Approche bayésienne pour la prédiction de la composition corporelle

Simiao TIAN

Doctorant à l'INRA Jouy-en-josas et Clermont-Ferrand
(Jean-Baptiste Denis, Estelle Kuhn, Laurence Mioche, Béatrice Morio)

Applibugs, Paris, 9 Décembre 2011

Plan

Qu'est ce que c'est la composition corporelle ?

La composition corporelle est définie par

- la contribution des différents tissus et organes dans l'organisme comme la masse grasse, la masse maigre, le squelette
- selon différents compartiments du corps :
  - tronc (**t**)
  - jambes (**l**)
  - bras (**a**)
  - corps entier (**b**)

Intérêt de l'étude de la composition corporelle

La composition corporelle est un indicateur de santé publique très important, elle permet :

1. d'apprécier globalement l'état nutritionnel et physio-pathologique des individus
2. d'analyser des variations telles que la dénutrition, la croissance, le vieillissement
3. d'interpréter le métabolisme énergétique
4. de prédire le risque métabolique d'un individu, ...

Méthodes de mesure

Diverses méthodes de mesure directes et indirectes :

(I) directes

- la radioactivation neutronique
- la DEXA (*Dual energy X-ray absorptiometry*), ...

(II) indirectes

- les mesures anthropométriques (*e.g.* BMI)
- l'impédancemétrie, ...

Bases disponibles

(I) Base de référence (USA) : **NHANES** (National Health and Nutrition Examination Survey) 10394 individus (5395 hommes et 4999 femmes)

(II) Bases françaises provenant d'organisations publiques françaises

- St-Brieuc $\sim$ 23308 individus (11453 hommes et 11855 femmes)
- Anzin $\sim$ 4696 individus (738 hommes et 3955 femmes)
- CHU $\sim$ 1096 individus (527 hommes et 569 femmes)
- Primefat $\sim$ 10534 individus, uniquement les hommes
- Sportifs/Toulouse $\sim$ 1078 individus (783 hommes et 295 femmes)
- SUVIMAX $\sim$ 359 individus (173 hommes et 186 femmes)

NHANES

	ID	SEX	AGE	HGT	WGT	aF	aL	aB	lF
1	22193	M	26	179.40	217.30	13.04	13.78	0.66	30.36
2	28316	M	21	183.20	214.90	12.07	16.25	0.73	27.42

	lL	lB	tF	tL	tB	bF	bL	bB	WAI
1	44.08	1.92	45.91	57.23	1.29	91.34	121.48	4.51	
2	46.00	2.09	41.41	58.48	1.44	82.87	127.12	4.88	

	BASE	bFF	BMI	%aF	%lF	%tF	%bF	%aL	%lL
1	NHA	125.99	67.52	6.00	13.97	21.13	42.03	6.34	20.29
2	NHA	132.00	64.03	5.62	12.76	19.27	38.56	7.56	21.41

	%tL	%bL	%aB	%lB	%tB	%bB	%bFF
1	26.34	55.91	0.30	0.88	0.59	2.08	57.98
2	27.21	59.15	0.34	0.97	0.67	2.27	61.42

Récapitulatif des bases

Base	Hommes	Femmes	Avec BC	Avec WAI
NHANES	5395	4999	Oui	Oui
SUVIMAX	173	186	Oui	Oui
Anzin	738	3955	Oui	Non
CHU	527	569	Oui	Non
Sportifs/Toulouse	783	295	Oui	Non
Primefat	10534	0	Non	Oui
Saint Brieuc	11453	11855	Non	Oui

Compartiments élémentaires

Table: Recensement de tous les compartiments envisageables

	F(at)	**L**(ean)	**B**(one)	**F**(at)**F**(ree)	**W**(eight)
h(ead)	hF	hL	hB	hFF	hW
a(rm)	aF	aL	aB	aFF	aW
t(ronc)	tF	tL	tB	tFF	tW
l(eg)	lF	lL	lB	lFF	lW
ap(pendice)	apF	apL	apB	apFF	apW
b(ody)	bF	bL	bB	bFF	bW

Covariables et Variables d'intérêt

(I) Covariables

- AGE : l'âge d'individu
- HGT : la taille d'individu
- WGT : le poids d'individu
- WAI : la tour de taille d'individu

(II) Variables d'intérêt

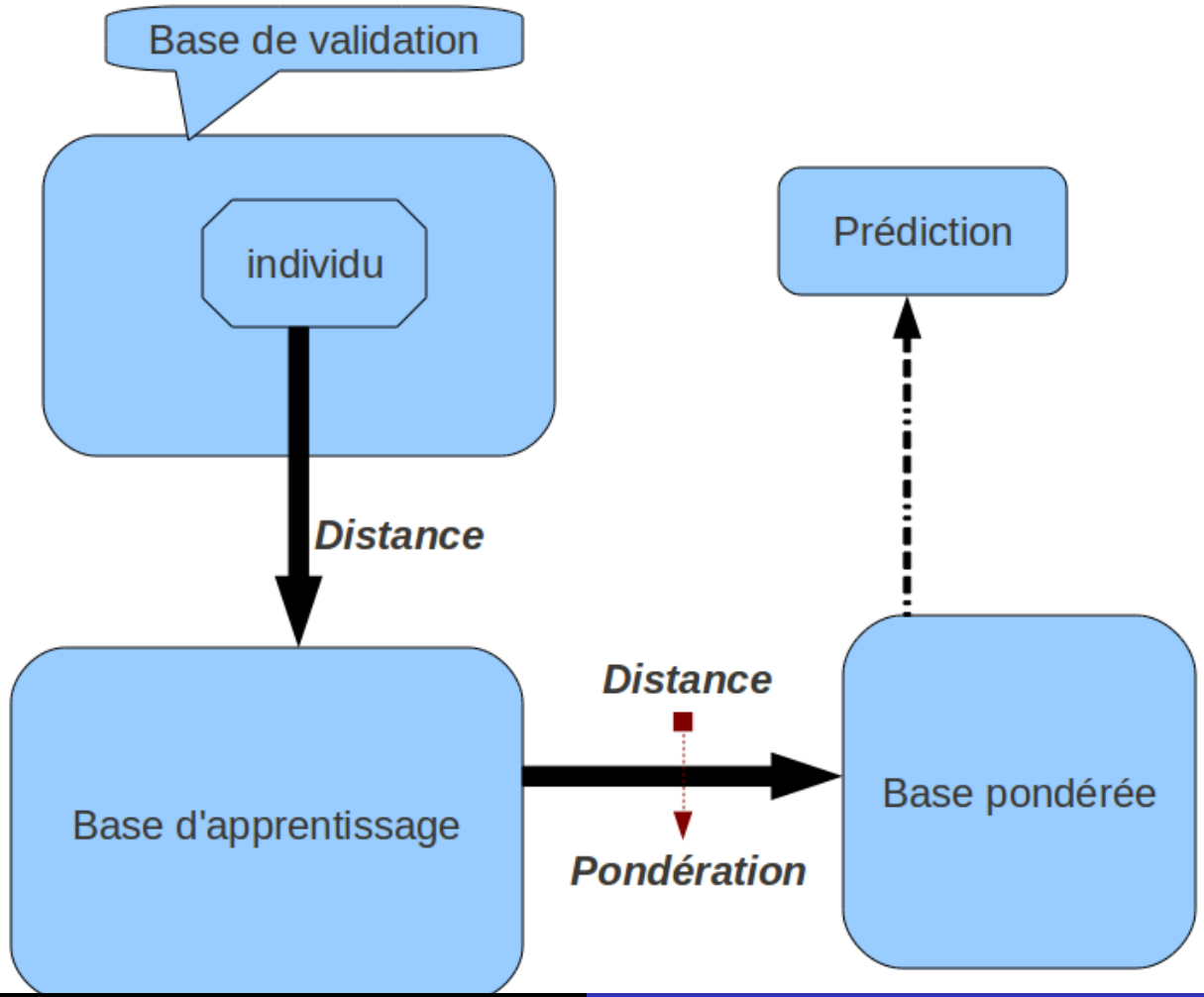
- tF, bF,
- aL, tL, lL, bL,
- bB,
- apL, bFF.

Régression Bayésienne Locale (RBL)

La Régression Bayésienne Locale est composée de 2 phases :

1. construire le *Fuzzy set* :
   - définir la distance d ;
   - définir la fonction d'appartenance $\omega = f(d)$.
2. appliquer une procédure Bayésienne
   - concevoir le modèle ;
   - proposer les lois priores.

Schéma du modèle local



Distances

Pour diversifier la manière d'extraire des sous-ensembles, de différentes formulations des distances sont proposées :

1. Distance 1 :

$$\max\{w_i|x_j - x_i|\}$$

2. Distance 2 :

$$\sum_{i=1}^{n} w_i|x_j - x_i|$$

3. Distance 3 :

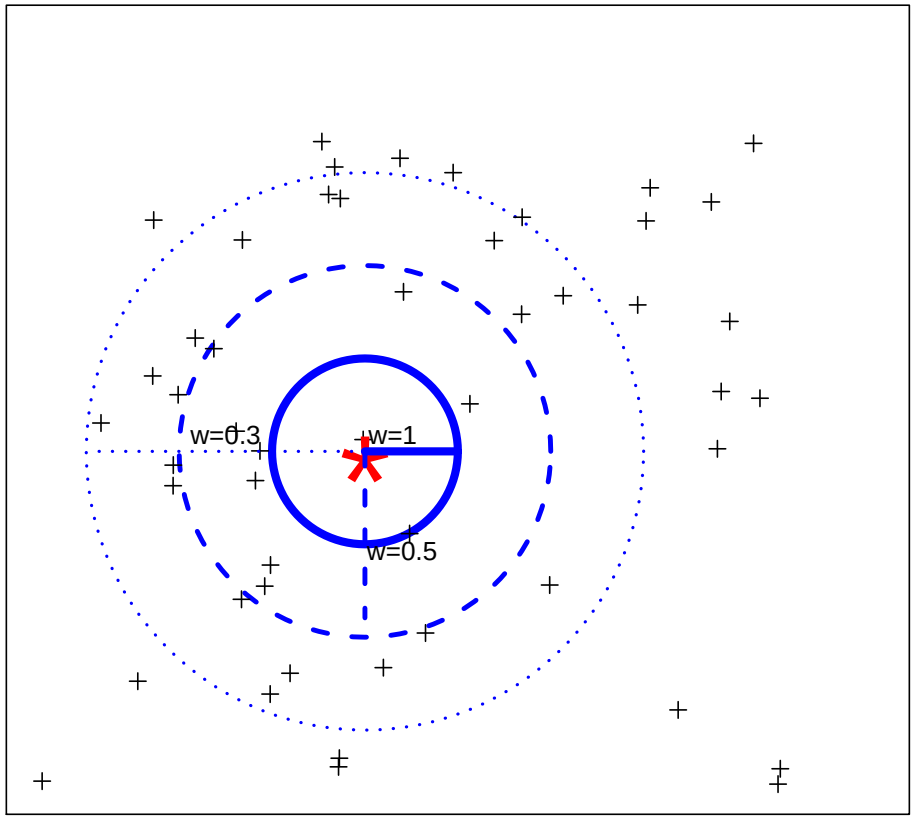
$$\sqrt{\sum_{i=1}^{n} w_i|x_j - x_i|^2}$$

Cas particulier de :

$$[\sum_{i=1}^{n} w_i|x_j - x_i|^k]^{\frac{1}{k}}$$

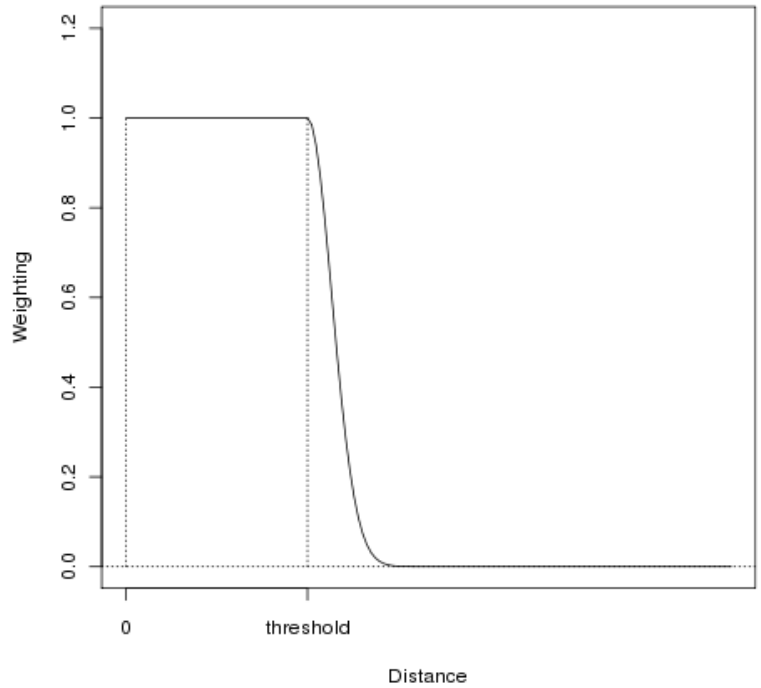
Introduction de *fuzzy set*

une distance $d \rightarrow$ une pondération ω



Objectif : définir une fonction $\omega = f(d)$ telle que le dégré d'appartenance est déterminé par la distance

$$\omega = \begin{cases} 1 & \text{si } d \leq \varepsilon \\ exp(-\pi(d-\varepsilon)^2) & \text{sinon} \end{cases}$$



Rappel

La statistique bayésienne :

$$P[\theta|Y] = \frac{P[\theta] \times P[Y|\theta]}{P[Y]}$$

$$\underset{P[\theta]}{\text{Information priore}} \xrightarrow{\text{data(Y)}} \underset{P[\theta|Y]}{\text{Information postériore}}$$

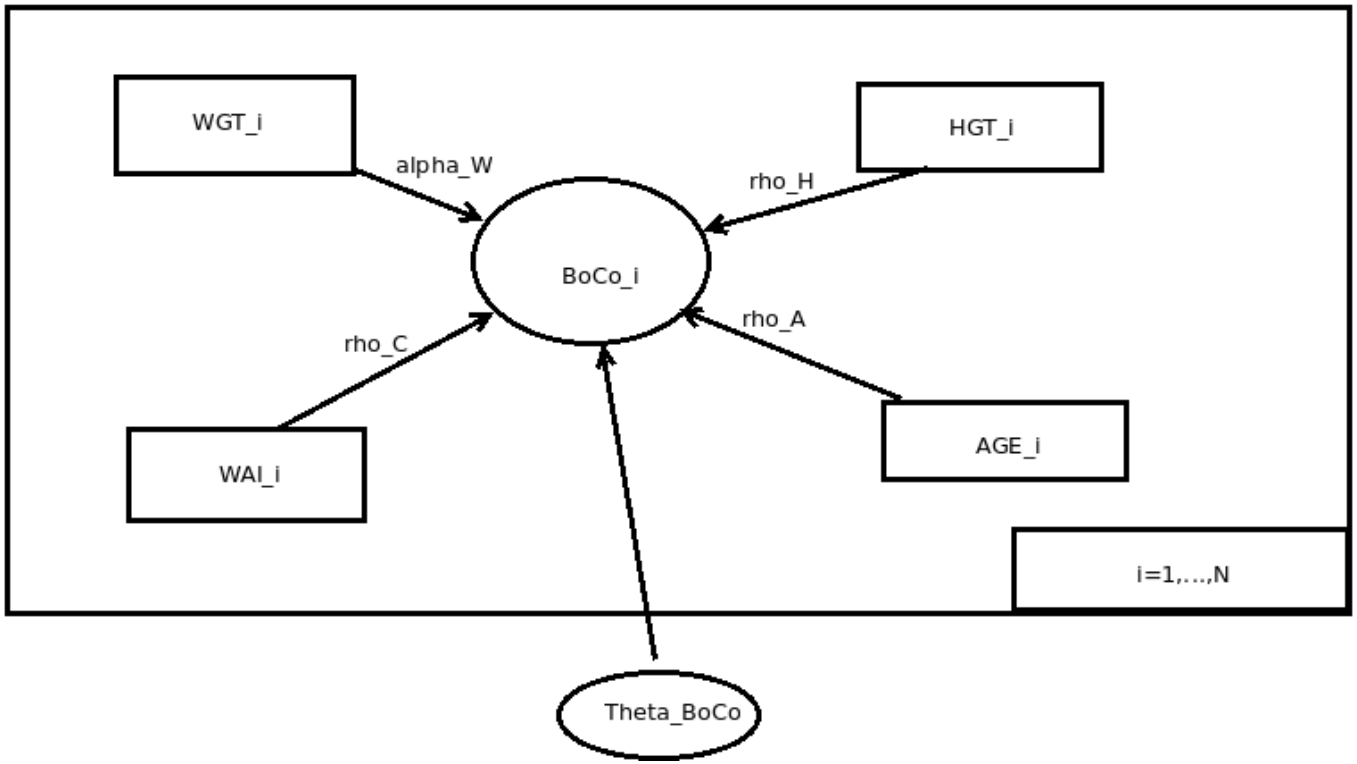
Le modèle linéaire bayésien :

$$Y|X_1, \ldots, X_p \sim \mathcal{N}\left(\mu(\beta, X_1, \ldots, X_p), \sigma^2\right)$$

avec

$$\begin{aligned} \mu(\beta, X_1, \ldots, X_p) &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j X_j \end{aligned}$$

DAG


WGT_i
HGT_i
alpha_W
rho_H
BoCo_i
rho_C
rho_A
WAI_i
AGE_i
i=1,...,N
Theta_BoCo

Modélisation

Dans notre cas, on va proposer un modèle linéaire bayésien :

$$\begin{aligned} \mu_Y &= \alpha_W \cdot WGT + \rho_A \cdot AGE \\ &\quad + \rho_H \cdot HGT + \rho_C \cdot WAI \\ Y &\sim \mathcal{N}(\mu_Y, \frac{1}{\sigma_Y^2}) \end{aligned}$$

- Y un compartiment à prédire ;
- $\alpha_W, \rho_A, \rho_H, \rho_C$ et σ_Y^2 paramètres.

Loi priore des paramètres

Il faut spécifier la loi priore des paramètres :

- $\alpha_W \sim \mathcal{N}(\mu_{\alpha_W}, s^2_{\alpha_W})$
- $\rho_A \sim \mathcal{N}(\mu_{\rho_A}, s^2_{\rho_A})$
- $\rho_H \sim \mathcal{N}(\mu_{\rho_H}, s^2_{\rho_H})$
- $\rho_C \sim \mathcal{N}(\mu_{\rho_C}, s^2_{\rho_C})$
- $\sigma_Y \sim \mathcal{U}(a, b)$

Comment donner une loi priore ?

1. demander à l'expert du domaine
2. mettre de grandes variances

Proposition des priores

Table: Propostion des distributions des paramètres pour bF

Noeud	Parents	paramètre associé	loi priore
bF	WGT	α_W	$\mathcal{N}(0.5, \frac{1}{0.1^2})$
	AGE	ρ_A	$\mathcal{N}(0, \frac{1}{0.04^2})$
	HGT	ρ_H	$\mathcal{N}(0, \frac{1}{0.05^2})$
	WAI	ρ_C	$\mathcal{N}(0, \frac{1}{0.06^2})$
		σ_{bF}	$\mathcal{U}(0, 4)$

Régression Bayésienne Locale

On suppose que le compartiment Y se modélise en fonction des covariables :

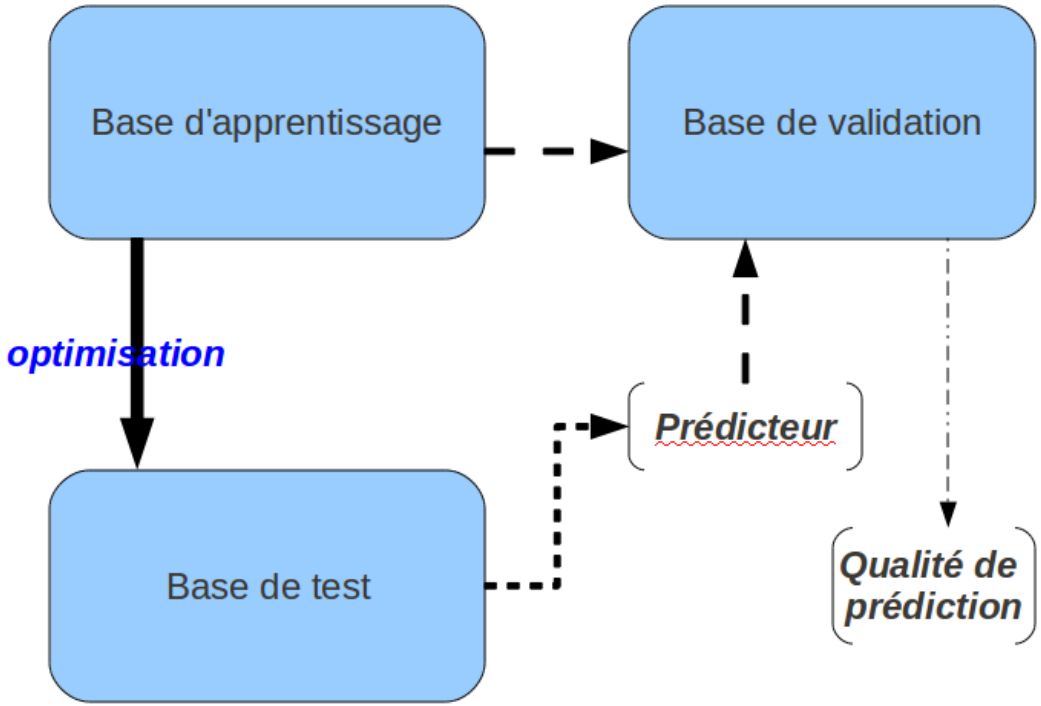
1. $\mu_Y \simeq RBL(WGT, AGE, HGT, WAI)$

$$\mu_Y = \alpha_W \cdot WGT + \rho_A \cdot AGE + \rho_H \cdot HGT + \rho_C \cdot WAI$$

2. les paramètres α_W, ρ_A, ρ_H et ρ_C suivent une loi priore ;
3. le compartiment Y_i d'un individu i suit une loi priore avec une pondération :

$$Y_i \sim \mathcal{N}(\mu_Y, \frac{\omega_i}{\sigma_Y^2})$$

Structure des bases



Critères de qualité

Pour comparer plusieurs modèles, certains critères sont mis en place :

- SEP1 :

$$SEP1 = \frac{1}{n}\sum_{i=1}^{n}|obs_i - pred_i|$$

- SEP2 :

$$SEP2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(obs_i - pred_i)^2}$$

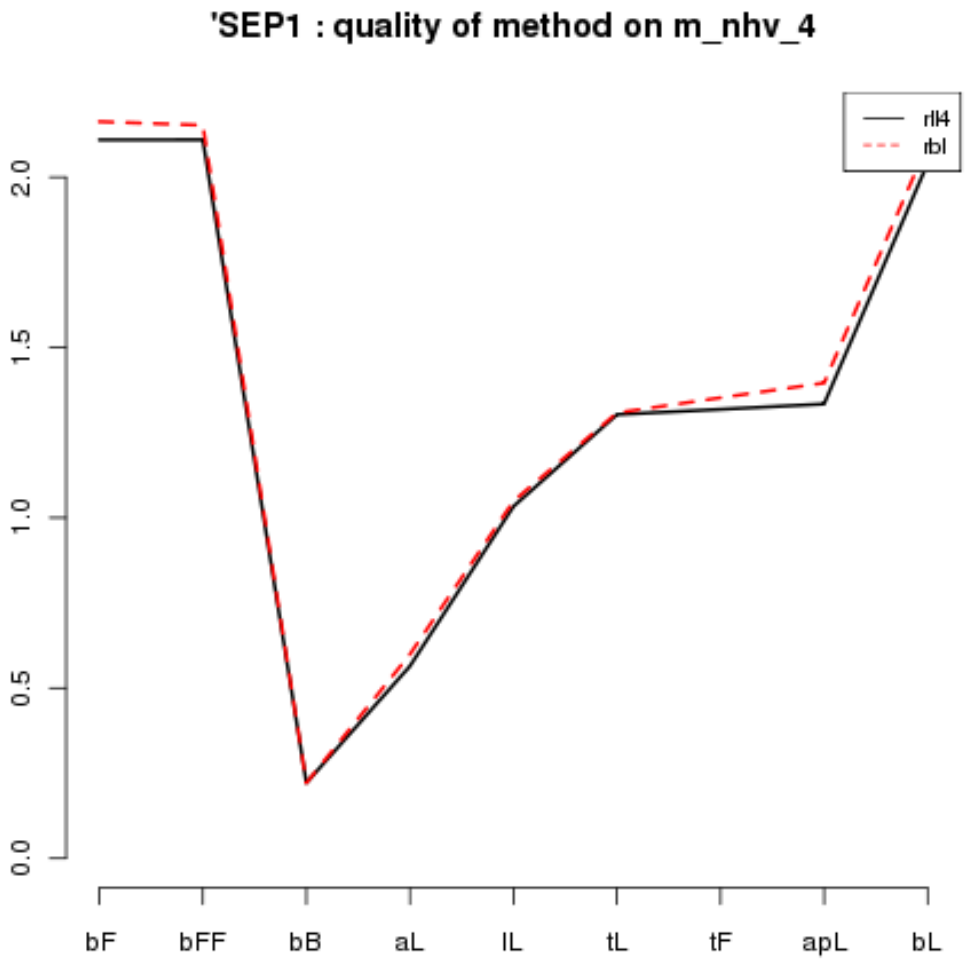
- REP1 :

$$REP1 = \frac{1}{n}\sum_{i=1}^{n}|\frac{obs_i - pred_i}{obs_i}|$$

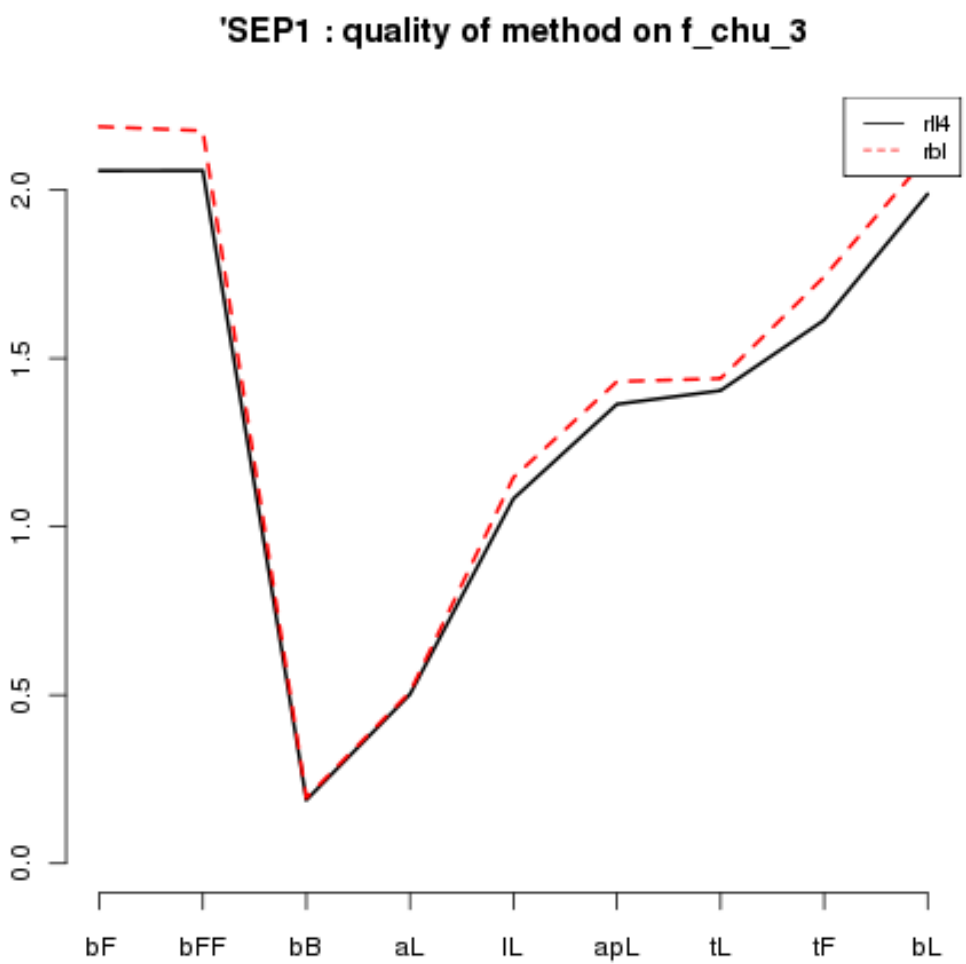
- REP2 :

$$REP2 = \sqrt{\frac{\sum_{i=1}^{n}(\frac{obs_i - pred_i}{obs_i})^2}{n}}$$

Résultat global - NHANES chez les hommes



Résultat global - CHU chez les femmes



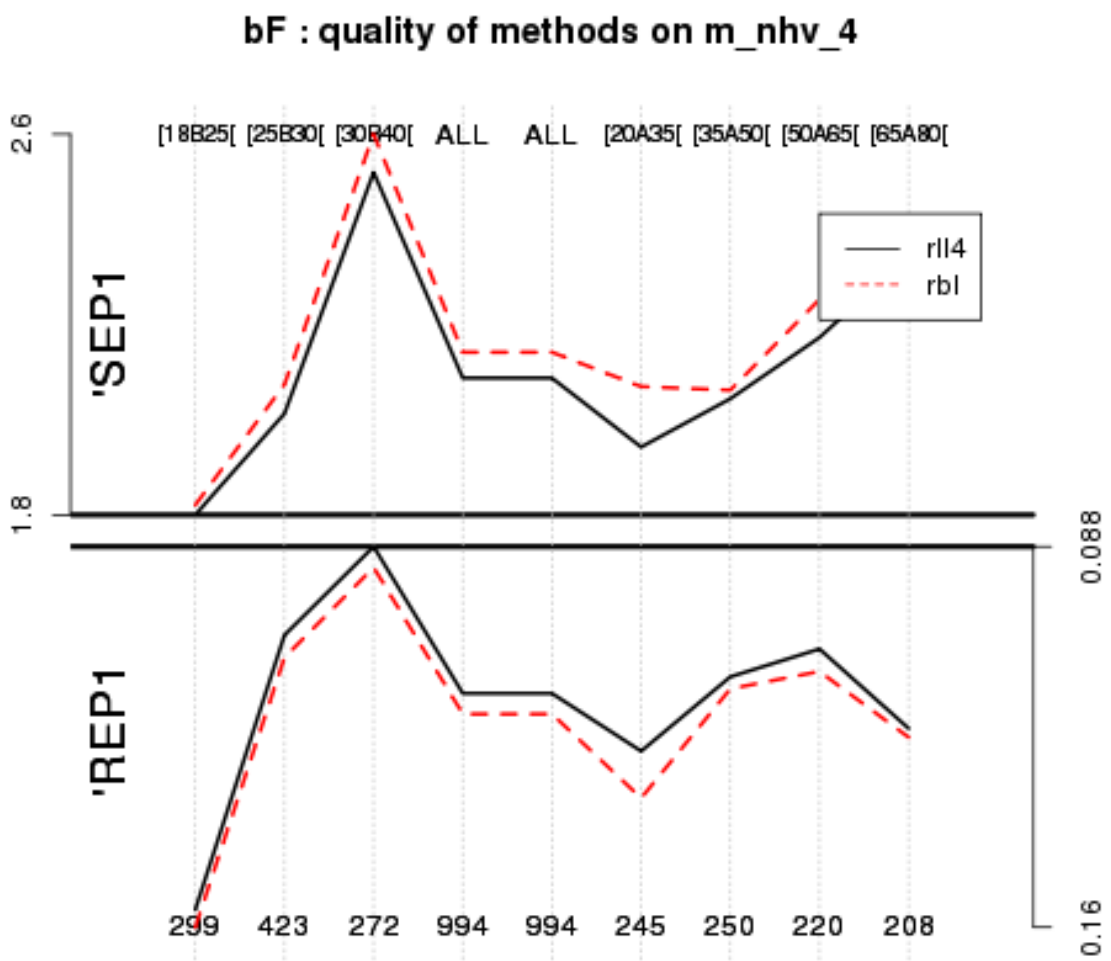
Conclusions

Conclusions, Questions :

1. RLL est meilleur que RBL ;
2. RBL est plus coûteux en temps de calcul ;
3. Néanmoins, RBL est plus flexible :
   - prédire non pas une valeur, mais une distribution ;
   - dépasser la contrainte des valeurs manquantes, ...

Toutes les remarques sont bienvenues.

Résultat par catégories - NHANES chez les hommes



Résultat par catégories - CHU chez les femmes