# ‘*Package ROP’: Arbres à  Noeuds Multivariés*

*Jean-Michel NGUYEN*

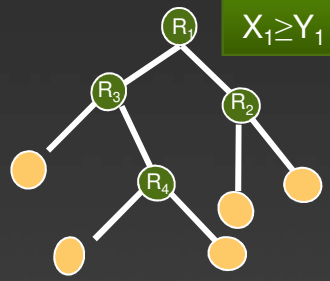
*R Rencontres*

*Rennes - 05 Juillet 2018*

CRCINA
Centre de Recherche en Cancérologie et Immunologie
Nantes Angers

# Les arbres de décisions « Classiques »

## ARBRES A NŒUDS MONOVARIES




$X_1 \geq Y_1$

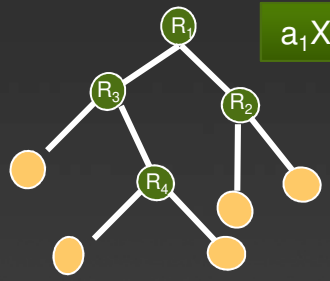
A chaque nœud, une seule variable est utilisé pour départager en n groupes l'échantillon.

Il existe une hiérarchie très importante des variables. La variable du premier nœud est déterminante et conditionne le reste de l'arbre.

Il n'y a pas de retour possible en arrière

Performances diagnostiques « modestes »

## ARBRES A NŒUDS MULTIVARIES




$a_1X_1+a_2X_2+\ldots \geq Y_1$

A chaque nœud, plusieurs variables sont utilisées pour départager en n groupes l'échantillon.

Pb d'estimation des coefficients

Il n'y a pas de retour possible en arrière

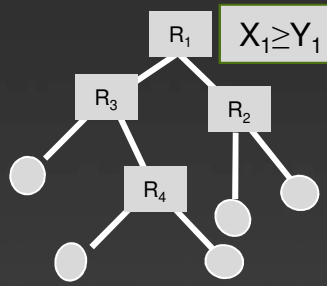


*Jean-Michel NGUYEN*

*05 Juillet 2018*

# Les innovations de ROP (Régression OPtimisée)

**ARBRES A NŒUDS MONOVARIES**



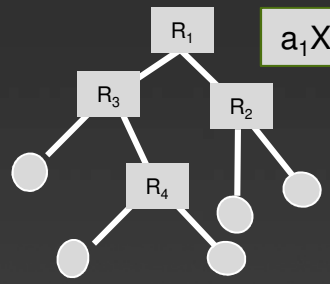

A chaque nœud, une seule variable est utilisé pour départager en n groupes l'échantillon.

Il existe une hiérarchie très importante des variables. La variable du premier nœud est déterminante et conditionne le reste de l'arbre.

Il n'y a pas de retour possible en arrière

Performances diagnostiques « modestes »

**ARBRES A NŒUDS MULTIVARIES**




A chaque nœud, plusieurs variables sont utilisées pour départager en n groupes l'échantillon.

Pb d'estimation coefficients

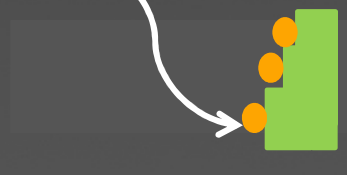
Il n'y a pas de retour possible en arrière

**2 INNOVATIONS MAJEURES**

Simulation d'entiers relatifs ex: [-5;+5] déterminés par analyse combinatoire exhaustive. Optimisation sur plusieurs critères (Se+Sp) , AUC…

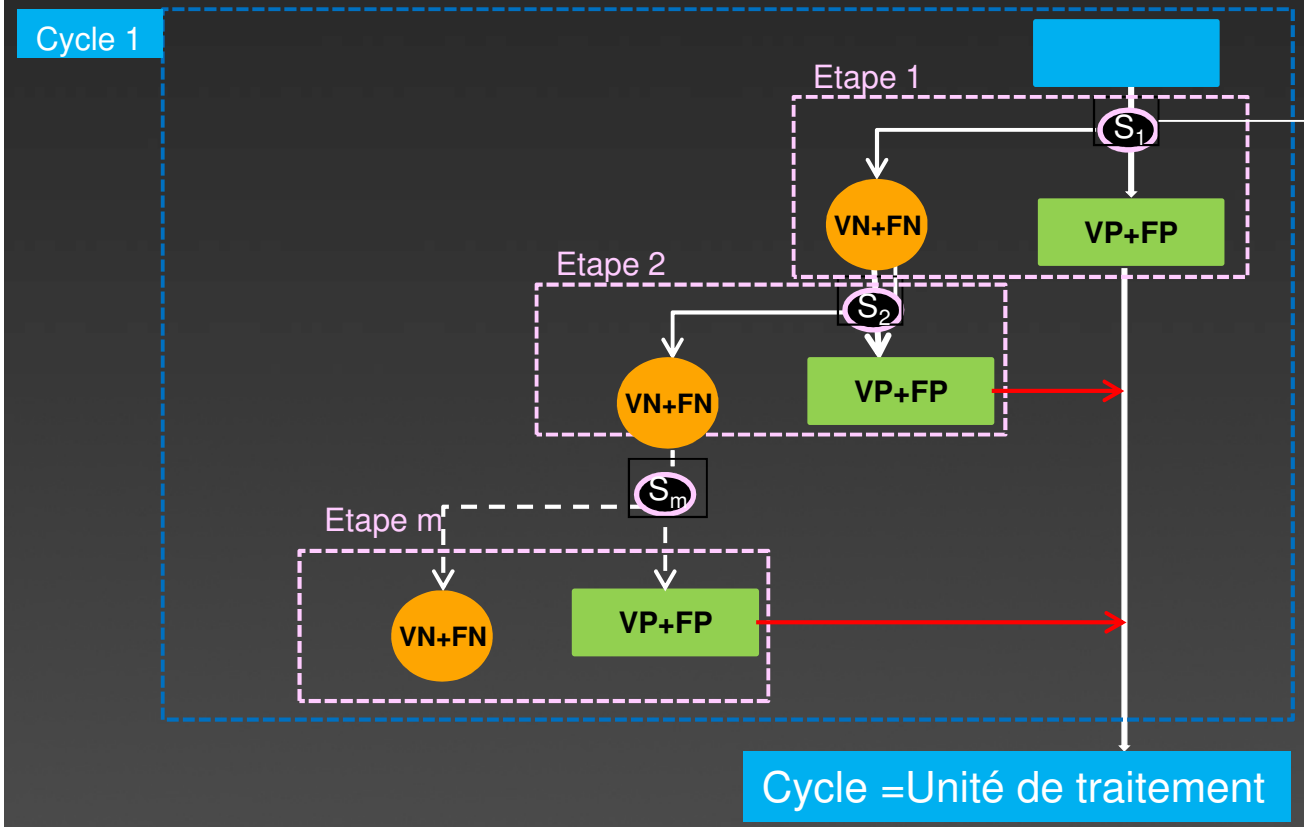
Lecture directe des effets des variables sur l'état Y à prédire :
Coef > 0 ⟷ Augmentation du risque
Coef < 0 ⟷ Diminution du risque
Coef = 0 ⟷ Facteur de confusion

Une partie des feuilles est réinjectée dans le tronc
= Effet feed-Back
= nouveau arbre de décision 'HCT'
= Half Christmas Tree

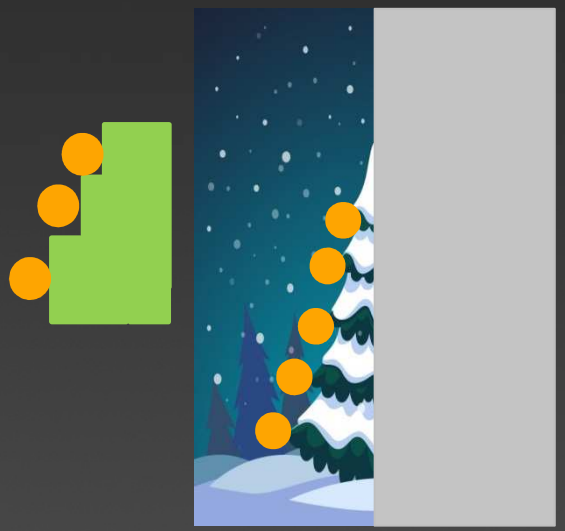


# ROP : UNE STRUCTURE EN DEMI-SAPIN DE NOEL
## Cycles et Etapes

Arbres ROP $C_nE_m$ à nm noeuds



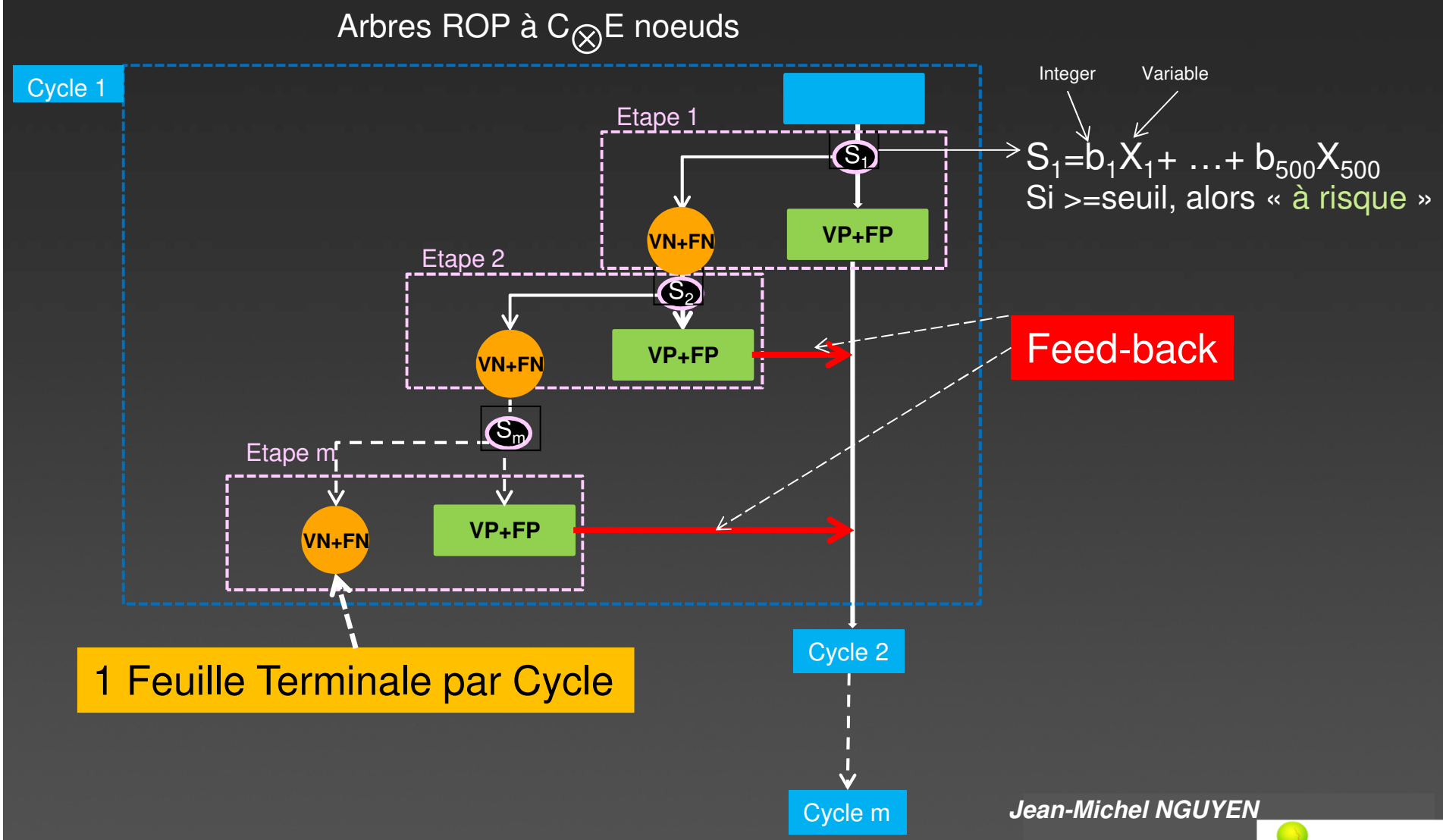

$S_1=b_1X_1+ \ldots + b_{500}X_{500}$
Si >=seuil, alors « à risque »



*Jean-Michel NGUYEN*

*05 Juillet 2018*

# ROP : UNE STRUCTURE EN DEMI-SAPIN DE NOEL
## Cycles et Etapes



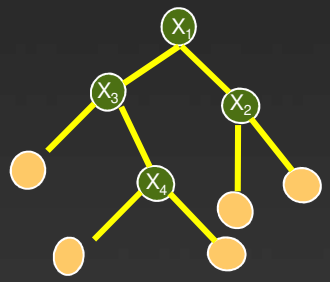



*Jean-Michel NGUYEN*

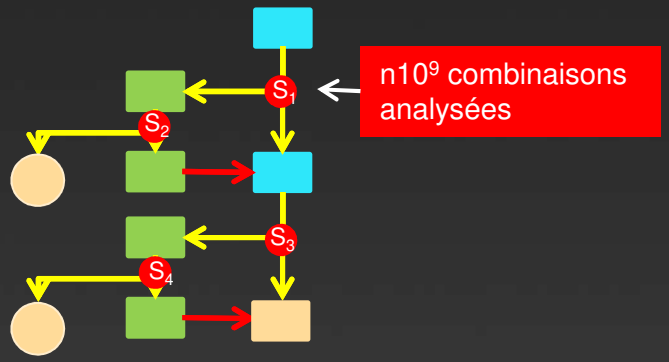
*05 Juillet 2018*

# Au total pour le prix de 4 nœuds….

## Arbre classique




## Arbre ROP




Le modèle ROP est un arbre cumulant **2 innovations structurelles majeures**, en plus des innovations algorithmiques.

- Arbre multivarié: les nœuds sont des scores de risque incluant plusieurs variables.
- Structure : une partie des feuilles terminales sont réinjectées dans le tronc principal.
- Performances supérieures à la régression logistique pour une robustesse équivalente

Les scores de risques sont calculés en analysant l'exhaustivité des combinaisons de toutes les variables et de tous les coefficients de risque de chaque variables.
Exemple : 10 variables, risque attribuable précision de 1 unité [-2;+2] = $5^{10}$ combinaisons analysées à chaque nœud. Pour un arbre à 4 nœuds, 200 milliards de combinaisons seront analysées et classées.

***Jean-Michel NGUYEN***

***05 Juillet 2018***

# Pour quelles performances ?

## Validation interne par boostrap (Steyerberg EW, Journal of Clinical Epidemiology 54 (2001) 774–781)

| | Variables Included | Apparent-Optimism | ROP | Logistic Regression | | CART | Discriminant Analysis |
|---|---|---|---|---|---|---|---|
| TITANIC | AGE,SEX,CLASS | Sensitivity Estimated | 0.854530475 | 0.844908201 | | 0.845924113 | 0.822594859 |
| | | Specificity Estimated | 0.672800434 | 0.681503006 | | 0.670681363 | 0.696052104 |
| | | Global Performance Es | 😎1.527 | 🔫1.526 | | 🔫1.517 | 🔫1.519 |
| | | | | | | | |
| ICU | SER,GENDER,RACE, CRN,SYS,PRE | Sensitivity Estimated | 0.697 | 0.631 | | 0.419 | 0.586 |
| | | Specificity Estimated | 0.733 | 0.779 | | 0.887 | 0.759 |
| | | Global Performance Est | 😎1.429 | 🔫1.410 | | 🔫1.306 | 🔫1.345 |
| | | | | | | | |
| LOWBWT | SMOKE,AGE,LWT,P TL,HT | Sensitivity Estimated | 0.708 | 0.618 | | 0.513 | 0.643 |
| | | Specificity Estimated | 0.657 | 0.721 | | 0.803 | 0.683 |
| | | Global Performance Est | 😎1.365 | 🔫1.338 | | 🔫1.316 | 🔫1.325 |
| | | | | | | | |
| PHARYNX | COND, SEX, TX, GRADE, T_STAGE | Sensitivity Estimated | 0.652 | 0.615 | | 0.644 | 0.614 |
| | | Specificity Estimated | 0.598 | 0.625 | | 0.565 | 0.612 |
| | | Global Performance Est | 😎1.250 | 🔫1.240 | | 🔫1.209 | 🔫1.226 |
| | | | | | | | |
| PROSTATE | RAYONX, AGE, ACIDE, TAILLE, GRADE | Sensitivity Estimated | 0.668 | 0.615 | | 0.644 | 0.614 |
| | | Specificity Estimated | 0.579 | 0.625 | | 0.565 | 0.612 |
| | | Global Performance Est | 😎1.247 | 🔫1.240 | | 🔫1.209 | 🔫1.226 |



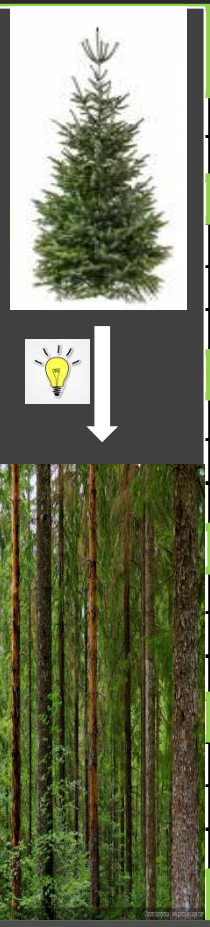
*Jean-Michel NGUYEN*

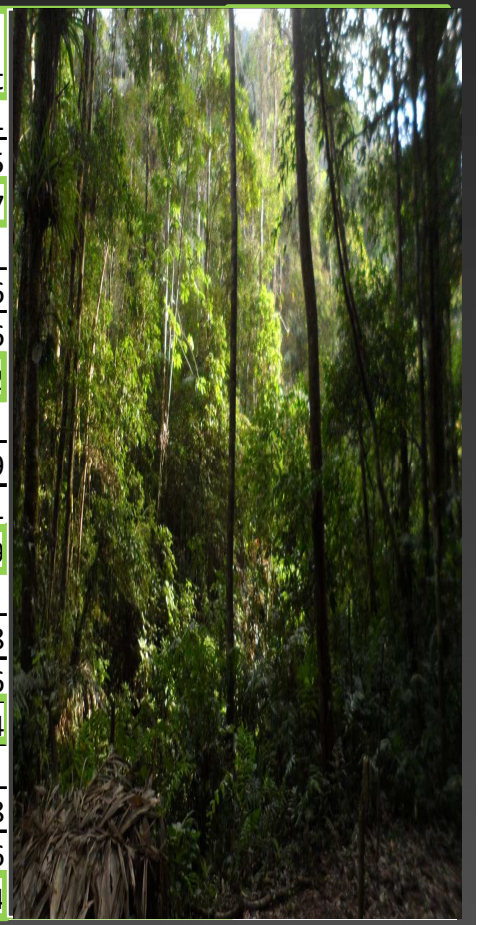
*05 Juillet 2018*

# On ne peut pas comparer un arbre ROP à une Forêt Aléatoire
# Mais une forêt de ROP à une Forêt Aléatoire

## Validation interne par boostrap (Steyerberg EW, Journal of Clinical Epidemiology 54 (2001) 774–781)

| | Variables Included | Apparent-Optimism | ROP | Random Forest |
|---|---|---|---|---|
| TITANIC | AGE,SEX,CLASS | Sensitivity Estimated | 0.854530475 | 0.846168911 |
| | | Specificity Estimated | 0.672800434 | 0.681242485 |
| | | Global Performance Es | 1.527 | 1.527 |
| | | | | |
| ICU | SER,GENDER,RACE, | Sensitivity Estimated | 0.697 | 0.766 |
| | CRN,SYS,PRE | Specificity Estimated | 0.733 | 0.806 |
| | | Global Performance Est | 1.429 | 1.572 |
| | | | | |
| LOWBWT | SMOKE,AGE,LWT,P | Sensitivity Estimated | 0.708 | 0.779 |
| | TL,HT | Specificity Estimated | 0.657 | 0.851 |
| | | Global Performance Est | 1.365 | 1.629 |
| | | | | |
| PHARYNX | COND, SEX, TX, | Sensitivity Estimated | 0.652 | 0.718 |
| | GRADE, T_STAGE | Specificity Estimated | 0.598 | 0.596 |
| | | Global Performance Est | 1.250 | 1.314 |
| | | | | |
| PROSTATE | RAYONX, AGE, | Sensitivity Estimated | 0.668 | 0.718 |
| | ACIDE, TAILLE, | Specificity Estimated | 0.579 | 0.596 |
| | GRADE | Global Performance Est | 1.247 | 1.314 |







*Jean-Michel NGUYEN*

*05 Juillet 2018*

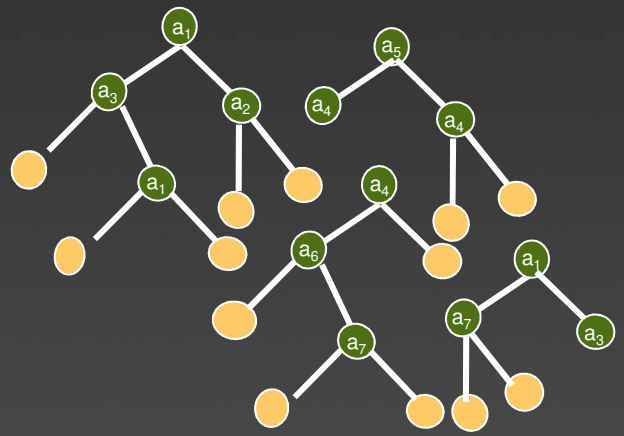
# PERSPECTIVES : FORETS D'ARBRES MULTIVARIES DE ROP

## SOLUTION 1 : Forêts Aléatoire à la « BREIMAN »

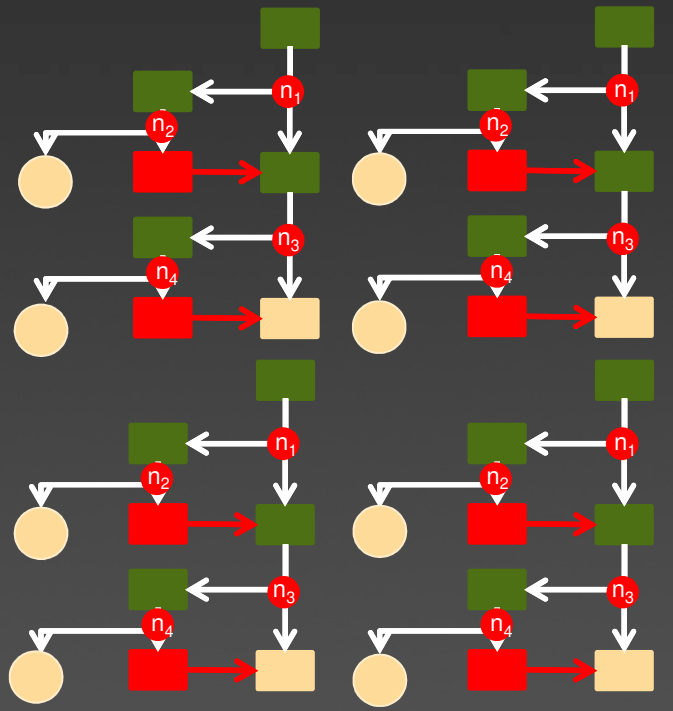
On remplace les arbres CART par des arbres ROP dans l'algorithme des Forêts Aléatoires :

Projet avec l'Ecole Centrale de Nantes

CART




**FAM de ROP**




*Jean-Michel NGUYEN*

*05 Juillet 2018*

# PERSPECTIVES : FORETS D'ARBRES MULTIVARIES DE ROP

## SOLUTION 2 : « Forêts d'Arbres Parfaits» : FAP de ROP

A partir de 15 variables, la probabilité d'avoir un arbre parfait (Se=Sp=AUC=100%) est « importante »

Wisconsin cancer du sein : 30 variables, 569 observations complètes, 30% d'arbres parfaits avec 15 variables TAS

Données omiques : >15000 variables, 82 observations, >50% d'arbres parfaits

VOIR POSTER….



*Jean-Michel NGUYEN*

*05 Juillet 2018*

# PACKAGE 'ROP' pour R CRAN

Auteurs : Nguyen Jean-Michel, Antonioli Daniel

Concurrent direct de la Régression Logistique
Y = Etat Binaire
X= variables explicatives
Qualitative codage binaire (1/0)
Quantitative ordonnée ou continue

**Syntaxe: 5 paramètres à saisir**
rop(fic, mini, maxi, nbCycles, typesVariables)

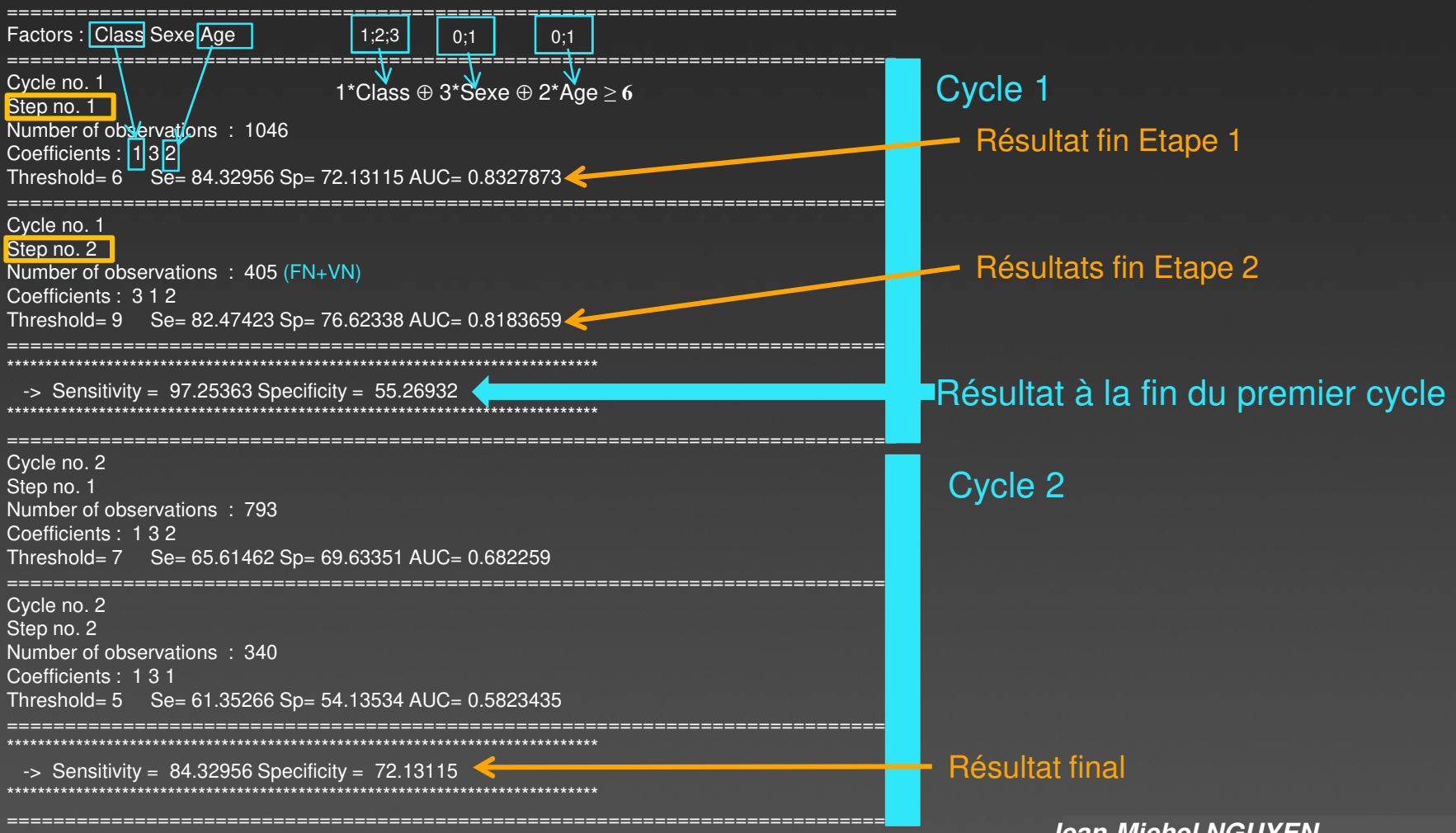
fic= nom du fichier en format csv, exemple: « *titanic* »

min= borne inférieure (entier relatif) de l'étendu des coefficients de risque, ex: « *-5* »
max= borne supérieure (entier relatif) l'étendu des coefficients de risque, ex: « *5* »

nbCycles= nombre de cycles, ex: « *3* ». Le nombre d'étapes est bloqué à 2

typesVariables= « *T* » si quantitative; « *F* » si binaire à 2 classes

```
rop(system.file("extdata", "titanic.csv", package = "ROP"), -3, 3, 2, c(FALSE, FALSE, FALSE)))
```

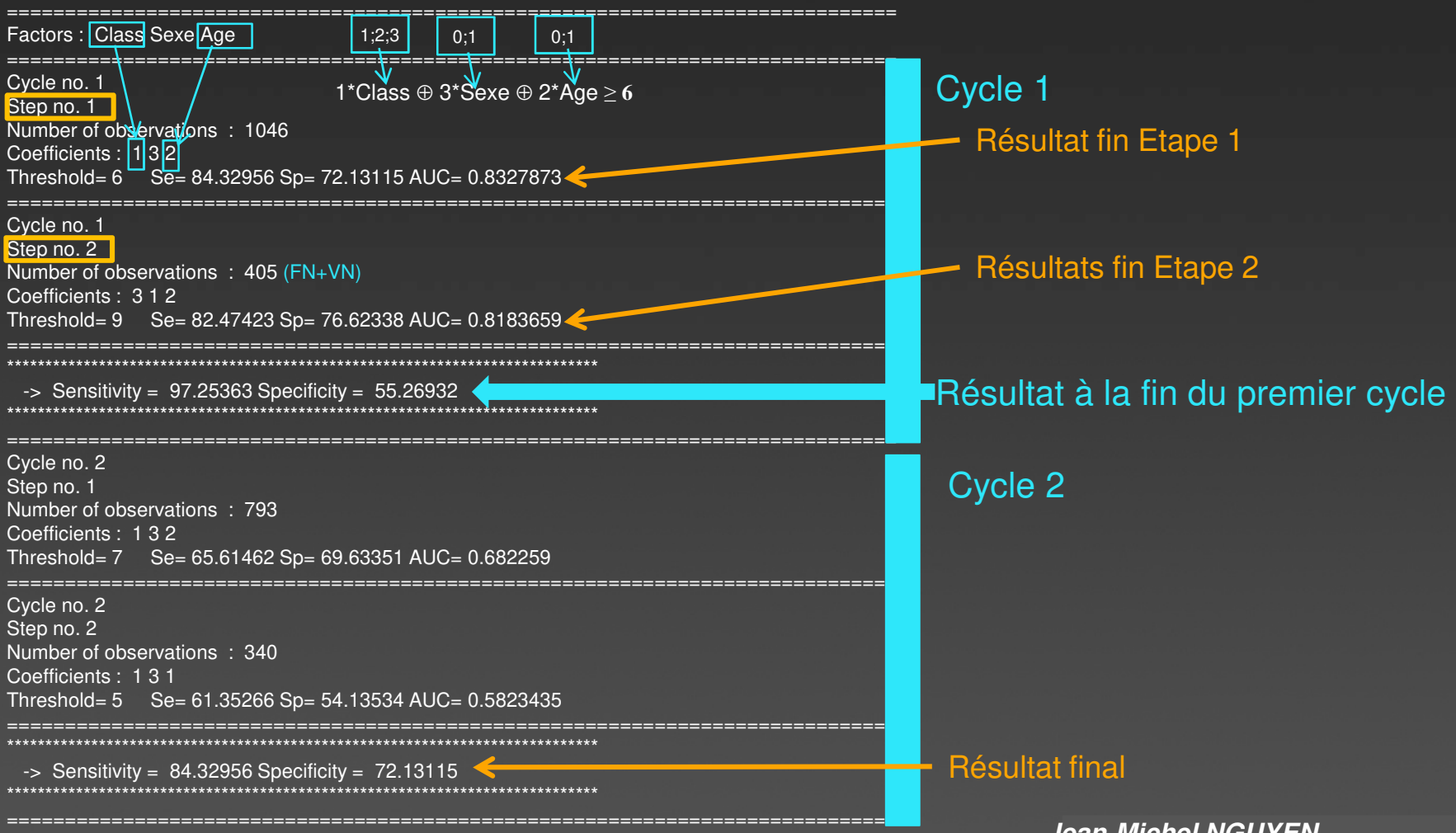





*Jean-Michel NGUYEN*

*05 Juillet 2018*

```
rop(system.file("extdata", "titanic.csv", package = "ROP"), -3, 3, 2, c(FALSE, FALSE, FALSE)))
```






*Jean-Michel NGUYEN*

*05 Juillet 2018*

# PRINCIPALES REFERENCES ROP

Références ROP


1. Nguyen Jean-Michel, Gaultier Aurélie, Antonioli Daniel. A Combined non Parametric Regression and Classification Model, 7th International Meeting - Statistical Methods in Biopharmacy, Paris 16-17 sept 2013.
2. Nguyen JM, Gaultier A, Antonioli D. A Combined NonParametric Regression and Classification Model for Subgroup Selection. ISBC, Munich, 25-29 August 2013.
3. Nguyen Jean-Michel. Blind man's bluff test using ROP- WIN symposium, Paris, July 2013.
4. Nguyen JM, Gaultier A, Antonioli D. How to identify an unknow factor in targeted therapies. P04-074-2nd International Symposium of the Cancer Research Center of Lyon; 21st -23rd September 2015.
5. Nguyen JM, Gaultier A, Antonioli D. Le Titanic revu par ROP, une nouvelle méthode de régression non paramétrique combinée à une classification, Revue d'Epidémiologie et de Santé Publique - Vol. 62 - N° S2 - p. 45-46 - février 2014
6. Nguyen JM, Gaultier A, Antonioli D. Données Fantômes et Régression OPtimisée (ROP). Revue d'Epidémiologie et de Santé Publique. Volume 64, Supplement 3, May 2016, Pages S155–S156. 10e Conférence Francophone d'Épidémiologie Clinique.
7. Nguyen JM, Knol AC,Saint-Jean M, Antonioli D, Gaultier A, Khammari A, Dreno B. Facteurs immunologiques associés à un effet protecteur de l'effraction capsulaire contre la récidive à 6 mois après curage des mélanomes de stade IIIB Ann. Dermatol. Venerol, Vol 143, Supp, Dec 2016, S210.
8. Vildy S, Nguyen JM, Gaultier A, Khammari A, Dreno B . Influence du délai de prise en charge chirurgicale des métastases ganglionnaires sur la survie dans le melanome. Ann. Dermatol. Venerol, Vol 143, Supp, Dec 2016, S373
9. Nguyen JM et all. ROP model in translational research. BIOREGATE,European Regenerative Medicine Forum 8 - 9 september, Nantes, France.
10. Nguyen JM, Gaultier A, Antonioli D. Modélisation des trous de données. RESP, Volume 65, Supplement 2, May 2017, Pages S99-S100
11. Nguyen JM, Gaultier A, Antonioli D. Abilities of Statistical Models to Identify Subjects with Ghost Prognosis Factors. J Health Edu Res Dev. 2015;3(141):2.
12. Castillo JM, Knol AC, Nguyen JM, Khammari A, Saint Jean M, Dreno B. Immuno-histochemical markers in advanced Basal-Cell Carcinoma: CD56 is associated with an absence of response to Vismodegib. Eur J Dermatol. 2016 Oct 1;26(5):452-459.
13. Vildy S, Nguyen JM, Gaultier A, Khammari A, Dreno B. Influence of delay time between lymph node recurrence and lymphadenectomy on survival of melanoma patients, Eur J Dermatol. 2017 Mar 2. doi:0.1684/ejd.2016.2955.
14. Nguyen JM, Gaultier A, Antonioli D. Trees with multvariate knots. BS annual conference Barcelona 2018 ; ISCB annual conference Melbourne 2018.
15. Nguyen JM, Gaultier A, Antonioli D. Forests of Perfect Trees. IBS annual conference Barcelona 2018 ; ISCB annual conference Melbourne 2018.
16. Nguyen JM, Antonioli D, ROP package, RCRAN


## La suite au Poster des Forêts d'Arbres Parfaits


***Jean-Michel NGUYEN***


***05 Juillet 2018***

CRCINA