Bayesian Inference　Chapter 8 Hidden Data Models pp.176-186
“＃”ではじまるコメントは安田が挿入したものです。

8.3 Occupancy models as hierarchical logistic regressions

MacKenzie et al.(2002)が検討した占有モデル（occupancy model）は特に生態学者にとって興味ある間違えのあるデータ問題（＃偽陰性あり、欠損値あり）の種類を扱う。この基本的問題は、検出方法が不完全なときに、あるエリア内での複数種のあるなしをモデル化することである。

１種のデータを考えることから始める。X_iはサイトiが”占有されている”（種がいる）ことを示すとしよう。そして、$Y_{ij}=1$は調査時期j、サイトiで種が検出されたことを示すとしよう。誤検出はないと仮定する：もし$X_i=1$でないかぎり$Y_{ij}=1$とならない。

任意のjで$Yij=1$なら$X_i=1$であるという点で、X_iは部分的に観察される。しかしながら、すべての調査の最中にその種の検出をし損ねるなら$X_i=0$か$X_i=1$となり、たとえ出現していたとしても種の検出をし損ねることになる。つまり、偽陰性（＃１なのに０としてしまう）となりうる。この問題はまさにSection 8.2で検討したrandomized response problemの拡張に過ぎない：

- “場所（Sites）”は“個体（Individuals）”に相当する。
- “占有（Occupancy）”は行動χに決して関係していない状態に相当する。
- “非検出（Not Detected）”は“Yes”の応答 – あいまいな応答に相当する。行動Xがないかもしれないし、あるかもしれない；同様に、非検出は１つのサイトが占有されていなかったことを意味するかもしれないが、実際に占有されていたサイトで占有されていないとされたのかもしれない。両者のケースで、真の状態はベルヌーイ試行で隠されている（一方でコインフリップし、他方でと成功観測の試み）。
- “検出（Detected）”は“No”の応答 － 明白な応答に相当する。“No”応答は行動χがないという確信を与える。種の検出が占有していたことを確信させるように。

占有データとの違いはある１つのサイトで多くの応答が得られることであり、まるで独立したコインフリップごとに行動χについて繰り返し訪ね、個人ごとの回答を記録し続けるるようなもの。多応答によって検出確率（detection probability）の推定が与えられるので、行動χ問題に関して、これは必要ない。なぜなら$\pi=1/2$とわかっていたから。

パラメータψ_i = Pr(サイト i が占有されている)（＃占有確率）をもつベルヌーイ型の確率変数（Bernoulli random variable）と検出確率(detection probability)

p_{ij} = Pr(調査時期j、サイトiで検出｜サイトiが占有)

でモデル化するときに、MacKenzie et al.(2002)が記載したODL（observed data likelihood）に関するモデルは、少なくとも１つのjで$Y_{ij}=1$だったサイトiでの尤度

$$\Psi_i \prod_{j=1}^{t} p_{ij}^{Y_{ij}} (1 - p_{ij})^{(1-Y_{ij})}$$

すべての j で Y_{ij}=0 だったサイト i での尤度

$$(1 - \Psi_i) + \Psi_i \prod_{j=1}^{t} (1 - p_{ij})$$

識別可能性（＃over, just,underidentified のことか？）に関して、制約は p_{ij} と ψ_i に課されなければならない。たとえば、サイトにわたって確率は同じである、あるいは観察された共変動を使ったこれらのモデリングである。

これらのモデルを考える有効なやり方は部分的な観察あるいは間違えのあるデータに関する回帰モデルとして扱うことである。検出指標のベクトル $\boldsymbol{Y}_i$ ＝（Y_{i1}, Y_{i2},…,Y_{it}）′はサイト i で t 回調査したことを示す。$\boldsymbol{p}_i$ = (p_{i1}, p_{i2},…, p_{it})'は共変量ベクトル $\boldsymbol{W}_i$ とパラメータベクトル $\boldsymbol{\gamma}$ によって決定されるとする（＃関数的依存）。さらに占有確率 ψ_i は共変量ベクトル $\boldsymbol{Z}_i$ とパラメータベクトル $\boldsymbol{\beta}$ によって決定されるとする。最終的に $\boldsymbol{Y}$（添え字なし）は Y_i の集合として定義され、そして同様に $\boldsymbol{W}$、$\boldsymbol{Z}$、$\boldsymbol{p}$、$\boldsymbol{\psi}$、$\boldsymbol{X}$ が定義される。このとき、モデルの CDL(Complete data likelihood)表記は以下によって与えられる

$$[Y \mid X, \gamma, W][X \mid \beta, Z] \tag{8.5}$$

[$\boldsymbol{X}|\boldsymbol{\beta},\boldsymbol{Z}$]は（たとえば）部分的に観察された占有の確率変数 $\boldsymbol{X}_i$ に関するロジスティック回帰で記載されたり（＃たとえば種の在不在と環境要因の関係）、[$\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\gamma},\boldsymbol{W}$]はデータにノイズが入る機構（data corruption mechanisms）を表す（＃$\boldsymbol{W}$ はその時の天気とか）。

ベイズ推測に関する問題として、Eq.（8.5）の CDL 表記は多くの長所を提供する：

(i) 生物学者の興味ある適切なレベルでの推測に焦点を合わせるのに役立つ。
つまりはモデル[$\boldsymbol{X}|\boldsymbol{\beta},\boldsymbol{Z}$]。

(ii) 事後予測分布（5.1.2）を基礎として、観察されていない $\boldsymbol{X}_i$ の推測の基礎として役に立つ。

(iii) 明確な完全条件付分布に関する MCMC サンプラーの構築を促進する。

これら３つについて、(i)と(ii)はサイエンティフィックな推測過程でとても重要である。しかし、これら恩恵は問題にモデルをフィットさせる我々の能力を必要とする。３つめの点について、ギブスサンプラーを使って示してみよう。

Gibbs Sampler for Occupancy Model {p_t,ψ} **占有モデル{p_t,ψ}のギブスサンプラー**

モデル{p_t,ψ}のもとでの完全条件付分布を記載する。つまり、ψはサイトにわたって一定であり（＃E（ψ）が一定？）、種の検出はサイト間で一定だけど調査時期に依存すると仮定している。調査時期 j でサイト i を調査していないときには Y_{ij}=NA（＃not available、欠損値）とする。このとき、

(i) p_j に事前分布(prior)としてベータ分布 Be(α_j、β_j)を設定すると、完全条件付分布[p_j|・]はベータ分布 Be(a_j,b_j)となる。

$$a_j = \alpha_j + \sum_{i:Y_{ij} \neq NA} Y_{ij}$$

$$b_j = \beta_j + \sum_{i:Y_{ij} \neq NA} X_i (1 - Y_{ij})$$

a_j の定義に使われている合計部分は時間 j で調査したサイトの中で検出サイト数、b_j の定義に使われている合計部分は時間 j で調査したサイトの中で非検出サイト数である。

(ii) ψの事前分布としてベータ分布 Be(a_ψ、β_ψ)を設定すると、完全条件付分布[ψ|・]は Be（a_ψ、b_ψ）となる。$a_\psi = \alpha_\psi + \sum_i X_i, \quad b_\psi = \beta_\psi + n - \sum_i X_i$

(iii) 非検出サイト（調査したけど NA を除いて全部 0）での X_i はわからない。ギブスサンプリングの実装には完全条件付分布からのサンプリングが必要で、以下のパラメータを持つベルヌーイ分布を考える。

$$\pi_i = \frac{\psi \prod_j q_{ij}}{(1-\psi) + \psi \prod_j q_{ij}}$$

ここで $q_{ij} = (1 - p_j)^{\mathbf{I}(Y_{ij} \neq NA)}$ （＃見つけられない確率）。注意として、πi は実際は占有サイトだったけれども、非検出と判断されたサイトの占有確率である。

このギブスサンプラーを適用できるサンプルデータセットを示す。

8.3.1 American Toads in Maryland

Mackenzie et al.(2002)はアメリカヒキガエル（American toads, *Bufo amerianus*）の研究結果を報告している。USA のメリーランド州にある 29 の湿地で、全調査日数は 2000 年 3 月 9 日から 5 月 30 日ま 82 日であった。Table8.2 はデータの制限バージョンで、14 日ごとでまとめた５つのデータと最後の 12 日でまとめた観察結果である（＃観測回数に関わらず、1 回でもありなら占有としている=Y_{ij}）。

これらデータの解析に挑戦してみると、結果は選択（chois＃調査時期のことか？）をモデル化することに鋭敏であるようだった。モデル{p_t,ψ}のもとで最尤推定値と asymptotic standard errors(ASE、＃n が十分大きければ（∞）、最尤推定値の分布は正規分布で近似できるとして、そこから推定した標準誤差)は Table 8.3 にある。注意として一つの推定値はパラメータ空間のエッジ（縁）である（p_2^=1）。そして、信頼区間 CI=MLE±1.96ASE は 0 よりしたへ伸びている（p_1, p_5, p_6）；p_3 の 95%CI はパラメータが取りうる範囲全体に広がっている。少なくとも Tab 8.3 で示されている PLI(#すいませんよくわかりません)はパラメータのレンジと矛盾ない良い点がある。しかし、これらの coverage rate は漸近的な近似をあてにして

いる。

検出確率と関連したこの制約的な精度を扱う一つのアプローチは p_t の時間的変動(調査時期による変動)を含まないモデル$\{p,\psi\}$をフィットすることである。そのとき、ψの推定値は 0.43（SE=0.12）まで増加する。$p_1=p_3=\cdots=p_6$ のもとで p_2 が 1 と推定されるなら、ψの MLE は変化せず、小数点以下 3 桁まで正確で、0.352 である。

Mackenzie et al.(2002)が示したように、より良い方向性は関連する共変量の関数として検出確率を記載することである。Site- や site*sample-特有の共変量はアメリカヒキガエル研究で集められた。サイト特有の共変量はサイト占有に関わる要因を説明し、この研究の焦点でもあった。Site*sample 共変量は検出可能性へフィットするより複雑なモデルを可能にし、おそらく厄介な（あまり興味ない）パラメータのより良く扱えるだろう。

我々の考えでは、これらすべての解析はベイズ理論の文脈で最も効果的に実行され、疑わしい漸近的な結果に依存する必要はない。仮に、モデルパラメータに使う事前分布（prior）について議論があったとしても、少なくとも選択でき、数学的に導かれる事後分布が存在する。事後分布を評価するために MCMC を使わなくてはならないときでも、結果はより長い chain でサンプルすることで、任意の精度が得られる。さらに、モデルを拡張するのに必要な BUGS コードの修正は簡単である。

モデル$\{p_t,\psi\}$のベイズ推測はこのセクションの始めで書いたギブスサンプラーを使って実行される。しかしながら、占有問題の BUGS コードは驚くほどシンプルである（Panel 8.5）。データの記載（data statement 含まれていない）は Y_{ij} の行列、そして X_i は含める必要がない。このモデルは Xi=1 を必要とするが、X_i が 1 に等しいことは観測 Y_{ij}=1 に由来しているからである。分からない Xi は他のすべての未知量として同じように扱われ、事後分布からサンプルして終わる。サイトが調べられていないときに（＃つまり NA）、同じことが Y_{ij} に関して生じる：私たちは調査が行われたときに、動物は検出されるかどうかを予測できるだろう。

また注意として、私たちの興味は**私たちがサンプルしたこれらの間での**占有サイト数についてであるかもしれない。仮想的な全体の占有確率ψよりも。モデル$\{p_t,\psi\}$での MLE（最尤推定値）は 35.2%であり、ヒキガエルが検出されたサイトの割合とそれほどかけ離れていない。経験の少ない分析者はこの推定値を見つける詳細な検討を冷やかすかもしれない。注意してほしいのは、カエルは 10 サイトで検出され、10/29=34.5%である。モデリング過程で最も魅力的なプロダクツの 1 つがベイズ理論の枠組みの下で最も容易に得られる：首尾一貫して疑わしいような数学的近似なしに、29 サイトに確率を割り当てることができるのだ。状態が分からない 19 サイト中 18 に対してこの確率（＃占有確率ψ）は 0.9~2.7%であった。そしてサイト 16 でそれは 19.4%だった（9. サイト 16 は 3 回しか調査していなく、検出確率が高かったときの 2 回目に調査していないのに。＃検出確率が高い調査時期で 0 とでていると占有していない確率が高い。）。また、非検出（調査したけど 0）占有サイト数は 0,1,2 でその確率はそれぞれ 0.671,0.252,0.052 だったということもできる。そして合計が 3,それ以上はたった 2.5%であった。

8.3.2 More Complex Occupancy Models

生物学的観点からすると、占有に関する研究は CDL の$[\boldsymbol{X}|\boldsymbol{\beta},\boldsymbol{Z}]$に焦点があてられる。生物学的に無関係なサンプリングの性質と興味のあるモデルの性質を区別するには Fig 8.5 で示したようなグラフィカルな表現が助けとなる。

ここまで X_i はスカラー値をもった確率変数である占有モデルを考えてきた。多くの豊富なモデルクラスが考えられる。たとえば、次のように仮定することもできる。X_i は 2 次元配列で、1 つは種でもう一つは時間（あるいは季節）と。これまで考えてきたモデルでは調査期間を通じて、占有状況は変化しないと仮定している。それは X_i を 1 species * 1 time としていること。

X_i が 1 species * k times の配列の場合には MacKenz*i*e et al.（2002）の単一種で多シーズンモデルが導かれ、これは対象種によるサイト i の占有が季節的に変化する場合である。モデルのグラフィカル表示でいうと、わずかな変化は beta の定義で、現在絶滅と移入のモデルに関する要因を含んでいる。もうひとつは X で、現在は各サイトでの観測をベクトル値で含めている。ほかの拡張としては X_i を s species * 1 time 配列とおき、多種単一時期モデルとして、サイト間での種間関係を記載した要因をベータに含めることである。もっとも一般化されたケースは X_i が s species * k times の配列であり、多種多時期モデルが導かれる。

多種多時期モデルにおいて、X_i は（$\#X_i$ がある、ないの 2 状態だから）2*s*k のクロス分類で表すことができ、これは特に共変量がクロス分類に影響しているような場合に、多数のパラメータをもつ可能性がある。加えて、データ改変処理したモデリングは$[\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\gamma},\boldsymbol{W}]$に関して多くのモデルとパラメータを導いてしまい、$\gamma$ はやっかいなパラメータとなる。挑戦として研究上の仮説の観点から興味ある本質的な特徴をつかむモデルセットを識別することである。手におえない多くのモデルセットを導くやっかいな要素なしに。

占有モデルのような状況は多くのパラメータとパラメータセットがサンプリングの厄介な側面を描写している mark-recapture models を反映している。占有モデルに関して、クロス集計したデータの解析に対数線形モデル（loglinear modeling method）を適用する可能性について知っており、クロス集計されたデータでは、種間と季節間の制約が比較的複雑な関係を持つモデルに必要とされるパラメータ数を減少させてくれる。我々は占有の動態や種間関係の記載に関して対数線形モデルを使うこと、そして、モデルの厄介な側面の次元縮小に関する戦略として、厄介なパラメータに関する階層モデルの使用を提唱する。

8.4 DISTANCE SAMPLING

占有モデルでは、検出確率の推測がサイトごとの繰り返し調査することで行えた。この点について、Chapter 9 で議論する mark-recapture models と近い関係がある。ここでは、我々の関心を distance sampling へ向け、検出確率はもっぱら異なった平均によって推定する。

Distance sampling models は一つのとても特異的な共変量の重要性を強調する：調査している対象（object）間の距離と相対的な位置である。ラインとランセクトでは、距離はトランセクトのセンターラインから測定され、円形プロットサンプリングや trapping webs ではポイントセンターから測定される。検出確率

と距離との間に関数的関係の仮定をおき、さらに対象の空間分布の仮定をおくことで、これらのモデルは欠損値を含んだ対象の数について推測が行えるようになる。

“distance sampling”の名前に秘められているのはモデル化される基礎データは距離の集合 X という考え方である。n 個の距離の集合（集まり）はすでに知られている確率分布 $\pi(X)$からのサンプルとみなされ、典型的には（その必要はないが）一様分布（uniform distribution）である。n はわからない状態で、すべての距離が測定されているとは限らない。目的は n について、あるいは n が部分集合である個体群サイズ N についての推測を行うことである。

距離 x_i は検出確率 $g_\theta(x_i)$と関係していると仮定することで推測が可能となる。$g_\theta(\cdot)$は未知のパラメータ θ を除いて、既知の関数である。ベクトル $\boldsymbol{x}$ は$(\boldsymbol{x}^{obs}, \boldsymbol{x}^{mis})'$に分割され、慣習的なモデリングは観測された一部の $\boldsymbol{x}^{obs}$ に基づいて計算された確率分布を基礎とする。それなら最初は観測された距離で分布計算してみる。つまり、

$$f(x \mid \text{Object is detected}) = \frac{\Pr(\text{Object is detected} \mid \text{Distance} = x)\pi(x)}{\Pr(\text{Object is detected})} = \frac{g_\theta(x)\pi(x)}{\int g_\theta(x)\pi(x)dx} \qquad (8.6)$$

推測は ODL を基礎として、

$$L(\theta; x^{obs}) \propto \prod f(x_i^{obs} \mid \text{Object is detected}).$$

式（8.6）は見覚えがあるはず：ベイズの定理（Bayes theorem）を経て$[x|D=1]$と計算したもので、距離 x が検出 D と関係して、距離の事前分布（”prior”） $\pi(x)$としている。そしてベルヌーイ型のデータ分布は

$$\Pr(D \mid x) = g_\theta(x)^D(1 - g_\theta(x))^{(1-D)}。$$

もっと自然なアプローチは line-transect sampling に関する階層モデルを考えることであると信じており、データがモデル化されるように検出（*detection*）を扱い、そして共変量として距離 distance を扱う。Distance sampling はまさに対象のサンプリングであり、距離は本質的な生物学的関心ではない。距離のたった一つの役割はモデルの厄介な側面の扱いをヘルプすることで、すべての対象を検出できない私たちの無力さを記述してくれる。“modeling the data You wish You had”の精神からスタートしたのと同じように、対象中心のアプローチ（the object-centerd approach）は distance sampling を close-population mark-recapture modeling や section 8.3 の占有モデルのような階層的な枠組みにする。

A Line-Transect Survey（ラインとランセクト調査）

このセクションではラインとンセクト調査を扱うが、ここで議論される考えは、容易にほかの形の distance sampling へ一般化される。ここではサイズ A の範囲にランダムに幅 $2w$、長さ L のとランセクト 1 つを設置した状況を考える。ラインとランセクト（”コドラート”、方形枠）は面積 $a = 2wL$。対象(object)として”動

物”を想定するが、ラインとランセクトサンプリングはさまざまな目的に適用できる（10.方法論的評価は木杭、ヒューロン湖（#五大湖の１つ、2 番目に大きい）の煉瓦、ポリスチレン製のカメを対象にした調査がある。またフィールドワークに参加する学生への巧妙な刺激としてはビール缶。）ここでは未知のサイズ N の個体群を考える。

野外調査で動物が記載されない場合に２つのことが想定される。これらは階層モデルの２つのパートで定義される。第 1 に、サンプルされるコドラート内に動物がいないかもしれない。ここでコドラート内の個体数を n としよう；単にコドラート内に彼らがいないため N-n 匹の動物は記載されない。第 2 に、コドラート内の n 匹すべてが記録されるわけではなく、よくあるケースである。”the data we wish we had”の点から野外調査を記述するため、CDL を使って、n と未知の N を関連付け、コドラート内で得られるデータを未知の n と関連付ける。

あらゆる位置で 1 匹を見つける確率が個体群が生息するエリア内(population area)のすべての位置で等しく、かつ、ある 1 個体の存在が他個体に影響しないならば（＃ランダム分布）、範囲 A 内の個体群サイズ N から面積 a のコドラートに出現する個体数 n は二項分布型の確率変数（binomial random variable）でモデル化できる（Seber, 1987,22）：

$$[n \mid N] = \binom{N}{n} p_c^n (1 - p_c)^{N-n} \qquad (8.7)$$

ここで $p_c = a/A$.

コドラートによって定義された範囲内での存在という条件のもとでは、動物の検出は独立なベルヌーイ型の確率変数（Bernoulli random variables）であり、センターラインからの距離によって決められる成功パラメータをもつ。距離 x_i にいる動物 i に関して、このベルヌーイ試行は D_i でラベルされ、その成功パラメータは

$$g_\theta(x_i) = \Pr(\text{Detection} \mid \text{Distance} = x_i)$$

となり、θは検出プロセスに支配される未知のパラメータを表す。$\boldsymbol{D} = (D_1, D_2, \ldots, D_n)'$、$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$とする。

個体群サイズを推測するために $g_\theta(0)$は既知であることを仮定しなければならない。たいていセンターライン上にいるすべての動物は検出されるから、$g_\theta(0) = 1$ とする。

次のようにコドラート内の動物データをモデル化する

$$[\mathbf{D}, \mathbf{x} \mid n, \theta, \gamma] = [\mathbf{D} \mid \mathbf{x}, n, \theta][\mathbf{x} \mid n, \gamma] \qquad (8.8)$$

ここで

$$[\mathbf{D} \mid \mathbf{x}, n, \theta] = \binom{n}{m} \prod_{i=1}^{n} g_\theta(x_i)^{D_i} (1 - g_\theta(x_i))^{1-D_i} \quad 。$$

ここで、m は検出される動物個体数であり、$[\mathbf{x} \mid n,\gamma] = \prod_{i=1}^{n} [x_i \mid \gamma]$ は動物配置の同時確率に関するモデルを記載している。x のモデルは必要であり、なぜならコドラート内で記載されなかった動物に関して距離は測定されていないためである。上記における注意として、一般的な仮定として$[x_i|n,\gamma]$は一様分布$U(0,w)$の密度関数であるとしている、任意のほかのモデルも考えられるけど。

2 つの構成要素（componet）を一緒に置くことで、階層ラインとランセクトモデルの CDL を得る

$$[\mathbf{D},\mathbf{x},n \mid N,\theta,\gamma] = [\mathbf{D} \mid \mathbf{x},\theta,n] \times [\mathbf{x} \mid n,\gamma] \times [n \mid N] \qquad (8.9)$$

Fig. 8.6 に DAG を表示した。$[n|N]$と対応する右中央の DAG の部分は想定しているデータのモデルであり、コドラート内の個体数である。もし多数の調査をしたら、この部分は higher-level modeling へ拡張でき、たとえば時間あるいは空間にわたった個体数(#の推測?）を容易にする。Fig.8.6 のほかの構成要素はモデルの厄介な側面をモデル化しており、それ自体興味ないけど、確かな推測には本質的なものである。

CDL はベイズ理論による解析で容易に扱われる。頻度論に立った解析は ODL を使う。ODL は記載されていない動物の構成要素（missing components）x^{mis} を除いて統合した CDL から得られる。距離の密度を表すπ(x)を持つ式（8.9）の式（8.7）と（8.8）を取り換えて、x^{min} を統合して（？）、

$$\binom{N}{m}(1 - p_c E_x[g_\theta(x)])^{N-m} \prod_{i=1}^{m} g_\theta(x_i^{obs})\pi(x_i^{obs})$$

をえる。この尤度は Borchers et al.(2002)により与えられた。

Example

Buckland et al.(1993)(section 8.3)は 150 本の木杭のラインとランセクト研究からデータ解析を議論している。対象エリア内に 1 つのトランセクトを設置、これを 11 人の観察者で繰り返しサンプルした。ここでは、最初の観察者のデータを解析する。検出関数として半正規分布(half-normal model) $p(x,\sigma^2)=\exp(-x^2/2\sigma^2)$をつかい、$0 < X < w = 20$ はトランセクトの幅。N とσ^2の完全条件付分布は自らのソフトでコードすることで容易に得られる。CDL をフィットすることは BUGS を使うことで簡単である。そのコードは Panel 8.6 と 8.7 にある。

Panel 8.6 を使って、トランセクト内で出現する杭の数 n の予測から始める。Panel 8.6 の興味深い特徴は n をモデル化するためのデータ拡大（data augmentation）の使用である。n は未知。そのためランダムとして扱わなければならない。そのため、未観測量の数は確率変数である。たとえば、n=m=172 では未観測の距離 x_i はない。しかし n=172 では、事後分布からサンプリングしなければならない 100 点の未観測な x_i がある。この状況は reversible jump MCMC を使った multimodel の実装と似ている。パラメータ次元のパレットが必要となり、少なくとも我々のモデルセットの中のモデルの最も高い次元と等しいとする(section 7.3.2)。

n について離散的一様分布(discreate uniform) $DU(\{1,2,\ldots,200\})$の事前分布を選択し、[n|

ψ]=*B*(200,ψ), [ψ]=*U*(0,1)でモデルして実行する。結果は｛0,1,…,200｝の *n* に関する事前離散的一様分布（discrete uniform prior）であり、示しやすい。Panel 8.6 のデータの記述は *I*[*i*]と *x*[*i*]について、128（＝200-*m*）の欠損値を含み、128 個の 0 が D_i と対応する。

コドラート内に 200 の杭を記載するとした解析と考えられ、現実的なようだし、一方で架空的。*I*[*i*]は杭があるかどうかの指標であり、*i* 番目のモデル化された杭は実際のものである。*I*[1]=*I*[2]=…=*I*[72]=1 とし、観察された杭と対応する。注意としては *D*[73]=…*D*[200]=0 であり、これらのいくつかは検出されていない実際の杭と対応する。同時に検出されえない架空の杭とも対応する。ノード n<-sum(*I*[1:200])はコドラート内の"実際の杭"の合計であり、推定したいと思っていたもの。

もし n がｍよりずっと大きいとき、それは BUGS コード内で、多くの"架空の"ノードを含む必要があるだろう。そして run time は十分に増やすのがいいだろう。良い方針は n についてさまざまな上限 M を試し、どれぐらい大きな値がデータによって支持されうるかを見ることである。今のケースではもともとM=1500にセットしたが、長さ100,000回の１マルコフ連鎖でｎの最大観察値は198であり、そしてサンプルされた値の0.5%は 167 を決して越えなかった。なので、ｎの事前分布を DU({1,2,…,1500})から DU({1,2,…200})へ変更して、ほどよくなったように見える。この選択は推測に影響しないが、おおよそ 90% run time が減少した。

Panel 8.1 や Panel 8.5(カエルデータ）のように、Panel 8.6 の BUGS コードは容易に別のモデルへ修正される。たとえば、検出を説明する共変量を含むには p[*i*]の定義を単に修正する必要がある。だけれども、注意として、未観測値があるため、共変量の分布に関するモデルが必要となる。BUGS コードは多数の調査データへ容易に修正されることや、またコドラートを越えて拡張することで既知の範囲の個体群サイズ N の予測することもできる。

Panel 8.7 は Panel 8.6 と比べ２つの違いを組み入れている。１つめは.obs と.mis への拡張によってノードを区別した。観測値に関するループでデータ値の特定を行い、データ記述にたくさんの *NA* を書くことを避けた。もう１つの違いは２つの指示変数（indicator variable）を特定したことであり、サイズ *N* の個体群の算入では *I.N* を、コドラートにおける個体数では *I.n* とした。この解析に関して、コドラートが範囲 *A* の８０%をカバーしていると仮定した。つまり Pc=a/A=0.80。*x, I, D* は２つの for ループで区別されているけど、これらの分布を支配するパラメータはそうでない。つまり両パート内で同じ psi と tau である。また、注意してほしいのは観察された *D* はすべて１であり、観察されていない *D* はすべて０であり、観察された *I* は１、観察されていない *I* は０あるいは１である。