Introduction à l'inférence bayésienne

Robin J. Ryder

CEREMADE – Université Paris-Dauphine PSL

21 décembre 2023

Plan

Cadre général de la statistique

En statistique, on cherche généralement à inférer de l'information sur un paramètre inconnu θ. Par exemple, on observe $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$ et on souhaite :

- Obtenir une estimation (ponctuelle) $\hat{\theta}$ de θ, par ex. $\hat{\theta} = 1.3$.

Cadre général de la statistique

En statistique, on cherche généralement à inférer de l'information sur un paramètre inconnu θ. Par exemple, on observe $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$ et on souhaite :

- Obtenir une estimation (ponctuelle) $\hat{\theta}$ de θ, par ex. $\hat{\theta} = 1.3$.
- Mesurer l'incertitude de notre estimateur, à l'aide d'un intervalle ou d'une région de valeurs plausibles, par ex. $[0.9, 1.5]$ est un intervalle de confiance à 95% pour θ.

Cadre général de la statistique

En statistique, on cherche généralement à inférer de l'information sur un paramètre inconnu θ. Par exemple, on observe $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$ et on souhaite :

- Obtenir une estimation (ponctuelle) $\hat{\theta}$ de θ, par ex. $\hat{\theta} = 1.3$.
- Mesurer l'incertitude de notre estimateur, à l'aide d'un intervalle ou d'une région de valeurs plausibles, par ex. $[0.9, 1.5]$ est un intervalle de confiance à 95% pour θ.
- Faire du choix de modèle / du test d'hypothèses, par ex. décider entre $H_0 : \theta = 0$ et $H_1 : \theta \neq 0$ ou entre $H_0 : X_i \sim \mathcal{N}(\theta, 1)$ et $H_1 : X_i \sim \mathcal{E}(\theta)$.

Cadre général de la statistique

En statistique, on cherche généralement à inférer de l'information sur un paramètre inconnu θ. Par exemple, on observe $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$ et on souhaite :

- Obtenir une estimation (ponctuelle) $\hat{\theta}$ de θ, par ex. $\hat{\theta} = 1.3$.
- Mesurer l'incertitude de notre estimateur, à l'aide d'un intervalle ou d'une région de valeurs plausibles, par ex. $[0.9, 1.5]$ est un intervalle de confiance à 95% pour θ.
- Faire du choix de modèle / du test d'hypothèses, par ex. décider entre $H_0 : \theta = 0$ et $H_1 : \theta \neq 0$ ou entre $H_0 : X_i \sim \mathcal{N}(\theta, 1)$ et $H_1 : X_i \sim \mathcal{E}(\theta)$.
- Utiliser cette inférence dans du post-traitement : prédiction, décision, entrée d'un autre modèle...

Pourquoi être bayésien ?

Certains domaines d'applications utilisent beaucoup l'inférence bayésienne, parce que :

- Les modèles sont complexes
- L'estimation de l'incertitude est essentielle
- La sortie d'un modèle est utilisée comme entrée d'un autre
- On s'intéresse à une fonction complexe de nos paramètres

Statistique fréquentiste

- L'inférence statistique cherche à estimer un paramètre inconnu θ au vu de données D.
- Dans le cadre fréquentiste, θ a une vraie valeur fixe (déterministe, inconnue).
- L'incertitude est mesurée par des intervalles de confiance, dont l'interprétation n'est pas intuitive : si j'obtiens un IC à 95% de $[80\ ;\ 120]$ (i.e. 100 ± 20) pour θ, je ne peux pas dire que θ appartient à l'intervalle $[80\ ;\ 120]$ avec probabilité 95%.

Statistique fréquentiste

- L'inférence statistique cherche à estimer un paramètre inconnu θ au vu de données D.
- Dans le cadre fréquentiste, θ a une vraie valeur fixe (déterministe, inconnue).
- L'incertitude est mesurée par des intervalles de confiance, dont l'interprétation n'est pas intuitive : si j'obtiens un IC à 95% de $[80\ ;\ 120]$ (i.e. 100 ± 20) pour θ, je ne peux pas dire que θ appartient à l'intervalle $[80\ ;\ 120]$ avec probabilité 95%.
- En statistique fréquentiste, on utilise souvent l'estimateur au maximum de vraisemblance : quelle valeur de θ rend les données les plus plausibles (selon notre modèle) ?

$$L(\theta|D) = P[D|\theta]$$

$$\hat{\theta} = \arg\max_{\theta} L(\theta|D)$$

Règle de Bayes

Rappelons la règle de Bayes : pour deux événements A et B, on a

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A] \cdot \mathbb{P}[A]}{\mathbb{P}[B]}.$$

Règle de Bayes

Rappelons la règle de Bayes : pour deux événements A et B, on a

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A] \cdot \mathbb{P}[A]}{\mathbb{P}[B]}.$$

Autre écriture, avec des densités marginales et conditionnelles :

$$\pi(y|x) = \frac{\pi(x|y)\pi(y)}{\pi(x)}.$$

Statistique bayésienne

- Dans le cadre bayésien, le paramètre θ est vu comme étant intrinsèquement aléatoire : il a une loi.
- Avant de voir les données, j'ai une loi *a priori* $\pi(\theta)$, généralement peu informative.
- Une fois que je prends les données en compte, j'obtiens une loi *a posteriori*, dont on peut espérer qu'elle est plus informative.
  Par la règle de Bayes,

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}.$$

Statistique bayésienne

- Dans le cadre bayésien, le paramètre θ est vu comme étant intrinsèquement aléatoire : il a une loi.
- Avant de voir les données, j'ai une loi *a priori* $\pi(\theta)$, généralement peu informative.
- Une fois que je prends les données en compte, j'obtiens une loi *a posteriori*, dont on peut espérer qu'elle est plus informative.
  Par la règle de Bayes,

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}.$$

  Par définition, $\pi(D|\theta) = L(\theta|D)$. La quantité $\pi(D)$ est une constante de normalisation ne dépendant pas de θ : généralement, on ne l'inclut pas et on écrit

$$\pi(\theta|D) \propto \pi(\theta)L(\theta|D).$$

Statistique bayésienne

$$\pi(\theta|D) \propto \pi(\theta)L(\theta|D)$$

- Différente personnes ont des lois a priori différentes, donc des lois a posteriori différentes. Mais avec assez de données, le choix de la loi a priori a peu d'impact.
- Il est maintenant légitime de faire des assertions probabilistes sur θ, comme "il y a une probabilité de 95% que θ appartienne à l'intervalle [78 ; 119]" (intervalle de crédibilité).

Avantages et inconvénients de la statistique bayésienne

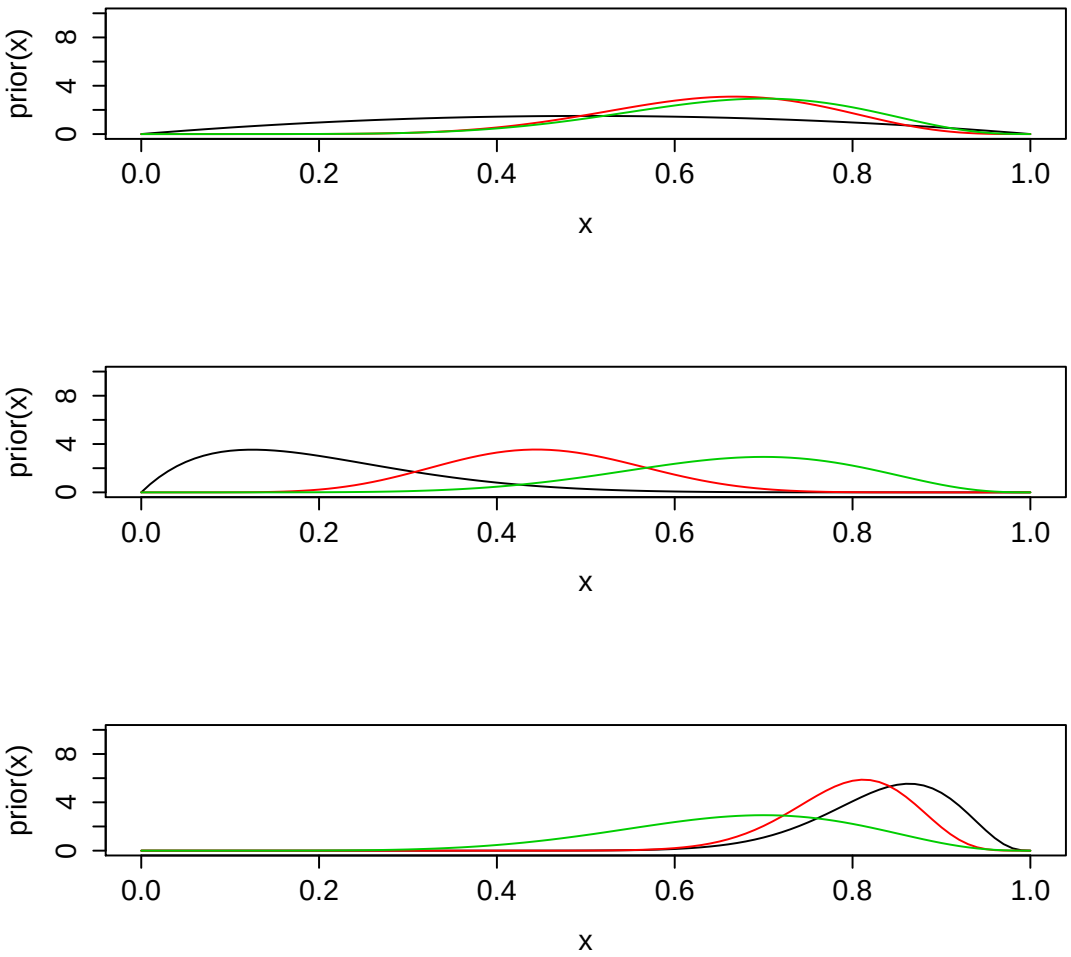
- Interprétation plus intuitive des résultats
- Plus facile de réfléchir sur l'incertitude
- Dans un cadre hiérarchique, on peut mieux prendre en compte toutes les sources de variation
- Spécification de la loi a priori : il faut vérifier que changer de loi a priori ne change pas le résultat
- Coût computationnel

Plan

Effet de la prior

Exemple : modèle de Bernoulli. θ=probabilité de succès.
On observe $S_n = 72$ succès en $n = 100$ tentatives.
Estimation fréquentiste : $\hat{\theta} = 0.72$
IC à 95% : [0.63 0.81].
Estimation bayésienne : dépend de la prior.

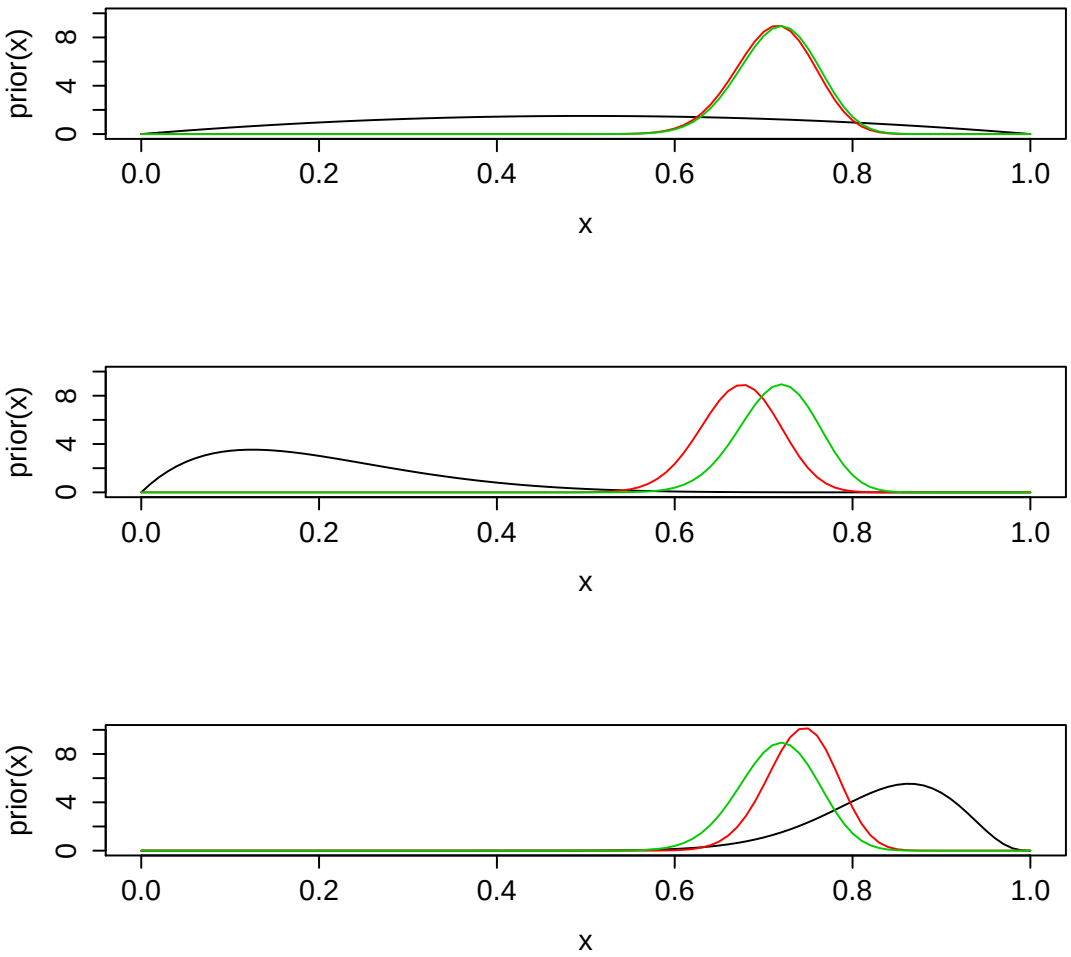
Effet de la prior ($n = 10$)



$S_n = 7,\ n = 10$

Noir : prior ; vert : vraisemblance, rouge : posterior.

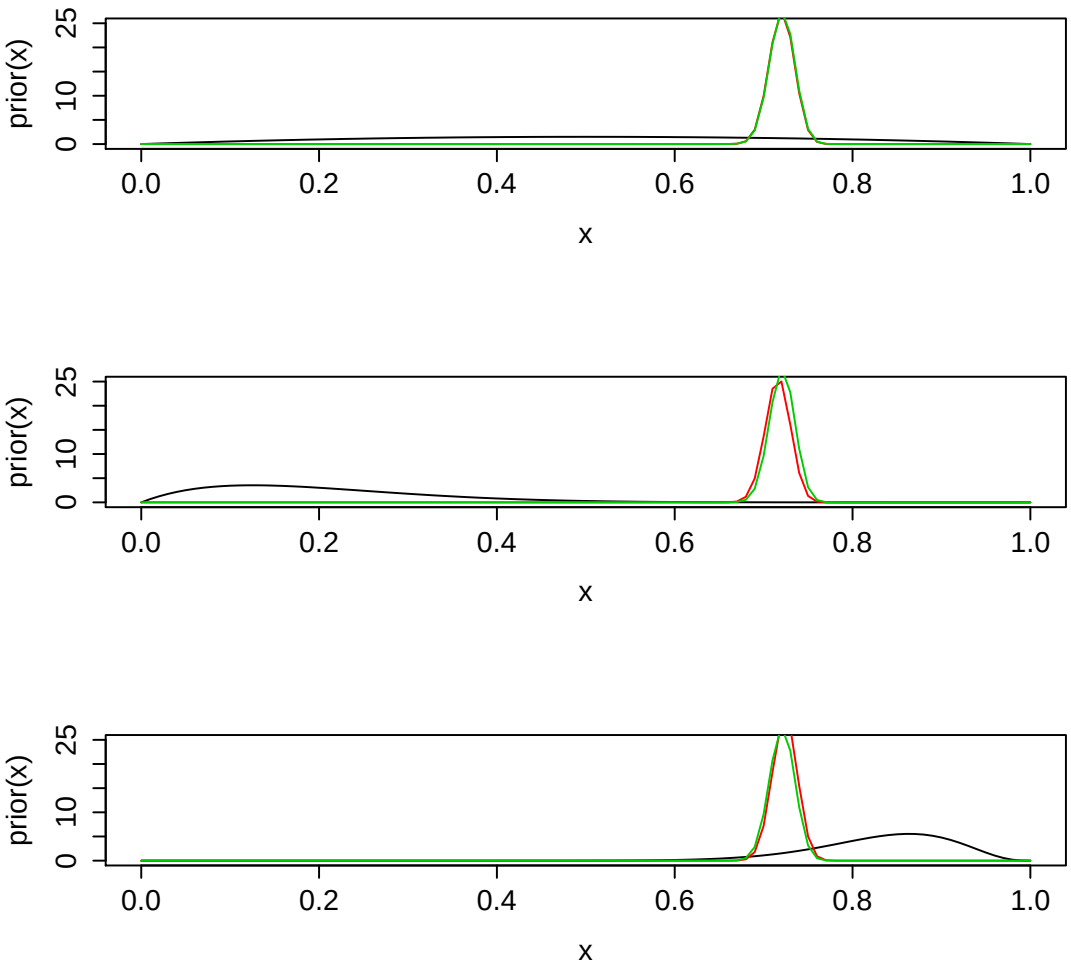
Effet de la prior ($n = 100$)



$S_n = 72,\ n = 100$

Noir : prior ; vert : vraisemblance, rouge : posterior.

Effet de la prior ($n = 1000$)



$S_n = 721,\ n = 1000$

Noir : prior ; vert : vraisemblance, rouge : posterior.

Choix de la prior

Le choix de la loi a priori peut avoir un impact important, surtout si les données sont de taille petite ou modérée. Comment choisir la prior ?

- Avis d'experts de l'application
- Expérience antérieure
- Prior conjuguée, donc pratique mathématiquement, avec des moments choisi par des experts
- Prior non-informative
- ...

Choix de la prior

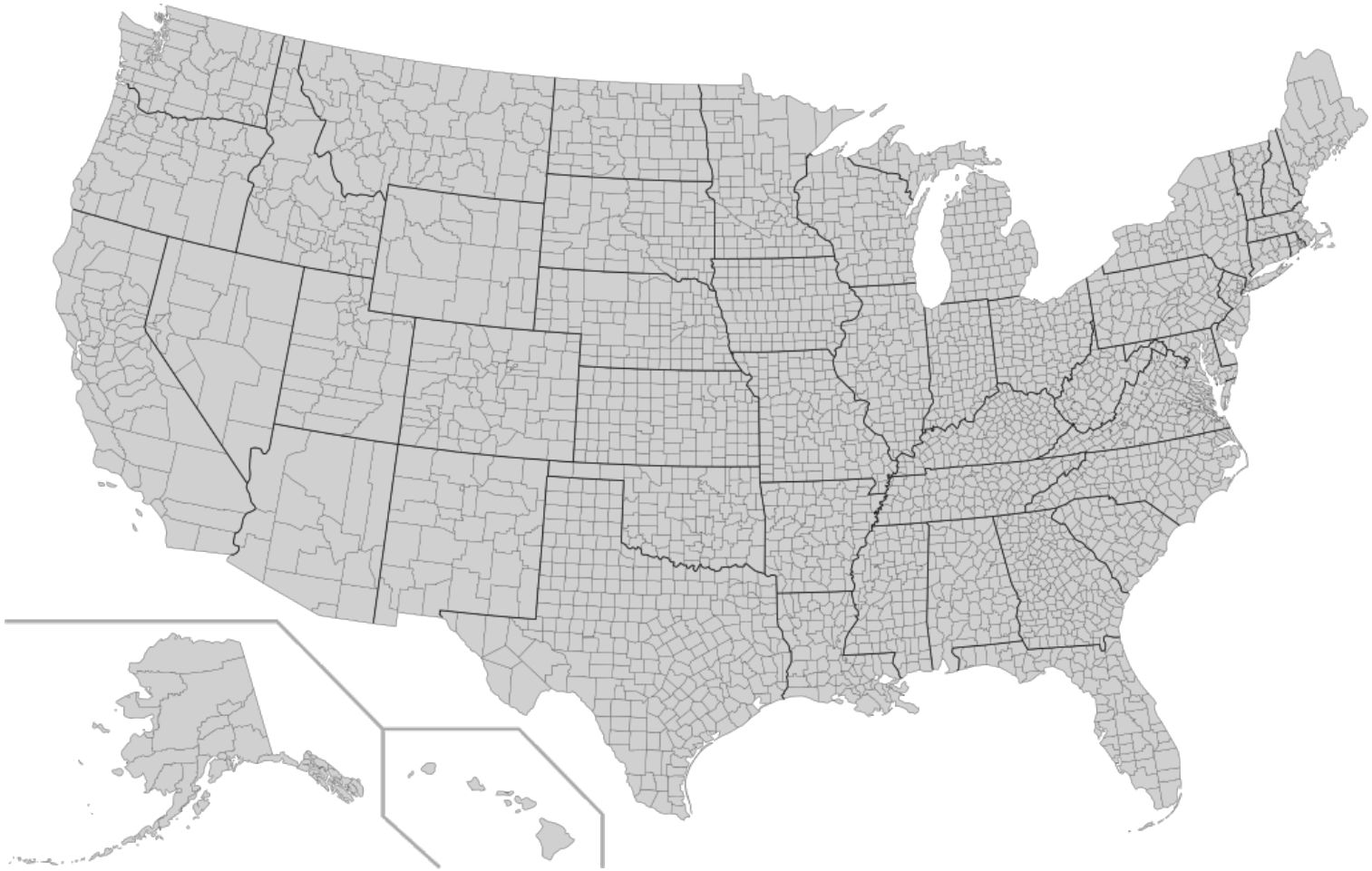
Le choix de la loi a priori peut avoir un impact important, surtout si les données sont de taille petite ou modérée. Comment choisir la prior ?

- Avis d'experts de l'application
- Expérience antérieure
- Prior conjuguée, donc pratique mathématiquement, avec des moments choisi par des experts
- Prior non-informative
- ...

Dans tous les cas, le mieux est d'essayer différentes priors, et de voir si les posteriors s'accordent : les données sont-elles suffisantes pour mettre d'accord des experts qui étaient en désaccord a priori ?

Plan

Exemple : Géographie du cancer du rein aux États-Unis



Estimation fréquentiste

On s'intéresse à la variation géographique du taux de cancer du rein aux États-Unis.
Modélisation intuitive : dans le comté j ($j = 1, \ldots, 3242$), l'incidence du cancer est θ_j. On observe $k_j \sim Poisson(\theta_j N_j)$ cancers parmi une population de taille N_j.
On estime par maximum de vraisemblance $\hat{\theta}_j = \frac{k_j}{N_j}$.
On s'intéresse aux comtés j tel que $\hat{\theta}_j$ est élevé.

Estimation fréquentiste

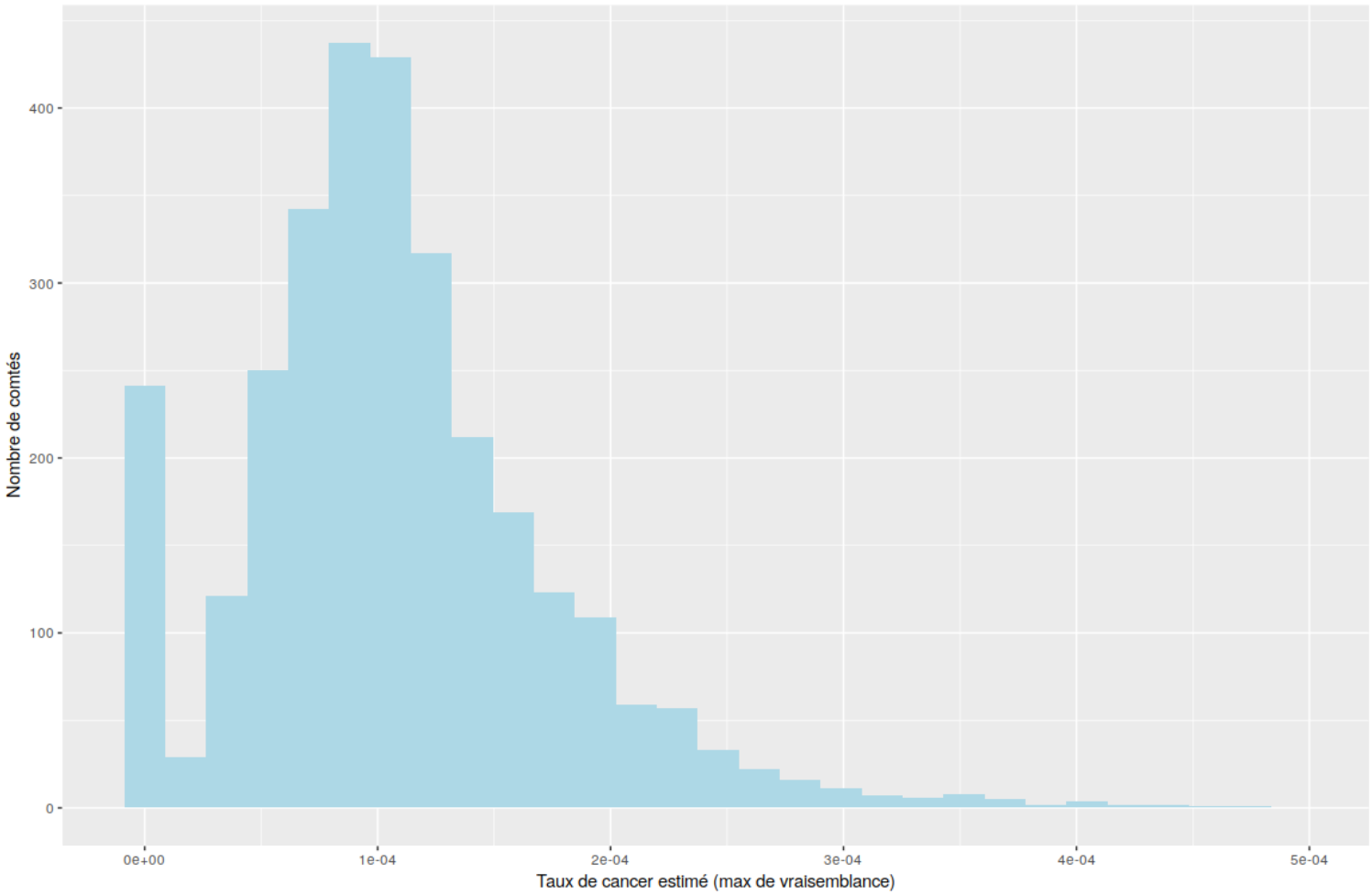


FIGURE – Histogramme des valeurs de $\hat{\theta}_j$.

Comtés avec beaucoup de cancers

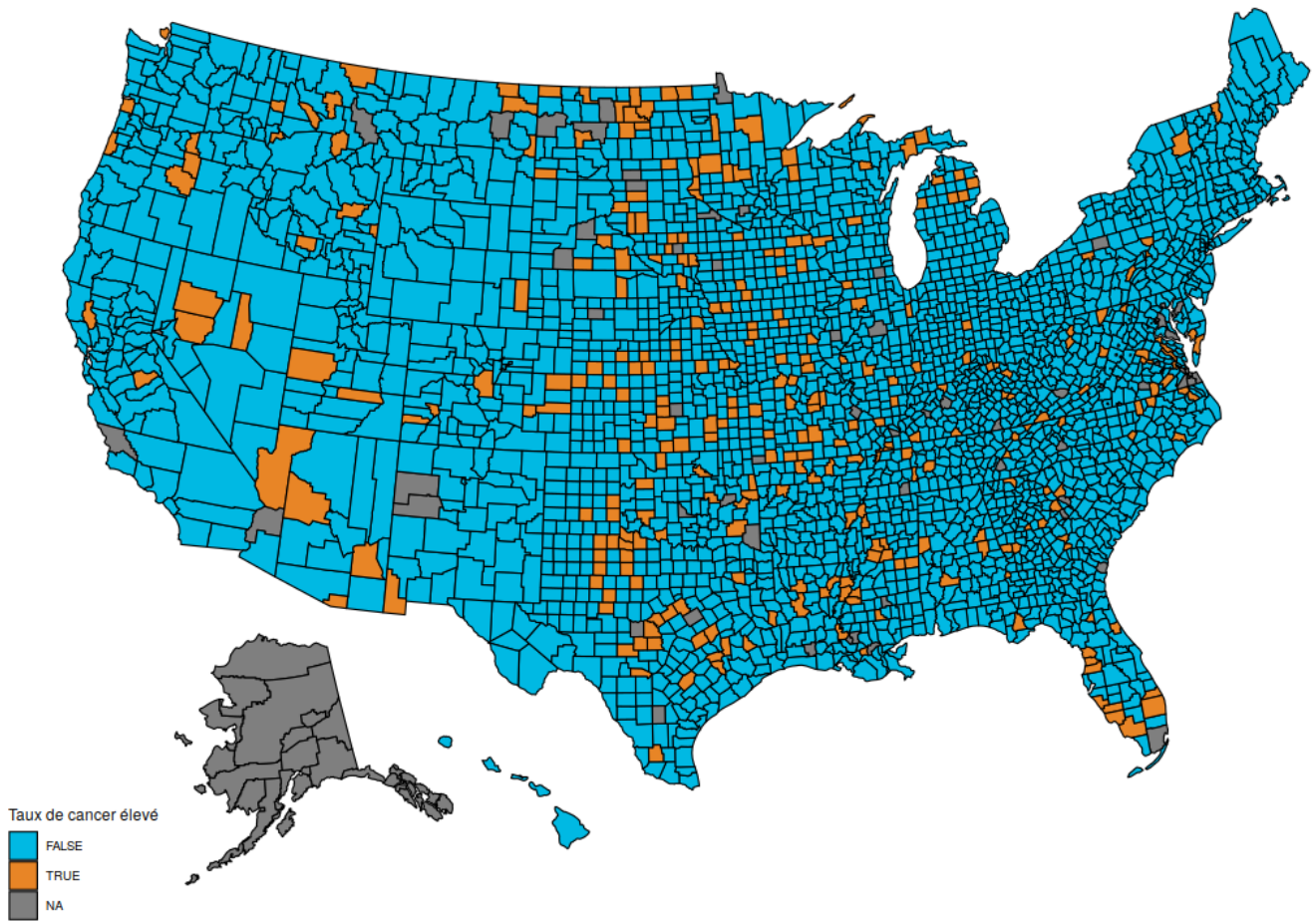


FIGURE – Comtés des États-Unis dont le taux de cancer du rein est dans le dernier décile (taux élevé).

Comtés avec peu de cancers

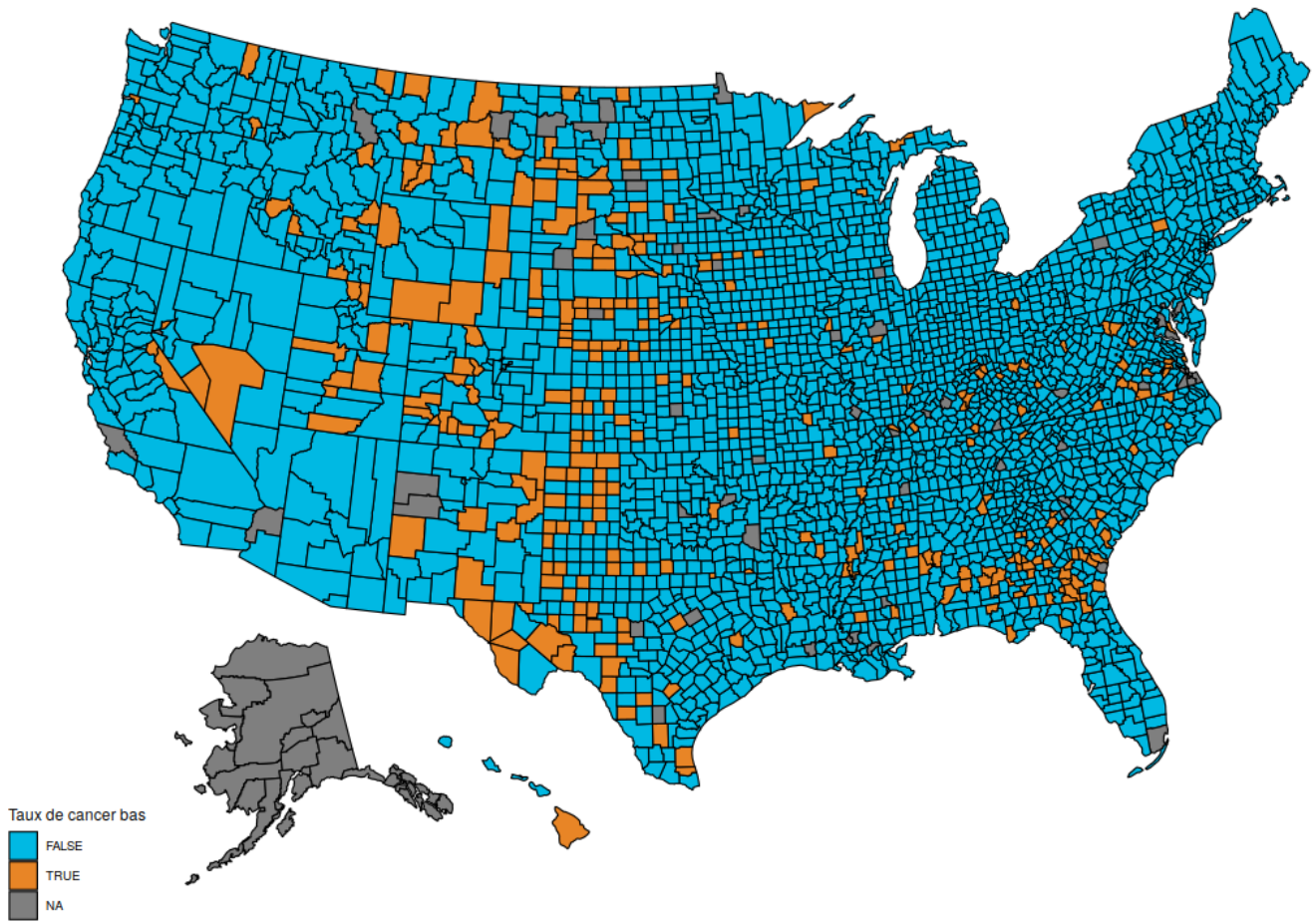


FIGURE – Comtés des États-Unis dont le taux de cancer du rein est dans le premier décile (taux bas).

Tous les comtés extrémaux sont des comtés de faible population (N_j faible).
Le taux moyen de cancer du rein est de 0.0001 en 10 ans.
Prenons un comté à 1000 habitants. Il a de fortes chances de n'avoir aucun cancer du rein : $k_j = 0$, $\hat{\theta}_j = 0$: il est dans le premier décile.
Si par hasard il a 1 cancer du rein, $k_j = 1$, $\hat{\theta}_j = 0.001$: il est dans le dernier décile.

Inférence bayésienne

Mettons une loi a priori $\theta_j \sim \Gamma(\alpha, \beta)$.
Des valeurs possibles sont $\alpha = 10$, $\beta = 100\,000$, avec alors une espérance a priori de $1 \cdot 10^{-4}$ et un écart-type de $3 \cdot 10^{-5}$.
A posteriori, l'espérance est de :

- si $N_j = 1000$ et $k_j = 0$, $E[\theta_j | k_j] = 0.99 \cdot 10^{-4}$
- si $N_j = 1000$ et $k_j = 1$, $E[\theta_j | k_j] = 1.1 \cdot 10^{-4}$
- si $N_j = 10^6$ et $k_j = 50$, $E[\theta_j | k_j] = 0.55 \cdot 10^{-4}$
- si $N_j = 10^6$ et $k_j = 200$, $E[\theta_j | k_j] = 1.9 \cdot 10^{-4}$

Histogramme des espérances a posteriori

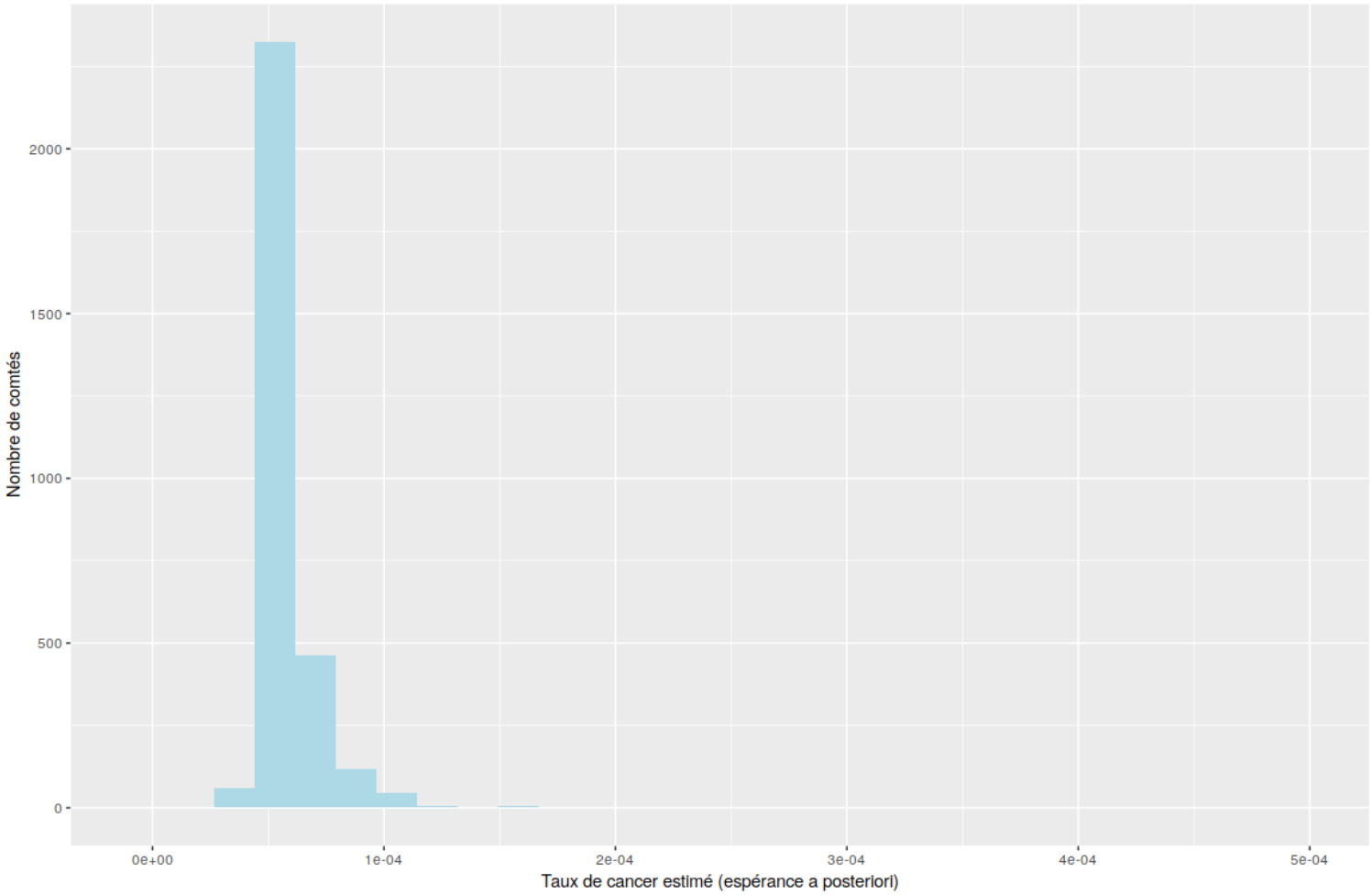


FIGURE – Histogramme des valeurs de $E[\theta_j|k_j]$.

Carte des espérances a posteriori

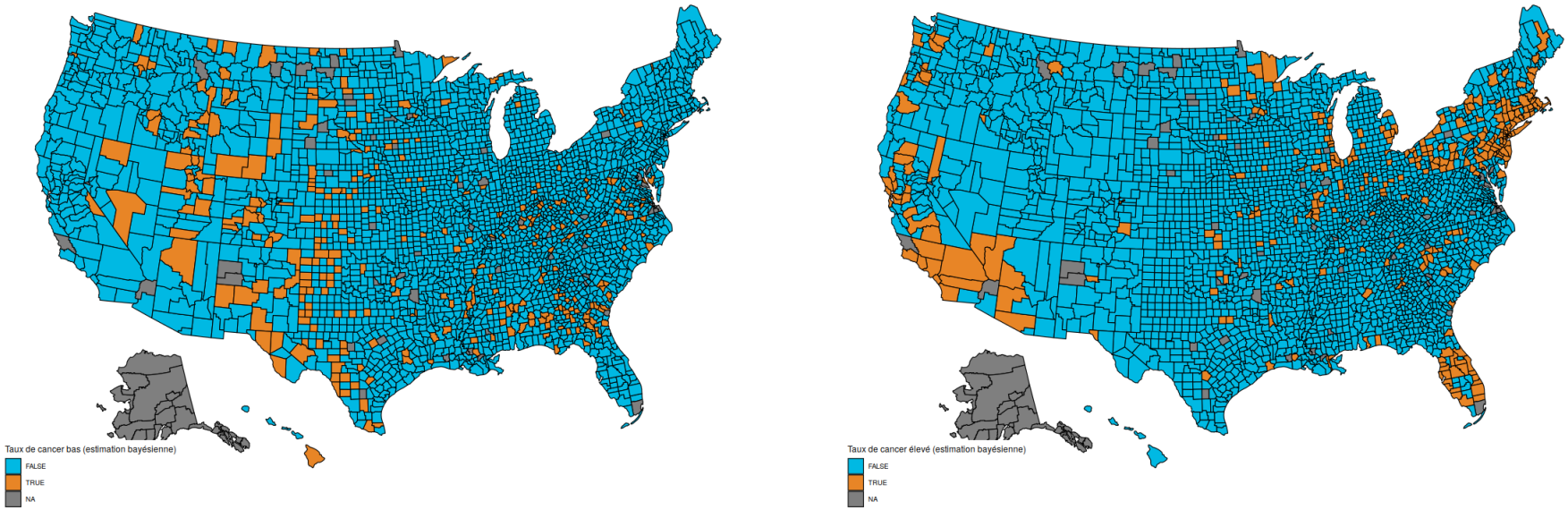


FIGURE – Comtés des États-Unis dont le taux de cancer du rein est dans le premier/dernier décile en espérance a posteriori.

Pourquoi être bayésien ?

- Information a priori disponible
- Les données doivent prendre le dessus uniquement quand elles sont suffisamment nombreuses
- Utilisation de la mesure de l'incertitude pour des prises de décision.

Plan

Exemple : Vol 447 d'Air France

Cette section et ses figures sont tirées de Stone et al. (Statistical Science 2014).

Le vol Air France 447 a disparu au dessus de l'Atlantique le 1erjuin 2009, en route de Rio de Janeiro vers Paris ; les 228 personnes à bord sont décédées.

Les trois premières tentatives de récupérer l'épave et les boîtes noires ont été infructueuses.

En 2011, une quatrième tentative a été lancée, basée sur une recherche bayésienne.

Exemple : Vol 447 d'Air France

Cette section et ses figures sont tirées de Stone et al. (Statistical Science 2014).
Le vol Air France 447 a disparu au dessus de l'Atlantique le 1erjuin 2009, en route de Rio de Janeiro vers Paris ; les 228 personnes à bord sont décédées.
Les trois premières tentatives de récupérer l'épave et les boîtes noires ont été infructueuses.
En 2011, une quatrième tentative a été lancée, basée sur une recherche bayésienne.

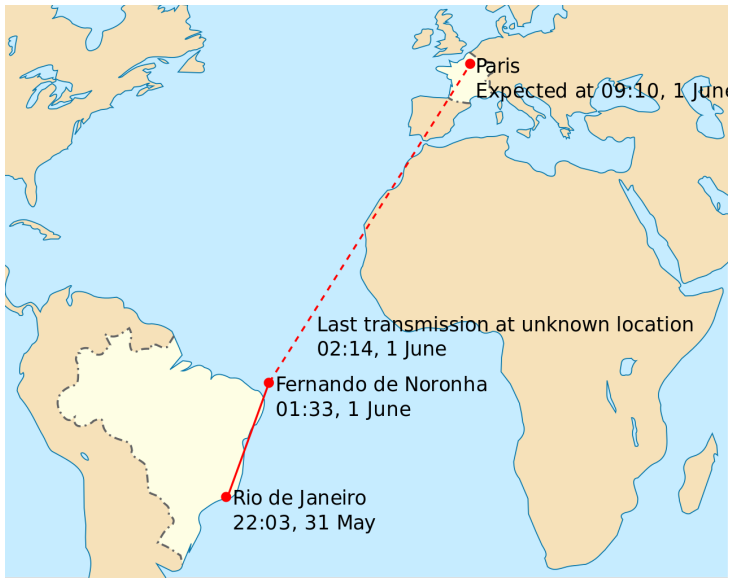
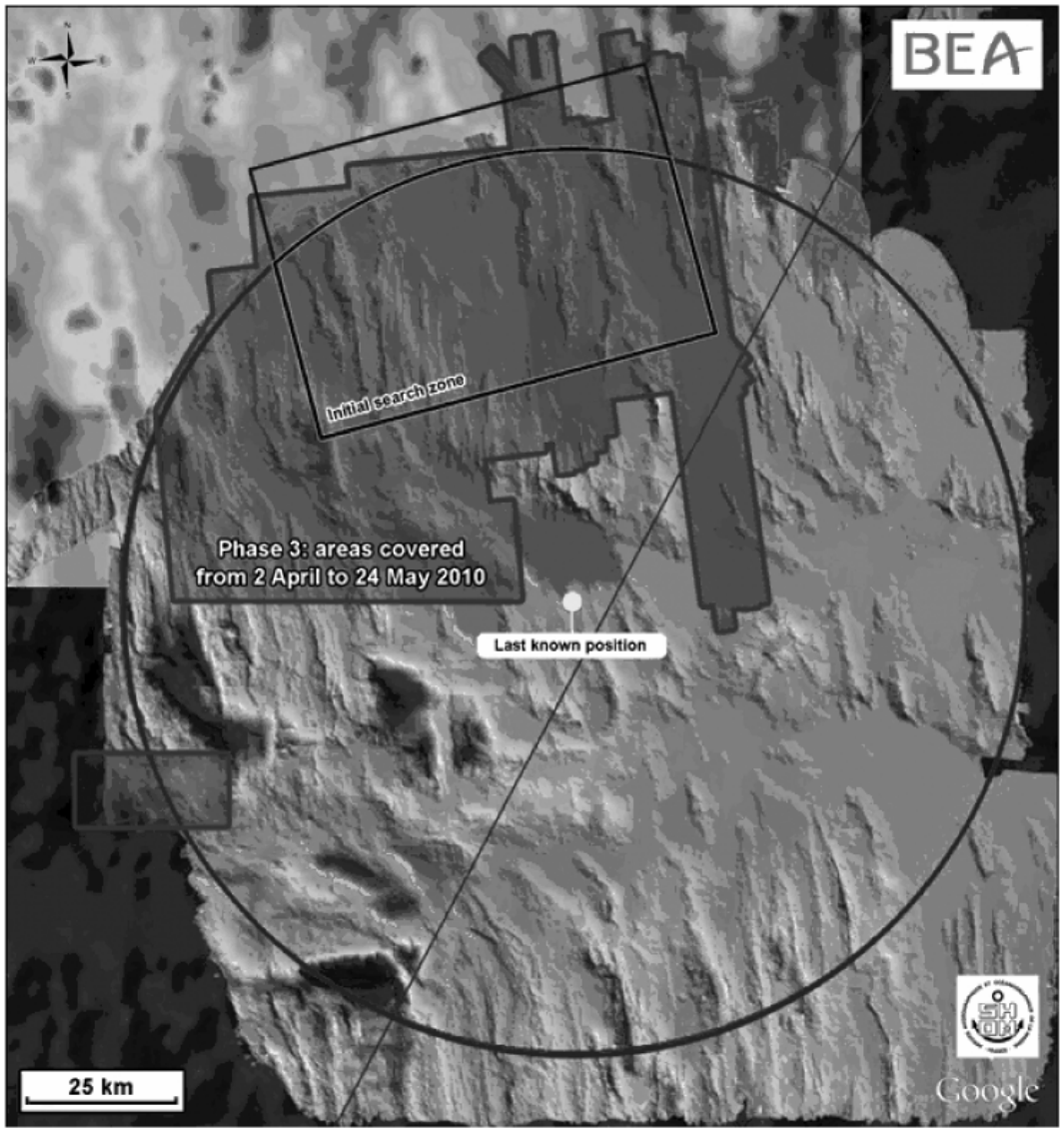


FIGURE – Route du vol. Image de Mysid, Domaine public.

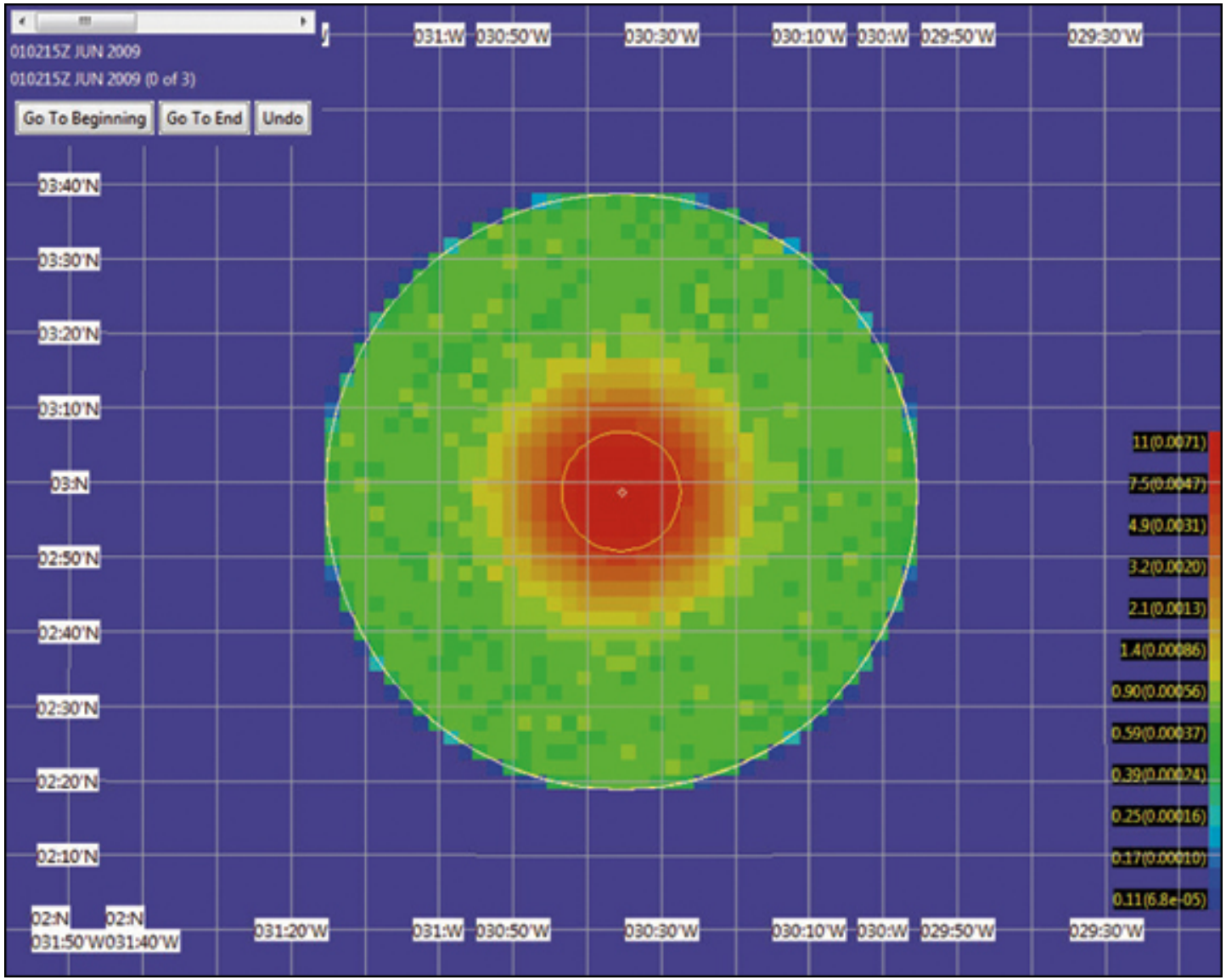
Pourquoi être bayésien ?

- Nombreuses sources d'incertitude
- Probabilités subjectives
- L'objet d'intérêt est une loi de probabilité
- Le formalisme fréquentiste ne s'applique pas (événement unique)

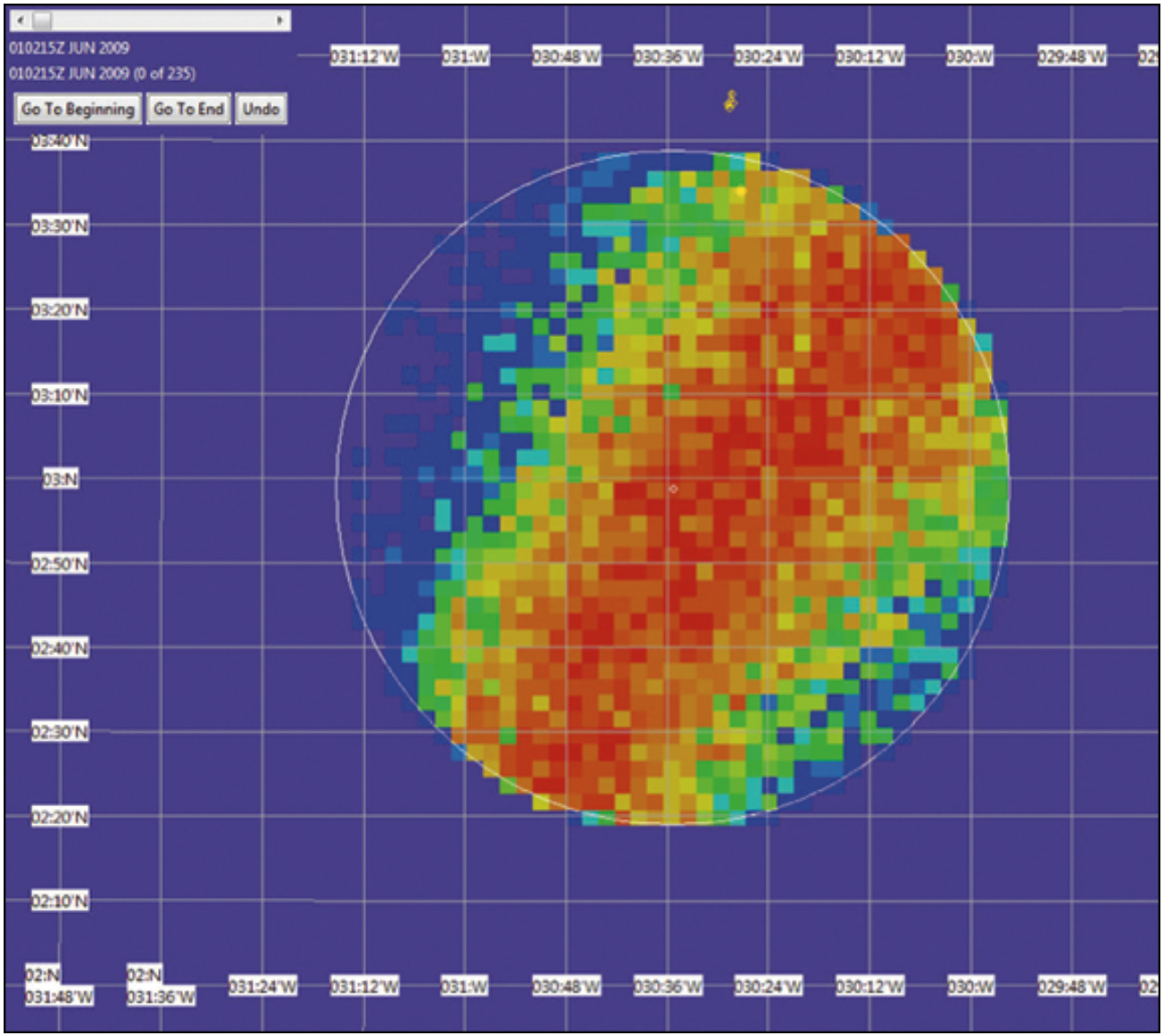
Tentatives précédentes



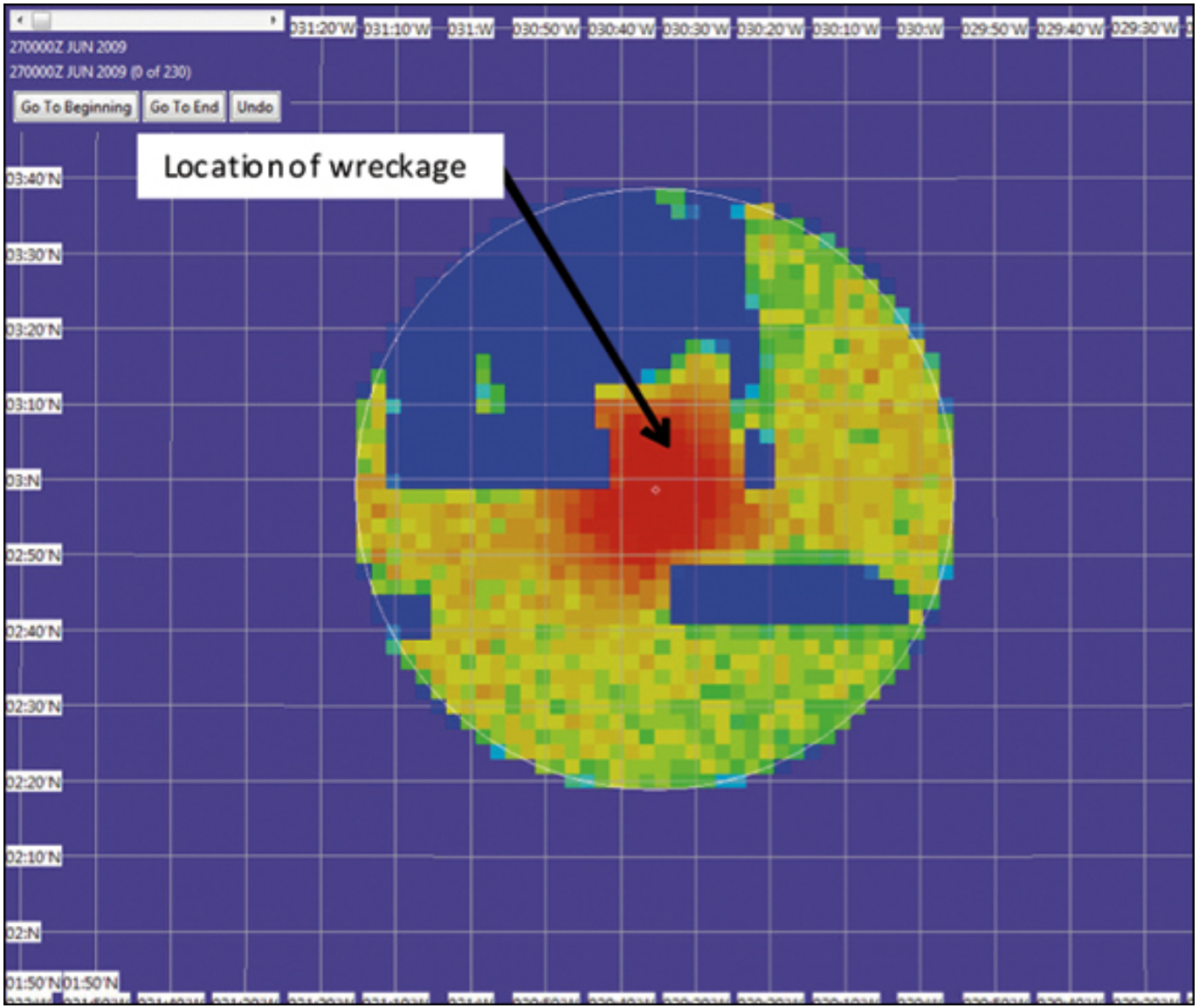
Prior basée sur la dynamique du vol



Vraisemblance basée sur la dérive



Posterior


270000Z JUN 2009
270000Z JUN 2009 (0 of 230)
Go To Beginning
Go To End
Undo
Location of wreckage
031:20'W
031:10'W
031:W
030:50'W
030:40'W
030:30'W
030:20'W
030:10'W
030:W
029:50'W
029:40'W
029:30'W
03:40'N
03:30'N
03:20'N
03:10'N
03:N
02:50'N
02:40'N
02:30'N
02:20'N
02:10'N
02:N
01:50'N
01:50'N

Conclusions

- Une fois la posterior connue, la recherche a été organisée en commençant par les zones de haute probabilité a posteriori
- Il y avait en fait plusieurs posteriors, car plusieurs modèles ont été pris en considération
- L'épave a été trouvée en une semaine.

Plusieurs modèles

Supposons qu'on ait plusieurs modèles $m_1, m_2, \ldots, m_k$. On peut voir l'indice du modèle comme un paramètre.
Prenons une loi a priori uniforme (par exemple) :

$$P[\mathcal{M} = m_j] = \frac{1}{k}.$$

La loi a posteriori nous donne la probabilité associée à chaque modèle, au vu des données.
On peut utiliser cette loi pour faire du choix de modèle (mais il existe d'autres techniques plus sophistiquées), mais aussi pour faire de l'estimation ou de la prédiction en intégrant sur l'incertitude sur le modèle.

Exemple : choix de variables pour la régression linéaire

Un modèle est un choix de covariables à inclure dans la régression. Avec p covariables, il y a 2^p modèles.
Cadre classique (fréquentiste) :

- On choisit des variables, à l'aide d'une fonction de pénalité (par ex AIC), ce qui nous donne un modèle
- On fait de l'estimation et de la prédiction au sein de ce modèle
- Si on veut une mesure de l'incertitude, c'est possible, mais seulement au sein de ce modèle

Exemple : choix de variables pour la régression linéaire

Un modèle est un choix de covariables à inclure dans la régression. Avec p covariables, il y a 2^p modèles.

Cadre classique (fréquentiste) :

- On choisit des variables, à l'aide d'une fonction de pénalité (par ex AIC), ce qui nous donne un modèle
- On fait de l'estimation et de la prédiction au sein de ce modèle
- Si on veut une mesure de l'incertitude, c'est possible, mais seulement au sein de ce modèle

Cadre bayésien :

- On explore l'espace de tous les modèles
- On obtient des probabilités a posteriori
- On fait de l'estimation et de la prédiction dans chaque modèle (ou, en pratique, dans chaque modèle de probabilité non négligeable)
- On pondère ces estimations / prédictions par la probabilité a posteriori de chaque modèle

On prend donc en compte l'incertitude sur le modèle.

Conclusions

- L'inférence bayésienne est un outil puissant pour prendre en compte complètement toutes les sources d'incertitude
- Difficulté du choix de la loi a priori
- Les principales limitations sont computationnelles