Bayesian approach using expert's opinion : Impact of rainfall on production of Périgord black truffles

Paul-Marie Grollemund [1,*,†] & Christophe Abraham [2,†] &
Meïli Baragatti [3,†]

[*] *UMR IMAG CNRS 5149, Université de Montpellier, Place E. Bataillon 34095, Montpellier Cedex, France*
[†] *UMR Mistea, Montpellier SupAgro-INRA, 2 place Pierre Viala, 34060 Montpellier Cedex 2, France*
[1] *paul-marie.grollemund@umontpellier.fr*
[2] *christophe.abraham@supagro.inra.fr*
[3] *meili.baragatti@supagro.inra.fr*

Résumé. Un point important de la modélisation bayésienne est de construire une distribution *a priori* des paramètres du modèle. Il est possible de construire une distribution qui prennent en compte des informations des experts du domaine d'application. L'extraction de ces informations est une tâche compliquée parce qu'elle consiste à traduire en termes probabilistes les avis des experts. Durant cet exposé, nous présenterons deux approches pour éliciter l'avis des experts à propos du modèle Bliss, cas particulier du modèle de régression linéaire fonctionnelle. Nous appliquerons ensuite les méthodologies proposées pour estimer l'impact des précipitations sur la production du truffe noire du Périgord.

Mots-clés. Méthodes bayésienne, Élicitation

Abstract. An important point of Bayesian statistical modeling is to specify the prior distribution of model's parameters. When prior knowledge is available, it is possible to build one which considers prior informations from subject-matter experts. The extraction of such informations is a complex task because the statistician have to state the expert's beliefs in probabilistic terms. In this talk, we will present two approaches to elicit expert's beliefs about the Bliss model which is a particular case of the Functional Linear Regression model. The proposed methodologies will be applied to estimate the influence of the rainfall on the production of the Périgord black truffles.

Keywords. Bayesian methods, Elicitation

1 Introduction

L'élicitation est une part importante de la modélisation bayésienne lorsque le nombre de données est limité. Dans ce cas, les données peuvent ne pas informer suffisamment le

modèle si bien que l'inférence statistique peut ne pas être fiable (Manel et al., 2001). Il est alors possible d'avoir recours à une inférence subjective en prenant en compte des avis d'experts.

La littérature concernant l'élicitation est vaste et un certain nombre d'analyses bibliographique sont disponibles (par exemple Jenkinson, 2005). Les méthodologies proposées s'intéressent principalement à la mise en place d'une procédure d'élicitation et à modéliser les avis des experts en termes de probabilités. Une première difficulté est que les experts ne sont pas familiers avec certains concepts statistiques et probabilistes. Pour des échanges productifs avec eux, l'élicitation doit être la plus simple possible (voir Crowder, 1992). Une seconde difficulté est que les informations obtenues doivent être adaptées pour être incluses dans un modèle statistique. C'est pourquoi certains auteurs ont développé des protocoles pour établir une procédure d'élicitation (voir par exemple Low-Choy et al., 2009).

Dans le cadre du modèle de régression linéaire, l'élicitation a été beaucoup étudiée par des auteurs comme Garthwaite et al. (2005). Dans ce cadre, James et al. (2010) souligne la difficulté de demander l'avis des experts à propos des coefficients de régression. En effet, les avis que peuvent avoir les experts reposent sur l'interprétation du modèle. Or donner un ordre de grandeur pour des coefficients n'est pas évident, *a fortiori* quand le *design* n'est pas orthogonal. Ainsi, l'élicitation ne peut pas être de renseigner directement la valeur d'un paramètre. Une alternative est alors de questionner les experts sur des quantités observables.

Dans ce papier, nous aborderons le problème de l'élicitation dans le cadre du modèle Bliss (Bayesian functional LInear regression with Sparse Step functions, voir Grollemund et al., 2017) qui est un cas particulier du modèle de régression linéaire fonctionnelle. Soit y une variable réponse réelle dépendante d'une covariable fonctionnelle $x(t)$ pour $t \in [0,1]$. Le modèle de régression linéaire fonctionnelle est donné par

$$y_i|\mu,\beta,\sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu + \int_0^1 x_i(t)\beta(t)\mathrm{d}t \ , \ \sigma^2\right), \tag{1}$$

voir Reiss et al. (2016) pour une revue des méthodes d'ajustement de ce modèle.

Le but de l'approche Bliss est de dégager les intervalles pour lesquelles la covariable fonctionnelle $x(\cdot)$ a un effet sur y. L'objectif est donc de fournir une estimation de la fonction coefficient qui soit parcimonieuse, *i.e.* obtenir $\hat{\beta}(\cdot)$ non nulle uniquement sur quelques intervalles. Cette approche repose sur une décomposition adaptative de β sur un ensemble de K fonctions étagées :

$$\beta(t) = \sum_{k=1}^{K} \beta_k \frac{1}{|\mathcal{I}_k|} \mathbf{1}\left\{t \in \mathcal{I}_k\right\},$$

où les β_k sont des nombres réels et les $\mathcal{I}_k$ sont des intervalles. Le modèle Bliss s'écrit

$$y_i|\mu,\beta,\sigma^2,\mathcal{I} \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu + x_i(\mathcal{I})\beta \ , \ \sigma^2\right) \tag{2}$$

où $x_i(\mathcal{I})$ est un vecteur dont le k^{e} élément est $\frac{1}{|\mathcal{I}_k|}\int_{\mathcal{I}_k} x_i(t)\mathrm{d}t$. Voir Grollemund et al., 2017 pour une spécification de la distribution *a priori* et des estimateurs. Ce modèle est appliqué pour étudier l'impact des précipitations (covariable) sur la production de truffes noires du Périgord (jeu de données fourni par J. Demerson). Pour cette étude, peu de données sont disponibles alors que l'estimation de la fonction coefficient est un problème compliqué (13 années d'observation pour estimer 11 paramètres). Dans ce cas, la prise en compte d'avis d'experts est importante afin d'obtenir des résultats plus pertinents et plus précis. De plus, les experts en trufficulture ont des avis solides sur la question, s'appuyant sur la connaissance de la croissance de la truffe et de ses mécanismes complexes de reproduction.

Nous proposons deux approches pour éliciter l'avis d'experts dans ce cadre. Premièrement, nous éliciterons l'avis des experts en leur demandant de construire des jeux de données plausibles d'après eux. Nous présenterons un modèle prenant en compte les données observées et les données élicitées. Deuxièmement, nous proposerons d'éliciter des informations concernant des caractéristiques de la fonction coefficient. Dans le cadre d'une collaboration avec F. Le Tacon, C. Murat, J. Gravier, P. Montpied, J.-L. Dupouey, ces méthodes seront appliquées pour étudier l'impact des conditions météorologiques sur la récolte de truffes noires du Périgord (voir Le Tacon et al., 2014 pour une présentation des données). L'avis de différents types d'experts (chercheurs et trufficulteurs) sera pris en compte.

2 Utiliser des données élicitées

L'elicitation de la fonction coefficient du modèle (2) est une tâche compliquée. En effet, comme son interprétation n'est pas simple, il n'est pas évident de proposer une valeur pour $\beta(\cdot)$ pour un temps t ou pour une intervalle donnée. Ainsi, il apparaît plus approprié de demander l'avis des experts concernant des quantités observables (voir Albert et al., 2012). En suivant les conseils de Crowder (1992), nous tirons l'information des experts en leur demandant de renseigner des pseudo données pour un *design* fixé. Cette méthode d'élicitation semble être simple pour les experts et permet de capturer leur avis sur le lien entre les deux variables.

Dans la suite, les observations de y seront notées par $y^0 = (y_1^0, \ldots, y_n^0)$ et les pseudo données de l'expert e par $y^e = (y_1^e, \ldots, y_n^e)$ pour $e = 1, \ldots, E$. Il est important de garder à l'esprit que les données sont différentes par nature (observées et élicitées), mais nous les considérons comme faisant partie d'un même jeu de données. Cependant l'incertitude des données élicitées y^e n'est pas la même que celle des données observées. Nous choisissons alors de modéliser y^0 et les y^e avec des variances différentes. Par exemple, si nous attribuons un poids faible à l'expert e, nous attribuerons une variance élevée à y^e. Nous

proposons de modéliser ces données par

$$y_i^e|x,\mu,\beta,\sigma^2,\mathcal{I} \sim \mathcal{N}\left(\mu + x_i(\mathcal{I})\beta, \frac{(\sigma^2)^{w^e}}{w_i^e}\right) \quad \text{pour } i=1,\dots,n \text{ et } e=0,\dots,E \tag{3}$$

où w_i^e est le poids de la données i de l'expert e et w^e est la moyenne des poids w_i^e. Le poids des données observées est w_i^0 =1. En utilisant la paramétrisation des intervalles par leurs milieux et leurs longueurs $\mathcal{I}_k = [m_k \pm \ell_k]$, la distribution *a priori* est la même que dans Grollemund et al. (2017).

La densité de la distribution *a posteriori* est alors donnée par

$$\pi(\mu,\beta,\sigma^2,\mathcal{I}|y^0,\dots,y^E) \propto (\sigma^2)^{-\frac{1}{2}(n+n\sum_e^E w^e+p+1)-(a+1)} \times \exp\left\{-\frac{1}{2}RSS(w) - \frac{1}{2\sigma^2}(\mu^2 v_0^{-1} + \beta^T\Sigma^{-1}\beta + 2b)\right\}\pi(\ell),$$

où $RSS(w)$ est la somme des carrés des résidus qui dépend des poids w_i^e pour $i=1,\dots,n$ et $e=0,\dots,E$. Notons que si les w_i^e sont tous nuls pour $e=1,\dots,E$, alors la précédente densité ne dépend pas des données élicitées.

De plus, l'espérance *a posteriori* de β est

$$\mathbb{E}\beta|\sigma^2,\mathcal{I},y^0,\dots,y^E = \frac{1}{\sigma^2}M_{\sigma^2,w}^{-1}\, x(\mathcal{I})^T y + \sum_{e=1}^{E} M_{\sigma^2,w}^{-1}\, x(\mathcal{I})^T {W_e}^{-1} y^e, \tag{4}$$

où W_e est une matrice diagonale dont les éléments sont $(\sigma^2)^{w^e}/w_i^e$, $M_{\sigma^2,w}$ est une matrice qui dépend de σ^2 et d'une combinaison de matrice de covariance de β *a priori* et du *design* pondéré par les poids w_i^e. Ainsi, (4) est une combinaison des données observées et des données élicitées, où le poids de chaque avis d'experts y^e est pondéré par la matrice W_e. Notons que lorsque tout les w_i^e sont nuls pour $e=1,\dots,E$, (4) ne dépend pas de l'avis des experts.

L'inférence se fait en utilisant une méthode numérique pour échantillonner la distribution *a posteriori*. Dans Grollemund et al. (2017), les auteurs proposent d'utiliser un *Gibbs sampler* mais pour ce modèle la distribution conditionnelle de σ^2 n'est pas conjuguée, nous proposons donc d'utiliser un *Metropolis-Within-Gibbs*.

3 Avis d'experts concernant la fonction coefficient

Pour l'étude du jeu de données des truffes noires du Périgord (Le Tacon et al., 2014), il est raisonnable pour les experts d'avoir un avis à propos 1) des périodes pour lesquelles les précipitations ont (ou non) un impact sur la production des truffes, et 2) si l'effet est positive ou négatif. Comme nous associons un effet positif sur une intervalle T à l'évènement "$\beta(\cdot)$ est positive sur T", nous nous définissons la quantité $\beta^s(\cdot)$ définie comme

$$\beta^s(t) = \mathbf{1}\{\beta(t) > 0\} - \mathbf{1}\{\beta(t) < 0\}.$$

Notons que $\beta^s(t)$ désigne simplement le signe de $\beta(t)$. Nous élicitons de chaque expert e, une fonction $\beta^s(\cdot)$ que nous noterons $\beta_e^s(\cdot)$. Pour un groupe de E experts, la distribution *a priori* de β que nous proposons est

$$\pi(\beta|\sigma^2, \mathcal{I}; \tau) \propto \pi_0(\beta|\sigma^2, \mathcal{I}) \times \prod_{e=1}^{E} \exp\left\{-\tau \operatorname{dist}(\beta^s, \beta_e^s; g_e)\right\}, \tag{5}$$

où *dist* est une distance sur l'ensemble des β^s, g_e est la fonction positive qui désigne la confiance de l'expert e et $\pi_0(\beta|\sigma^2, \mathcal{I})$ est la distribution *a priori* de β donnée dans Grollemund et al. (2017). En ce qui concerne l'hyperparamètre de régularisation τ, nous choisissons de le fixer avec une procédure de validation croisée bayésienne (voir Vehtari and Ojanen, 2012). L'idée de cette distribution *a priori* est de pénaliser les valeurs de paramètre pour lesquelles β^s est éloigné de β_e^s pour chaque expert. Ainsi, le produit dans la formule (5) peut être vu comme un terme de pénalisation.

Dans ce papier, nous choisissons la distance L^2 pour *dist* :

$$\operatorname{dist}(\beta^s, \beta_e^s; g_e) = \int (\beta^s(t) - \beta_e^s(t))^2 g_e(t) \mathrm{d}t.$$

Dans ce cas, la distribution *a priori* de β se réécrit comme

$$\pi(\beta|\sigma^2, \mathcal{I}; \tau) \propto \pi_0(\beta|\sigma^2, \mathcal{I}) \times \exp\left\{-\tau \operatorname{dist}(\beta^s, \bar{\beta}^s; \bar{g})\right\}$$

où $\bar{g}$ désigne la confiance global des experts $\sum_{e=1}^{E} g_e(t)$ et

$$\bar{\beta}^s(t) = \sum_{e=1}^{E} \frac{g_e(t)}{\bar{g}(t)} \beta_e^s(t),$$

qui est l'avis moyen des experts. La distribution *a posteriori* de β^s est alors tirée vers l'avis moyen des experts $\bar{\beta}^s$. Nous utiliserons un *Metropolis-Within-Gibbs* pour échantillonner suivant la distribution *a posteriori*.

4 Application et discussion

Nous proposons deux approches pour inclure de l'information *a priori* dans le modèle (2). Durant l'exposé, nous présenterons comment ont été effectué les procédures d'élicitation auprès des experts. Nous appliquerons les approches proposées à des données simulées pour évaluer la sensibilité du poids attribué à l'information des experts. Ensuite, nous présenterons les résultats obtenus sur les données des truffes noires du Périgord. Enfin, les résultats seront discutés.

Références

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., and Rousseau, J. (2012). Combining Expert Opinions in Prior Elicitation. *Bayesian Analysis*, 7.

Crowder, M. (1992). Bayesian priors based on a parameter transformation using the distribution function. *Annals of the Institute of Statistical Mathematics*, 44(3) :405–416.

Garthwaite, P., Kadane, J., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions.

Grollemund, P.-M., Abraham, C., Baragatti, M., and Pudlo, P. (2017). Bayesian functional linear regression with sparse step functions. *Statistics, Methodology arXiv :1604.08403.*

James, A., Low-Choy, S., and Mengersen, K. (2010). Elicitator : An expert elicitation tool for regression in ecology. *Environmental Modelling & Software*, 25.

Jenkinson, D. (2005). The elicitation of probabilities - A review of the statistical literature. *Technical Report.*

Le Tacon, F., Marçais, B., Courvoisier, M., Murat, C., Montpier, P., and Becker, M. (2014). Climatic variations explain annual fluctuations in french périgord black truffle wholesale markets but do not explain the decrease in black truffle production over the last 48 years. *Mycorrhiza*, 24.

Low-Choy, S., O'Leary, R., and Mengersen, K. (2009). Elicitation by Design in Ecology : Using Expert Opinion to Inform Priors for Bayesian Statistical Models. *Ecology*, 90.

Manel, S., Williams, C., and Ormerod, S. (2001). Evaluating presence-absence models in ecology : the need to account for prevalence. *Journal of Applied Ecology*, 38.

Reiss, P., Goldsmith, J., Shang, H., and Ogden, T. R. (2016). Methods for scalar-on-function regression. *International Statistical Review.*

Vehtari, A. and Ojanen, J. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statist. Surv.*, 6 :142–228.