

Projet HAB904B

Basile Pajot (DARWIN), Marion Themeze-Leroy (ECOSYSTEMES),

2023-12-21

Contents

1	Lecture et exploration des données	1
1.1	La variable à expliquer	1
1.2	Les variables explicatives	1
1.3	Exploration des données	2
2	Ajustement d'un modèle simple	5
3	Comparaison de modèles	9
4	Inférence et interprétation des résultats	11
4.1	Prévalence discontinue	12
5	Discussion	14

1 Lecture et exploration des données

1.1 La variable à expliquer

La variable d'intérêt à expliquer, est **Shells**, soit le nombre de carapaces de tortues récentes trouvées lors des relevés sur le terrain. Cette variable est un proxy pour estimer le nombre de tortues mortes d'une année sur l'autre.

1.2 Les variables explicatives

- **Prev** est une variable explicative qualitative qui correspond à la prévalence pour *Mycoplasma agassizii*, soit le rapport entre le nombre de tortues séropositives sur l'effectif total de tortues par année pour chaque site.
- **Site** est une variable qualitative qui correspond au site d'échantillonnage. Elle a 10 modalités : le parc national *Big Shoals* (*BS*), l'aire de gestion de la faune sauvage *Camp Blanding* (*CB*), l'aire de gestion de la faune sauvage et de l'environnement *Cecil Field/Branan Field* (*CF*), une propriété privée en Floride centrale (*Cent*), le parc national *Fort Cooper* (*FC*), l'aire de gestion de la faune sauvage *Flying Eagle* (*FE*), le parc national *Gold Head Branch* (*GH*), l'aire de gestion de la faune sauvage et de l'environnement *Perry Oldenburg* (*Old*), la station biologique *Ordway-Swisher* (*Ord*), l'aire de gestion de la pêche *Tenoroc Fish* (*TE*).

- **Area** est une variable quantitative qui correspond à l'aire couverte par site lors des relevés.
- **Year** est une variable qualitative qui correspond à l'année pour laquelle les relevés ont été faits. Elle a 3 modalités : 2004, 2005, 2006.

1.3 Exploration des données

Nous regardons un résumé statistique des variables de notre jeu de donnée.

```
##      Site      year      shells      type      Area
## BS      : 3  Length:30      Min.    :0.00  Fresh:30  Min.    : 5.30
## CB      : 3  Class :character  1st Qu.:0.00      1st Qu.:15.20
## Cent    : 3  Mode  :character  Median :1.00      Median :27.30
## CF      : 3      Mean  :1.80      Mean  :29.02
## FC      : 3      3rd Qu.:2.75      3rd Qu.:43.20
## FE      : 3      Max.    :9.00      Max.    :61.00
## (Other):12
##      density      prev      total_turtle      standprev
## Min.    : 1.80  Min.    : 1.00  Min.    : 44.80  Min.    : -0.9163
## 1st Qu.: 2.50  1st Qu.: 1.20  1st Qu.: 74.75  1st Qu.: -0.9088
## Median : 3.50  Median :15.95  Median :104.22  Median : -0.3559
## Mean    : 8.76  Mean    :25.45  Mean    :119.72  Mean    : 0.0000
## 3rd Qu.: 4.80  3rd Qu.:42.05  3rd Qu.:174.90  3rd Qu.: 0.6223
## Max.    :33.00  Max.    :80.70  Max.    :200.79  Max.    : 2.0709
##
##      H      Cov_2004      Cov_2005      Cov_2006
## Min.    :0.0000  Min.    :0.0000  Min.    :0.0000  Min.    :0.0000
## 1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.0000
## Median :0.0000  Median :0.0000  Median :0.0000  Median :0.0000
## Mean    :0.4333  Mean    :0.3333  Mean    :0.3333  Mean    :0.3333
## 3rd Qu.:1.0000  3rd Qu.:1.0000  3rd Qu.:1.0000  3rd Qu.:1.0000
## Max.    :1.0000  Max.    :1.0000  Max.    :1.0000  Max.    :1.0000
##
```

Nous avons un plan d'expérience équilibré avec un même nombre d'observations par site et par année.

Pour la prevalence **prev** la moyenne est supérieure à la médiane, c'est-à-dire que plus de 50% des valeurs sont inférieures à la moyenne. Il en est de même pour le nombre de carapces **shells**. De plus, pour la prévalence, la différence entre le troisième quartile et le minimum est d'environ 40, tout comme la différence entre le maximum et le 3ème quartile. Ainsi, la gamme de valeurs prise par 25% des données est égales à celle prise par 75% des données. Pour le nombre de carapaces, la gamme de valeurs prise par 25% des données est plus de trois fois supérieur à celle prise par 75% des données. Ceci est illustré par les *figures 1 et 2*.

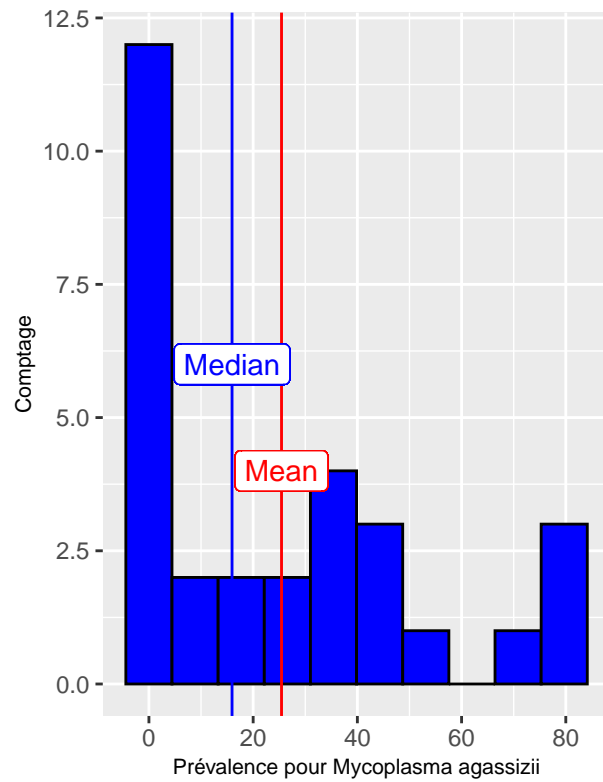


Figure 1 : Distribution de la prévalence pour Mycoplasma agassizii

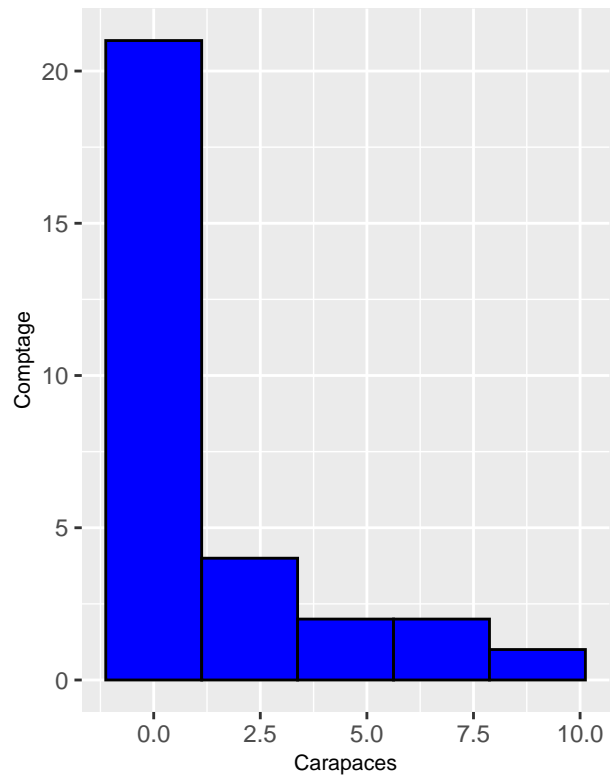


Figure 2 : Distribution du nombre de carapaces

La distribution du nombre de carapaces ressemble à une distribution de Poisson.

La figure 3 donne plusieurs informations sur notre jeu de données.

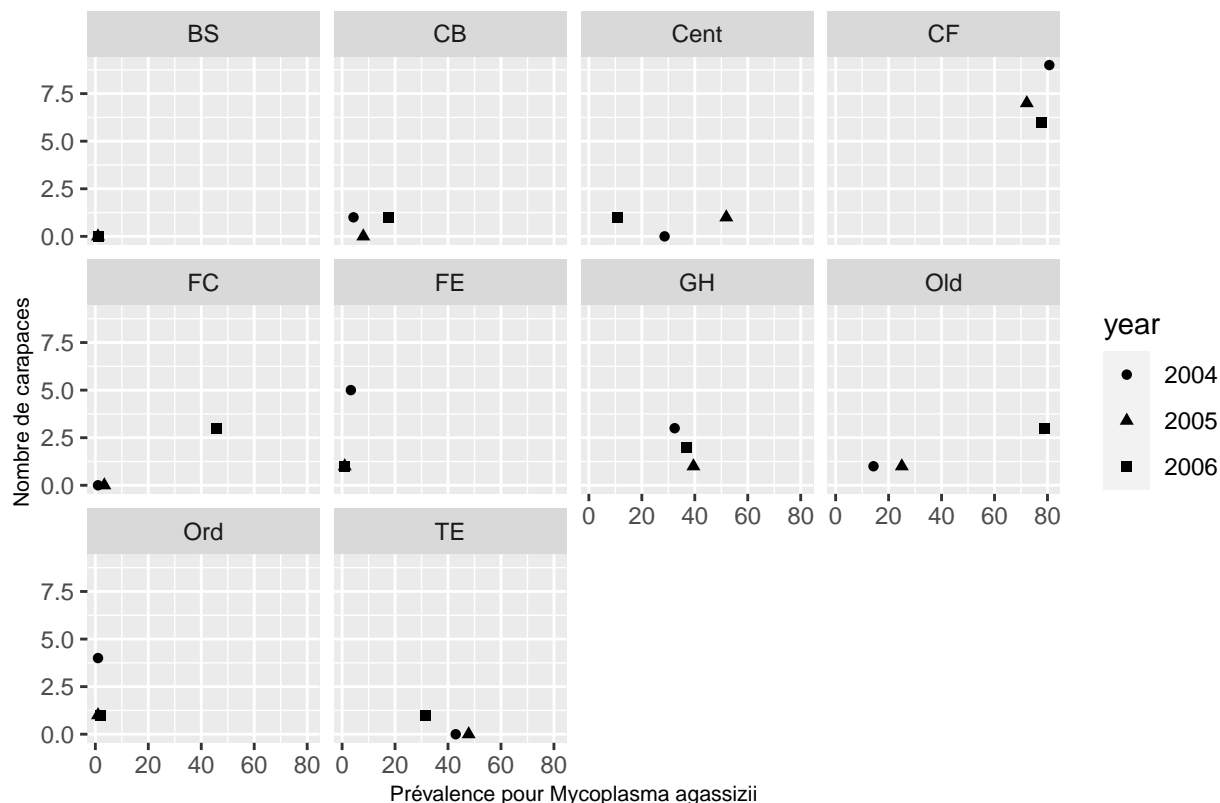


Figure 3 : Le nombre de carapaces en fonction de la prevalence par site par année

Tout d'abord, le nombre de carapaces récentes trouvées varie ou non en fonction des années et cette variation n'est pas la même en fonction des sites. Le nombre de carapaces récentes par rapport à l'année précédente reste constante augmente ou diminue. On observe des évolutions différentes pour les sites : on observe une diminution du nombre de carapaces pour le site CF sur les trois ans ou un changement de tendance se traduisant par une diminution puis une augmentation pour le site CB.

Ensuite, la prévalence en fonction des sites peut également varier en fonction des années. Comme précédemment, cette variation n'est pas la même en fonction des sites. La prévalence reste constante, augmente ou diminue. La variation peut être globale sur les trois années d'étude (augmentation de la prévalence pour le site Old) ou changer (augmentation puis diminution pour le site Cent). Lorsque la prévalence augmente d'une année à l'autre $prev[n] < prev[n+1]$, le nombre de carapaces récentes trouvées l'année suivante augmente $shells[n+2] > shells[n+1]$ (sites CB, Old), et inversement (sites CF). Ainsi, les variations du nombre de carapaces récentes trouvées pourrait être expliquée par les variations de la prévalence.

Il apparaît également que certains sites ont de faibles prévalences (BS, Ord, FE) quelque soit l'année et d'autres des prévalences élevées (CF, GH, TE). Ceci concourt avec les observations faites précédemment avec le résumé de la variable `prev` et la distribution de la prévalence. Ainsi, nous pourrions séparer les sites en deux catégories, ceux à faible ou forte prévalence. Cette variable sera donc traitée de deux manières : de manière continue et de manière discontinue avec deux catégories :

- faible prévalence (0) : `Prev < 0.25`
- haute prévalence (1) : `Prev > 0.25`

Nous allons donc essayer de déterminer si la prévalence et l'année permettent d'expliquer les variations du nombre de carapaces. Nous avons vu que la prévalence et le nombre de carapaces trouvées diffère entre les sites et entre les années. Afin de pouvoir nous concentrer sur l'effet de la prévalence, nous allons mettre un effet aléatoire sur la variable `Site`. Nous pourrions faire de même pour la variable `Année` mais par souci de simplification, nous allons garder cette variable en effet fixe.

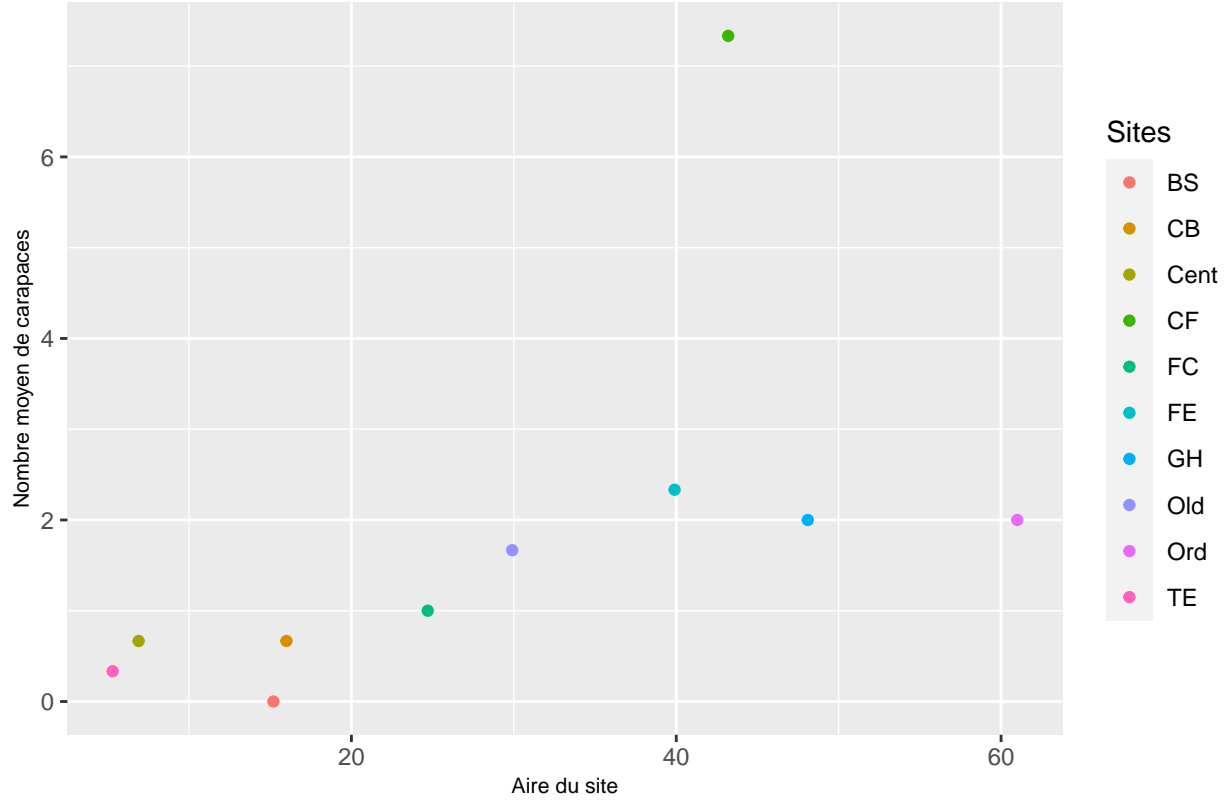


Figure 4 : Le nombre moyen de carapces en fonction de l'aire du site d'étude

D'après la *figure 4* les sites n'ont pas tous la même aire et il semble qu'un plus grand nombre de carapces sont trouvés sur les sites avec une plus grande aire. Afin de pouvoir comparer les sites entre eux, nous allons prendre le rapport entre le nombre de carapces trouvées par site et l'aire du site.

D'après les observations faites précédemment, nous souhaitons donc déterminer si :

- le nombre de carapces récentes trouvées est corrélée avec la prevalence pour *Mycoplasma agassizii* pour une année donnée.
- le nombre de carapces récentes trouvées est plus grand dans les sites à haute prévalence par rapport aux sites à basse prévalence.

2 Ajustement d'un modèle simple

Nous commençons par un modèle simple M1 en considérant uniquement l'effet de la prévalence sur le nombre de carapces récentes trouvées.

L'équation (approche fréquentiste) du modèle linéaire simple est la manière suivante :

$$\frac{shells}{aire_{site}} = \mu_0 + \beta * prev$$

Ceci se traduit en approche bayésienne par un modèle considérant les hypothèses suivantes :

- **shells** suit un loi de poisson de paramètre λ (*cf figure 2*), c'est-à-dire que c'est une variable discrète de comptage dans un intervalle de temps et un espace donnés ; avec une variance égale à la moyenne $E(shells) = V(shells) = \lambda$

- toutes les observations de **shells** sont **indépendantes**
- le logarithme de la moyenne de **shells** peut être exprimée comme la combinaison linéaire des variables explicatives sélectionnées.
- que les paramètres à estimer (ordonnée à l'origine et coefficients de regression) suivent des lois connues, explicités ci-après.

Nous avons donc :

$$Shells_i \stackrel{i.i.d}{\sim} Pois(\lambda_i) \text{ avec } i = 1, \dots, 30 \text{ le nombre d'observations}$$

$$\log(\lambda_i) = \log(aire_i) + \mu_0 + \beta * prev_i, \text{ la fonction de lien}$$

Nous utilisons comme priors les distribution suivantes :

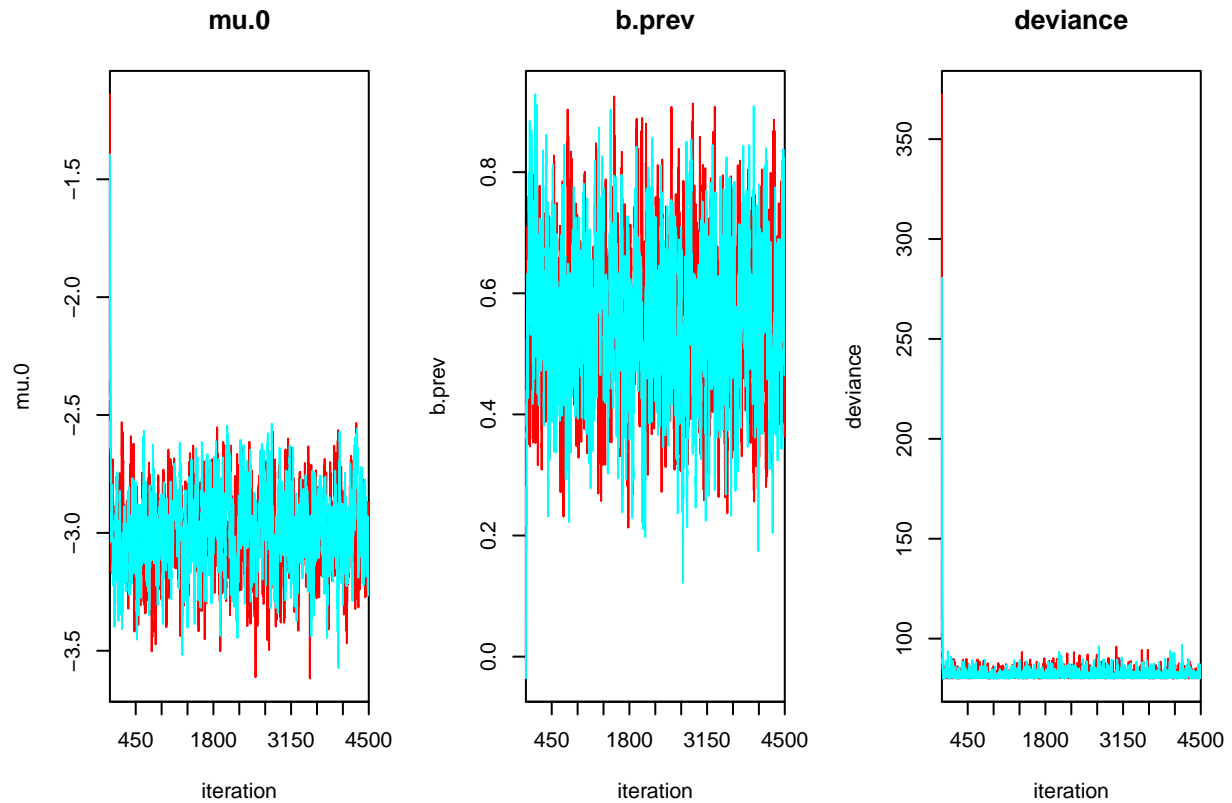
$$\mu_0 \sim \mathcal{N}(0, 100)$$

$$\beta \sim \mathcal{N}(0, 100)$$

Pour le modèle M1 nous obtenons les résultats suivants.

```
## Inference for Bugs model at "/var/folders/l9/dzz054qd4_q8g63vvgsts36w0000gn/T//Rtmp0sfhvx/modelba155
## 2 chains, each with 9000 iterations (first 4500 discarded)
## n.sims = 9000 iterations saved
##          mu.vect sd.vect  2.5%   25%   50%   75%  97.5%  Rhat n.eff
## b.prev      0.563   0.116  0.337  0.485  0.561  0.641  0.795 1.002 1700
## mu.0        -3.001   0.167 -3.318 -3.111 -3.001 -2.894 -2.681 1.006 270
## deviance    82.188   4.824 80.153 80.651 81.424 82.844 87.519 1.011 3800
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 11.6 and DIC = 93.8
## DIC is an estimate of expected predictive error (lower deviance is better).
```

Nous vérifions que le modèle a bien convergé.



Les 2 chaînes se mélangent bien et convergent toutes deux. Ceci est aussi confirmé par la statistique de Gelman-Rubin \hat{R} qui est inférieure à 1.1 pour chaque paramètre estimé. Nous notons aussi que `n.eff` est supérieur à 100.

Nous avons donc des estimations de nos paramètres qui sont stables et des chaînes peu autocorrélées.

Nous regardons maintenant la distribution *a posteriori* et le moyenne (λ) du nombre de carapaces à partir des paramètres estimés par notre modèle.

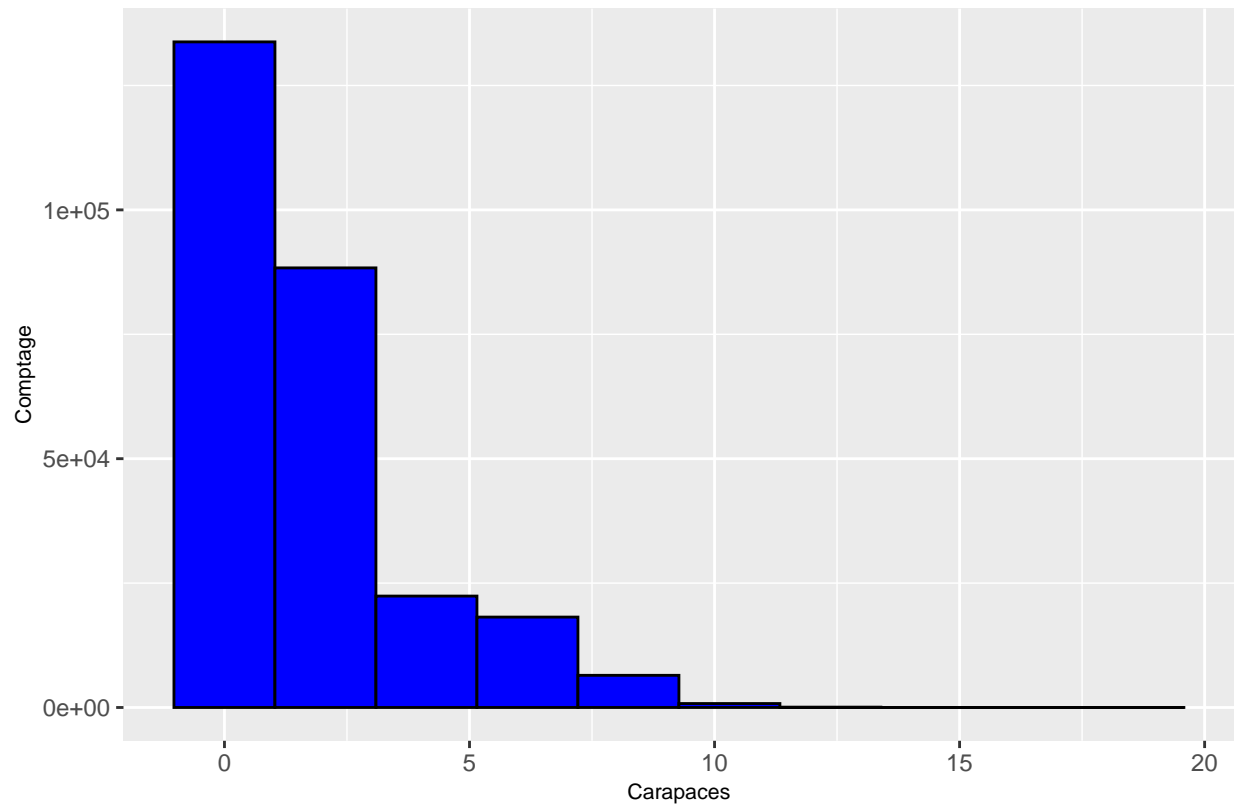


Figure 5 : Distribution du nombre de carapaces pour le modèle 1

```
## [1] 1.810626
```

Les valeurs estimées du modèle se prettent assez bien à prédire la distribution du nombre de carapaces. Regardons ce qu'il en est sur le nombre de carapaces en fonction de la prévalence en fonction du site et de l'année.

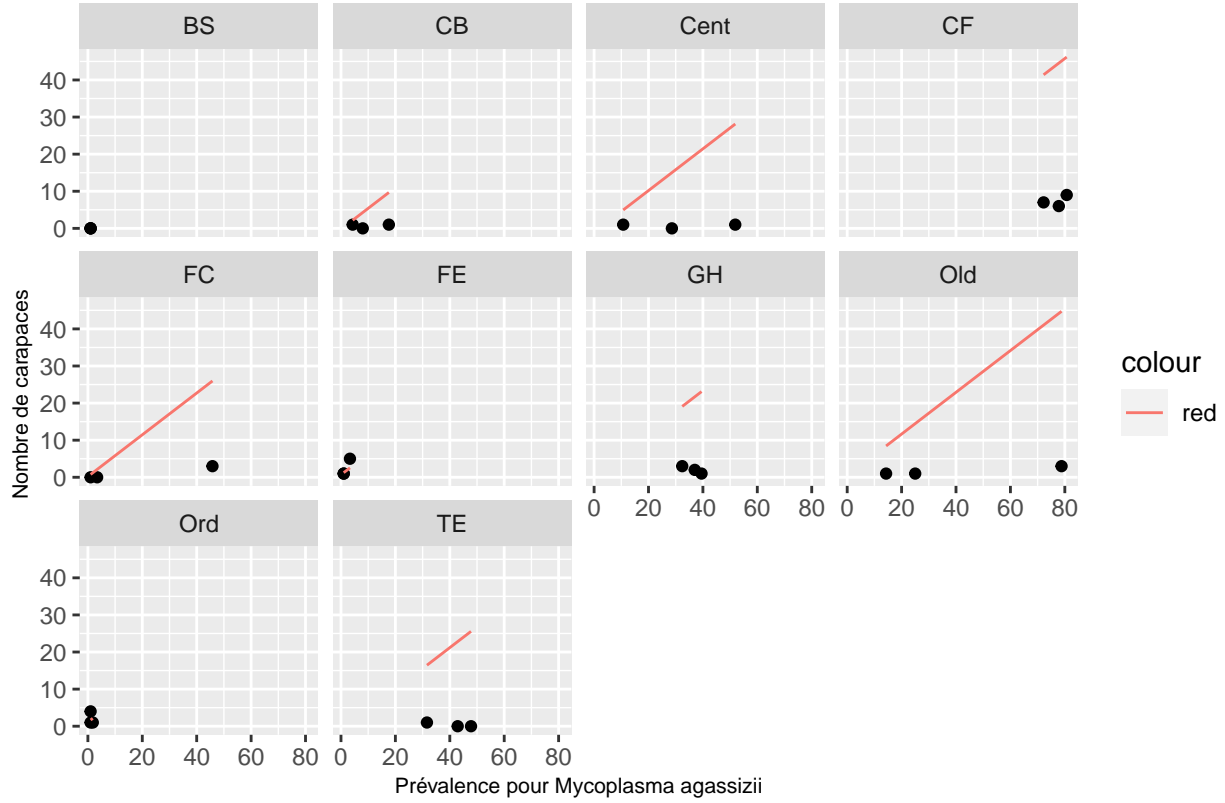


Figure 6 : Le nombre de carapaces en fonction de la prevalence par site par année

Notre modèle M1 prédit bien la distribution des carapaces (*cf* figure 5) mais elle ne prend pas en compte la variabilité des sites comme le montre les droites de regressions construites à partir de le moyenne de distributions postérieures de nos paramètres (*cf* figure 6).

3 Comparaison de modèles

D'après nos observations de la *partie 1*, nous pouvons faire l'hypothèse que le nombre de carapaces récentes trouvées peut être expliqué par un effet de l'année, par la prévalence pour *Mycoplasma agassizii* et par un effet du site.

Bien qu'il y ait une variabilité entre sites, nous souhaitons avant tout connaître l'effet de la prévalence sur le nombre de carapaces récentes d'une année sur l'autre, quel que soit le site. Afin de pouvoir généraliser à tout site, nous allons mettre un effet aléatoire du site sur l'ordonnée à l'origine pour prendre en compte cette variabilité, sans pour autant l'estimer.

Le modèle le plus complet qui répond à cette hypothèse est le suivant :

$$\log(\lambda_i) = \log(aire_i) + \mu_0 + \gamma_{i,j} + \alpha_0 * Cov_{2004,i} + \alpha_1 * Cov_{2005,i} + \alpha_2 * Cov_{2006,i} + \beta * prev_i, \text{ avec } i = 1, \dots, 30 \text{ et } j = 1, \dots, 10$$

Il est possible que seulement l'une de ces variables ou des combinaisons deux par deux expliquent le mieux la variabilité du nombre de carapaces récentes. Nous allons donc ajuster différents modèles qui prennent en compte ces trois variables individuellement ou en les combinant pour étudier uniquement l'effet de l'année, l'effet de l'année et du site, ou l'effet de l'année et de la prévalence. Dans les modèles avec la prévalence comme variable continue, celle-ci est standardisée.

Pour répondre à notre deuxième question concernant l'influence des sites à haute prévalence par rapport aux sites à faible prévalence, nous allons étudier trois modèles utilisant la prévalence en tant que covariable

discontinue H , avec $H = \begin{cases} 1 & \text{si prevalence} > 25 \\ 0 & \text{sinon} \end{cases}$.

Le premier avec l'effet de la prévalence uniquement, un deuxième avec l'effet de l'année et de la prévalence, et un troisième avec l'effet de l'année, de la prévalence et un effet aléatoire du site sur l'ordonnée à l'origine.

Pour l'ensemble des modèles, nous utilisons comme priors les distributions suivantes :

- $\mu_0 \sim \mathcal{N}(0, 100)$, avec μ_0 l'ordonnée à l'origine
- $\beta \sim \mathcal{N}(0, 100)$, avec β la pente de regression associée à la prévalence
- $\alpha_0 \sim \mathcal{N}(0, 100)$, avec α_0 le coefficient associé à l'année 2004
- $\alpha_0 \sim \mathcal{N}(0, 100)$, avec α_0 le coefficient associé à l'effet spécifique de l'année 2004
- $\alpha_1 \sim \mathcal{N}(0, 100)$, avec α_1 le coefficient associé à l'effet spécifique de l'année 2005
- $\alpha_2 \sim \mathcal{N}(0, 100)$, avec α_2 le coefficient associé à l'effet spécifique de l'année 2006
- $\gamma_{i,j} \sim \mathcal{N}(0, \tau)$, avec $\gamma_{i,j}$ le coefficient spécifique sur l'intercept associé à l'effet du $j^{\text{ème}}$ site pour la $i^{\text{ème}}$ observation, avec $\tau = \sigma_s^2$
- $\sigma_s \sim \mathcal{U}(0, 100)$

Nous considérons ici les modèles suivants:

Tableau 1 : Modèles construits et testés

Modèles	Formules
Continus	
M1	$\log(\lambda) = \log(aire_i) + \mu_0 + \beta * prev_i$
M2	$\log(\lambda) = \log(aire_i) + \mu_0$
M3	$\log(\lambda) = \log(aire_i) + \mu_0 + \alpha_0 * Cov_{2004,i} + \alpha_1 * Cov_{2005,i} + \alpha_2 * Cov_{2006,i}$
M4	$\log(\lambda) = \log(aire_i) + \mu_0 + \alpha_0 * Cov_{2004,i} + \alpha_1 * Cov_{2005,i} + \alpha_2 * Cov_{2006,i} + \beta * prev_i$
M5	$\log(\lambda) = \log(aire_i) + \mu_0 + \alpha_0 * Cov_{2004,i} + \alpha_1 * Cov_{2005,i} + \alpha_2 * Cov_{2006,i} + \gamma_{i,j}$
M6	$\log(\lambda) = \log(aire_i) + \mu_0 + \alpha_0 * Cov_{2004,i} + \alpha_1 * Cov_{2005,i} + \alpha_2 * Cov_{2006,i} + \gamma_{i,j} + \beta * prev_i$
Discontinus	
M7	$\log(\lambda) = \log(aire_i) + \mu_0 + \alpha_0 * Cov_{2004,i} + \alpha_1 * Cov_{2005,i} + \alpha_2 * Cov_{2006,i} + \gamma_{i,j} + \beta * H_i$
M8	$\log(\lambda) = \log(aire_i) + \mu_0 + \alpha_0 * Cov_{2004,i} + \alpha_1 * Cov_{2005,i} + \alpha_2 * Cov_{2006,i} + \beta * H_i$
M9	$\log(\lambda) = \log(aire_i) + \mu_0 + \alpha_0 * Cov_{2004,i} + \alpha_1 * Cov_{2005,i} + \alpha_2 * Cov_{2006,i} + \beta * H_i + \gamma_{i,j}$

Les DICs de ces modèles sont les suivants : Tableau 2 : DICs des modèles

##	DICs
## Continus	
## M1	93.8217768508141
## M2	105.645128880156
## M3	119.757702054292
## M4	91.0657159892484
## M5	94.3737399231717
## M6	86.6294258915715
## Discontinus	
## M7	112.329866344267
## M8	96.7037156588047
## M9	91.1192586167698

4 Inférence et interprétation des résultats

Sur la base du meilleur modèle, donnez les estimations des paramètres ainsi qu'une mesure de l'incertitude associée. Interprétez vos résultats. ## Prévalence continue | D'après le tableau des DIC, le modèle 6 est le meilleur avec les résultats de simulations ci-dessous.

```
## Inference for Bugs model at "/var/folders/l9/dzz054qd4_q8g63vvgsts36w0000gn/T//Rtmp0sfhvx/modelba155"
## 2 chains, each with 9000 iterations (first 4500 discarded)
## n.sims = 9000 iterations saved
##      mu.vect sd.vect      2.5%      25%      50%      75%      97.5%      Rhat      n.eff
## alpha_0    0.296   5.631 -10.919 -3.486   0.336   4.057  11.317  1.001   9000
## alpha_1   -0.387   5.625 -11.608 -4.175 -0.350   3.381  10.655  1.001   9000
## alpha_2   -0.121   5.632 -11.300 -3.929 -0.049   3.665  10.894  1.001   9000
## b.prev     0.579   0.172   0.251   0.464   0.576   0.685   0.933  1.007   5900
## mu.0      -3.048   5.629 -14.068 -6.811 -3.116   0.739   8.227  1.001   9000
## sd.s       0.356   0.286   0.011   0.161   0.304   0.474   1.075  1.021   2400
## deviance  79.785   3.700  73.999  77.256  79.272  81.750  88.387  1.001   3000
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 6.8 and DIC = 86.6
## DIC is an estimate of expected predictive error (lower deviance is better).
```

D'après la statistique de Gelman-Rubin \hat{R} qui est inférieure à 1.1 pour chaque paramètre estimé, les chaînes convergent bien. Le `n.eff` de chaque paramètre est supérieur à 100.

Nous avons donc des estimations de nos paramètres qui sont stables et des chaînes peu autocorrélées.

Regardons maintenant, la distribution *a posteriori* et la moyenne du nombre de carapaces à partir des

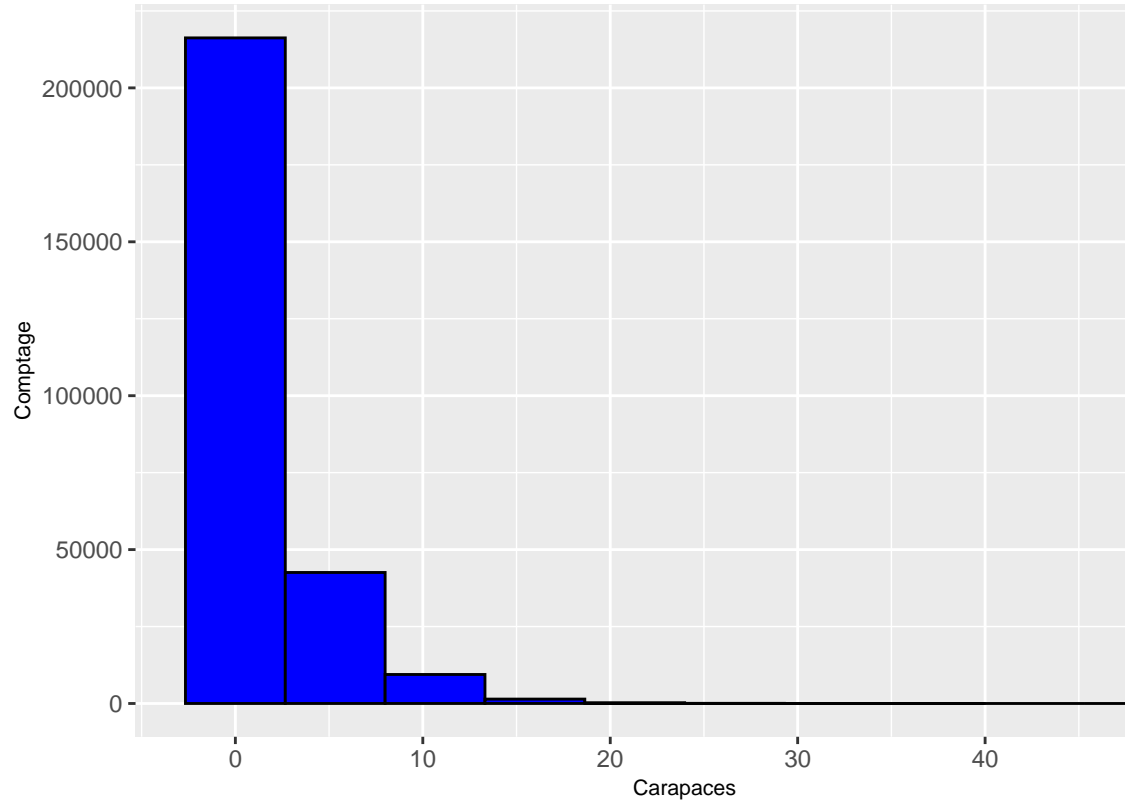


Figure 7 : Distribution du nombre de carapaces pour le modèle 6

paramètres estimés par M6.

[1] 1.8085

La moyenne a posteriori pour μ_0 est de -2.99 et de β est de 0.57. Nous avons, toutes choses égales par ailleurs, une augmentation de $5 * e^{-2.99-0.42+\gamma} = 0.19$ carapaces (avec $\gamma \sim \mathcal{N}(0, 0.32^2)$) à $5 * e^{-2.99+5*0.57-0.42+\gamma} = 3$ carapaces (avec $\gamma \sim \mathcal{N}(0, 0.32^2)$) pour une augmentation de 0% à 5% de la prévalence pour un site d'aire 5 unités de surface, l'année 2005. La moyenne a posteriori de l'écart-type pour l'effet aléatoire du site est de $\sigma_s = 0.32$, soit un effet plutôt faible. Aucun des intervalles de confiance ne contiennent 0, à part ceux pour le paramètre **année**. Ainsi, à part l'année qui pourrait ne pas avoir d'effet, la prévalence explique bien les variations du nombre de carapaces.

4.1 Prévalence discontinue

D'après le tableau des DIC, le modèle 9 est le meilleur modèle avec la prévalence discontinue. Les résultats du modèle sont ci-dessous.

```
## Inference for Bugs model at "/var/folders/l9/dzz054qd4_q8g63vvgsts36w0000gn/T//Rtmp0sfhvx/modelba151
## 2 chains, each with 9000 iterations (first 4500 discarded)
## n.sims = 9000 iterations saved
##      mu.vect sd.vect   2.5%   25%   50%   75%  97.5%  Rhat  n.eff
## alpha_0  -0.578   5.035 -10.435 -4.095 -0.572  2.891  9.247 1.001  9000
## alpha_1  -1.247   5.038 -11.092 -4.725 -1.237  2.236  8.564 1.001  9000
## alpha_2  -0.951   5.039 -10.770 -4.428 -0.953  2.493  8.869 1.001  9000
## b.prev    1.113   0.443   0.241  0.822  1.109  1.395  2.008 1.005   350
## mu.0     -2.636   5.036 -12.449 -6.097 -2.652  0.837  7.241 1.001  9000
```

```
## sd.s      0.552  0.288  0.093  0.359  0.506  0.699  1.233 1.009  7900
## deviance 80.804  4.543 73.837 77.494 80.048 83.492 91.182 1.001  2600
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 10.3 and DIC = 91.1
## DIC is an estimate of expected predictive error (lower deviance is better).
```

D'après la statistique de Gelman-Rubin \hat{R} qui est inférieure à 1.1 pour chaque paramètre estimé, les chaînes convergent bien. Le `n.eff` de chaque paramètre est supérieur à 100.

Nous avons donc des estimations de nos paramètres qui sont stables et des chaînes peu autocorrélées.

Regardons maintenant, la distribution *a posteriori* et la moyenne du nombre de carapaces à partir des

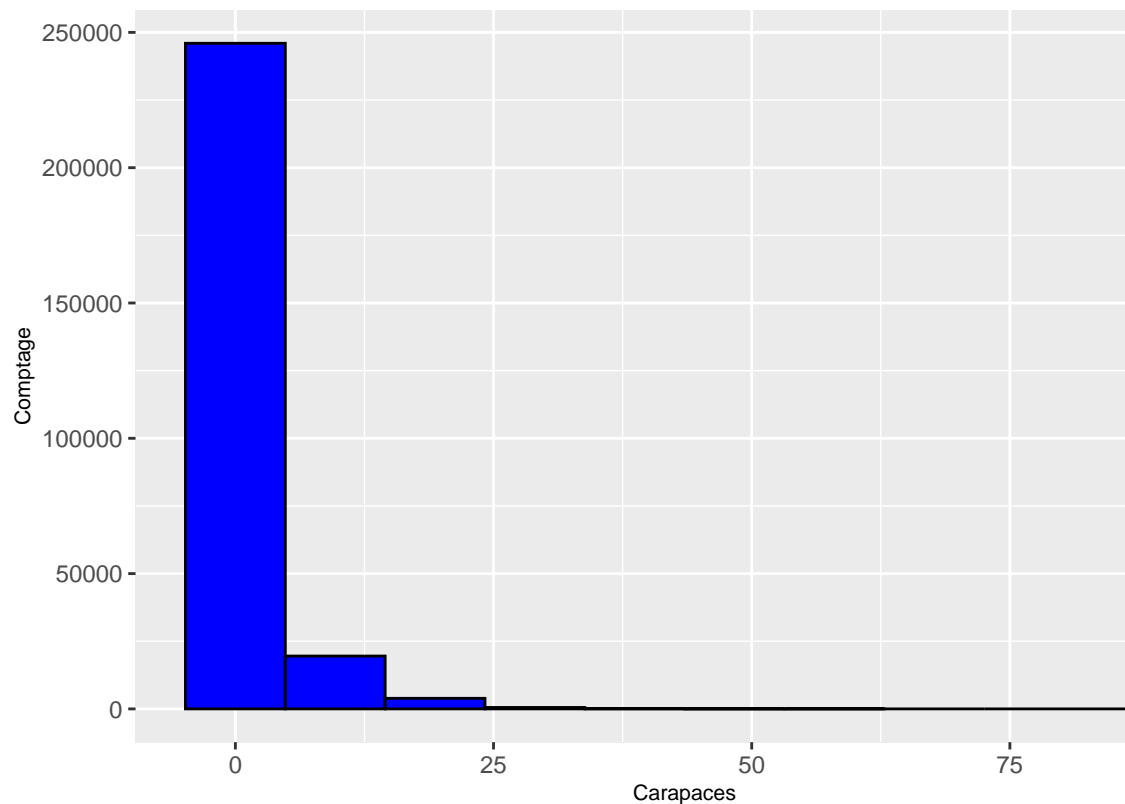


Figure 8 : Distribution du nombre de carapaces pour le modèle 9

paramètres estimés par M9.

```
## [1] 1.966844
```

La moyenne *a posteriori* pour β est de 1.16 ce qui correspond, toutes choses égales par ailleurs à une augmentation moyenne de $e^{1.16} = 3.2$ carapaces pour les sites à haute prévalence. La moyenne *a posteriori* de l'écart-type pour l'effet aléatoire du site est de $\sigma_s = 0.58$, soit un effet plus marqué que pour M6 mais qui demeure relativement faible. Aucun des intervalles de confiance ne contiennent 0, à part ceux pour le paramètre **année**. Ainsi, à part l'année qui pourrait ne pas avoir d'effet, il y aurait bien une différence entre les sites à haute et faible prévalence.

5 Discussion

Comme dans l'article, nous avons eu des DIC proches entre **M6** avec la prévalence continue et **M9** avec la prévalence discontinue. Nous avons également eu des résultats semblables à ceux de l'article en ce qui concerne l'interprétation. Cependant, les estimateurs de nos paramètres sont plus élevés que ceux de l'article. Par exemple, pour le modèle 6, nous avons $\mu_0 = -2.9 > a_0 = -4.8$ et $\beta = 0.57 > a_1 = 0.02$. Par contre, notre estimation de l'écart-type pour l'effet aléatoire du site est similaire (autour de 0.30).

Selon nous, plusieurs choses pourraient éventuellement expliquer ces différences :

- le choix de l'offset : nous avons pris l'aire des sites comme offset, alors que dans l'article ils utilisent le nombre de tortues totales par site.
- des différences d'implémentation de la covariable **année** car nous ne mettons pas l'année en effet aléatoire
- le choix des priors pour les effets aléatoires; car les auteurs ont testé deux implémentations possibles et nous ne savons pas s'ils ont retenu celle où la variance inverse suit une loi uniforme ou celle où l'écart-type suit une loi uniforme.

Eventuellement, mais moins probable selon nous, le f