Mixed effects modeling of clustered extreme values

鹿児島大学大学院　桃木 光輝　　鹿児島大学大学院　吉田 拓真

1　はじめに

気象データから災害リスクを定量化することは，災害対策計画や都市開発の観点から重要である．言い換えると，それは災害を引き起こすような極端な値の発生確率を予測することに相当する．しかしながら，稀な事象を対象とするこの分野では，利用可能なデータが決して多くないという問題がある．極値統計学では，数学的に保証されたパラメトリックな分布を活用することで，このような状況下でもより精度の高い予測を実現する．一方，気象データの特徴は複数の気象観測所でデータが蓄積されている点である．本研究で着目する混合効果モデルは，全ての観測所のデータを一つのモデリングに統合してデータ全体の共通情報を引き出しつつ，各観測所のデータの分布の差異を効率的・効果的に予測することができる (Sugasawa and Kubokawa 2020)．そのため，混合効果モデルは，利用可能なデータが不足しがちな極値統計学との相性が非常に良い．本研究では，混合効果モデルを応用した新たな極値統計モデルを提案し，予測の不確実性を測るための漸近理論を確立する．

極値統計学において，最も重要視されているのは，極値指数（extreme value index）と呼ばれる，分布の裾の重さに関連するパラメータの予測である．先行研究には，ベイズ階層モデルを応用した手法があるが，実データ解析において極値指数の予測結果の大きな不確実性が指摘されている (Dyrrdal et al. 2015)．本研究では，アプリケーションの範囲をリスクの予測が難しいとされるものに限定することで，極値指数のより詳細な解析が可能な手法を実現する．

また，その他の手法には，多変量極値統計学やコピュラなどがある（Davison et al. 2012）．これらは全ての観測所のデータの同時分布をパラメトリックに解析する手法であり，主に，観測所間のリスクの依存構造に焦点を当てている．しかしながら，多変量極値統計学の場合，多くの観測所のデータを同時にモデリングすることが困難である（Huster and Wadsworth 2020）．また，災害は複雑な要因により引き起こされると考えられるが，これらの手法は位置情報以外の共変量情報をモデリングに組み込むことが困難である．本研究では，災害リスクの地理的な依存構造よりも，災害の要因を明らかにするためのモデリングの開発に専念する．

2　モデル

本研究はクラスターデータ（clustered data）

$$\left\{(Y_{ij}, \boldsymbol{X}_{ij}) \in \mathbb{R}^{+} \times \mathbb{R}^{p},\ i = 1, 2, \cdots, n_j,\ j = 1, 2, \cdots, J\right\}$$

を対象とする．気象データはその一例である．J がクラスターの個数，n_j が各クラスター内のデータ数であり，$(Y_{ij}, \boldsymbol{X}_{ij})$ は j 番目のクラスターにおける i 番目の観測である．ただし，Y_{ij} は目的変数であり，$\boldsymbol{X}_{ij}$ は説明変数である．

U_j を未知の分散 σ_0^2 を持つ正規分布 $N(0, \sigma_0^2)$ に従う未観測の確率変数とする．このとき，$\boldsymbol{X}_{ij} = \boldsymbol{x}$ と $U_j = u_j$ が与えられた下での Y_{ij} の条件付き分布関数 $F(y|\boldsymbol{x}, u_j) = P(Y_{ij} \leq y|\boldsymbol{X}_{ij} = \boldsymbol{x}, U_j = u_j)$ に，パレート型分布（Pareto-type distribution）

$$1 - F(y|\boldsymbol{x}, u_j) = y^{-1/\gamma(\boldsymbol{x}, u_j)} \mathcal{L}(y, \boldsymbol{x}, u_j)$$

を仮定する．ここで，$\gamma(\boldsymbol{x}, u) > 0$ が極値指数であり，$\mathcal{L}(y, \boldsymbol{x}, u)$ は任意の $s > 0$ に対して $\lim_{y\to\infty} \mathcal{L}(ys, \boldsymbol{x}, u)/\mathcal{L}(y, \boldsymbol{x}, u) \to 1$ を満足するような関数である．この分布族には t 分布やパレート分布などの裾の重い分布が多く含まれる．分布の裾における挙動を決定する極値指数の予測が本質的に重要である．本研究では，極値指数に混合効果モデル

$$\log\left\{\gamma(\boldsymbol{x}, u_j)^{-1}\right\} = \alpha_0 + \boldsymbol{\beta}_0^{\top}\boldsymbol{x} + u_j, \quad j = 1, 2, \cdots, J$$

を仮定する．ここで，$\alpha_0 \in \mathbb{R}$ と $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ は未知の回帰係数である．$U_j \equiv 0$ の古典的な線形モデルは Wang and Tsai (2009) で研究され，提案モデルはその拡張となっている．α_0 と $\boldsymbol{\beta}_0$ は全クラスター共通のパラメータである．$U_j = u_j$ は変量効果と呼ばれ，これにより説明変数 $\boldsymbol{X}_{ij} = \boldsymbol{x}$ では説明のつかない影響，すなわち，クラスター毎の分布の違いを考慮できる．また，その実現値そのものを予測することも可能である．

講演では，研究を通して得られた提案モデルの数学的性質と数値パフォーマンスについて報告する．

参考文献

[1] Davison, A.C., Padoan, S.A., and Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical Science*. **27** 161–186.

[2] Dyrrdal, A.V., Lenkoski, A., Thorarinsdottir, T.L., and Stordal, F. (2015) Bayesian hierarchical modeling of extreme hourly precipitation in Norway. *Environmetrics*, **26** 89–106.

[3] Sugasawa, S., and Kubokawa, T. (2020) Small area estimation with mixed models: a review. *Japanese Journal of Statistics and Data Science*, **3** 693–720.

[4] Wang, H., and Tsai, C. L. (2009). Tail index regression. *Journal of the American Statistical Association*. **104** 1233–1240.