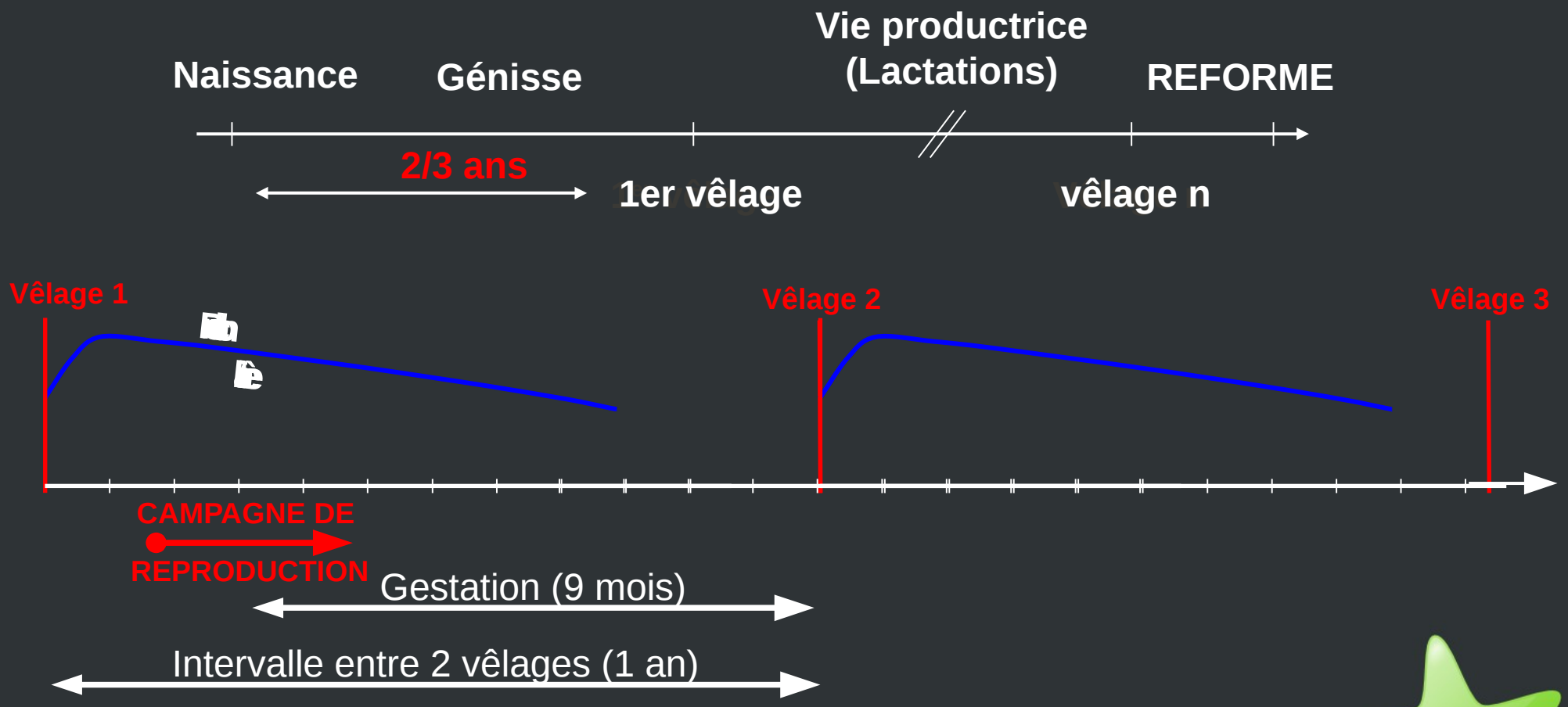
Prédiction d’un événement binaire à partir de données fonctionnelles : Application aux bovins laitiers

C. Sauder, C. Disenhaus, Y. Le Cozler , H. Cardot

Thèse co-financée INRA - Région Bretagne

A dairy cow life

Durabilité animale

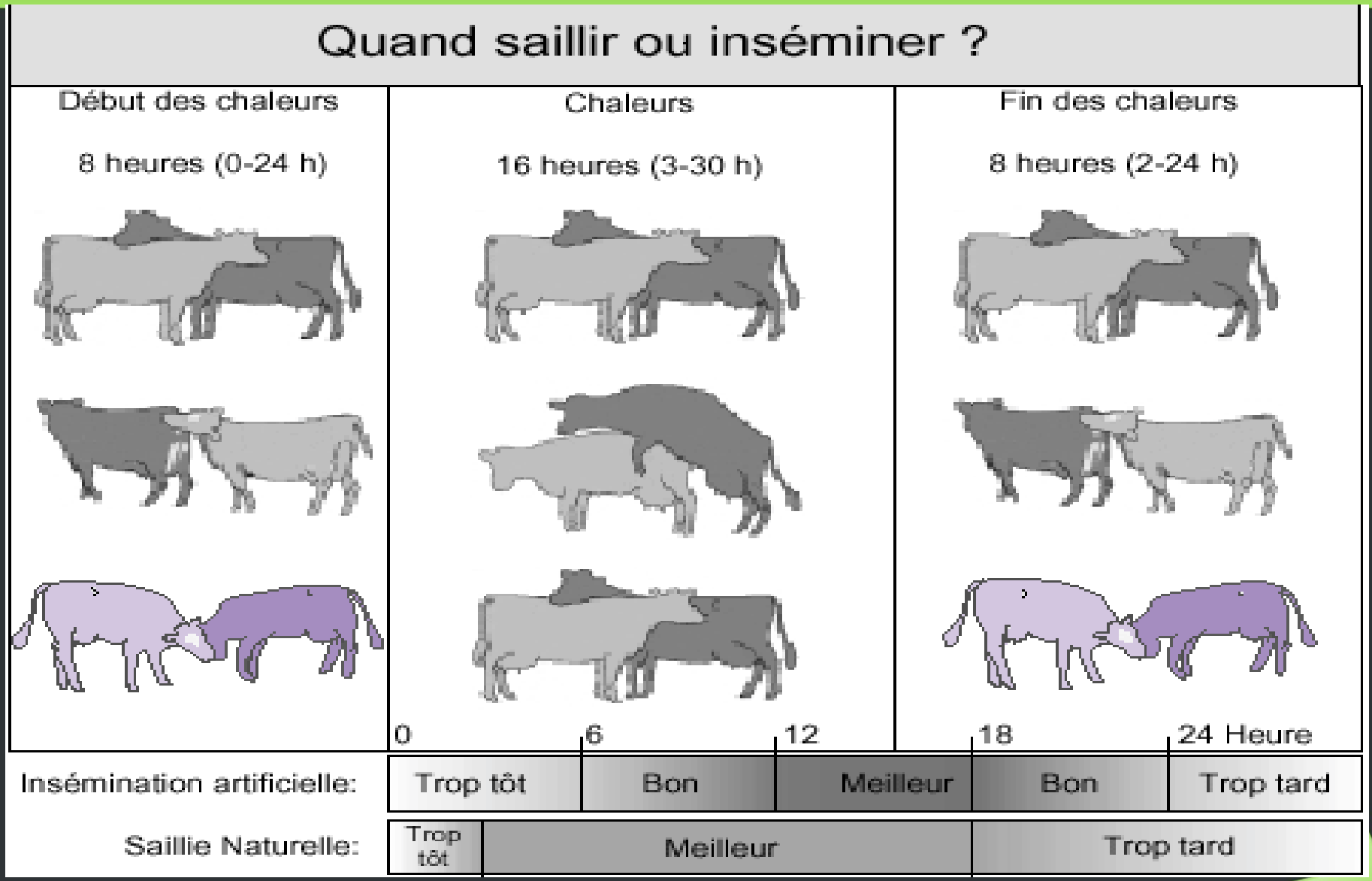


Question

- Peut-on prédire si la vache va être gestante (pregnant) à la première insémination grâce à sa courbe de lactation avant insémination (42 jours) ?

Sexy cows 1.0

Sexy cows 2.0


Quand saillir ou inséminer ?
Début des chaleurs
8 heures (0-24 h)
Chaleurs
16 heures (3-30 h)
Fin des chaleurs
8 heures (2-24 h)
0
6
12
18
24 Heure
Insémination artificielle:
Trop tôt
Bon
Meilleur
Bon
Trop tard
Saillie Naturelle:
Trop tôt
Meilleur
Trop tard

Problématique

- Prédire si Y=0 ou 1 à partir de X(t)
  - Y = 1 : succès de l'insémination
  - X(t) : courbe de lactation (t=1,...., 42)

Logistic regression

- GLM, binomial function, logit link
- Success probability $\pi_i = P(Y = 1 | X = x_i(t); t \in T)$

- 'Classical '

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j \in T} \beta_j x_{ij}$$

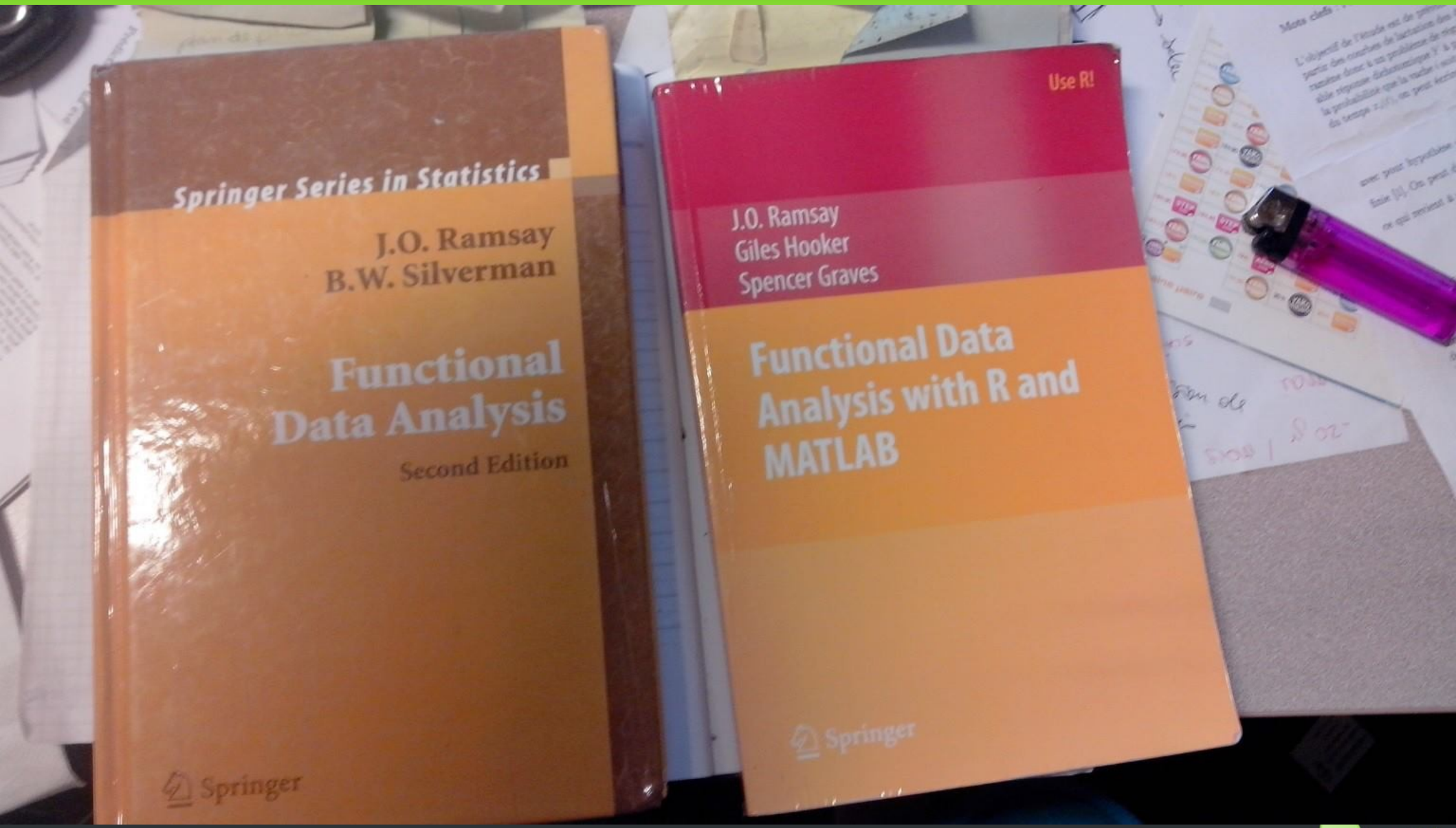
- Functional

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

'Classical' logistic regression

```
17  ypred=c()
18  foreach(i=1:362) %do% {
19  glmC<-glm(gIA1_42~., family=binomial, data=tab2[-c(i),])
20  pred=predict(glmC,newdata=tab2[i,-43],  type='response')
21  ypred=c(ypred,pred )
22  }
23
24  pred=ifelse(ypred>=0.5, 1,0)
```

FDA


Springer Series in Statistics
J.O. Ramsay
B.W. Silverman
Functional Data Analysis
Second Edition
Springer
Use R!
J.O. Ramsay
Giles Hooker
Spencer Graves
Functional Data Analysis with R and MATLAB
Springer

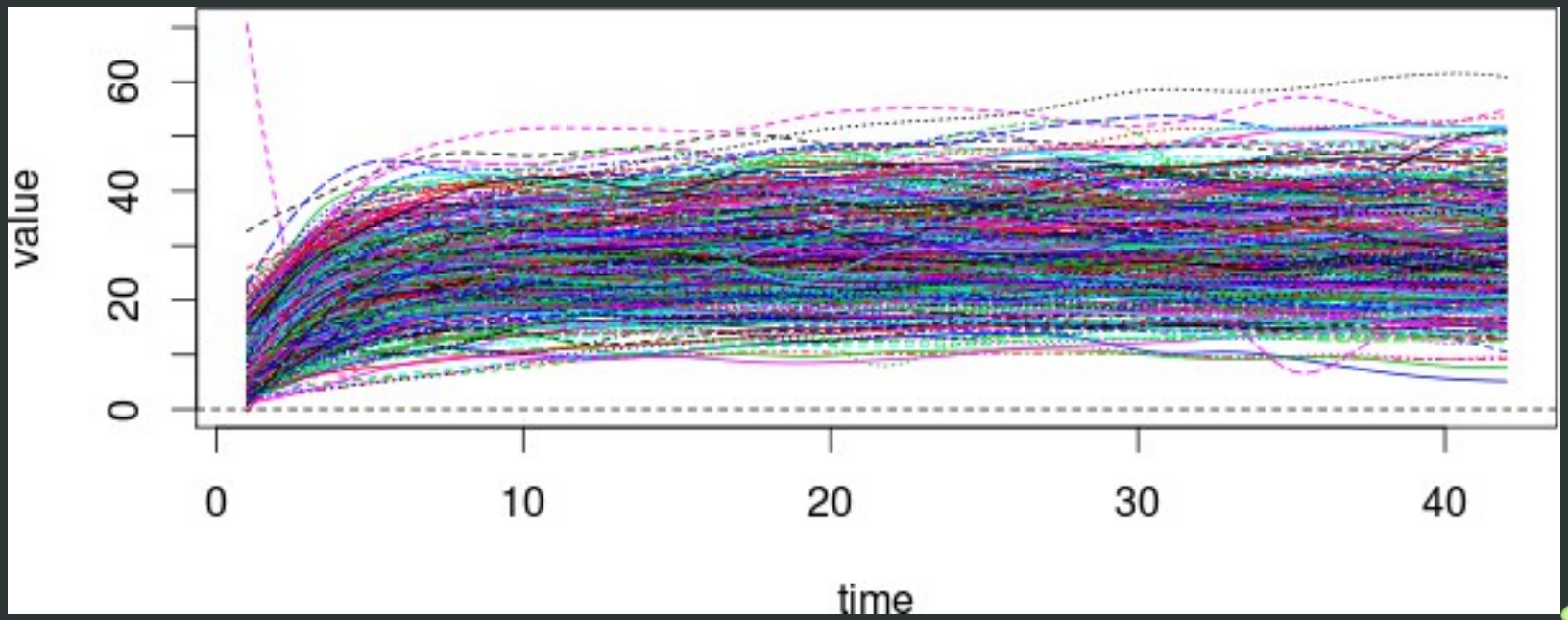
fda et fda.usc

	fda : Functional Data Analysis	fda.usc : and Utilities for Statistical Computing
authors	J. O. Ramsay, Hadley Wickham, Spencer Graves, Giles Hooker	Manuel Febrero-Bande, Manuel Oviedo de la Fuente
R class	« fd » object	« fdata » object

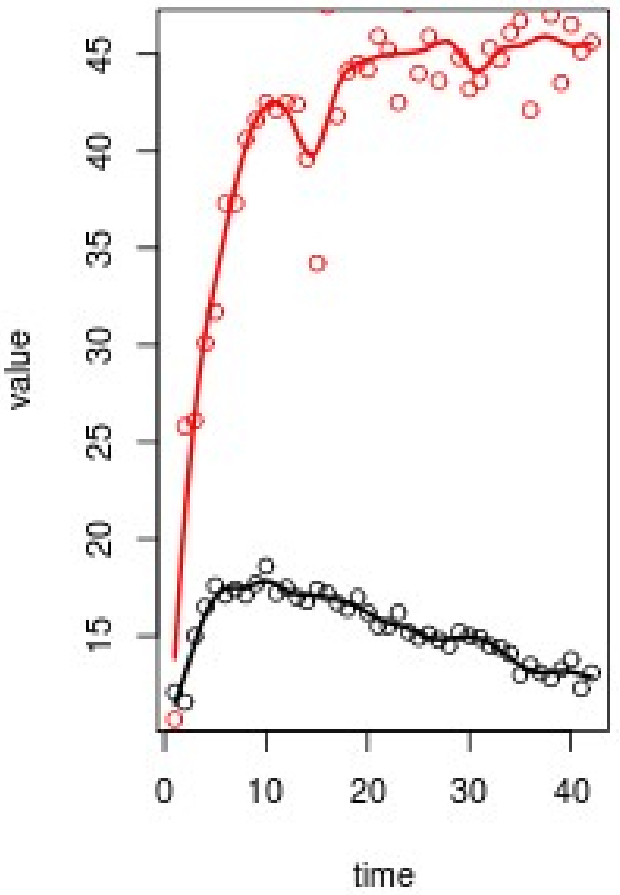
fda

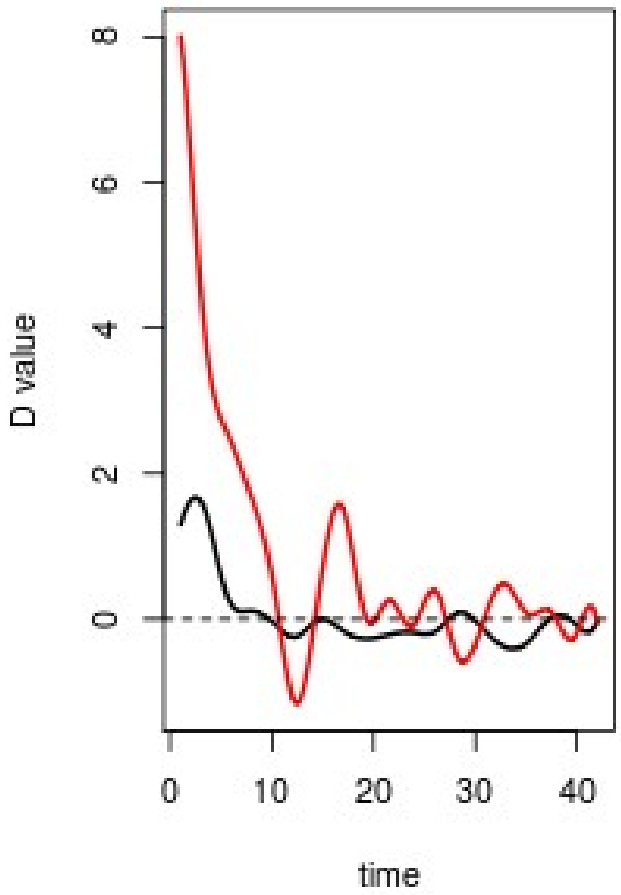
1. basis = create.bspline.basis(
   c(1,42), nbasis=36, norder=4)

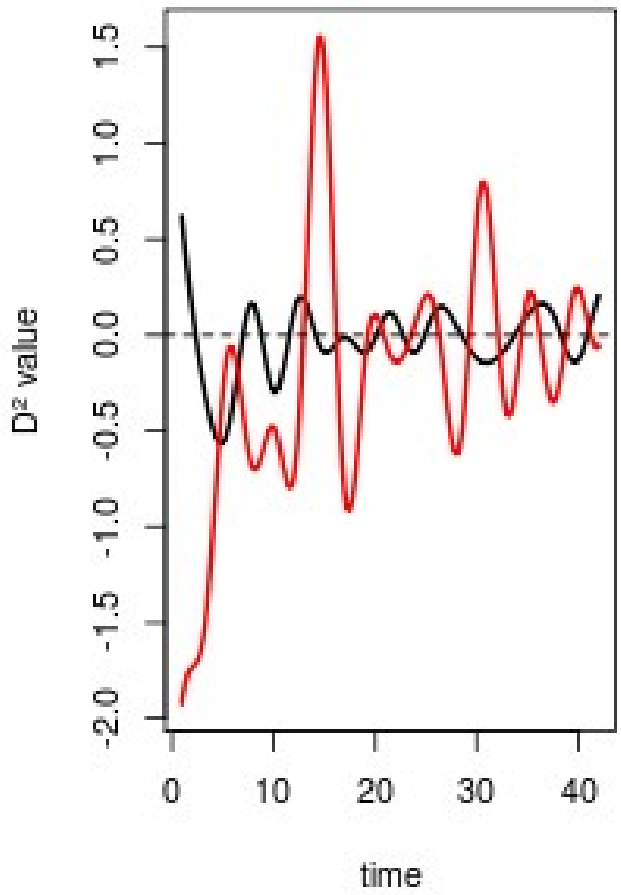
2. Smoothing data with a roughness penalty :
   fdParobj = fdPar(fdobj=basis, Lfdobj=2, lambda=1)
   plfd = smooth.basis(jr,plmat,fdParobj=fdParobj)$fd



Dérivées : deriv.fd()

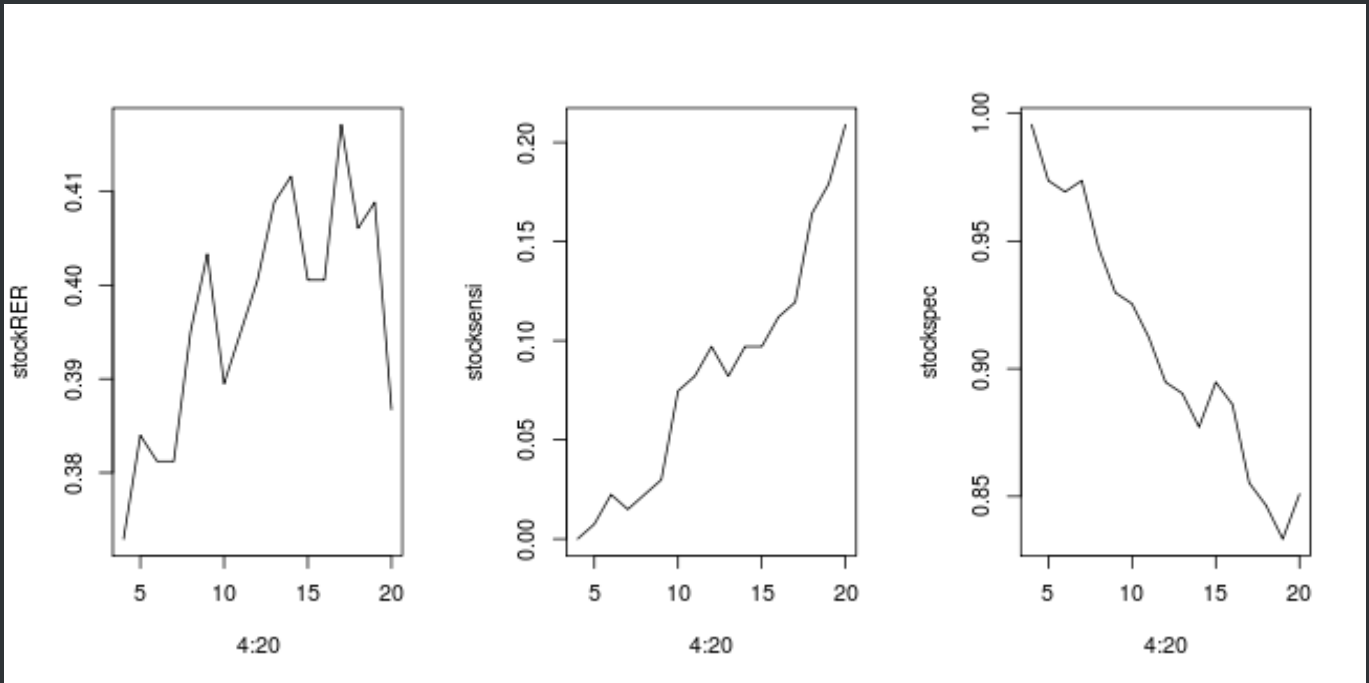




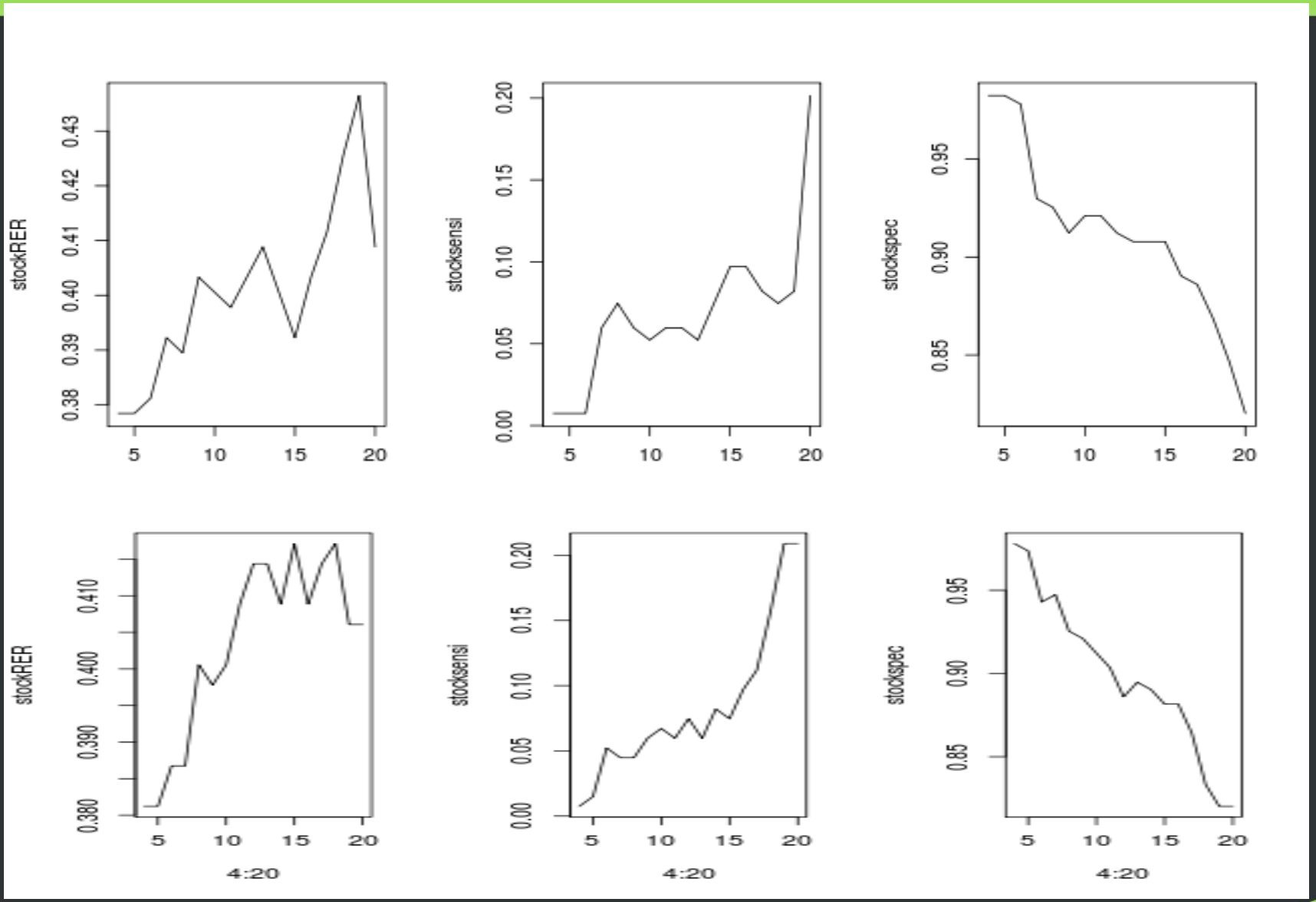


Functional logistic regression

- fregre.glm() (package fda.usc)
- Trouver la meilleure base pour la regression
  - CV leave one out pour 4 à 20 splines pour la base



First and second derivatives



Comparison results

	Functional GLM			GLM
	y~x	y~x'	y~x''	y~x1+...+x42
RER	38.8	40.9	40.6	42.3
Sensibility	22.2	20.1	20.9	21.6
Specificity	85.1	82.0	82.0	78.9

y.pred	y 0	1
0	194	106
1	34	28

Vpred	0	1
0	187	41
1	107	27

Vpred	0	1
0	187	41
1	106	28

pred	0	1
0	180	48
1	105	29

Conclusion

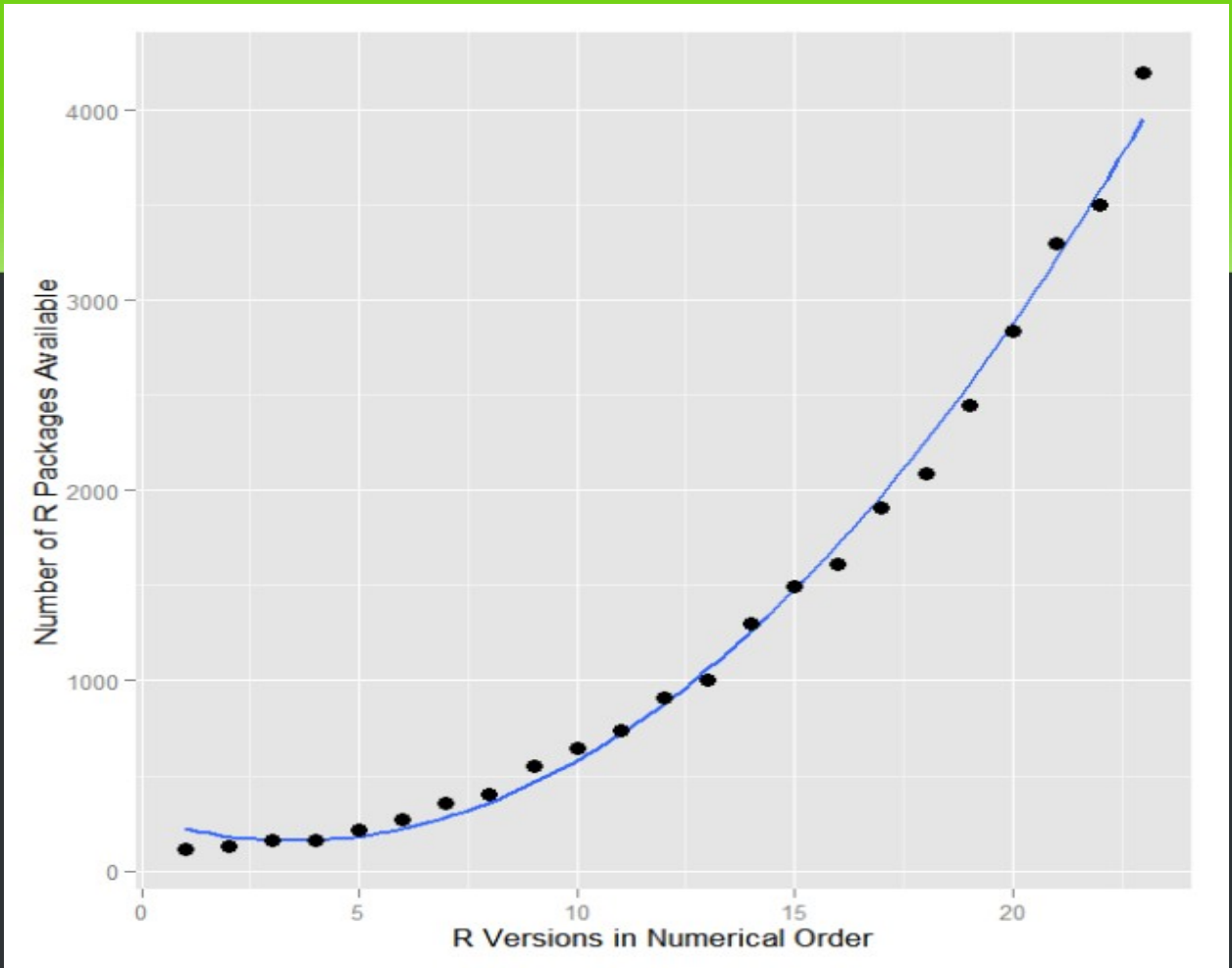
- Only milk yield curves is not sufficient to predict this binary event
- Too many R packages (>4000)

Merci de votre attention

References

[1] Manuel Febrero-Bande and Manuel Oviedo de la Fuente. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012.

[2] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2005.



Number of R packages plotted for each major release of R. The last value on the x-axis represents version 2.15.2, the final release in 2012.

Robert A. Muenchen, The Popularity of Data Analysis Software, 2013
http://r4stats.com/

Warning

- Simplifier / valider packages existants plutôt qu'en construire de nouveaux parce qu'on est pas sûr de ce qu'il y a dedans et qu'on préfère tout reprogrammer pour être sûr de ce qu'on fait.
- Ok mais le but est que des utilisateurs puissent se servir des packages, pas seulement la personne qui les a construit
- Pourquoi tant de packages ?
    - Publication et reconnaissance scientifique

Cross validation 70/30

```
rForest=randomForest
```

```
        lm    glm  rpart tree REEMtree ctree rForest cforest
1 RER  41.87 41.94 44.73 45.08  33.86  37.02   40.84   38.30
2 Sens 25.29 26.95 36.95 39.09  52.02   0.00   19.48   13.17
3 Spec 77.27 76.17 66.05 64.15  74.41 100.00   82.20   89.89
```

Cross validation 70/30

```
rForest=randomForest
```

```
        lm    glm  rpart tree REEMtree ctree rForest cforest
1 RER 41.87  41.94 44.73 45.08   33.86  37.02   40.84   38.30
2 Sens 25.29 26.95 36.95 39.09   52.02   0.00   19.48   13.17
3 Spec 77.27 76.17 66.05 64.15   74.41 100.00   82.20   89.89
```

Regression Trees with Random Effects for Longitudinal (Panel) Data