漁業統計検討会 (清水)

「統計モデリングセミナー」 (2012 年 12 月) 投影資料

全部で 7 回中の 5 回目

階層ベイズモデルの応用

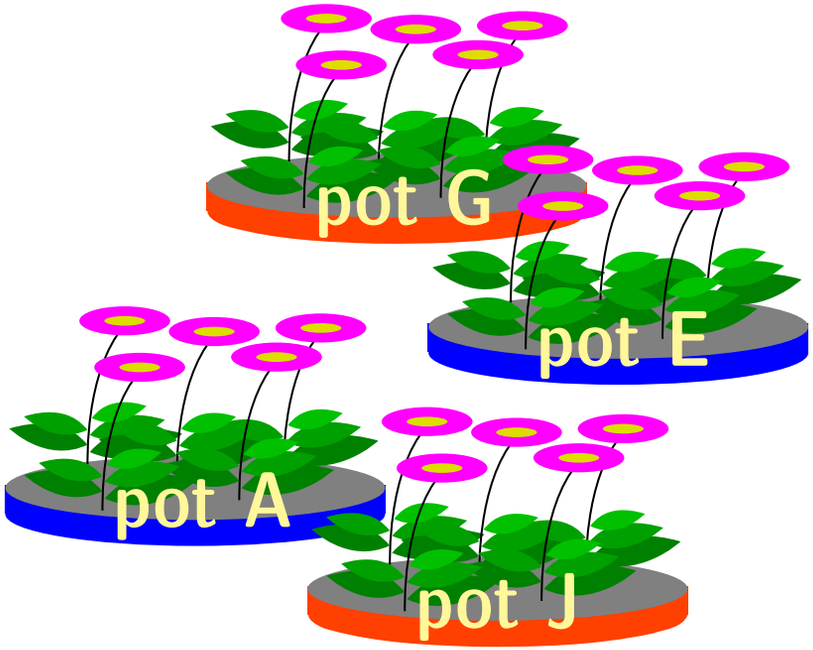
空間・時間構造などをあつかう

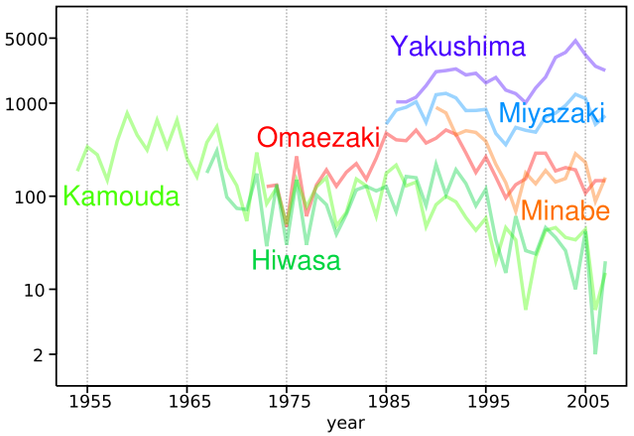
久保拓弥 kubo@ees.hokudai.ac.jp

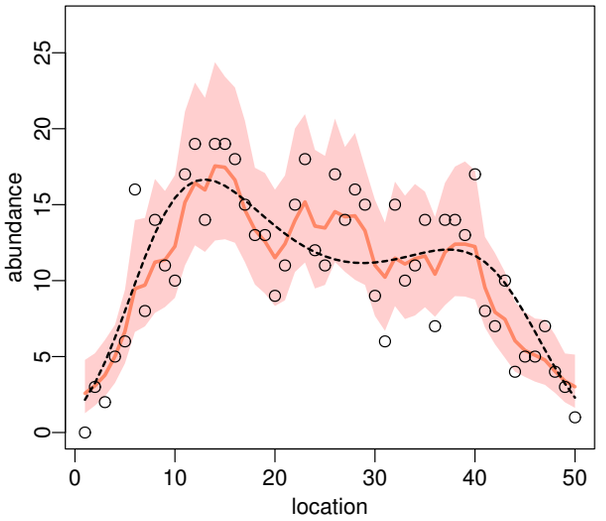
http://goo.gl/OyB2k

今回のハナシ: いろいろな階層ベイズモデル

1. 個体差 + ブロック差というネストしたランダム効果
2. 「隣と似ている」空間相関のあるランダム効果
3. 時間変化する潜在変数: ウミガメ上陸数の統計モデル







復習: ベイズ用語の整理

- ベイズ統計モデルでは (事後分布を) 推定したいパラメーターに事前分布を設定
- キモチとしては「推定したいパラメーターの範囲はこのへんでしょう」を表現している

(事後分布) $\propto$ (尤度) $\times$ (事前分布) $\times$ (超事前分布)

- 階層ベイズモデル

  $p(\beta, r, \tau|y) \propto p(y|\beta, r)p(r|\tau)p(\tau)p(\beta)$

  - 推定計算方法: **Markov Chain Monte Carlo (MCMC) 法**
    - MCMC 計算わざ 1: **Metropolis-Hastings 法**
    - MCMC 計算わざ 2: **Gibbs sampler**

(上のふたつは本質的には同じもの)

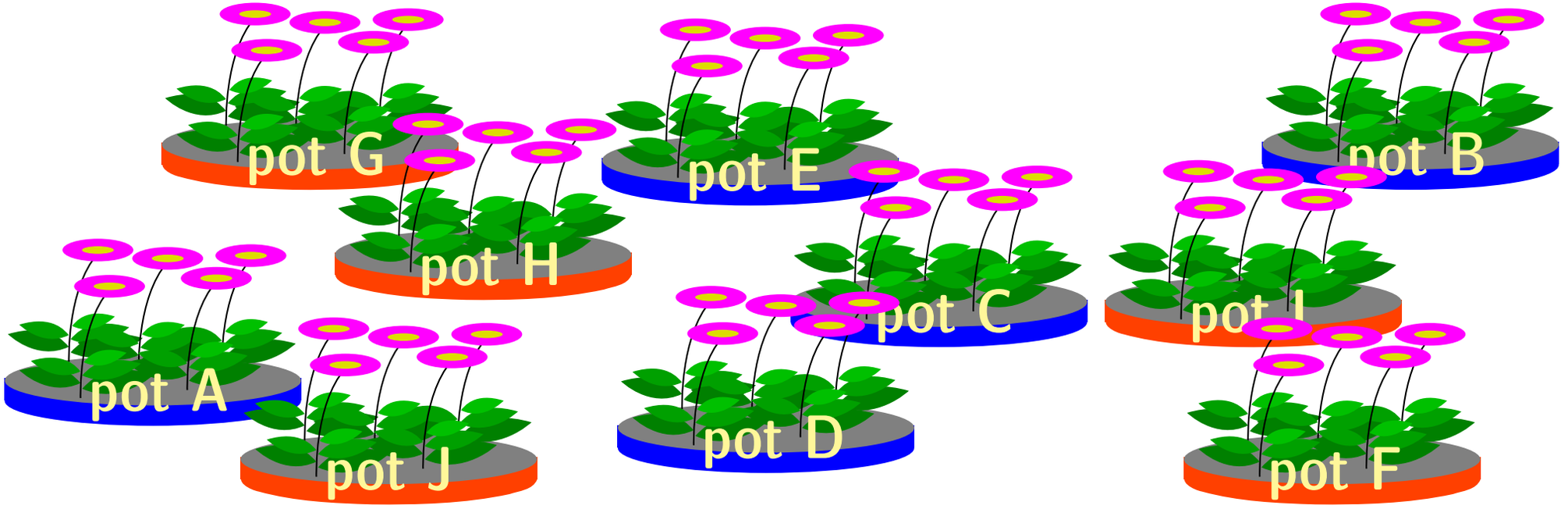
復習: よく使われる 3 種類の事前分布

1. 主観的な事前分布
   - 解析者が恣意的に決める事前分布
   - 評判が悪いので，可能なかぎり使わずにすませたい
2. 無情報事前分布
   - 「このパラメーターはどんな値でもいいんです」と一見中立的な立場を表現する
   - パラメーターの事前分布は可能なかぎりこれですませたい
3. 階層的な事前分布
   - 観測データを参照しつつ超パラメーターを変化させる
   - 個体差・場所差などなどを表現するときに必須の事前分布

複数ランダム効果の統計モデル

— 個体差 + 場所差 —

架空植物の例題: またまた種子数データ



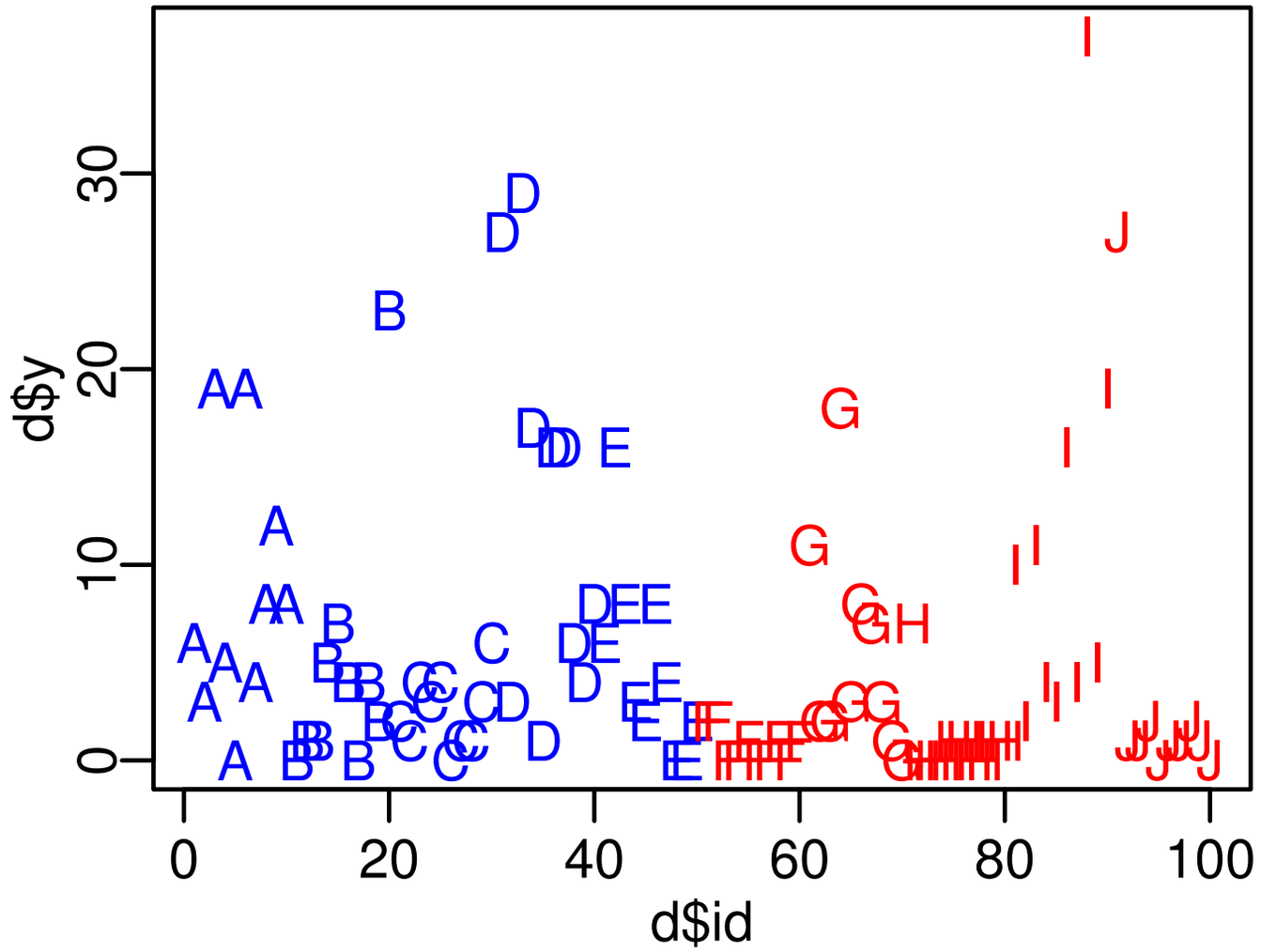
- 肥料をやったら個体ごとの種子数 y_i が増えるかどうかを調べたい
- 植木鉢 10 個，各鉢に 10 個体の架空植物 (合計 100 個体)
  - コントロール ($f_j =$ C) 5 鉢 (合計 50 個体)
  - 肥料をやる処理 ($f_j =$ T) 5 鉢 (合計 50 個体)

データはこのように格納されている

```
> d <- read.csv("d1.csv")
> head(d)
  id pot f  y
1  1   A C  6
2  2   A C  3
3  3   A C 19
4  4   A C  5
5  5   A C  0
6  6   A C 19
```

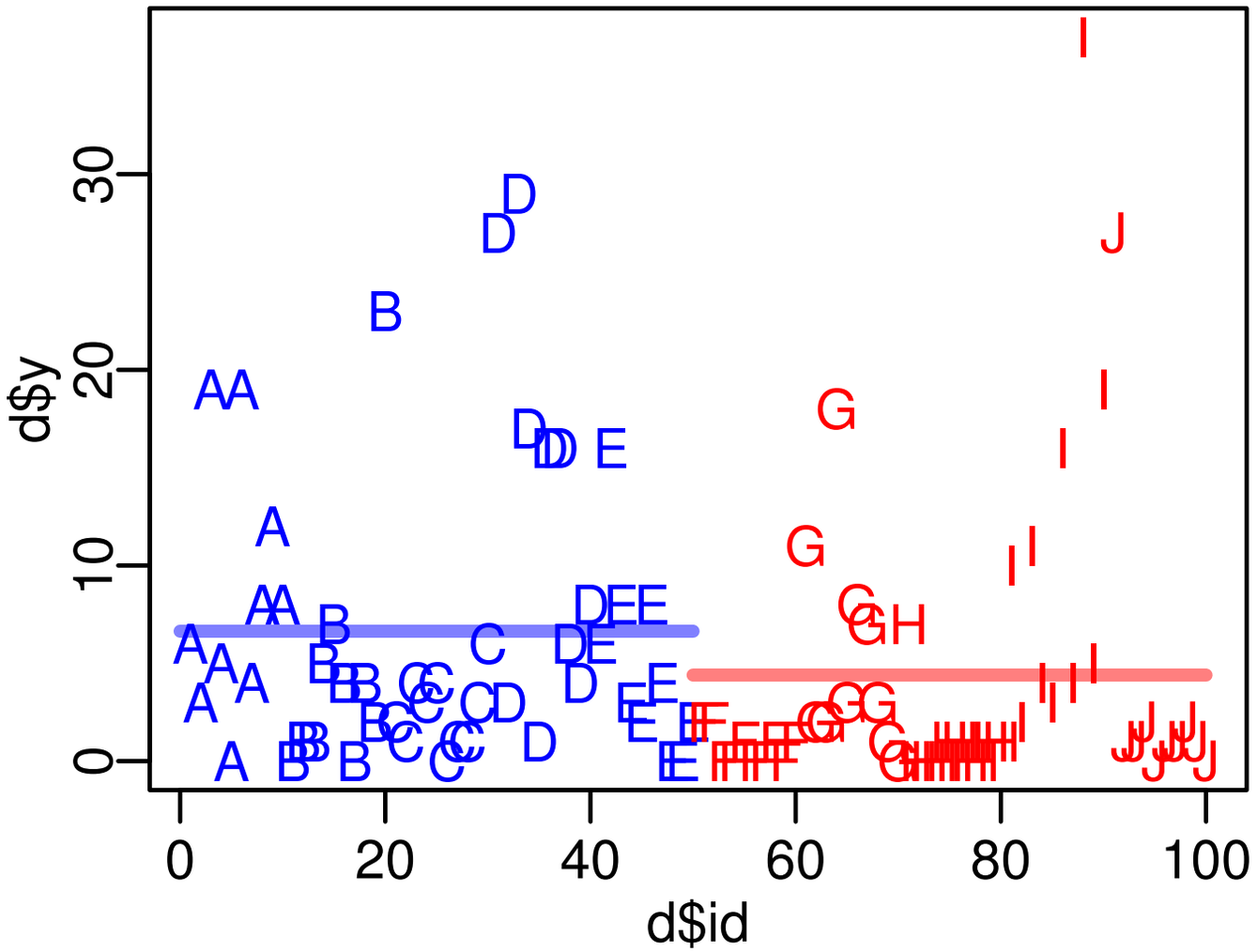
- id 列: 個体番号 $\{1, 2, 3, \cdots, 100\}$
- pot 列: 植木鉢名 $\{\mathsf{A}, \mathsf{B}, \mathsf{C}, \cdots, \mathsf{J}\}$
- f 列: 処理: コントロール C, 肥料 T
- y 列: 種子数 (応答変数)

データはとにかく図示する!!



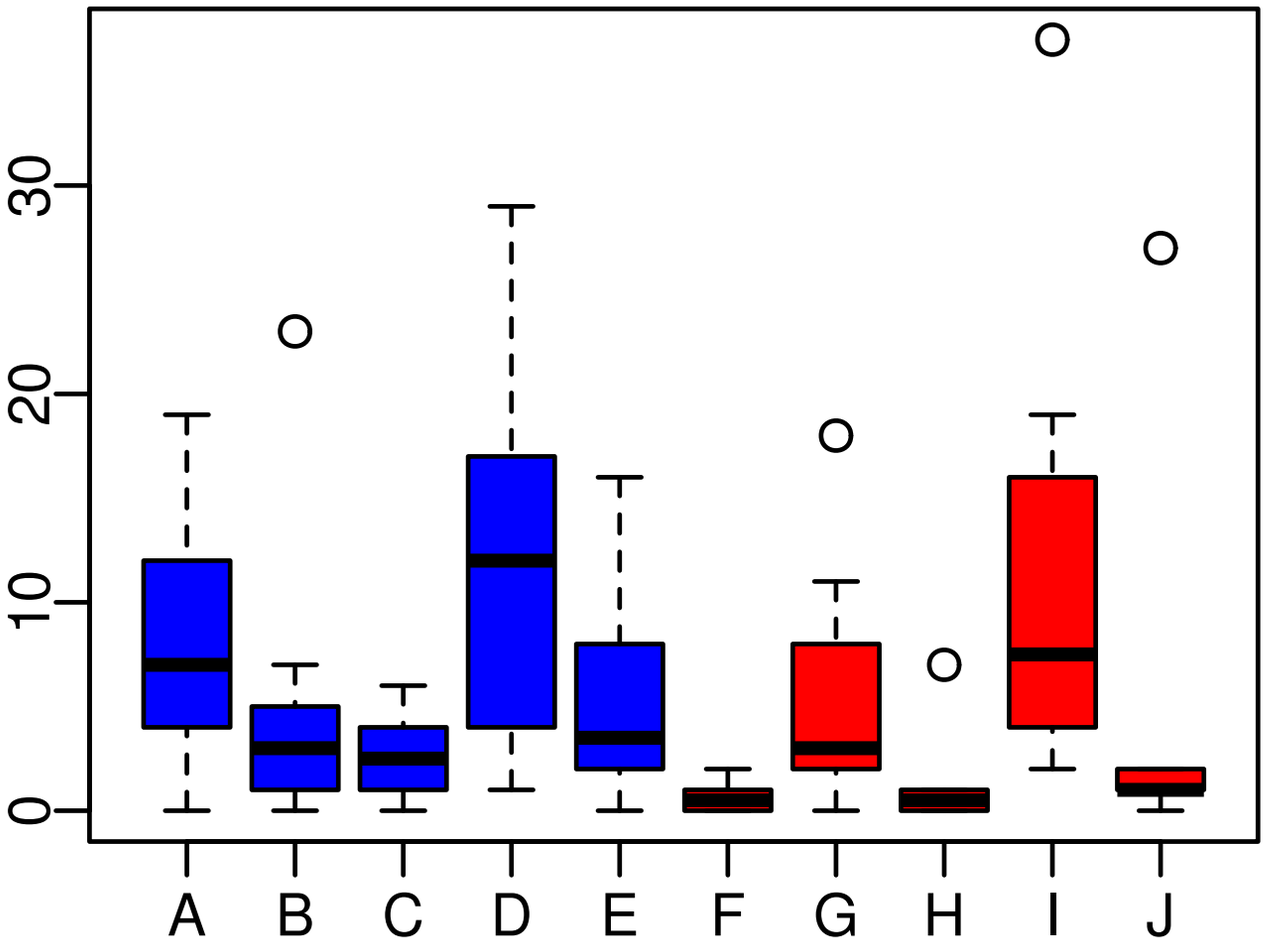
- `plot(d$id, d$y, pch = as.character(d$pot), ...)`
- コントロール・処理 でそんなに差がない?

処理ごとの平均も図に追加してみる



- むしろ 処理 のほうが平均種子数が低い?

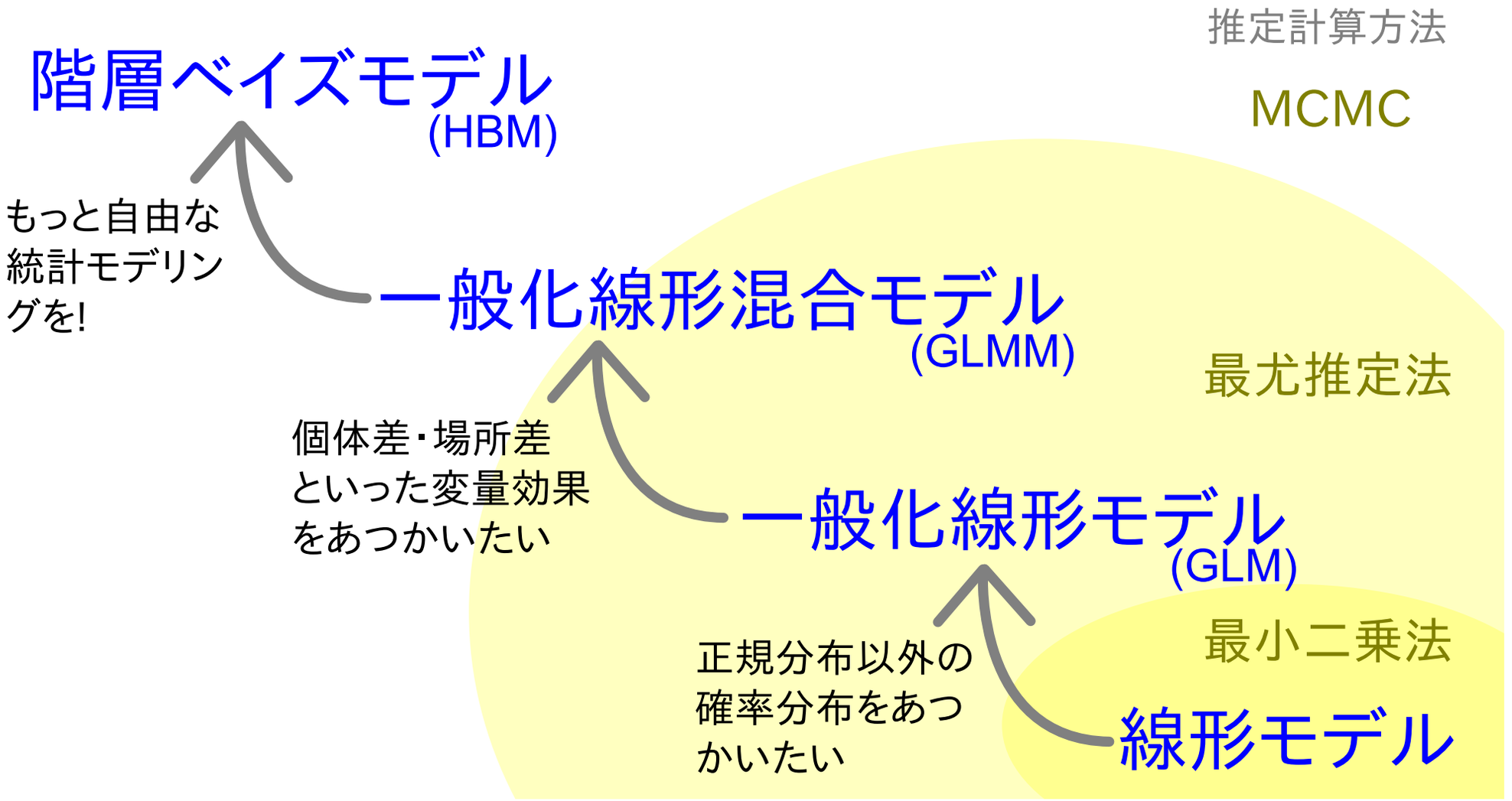
個体差だけでなく植木鉢差も？ overdispersion!



- `plot(d$pot, d$y, col = rep(c("blue", "red"), each = 5))`
- 植木鉢由来の random effects みたいなものはブロック差と呼ばれる

(一般化な) 線形モデルのわくぐみで，とりあえず考えてみる

線形モデルの発展



GLM: 個体差もブロック差も無視

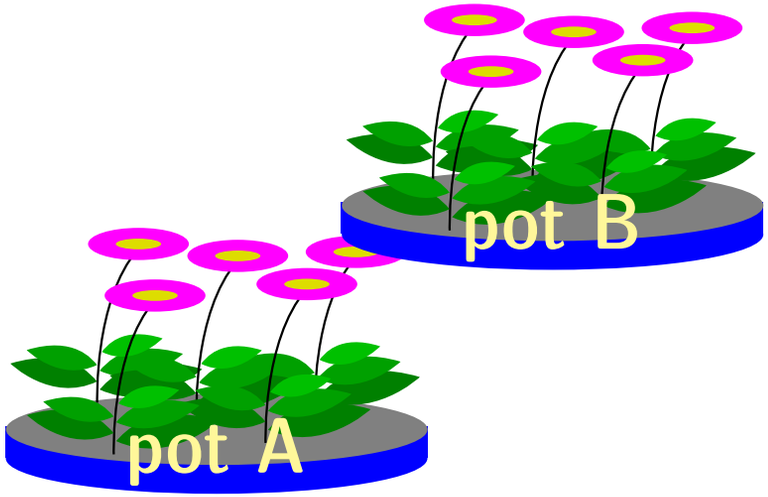
```
> summary(glm(y ~ f, data = d, family = poisson))
...(略)...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8931     0.0549   34.49  < 2e-16
fT           -0.4115     0.0869   -4.73  2.2e-06
...(略)...
```

- 肥料をやる処理 (f) をすると，平均種子数が下がる?
- AIC でモデル選択しても同じような結果に

GLMM: ブロック差は無視

```
> library(glmmML)
> summary(glmmML(y ~ f, data = d, family = poisson,
+ cluster = id))
...(略)...
                coef se(coef)     z Pr(>|z|)
(Intercept)  1.351    0.192  7.05  1.8e-12
fT          -0.737    0.280 -2.63  8.4e-03
...(略)...
```

- やっぱり同じ?
- むしろ肥料処理の悪影響が強い?



個体差 + ブロック差を考える階層ベイズモデル

- ここでは log リンク関数を使う
- 平均の対数 $\log(\lambda_i) = a + bf_i +$ (個体差) + (ブロック差)
- 事前分布の設定
  - 切片 a と f_i の係数 b は無情報事前分布 (すごく平らな正規分布)
  - 個体差とブロック差は階層的な事前分布 (それぞれ標準偏差 σ_1, σ_2 の正規分布，平均はゼロ)
  - 標準偏差 σ_* は無情報事前分布 ($[0, 10^4]$ の一様分布)

個体差 + ブロック差のあるポアソン回帰の BUGS code (1)

```
model
{
  for (i in 1:N.sample) {
    Y[i] ~ dpois(lambda[i])
    log(lambda[i]) <- a + b * F[i] + r[i] + rp[Pot[i]]
  }
  # 次のページの事前分布の定義につづく
```

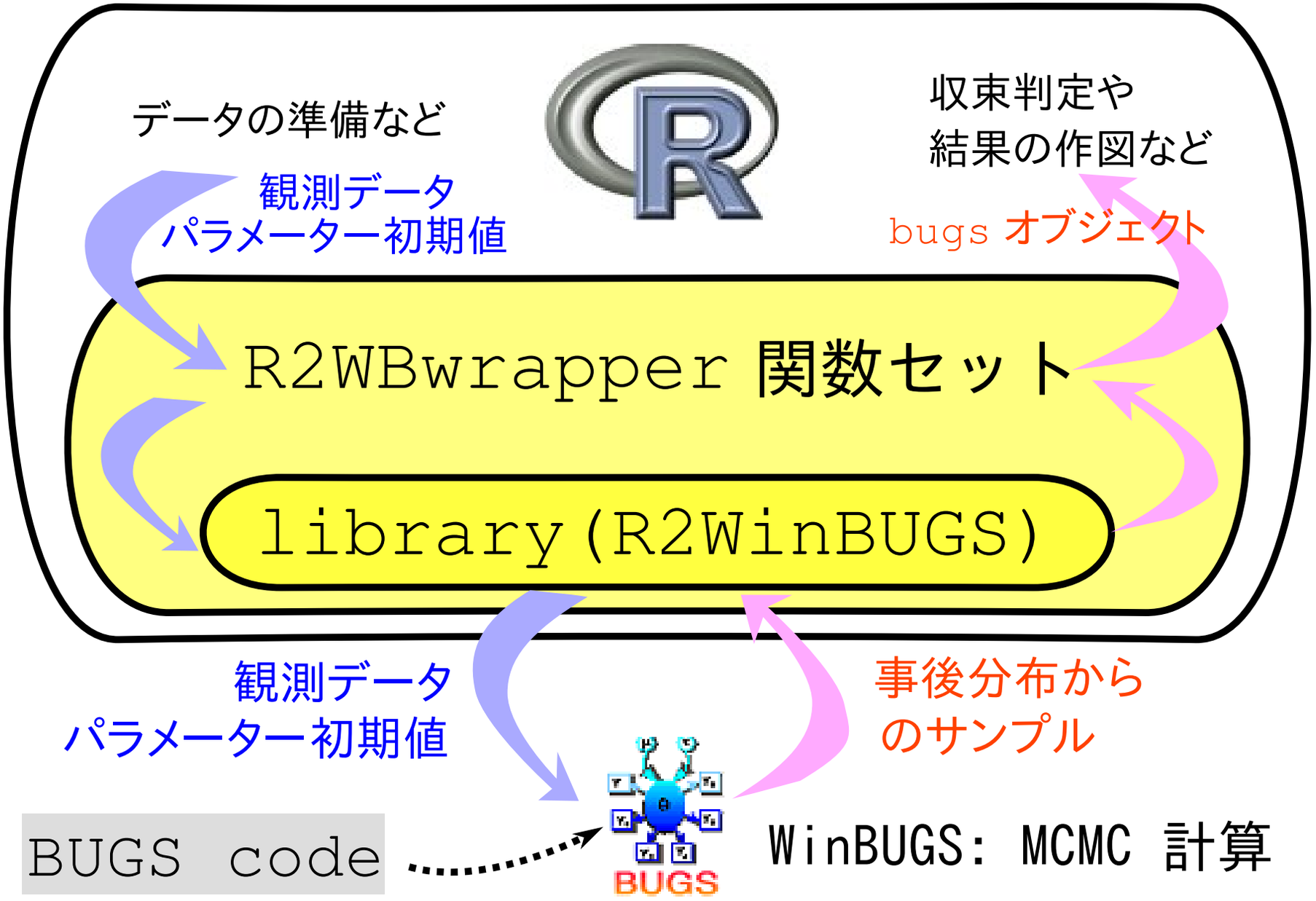
ここでの BUGS coding のポイント

- 因子型の説明変数 $f_i \in \{\mathtt{C}, \mathtt{T}\}$ は，それぞれ `F[i]` を `0`，`1` と置きかえる
- `Pot[i]` は `1`，`2`，`...`，`10` と数字になおした植木鉢名をいれておいて，植木鉢の効果 `rp[...]` を参照させる

個体差 + ブロック差のあるポアソン回帰の BUGS code (2)

```
  # 前のペイジからのつづき
  a ~ dnorm(0, 1.0E-4) # 切片
  b ~ dnorm(0, 1.0E-4) # 肥料の効果
  for (i in 1:N.sample) {
    r[i] ~ dnorm(0, tau[1])  # 個体差
  }
  for (j in 1:N.pot) {
    rp[j] ~ dnorm(0, tau[2]) # 植木鉢の差（ブロック差）
  }
  for (k in 1:N.tau) {
    tau[k] <- 1.0 / (sigma[k] * sigma[k]) # 個体・植木鉢のばらつき
    sigma[k] ~ dunif(0, 1.0E+4)
  }
}
```

R2WBwrapper 経由で WinBUGS を使う

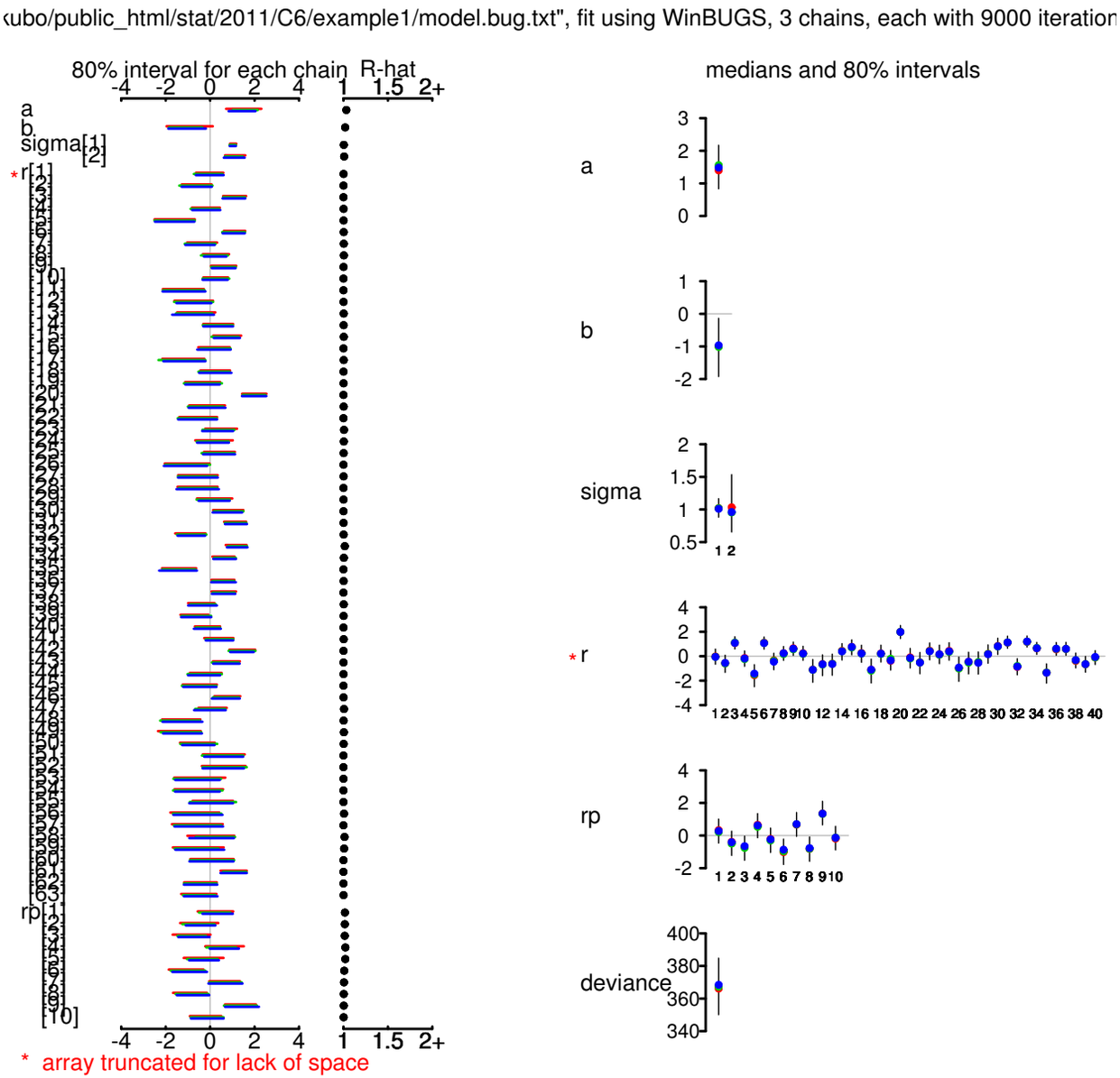


MCMC サンプリングの設定

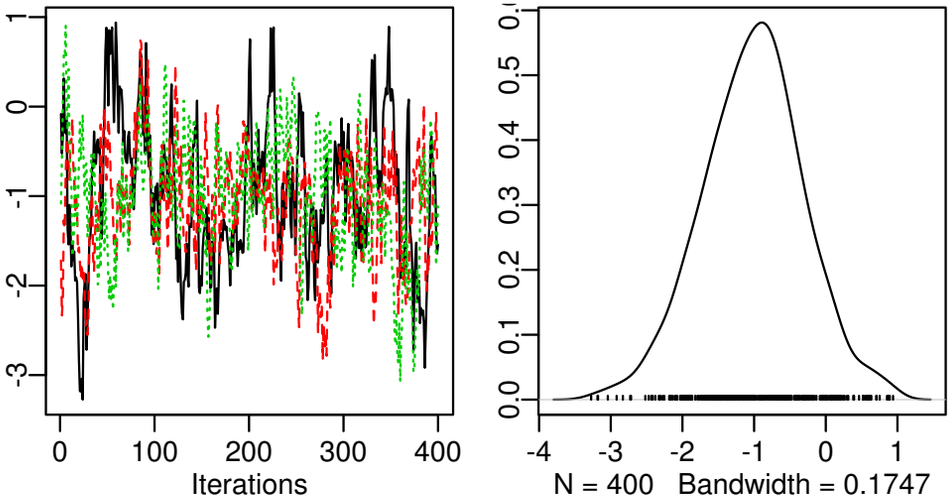
```
post.bugs <- call.bugs(
  file = "model.bug.txt",
  n.iter = 9000, n.burnin = 1000, n.thin = 20
)
```

- (default で) 3 chain
- `n.iter = 9000`: ひとつの chain は 9000 MCMC step
- `n.burnin = 1000`: 最初の 1000 MCMC step は捨てる
- `n.thin = 20`: 20 MCMC step とばしで
- サンプル数は 400×3

WinBUGS による事後分布の推定，R で収束判定



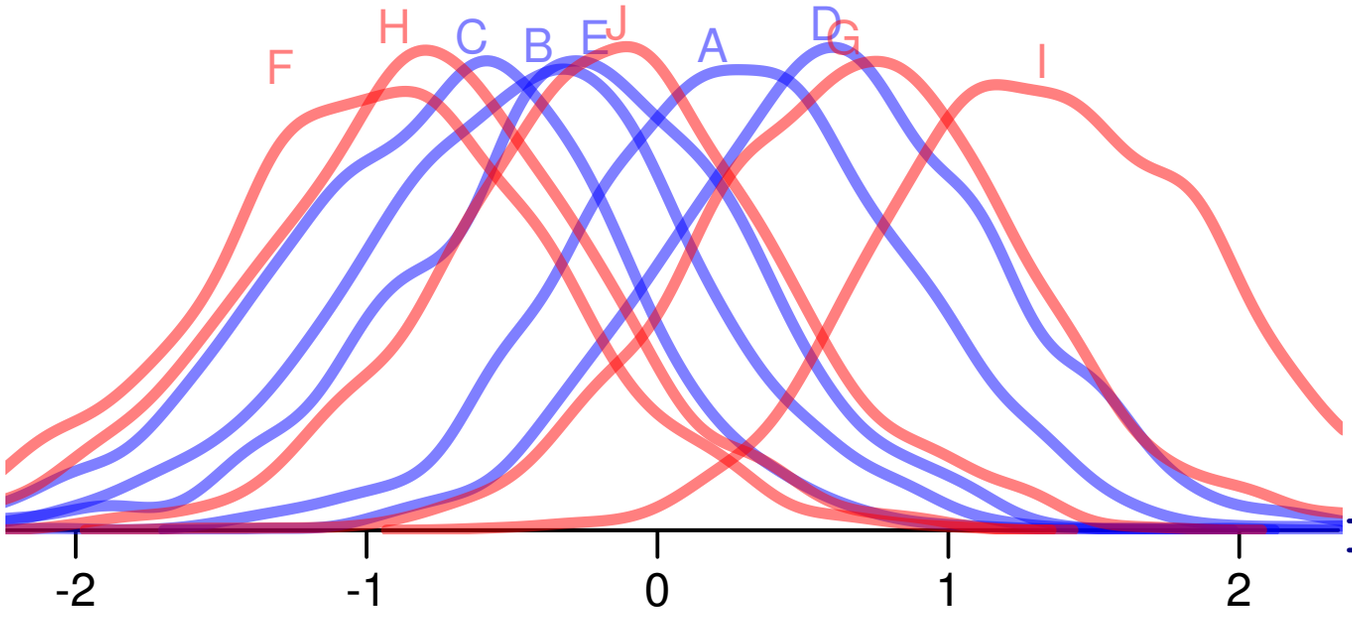
肥料の効果 (パラメーター b) はなさそう?



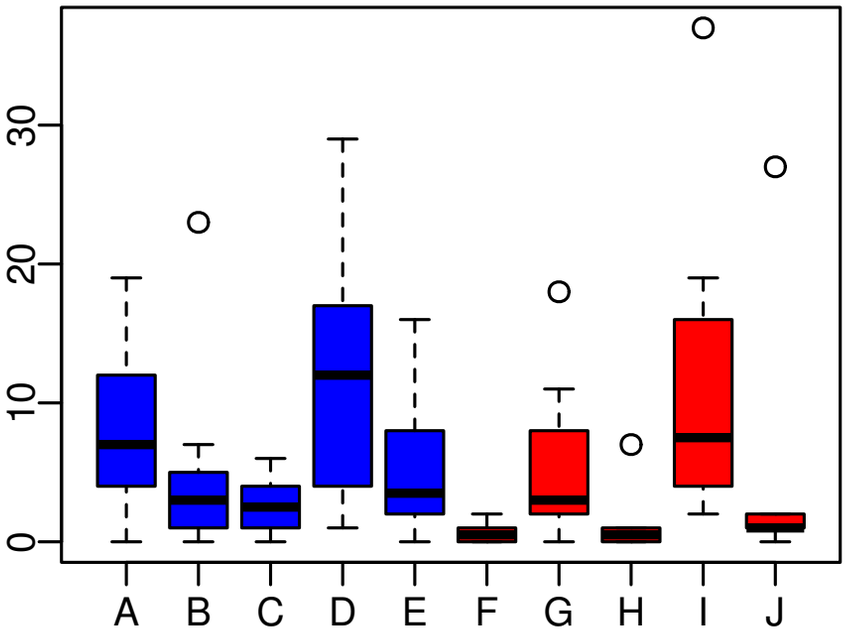

```
> print(post.bugs, digits.summary = 3)
...(略)...
             mean     sd    2.5%     25%     50%     75%   97.5%  Rhat n.eff
a           1.501  0.529   0.482   1.157   1.493   1.852   2.565 1.032   240
b          -1.016  0.706  -2.436  -1.476  -0.993  -0.565   0.395 1.019   450
sigma[1]    1.020  0.114   0.822   0.939   1.014   1.089   1.265 1.004   510
...(略)...
```

この架空データを生成した種子数シミュレイションでは，肥料の効果はまったく無いと設定していた

推定された植木鉢の差 (ブロック差)



ブロック差
rp[j]



こういう結果の解釈の注意

- 個体差・ブロック差が大きい
- もしブロック差を人為的に小さくできないなら，ブロック数をもっと増やして，より正確な「植木鉢の効果のばらつき」を正確に推定するしかない
- 肥料の効果 (パラメーター b) については,「効果があるともないとも言えない」ぐらいが妥当だろう
- **random effects** の影響が大きいときには，**fixed effects** の大きさが見えにくくなる— ニセの「効果」が見えることもあれば，見えるはずの傾向が隠されることも

WinBUGS で空間相関モデリング

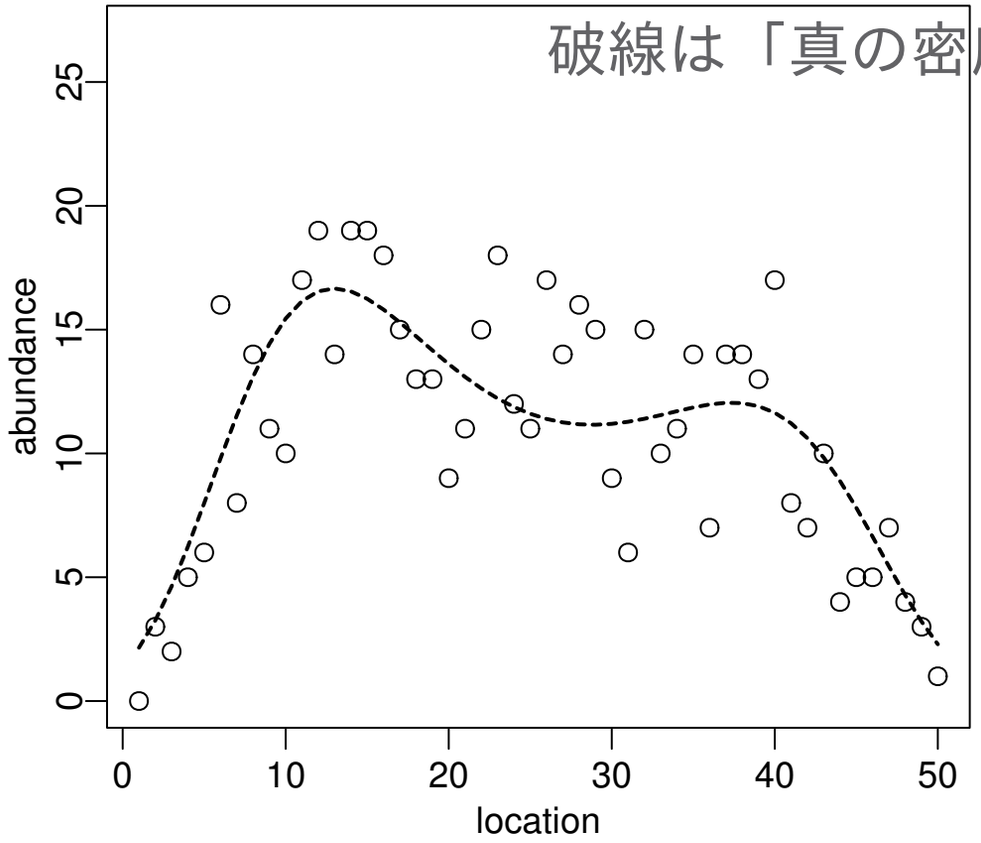
— 一次元の架空データ —

架空の例題: 個体数データ，一次元空間データ

欠測データなし

環境は均質 (に見える)

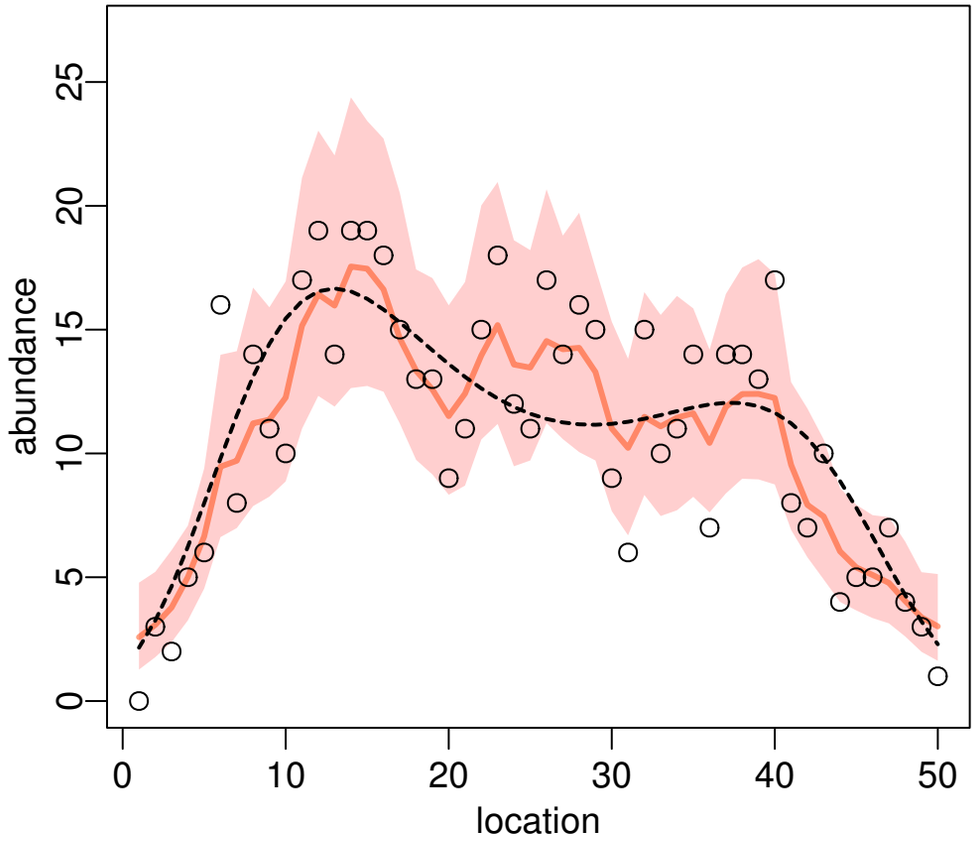
破線は「真の密度」



問: 空間自己相関を考慮して生物個体の密度推定

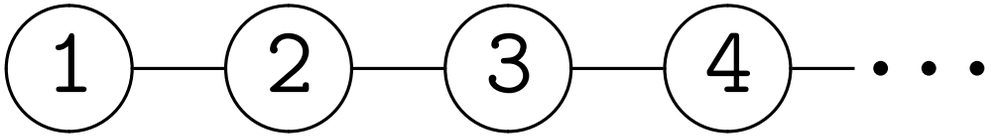
解析の目的: まずはこんな推定をしてみたい

空間相関を考慮するモデル
欠測データなし



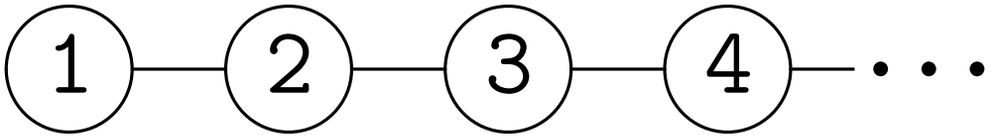
(彩色された領域は平均値の事後分布の 95% 区間，曲線は中央値)

空間相関のある「場所差」階層ベイズモデル



- 地点 i の観測個体数は平均 λ_i のポアソン分布にしたがう:
  $y_i \sim \text{Poisson}(\lambda_i)$
- 平均 λ_i の対数は (全体の平均) + (場所差) と分割する:
  $\log \lambda_i = \beta + r_i$
- ベイズモデルとしてあつかいたいので，推定したいパラメーターの事前分布 を決めてやらなければならない
  - 事前分布 についてはあとで説明
- 全体の平均 β は無情報事前分布にしたがう:
  $\beta \sim \text{Normal}(0, 10^2)$,

空間相関のある「場所差」階層ベイズモデル (続)



- Conditional Autoregressive (CAR) モデルにおける場所差 r_i の条件つき事前分布(N_i は i の近傍場所数, $\mathbf{J}_i$ は i の近傍場所):

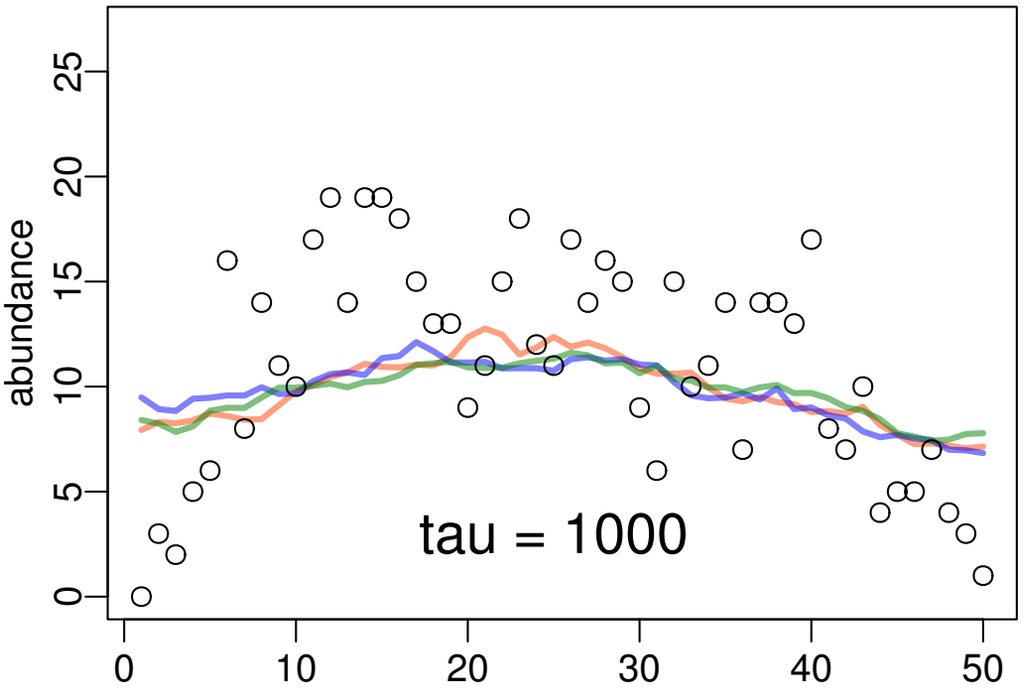
$$r_i \sim \text{Normal}\left(\frac{\sum_{j \in \mathbf{J}_i} r_j}{N_i}, \frac{\sigma}{N_i}\right)$$

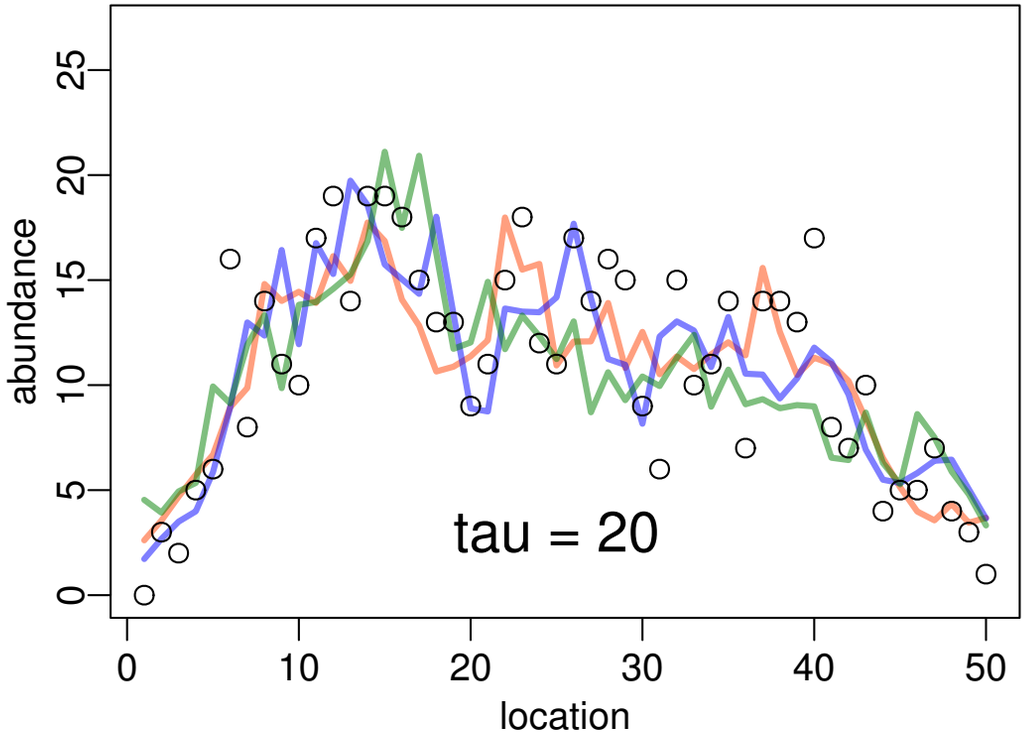
- σ は無情報事前分布にしたがう:
  $\tau = 1/\sigma^2 \sim \text{Gamma}(1.0^{-2}, 1.0^{-2})$
- ベイズの定理 $\rightarrow$ 事後分布の導出

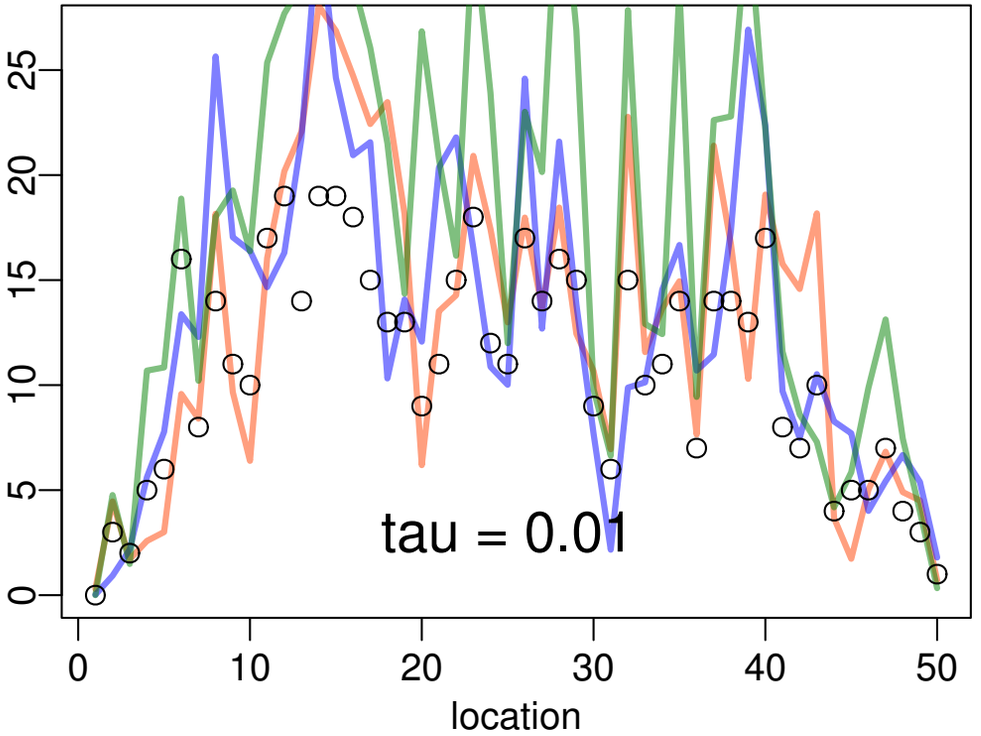
$$p(\beta, \{r_i\}, \tau \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \beta, \{r_i\}, \tau) \times (\text{事前分布あれこれ})}{\int\int \cdots \int (\uparrow \quad \text{分子}) d\beta dr_1 \cdots dr_{50} d\tau}$$

超パラメーター τ が決める 隣との類似度

- τ が大 (σ が小) だと隣と似ている
- τ が小 (σ が大) だと隣と似てない
- ベイズ推定によって適切な τ の範囲 (事後分布) が得られる

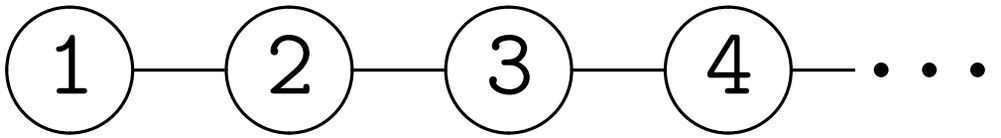






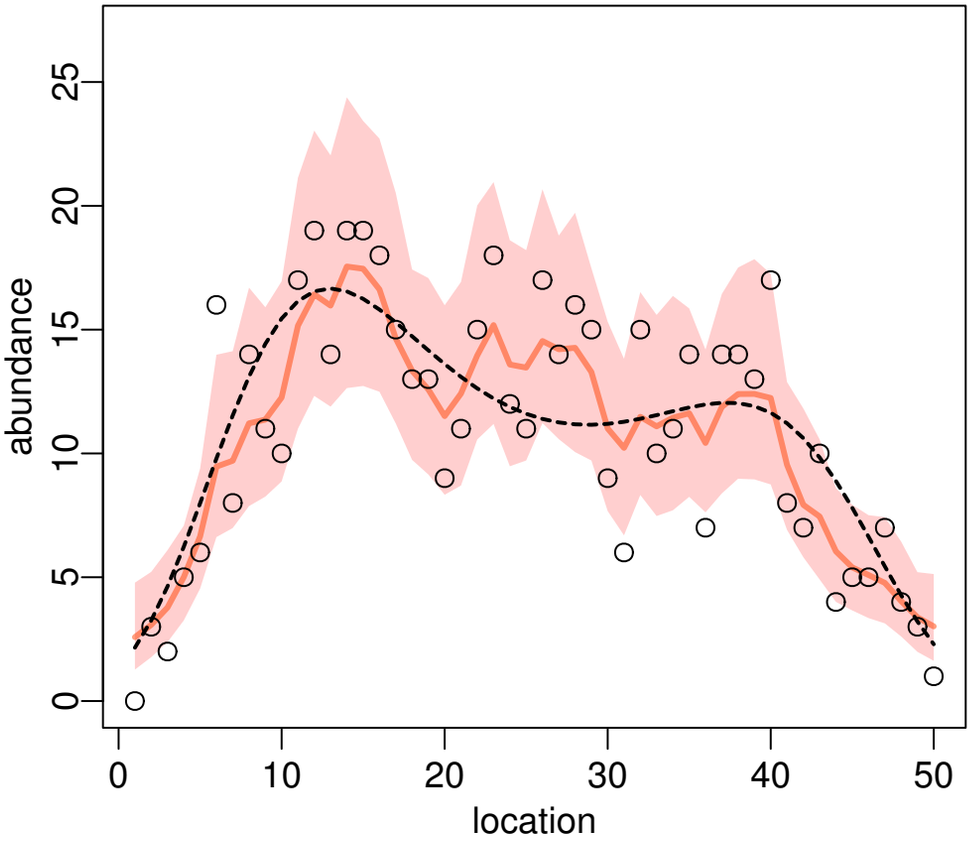
BUGS 言語: ベイズモデルを記述する言語

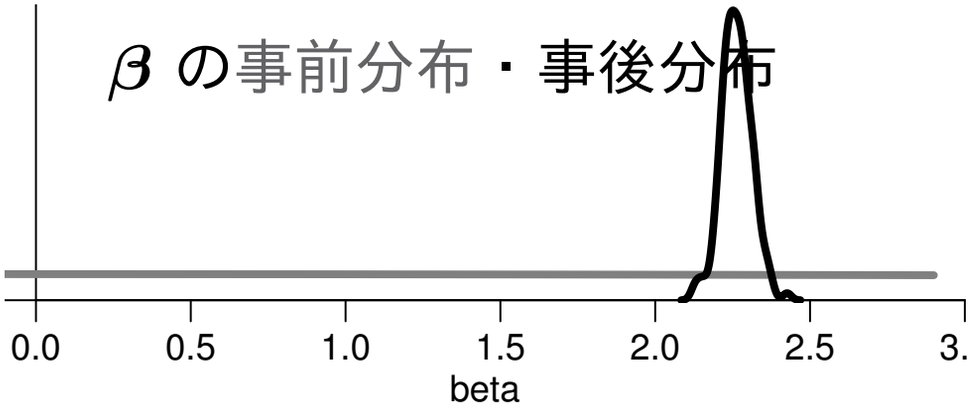
```
model { # BUGS コードで定義された階層ベイズモデルの例
  for (i in 1:N.site) {
    Y[i] ~ dpois(mean[i])           # 観測データと密度の関係
    log(mean[i]) <- beta + re[i]  # (全体の平均) + (場所差)
  }
  # 場所差 re[i] を CAR model で生成
  re[1:N.site] ~ car.normal(Adj[], Weights[], Num[], tau)
  beta ~ dnorm(0, 1.0E-2)       # 全体の平均は無情報事前分布
  tau ~ dgamma(1.0E-2, 1.0E-2) # 場所差のばらつきは無情報事前分布
}
```

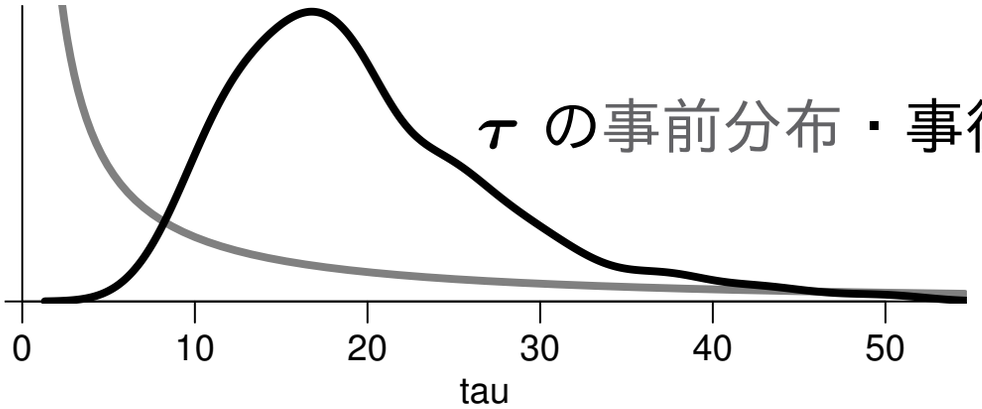


空間相関のある「場所差」モデルの推定結果

空間相関を考慮するモデル
欠測データなし



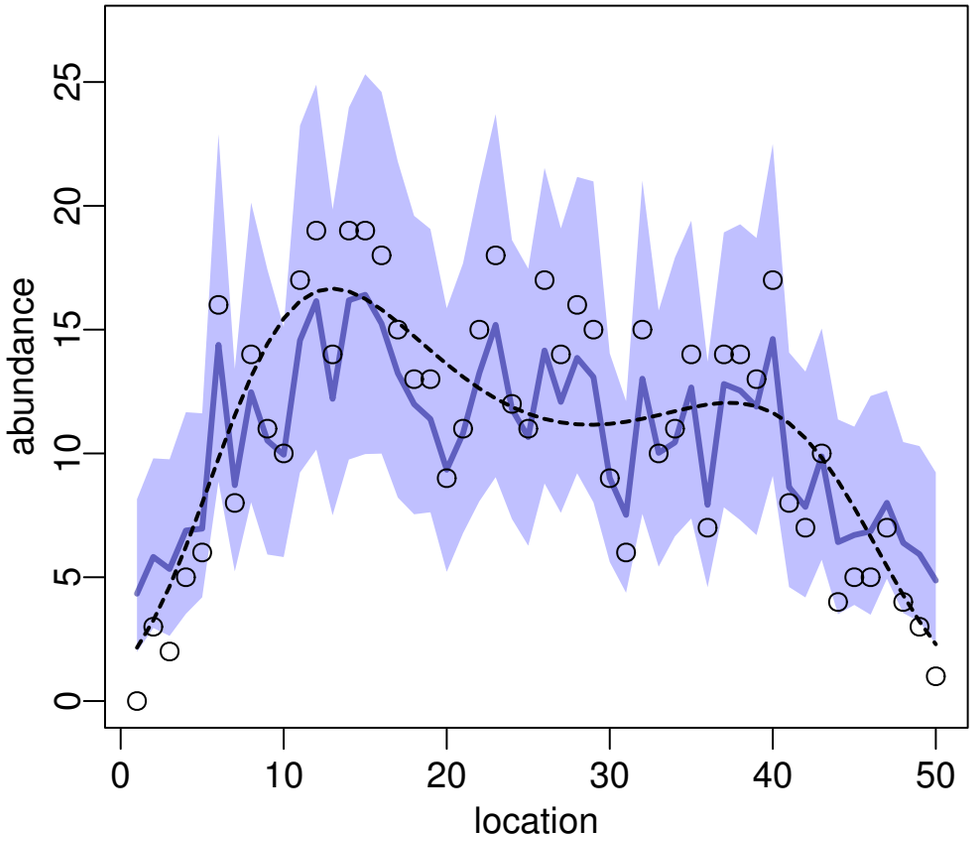




(彩色された領域は平均値の事後分布の 95% 区間，曲線は中央値)

空間相関を考慮しないベイズモデルの推定結果

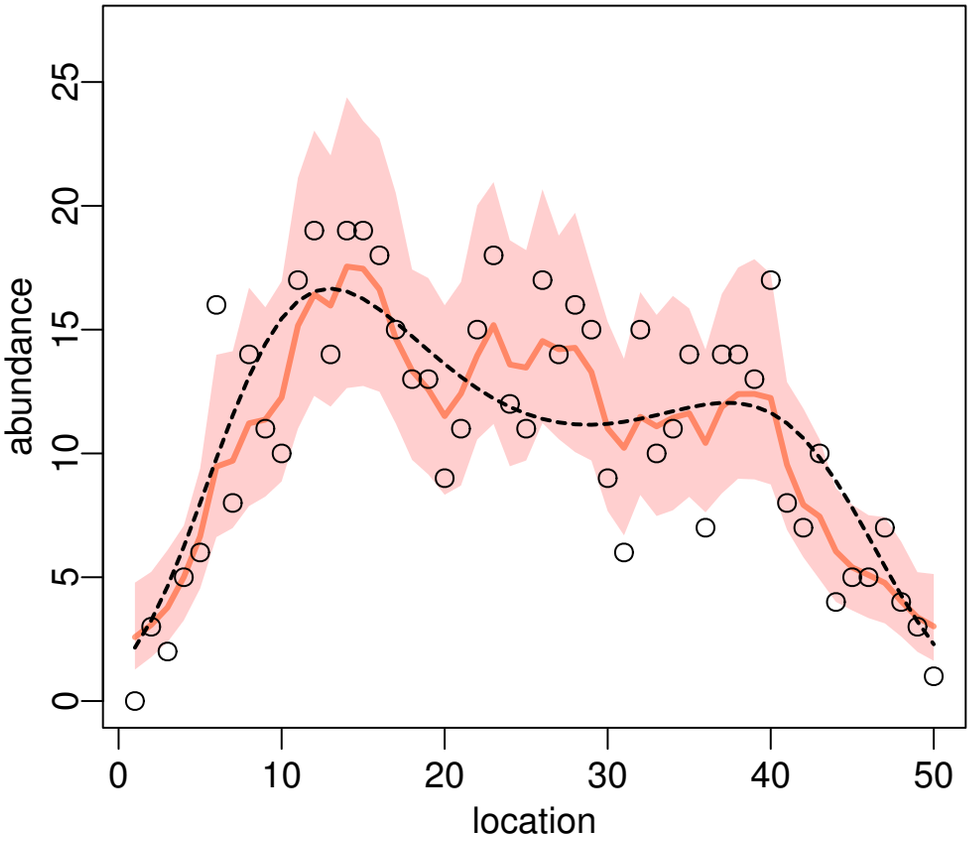
空間相関を考慮しないモデル
欠測データなし



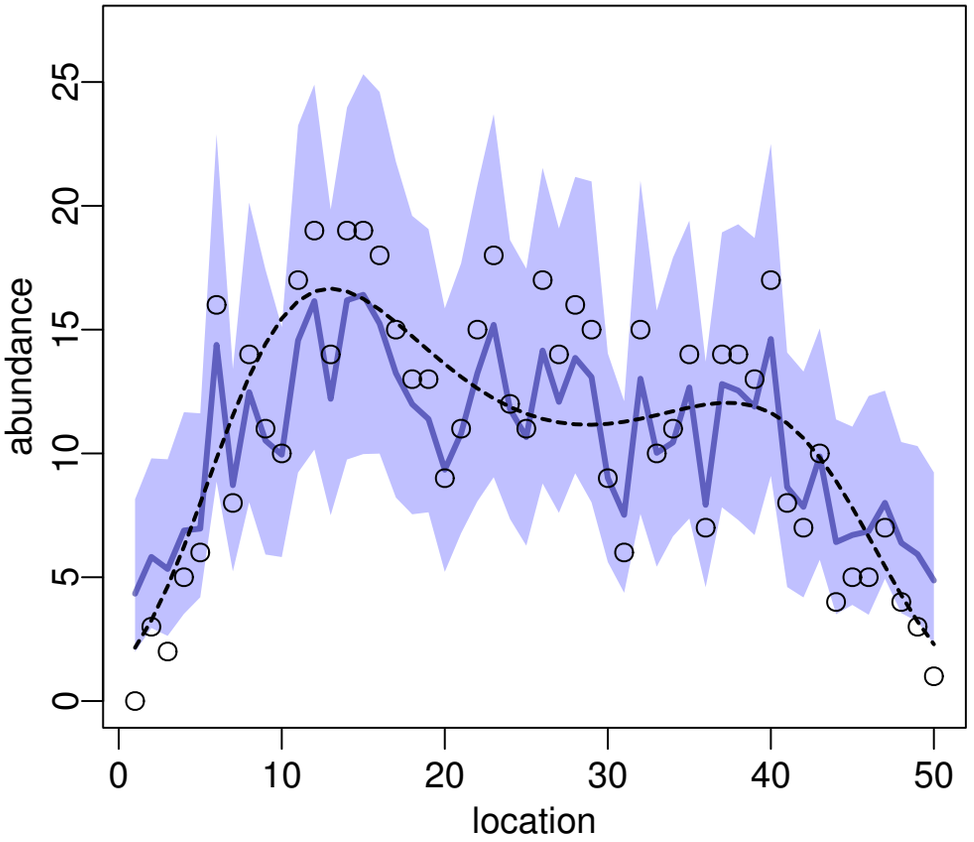
空間相関とか考えない GLMM 的なモデルでも OK?

空間相関を考慮する vs しないモデル

空間相関を考慮するモデル
欠測データなし



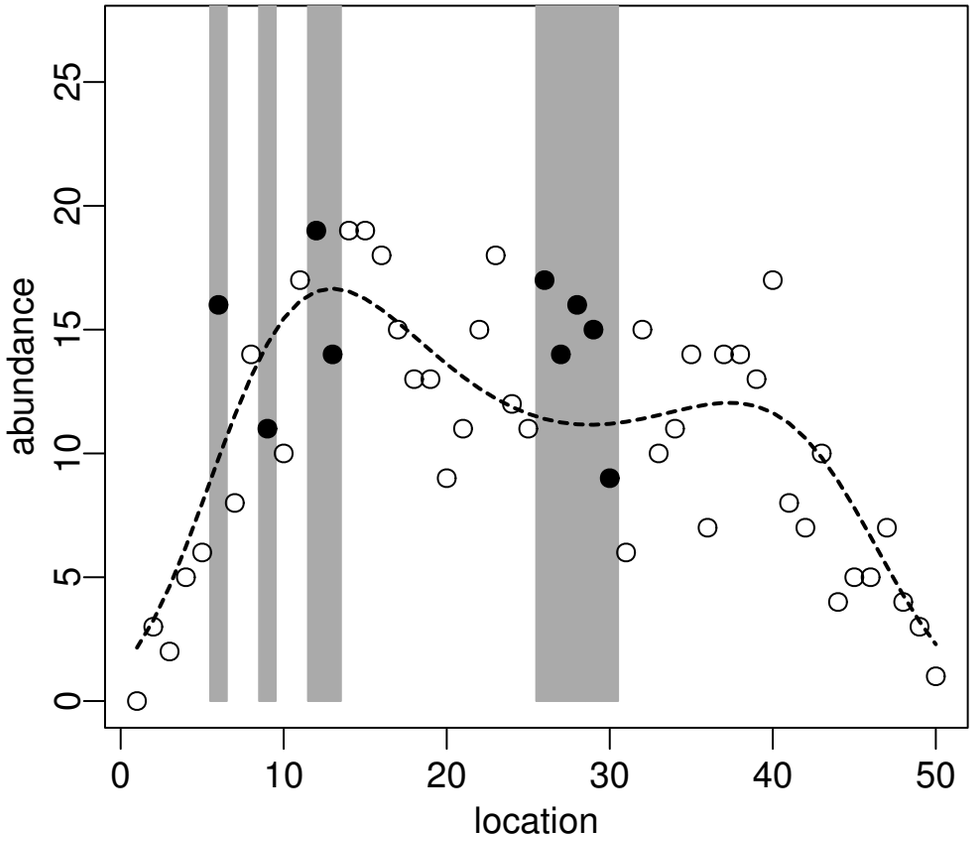
空間相関を考慮しないモデル
欠測データなし



空間相関を考慮する必要があるのだろうか?

架空の例題 (続): 欠測がある場合は?!

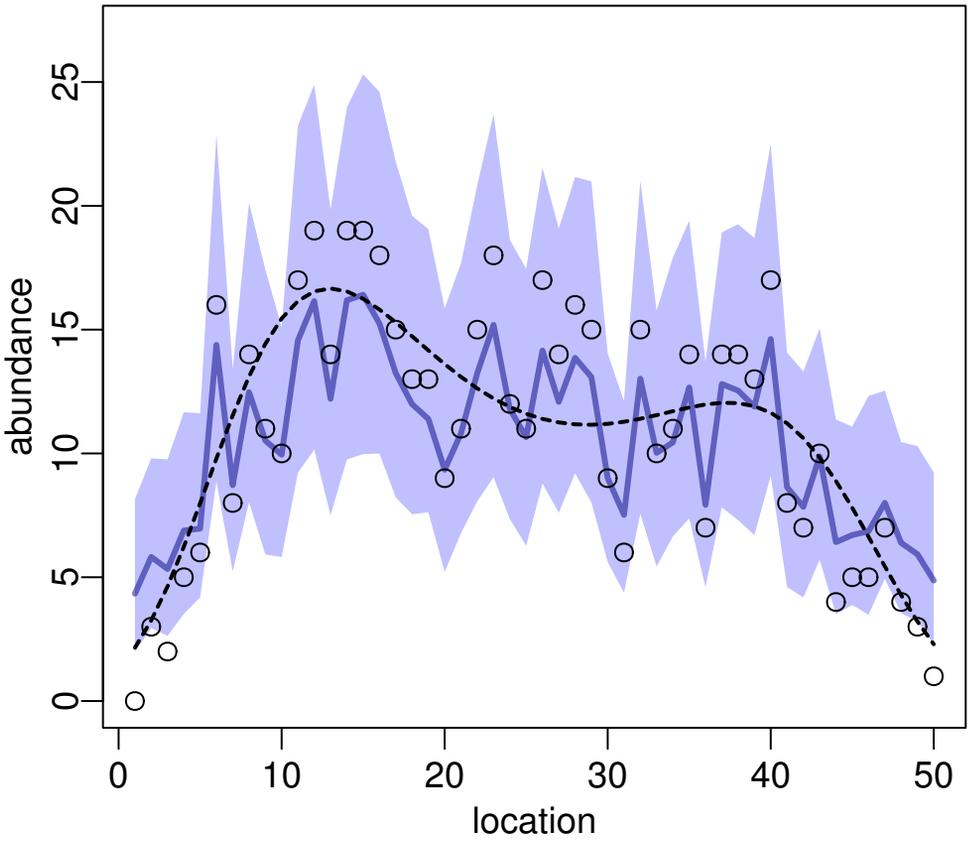
欠測あり　　欠測値の予測!



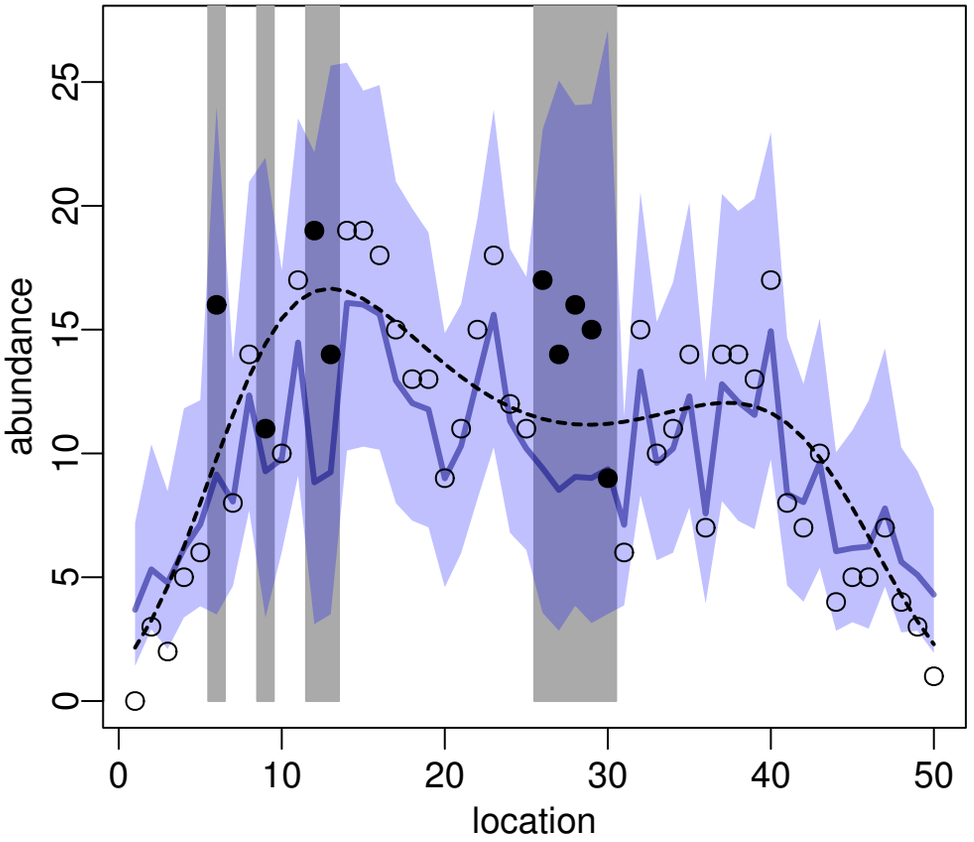
灰色の領域で観測できなかった (●は観測できなかった点)

空間相関を考慮しないベイズモデルは欠測にヨワい

空間相関を考慮しないモデル
欠測データなし



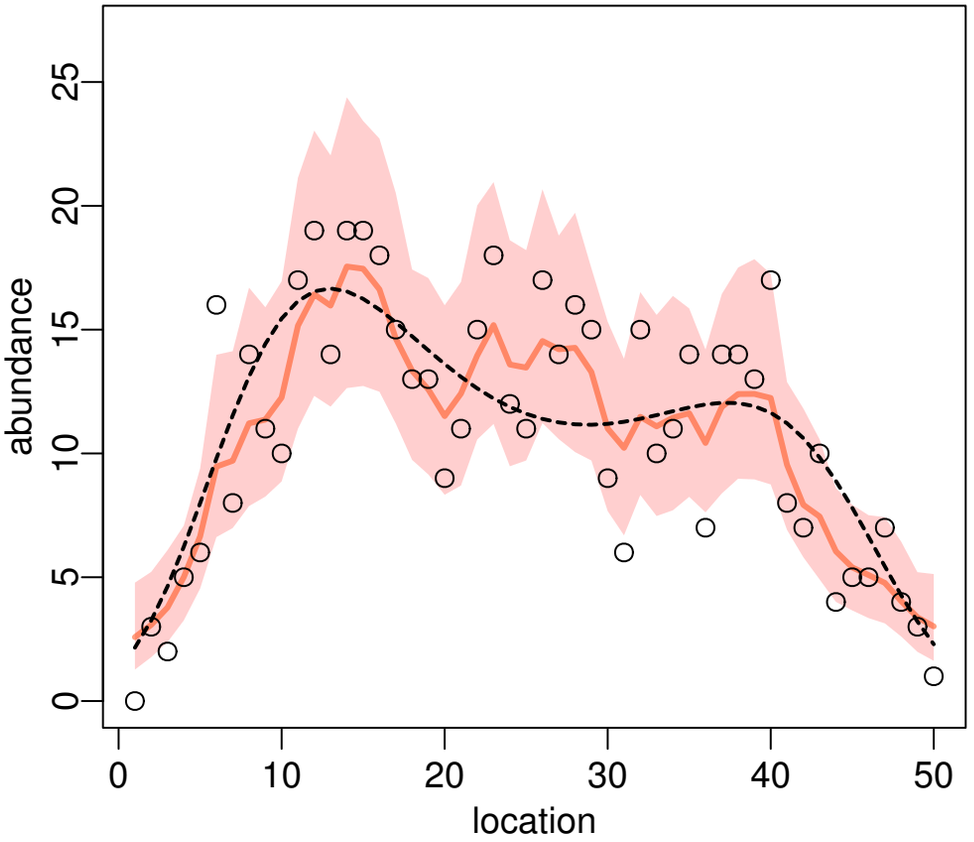
空間相関を考慮しないモデル
欠測あり



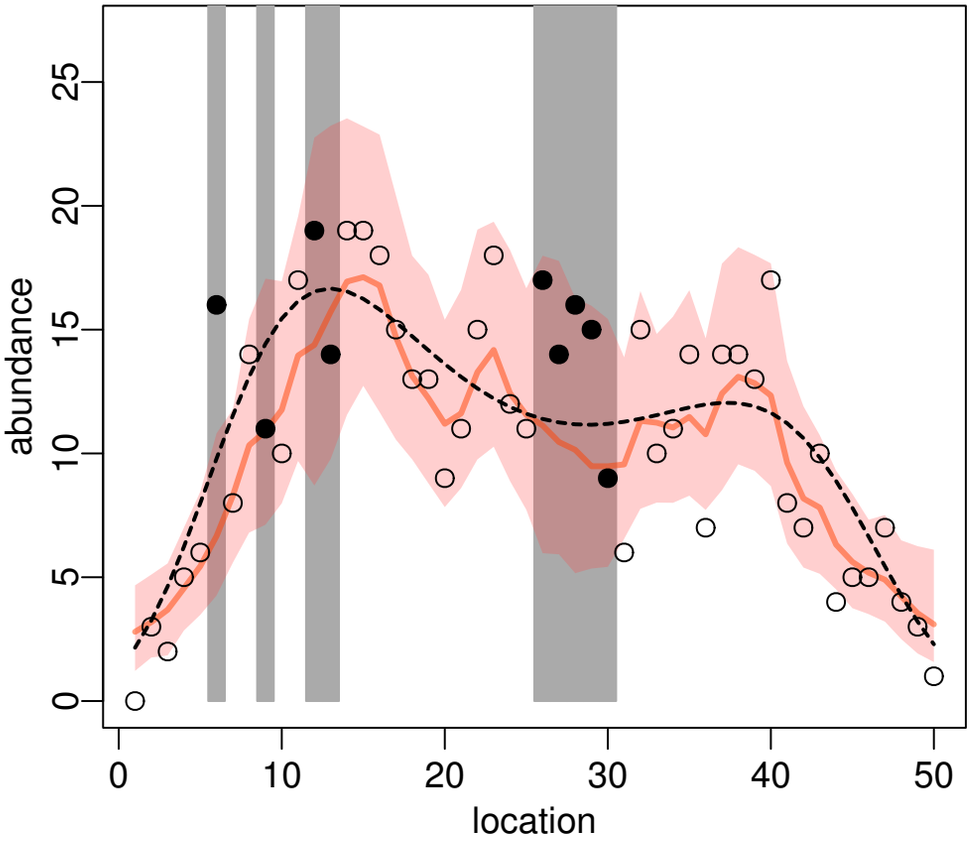
欠測領域で事後分布がひろがる!

空間相関を考慮するモデルは欠測に頑健

空間相関を考慮するモデル
欠測データなし



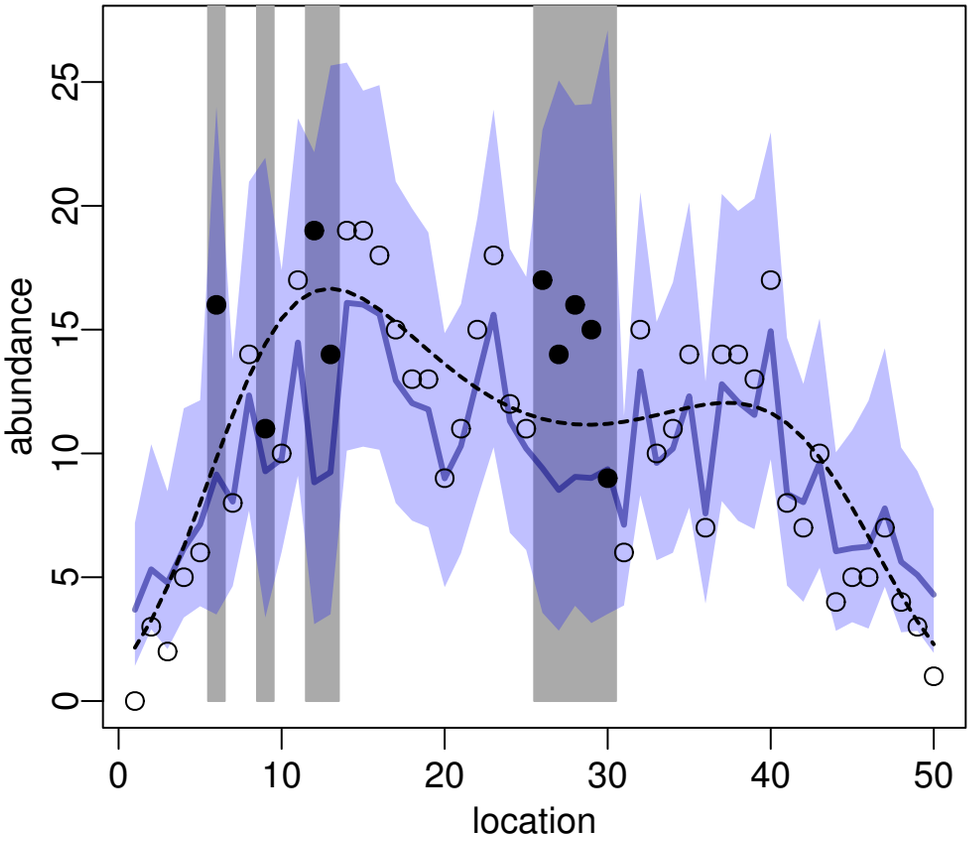
空間相関を考慮するモデル
欠測あり



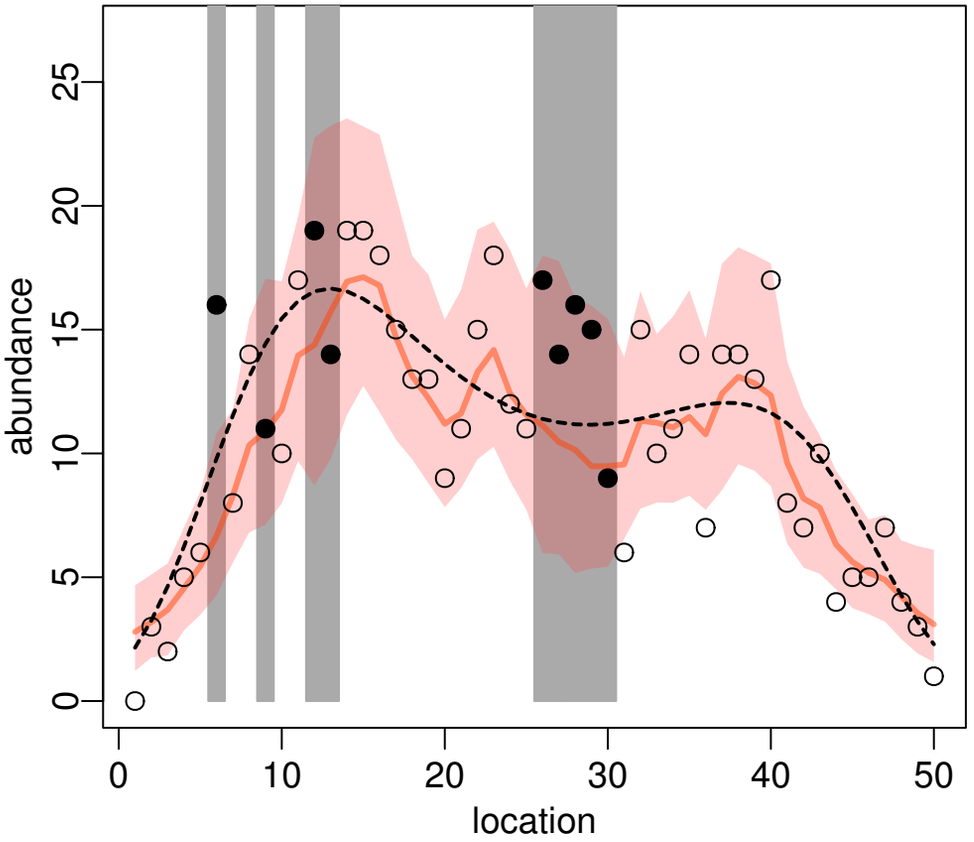
CAR 階層ベイズモデルで「隣は似てるよ」効果を表現

ベイズモデルの御利益: 空間的・時間的な欠測にも対処可能

空間相関を考慮しないモデル
欠測あり



空間相関を考慮するモデル
欠測あり



この単純な例題を拡張して環境要因などをくみこめる

まとめ: 空間構造のあるランダム効果

- ガウス確率場 (**Gaussian random field**) で「隣と似ている」ランダム効果を表現する
- 各地点独立と仮定するランダム効果でも，それっぽい推定はできないこともない
- しかし欠測のあるデータセットの解析においては，空間相関を考慮したベイズ統計モデルが威力を発揮するだろう
- ガウス確率場のモデリングはさらにいろいろと工夫できる— よりなめらかに変化させるような方法もある

ウミガメ上陸数の統計モデリング

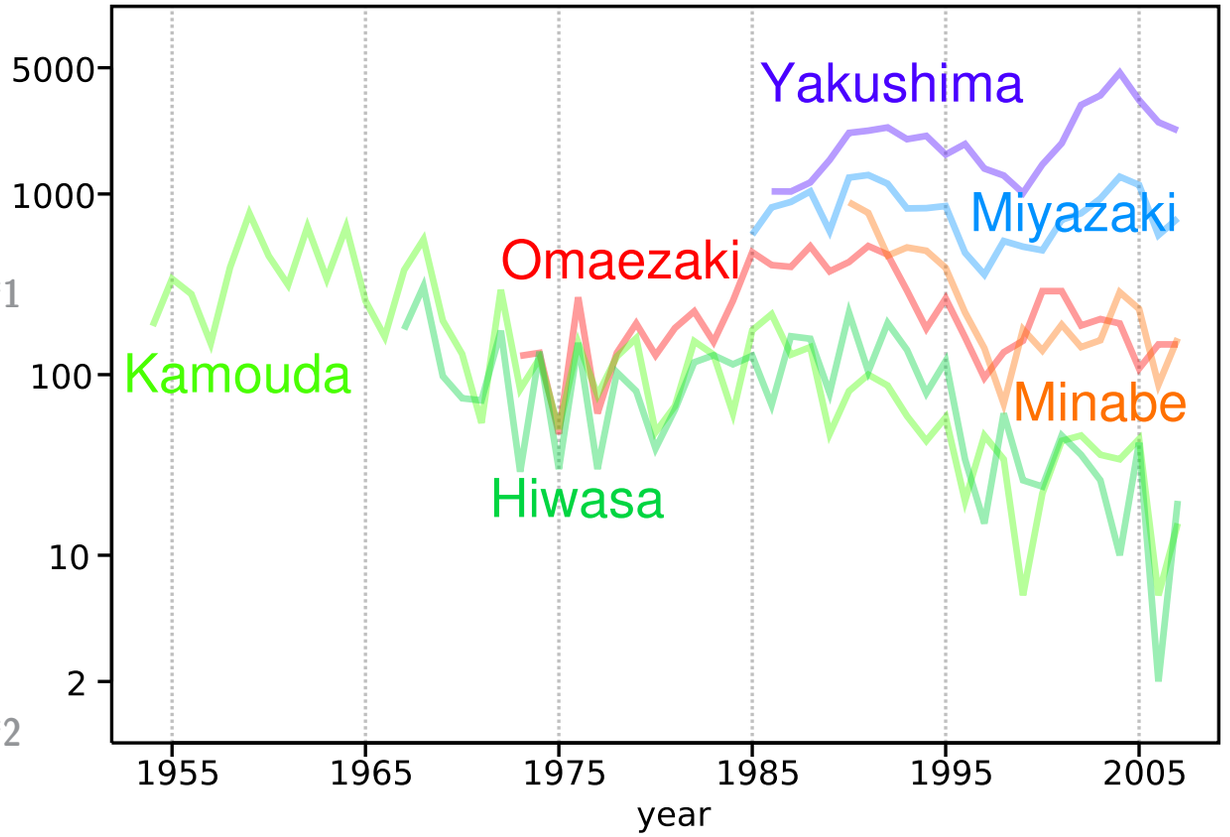
階層ベイズモデルを応用した時系列データ解析

(重田麻衣氏・亀崎直樹氏との共同研究)

ウミガメ上陸数の観測データ (1954–2007)

*1

*2



Omaezaki (静岡), Minabe (和歌山), Kamouda (徳島)

Hiwasa (徳島), Miyazaki (宮崎), Yakushima (鹿児島)

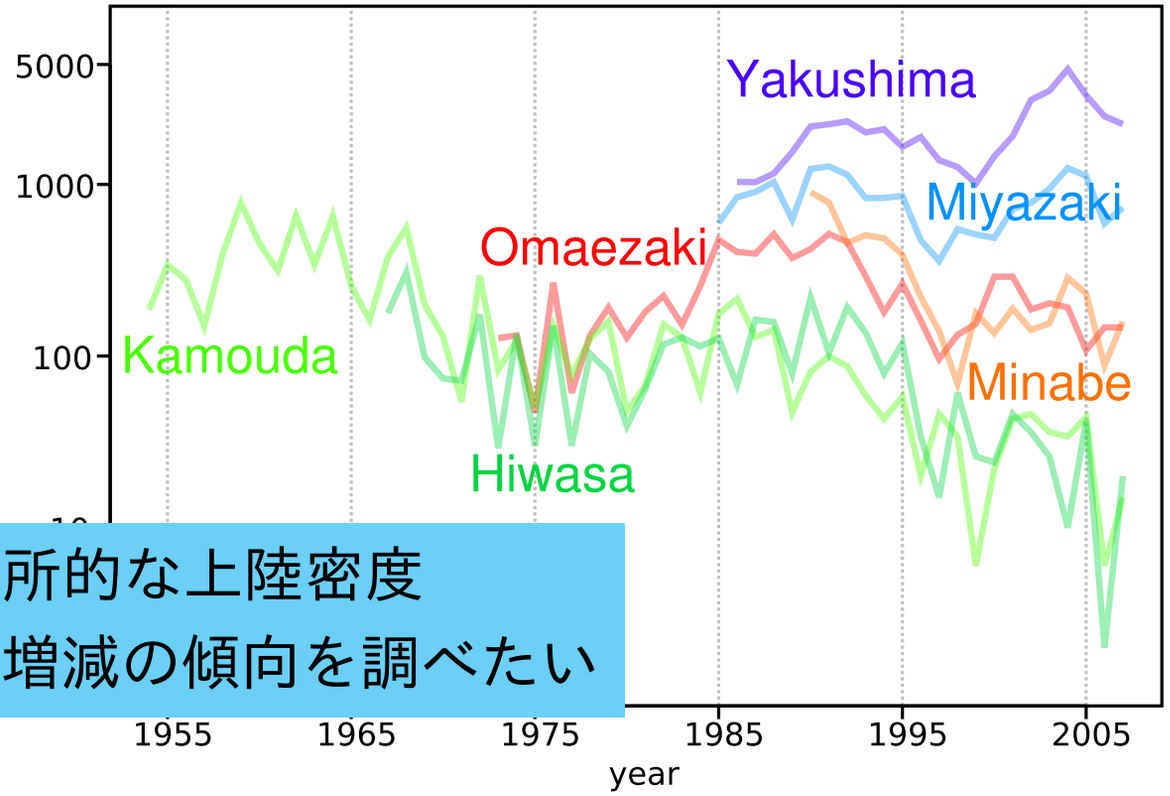
*1 © Mike Gonzalez, October 14, 2007. Wikimedia Commons. *2 © OpenCage, February 2, 2006. Wikimedia Commons.

ウミガメ上陸数の観測データ，その内容

- 各地のウミガメ上陸海岸ごとに，ウミガメ上陸観察団体 (アマチュア研究者が主体) があって，長年にわたったデータをとっている
- 全国的な組織であるウミガメ協議会はそのデータの所在を把握している
- 統一的な方法で調査しているわけではないので，海岸間の比較が難しいところがある
  - たとえば上陸数の大小比較は意味がないかも

問: 産卵場所としての利用が減少している海岸は?

産卵場所として不適 → 上陸数の減少となっている?



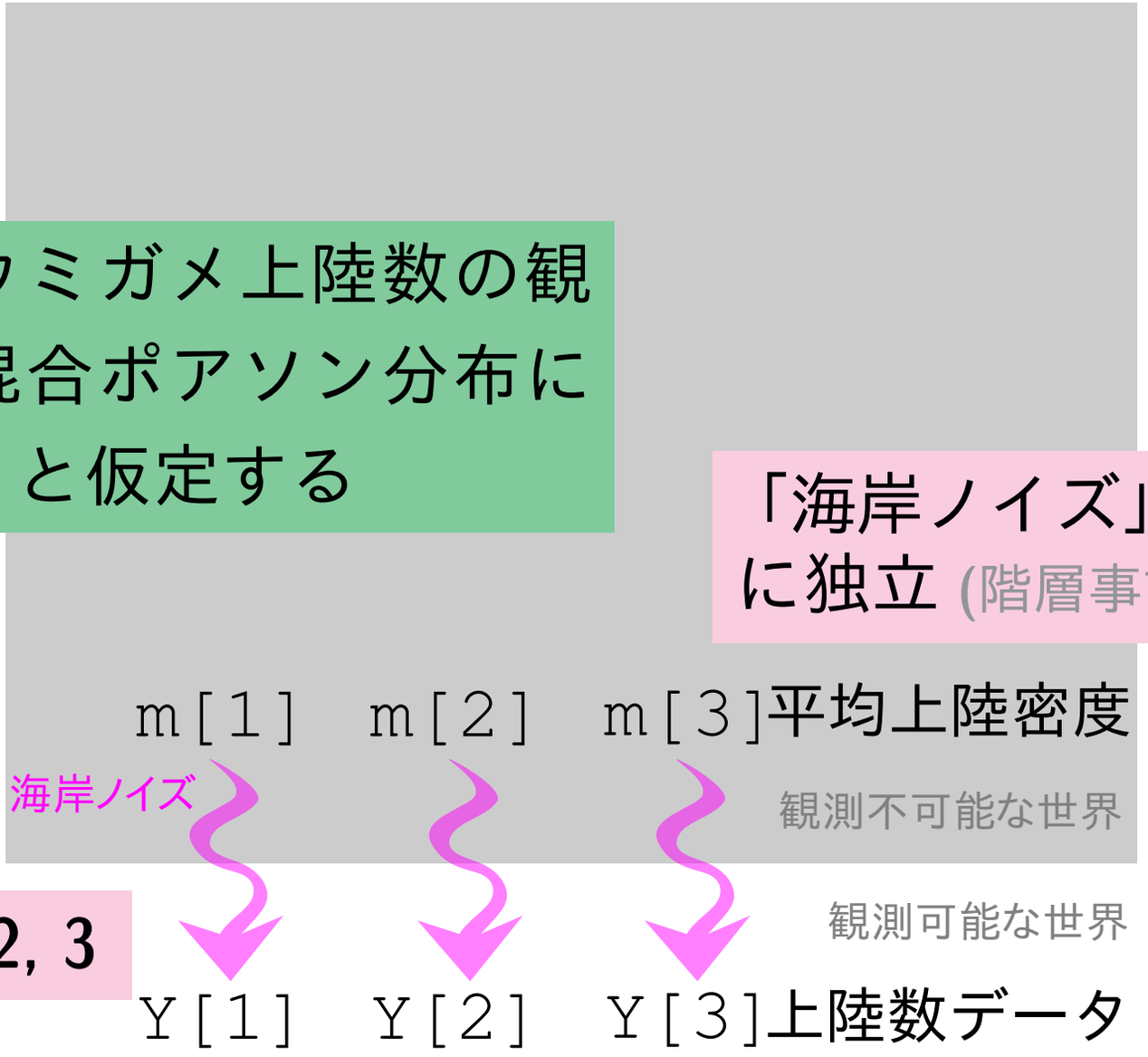
局所的な上陸密度
の増減の傾向を調べたい

- 上陸密度 ⟺ 産卵地としての海岸の良さ?
- 上陸数変動の状態空間モデル (階層ベイズモデルの一種)

ウミガメ上陸数の観測モデル

各年のウミガメ上陸数の観測値が混合ポアソン分布にしたがうと仮定する

「海岸ノイズ」は年ごとに独立 (階層事前分布)

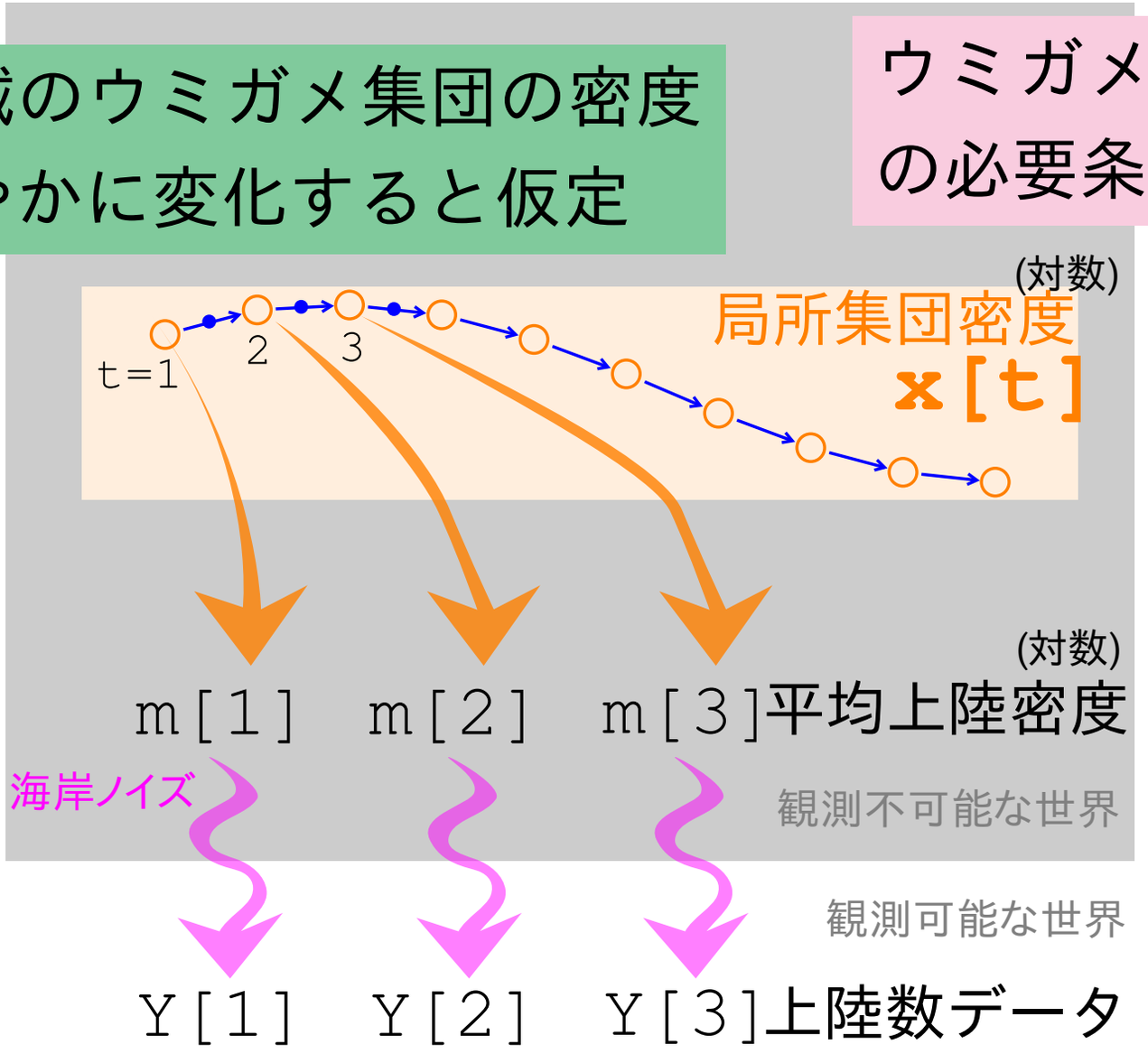


観測年 1, 2, 3

ゆっくり変化させる地域集団密度 (状態モデル)

ある海域のウミガメ集団の密度がゆるやかに変化すると仮定

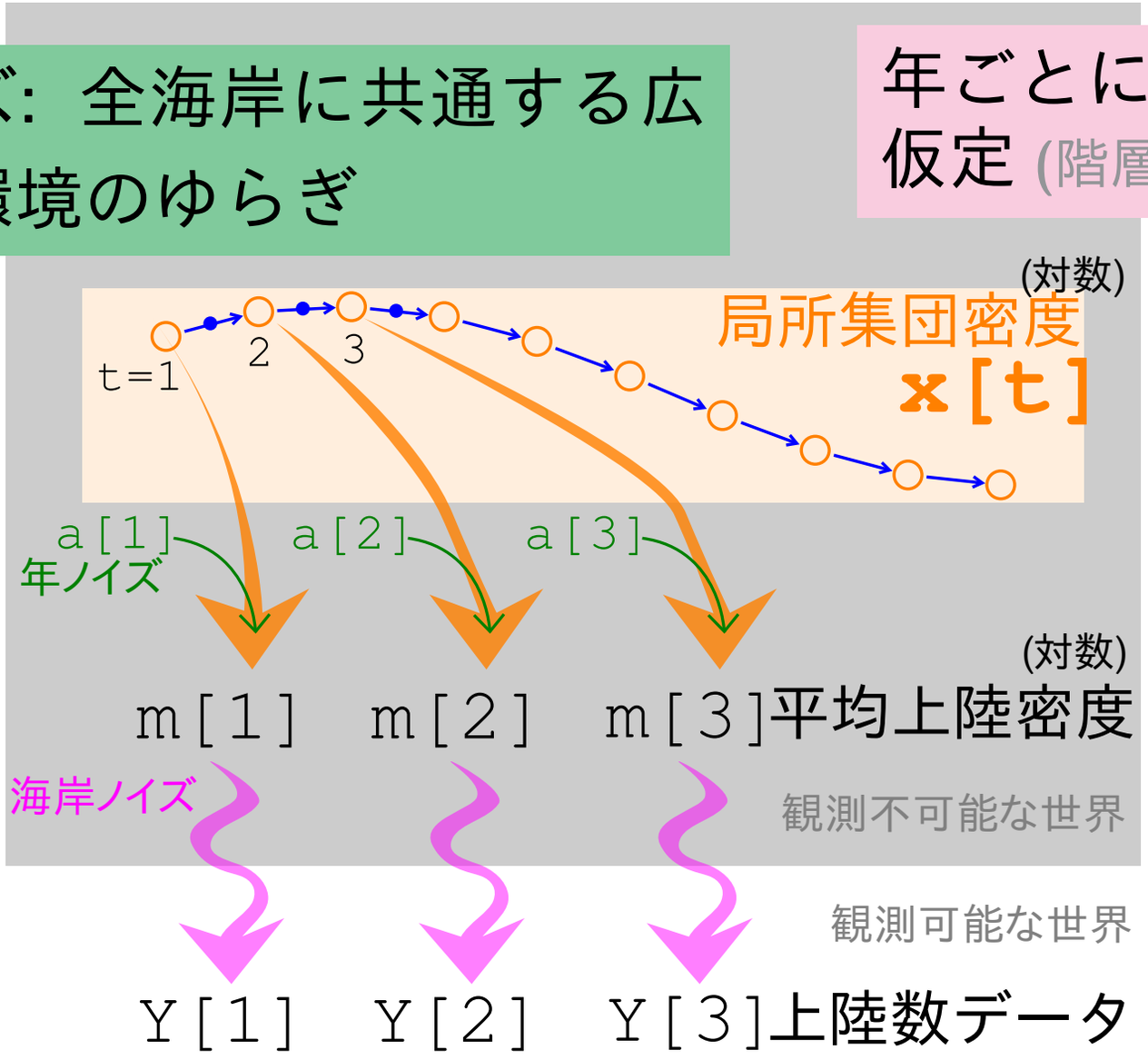
ウミガメ集団維持の必要条件?



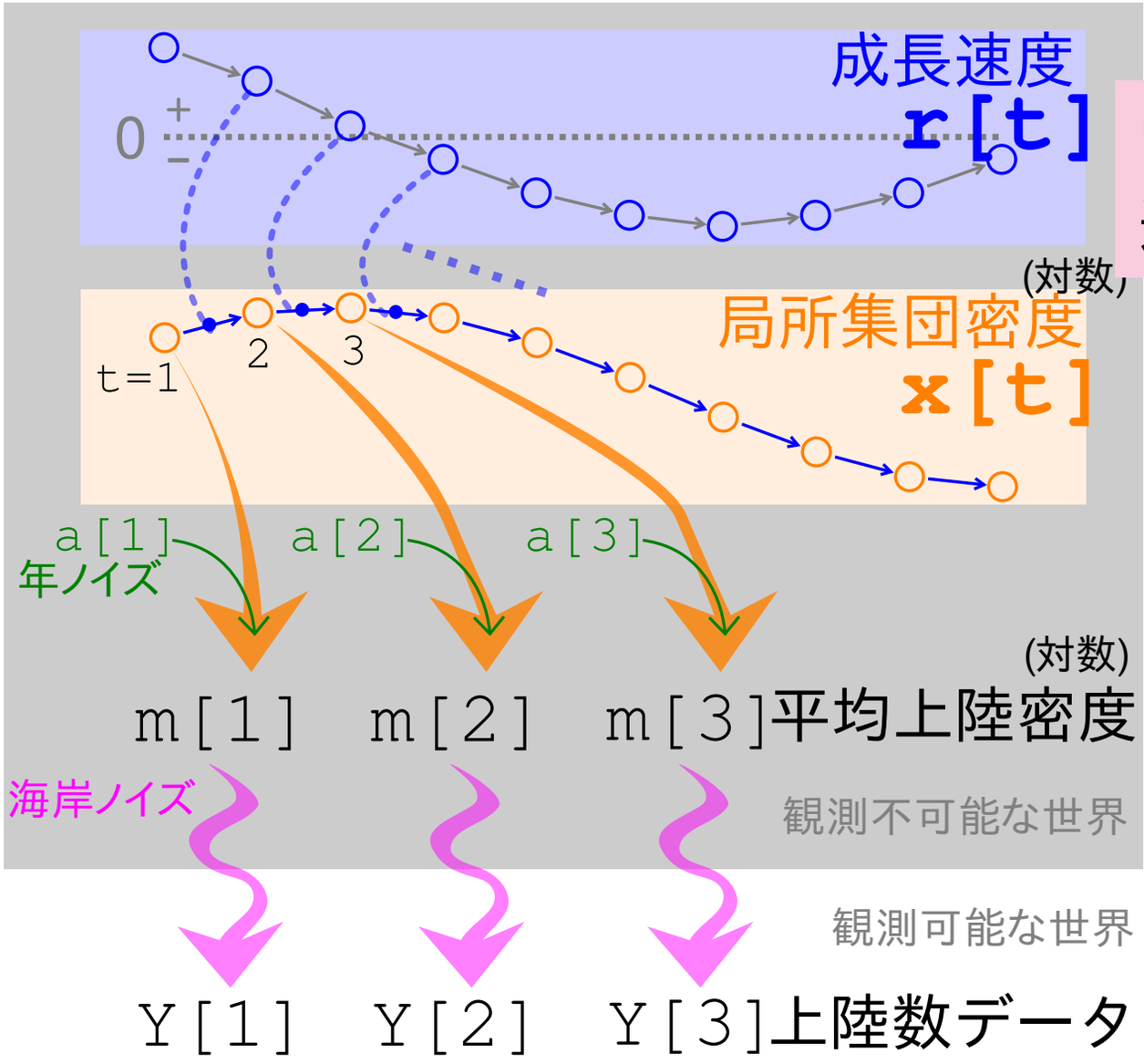
年ノイズ: 年ごとに気象条件などの影響など

年ノイズ: 全海岸に共通する広域的な環境のゆらぎ

年ごとに独立を仮定 (階層事前分布)



ウミガメ上陸数の状態空間モデル



ゆっくりした
速度変化

状態空間モデルの (擬似) BUGS code 実装

カウントデータはポアソン分布にしたがう

```
Y[t] ~ dpois(exp(log.y[t]))
```

平均対数上陸密度 m[t]に海岸ノイズを加える

```
log.y[t] ~ dnorm(m[t], tau)

m[t] <- x[t] + a[t]
```

年ノイズは階層的な事前分布にしたがう

```
a[t] ~ dnorm(0.0, tau.a)
```

局所集団密度の変化にもノイズを加える

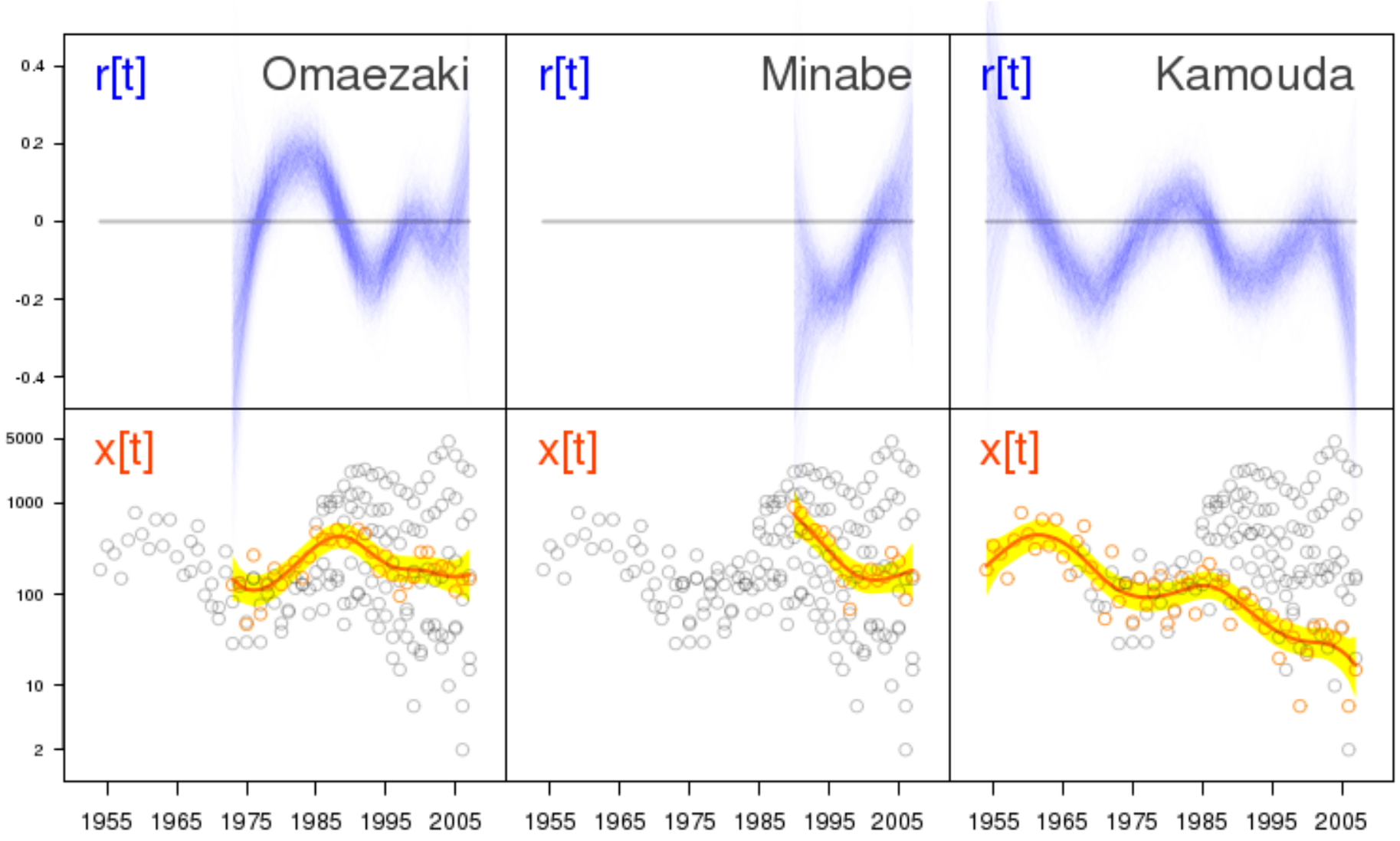
```
x[t] ~ dnorm(x[t-1] + r[t], tau.x)
```

局所集団の平均増殖速度はランダムに変化する

```
r[t] ~ dnorm(r[t-1], tau.r)
```

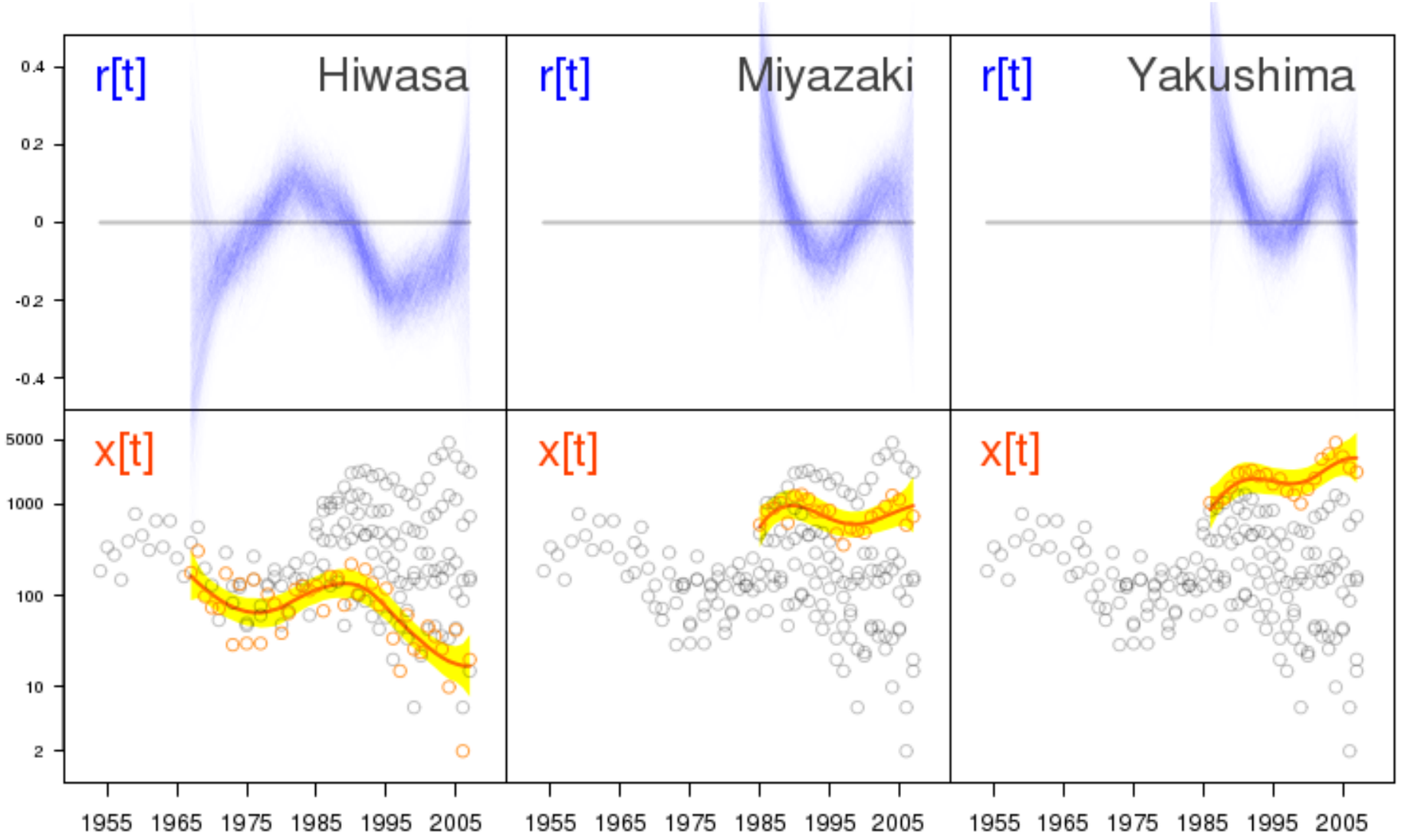
(上は概要のみ ……実際はもっと複雑怪奇なモノになります)

上陸密度増減の事後分布 (静岡, 和歌山, 徳島)



上段: 平均増殖速度; 下段: 上陸密度

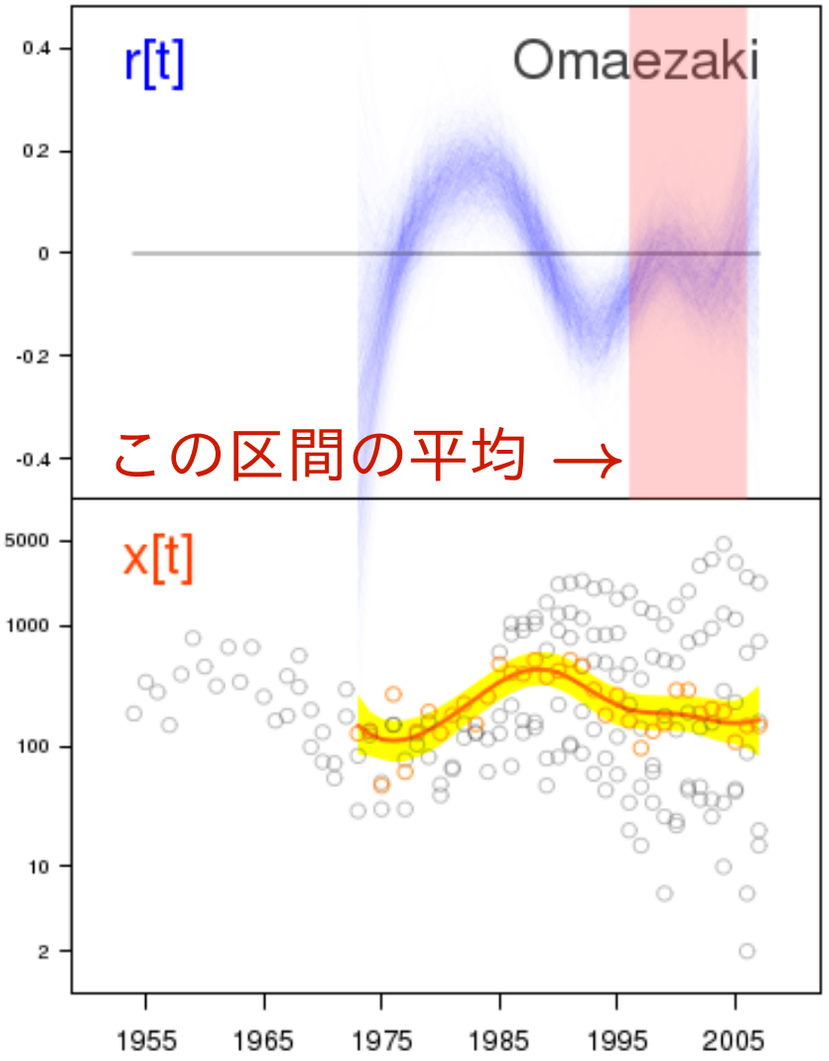
上陸密度増減の事後分布 (徳島, 宮崎, 鹿児島)



上段: 平均増殖速度; 下段: 上陸密度

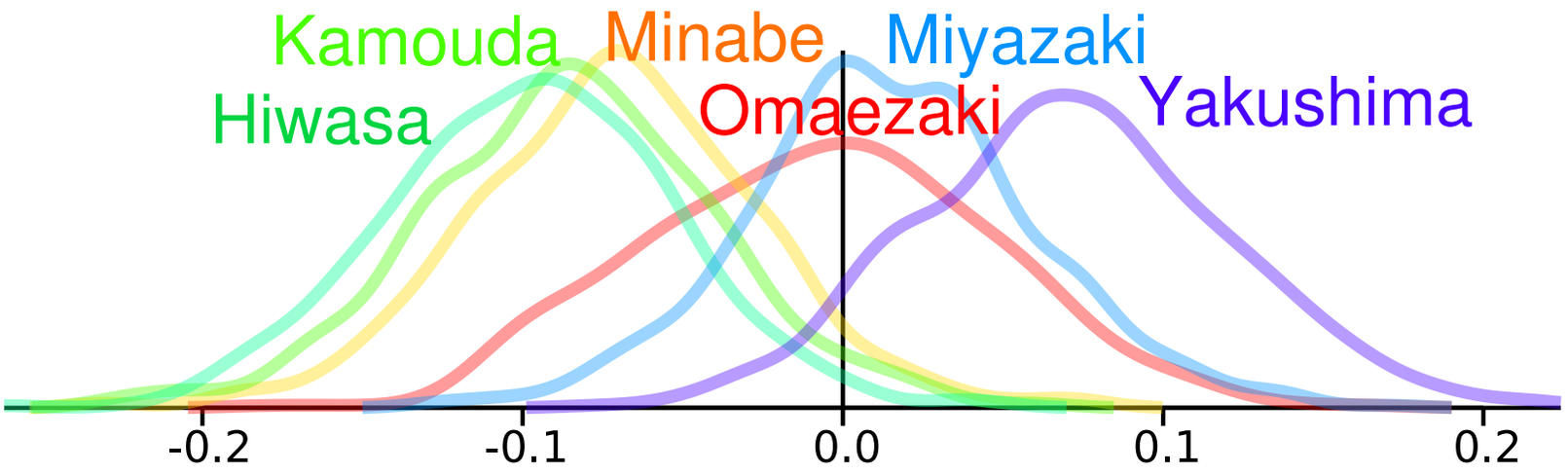
「産卵地として選ばれなくなってきた海岸」はどこ?

- どの海岸で上陸数が減少しているか?
- 平均成長速度: 海岸ごとの成長速度 r[t] の最近 11 年間の平均を指標とした
- これは個体群生態学でよく使われる，個体群成長速度の幾何平均に該当する
- r[t] の事後分布を使って，平均成長速度の予測分布を評価する



平均成長速度による産卵地としての海岸評価

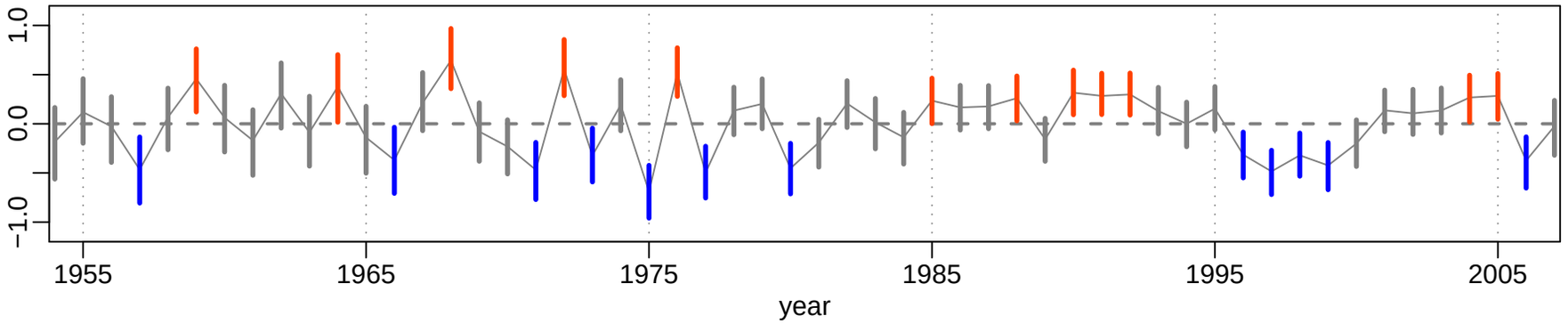
- 1996–2006 年の平均成長速度の予測分布 (海岸ごとに)

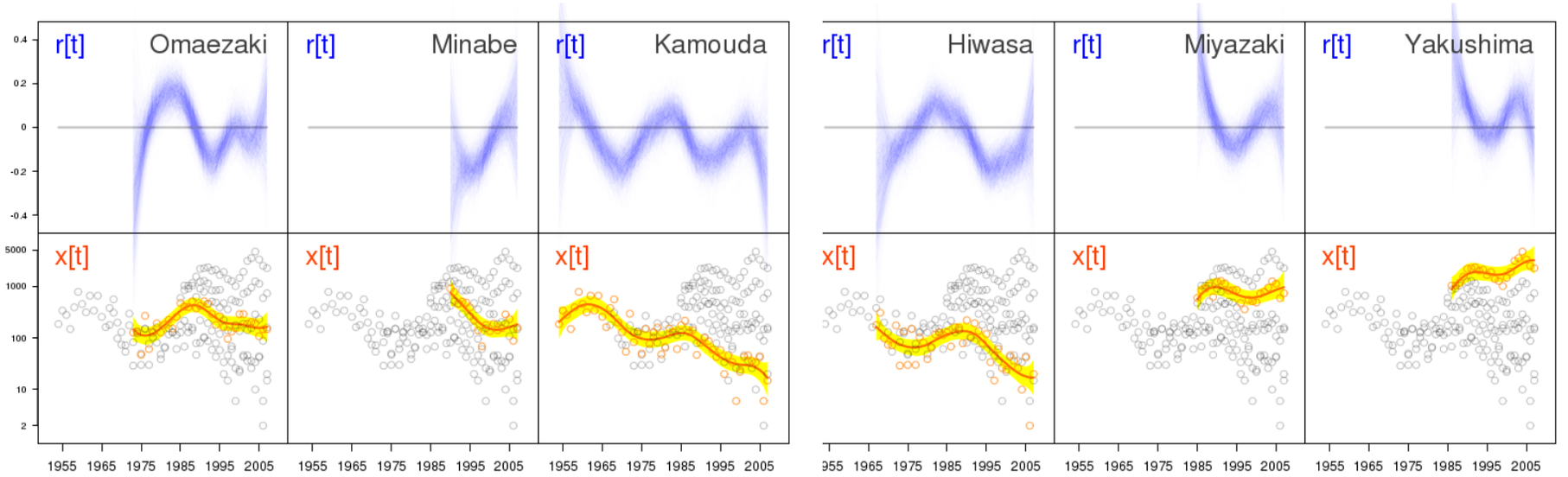


- 徳島県の 2 海岸 (Hiwasa, Kamouda) では，予測分布の 95% 区間にゼロを含まない
  - 産卵地として選ばれなくなりつつある?
- 他の海岸では何とも言えない?

全海岸に共通する年変動 a[t]

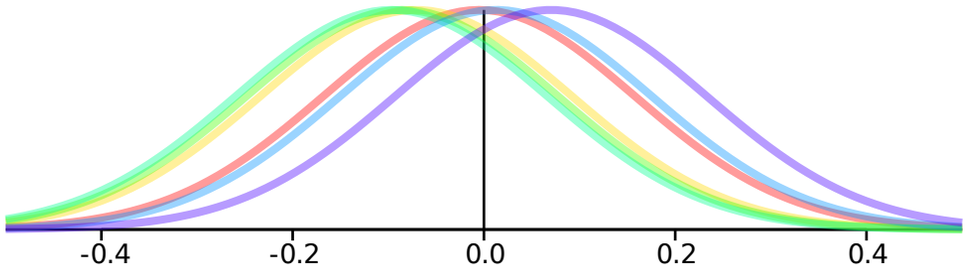
- 年ごとに独立ではなく，ゆるやかに変動?





- ウミガメの繁殖活性と気象要因との関連?

10 年後のウミガメ上陸数は?



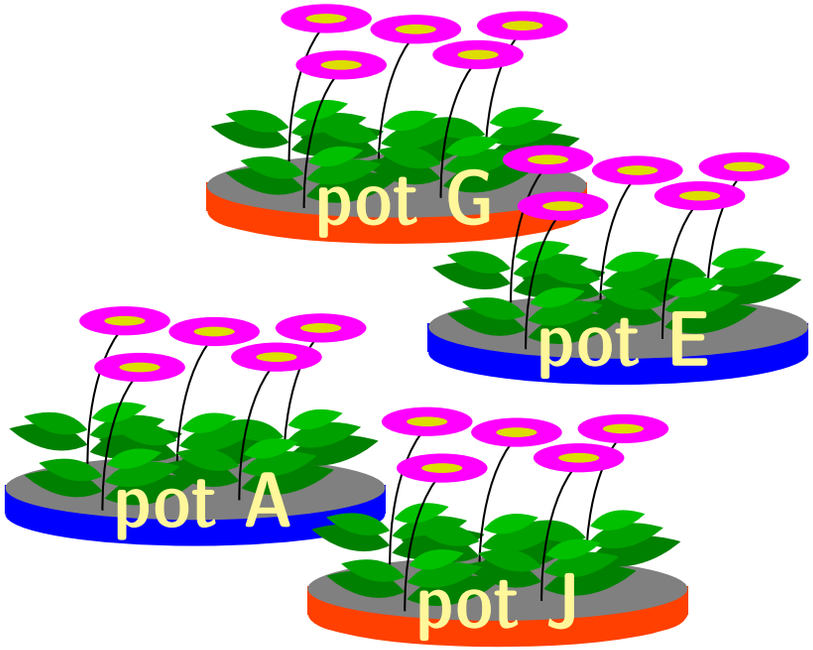
- 海岸ごとの平均成長速度 r[t] の 10 年間の変動幅を予測した
- 2006 年時点での平均成長速度 r[t] を左で推定した平均成長速度の平均の分布の中央値に等しいと仮定
- r[t] の年変動の SD は上で推定した 0.052 を仮定して 10 年間の random walk の範囲を示した

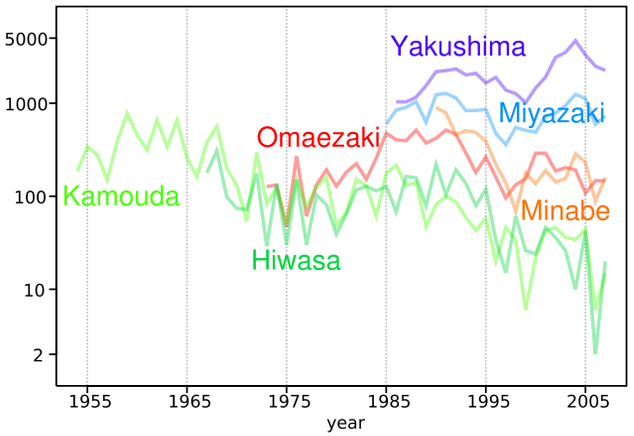
海岸	2016 年に r[t] がプラスとなる確率
Omaezaki	49%
Minabe	33%
Kamouda	30%
Hiwasa	28%
Miyazaki	53%
Yakushima	67%

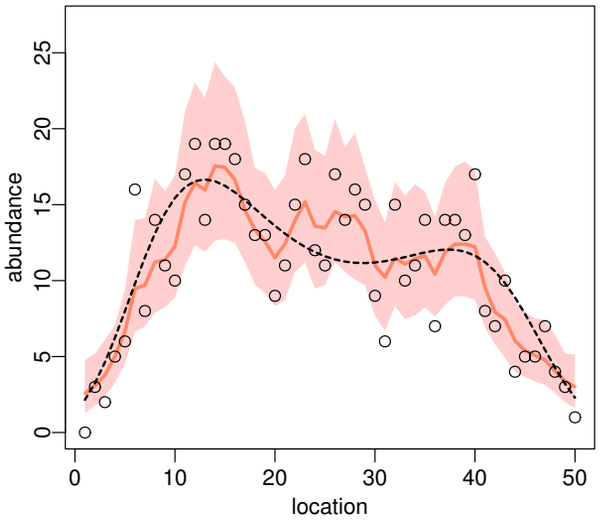
このあたりで終了

今回のハナシ: いろいろな階層ベイズモデル

1. 個体差 + ブロック差というネストしたランダム効果
2. 「隣と似ている」空間相関のあるランダム効果
3. 時間変化する潜在変数: ウミガメ上陸数の統計モデル







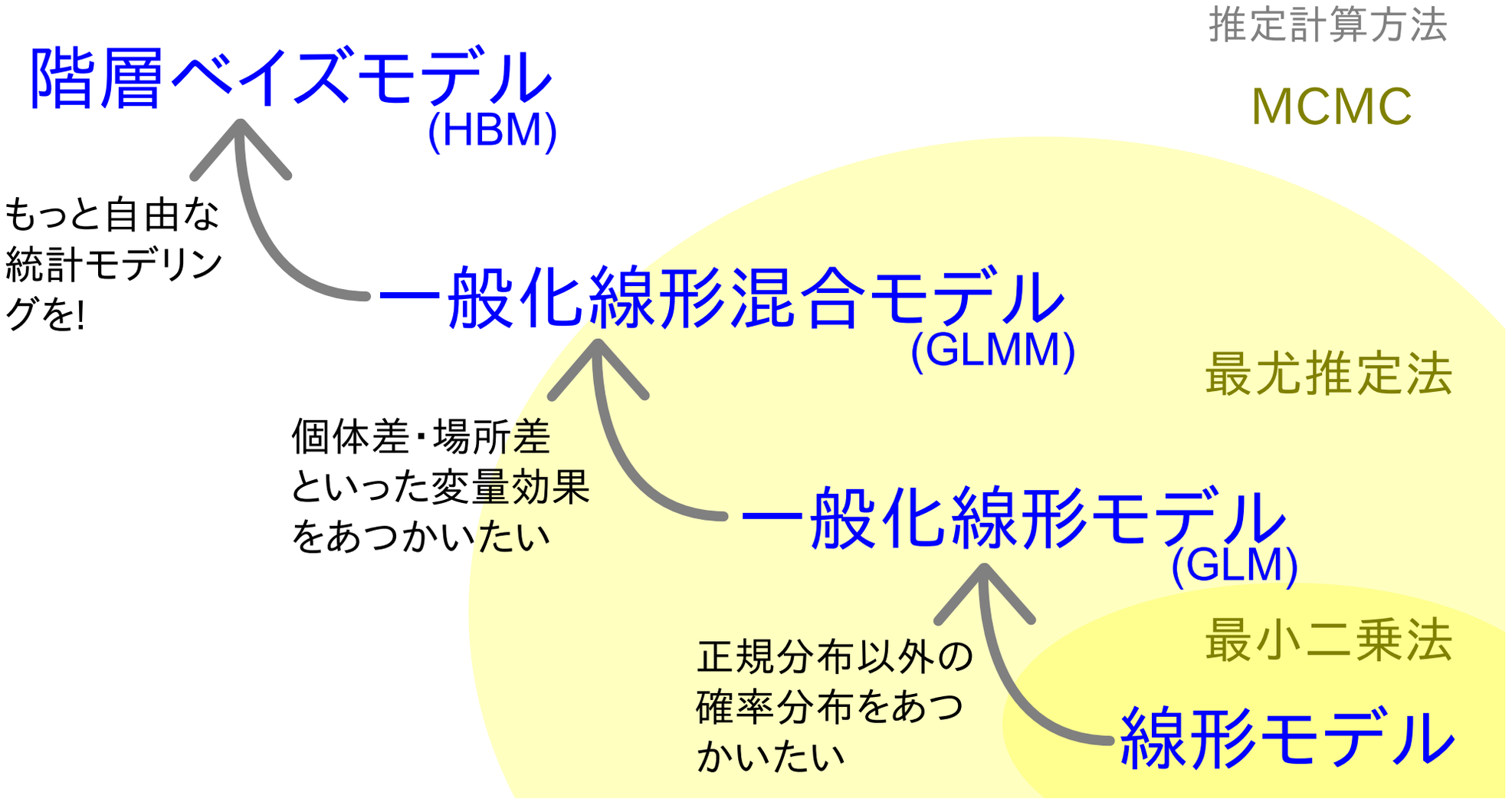
今日の話のまとめ: 階層ベイズモデルと生態学

今まで「無いこと」にされていた要因
階層ベイズモデルで明示的に組みこめるように

- 複数の **random effects** (個体差・ブロック差・……)
- 「隠れた」状態をあつかうモデル
  - 例: 時系列モデリングや「欠側値を補う」処理
- あるいは空間構造のある観測データのモデル化も
  - 例: 「隣は似てるよ」効果 – **Gaussian Random Field**

現実のあれこれを反映させられる統計モデルは
モデルによる現象の説明を促進するだろう

線形モデルの発展


推定計算方法
階層ベイズモデル
(HBM)
MCMC
もっと自由な
統計モデリン
グを!
一般化線形混合モデル
(GLMM)
最尤推定法
個体差・場所差
といった変量効果
をあつかいたい
一般化線形モデル
(GLM)
最小二乗法
正規分布以外の
確率分布をあつ
かいたい
線形モデル