Calibration of a pollination model using Approximate Bayesian Computation

joint work with Ullrika Sahlin, Yann Clough and Henrik G. Smith (from Lund University)

StatMathAppli 2023
September 18th, 2023

Charlotte Baey

Université de Lille

Context

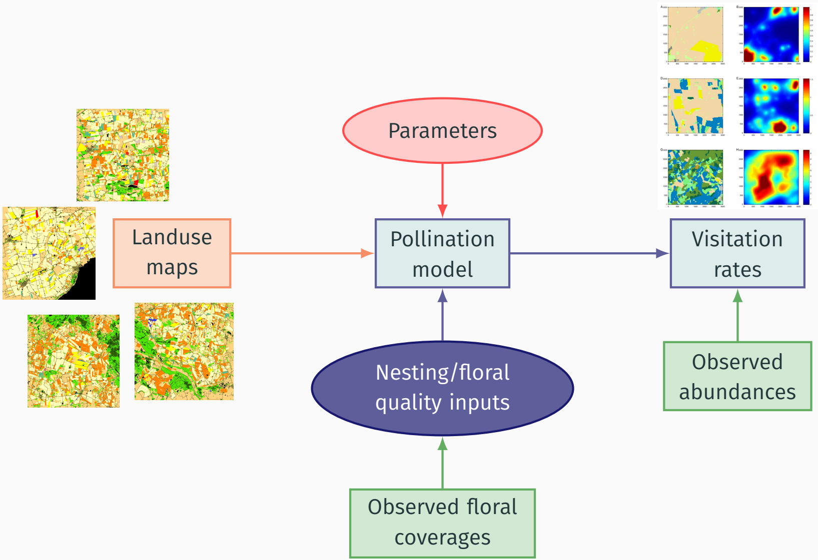
Context

- Evaluate the impacts of different changes on ecosystems and ecosystem services
  → *the benefits humans obtain from ecosystems (e.g. : crop pollination, oxygen production by plants, carbon sequestration, ...)*
- To this aim, some models for ecosystem services have been developed
- But they are often complex (black-box models, time-consuming, ...) and rarely calibrated on experimental data (rely on expert judgment, literature data, ...)
- **Objective:** propose a general methodology to calibrate these models

Model and data

Pollination model: Central Place Foragers (CPF) model

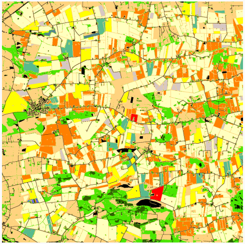
Pollination model for bumble bees based on central foraging theory:



Model inputs

For each sampling site i, each year j and each period k:

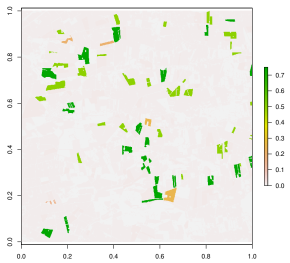
A landscape map



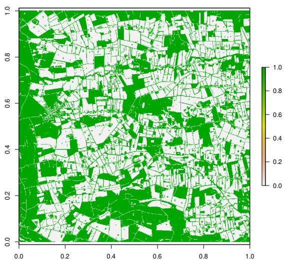
$\Downarrow$

denoted by $\mathcal{M}_{ijk}$

A "floral quality" map



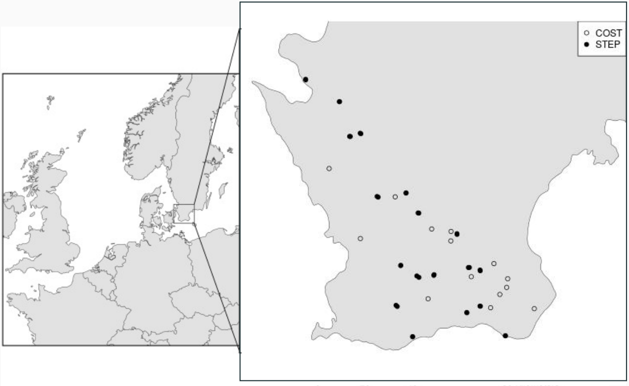
A "nesting" map



$\underbrace{\hspace{10em}}$

informed by expert judgement or literature data

Data



- Two studies on pollinator abundances in southern Sweden
- Data collected in four different years, several times a year (covering 3 different periods of bumblebees life cycle) → **790 data points**
- Number of bees flying or foraging in a given transect for a given period of time was recorded

Statistical model - Bayesian formulation

- y_{ijk}: observed nb of bees on site i, year j and period k.

Statistical model - Bayesian formulation

- y_{ijk}: observed nb of bees on site i, year j and period k.
- **Likelihood**

$$\begin{cases} y_{ijk} \mid \lambda_{ijk}, \theta & \sim \mathcal{P}(c_i \cdot \lambda_{ijk}) \\ \log \lambda_{ijk} & = \log \nu_i(\theta, \mathcal{M}_{jk}) + \beta_k + \varepsilon_{ijk} \\ \varepsilon_{ijk} & \sim \mathcal{N}(0, \sigma^2). \end{cases}$$

  - c_i a known scaling parameter,
  - λ_{ijk} the real intensity of the visitation rates,
  - $\nu_i(\theta, \mathcal{M}_{ijk})$ is the predicted visitation rates,
  - β_k a period-specific parameter
- Complete vector of parameters $\psi = (\tau_0, f_0, a, b, \beta_1, \ldots, \beta_K, \sigma^2)$

Statistical model - Bayesian formulation

- y_{ijk}: observed nb of bees on site i, year j and period k.
- **Likelihood**

$$\begin{cases} y_{ijk} \mid \lambda_{ijk}, \theta & \sim \mathcal{P}(c_i \cdot \lambda_{ijk}) \\ \log \lambda_{ijk} & = \log \nu_i(\theta, \mathcal{M}_{jk}) + \beta_k + \varepsilon_{ijk} \\ \varepsilon_{ijk} & \sim \mathcal{N}(0, \sigma^2). \end{cases}$$

  - c_i a known scaling parameter,
  - λ_{ijk} the real intensity of the visitation rates,
  - $\nu_i(\theta, \mathcal{M}_{ijk})$ is the predicted visitation rates,
  - β_k a period-specific parameter
- Complete vector of parameters $\psi = (\tau_0, f_0, a, b, \beta_1, \ldots, \beta_K, \sigma^2)$
- **Priors**

$$\tau_0 \sim \mathcal{LN}_{[0,1000]}(\log(1000), 1) \quad f_0 \sim \mathcal{LN}(\log(0.1), 1)$$

$$a \sim \mathcal{U}([100, 1000]) \quad b \sim \mathcal{U}([100, 1000])$$

$$\beta_k \sim \mathcal{N}(0, 100), \quad k = 1, \ldots, K$$

$$\sigma^2 \sim \mathcal{IG}(1, 1)$$

- In a Bayesian context, we are now interested in the **posterior** distribution of the parameters:

$$\pi(\psi \mid y) \propto \underbrace{f(y \mid \psi)}_{\text{likelihood}} \underbrace{p(\psi)}_{\text{prior}}$$

- But here the likelihood is intractable:

$$\begin{aligned} f(y \mid \psi) &= \int f(y, \lambda \mid \psi) d\lambda = \int f(y \mid \lambda, \psi) f(\lambda \mid \psi) d\lambda \\ &= \prod_{ijk} \frac{1}{\sqrt{2\pi}\sigma y_{ijk}!} \int_0^{+\infty} e^{-\lambda} \lambda^{y_{ijk}-1} \exp\left(-\frac{(\log \lambda - \log \nu_i(\theta, \mathcal{M}_{ijk}) - \beta_k)^2}{2\sigma^2}\right) d\lambda \end{aligned}$$

- We rely on approximate Bayesian computation (ABC)

Approximate Bayesian Computation

Approximation Bayesian computation (ABC)

- Introduced at the end of the 1990 in the area of population genetics

Approximation Bayesian computation (ABC)

- Introduced at the end of the 1990 in the area of population genetics

ABC rejection sampling (Tavaré et al. 1997)

Input: a threshold ε and a distance d on the set of observations

For $m = 1, \ldots, M$:

1. draw a sample $\psi^{(m)}$ from the prior distribution
2. generate a set of observations $y^{(m)}$ using $p(y \mid \psi)$
3. if $d(y_{obs}, y^{(m)}) \leq \varepsilon$, keep $\psi^{(m)}$
4. **Output**: a sample of size M_ε with all the accepted sets of parameters $\psi^{(m)}$

- Introduced at the end of the 1990 in the area of population genetics

ABC rejection sampling (Tavaré et al. 1997)

Input: a threshold ε and a distance d on the set of observations

For $m = 1, \dots, M$:

1. draw a sample $\psi^{(m)}$ from the prior distribution
2. generate a set of observations $y^{(m)}$ using $p(y \mid \psi)$
3. if $d(y_{obs}, y^{(m)}) \leq \varepsilon$, keep $\psi^{(m)}$
4. **Output**: a sample of size M_ε with all the accepted sets of parameters $\psi^{(m)}$

- Curse of dimensionality: increase M or ε to get a reasonable value M_ε

Approximate Bayesian computation (ABC)

Several extensions to the original algorithm have been proposed:

- introduction of summary statistics $s(\cdot)$ of dimension $q < n \rightarrow$ samples from $\pi(\psi \mid s_{obs})$ instead of the posterior $\pi(\psi \mid y_{obs})$ (Blum et al. 2013)
- replace crude rejection by kernel smoothing $\rightarrow$ each sample is used, with a weight $w_m = K(d(y_{obs}, y^{(m)}))$
- produce adjusted samples using the relationship between parameters and summary statistics (Blum et François, 2010)
- approaches focusing on the estimation of one-dimensional quantities from the ABC posterior (Raynal et al. 2018)

Summary of our approach

Initial `abcTable`

Parameters				Summary statistics			
$\psi_1^{(1)}$	$\psi_2^{(1)}$	$\dots$	$\psi_p^{(1)}$	$s_1^{(1)}$	$s_2^{(1)}$	$\dots$	$s_q^{(1)}$
$\psi_1^{(2)}$	$\psi_2^{(2)}$	$\dots$	$\psi_p^{(2)}$	$s_1^{(2)}$	$s_2^{(2)}$	$\dots$	$s_q^{(2)}$
$\vdots$							
$\psi_1^{(M)}$	$\psi_2^{(M)}$	$\dots$	$\psi_p^{(M)}$	$s_1^{(M)}$	$s_2^{(M)}$	$\dots$	$s_q^{(M)}$

define weights for each sample

$$w_m = \frac{3}{4}\left(1 - \frac{(s^{(m)} - s_{\text{obs}})^2}{q_\varepsilon}\right) \mathbb{1}_{(s^{(m)} - s_{\text{obs}})^2 \leq q_\varepsilon}$$

Regression adjustment techniques (section 2.3.4)

1. regression of parameters on summary statistics

$$\psi_i^{(m)} = m_i\left(s^{(m)}\right) + \sigma_i\left(s^{(m)}\right)\varepsilon_{im}$$

2. adjusted samples from estimated regression model

$$\psi_i^{*(m)} = \hat{m}_i(s_{\text{obs}}) + (\psi_i^{(m)} - \hat{m}_i(s^{(m)}))\,\frac{\hat{\sigma}_i(s_{\text{obs}})}{\hat{\sigma}_i(s^{(m)})}$$

Output

$\psi_i^{*(m)}, i = 1, ..., p, m = 1, ..., M$
adjusted ABC posterior samples

Machine learning approaches (section 2.3.5)

1. quantile regression using random forest
2. quantile regression using boosting methods

Output

Scalar predictions from ABC posterior:
posterior mean and median, 95% CI

Methods based on regression adjustment

- **Main idea:** build a relationship between the parameter values and the summary statistics values, e.g. via regression techniques.

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

  Then, samples from $\pi_{ABC}(\psi \mid s_{obs})$ are obtained via:

$$\psi_i^{*(m)} = \hat{m}_i(s_{obs}) + \hat{\sigma}_i(s_{obs}) \frac{\left(\psi_i^{(m)} - \hat{m}(s^{(m)})\right)}{\hat{\sigma}_i(s^{(m)})}$$

- several choices for m_i and σ_i to handle nonlinearity and heteroscedasticity

Methods based on regression adjustment

We compared:

Regression adjustment methods

- local linear heteroscedastic model (Beaumont et al. 2002) [LocLH]

With these methods, we get as outputs a sample of the ABC posterior distribution.

Methods based on regression adjustment

We compared:

Regression adjustment methods

- local linear heteroscedastic model (Beaumont et al. 2002) [LocLH]
- local nonlinear heteroscedastic model (Blum and François 2010) [LocNLH]

With these methods, we get as outputs a sample of the ABC posterior distribution.

Methods based on regression adjustment

We compared:

Regression adjustment methods

- local linear heteroscedastic model (Beaumont et al. 2002) [LocLH]
- local nonlinear heteroscedastic model (Blum and François 2010) [LocNLH]
- adaptive nonlinear heteroscedastic model (Blum and François 2010) [ANLH]
  → two-step procedure:

With these methods, we get as outputs a sample of the ABC posterior distribution.

Methods based on regression adjustment

We compared:

Regression adjustment methods

- local linear heteroscedastic model (Beaumont et al. 2002) [LocLH]
- local nonlinear heteroscedastic model (Blum and François 2010) [LocNLH]
- adaptive nonlinear heteroscedastic model (Blum and François 2010) [ANLH]
  → two-step procedure:
  1. perform a LocNLH regression and estimate the distribution support D of the adjusted values

With these methods, we get as outputs a sample of the ABC posterior distribution.

Methods based on regression adjustment

We compared:

Regression adjustment methods

- local linear heteroscedastic model (Beaumont et al. 2002) [LocLH]
- local nonlinear heteroscedastic model (Blum and François 2010) [LocNLH]
- adaptive nonlinear heteroscedastic model (Blum and François 2010) [ANLH]
  → two-step procedure:
  1. perform a LocNLH regression and estimate the distribution support D of the adjusted values
  2. perform a second LocNLH regression using parameters values samples from p_D, the conditional prior of the parameters given that they fall in D

With these methods, we get as outputs a sample of the ABC posterior distribution.

Methods based on regression adjustment

We compared:

Regression adjustment methods

- local linear heteroscedastic model (Beaumont et al. 2002) [LocLH]
- local nonlinear heteroscedastic model (Blum and François 2010) [LocNLH]
- adaptive nonlinear heteroscedastic model (Blum and François 2010) [ANLH]
  → two-step procedure:
  1. perform a LocNLH regression and estimate the distribution support D of the adjusted values
  2. perform a second LocNLH regression using parameters values samples from p_D, the conditional prior of the parameters given that they fall in D
- nonlinear homoscedastic regression via random forest (Bi et al. 2022) [RFA]

With these methods, we get as outputs a sample of the ABC posterior distribution.

Methods based on quantile regression

- Sometimes we are only interested in some quantities from the posterior distribution (e.g. quantiles, mean, ...)

Methods based on quantile regression

- Sometimes we are only interested in some quantities from the posterior distribution (e.g. quantiles, mean, ...)

→ what if we try to approximate these quantities using ABC instead of the whole posterior ?

Methods based on quantile regression

- Sometimes we are only interested in some quantities from the posterior distribution (e.g. quantiles, mean, ...)

→ what if we try to approximate these quantities using ABC instead of the whole posterior ?

Methods based on quantile regression

- Sometimes we are only interested in some quantities from the posterior distribution (e.g. quantiles, mean, ...)

→ what if we try to approximate these quantities using ABC instead of the whole posterior ?

Quantile regression methods

- Quantile regression using random forests (Raynal et al. 2016) [qRF]
- Quantile regression using gradient boosting [qGBM]

With these methods, we get as outputs the mean, the median, and the 2.5% and 97.5% quantiles of the ABC posterior distribution.

Choice of the summary statistics

We used the interquartile range and the number of 0's:

1. per site, per period and per year, all habitat types combined
2. per habitat type, per period and per year, all sites combined

- aggregation across habitats accounts for differences in population sizes between landscapes,
- habitat-specific summaries captures joint effect of population size and relative attractiveness of the habitats

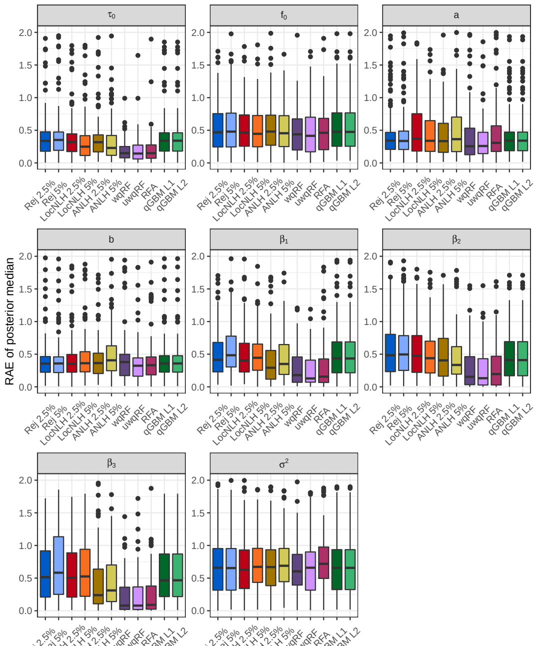
→ first reduction of the dimension, from 790 data points to 404 summary statistics

Results

Simulation study

- $M = 100\,000$ parameter samples from the prior $\longrightarrow$ M datasets
- 100 datasets were randomly chosen as reference datasets
- ABC posterior samples and quantiles were estimated on these 100 datasets using the remaining 999 900 datasets.
- Two values for the threshold q_ε in the weighting kernel (2.5% or 5% of the data)
- Comparison of the relative absolute error between posterior median and true value, empirical coverage of the CI

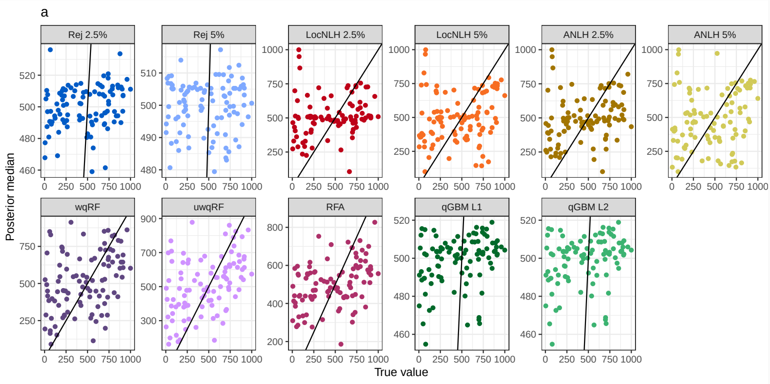
Results – RAE



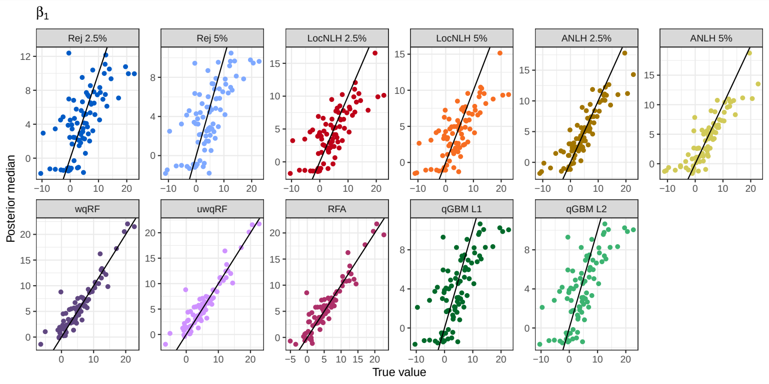
τ_0
f_0
a
b
β_1
β_2
β_3
σ^2
RAE of posterior median
Rej 2.5%
Rej 5%
LocNLH 2.5%
LocNLH 5%
ANLH 2.5%
ANLH 5%
wqRF
uwqRF
RFA
qGBM L1
qGBM L2

Results – MAP estimate

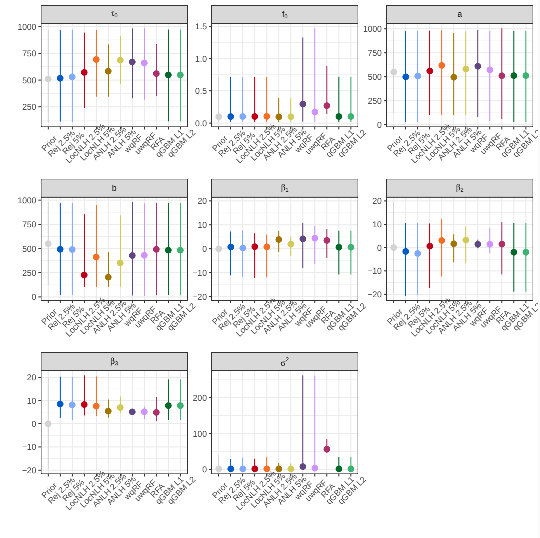
Extracted results for parameters a and β_1:



Extracted results for parameters a and β_1:

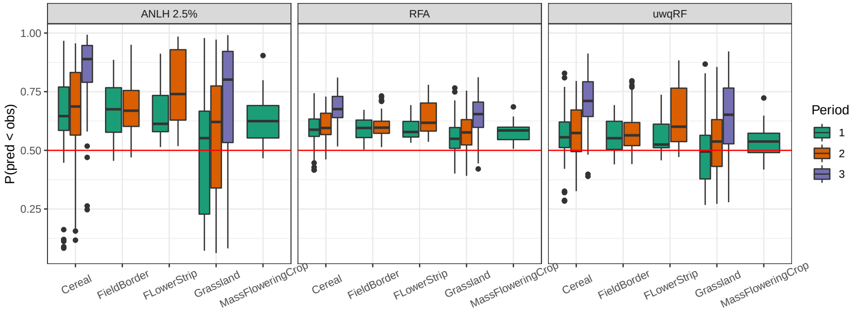


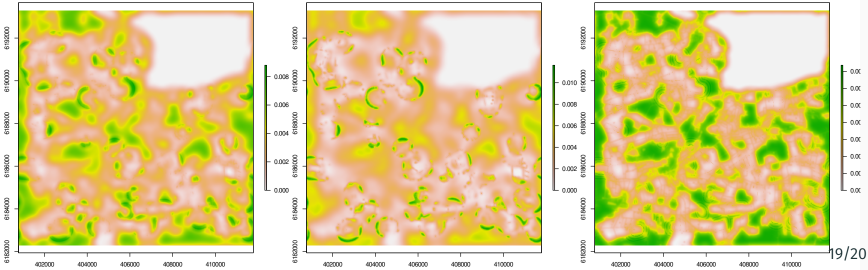
Results on real data



- 95% CI narrower than prior for most parameters using the best identified methods
- Some parameters are difficult to estimate
- σ^2 is overestimated by some methods

Results – predictions





Conclusion and perspectives

Conclusion and perspectives

Conclusion

- Posterior distributions were narrower than the prior for most parameters
- But, some parameters were difficult to estimate (CPF parameters vs. observation parameters) → identifiability issues?
- Predicted values tend to be overdispersed
- Results are conditional on the floral and nesting maps

Perspectives

- Use the estimated ABC posterior distribution to tune likelihood-free MCMC algorithms (initialization of the chain, choice of the proposal distribution) (e.g. Wegmann 2009)
- Evaluate the influence of the input maps
- Perform model comparison