Calibration d'un modèle de pollinisation à l'échelle du paysage par des méthodes de type Approximate Bayesian Computation

Charlotte Baey [1] & Henrik G. Smith [2,3] & Maj Rundlöf [2] & Ola Olsson [2] & Yann Clough [2] & Ullrika Sahlin [2]

[1] *Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France*
charlotte.baey@univ-lille.fr

[2] *Lund University, Department of Biology, SE-223 62 Lund, Sweden*
[3] *Lund University, Centre for Environmental and Climate Science, SE-223 62 Lund, Sweden*

Résumé. La modélisation des services écosystémiques passe souvent par la construction de modèles mécanistes parfois complexes, dont la calibration peut s'avérer délicate. Dans ce travail, on s'intéresse à un modèle de pollinisation spatialement explicite, appliqué au bourdon terrestre (*Bombus terrestris*). La vraisemblance n'étant pas calculable analytiquement, nous proposons une approche de type Approximate Bayesian Computation (ABC) pour l'estimation des paramètres du modèle. Nous comparons différentes méthodes ABC permettant de prendre en compte la grande dimension des observations. La première étape consiste à définir un ensemble de statistiques résumées sur laquelle les méthodes ABC seront appliquées. Nous considérons ensuite deux stratégies, l'une reposant sur des méthodes de régression pour ajuster les échantillons obtenus selon la loi a posteriori ABC, et l'autre reposant sur des méthodes de type machine learning pour approcher certaines caractéristiques de la loi a posteriori ABC (e.g. moyenne et quantiles). Les résultats obtenus sur données simulées montrent que certains paramètres sont plus faciles à estimer que d'autres, et les approches basées sur des forêts aléatoires s'avèrent plus performantes. L'application aux données réelles montre l'intérêt de l'approche ABC dans le contexte de modèles complexes tout en mettant en évidence les difficultés liées au choix des statistiques résumées et de la méthode.

Mots-clés. Statistique bayésienne, méthode ABC, calibration, modèle de pollinisation

Abstract. Modelling ecosystem services often requires building mechanistic models which might be complex and which calibration can be challenging. In this work, we are interested in a spatially explicit foraging model for *Bombus terrestris*, accounting to bee distribution in the landscape. The likelihood of the model being intractable, we rely on Approximate Bayesian Computation (ABC) for the estimation of the model parameters. We compare different ABC methods to handle the high dimension of our observations. The first step consists in the definition of a set of summary statistics on which ABC is then applied. Two ABC strategies were then studied. The first one rely on the use of regression adjustment methods, to produce ABC posterior samples. The second one rely on the use of machine learning approaches to approximate key quantities of the ABC posterior distribution (e;g. mean and quantiles). Results from simulated data show that some parameters are easier to calibrate than others, and that approaches based on random forests performed better.

Results on real data show how appealing the methodology can be, even though tuning ABC can be challenging, especially regarding the choice of the summary statistics.

Keywords. Bayesian statistics, ABC method, calibration, pollination model

1 Introduction

L'étude et l'évaluation des services écosystémiques, que ce soit d'un point de vue écologique ou socio-économique, est un enjeu majeur des politiques publiques. Parmi ces services écosystémiques, on retrouve notamment la pollinisation par les insectes, qui permet entre autres de maintenir la biodiversité des plantes sauvages ainsi que la production d'un grand nombre de plantes cultivées. Dans un contexte de déclin des populations d'insectes pollinisateurs, il est alors crucial de pouvoir estimer avec le plus de précision possible le statut de ces populations, mais aussi de pouvoir prédire l'effet de certaines pratiques agricoles ou paysagères sur ces populations. Ces questions peuvent être explorées à l'aide de modèles écologiques, et dans le cas qui nous intéresse, de modèles spatialement explicites prenant en compte la distribution des insectes dans le paysage.

Les modèles développés dans ce cadre là sont des modèles dits mécanistes, qui sont souvent complexes et hautement non linéaires, produisant des sorties de grande dimension qui peuvent elle-mêmes posséder des structures complexes. La calibration des paramètres de ces modèles peut alors s'avérer délicate. Par exemple, le temps d'exécution d'une instance du modèle peut être élevé, rendant coûteuse toute procédure d'estimation nécessitant des appels répétés au modèle. De plus, ils s'écrivent souvent comme un ensemble de relations hiérarchiques faisant intervenir des variables latentes qui peuvent être également de grande dimension.

Dans ce travail, nous proposons une approche de type *Approximate Bayesian Computation* (ABC), qui permet de contourner les difficultés évoquées plus haut. Nous comparons différentes méthodes de type ABC, reposant sur deux types de stratégies pour la prise en compte de données en grande dimension. La première utilise des méthodes de régression pour ajuster les échantillons obtenus sous la loi a posteriori ABC, et la deuxième utilise des approches de type machine learning pour estimer des quantités clés de la distribution a posteriori ABC. Les méthodes sont comparées sur des données simulées puis appliquées sur un jeu de données réelles.

2 Modèle de pollinisation

2.1 Modèle mécaniste de type 'Central Place Forager'

Le modèle de pollinisation utilisé (Olsson et Bolin, 2014) est basé sur la théorie du 'central place foraging' qui permet de décrire le comportement d'animaux qui partent à la recherche de nourriture à partir d'un nid ou d'un habitat central. C'est notamment le cas du bourdon

terrestre. A partir de la localisation des nids de bourdons et des sources de nourriture dans le paysage, on peut alors déterminer les ressources qui seront visitées par les bourdons d'un nid donné en fonction de la qualité de la resource et de sa distance au nid.

Tout d'abord, le paysage est divisé en cellules carrées, auxquelles on associe un type d'habitat (ex. : champ de colza, forêt, zone urbaine, ...), puis une qualité de resource et une présence ou absence de nid en fonction du type d'habitat. Pour un bourdon dont le nid est situé sur la cellule i et une resource située sur la cellule j, on mesure la différence entre la distance maximale que l'insecte est prêt à parcourir pour une resource de cette qualité, et la distance réelle entre i et j par la quantité suivante : $\Delta_{ij} = \tau_0 \left(1 - \frac{f_0}{f_j}\right) - d_{ij}$, où τ_0 est la distance maximum qu'un bourdon peut parcourir, et f_0 la qualité minimale de resource exigée par l'insecte. Puis on définit la qualité d'un nid situé en i par : $s_i = \sum_j \Delta_{ij} \mathbf{1}_{\Delta_{ij}>0}$. Plus il y a de sites avec des ressources de bonne qualité à proximité de la cellule i, plus la qualité d'un nid situé en i sera élevée.

En pratique, un bourdon dont le nid est entouré de ressources de bonne qualité aura tendance à rester à proximité de son nid et à peu exploiter les sites situés loin de son nid. On peut alors définir la distance maximale qu'un bourdon est prêt à parcourir à partir de son nid, en fonction de la qualité de celui-ci :

$$\tau_i = \frac{\tau_0}{1 + \exp((\sqrt{s_i} - a)/b)}.$$

Comme pour la définition de Δ_{ij}, on peut définir Δ^*_{ij} en remplaçant τ_0 par τ_i :

$$\Delta^*_{ij} = \tau_i \left(1 - \frac{f_0}{f_j}\right) - d_{ij}.$$

L'intensité de visites de bourdons sur le site j est alors définie par :

$$\nu_j(\theta, \mathcal{M}) = \sum_{i=1}^{I} q_i \frac{\Delta^*_{ij}}{\sum_{j=1}^{J} \Delta^*_{ij}},$$

où q_i est une variable binaire indiquant la présence ou l'absence de nid sur la cellule i.

2.2 Données

On dispose de deux jeux de données provenant du sud de la Suède, sur la surveillance des insectes pollinisateurs dans différents types de paysage présentant un gradient de ressources florales. Les deux études recouvrent 4 années d'observation, et plusieurs observations ont été faites au cours de chaque année d'observation, correspondant à différentes périodes du cycle de vie du bourdon. Au total, on dispose d'un ensemble de 790 observations du nombre de bourdons, relevées dans des transects de longueur fixée, et dans différents types d'habitat : champ de colza, prairie (semi-naturelle), bordure de champ, champ de céréales.

2.3 Modèle statistique

On note alors $y_{ijk}, i = 1, \ldots, n, j = 1, \ldots, J, k = 1, \ldots, K$ les observations du nombre de bourdons sur le site i, l'année j et à la période k.

Vraisemblance. On note λ_{ijk} l'intensité réelle du taux de visites sur le site i, l'année j et la période k. Le modèle est alors décrit sous la forme hiérarchique suivante :

$$y_{ijk} \mid \lambda_{ijk}, \theta \sim \mathcal{P}(c_i \cdot \lambda_{ijk}),$$

où c_i est une constante connue permettant de prendre en compte la surface de la zone d'observation et la durée d'observation et θ représente l'ensemble des paramètres du modèle de pollinisation.

$$\log \lambda_{ijk} = \log \nu_i(\theta, \mathcal{M}_{ijk}) + \beta_1 + \sum_{l=2}^{K} \beta_l \mathbf{1}_{l=k} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2).$$

Les paramètres $\beta_l, l = 1, \ldots, K$ permettent de prendre en compte l'évolution de la taille de la population au cours de la saison.

On note $\psi = (\theta, \beta_1, \ldots, \beta_K, \sigma^2)$ l'ensemble des paramètres du modèle. Le modèle ainsi décrit conduit à une vraisemblance de type Poisson-lognormal, qui n'est pas calculable analytiquement.

Lois a priori. Des lois a prioris non informatives ont été choisies en l'absence d'information biologiques disponible sur les processus correspondants. Pour les paramètres τ_0 et f_0 dont l'interprétation biologique est plus aisée, l'avis d'experts a été pris en compte.

$$\begin{aligned}
\tau_0 &\sim \mathcal{LN}_{[0,1000]}(\log(1000), 1) \\
f_0 &\sim \mathcal{LN}(\log(0.1), 1) \\
a &\sim \mathcal{U}([100, 1000]) \\
b &\sim \mathcal{U}([100, 1000]) \\
\beta_k &\sim \mathcal{N}(0, 100), \quad k = 1, \ldots, K \\
\sigma^2 &\sim \mathcal{IG}(1, 1)
\end{aligned}$$

3 Estimation par la méthode ABC

3.1 Choix des statistiques résumées

Une première étape consiste à construire un ensemble de statistiques résumées, afin de procéder à une première réduction de la dimension. Le choix suivant a été fait, en concertation avec les biologistes : intervalle interquartile et nombre de 0 observés par site, par

période et par an, tout types d'habitat confondus, et intervalle interquartile et nombre de 0 observés par type d'habitat, par période et par an, tous sites confondus. Ce choix a permis de passer de 790 observations à 404 statistiques résumées.

Puis, afin de définir la distance entre les statistiques résumées observées et celles qui seront calculées lors des procédures ABC, on utilise un noyau d'Epanechnikov, dont l'échelle est réglée de sorte à ne conserver qu'une proportion ϵ des valeurs simulées qui sont les plus proches des valeurs observées.

3.2 Méthodes pour l'approximation de la loi a posteriori ABC

Le premier point de vue adopté s'attache à construire des échantillons suivant la loi a posteriori ABC, en utilisant des méthodes de régression pour ajuster les valeurs de paramètres associées à des statistiques résumées situées au voisinage des statistiques résumées observées. L'idée consiste à construire tout d'abord un modèle de régression des paramètres du modèle sur les statistiques résumées (Blum et François, 2010) :

$$\psi_i^{(m)} = m_i\left(s^{(m)}\right) + \sigma_i\left(s^{(m)}\right)\varepsilon_{im}, \quad i = 1, \ldots, p \tag{1}$$

où les ε_{im} sont des variables iid centrées, et où la fonction σ_i permet de prendre en compte une éventuelle hétéroscédasticité. Une fois le modèle de régression estimé, on peut construire des valeurs ajustées pour les paramètres retenus de la façon suivante :

$$\psi_i^{*(m)} = \hat{m}_i(s_{\text{obs}}) + (\psi_i^{(m)} - \hat{m}_i(s^{(m)}))\ \frac{\hat{\sigma}_i(s_{\text{obs}})}{\hat{\sigma}_i(s^{(m)})}, \quad i = 1, \ldots, p.$$

Quatre approches ont été comparées : i) m_i linéaire et erreurs hétéroscédastiques, ii) m_i non linéaire et erreurs hétéroscédastiques, iii) approche en deux étapes où (1) est appliquée une première fois pour obtenir des échantillons ajustés à partir desquels on estime le support de la loi a posteriori ABC, puis on ré-applique (1) sur les échantillons ajustés appartenant à ce support, iv) m_i non linéaire et erreurs homoscédastiques, estimé par un modèle de forêt aléatoire.

3.3 Méthodes pour l'approximation de caractéristiques de la loi a posteriori ABC

Le second point de vue consiste à approcher certaines quantités d'intérêt de la loi a posteriori ABC, sans chercher à reconstruire toute la distribution. On s'est intéressé à l'estimation de la moyenne, de la médiane et des quantiles d'ordre 0.025 et 0.975 de la loi a posteriori ABC.

Deux approches ont été testées : i) régression (classique et quantile) par forêt aléatoire (voir Raynal et al. 2018), ii) régression par méthode de boosting. Dans le premier cas, on a comparé les forêts aléatoires pondérées ou sans poids, en utilisant les poids donnés par le noyau d'Epanechnikov, et dans le deuxième cas on a comparé deux fonctions de perte, la perte L_1 et la perte L_2.

4 Résultats

4.1 Données simulées

Les méthodes ont d'abord été comparées sur un jeu de données simulées, selon la procédure suivante : (i) $M = 100\ 000$ valeurs de paramètres ont été échantillonnées à partir des lois a priori, et M jeux de données simulées ont été construits à l'aide de ces valeurs de paramètres, (ii) 100 jeux de données ont été choisis aléatoirement parmi les 100 000, pour jouer le rôle de jeux de données de référence, et (iii) les différentes méthodes ont été appliquées sur chacun des 100 jeux de données de référence pour obtenir soit des échantillons selon la loi a posteriori ABC soit des quantités unidimensionnelles de cette loi a posteriori, en utilisant les 999 900 jeux de données restants.

Les performances des méthodes ont été comparées à l'aide de l'erreur relative absolue (RAE) entre la médiane a posteriori et la vraie valeur du paramètre (voir Figure 1), et à l'aide de la couverture empirique des intervalles de crédibilité obtenus.

Les méthodes basées sur les forêts aléatoires sont associées aux valeurs les plus faibles de l'erreur relative absolue, quel que soit le paramètre considéré. La méthode de régression adaptative fonctionne en général mieux que les méthodes non adaptatives, et les méthodes basées sur une régression linéaire locale produisent des lois a posteriori aux supports trop larges, ne respectant pas en particulier le support de la loi a priori. En pratique, ces approches n'ont parfois pas abouti à cause de problèmes de convergence.

4.2 Données réelles

Les résultats obtenus sur les données réelles diffèrent selon le paramètre considéré. D'un côté, pour les paramètres du modèle mécaniste de pollinisation, i.e. $\theta = (\tau_0, f_0, a, b)$, les méthodes de type rejet ou utilisant le gradient boosting produisent des intervalles identiques à ceux que l'on obtiendrait avec la loi a priori alors que les autres approches produisent des intervalles sensiblement différents. D'un autre côté, pour les paramètres du modèle d'observation, i.e. $\omega = (\beta_1, \beta_2, \beta_3, \sigma^2)$, toutes les approches produisent des intervalles de crédibilité qui diffèrent de ceux obtenus avec la loi a priori. Pour les méthodes basées sur des forêts aléatoires, les résultats obtenus pour le paramètre σ^2 indiquent une sur-estimation de ce paramètre, ce qui pourrait être dû à une sous-estimation de l'intensité sous-jacente du processus de Poisson.

5 Conclusion

L'objectif de ce travail était de proposer une méthodologie pour la calibration de modèles complexes pour lesquels la vraisemblance n'est pas calculable analytiquement. Différentes méthodes de type ABC ont été comparées, et les avantages et inconvénients de chacune d'entre elles ont été discutés. Il ressort de cette étude que les approches basées sur les forêts aléatoires, se concentrant sur l'approximation de quantités unidimensionnelles de la

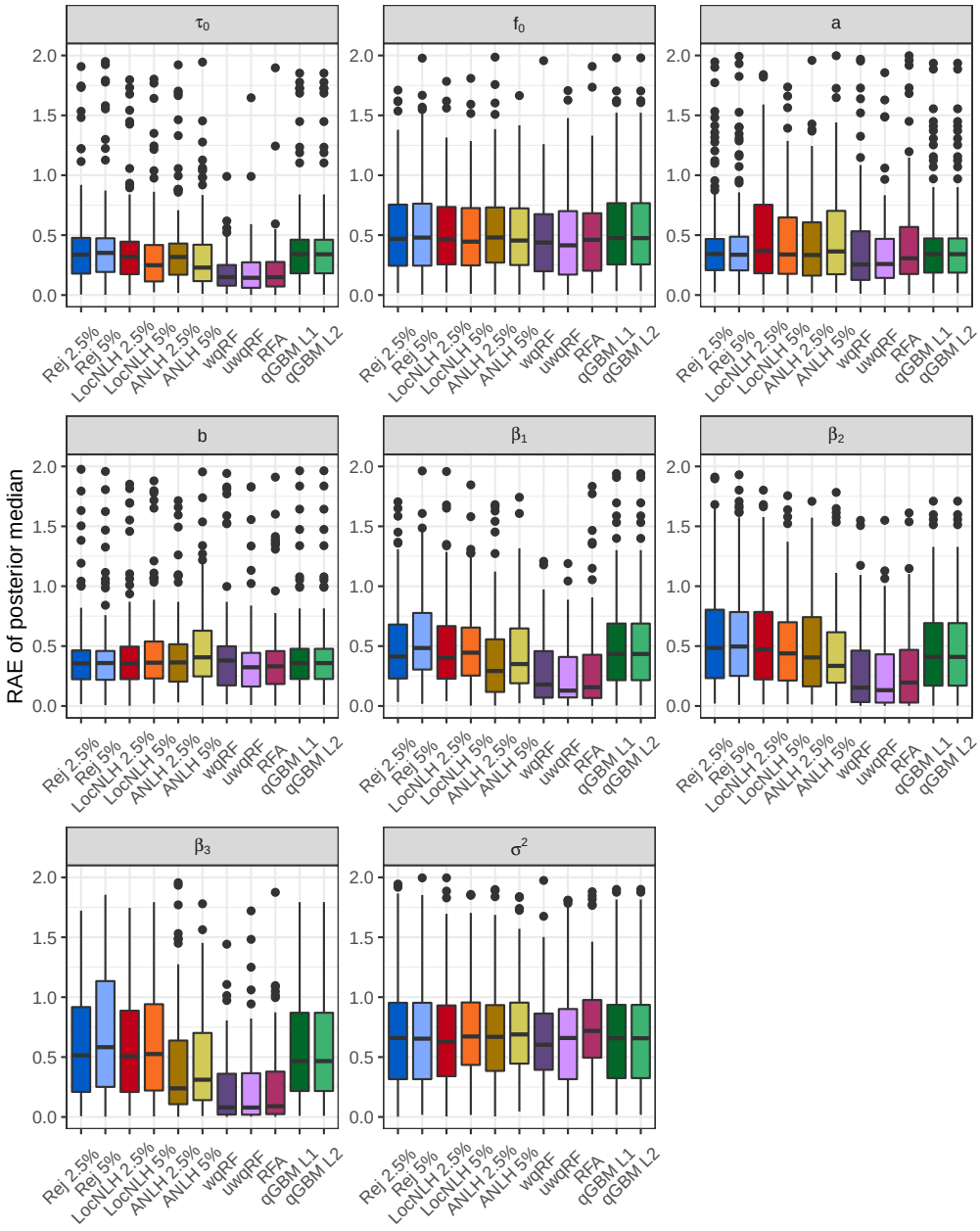

Figure 1: Erreur relative absolue en utilisant la médiane a posteriori

loi a posteriori ont de meilleures performances sur notre modèle. Elles sont particulièrement simples à mettre en place.

En revanche, certains paramètres restent difficiles à estimer, quelle que soit la méthode utilisée, ce qui souligne les difficultés auxquelles on peut faire face dans ce type de contexte. Plusieurs situations peuvent être à l'origine de cette difficulté d'estimation : la loi a priori est mal choisie, le modèle n'est pas identifiable, les données ne sont pas suffisamment informatives, le choix des statistiques résumées n'est pas adapté. Chacune de ces pistes peut être poursuivie afin d'améliorer le modèle, ou le recueil des données.

Bibliographie

C Baey, H. G; Smith, M. Rundlöf, O. Olsson, Y. Clough, U. Sahlin (2023), Calibration of a bumble bee foraging model using Approximate Bayesian Computation, *Ecological Modelling*, 477: 110251.

M. G. Blum and O. François (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and computing*, 20(1):63–73.

O. Olsson and A. Bolin (2014). A model for habitat selection and species distribution derived from central place foraging theory. *Oecologia*, 175(2):537–548.

L. Raynal, J.-M. Marin, P. Pudlo, M. Ribatet, C. P. Robert, and A. Estoup (2018), ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728.

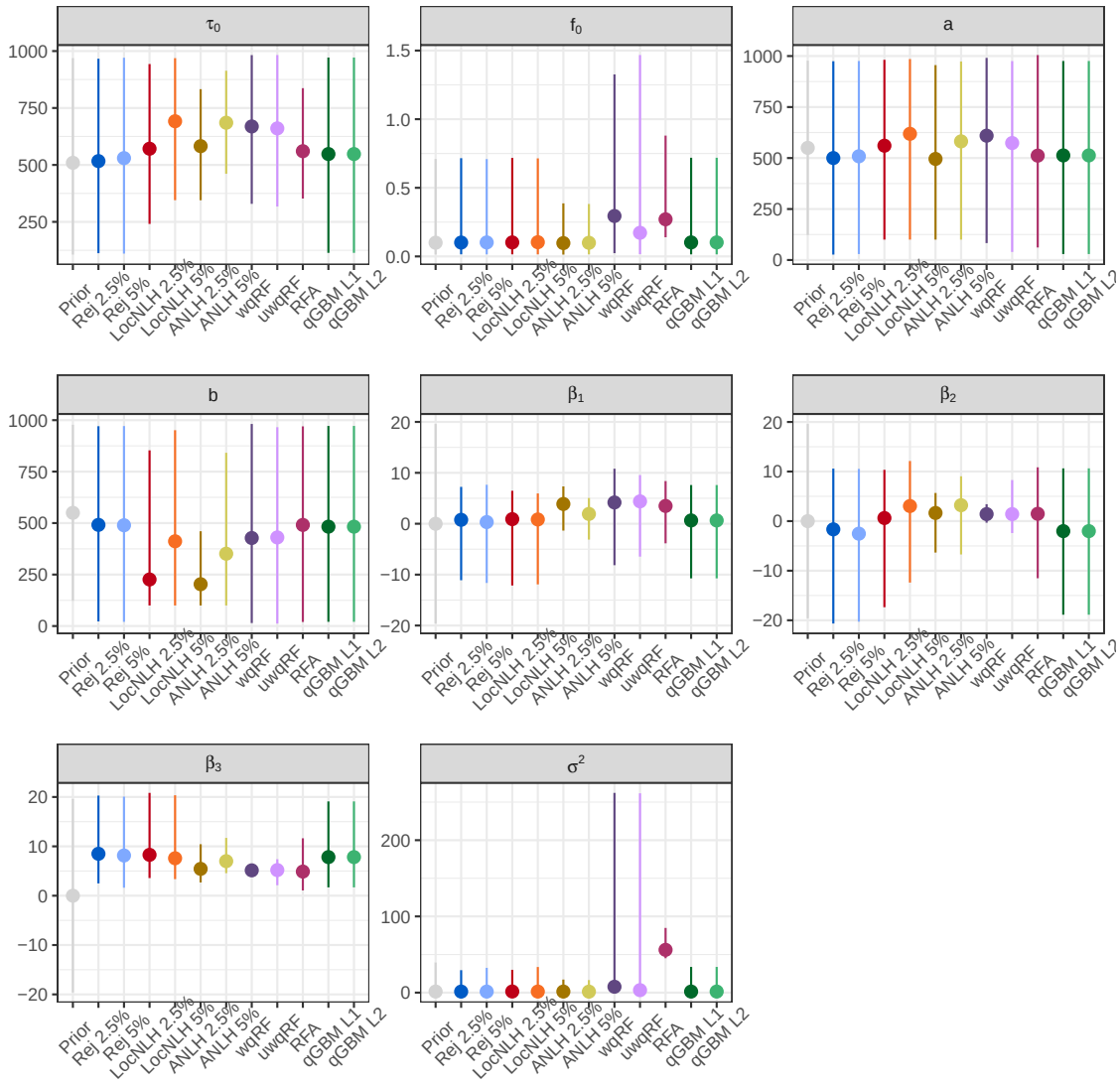

Figure 2: Intervalle de crédibilité à 95% sur les données réelles.