

**Prédiction d'un événement binaire à partir de données fonctionnelles :  
Application aux bovins laitiers**

C. Sauder<sup>a</sup> and H. Cardot<sup>b</sup>

<sup>a</sup>UMR 1348 PEGASE  
INRA-Agrocampus Ouest  
65 rue de Saint-Brieuc, Rennes  
cecile.sauder@rennes.inra.fr

<sup>b</sup>Institut de Mathématiques  
Université de Bourgogne  
9 av. Alain Savary, Dijon  
herve.cardot@u-bourgogne.fr

**Mots clefs** : Prédiction, régression logistique, données fonctionnelles.

L'objectif de l'étude est de prévoir le succès (ou non) à la première insémination d'une vache à partir des courbes de lactation des 42 premiers jours suivant le vêlage précédent. Le problème se ramène donc à un problème de régression logistique fonctionnelle où on cherche à prédire une variable réponse dichotomique  $Y$  à partir de courbes  $X$ . En posant  $\pi_i = P(Y = 1 | X = x_i(t); t \in T)$  la probabilité que la vache  $i$  soit gestante sachant l'évolution de sa production laitière en fonction du temps  $x_i(t)$ , on peut écrire le modèle suivant, pour  $i = 1, \dots, n$ ,

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \int_0^T \beta(t)x_i(t)dt$$

avec pour hypothèse que les  $\beta(t)$  et les courbes  $x_i(t)$  sont dans le même espace de dimension finie [1]. On peut donc les écrire ainsi  $\beta(t) = \sum_{q=1}^p b_q \Psi_q(t) = \mathbf{b}'\Psi$  et  $x_i(t) = \sum_{q=1}^p c_{iq} \Psi_q(t) = \mathbf{c}_i'\Psi$  ce qui revient à un problème de régression fonctionnelle classique

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 \mathbf{I} + \mathbf{C}\Phi \mathbf{b}$$

avec  $\mathbf{C} = (c_{iq})$  et  $\Phi = \left(\Phi_{kq} = \int_T \Psi_k(t)\Psi_q(t)dt\right)$ .

Différents packages R sont disponibles pour réaliser ce type d'analyse et nous présentons ici en détail les packages **fa** [2] et **fa.usc** [3]. Nous commençons par présenter le format des données requis pour l'utilisation de ces packages, puis nous détaillons l'utilisation de la fonction de régression logistique fonctionnelle *fregre.glm*. Le package est illustré sur les données des bovins et les résultats sont comparés à ceux obtenus par régression logistique classique (non fonctionnelle), avec sélection de variables, qui est implémentée dans la fonction *glm* du package **stats**. Les résultats de 50 validations croisées sont comparés selon 3 critères, le taux global de bien classés, la sensibilité et la spécificité.

## Références

- [1] Cardot, H., Faivre, R., Goulard, M., (2003). Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. *Journal of Applied Statistics*, Vol. 30, 1185-1199.
- [2] Ramsay, J. O., Silverman, B. W. (2006), *Functional Data Analysis*, 2nd ed., Springer, New York.
- [3] Febrero-Bande, M., Oviedo de la Fuente, M., (2012), Statistical Computing in Functional Data Analysis : The R Package *fda.usc*, *Journal of Statistical Software*, 51(4), 1-28.