

統計モデリング入門 新潟大 2015 (7a)
階層ベイズモデルの応用: 場所差と時系列

久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

新潟大学集中講義 <http://goo.gl/m8HSBM>

2015-05-27

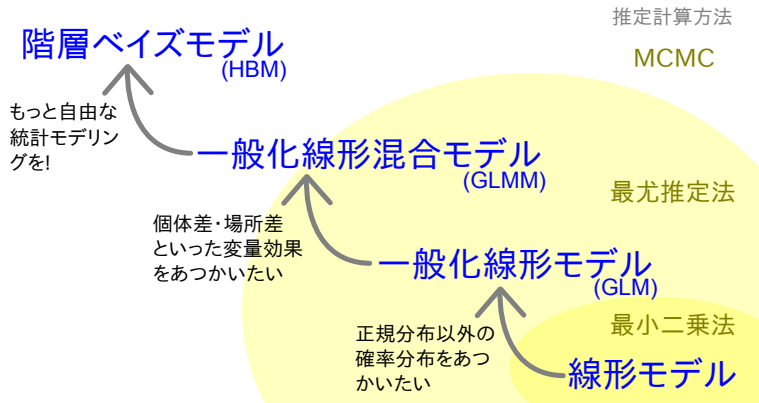
ファイル更新時刻: 2015-05-26 17:53

この時間に説明したいこと

1. 個体差 + 場所差の階層ベイズモデル (ファイル 7a)
2. 空間構造のある階層ベイズモデル (ファイル 7a)
3. 時間構造のある階層ベイズモデル (ファイル 7b)

“統計モデリング入門” に登場する統計モデル

線形モデルの発展



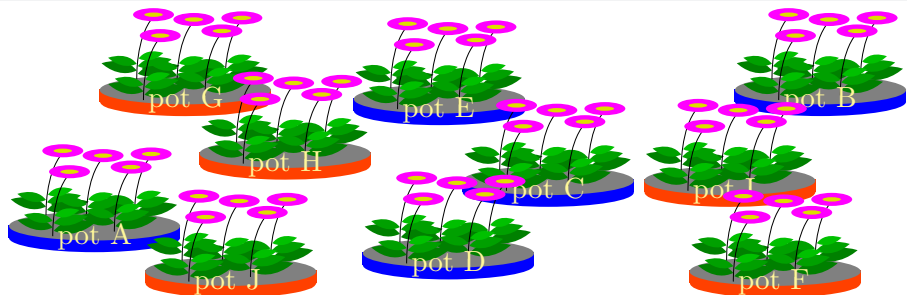
データの特徴にあわせて線形モデルを改良・発展させる

1. 複数ランダム効果の階層ベイズモデル

個体差 + グループ差, など

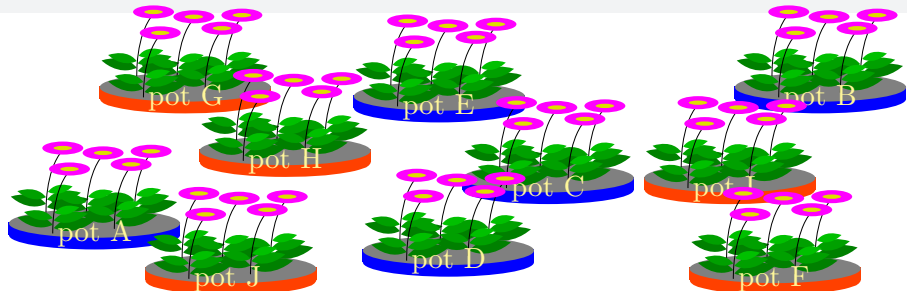
そして “てぬき” モデリングの危なさについて

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 ($f_j = \mathbf{C}$) 5 鉢 (合計 50 個体)
 - 肥料をやる処理 ($f_j = \mathbf{T}$) 5 鉢 (合計 50 個体)

このへんてこな例題, 言い換えると.....?



- 教育法をやったら児童ごとの算数の得点 y_i が増えるかどうかを調べたい
- 小学校 10 校, 各校に 10 人の児童に算数のテスト (合計 100 児童)
 - コントロール ($f_j = \mathbf{C}$) 5 校 (合計 50 児童)
 - 教育法実施 ($f_j = \mathbf{T}$) 5 校 (合計 50 児童)

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

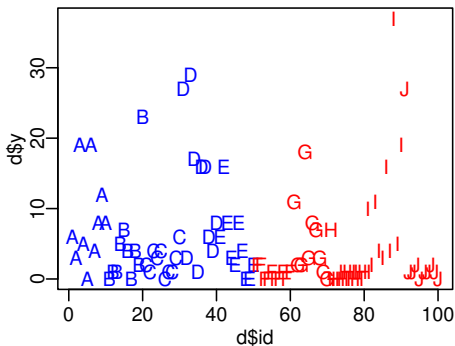
{1, 2, 3, ..., 100}

- pot 列: 植木鉢名 {A, B, C, ..., J}

- f 列: 処理: コントロール C, 肥料 T

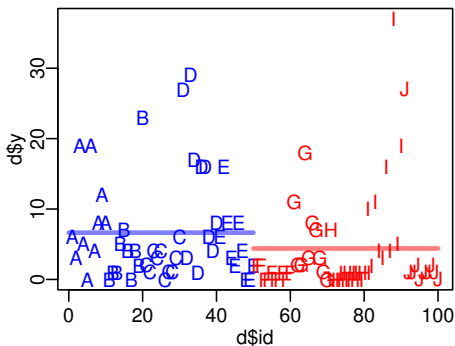
- y 列: 種子数 (応答変数)

データはとにかく図示する!!



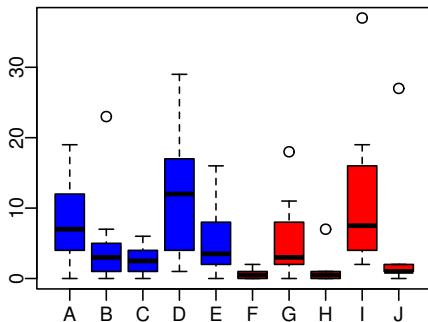
- `plot(did, dy, pch = as.character(d$pot), ...)`
- **コントロール**・**処理** でそんなに差がない?

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?
- (注) この架空データは **肥料の効果はゼロ** と設定して生成した

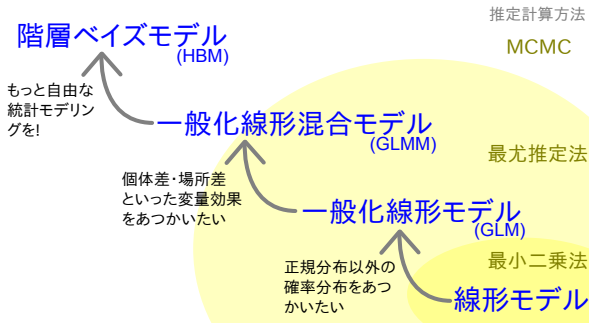
個体差だけでなく植木鉢差もありそう?



- `plot(dpot, dy, 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: 個体差だけ考慮, ブロック差は無視

```
> library(glmML)
> summary(glmML(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

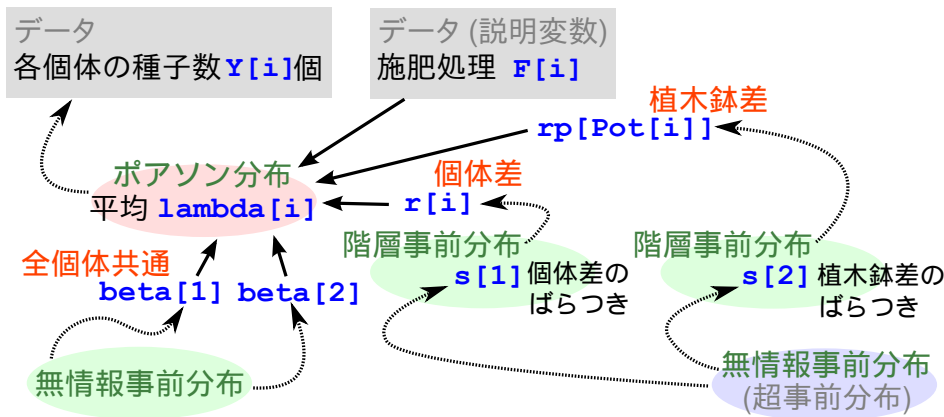
...(略)...

- 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 $\log(\lambda_i) = a + bf_i + (\text{個体差}) + (\text{ブロック差})$
- 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)

植木鉢問題の階層ベイズモデルの図示



個体差 + ブロック差のあるポアソン回帰の BUGS code (1)

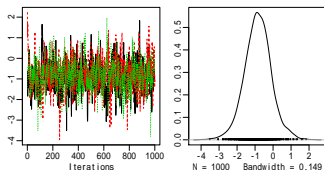
```
model
{
  for (i in 1:N.sample) {
    Y[i] ~ dpois(lambda[i])
    log(lambda[i]) <- a + b * F[i] + r[i] + rp[Pot[i]]
  }
  # 次のページの事前分布の定義につづく
```

ここでの BUGS coding のポイント

- 因子型の説明変数 $f_i \in \{C, T\}$ 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)

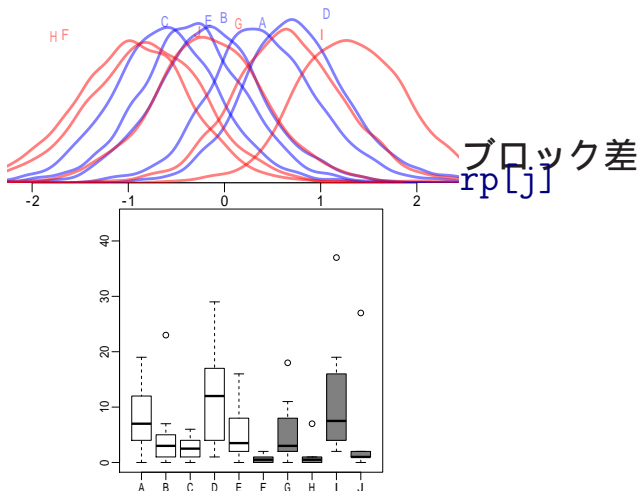
```
# 前のページからのつづき
a ~ dnorm(0, 1.0E-4) # 切片
b ~ dnorm(0, 1.0E-4) # 肥料の効果
for (i in 1:N.sample) {
  r[i] ~ dnorm(0, tau[1]) # 個体差
}
for (j in 1:N.pot) {
  rp[j] ~ dnorm(0, tau[2]) # 植木鉢の差 (ブロック差)
}
for (k in 1:N.tau) {
  tau[k] <- 1.0 / (sigma[k] * sigma[k]) # 個体・植木鉢のばらつき
  sigma[k] ~ dunif(0, 1.0E+4)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

肥料の効果 (パラメーター b) はなさそう?

	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat
a	1.501	0.529	0.482	1.157	1.493	1.852	2.565	1.00
b	-1.016	0.706	-2.436	-1.476	-0.993	-0.565	0.395	1.00
sigma[1]	1.020	0.114	0.822	0.939	1.014	1.089	1.265	1.00
...(略)...								

この架空データを生成した種子数シミュレーションでは、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)



統計モデリングの手ぬきは危険!

- **random effects** つまり 個体差・ブロック差が大きい
- **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
 - 個体差・ブロック差の階層ベイズモデルが必要!
- 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

階層ベイズモデルと GLMM の関係は?

線形モデルの発展



一般化線形混合モデル
(Generalized Linear Mixed Model) は階層ベイズモデルのひとつ

- GLMM では個体差・植木鉢差といった local parameter は積分して消去してしまう
- 階層ベイズモデルでは, 何もかも事後分布として推定してしまう

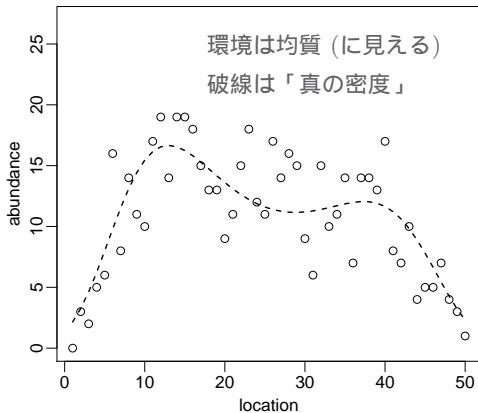
2. 複数ランダム効果の階層ベイズモデル

個体差 + グループ差, など

そして “てぬき” モデリングの危なさについて

架空の例題: 個体数データ, 一次元空間データ

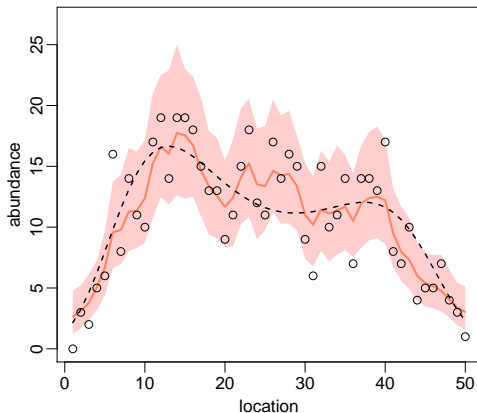
欠測データなし



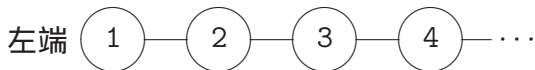
問: 空間自己相関を考慮して生物個体の密度推定

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

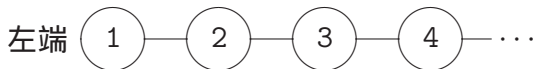


空間相関のある「場所差」階層ベイズモデル



- 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):

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

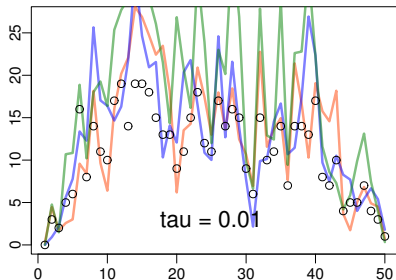
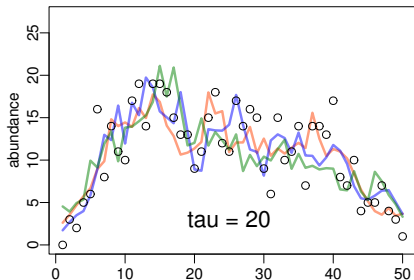
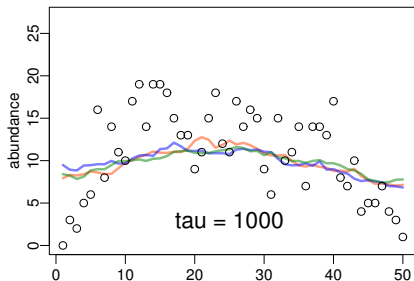
- σ は無情報事前分布にしたがう: $\tau = 1/\sigma^2 \sim \text{Gamma}(1.0^{-2}, 1.0^{-2})$
- ベイズの定理 → 事後分布の導出

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

超パラメーター τ が決める

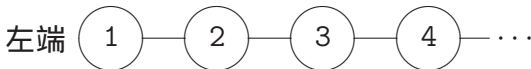
隣との類似ぐあい

- 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



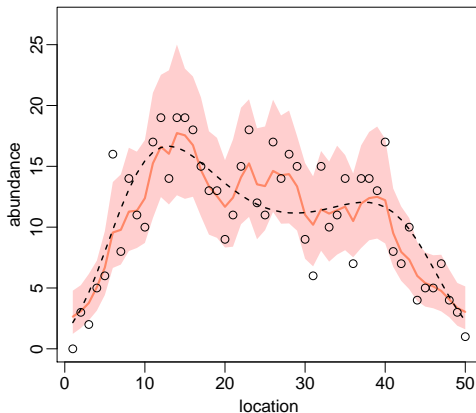
BUGS 言語: ベイズモデルを記述する言語

```
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)   # tau の無情報事前分布
}
```



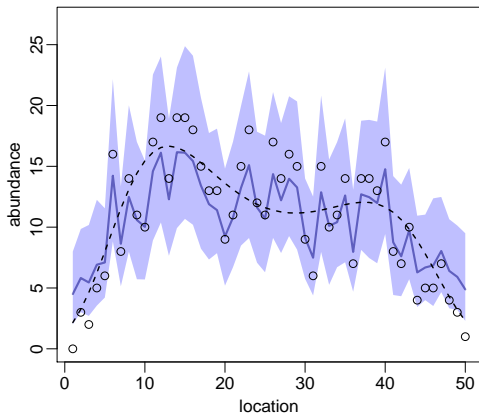
空間相関のある「場所差」モデルの推定結果

空間相関を考慮するモデル
欠測データなし



空間相関を考慮しないベイズモデルの推定結果

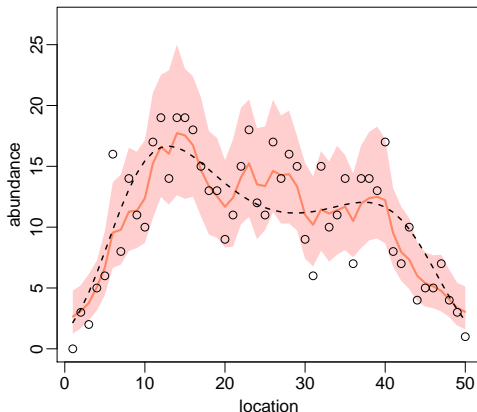
空間相関を考慮しないモデル
欠測データなし



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

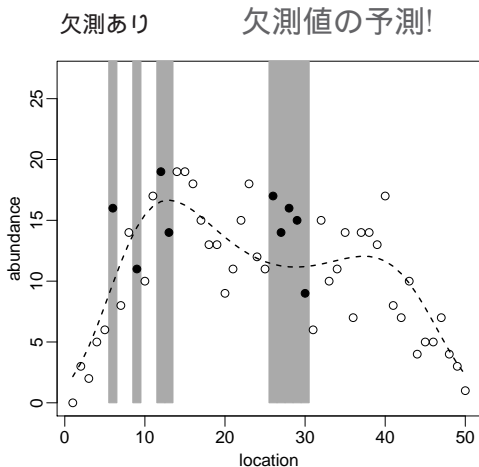
欠測データなし



空間相関を考慮しないモデル

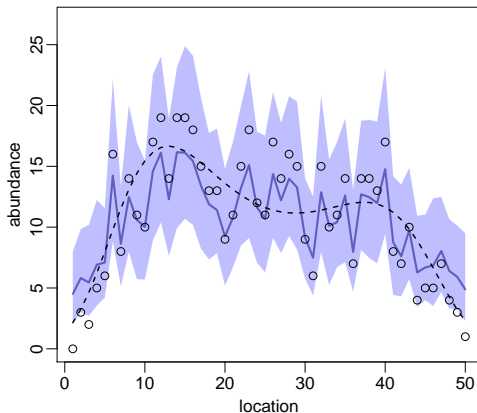
欠測データなし

架空の例題 (続): 欠測がある場合は?!



空間相関を考慮しないベイズモデルは欠測にヨワイ

空間相関を考慮しないモデル
欠測データなし

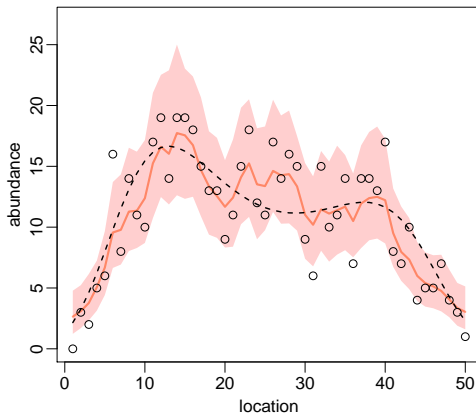


空間相関を考慮しないモデル
欠測あり

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

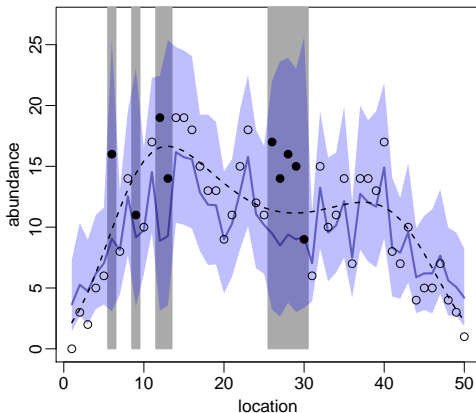


空間相関を考慮するモデル

欠測あり

ベイズモデルのご利益: 欠測をうまく対処

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

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