Modélisation de risque individuel basée sur le modèle de décompression dynamique calibré en population

Asya METELKINA
travail en commun avec L. Pronzato and J. Rendas

2 novembre 2015
Modélisation en biologie et statistique de données biomédicales

Observations

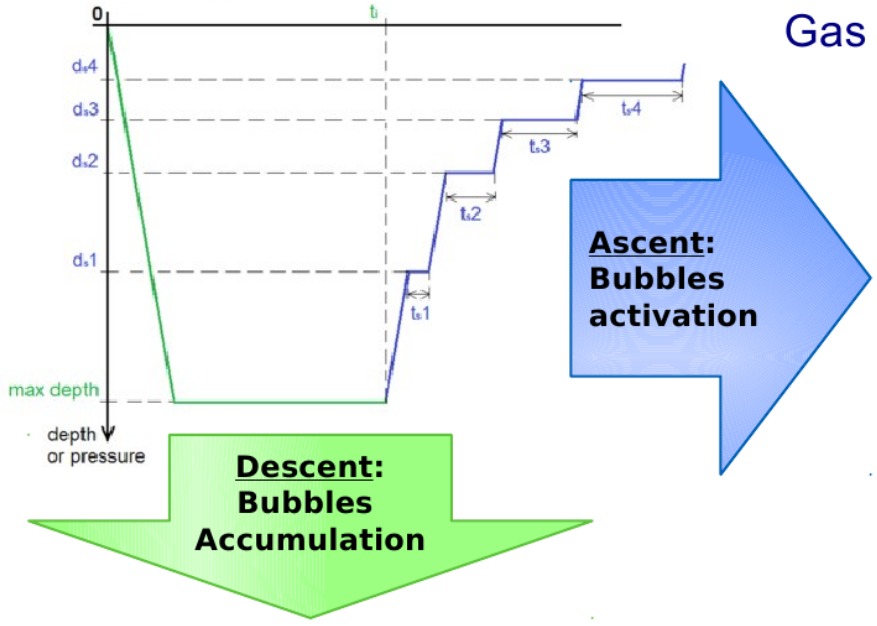
On observe $\mathbb{X}_K = \{X_1, \ldots, X_K\}$ où $X_k = (S_k, N_k, A_k)$ avec:

- le profil de pression $S_k : [0,T] \to \mathbb{R}_+$,
- le nombre de plongées $N_k \in \mathbb{N}$ avec le profil S_k,
- le nombre d'accidents de décompression $A_k \in \{1, \ldots, N_k\}$.

Notre but

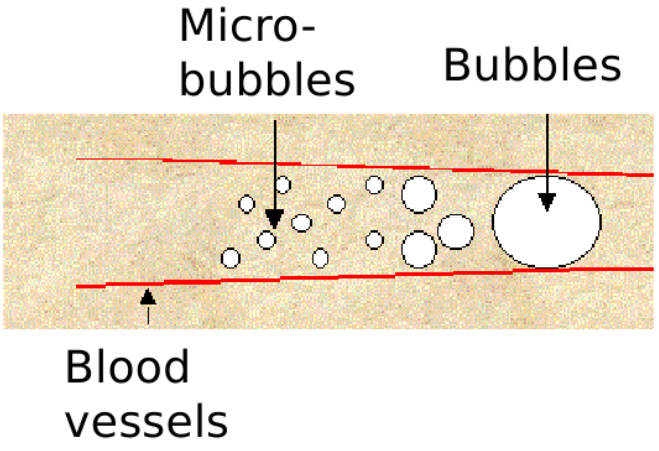
- Modéliser la probabilité $\mathbb{P}(Y = 1|S_k)$ d'accident de décompression $Y \in \{0,1\}$ conditionnellement au profil de pression $S_k : [0,T] \mapsto \mathbb{R}_+$.
- Prédire $\mathbb{P}(Y = 1|S)$ pour un **nouvel** profil de pression S.

Dive profile



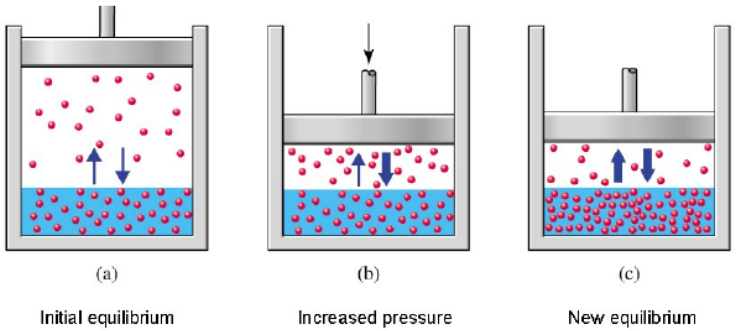
Effect of decompression :

Gas is transfered from tissues to blood and gas bubbles are growing

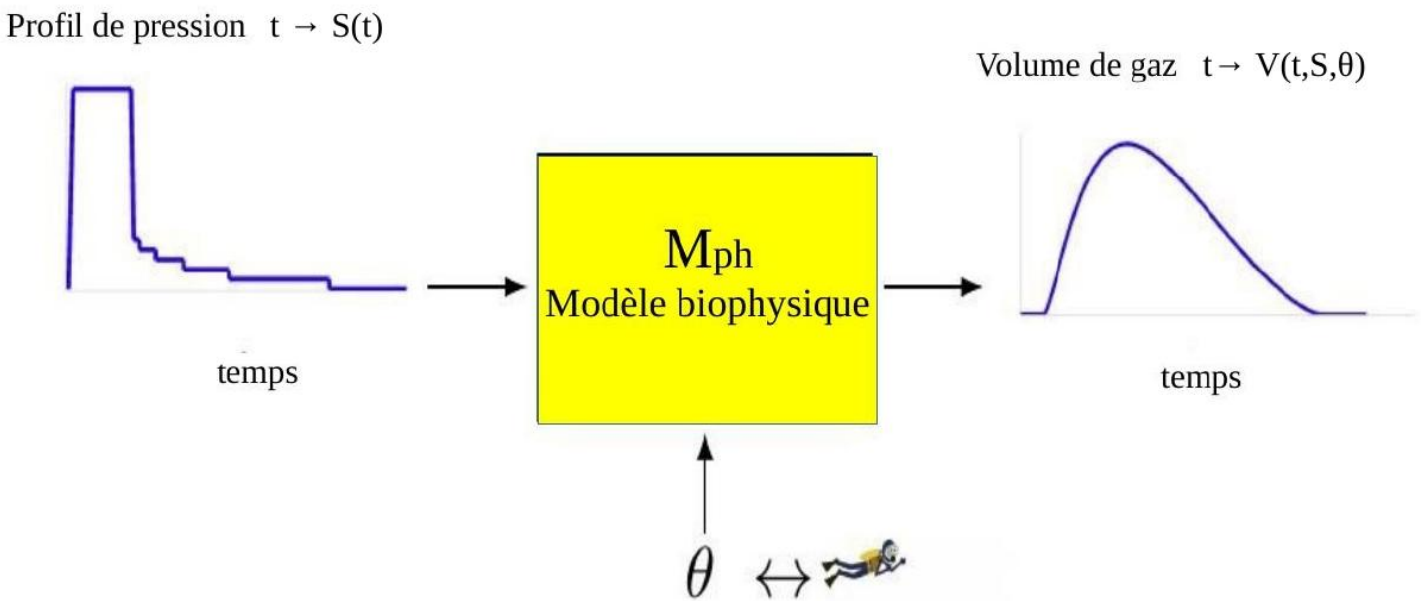


Effect of compression :

Gas is dissolved in tissues



- Les bulles de gaz dans le sang à l'origine des accidents.
  Ces bulles ne sont pas directement observables, c'est un **processus latent** $t \mapsto V(t)$.
- Des mesures indirectes $m_S(t_i) \in \{0,1,2,3,4\}$ de $V(t_i)$ avec $t_i \in [0,T]$ ont montré une grande variabilité individuelle pour un même S.
- V dépend de S et des paramètres θ d'un individu.



Modèle dynamique de décompression (J. Hugon 2010)

Un modèle dynamique $\mathcal{M}_{bph}(\theta)$, système des équations différentielles ordinaires résolue numériquement, calcule $V : t \mapsto V(t, S, \theta)$ pour un plongeur des paramètres $\theta \in \Theta$, où Θ est un compact de $\mathbb{R}^d$

$$S \rightarrow \boxed{\mathcal{M}_{bph}(\theta)} \rightarrow V(., S, \theta)$$

Distribution des θ dans la population (Y. Bennani, 2015)

Les θ ne peuvent pas être mesurés, **variables latentes**.
En supposant que $\theta \sim \pi_\theta$ sont i.i.d. dans la population, la distribution $\hat{\pi}_\theta$ a été estimée:

- Continue
- Non-gaussienne

Formalisation du problème d'estimation de risque

Soit $Y \in \{0,1\}$ une v.a. et $\theta \sim \pi_\theta$ i.i.d., on cherche un modèle paramétrique $\mathcal{M}_{hz}(\beta)$ de la distribution de risque en population:

$$\mathbb{P}(Y=1|S,\theta) = f_\beta(S,\theta), \quad \beta \in \mathbb{R}^d, \theta \sim \pi_\theta \quad \text{i.i.d.}$$

On va supposer que

- θ sont des paramètres de $\mathcal{M}_{bph}(\theta)$ et π_θ est $\hat{\pi}_\theta$ estimé.
- $\mathbb{P}(Y=1|S,\theta) = \mathbb{P}(Y=1|V(.,S,\theta)) = p_\beta(V(.,S,\theta))$ avec V est calculé par $\mathcal{M}_{bph}(\theta)$.
- β ne dépend pas de θ, π_θ et S.

Modèle paramétrique de risque en population

$$S \longrightarrow \boxed{\begin{array}{c}\mathcal{M}_{ph}(\theta), \\ \theta \sim \pi_\theta\end{array}} \longrightarrow V \sim \pi_V(S) \longrightarrow \boxed{\mathcal{M}_{hz}(\beta)} \longrightarrow p_\beta(V), V \sim \pi_V(S)$$

où $\pi_V(S)$ est la mesure image de π_θ sous l'application $\theta \mapsto V(.,S,\theta)$

Estimation par maximum de vraisemblance

Supposons que $p_\beta : \mathcal{V} \mapsto [0,1]$ sont fixés pour $\beta \in \mathcal{B}$, un compact de $\mathbb{R}^J$. On estime β par maximum de vraisemblance:

$$\hat{\beta} = \arg\max_{\beta \in \mathcal{B}} LL_{(S_k,N_k,A_k)_{k=1}^K}(\beta),$$

$$LL(\beta) = C + \sum_{k=1}^{K} \left[A_k \ln \mathbb{P}(Y=1|S_k) + (N_k - A_k) \ln(1 - \mathbb{P}(Y=1|S_k)) \right].$$

où $C = C((A_k,N_k)_{k=1}^K)$ et $\mathbb{P}(Y=1|S_k) = \mathbb{E}_{\pi_\theta}(p_\beta(V(,S_k,\theta)))$.

- Chaque évaluation de $LL_{(S_k,N_k,A_k)}(\beta)$ demande K calculs de $\mathbb{E}_{\pi_\theta}(p_\beta(V(,S_k,\theta)))$.
- π_θ est une distribution continue et non-gaussienne, donc $\mathbb{E}_{\pi_\theta}(p_\beta(V(,S_k,\theta)))$ n'a pas de forme analytique.

Approximation Monté-Carlo

$$\mathbb{E}_{\pi_\theta}(p_\beta(V(,S_k,\theta))) \approx \frac{1}{m}\sum_{i=1}^{m} p_\beta(V(.,S_k,\theta_i)), \quad \theta_i \sim \pi_\theta \quad \text{i.i.d.}$$

avec un nombre $m \in \mathbb{N}$ de tirages assez grand.

Le calcul de $LL_{(S_k,N_k,A_k)}(\beta)$ demande $K \times m$ calculs de $V(,S,\theta)$, *i.e.* $K \times m$ appels de simulateur de $\mathcal{M}_{bph}(\theta)$.

Une famille de modèles de variable binaire

On considère $p_\beta(V) = \phi(\beta^t h(V))$ avec $\beta \in \mathbb{R}^J$ et

- une fonction d'*extraction d'attributs* $h : \mathcal{V} \to \mathcal{H} \subset \mathbb{R}^J$,
- une fonction de *lien* $\phi : \mathcal{H} \to [0,1]$.

Exemple: modèle de hasard $\phi(x) = 1 - e^{-x}$ pour $\mathcal{H} \subset \mathbb{R}_+^d$ où régression logistique $\phi(x) = \frac{1}{1+e^{-x}}$ pour $\mathcal{H} \subset \mathbb{R}^d$.

Il suffit de connaître $h(V(.,S_k,\theta_i))$, $k = 1,\ldots,K$, $i = 1,\ldots,m$.

Soit $\mathcal{D} = \left\{\theta_i^0, \ldots, \theta_d^0\right\}$ une grille régulière ou un plan d'expérience *space filling* de d points dans $\Theta = \text{supp}(\pi_\theta)$.

Supposons qu'on a calculé $V(., S_k, \theta_i^0)$ pour tout $k = 1, \ldots, K$ et $i = 1, \ldots, d$ avec $K \times d$ appels de $\mathcal{M}_{bph}(\theta)$.

Processus Gaussien (krigeage)

Pour chaque $j = 1, \ldots J$ et $k = 1, \ldots K$, on considère l'application

$$g_{j,k} : \theta \mapsto h_j(V(., S_k, \theta))$$

comme une réalisation d'un processus Gaussien:

$$g_{j,k} \sim \mathcal{GP}(\mu_{j,k}(.), \sigma_{j,k}^2 K_{\rho_{j,k}}(.,.))$$

- de fonction moyenne $\mu_{j,k}(\theta)$,
- de noyau de covariance $K_{\rho_{j,k}}(\theta, \theta')$ de paramètre $\rho_{j,k}$

Choix du modèle de krigeage

- S'il n'y a pas de tendance évidente dans $\theta \mapsto g_{j,k}(\theta)$, on utilise le krigeage ordinaire avec $\mu_{j,k} \in \mathbb{R}$.
- Covariance isotrope Matérn de paramètre de régularité $\nu \in \mathbb{Z} + \frac{1}{2}$ fixé et de paramètre d'échelle $\rho_{j,k} \in \mathbb{R}_+$

$$K_{\rho_{j,k}}(r) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}r}{\rho_{j,k}} \right)^{\nu} \mathcal{K}_{\nu} \left(\frac{2\sqrt{\nu}r}{\rho_{j,k}} \right)$$

où $r = |\theta - \theta'|$ et $\mathcal{K}_{\nu}(.)$ est la fonction de Bessel modifiée d'ordre ν.

- Quand il n'y a pas de raison de compter que $g_{j,k}$ sont plus que continues, on fixe $\nu = \frac{3}{2}$.

Remarque: La covariance Matérn correspond à l'interpolation par les splines de Sobolev.

On construit l'interpolateur $\hat{g}_{j,k}$ par la méthode *plug-in*, *i.e.* utilisant des paramètres estimés:

Estimation des paramètres du modèle de krigeage

Estimateur MV $\hat{\mu}_{j,k}$ et estimateur MV restreinte $\hat{\rho}_{j,k}$ et $\hat{\sigma}_{j,k}$.

Meilleur prédicteur linéaire empirique (Santner, Williams & Notz)

L'interpolateur *plug-in* $\hat{g}_{j,k}$ est une solution du système de krigeage dual

$$\hat{g}_{j,k}(.) = \sum_{i=1}^{d} \alpha_{j,k}^{i} K_{\hat{\rho_{j,k}}}(., \theta_i^0) + \hat{\mu}_{j,k}$$

avec $(\alpha_{j,k}^1, \ldots, \alpha_{j,k}^d) \in \mathbb{R}^d$ vérifiant:

$$\begin{cases} \sum_{i=1}^{d} \alpha_{j,k}^{i} K_{\hat{\rho_{j,k}}}(\theta_l^0, \theta_i^0) + \hat{\mu}_{j,k} = g_{j,k}(\theta_l^0) \\ \sum_{i=1}^{d} \alpha_{j,k}^{i} = 0 \end{cases}$$

Les $\alpha_{j,k}^{i}$, $\hat{\mu}_{j,k}$ et $\hat{\rho}_{j,k}$ sont calculés une fois. Pour chaque interpolation $\hat{g}_{j,k}(.)$ on ne recalcule que $K_{\hat{\rho_{j,k}}}(., \theta_i^0)$

Espace des noyaux reproductibles

Si K est un noyau symétrique, le produit scalaire des fonctions $f_x(.) = K(.,x)$ avec $x \in X$

$$(K(.,x), K(.,y))_{\mathcal{K}} = K(x,y), \quad x,y \in X$$

engendre le produit scalaire sur l'espace de combinaisons linéaires finies

$$\mathcal{K}_0 = \left\{ \sum_{j=1}^{J} \alpha_j K(.,x_j) \quad \text{avec} \quad x_j \in X,\ j = 1,\ldots,J \right\}$$

Une fermeture de $\mathcal{K}_0$ est un espace de Hilbert naturel associé au noyau K, ou l'*espace des noyaux reproductibles* associé à K.

Espace des noyaux reproductibles associé au noyau Matérn

Le noyau de Matérn de paramètre de régularité ν génère l'espace de Sobolev $\mathcal{K} = W_2^{\nu-\frac{1}{2}}(\mathbb{R})$. Pour $\nu = \frac{3}{2}$, c'est l'espace de Sobolev $W_2^1(\mathbb{R})$ des fonctions $\mathbb{L}_2(\mathbb{R})$ ayant une dérivée dans $\mathbb{L}_2(\mathbb{R})$.

Erreur d'approximation par le krigeage (Wu & Schaback,1993)

- K_ρ un noyau de Matérn de régularité ν.
- $\mu \in \mathbb{R}$.
- Θ compact de $\mathbb{R}^m$.
- $g : \Theta \to \mathbb{R}$ telle que $g - \mu \in \mathcal{E}$ où $\mathcal{E}$ est l'espace des noyaux reproductibles généré par K_ρ.
- $\mathcal{D}$ une grille régulière de pas h dans Θ.
- $\hat{g}$ est le meilleur prédicteur linéaire de g associé au modèle gaussien $\mathcal{GP}(\mu, K_\rho(.,.))$.

Sous ces hypothèses, on a

$$|g - \hat{g}| \leq |g - \mu|_{\mathcal{E}} P(h) \quad \text{avec} \quad P(h) \leq C(\rho, \nu) h^{2\nu}$$

Pour $\nu = \frac{3}{2}$, on a $P(h) \leq C(\rho) h^3$.

Utilité de l'interpolation par krigeage

Une fois les attributs importants $(h_1(.), \ldots, h_K)$ sont sélectionnés et des interpolateurs $\hat{g}_{j,k}()$ sont construits, on peut:

- remplacer le modèle $\mathcal{M}_{hz}(\beta)$ par le modèle approché $\tilde{\mathcal{M}}_{hz}(\beta)$ pour des profils S_k, $k = 1, \ldots, K$:

$$p_\beta(V(.,S_k,\theta)) = \phi\left(\sum_{j=1}^{J} \beta_j h_j(V(.,S_k,\theta)\right) \approx$$

$$\tilde{p}_\beta(V(.,S_k,\theta) = \phi\left(\sum_{j=1}^{J} \beta_j \hat{g}_{j,k}(\theta))\right)$$

- calculer la vraisemblance approchée $\tilde{LL}(\beta)$.
- simuler des accidents dans la population virtuelle depuis $\tilde{\mathcal{M}}_{hz}(\beta)$.
- changer de distribution π_θ.

Approximation d'estimateur de β par MCKrige

Estimation par maximum de la vraisemblance approchée par la méthode MCKrige (Monte Carlo + krigeage):

$$\hat{\beta} = \arg\max_{\beta \in \mathcal{B}} \tilde{LL}(\beta)$$

avec

- $\tilde{LL}(\beta) = C + \sum_{k=1}^{K} \left[A_k \ln(\tilde{p}_\beta(S_k)) + (N_k - A_k) \ln(1 - \tilde{p}_\beta(S_k)) \right]$.
- $\tilde{p}_\beta(S_k) = \frac{1}{m} \sum_{s=1}^{m} \phi \left(\sum_{j=1}^{J} \beta_j \hat{g}_{j,k}(\theta_s) \right)$ avec $\theta_s \sim \pi_\theta$ i.i.d.

L'optimisation de la vraisemblance est réalisée en deux étapes:

- Recherche globale dans une grille $\mathcal{B}_0$: $\beta_0 = \arg\max_{\beta \in \mathcal{B}_0} \tilde{LL}(\beta)$,
- Recherche locale par la méthode de Powell (optimisation sans information de la dérivée) au voisinage de β_0: $\hat{\beta} = \arg\max_{\beta \in U(\beta_0)} \tilde{LL}(\beta)$.

Comment estimer les intervalles de confiance? La dépendance de $\tilde{LL}(\beta)$ de β est fortement non-linéaire.

Intervalles de confiance par le bootstrap paramétrique

On se donne des interpolateurs $\hat{g}_{j,k}()$ et l'estimateur MCKrige $\hat{\beta}$.
On fixe $N_{boot} \in \mathbb{N}$. Pour $b = 1 : N_{boot}$ on répète:

- Pour chaque $k = 1 : K$ tirer N_k paramètres $\theta_i \sim \hat{\pi}_\theta$ i.i.d.
- Calculer $\hat{p}_\beta(S_k, \theta_i) = \phi(\sum_{j=1}^{J} \beta_j g_{j,k}(\theta_i))$, $k = 1, \ldots, K$ et $i = 1 \ldots N_k$.
- Tirer indépendamment $Y_{i,k} \sim \mathcal{B}(\hat{p}_\beta(S_k, \theta_i))$. $A_k^b = \sum_{i=1}^{N_k} Y_{i,k}$.
- Estimer $\hat{\beta}^b = \arg\max_{\beta \in B} \tilde{LL}_{(S_k, N_k, A_k^b)_{k=1}^K}(\beta)$.

Le biais de $\hat{\beta}$ est estimé par $\frac{1}{N_{boot}} \sum_{b=1}^{N_{boot}} \hat{\beta}^b - \hat{\beta}$.
$\hat{IC}_\alpha(\hat{\beta}) = [\hat{\beta}^{b_{\alpha/2}}, \hat{\beta}^{b_{1-\alpha/2}}]$ avec $\hat{\beta}^{b_\alpha}$ les α-quantiles empiriques de $\hat{\beta}^b$.

Nous considérons $\mathcal{M}_{hz}$ correspondant modèles de hasard avec une intensité causale par rapport à V:

La forme de la probabilité $p_\beta(V)$

$$\mathbb{P}(Y = 1 \text{ avant } t|V) = 1 - \exp\left(-\beta \int_0^t f(s,V)ds\right)$$

avec l'intensité $f(t,V)$ causale par rapport à V, qui vérifie $f(t,V) \geq 0$ et $f(0,V) = 0$.

Pour ces modèles $p_\beta(V) = \phi(\beta^t h(V))$ avec

- $\phi = 1 - \exp(-x)$
- $h(V) = \int_0^T f(s,V)ds$

Matrice de Ficher approchée

$$\tilde{I}(\beta) = \sum_{k=1}^{K} N_k \frac{\left(\frac{1}{m}\sum_{i=1}^m (e^{-\beta\hat{g}_k(\theta_i)}\hat{g}_k(\theta_i)\right)^2}{\left(1 - \frac{1}{m}\sum_{i=1}^m e^{-\beta\hat{g}_k(\theta_i)}\right)\left(\frac{1}{m}\sum_{i=1}^m e^{-\beta\hat{g}_k(\theta_i)}\right)}$$

Etude en simulation

- $K = 131$, S_k $(k = 1, \ldots, K)$ des profils de plongée réels.
- N_k correspond à leur fréquence d'utilisation réelle:
  N_k varie de 1 à 6138 ce qui reflète la censure par dangereusité.
- Modèle $\mathcal{M}_{bph}(\theta)$ de J. Hugon avec $\Theta = [0,1]^2$.
- Plan d'expérience $\mathcal{D}$ est une grille régulière 15×15 dans Θ (de pas 0.0714)
- $\mathcal{M}_{hz}(\beta)$ avec $\phi(x) = 1 - \exp(-x)$ et $h(V) = \int_0^T V(t)dt$.

Jeu de paramètres de hasard utilisé

Les valeurs de $\beta = 0.45$, $\beta = 5$ et $\beta = 50$.

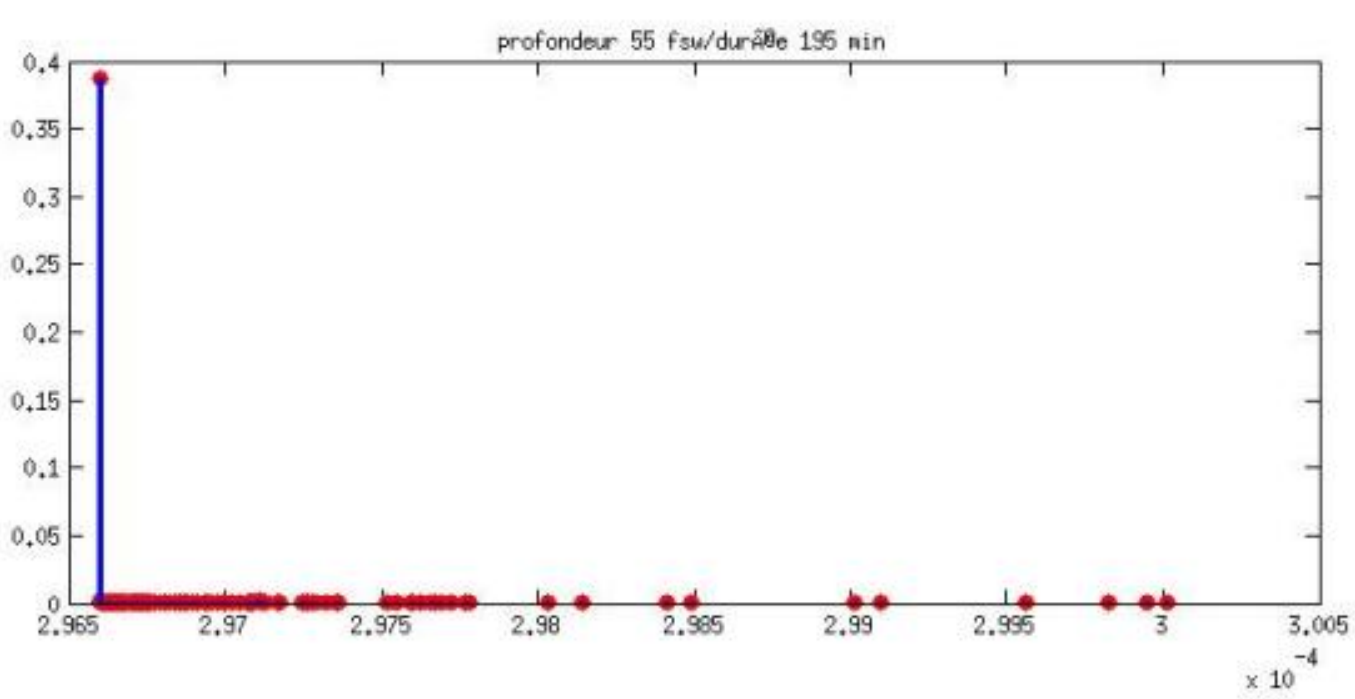
Résultats des simulations

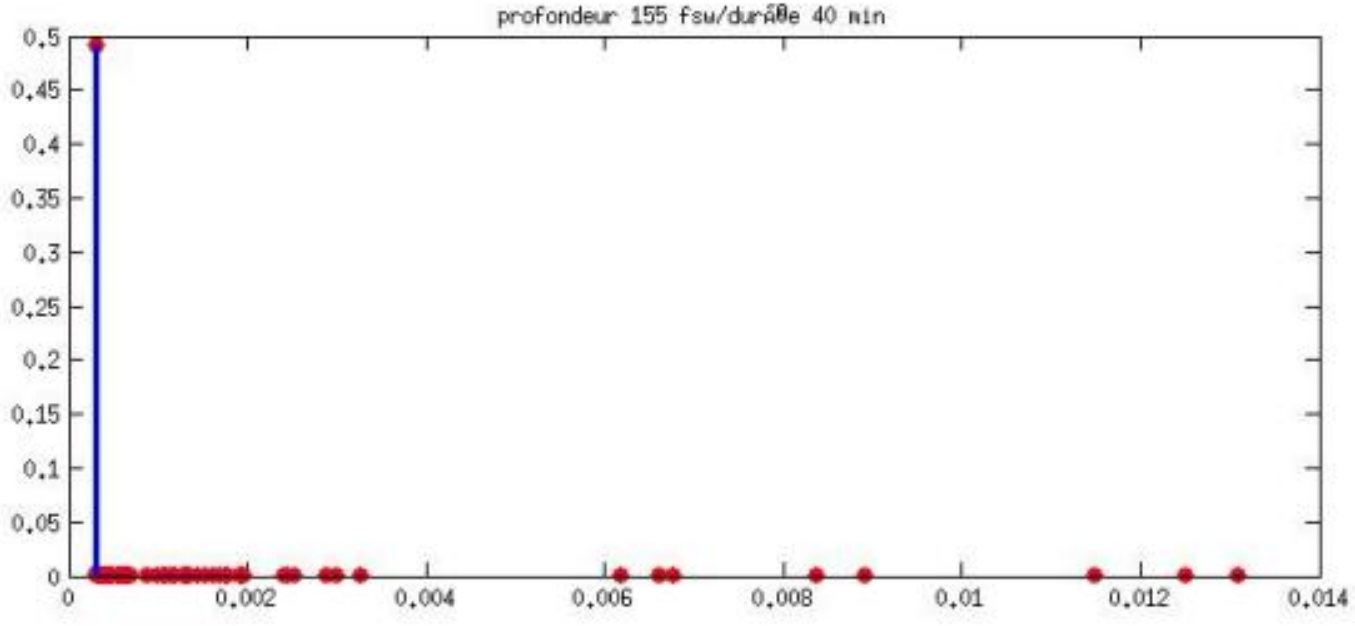
$\mathcal{M}_{hz}(\beta)$	$\hat{\beta}$	$I\hat{C}_{0.95}$	$\hat{\beta}^*$ $(I\hat{C}^*_{0.95})$
$\mathcal{M}_{hz}(0.45)$	0.48	$0.28 - 0.67$	$0.47\ (0.28 - 0.67)$
$\mathcal{M}_{hz}(5)$	5.4	$4.6 - 6.1$	$4.9\ (4.3 - 5.5)$
$\mathcal{M}_{hz}(50)$	50	$50 - 50$	$29\ (28 - 31)$

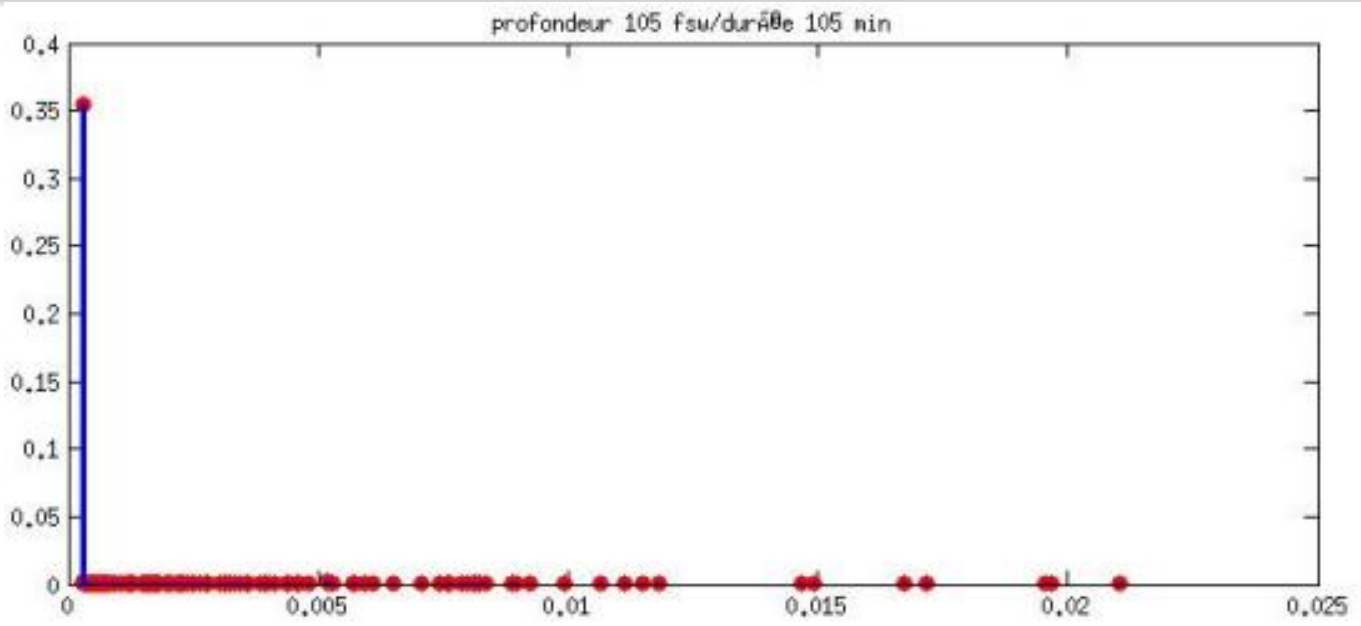
Ici $\hat{\beta}^*$ et $I\hat{C}^*_{95\%}$ sont estimés avec le modèle de *signal moyen*:

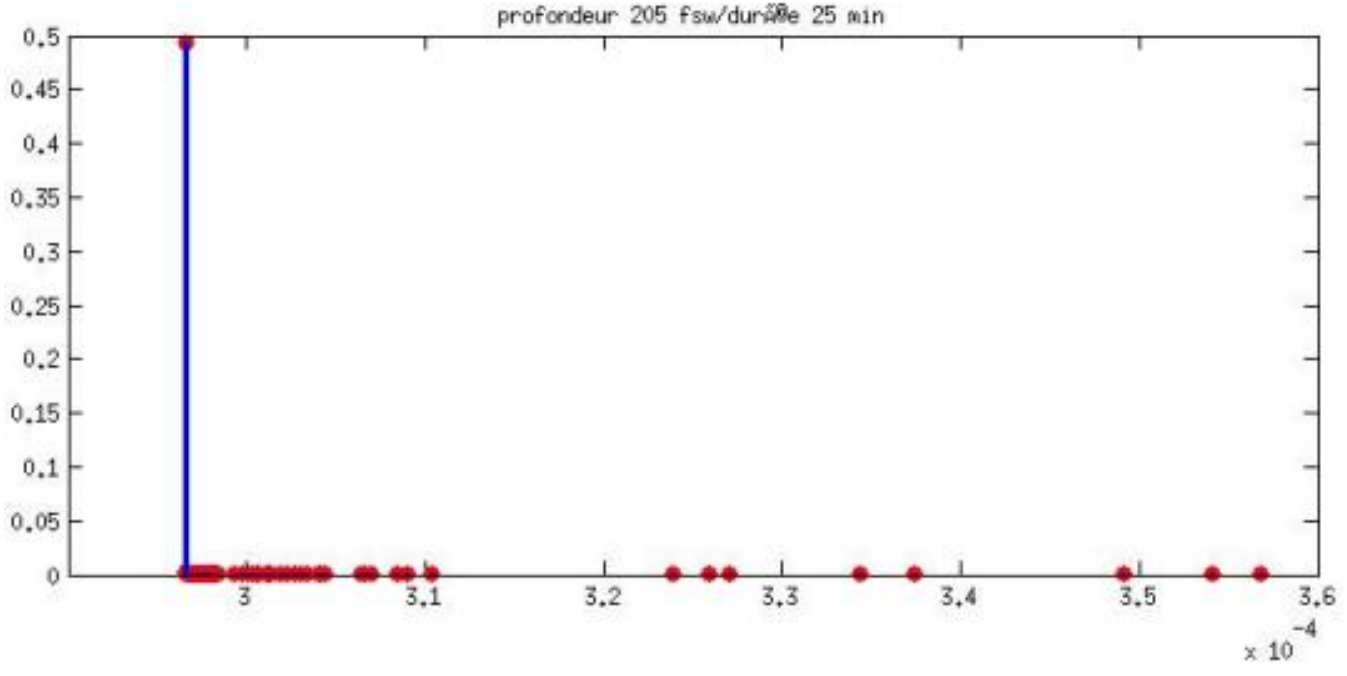
$$h(V(.,S_k)) = \mathbb{E}_{\pi_\theta}\left(\int_0^T V(t,S_k,\theta)dt\right).$$

- Si $\beta h(V) << 1$, alors $\hat{\beta} \approx \hat{\beta}^*$.
- Si $\beta h(V)$ est plus grand, $\hat{\beta}$ et $\hat{\beta}^*$ sont différents.
- Le modèle de signal moyen sous-estime le risque pour la partie de population au risque élevé.


profondeur 55 fsw/durÃ©e 195 min
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
2.965
2.97
2.975
2.98
2.985
2.99
2.995
3
3.005
x 10^-4


profondeur 155 fsw/durÃ©e 40 min
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014


profondeur 105 fsw/durÃ©e 105 min
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.005
0.01
0.015
0.02
0.025


profondeur 205 fsw/durÃ©e 25 min
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
3
3.1
3.2
3.3
3.4
3.5
3.6
x 10^-4

Conclusions

En présence de la variabilité individuelle de réponse au profil de pression S, les modèles de risque *moyen* sous-estiment le risque pour une partie de population.
Il est nécessaire de prendre cette variabilité en compte.

Travaux en cours et perspectives

- Etude théorique de l'erreur de l'estimation de β causée par MCKrige.
- Méthode bayesienne pour inclure l'information de mesures indirectes de bulles dans la prédiction de risque grâce au modèle joint de risque-mesure.
- Krigeage fonctionnel des volumes de bulles.
- Application de la méthodologie aux modèles PBPK.

Merci pour votre attention!