データ統合による外来生物の侵入リスク推定

本間　翔太　　総合研究大学院大学 統計科学コース 博士後期課程 2年

1. 概要

自然分布域から人為的に運ばれた先において，生物多様性や人の活動の脅威となる生物は，侵略的外来生物と呼ばれる．世界的に侵略的外来生物の拡散が加速しており，国際物流はその主な要因の一つと報告されている．日本の国際物流のうち，海上物流貨物は99%以上を占めているため，港湾は外来生物の侵入口である．特に外来生物は，既存の生態系が乱された人工環境において定着しやすいとも考えられており，港湾は外来生物侵入のホットスポットであると言えるだろう．

近年では，南米を原産とする陸生の外来生物であるヒアリ(*Solenopsis invicta*)やアカカミアリ(*Solenopsis geminata*)が，コンテナを経由して侵入していることが確認されている．この様な脅威に対して，継続的なモニタリングと発見された際の速やかな駆除が続けられているが，様々な制約の中で，こうした対策を継続していくためには，リスクの高い場所を特定し，適切に対策のエフォートを割り振ることが求められる．統計モデルを用いたリスク推定は，侵入管理策を計画する際に重要である．

しかし，外来生物の拡散現象は複雑であり，その要因の一つはコンテナが世界中を循環することであろう．また，調査データが十分に得られないことも正確なリスク評価を難しくさせている．こうした中で，空間統計モデルは，コンテナ物流の複雑さによる不確実性を捉える可能性がある．また，得られる周辺データを用いることで，少ないデータを最大限に活用できる可能性があり，データ統合（data integration）と呼ばれる手法は，近年，疫学や生態学分野における実データを用いた応用において，その有効性が実証されてきた(Miller et al., 2019)．空間構造の相関を用いたデータ統合の方法について，解析例とともに紹介する．

2. データ

ヒアリが国内で初めて発見された2017年以降，国際コンテナ航路を持つ港湾68港湾において，環境省および国土交通省によりトラップ等を用いたヒアリ類を対象としたモニタリング調査が行われている．このような調査データ（survey data）は，常時取得されるものではなく，年に2回程度の実施に留まっており，観測データは多くない．一方で，環境省は，コンテナヤードや倉庫等，様々な場所において，ヒアリおよびアカカミアリと専門家により同定されたデータ（presence-only data）を公開している．このようなデータは，survey dataを用いたリスク推定を改善するだろうか．

3. モデル

こうした問いに対して，階層ベイズモデルを用いた検証を行う．特定の種に着目する場合について記載する．

階層ベイズモデル

調査データ（survey data, $Y_{i,1}$ ）と報告データ（presence-only data, $Y_{i,2}$）は，以下のようにモデル化された：

$$Y_{i,1} \sim \text{Binomial}(n_i, p_i) \text{ for survey data (detect / not detect)},$$
$$Y_{i,2} \sim \text{Poisson}(\lambda_i) \text{ for presence-only data.}$$

ここで，$i = 1, \ldots, 68$で，n_iは調査実施回数，p_iは出現確率，λ_i は観測の期待発生数を表す．さらに，それぞれのパラメータは，linear predictor η_iとリンク関数（logit linkおよびlog link function）を用いて以下の様にモデル化された:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \eta_{i,1} = \beta_{0,1} + \sum_{m=1}^{M_1} \beta_{m,1} x_{i,m,1} + \theta_{i,1}$$
$$\log(\lambda_i) = \eta_{i,2} = \beta_{0,2} + \sum_{m=1}^{M_2} \beta_{m,2} x_{i,m,2} + \theta_{i,2}$$

ここで，M_1, M_2は説明変数の数，$x_{i,m,*}$ は説明変数，$\beta_{0,*}$は切片，$\beta_{m,*}$は説名変数の効果を表す．$\sum_{m=1}^{M_*} \beta_{m,*} x_{i,m,*}$の項は，固定効果と呼ばれる．コンテナ輸入量はこの現象を駆動する要因であり，固定効果としてモデルに組み込まれる．$\theta_{i,1}, \theta_{i,2}$はランダム効果と呼ばれ，空間構造およびデータセット間の相関を表すようにモデル化された．

空間モデルによる拡散のモデル化

コンテナ物流を介して侵入する外来生物は，貨物が輸入された港から，さらに国内各地へ運ばれる．そのため，輸入量の小さい港湾であっても，輸入量の大きな港湾と密接に結ばれている港湾では，リスクが高い可能性あり，その構造を考慮することは重要だろう．

この効果は，intrinsic Conditional Autoregressive (iCAR) モデルを用いて以下のようにモデル化された：

$$\theta_{i|-i} \sim \text{Normal}\left(\frac{\sum_{j=1}^{N} w_{ij}\,\theta_j}{\sum_{j=1}^{N} w_{ij}}, \frac{1}{\sum_{j=1}^{N} w_{ij}} \frac{1}{\tau}\right).$$

ここで，$\theta_{i|-i}$は，i以外のθで条件づけられた変数θ_iを意味し，また w_{ij}は港間の物流による結び付きの強さを表す行列$\boldsymbol{W}$の要素であり，港iとjでやり取りがある場合にはその量，結び付かない場合には0である．したがって，$\boldsymbol{W}$は，対角要素0の対象行列となる（Fig. 1）．つまり，θ_iの期待値は，物流のつながりの強さによる周囲の変数θ_{-i}の重み付き平均となる．τは変数$\boldsymbol{\theta}$の精度パラメータであり，分散の逆数である．

iCARモデルは，多変量正規分布を用いて以下で表される；

$$\boldsymbol{\theta} \sim \text{Normal}\left(\mathbf{0}, \frac{1}{\tau}(\mathbf{D} - \mathbf{W})^{-1}\right).$$

ここで，$\boldsymbol{D}$は対角要素に隣接要素数を持つ対角行列であり，$\boldsymbol{D}=diag(\sum_{j=1}^{N} w_{ij})$である．

このように，ある確率変数が正規分布で仮定され，その変数が関連すると仮定した変数のみに依存する場合は，Gauss Markov Random Field (GMRF)と呼ばる．iCARは，生態学において，空間相関構造をモデル化する際に従来から用いられてきた．しかし，このモデルのより一般的な概念はネットワークであり，侵入生態学における人的な拡散構造を表現するためのCARの適用は，空間モデルのより拡張的な適用と解釈できる．

潜在的な空間構造を考慮したデータ統合のモデル化

データの持つ空間構造を考慮しつつ，各データの潜在変数間の相関を推定するため，$\boldsymbol{\Theta} = (\boldsymbol{\theta_{i,1}}, \boldsymbol{\theta_{i,2}})$は，MCAR（Multivariate Conditional Autoregressive）モデルを用いて以下のようにモデル化された：

$$vec(\boldsymbol{\Theta}) = \begin{pmatrix} \theta_{11} \\ \vdots \\ \theta_{n1} \\ \theta_{12} \\ \vdots \\ \theta_{n2} \end{pmatrix} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Lambda}^{-1} \otimes (\boldsymbol{D} - \boldsymbol{W})^{-1}).$$

ここで，$vec(\cdot)$は，行列の列ベクトルへの変換の操作を表し，$\otimes$はクロネッカー積を表す．$\boldsymbol{\Lambda}^{-1}$はデータセット間の分散共分散行列であり以下でモデル化された：

$$\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} 1/\tau_1 & \rho/\sqrt{\tau_1\tau_2} \\ \rho/\sqrt{\tau_1\tau_2} & 1/\tau_2 \end{bmatrix}.$$

ここでτ_1, τ_2は各データセットの精度パラメータで，$\boldsymbol{\Lambda}^{-1}$の対角要素は分散を表している．$\rho$は相関係数であり，非対角要素は共分散である．2つのランダム変数は，クロネッカー積により結合され，データセット同士が持つ構造の相関が表現される(Fig. 2)．

パラメータ推定

推定には，Integrated Nested Laplace Approximations（INLA）を用いた (Rue et al., 2009)．階層ベイズモデルのパラメータ推定については，MCMCがよく用いられるが，潜在変数がガウス分布で仮定された場合の推定において，INLAはMCMCと同等の近似誤差で，高速に事後分布の近似を得ることが報告されている．INLAはRで計算可能なインターフェースが公開されており，高度で柔軟なモデリングが可能である．

4. 計算例

データ統合モデル（Joint）のDICおよびWAICを示す（Tab. 1）．コンテナ物流による港の繋がりを仮定したケースを示している（MCAR）．また，推定された各港のリスクを描画した（Fig. 3）．個別に推定されたモデルと比較することで，データの有効活用に繋げられるだろう．今後は，様々なモデルとの比較を実施する予定である．

Figure 1: Connectivity matrix ***W*** for domestic container logistics

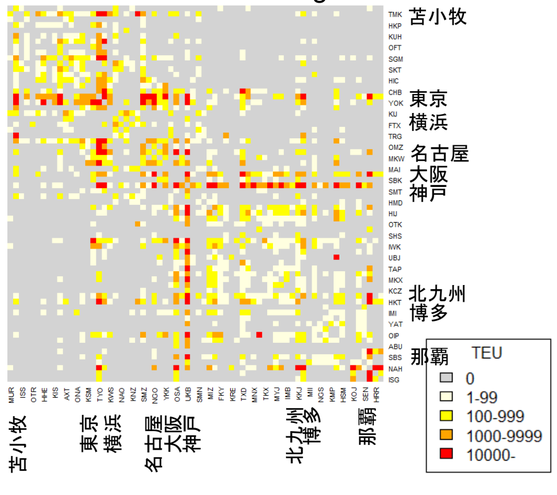

Figure 2: Graphical image for data integration

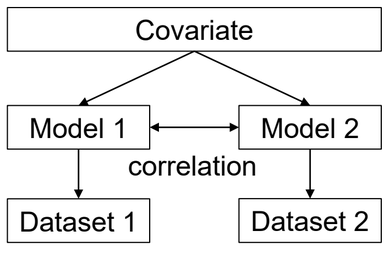

Figure 3: Posterior mean of p_i by the best model for *S. Invicta*.

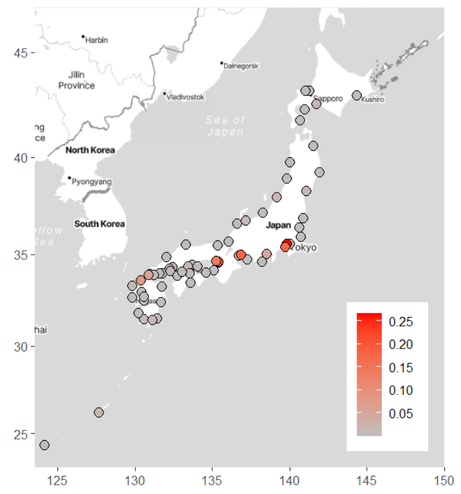

Table 1: 計算結果

	Model type	*S. invicta*		*S. geminata*	
		Survey	Presence-only	Survey	Presence-only
DIC	Joint (MCAR)	60.18	67.19	70.69	82.17
WAIC	Joint (MCAR)	61.32	68.75	71.34	85.52

5. 参考文献

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution / British Ecological Society*, *10*(1), 22–37.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *71*(2), 319–392.