

簡単な例題で理解する空間統計モデル

久保拓弥

北海道大学大学院地球環境科学研究院

An introduction to spatial statistical modeling for ecologists

要旨: 観測データの背後にある生態学的なプロセスを特定するときに、データの空間構造に由来する空間的自己相関 (空間相関) のある「場所差」はとりあつかいの難しい「ノイズ」である。空間相関のある「場所差」は random effects として統計モデルの中で表現するのがよい。近年よく使われている GLMM など簡単な階層ベイズモデルでは空間相関のある random effects をうまくあつかえない。そこで空間相関をうまく表現できる intrinsic Gaussian CAR model の概要を説明し、単純化した架空データから得られる推定結果を示す。また階層ベイズモデルが威力を発揮する、欠測のある観測データが与えられた状況での推定結果も示した。

キーワード: ベイズ統計モデル, 空間的自己相関, MCMC, R, WinBUGS

はじめに

観測データの背後にある生態学的なプロセスを特定するときに、データの空間構造に由来する空間的自己相関 (spatial autocorrelation; 以下、空間相関) のある「場所差」はとりあつかいの難しい「ノイズ」である。この記事では空間相関のある架空データを使った簡単な例題とその解法を紹介し、読者がデータを解析するときの参考になるようにしたい。^{1 2}

生態学研究のデータにあらわれる空間相関とは何なのだろうか? ひとことと言えば「ある場所の観測値とその近くの観測値は似ている、しかし遠方の観測値とは似ていない」現象である。野外調査地や空間構造のある実験場のデータをあつかった経験のある人の多くはこのようなパターンを観察したことがあるだろう。たとえば、「個体群密度は調査地内で均質というよりは粗密のムラがあるようだ」「このあたり一帯の群集組成は似ているけれど、少し離れたところは全然異なる」「遺伝的類似度が距離とともに減少する」などなどといった現象である。

このような「場所差」が生じる原因はさまざまであ

る (深澤ほか 2009)。観測者が測定しなかった、もしくはできない環境要因によるものかもしれない。あるいは生物の移動分散が限定されているために、生物の分布がランダムではなく集中分布になっているのかもしれない。

しかしこの記事では、空間相関のある「場所差」が何に由来するものなのか「よくわからない」ままデータ解析しなければならない状況にあるものとして話をすすめる。つまり空間相関の発生そのものにはそれほど関心はないけれど、しかしながら空間相関を「なかったこと」にしてデータ解析するのはちよつとまずいではなからうか、と不安になる場合の解決策を提示してみたい。

この「始めよう! ベイズ推定によるデータ解析」特集内の他の記事では実データを使った解析例が示されている。しかし、この記事ではあえて実データではなく単純化した架空データの例題をつかって、空間相関を考慮した統計モデリングと従来の統計モデルのつながりに重点をおいて解説している。

この記事であつかう例題のデータやプログラムなどのファイルは、この解説記事のサポート web page (<http://hosho.ees.hokudai.ac.jp/~kubo/ce/CarNormalExample.html> あるいは生態学 `car.normal` で検索) からダウンロードできる。ソフトウェアのインストール法から結果の作図法まで具体的に紹介しているので、空間統計モデリング

¹ ここでは 2008 年 3 月の生態学会大会 (福岡) のベイズ企画集会で久保が簡単な例題として紹介した空間相関を考慮した統計モデルを推定計算する方法について解説する。

² 出典: 久保拓弥, 2009. 簡単な例題で理解する空間統計モデル (特集: 始めよう! ベイズ推定によるデータ解析). 日本生態学会誌 59: 187-196

について勉強したいと考えている読者はこのサポート web page も利用していただきたい。

例題であつかう単純化した架空データ

空間構造のあるデータやその統計モデリングとはそもそも何なのか、といったことを説明するために、まずはここであつかう例題の説明から始めることにしたい。図 1 に示しているような観測データ (架空データ) があつたとしよう。

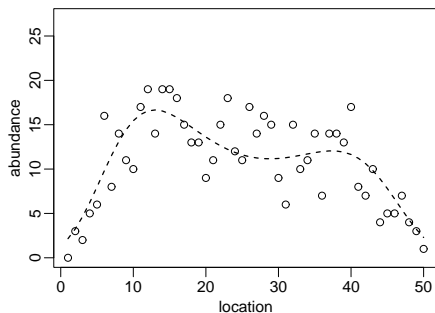


図 1: 例題の一次元空間上の架空データ。横軸は調査プロットの位置、縦軸は観測された個体数。

これは一次元空間上のいろいろな場所で観測された生物の個体数をあらわしている。ここでは、50 個の観測地点が一次元空間上に等間隔にならんでいるとしよう。(図 1 の横軸)。観測地点の名前 (location) である i は「左」から $1, 2, \dots, 50$ とする。各観測地点 i で生物の個体数 (abundance) $Y[i]$ を観測し (図 1 の縦軸)、図中の丸は観測地点ごとの観測値である。

現実のデータは二次元・三次元の空間構造をもつことが多い。しかしながら、ここでは空間統計モデリングの説明のための例題ということで、わかりやすい一次元の空間構造のデータをあつかうことにする。

これは架空データなので、(現実の世界では絶対に知りえない)「真の個体群密度」なるものが存在している。図 1 の破線は「地点 i の真の局所個体群密度」(以下では、局所個体群密度) である。この架空データは各地点の局所個体群密度を平均とするポアソン乱数で生成されたものである。しかしながら、この

架空世界の人間には局所個体群密度 (破線) は直接には観測できないものであり、さらに第 1–50 地点の環境はすべて同じ (均質な環境) に見える、としよう。

ただし図 1 に示されている観測値をみれば、「どうやらこの生物の密度は場所ごとに異なるらしい」といったことは見当はつくだろう。

問: 観測データを統計モデルで表現せよ

この架空データをつかった例題であつかう問としては「図 1 に示されている観測データを統計モデルとして表現せよ」ということにしたい。統計モデルとは観測データときちんと対応させられる数理モデルである。

むしろ現実の生態学のデータ解析では、多くの場合、ただ単に「統計モデル化しただけ」では不十分なことが多く、たとえば、「理由のよくわからない空間相関」だけでなく環境要因も考慮して説明せよとか、オス・メスで個体群密度の空間相関の異質性が異なるか調べよ、といった問題になる。

しかしながら、どのような生態学的な解析であっても、そもそも観測データを統計モデルとして表現することが必要条件となる。つまり、観測データにもとづく何かを主張するのであれば、まずはきちんと統計モデルを作ってパラメーターなどを推定をしたうえで、それらについて比較・考察しなければならない。³ 今回の例題はこの「データ解析の最初の一步」の部分だけにとりくんでいる、ということにしよう。

統計モデリングにとりくんでみる

さて、図 1 の観測データを統計モデル化するために、まずは基本的な検討からはじめてみよう。統計モデルの骨格となる部品は確率分布である。そこでまず最初に、各地点で観測された個体数がどんな確率分布で表現できるか、を考えなければならない。個体数は 1 個体、2 個体と数えられるカウントデータなので、これはポアソン分布で表現できそうだ、と考

³ データ解析というと「検定にかける」だけのことだと考えている人も多いが、「検定にかける」ときにも観測データの統計モデリングとそのパラメーター推定は常になされている。

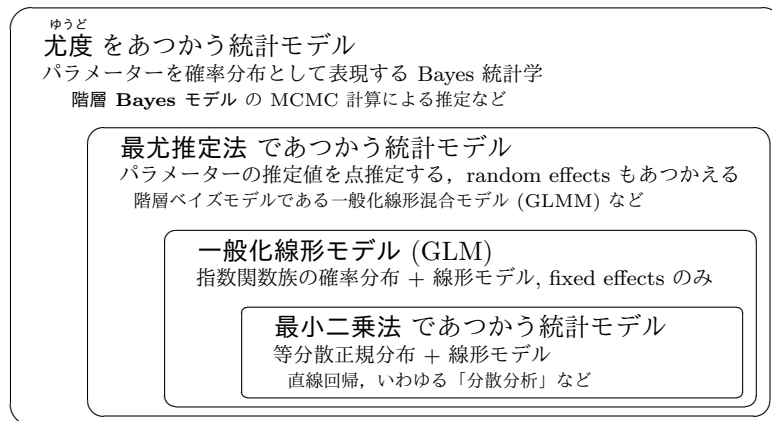


図 2: 一般化線形な統計モデルの世界。

えるのがもっとも簡単だろう。つまり図 2 で示している統計モデリングの世界地図の中でいちばん狭い部分から出て、一般化線形モデル (generalized linear model; GLM; 久保・粕谷 2006 や Crawley 2008 など参照) の世界に踏みこんでいる。

次に図 1 がポアソン分布で説明できるかどうか検討してみよう。観測者には図 1 の破線、つまり局所個体群密度は見えていないし、環境も一様であるかのように考えているので、どの場所でも個体群密度が等しい、と仮定せざるをえない。そこで、50 ケ所の観測地点で独立に観測値 $Y[i]$ が得られたものとして、この $Y[i]$ がある平均値のポアソン分布にしたがうとする。このとき、平均値は標本平均をつかえばよいので 10.9 となる (ネット上で公開しているこの例題データと R で計算できる)。このデータがポアソン分布から得られたのであれば分散も平均値と同じ 10.9 ぐらいになると期待されるのだが、 $Y[i]$ の標本分散は 27.4 であり、これは標本平均の 3 倍ちかい値である。この点から、このデータは過分散 (overdispersion; 久保・粕谷 2006; Crawley 2008) だと判断するのが妥当であり、単純なポアソン分布では統計モデル化できそうにない。そもそも、図 1 を見た時点で「各観測地点で独立」といった仮定は成立していそうにもない、とわかるだろう。

階層ベイズモデルの導入

図 1 のような観測データをうまく表現するためには、なんらかのかたちで「場所差」を考慮した統計モデリングが必要になりそうだ。このような「場所差」は random effects として統計モデルに組みこむのが良いかもしれない。というのも、fixed effects として組みこむと、50 個のデータを説明するために、観測地点ごとのパラメーター 50 個を最尤推定することになり、統計モデルとしては無意味なものになる (久保・粕谷 2006; 久保 2007)。⁴

われわれが良く使う統計モデルの多くでは、その確率分布の平均を線形モデルで定義する。この線形予測子は fixed effects と random effects に分けられる (久保・粕谷 2006; Crawley 2008)。fixed effects は平均値に影響をおよぼす要因で、しかも解析者が関心をもつ効果 (たとえば実験処理の効果) である。この例題にはそういう要因はないけれど、あえていえば全 50 観測地点の平均個体群密度が fixed effects に該当しそうだ。これに対して、random effects は「個体差」や「場所差」といったよくわからない (直接には観測されていない) 要因の効果で、確率分布の

⁴ random effects を使わないモデリングの例として、多項式を使えば「場所差」が表現できる、といったアイデアがある。しかしながら図 1 に示されているような「ぐねぐね」を表現するためにはかなり高次の多項式になるだろう。つまり多数のパラメーターの最尤推定が必要となるので、あまり良い方法ではない。なお、この記事では GLM や GLMM を話の出発点にしているので、やたらと fixed / random effects という語が登場するが、ベイズ統計ではこのあたりの区別はそれほど重要ではない。むしろ「そのパラメーターの事前分布は何か」に注意すべきである。

平均には影響を与えない。⁵

観察されたパターンを説明する要因を fixed & random effects を同時に考えるモデルを混合モデル (mixed model) とよび、GLM を mixed model にしたものが **GLMM** (generalized linear mixed model; 久保・粕谷 2006; Crawley 2008) である。混合モデルは階層ベイズモデルのひとつであり (石黒ほか 2004; 久保 2007)、その中でもっとも簡単なものは最尤推定法であつかえる (図 2)。

近年になって、生態学のデータ解析においてベイズ統計モデリングが普及しつつある (Clark 2005, 2007)。その理由のひとつは、「ベイズではパラメーターは確率分布として表現してよい」という特徴にある。このような概念にもとづく統計モデルはたいへん柔軟なものになるので、従来のモデルに比べると、たとえば「場所差」などを格段に容易に表現できるようになった。

非ベイズ的なわくぐみのもとでは、推定結果は「真の値」に近いと考えられる最尤推定値を計算しようとする (図 2)。ベイズ的なわくぐみでは、パラメーターの推定結果は確率分布つまり事後分布 (posterior) として表現され、必ずしも最尤推定値にはこだわらない。事後分布の解釈としては、事前分布 (prior) という制約をあたえたときに、統計モデル全体が観測データによくあうように選ばれたパラメーターを確率分布として表現したもの、と考えればよいだろう。もっと短くいえば「それっぽいパラメーターの確率分布」ということになる。われわれ統計学的ツールのエンドユーザーとしては、推定結果が確率分布であっても何も困らないし、推定の信頼区間の解釈はベイズのほうがわかりやすい (Clark and Gelfand 2006)。

この事前分布という概念についていくつか簡単に補足しておきたい。できるだけ値の範囲を制約しないカタチになっている事前分布は無情報事前分布 (non informative prior) とよばれる。こんにちのベイズ統計モデリングでは、fixed effects のパラメーターの事前分布は無情報にすることが多い。また「場所差」をあらわす random effects は、各地点で独立ばらばらな値をとらないように、全体に似ている部分があると仮定するので (久保・粕谷 2006; 久保 2007)、無情報ではない事前分布となる。ただしその場合でも

事前分布のカタチを決めるパラメーター (超パラメーターあるいは hyper parameter) の事前分布 (超事前分布) は無情報と設定することが多い。このように事後分布 – 事前分布 – 超事前分布 と階層化しているベイズモデルは階層ベイズモデルとよばれる。恣意的に事前分布を決めることを避け、いわば「できるだけ観測データに事前分布を決めさせる」方式である。

各地点は独立と仮定する GLMM

さて、もっとも簡単化した階層ベイズモデルである混合モデルでは random effects (この例題では全 50 観測地点の「場所差」) をあらわすパラメーターをいちいち最尤推定しない。たとえば場所番号 $i = 1$ の「場所差」をあらわすパラメーター `re[1]` の値が 1.2345... などと確定できるはずだ、とは考えない。つまり、ベイズでは「パラメーターの真の値」なるものは想定しないで、`re[1]` の値は -3 ぐらいかもしれないし $+0.5$ ぐらいかも、といった事後分布にしたがう確率変数である、と考える。

具体的に、どのように「場所差」パラメーターの確率分布 (50 ケ所ぶんの事後分布) は推定されるのだろうか? 上で説明したように (簡単版ベイズモデルである) GLMM では、場所ごとの観察データをうまく説明できるように「場所差」を強調しつつ、その一方で「場所差」がむやみやたらと大きくならないように制約しつつ、これらのバランスを統計学的な意味でとるようにしている (久保・粕谷 2006; 久保 2007)。

空間相関がない、つまり「場所差」が各地点で独立である観測データならば、このような GLMM を適用できる。たとえば、ひろく普及しているデータ解析ソフトウェアである R (R Development Core Team 2008; 久保ほか 2008) の GLMM 推定関数である `glm.nb()` 関数 (`library(MASS)`; Venables and Ripley 2002) や `glmmML()` 関数 (`family = poisson`; `library(glmmML)`; 久保 2007 などの解説参照) を使ってパラメーター推定ができる。

しかしながら、「各地点の random effects は独立」と仮定する GLMM では、この例題で考えているような、「場所差」があり、しかも「場所差」が隣とは似ている、といった空間相関がある状況には適用で

⁵ 厳密にいうと確率分布の中央値に影響を与えない効果。

きない。つまり図 1 に見られるような空間相関はうまく表現できないので、統計モデルをもっと改良しなければならない。

MCMC 計算法と WinBUGS

統計モデルをさらに改良するために必要になるので、ここで統計モデルの推定計算法について説明したい。

簡単なベイズモデル (GLMM) では、事前分布で積分した尤度を最大化するようなパラメータを推定している (階層ベイズモデルの最尤推定)。R の GLMM 推定関数はそのように数値的に最尤推定している。しかしながら、この例題のように random effects が「隣と似ている」といった複雑な状況になると数値的な最尤推定は計算技術的に難しいものになってくる。

そこで図 2 の一番外側の領域、尤度をあつかう (しかし最尤推定にこだわらない) 世界にふみこむことになる。より複雑な階層ベイズモデルによる推定計算には Markov chain Monte Carlo (MCMC) 法が使われることが多い (山道 2008 など; この特集の他の記事も参照)。

この MCMC 計算はもともとは統計物理学の分野で発展してきた手法である (伊庭 2003; 伊庭ほか 2005)。これは最尤推定法と同じく尤度 (統計モデルの観測データへのあてはまり) にもとづくものである。数値的な最尤推定法では試行錯誤によって尤度最大となるパラメータの値を推定する。ベイズモデルの MCMC による推定でもパラメータの値を試行錯誤で変化させる。しかしながら、MCMC 法ではこの試行錯誤 (以下、サンプリングとよぶ) で得られたパラメータの値のセットが事後分布からの無作為抽出標本となるような、巧妙な「値の変化のさせかた」のルールが適用されている。このようにしてサンプリングされた値のセットによって、われわれはパラメータの事後分布の性質 (事後平均値など) を推定できる。

MCMC 計算によって事後分布からサンプリングするソフトウェアは多数ある。ここでは、現時点ではもっともよく使われている WinBUGS (Spiegelhalter et al. 2004) を使った計算を紹介する。⁶ WinBUGS

⁶ WinBUGS は残念ながら Microsoft Windows 専用のソフト

では統計モデルを図 3 に示しているような BUGS 言語で定義する。図 3 の BUGS 言語のコードは図 1 の架空データを統計モデル化した例のひとつである。以下ではこのコードにそって話をすすめていこう。

空間相関を表現するガウス確率場

「場所差」をあらわす random effects である $re[i]$ が各地点で独立だと仮定するモデルでは図 1 のようなパターンをうまく表現できないようなので、 $re[i]$ が「隣」どうしでおたがい似ている (空間相関がある) ような統計モデルを構築したい。

ある観測地点 i での個体数の観測値 $Y[i]$ が i の局所密度でもある $mean[i]$ を平均とするポアソン分布にしたがっている、という関係を BUGS 言語では

$$Y[i] \sim dpois(mean[i])$$

と記述できる。この平均 $mean[i]$ の対数が「切片」 (全体の対数平均密度) である $beta$ と「場所差」である $re[i]$ の和になっているなら、

$$\log(mean[i]) \leftarrow beta + re[i]$$

と書けばよい (GLM の用語でいえば log link 関数)。

この「場所差」の random effects をあらわす $re[i]$ をどう表現すればよいのか、それがこの統計モデリングのかなめになる。

たとえば、R の `glmmML()` つまり GLMM をつかって統計モデリングした場合、 $re[i]$ は各地点独立で同じ正規分布 (平均ゼロ、分散の逆数 τ) にしたがう、と仮定することになるので、

$$re[i] \sim dnorm(0.0, \tau)$$

と記述することになる。GLMM はベイズモデルの一種であるので、上の関係を見て「 $re[i]$ の事前分布は平均ゼロ、分散の逆数が τ の場所 i ごとに独立した正規分布である」と言える。

「場所差」の random effects $re[i]$ が「各地点独立」ではなく「独立ではない、隣とは似ている」ソフトウェアである。Mac OS X や Linux 上でも工夫すれば使える。他の MCMC 計算ソフトウェアについてはサポートページを参照。

```

model
{
  Tau.noninformative <- 1.0E-2      # 無情報事前分布（正規分布）の分散逆数
  P.gamma <- 1.0E-2                # 無情報事前分布の（ガンマ分布）の分散逆数
  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]    # log(密度) は (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, Tau.noninformative) # 全体の平均は無情報事前分布にしたがう
  tau ~ dgamma(P.gamma, P.gamma) # 場所差のばらつきは無情報事前分布にしたがう
}

```

図 3: この架空データの例題を解決するための BUGS コード。 `car.normal()` の random effects 項をもつポアソン回帰になっている。

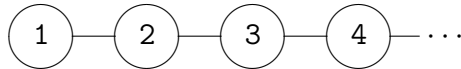


図 4: 例題の架空データの空間構造。

とを表現したいのであれば、それにふさわしい確率分布を選ばばよい、と考えるのは自然な発想だろう。この例題の架空データの空間構造は図 4 のようになっている。そこでこの $re[i]$ をまとめた $\{re[1], re[2], \dots, re[50]\}$ というベクトルの事前分布を 50 次元の多変量正規分布とすればよい。というのも、分散共分散行列で地点間の相関を定義できるからだ。

しかしながら、観測データから 50 次元の分散共分散行列を推定するのは、現代の計算機でもむずかしい。そこでいろいろな制約条件をつけて問題を単純化してみる:⁷ ある場所 i の「近傍」は有限である ($Num[i]$ 個; 図 3 の BUGS コードも参照); 近傍 j の影響はどれも等しく $1 / Num[i]$ もしくはゼロ (j に依存せず $Num[i]$ だけで決まる); $re[i]$ の分散の逆数は $Num[i]tau$ (tau はどの場所でも同じ)、など。このように条件を限定すると、各地点の random effects である $re[i]$ の条件つき確率分布を明示的に書くことができ、これは局所的な条件つき事前分布になっている。ある地点 $re[i]$ 以外のすべての

$re[*]$ が与えられているとしよう。この条件のもとでは、近傍 (つまりとなりの調査地) である $re[i - 1]$ と $re[i + 1]$ の平均値を平均とし、分散逆数が $tau \times Num[i]$ である正規分布が $re[i]$ の条件つき確率分布となる (Thomas et al. 2004)。たとえば図 4 の $i = 2$ の場所の $re[2]$ (近傍は 1 と 3 の 2 個) の条件つき確率分布は平均 $(re[1] + re[3]) \times 0.5$ で分散逆数が $2 tau$ の正規分布である。

このような (近傍の状態がとりあえず決まっているという条件のもとでの) 条件つき確率分布で各地点の random effects が定義されるモデルは確率場 (random field または Gaussian random field; 間瀬・武田 2001; Banerjee et al. 2004) とよばれ、このような確率場は条件つき自己回帰 (conditional autoregressive; CAR または Gaussian CAR) で生成される。とくに上のようにさまざまな制約をつけてパラメーター tau (と近傍指定だけで) 計算できるように単純化したものは intrinsic Gaussian CAR model とよばれる (深澤ほか 2009)。

MCMC 計算ソフトウェア WinBUGS では `car.normal()` 関数を使って計算できる (図 3; Thomas et al. 2004)。この場合の MCMC 計算では地点 i の周辺の $re[i - 1]$ や $re[i + 1]$ だけでなく、観測データ $Y[i]$ へのあてはまり (尤度) も考慮された条件つき確率分布から $re[i]$ の値がサンプリングされる。`car.normal()` 関数は、すべての地点の $re[i]$ について同じように逐次的にサンプリングしている (山道・角谷 2009)。

⁷ 空間相関をあらゆる多変量正規分布のあつかいを単純化する方法には他にもいろいろある。ここで紹介している、局所状態の確率分布を単純な正規分布で与える、という方法はじつは多変量正規分布とのつながりがわかりにくい。

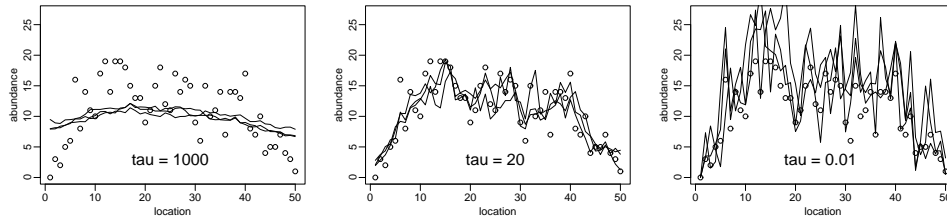


図 5: `car.normal()` の分散逆数 τ と random effects の関係。図 3 で τ をある値に固定して確率場を定義し、3 つの τ ごとに「場所差」のパラメーター $\{re[1], re[2], \dots, re[50]\}$ を 3 セットずつ WinBUGS で事後分布からサンプリングしたもの。分散の逆数 τ が大きければ「隣と似ている」度が高くて全体のばらつきも小さく (左)、 τ が小さければ「隣と似ている」度が小さくなり全体のばらつきも大きくなる (右)。

このようなモデリングによって空間相関を統計モデルとして表現できる。ただし空間相関の強弱をわかりやすく示す相関係数のようなパラメーターは存在せず、 τ の大小で「隣と似ている」度が表現される (深澤ほか 2009)。図 5 を見てもらうと「なんとなく」わかるように分散の逆数である τ が大きければ「隣と似ている」度が高くて全体のばらつきも小さく (図 5 の左)、 τ が小さければ「隣と似ている」度が小さくなり全体のばらつきも大きくなる (図 5 の右)。

推定計算の方法と準備のおおまかな説明

それでは図 3 で定義されている、空間相関のある random effects をくみこんでいる階層ベイズモデルを WinBUGS で推定計算する方法についておおまかな説明をしてみよう。もっと具体的で詳細な説明はこの解説記事のサポート web page に掲載されている。またこのサポートページからこの計算に必要なデータファイルとプログラムファイルがダウンロードできる。

WinBUGS はそれ単独でも動作するソフトウェアである。しかしながら WinBUGS 単体では何とも使いづらいので、R を介して使うことが多い。作業の流れとしては以下ようになる:

1. R で観測データ (ここでは架空データ) の準備、空間構造の定義、パラメーターの初期値などを設定

2. R が WinBUGS を呼び出してベイズモデルのパラメーター推定のための MCMC 計算させる
3. WinBUGS の推定計算結果を R がうけとる
4. R 内で収束診断やさまざまな作図・作表など

したがって、この推定計算のためには、まず R と WinBUGS をインストールする必要がある。これらのインストール方法は、この解説記事のサポート web page からリンクしているそれぞれの「インストール方法」解説ページを参照してもらいたい。

さらに R から WinBUGS を使うためには、R の追加 package である `library(R2WinBUGS)` をインストールする必要がある。これは R 内で `install.packages("R2WinBUGS")` とすればよい。

さらに必要なファイルをこの解説記事のサポート web page からダウンロードする必要がある。このときにダウンロードするファイルについて一点だけ補足説明しておく、`R2WBwrapper.R` という「R2WinBUGS をラップする関数群」を定義したファイルが含まれている。R2WinBUGS は (これまた) そのままでは何とも使いづらいので、より簡単にあつかうための関数を準備したものである。

空間相関を考慮した推定結果

サポート web page どおりにソフトウェアなどを動かすことができたとする、図 6 (A) のような結

果がえられる。⁸

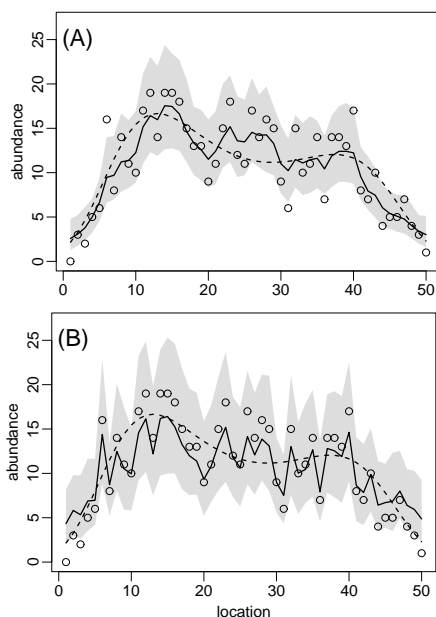


図 6: 空間相関ありモデルの推定結果。(A) 空間相関をいれた統計モデルから推定された「場所差」の事後分布。図 1 のデータと重ねている。(B) 各地点独立な random effects があると仮定して推定した結果。

図 6 (A) では、推定結果は全体の平均 β と各地点の random effects である $re[i]$ の和の事後分布の中央値 (黒線) と 80% 信頼区間 (グレイの領域) を示している。

事後分布の信頼区間の範囲はそれなりに大きいのが、中央値は観測データに「ひきよせられ」つつも、「隣と似ている」ように選ばれているので破線で示されている「真の個体群密度」に近いものになっている。

ベイズモデルを MCMC 計算によって推定すると、 β や τ といったパラメータの事後分布も推定結果として得られる (図 7)。これらのパラメータの事前分布は無情報事前分布 (図 3) と指定していた。観測データと事前分布を組み合わせることで、統計モデルのあてはまりがよくなるような事後分布が推定されている。図 5 では τ には固定した値を指定していたが、図 7 では「 τ がとりうるいろいろな

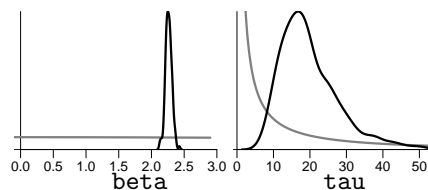


図 7: パラメータ β と τ の事後分布。黒の曲線で描かれているのが事後分布、グレイの曲線は無情報事前分布。

妥当な範囲」でサンプリングしていることになる。

各地点独立と仮定した推定結果

空間相関を考慮した統計モデルと比較するために、今度は空間相関を考慮していない random effects を使った統計モデルの推定結果を示してみよう。これは上で述べたように簡単なベイズモデルつまり GLMM であり `glmmML()` など R の推定関数で計算できる。

しかしここでは比較のため WinBUGS で推定計算する。GLMM を定義する BUGS code はこのように書ける。

```
model
{
  Tau.noninformative <- 1.0E-2
  P.gamma <- 1.0E-2
  for (i in 1:N.site) {
    Y[i] ~ dpois(mean[i])
    log(mean[i]) <- beta + re[i]
    re[i] ~ dnorm(0.0, tau) # ここに注目!
  }
  beta ~ dnorm(0, Tau.noninformative)
  tau ~ dgamma(P.gamma, P.gamma)
}
```

これは 図 3 の BUGS コードとほとんど同じであるが、「場所差」 random effects である $re[i]$ が独立同分布な正規分布になっているところが異なっている。

この推定結果は 図 6 (B) で示しているようなものとなる。定量的な差異はここでは示さないが、図 6 (A) と比べると場所ごとの「がたつき」が大きい推定結果のように見える。これは $re[i]$ が比較的自由に選べるようにモデル化されているためで、各地点の観測値にひっぱられた $re[i]$ が選ばれていて、

⁸ これぐらい簡単なベイズ推定であれば、計算時間はたいていの場合 10 秒以内に終了する。

破線で示されている「真の個体群密度」からの乖離が大きくなっている。

ベイズが強い状況: 欠測のある観測データ

図 6 (A) と (B) の推定結果を見くらべて、「まあ、GLMM でもそれっぽい結果がえられるんだから、べつに MCMC 計算や `car.normal()` なんか使わなくてもいいじゃない」といった意見もあるかもしれない。そこで空間相関を考慮したベイズ統計モデルの強みのひとつである「データに欠測がある状況」の簡単な例を示してみたい。

まず図 1 の架空データから何点かのデータを取りのぞいたデータセット (図 8) をつくる。つまり観測地点は一直線上にならんでいるのだけど、そのうち何点かでは (なんらかの理由で) 調査そのものできなかった、という想定ということにしよう。あるいは観測地点が等間隔に並んでいるのではなく、観測地点間の距離がばらばらである、という状況だと考えてもよい。これらは現実になりそうな状況だろう。

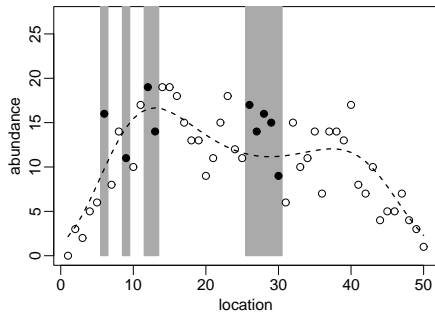


図 8: 例題の架空データ (欠測あり)。図 1 に示されているデータのうち、いくつかは欠測であったという想定架空データ。グレイの帯が「観測できなかった」場所であり、黒い丸が観測されなかった個体数。

この欠測ありの観測データに対して、先ほどの空間相関を考慮した・しなかった random effects を入れたベイズモデルを適用してみよう。WinBUGS は (R と同じく) 欠測をうまくあつかえるのでデータに欠測があっても BUGS 言語で書かれた統計モデルをとくに変更することなくそのまま推定計算できる。

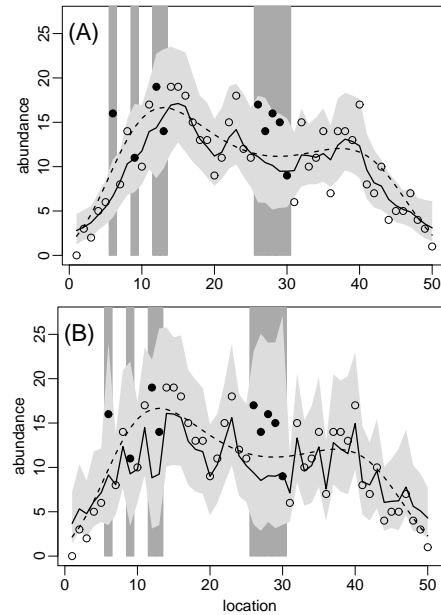


図 9: 欠測をふくむデータを使った推定結果。(A) 空間相関をいれた統計モデル (B) 各地点での独立性を仮定した統計モデルで推定した結果。

まず空間相関のあるモデル (`car.normal()` を使ったもの) の推定結果は図 9 (A) のようになった。これは全データがそろった状況で推定した図 6 (A) とあまりかわらないものになった。そもそもこの架空データの「真の個体群密度」(破線) がなだらかに変化するものであったので、「隣は似ている」と仮定する `car.normal()` を利用したベイズモデルが欠測観測地点の `re[i]` をうまく補間している。

これにて対して、各地点独立とした場合の推定結果は図 9 (B) のようになった。データが欠測した観測地点の `re[i]` は (上とは異なり地点間の「たすけあい」がないので) 大きくばらついていることがわかる。

統計モデリングが重要になる時代

この特集の他の記事でも示しているように、`car.normal()` などを使った階層ベイズモデルは柔軟で表現力があるので、生態学の観測データでしばしば見られる空間相関のモデリングが可能になった。この記事で紹介した例題とその解決はしごく単純なも

のであったが、より複雑な状況にも対応可能である。

たとえばこれらの調査地のうちランダムに選んだ何カ所かで何か実験処理をして個体群密度に影響がどうかどうかを調べたいとしよう。その場合、実験処理は fixed effects なので、この 図 3 で示している BUGS コードの beta まわりを改造して実験処理という要因に対処すればよい。空間相関のある random effects がある状況のもとでも、処理という fixed effects の強さをより正確に推定できる。

生態学の新しい知見はつねに観測データの中にある。観測データを統計モデルとして表現することで、やっかいな (あえて言えば) 「ノイズ」の背後にある生態学的なプロセスがよく理解できることがある。このように観測データにむきあって生態学的な現象を解明していこうとする立場は、やみくもに「検定にかけ」てひたすら $P < 0.05$ ばかりを探索する (ありがちな) 方法論とはちょっとちがっているようだと思われた読者も多いだろうし、あるいは現実の観測データとの定量的な比較方法が定義されていない数理モデルだけを使った研究とも異なっている。いまや科学の多くの分野で普及しつつあるベイズ統計学にもとづくデータ解析は、観察された現象と生態学的な理解をより巧妙に接続する理論であり、今後のデータ解析になくしてはならない道具となるだろう。

さて、生態学の道具としてこういった統計モデルを使っていこうという人たちは何をどう勉強していけばよいだろうか? 図 2 でいえば、まずは GLM 周辺をよく理解し、R や WinBUGS を操作する言語を習得しながら、自分のデータに適用してみる試行錯誤が必要になるだろう。ここで満足せずに、さらに観測データとモデルの推定結果をよく見くらべて、統計モデルの改善をくりかえしていけば、ベイズの領域にたどりつくことになる。この世界に入れば「この統計モデルでは〇〇ができない」といった制約が減るので、かなり自由なモデリングが可能になる。あとは文献を読んだり仲間と議論しながら「統計モデルによる現象のより良い表現」の技術を楽しく勉強できるだろう。

参考文献

- Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC, London
- Clark JS (2005) Why environmental scientists are becoming Bayesians. Ecology Letters 8:2-14
- Clark JS (2007) Models for Ecological Data. Princeton University Press, Princeton
- Clark JS and Gelfand AE (2006) Hierarchical Modeling for the Environmental Sciences. Oxford University Press, New York
- Crawley MJ (2008) 統計学 : R を用いた入門書 (野間口謙太郎・菊池泰樹訳). 共立出版, 東京
- 深澤圭太・石濱史子・小熊宏之・武田知己・田中信行・竹中明夫 (2009) 条件付自己回帰モデルによる空間自己相関を考慮した生物の分布データ解析. 日本生態学会誌 59: 171-186
- 伊庭幸人 (2003) ベイズ統計と統計物理. 岩波書店, 東京
- 伊庭幸人・種村正美・大森裕浩・和合肇・佐藤整尚・高橋明彦 (2005) 計算統計 II マルコフ連鎖モンテカルロ法とその周辺. 岩波書店, 東京
- 石黒真木夫・松本隆・乾敏郎・田邊國士 (2004) 階層ベイズモデルとその周辺. 岩波書店, 東京
- 久保拓弥・粕谷英一 (2006) 「個体差」の統計モデリング. 日本生態学会誌 56: 181-190 * ネット上で検索・ダウンロード可能
- 久保拓弥 (2007) 階層モデルで「個性」をとらえる. 数学セミナー 554: 16-22 (2007 年 11 月号) * ネット上で検索・ダウンロード可能
- 久保拓弥・竹中明夫・粕谷英一 (2008) R で改善する生態学のデータ解析. 日本生態学会誌 58: 219-224 * ネット上で検索・ダウンロード可能
- 間瀬茂・武田純 (2001) 空間データモデリング. 共立出版, 東京
- R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
- Spiegelhalter D, Thomas A, Best N, Lunn D (2004) WinBUGS User Manual (version 1.4)
- Thomas A, Best N, Lunn D, Arnold R, Spiegelhalter D (2004) GeoBUGS User Manual (version 1.2)
- Venables WN, Ripley BD (2002) Modern Applied Statistics with S (4th ed.). Springer, New York
- 山道真人 (2008) 階層ベイズとその反応拡散モデルへの応用. 数理生物学会ニュースレター 56: 10-15
- 山道真人・角谷拓 (2009) マルコフ連鎖モンテカルロ (MCMC) 法を用いたシミュレーションモデルのパラメーター推定: ベイジアンキャリブレーション入門. 日本生態学会誌 59: 207-216