Détection bayésienne d'outliers et ses applications en archélogie

Jean-Michel GALHARRET [1] & Anne PHILIPPE [1] & Norbert MERCIER [2]

[1] *Laboratoire Jean Leray , 2 rue de la Houssinière, 44322 Nantes Cedex 3 , jean-michel.galharret@univ-nantes.fr*

[2] *UMR 5060 CNRS, Université Bordeaux Montaigne, Maison de l'Archéologie, Esplanade des Antilles, 33607 Pessac Cedex, France*

Résumé. Nous nous intéressons à la détection d'outliers pour des questions de datation en archéologie. La méthode proposée est basée sur une extension du modèle d'événement proposé par Lanos et Philippe (2018). On exploite les hyperparamètres de ce modèle robuste pour identifier les outliers. Ces valeurs aberrantes sont alors supprimées de l'échantillon avant une ré-estimation du paramètre d'intérêt par une méthode non robuste. On applique cette procédure à la combinaison de dates et à l'estimation de l'âge d'un objet par luminescence. Dans les deux cas nous montrons par des simulations qu'il est préférable d'exclure les outliers détectés plutôt que d'utiliser la méthode d'estimation robuste. Les résultats sont meilleurs en terme d'exactitude et de précision.

Mots-clés. Modèle hiérarchique bayésien, valeurs aberrantes, statistique appliquée en archéologie.

Abstract. We are interested in the detection of outliers for dating methods in archaeology. The proposed method is based on an extension of the event model proposed by Lanos and Philippe (2018). The hyperparameters of this robust model are used to identify outliers. These outliers are then removed from the sample before the parameter of interest is re-estimated using a non-robust method. This procedure is applied to the combination of dates and the estimation of an OSL age. In both cases we show by simulations that it is better to exclude the detected outliers than to use the robust estimation method. The results are better in terms of accuracy and precision.

Keywords. Bayesian hierarchical model, outliers, statistics applied in archaeology.

1 Introduction

Dans les problèmes de datation, les experts sont souvent confrontés à des données aberrantes provenant d'erreur de prélèvement lors des fouilles archéologiques. Pour gérer la présence d'éventuels outliers, on peut utiliser des modèles de mélange à deux composantes l'une d'elles correspondant aux valeurs aberrantes. Cette approche a été implémentée dans le logiciel OxCal, référence en datation par le carbone 14 et développé par Bronk

Ramsey, C. (2009). Une alternative qui est basée sur un modèle robuste (appelé modèle d'événement) a été proposée par Lanos et Philippe (2018). Comme pour toutes les approches robustes l'inconvénient est une perte d'efficacité. Dans cet article nous proposons une méthode d'estimation en deux étapes : on commence par détecter les outliers en utilisant le modèle d'événement et ensuite on estime les dates sur le sous-échantillon sans ces données identifiées comme aberrantes.

2 Description de la procédure

On généralise le modèle d'événement proposé initialement pour des données gaussiennes. On considère une famille paramétrique $\{f(\cdot|\theta),\ \theta \in \Theta\}$ et un échantillon $X_1, ..., X_n$ de vraisemblance :

$$\prod_{i=1}^{n} f(X_i|\theta_i) \quad (\theta_1, \ldots, \theta_n) \in \Theta^n \tag{1}$$

Si aucune information n'est disponible pour distinguer les paramètres θ_i, nous choisissons une loi a priori échangeable sur $(\theta_1, \ldots, \theta_n)$ de la forme :

$$\pi(\theta_1, \ldots, \theta_n|\theta, \sigma) = \prod_{i=1}^{n} \pi(\theta_i|\theta, \sigma^2) \tag{2}$$

où θ, σ^2 sont inconnus. On suppose que la loi est paramétrisée par : $\theta = \mathbb{E}(\theta_i|\theta, \sigma^2)$ et $\sigma^2 = \mathrm{Var}(\theta_i|\theta, \sigma^2)$. Le paramètre σ^2 donne une mesure a priori de la variabilité des $(\theta_i)_i$. Néanmoins, lorsque X_i est un outlier θ_i sera éloigné de θ et un seul σ^2 ne va pas permettre de capter ces larges variations. Lanos et Philippe (2018) proposent d'intégrer des variances individuelles, autrement dit la loi a priori (2) devient :

$$\pi(\theta_1, \ldots, \theta_n|\theta, \sigma_1, \ldots, \sigma_n) \ = \ \prod_{i=1}^{n} \pi(\theta_i|\theta, \sigma_i^2) \tag{3}$$

où $\theta = \mathbb{E}(\theta_i|\theta, \sigma_i^2)$, $\sigma_i^2 = \mathrm{Var}(\theta_i|\theta, \sigma^2)$. Il reste à définir une loi a priori sur les paramètres $\theta, \sigma_1^2, \ldots, \sigma_n^2$,

$$\pi(\theta, \sigma_1^2, \ldots, \sigma_n^2) \ = \ \pi_0(\theta) \prod_{i=1}^{n} \pi_s(\sigma_i^2) \tag{4}$$

Pour la loi a priori π_s, nous suivons la même stratégie que Spiegelhalter, et al. (2004) qui proposent une loi de shrinkage uniforme. Cette loi est construite de la façon suivante : on fixe S_0^2 et on suppose que le shrinkage "moyen" $\dfrac{\sigma_1^2}{S_0^2 + \sigma_1^2}$ suit une loi uniforme. Le paramètre S_0^2 correspond ainsi à la médiane de la loi de σ_1^2.

Règle de décision : On propose une règle de décision basée sur la variance individuelle σ_i^2 pour détecter les outliers. A posteriori, pour un outlier X_i, la variance individuelle σ_i est grande puisque σ_i tend vers l'infini quand $|\theta_i - \theta|$ tend vers l'infini. Pour quantifier ce fait, nous allons comparer la médiane de la loi a posteriori de σ_i^2 avec la médiane S_0^2 de sa loi a priori π_s. Pour un seuil fixé k, l'observation X_i est un outlier lorsque

$$\mathbb{P}(\sigma_i > kS_0 | X_1, \ldots, X_n) > 0.50 \tag{5}$$

Le seuil k représente le facteur d'augmentation de la médiane par rapport à sa valeur a priori. La valeur de k est fixée pour que la procédure rejette au plus 5% des observations dans un échantillon sans outlier. En pratique la valeur de k est obtenue par une méthode de Monte Carlo.

Gestion des outliers : Notre stratégie est d'exclure les outliers en utilisant la règle de décision (5). Soit $(X_i)_{i\in J}$ le sous-échantillon de $(X_i)_{i\in\{1,\ldots,n\}}$ obtenu en retirant les outliers. Nous ré-estimons le paramètre d'intérêt θ avec le modèle (1) et (2) ou bien avec le modèle :

$$\begin{aligned} &\prod_{i\in J} f(x_i|\theta) \quad \theta \in \Theta \\ &\theta \sim \pi_0 \end{aligned} \tag{6}$$

qui correspond au cas limite du modèle (1), (2) en prenant $\forall i, \sigma_i = 0$. Cette estimation est calculée sur un plus petit échantillon que la version robuste mais nous verrons qu'elle est néanmoins plus efficace dans les deux applications qui seront développées.

3 Applications en archéologie:

Nous proposons deux applications de la procédure précédente.

3.1 Combinaison de dates

Comme étudié dans Lanos et Philippe (2018), nous appliquons la méthodologie précédente à une combinaison de mesures gaussiennes pour lesquelles on connait l'incertitude de mesure. Le modèle hiérarchique correspondant est le suivant :

$$\begin{aligned} X_i &= \theta_i + s_i\epsilon_i, \qquad \forall\ i = 1, ..., n \\ \theta_i &= \theta + \sigma_i\rho_i \end{aligned} \tag{7}$$

où $(\epsilon_1, ..., \epsilon_n, \rho_1, ..., \rho_n)$ sont centrés, gaussiens et indépendants, et $s_1^2, \ldots, s_n^2$ sont connus. Sur ce premier exemple, nous mettons en évidence l'apport de notre approche par rapport à l'estimation de l'âge donné par le modèle (7). Cette méthode d'estimation de l'âge a

été proposée par Lanos et Philippe (2018) comme une méthode robuste à la présence d'outliers. Pour comparer ces deux méthodes nous simulons des échantillons contaminés par des outliers et nous comparons les performances en fonction du taux de contamination choisi. Les résultats numériques mettent en évidence une réduction du biais et de la variance de l'estimateur de Bayes lorsque l'on suit notre procédure. A la lumière de ces résultats il semble donc préférable d'exclure les outliers et de ré-estimer l'âge sur un plus petit échantillon.

3.2 Détermination d'un âge par luminescence

Pour la méthode de datation par luminescence stimulée optiquement (OSL), nous proposons une méthode d'estimation bayésienne de l'âge (voir Mercier 2008 pour une description de cette méthode). Cette estimation de l'âge A est effectuée à partir du débit de dose D_R dans l'environnement et de la dose équivalente D_e mesurée sur l'échantillon prélevé. La relation fondamentale liant ces paramètres est :

$$D_e = AD_R \tag{8}$$

Les données disponibles pour estimer cet âge A sont :

- un échantillon simulé par l'expérimentateur suivant la loi de D_R en fonction de paramètres physiques mesurés sur le terrain. On choisit une loi de Cauchy pour modéliser la loi des D_R. On note μ_R, σ_R les paramètres de cette loi.
- un échantillon $(\widetilde{D}_e^j)_j$ de doses équivalentes mesurées sur chaque grain de quartz constituant l'échantillon à dater. Chaque dose équivalente D_e^j est connue à une erreur ε_j près de variance connue s_j^2. On a pour tout j :

$$\widetilde{D}_e^j = D_e^j + \varepsilon_j,\ \varepsilon_j \sim \mathcal{N}(0, s_j^2)$$

En notant $(a_j)_j$ les âges de chaque grain de l'échantillon, la relation (8) nous permet d'écrire la loi des D_e^j comme :

$$D_e^j \sim \mathcal{C}(a_j\mu_R, a_j^2\sigma_R^2)$$

où $\mathcal{C}$ est la loi de Cauchy de paramètre de position $a_j\mu_R$ et d'échelle $a_j\sigma_R$. La relation (3) nous donne le lien entre l'âge cherché A et les âges a_j :

$$a_j \ = \ A + \sigma_j\rho_j$$

On applique cette méthode sur des données réelles. Sur un échantillon constitué de 53 grains, on détecte 22.6% de valeurs aberrantes (cf Figure 1)

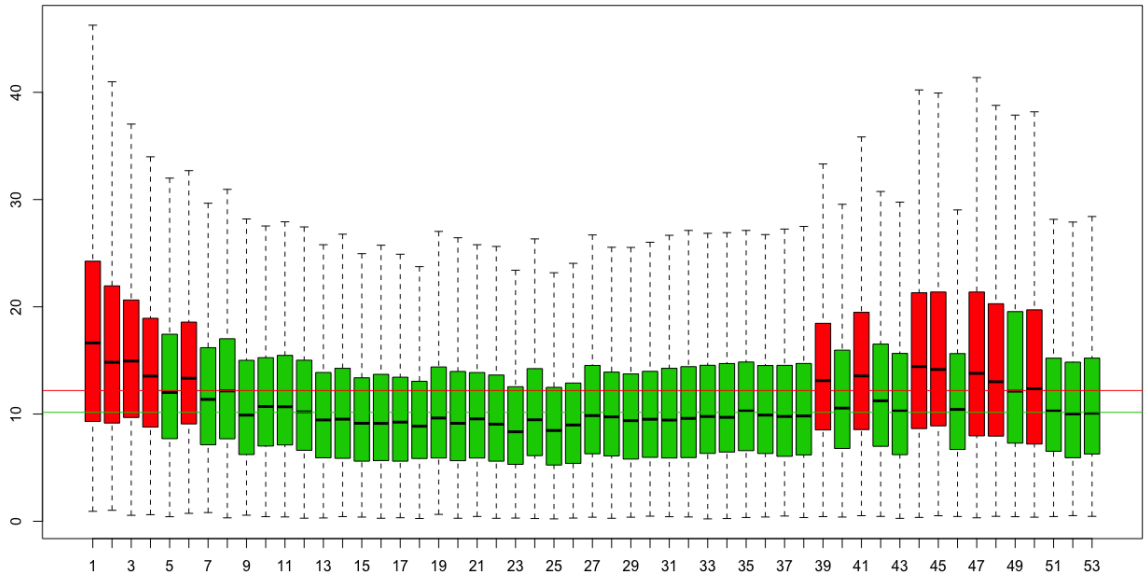

Figure 1: Boxplot des échantillons simulés par MCMC suivant les lois a posteriori des paramètres σ_j de chaque grain j. La ligne rouge représente kS_0, les boxplot en rouge correspondent aux données détectées comme aberrantes.

Après l'élimination des données aberrantes l'âge estimé est :

$\mathbb{E}(A\|data)$	LowerCI 95%	UpperCI 95%
58.29	52.80	64.62

Cette estimation de l'âge est concordante avec la connaissance de la période du site d'où l'échantillon provient. En effet, sur ce site, du matériel archéologique a aussi été daté par la méthode de datation par le carbone 14 qui fournit des âges contemporains au nôtre.

Bibliographie

Bronk Ramsey, C. (2009) *Dealing with outliers and offsets in radiocarbon dating.*, Radiocarbon, 51(3) :1023–1045.

Spiegelhalter, D.J., Abrams, K.R., and Myles, J. P., (2004) *BAyesian Approaches to Clinical Trials and Health-Care Evaluation*, Wiley, Chichester.

Lanos, P. and Philippe, A. (2018), *Event date model: a robust Bayesian tool for chronology building*, Communications for Statistical Applications and Methods, 25(2) :131-157.

Mercier, N. (2008). Datation des sédiments quaternaires par luminescence stimulée optiquement: un état de la question. Quaternaire. Revue de l'Association française pour l'étude du Quaternaire, 19(3), 15-204.