが多い料率算定において、 CARモデルは、一定の条件に対して納得案のある事前情報を持つ特性から、使い勝手のよい手法であると考えられる

地域間の影響を考慮したモデリング（その先）

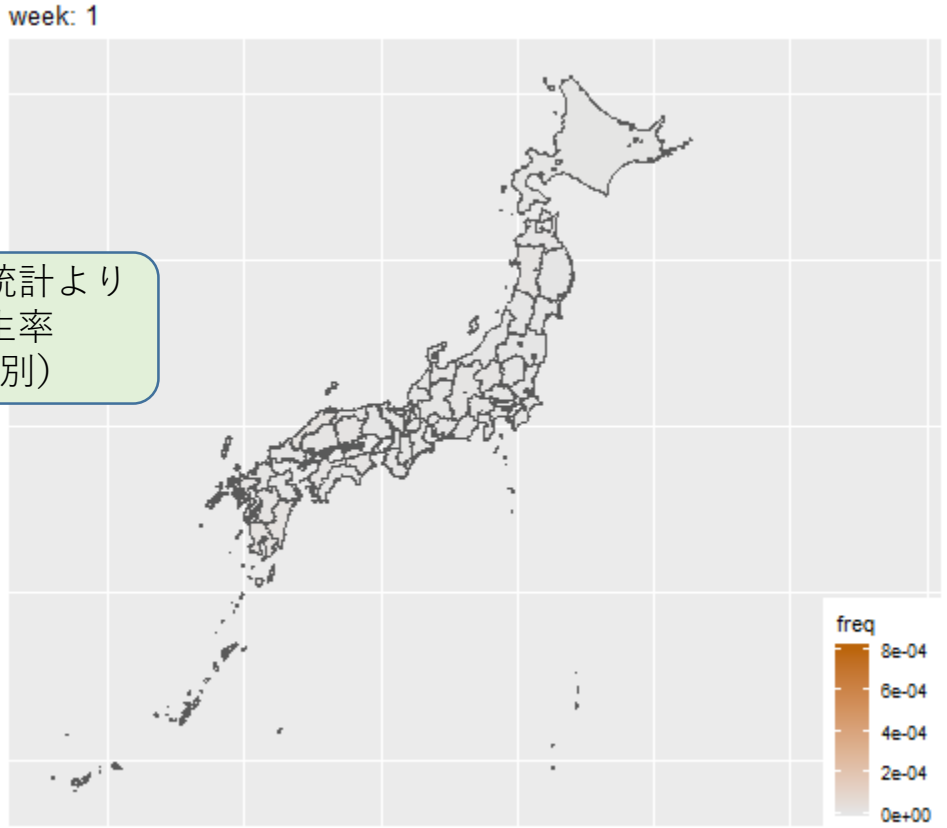
○ 地域間の従属性について、時間による変化を考慮する

・・・CARモデル×時系列モデル

- 変量効果に時系列を加える
- R Package「CARBayesST」が対応



国立感染症研究所統計より
手足口病の発生率
（都道府県・週別）



参考文献

- Shi, P., & Shi, K. (2017). TERRITORIAL RISK CLASSIFICATION USING SPATIALLY DEPENDENT FREQUENCY-SEVERITY MODELS. *ASTIN Bulletin, 47*(2), 437-465. doi:10.1017/asb.2017.7
- BRECHMANN, E. and CZADO, C. (2014) Spatial modeling. In Predictive Modeling Applications in Actuarial Science: Volume I, Predictive Modeling Techniques (eds. E.W. Frees, G. Meyers and R.A. Derrig), pp. 260–279. Cambridge: Cambridge University Press.
- C. Brunsdon, L. Comber, 湯谷啓明(訳), 工藤和奏(訳), 市川太祐 (訳) (2018) Rによる地理空間データ解析入門. 共立出版.
- Bivand R, Wong DWS (2018). “Comparing implementations of global and local indicators of spatial association.” *TEST*, 27(3), 716–748. https://doi.org/10.1007/s11749-018-0599-x.
- Duncan Lee (2013). CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors. Journal of Statistical Software, 55(13), 1-24. URL http://www.jstatsoft.org/v55/i13/.
- Lee D, Rushworth A, Napier G (2018). “Spatio-Temporal Areal Unit Modeling in R with Conditional Autoregressive Priors Using the CARBayesST Package.” _Journal of Statistical Software_, *84*(9), 1-39. doi: 10.18637/jss.v084.i09 (URL: https://doi.org/10.18637/jss.v084.i09).
- 金明哲(2010). Rで学ぶデータサイエンス 7 地理空間データ分析.共立出版.