Les arbres qui cachent les forêts?

le partitionnement récursif pour les SHS

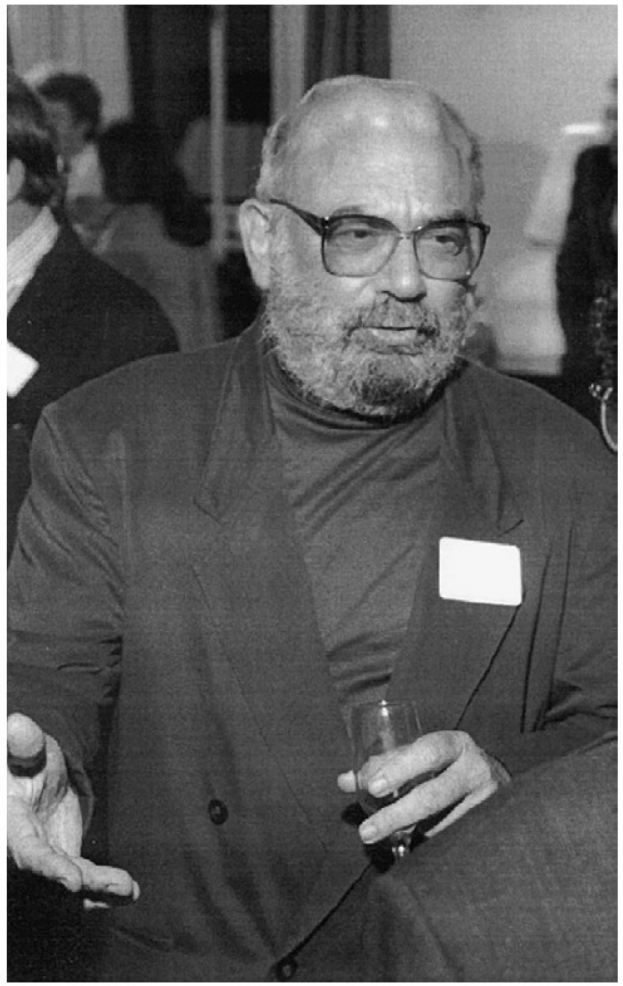
Séminaire R à l'usage des Sciences Sociales
EHESS, 23 mars 2017

Nicolas Robette
Laboratoire de Sociologie Quantitative
(CREST-ENSAE)

Machine learning (= *apprentissage automatique)*

- *Apprentissage supervisé*
    - « classification »
    - régression
- Clustering (*apprentissage non-supervisé)*
- réduction de dimensions
- …

Leo Breiman, 1928 - 2005



1954: PhD Berkeley (mathematics)

1960 -1967: UCLA (mathematics)

1969 -1982: Consultant

1982 - 1993 Berkeley (statistics)

1984 “Classification & Regression Trees”
(with Friedman, Olshen, Stone)

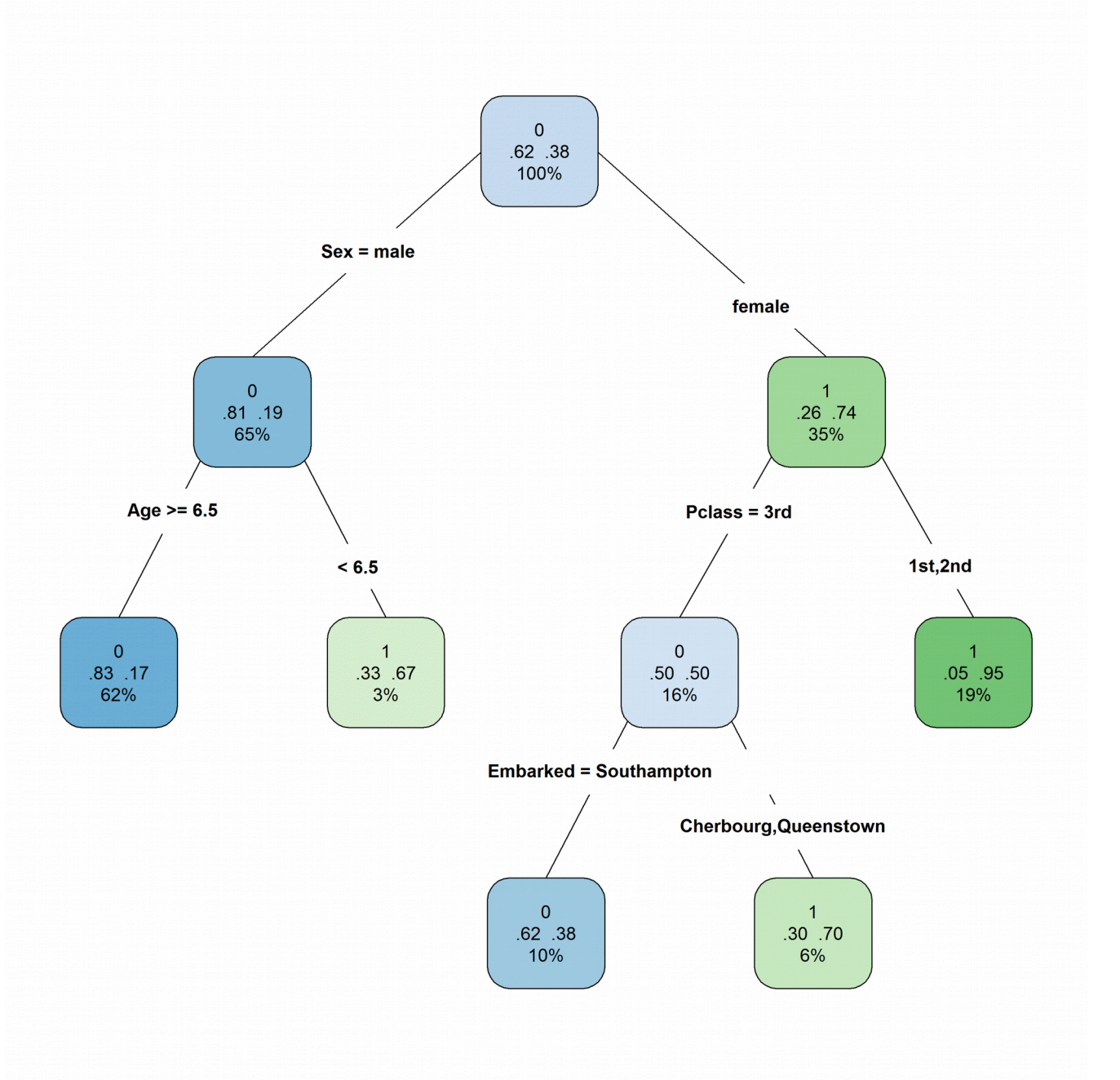
1996 “Bagging”

2001 “Random Forests”

Le cas Titanic

Base de données de passagers du **Titanic** (N=891), pour lesquels on dispose des données suivantes :

- **Survie** : oui / non (1/0), que l'on va chercher à expliquer à partir des caractéristiques individuelles
- **Sex** : female / male
- **Pclass** : classe du passager (1 / 2 / 3)
- **Age** : variable continue
- **Embarkment** : port d'embarquement (Cherbourg / Southampton / Queenstown)


0
.62 .38
100%
Sex = male
female
0
.81 .19
65%
1
.26 .74
35%
Age >= 6.5
< 6.5
Pclass = 3rd
1st,2nd
0
.83 .17
62%
1
.33 .67
3%
0
.50 .50
16%
1
.05 .95
19%
Embarked = Southampton
Cherbourg,Queenstown
0
.62 .38
10%
1
.30 .70
6%

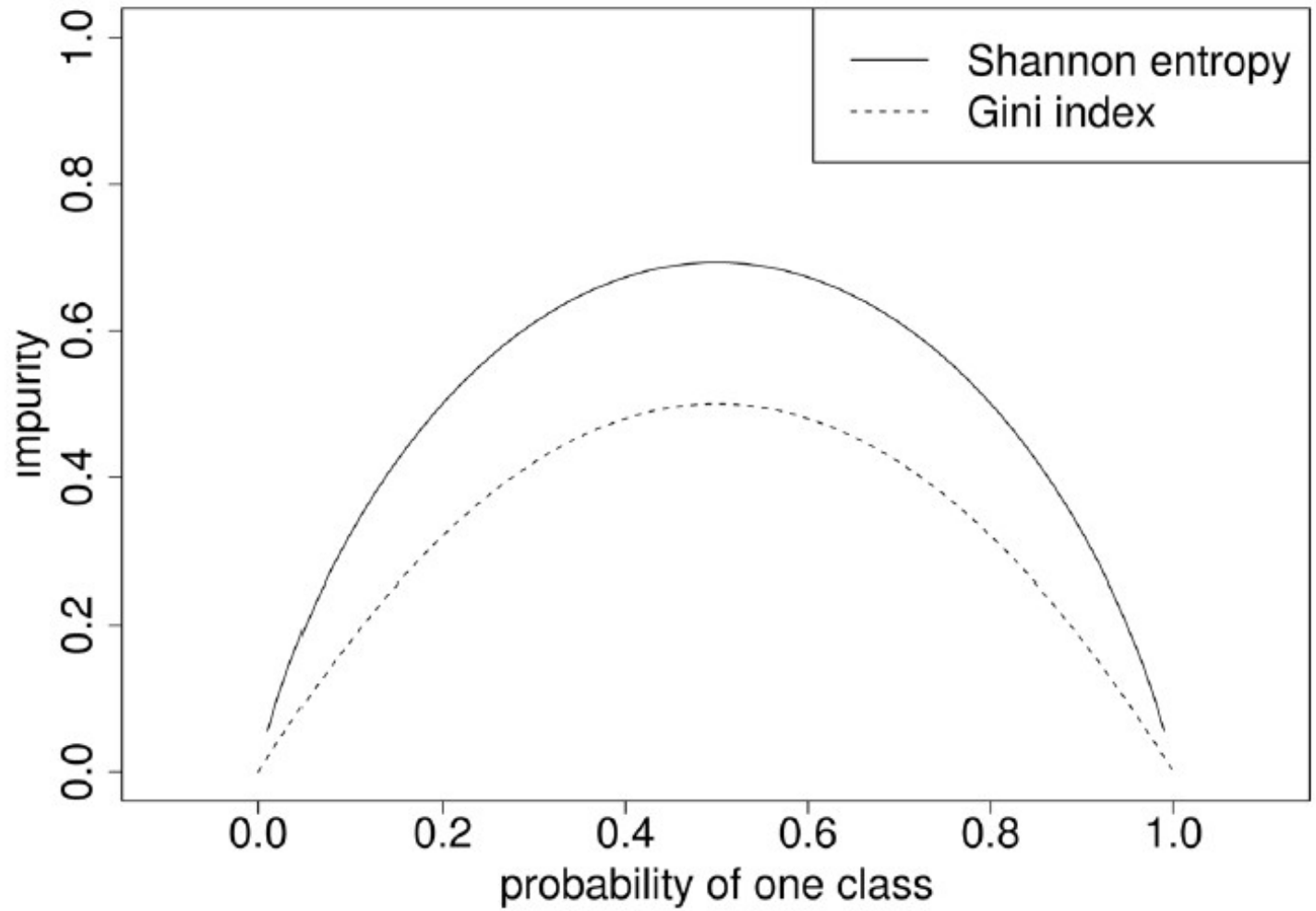
Classification & Regression Trees

- Au niveau du nœud initial (racine), on “découpe” / “segmente” / sépare (*split*) les individus en deux sous-groupes, qui forment des nœuds « filles » (*“daughter” nodes*). Puis chacun de ces sous-groupes est à son tour séparé, etc. L’objectif est de construire des sous-groupes les plus « homogènes » du point de vue de la variable à expliquer. Il s’agit donc d’un algorithme <u>récursif</u> de découpage / partition de l’espace des données en sous-régions homogènes en terme de classe.
- à résoudre :
  - sélection de variable et critère de segmentation (*splitting criteria*)
  - règle d’arrêt dans la construction de l’arbre
  - décision sur une feuille = post-élagage

Critère de segmentation

- Indices d'entropie (Gini, Shannon…), p-values…

→ Au final, on choisit la variable X telle qu'elle est la plus liée (corrélée) avec Y (ie réduction d'impureté maximale ou p-value la plus petite).


Shannon entropy
Gini index
impurity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
probability of one class

```
Node number 1: 891 observations,    complexity param=0.4444444
  predicted class=0  expected loss=0.3838384  P(node) =1
    class counts:   549   342
   probabilities: 0.616 0.384
  left son=2 (577 obs) right son=3 (314 obs)
  Primary splits:
      Sex      splits as  RL,       improve=124.426300, (0 missing)
      Pclass   splits as  RRL,      improve= 43.781830, (0 missing)
      Embarked splits as  RLL,      improve= 12.131190, (2 missing)
      Age      < 6.5  to the right, improve=  8.814172, (177 missing)
```

Splitting criteria

- **Regression**: residual sum of squares

$$RSS = \sum_{left} (y_i - y_L^*)^2 + \sum_{right} (y_i - y_R^*)^2$$

where y_L^* = mean y-value for left node

y_R^* = mean y-value for right node

- **Classification**: Gini criterion

$$Gini = N_L \sum_{k=1,\ldots,K} p_{kL} (1- p_{kL}) + N_R \sum_{k=1,\ldots,K} p_{kR} (1- p_{kR})$$

where p_{kL} = proportion of class k in left node

p_{kR} = proportion of class k in right node

Règles d’arrêt

a) Toutes les feuilles sont pures ; ou seuil de spécialisation

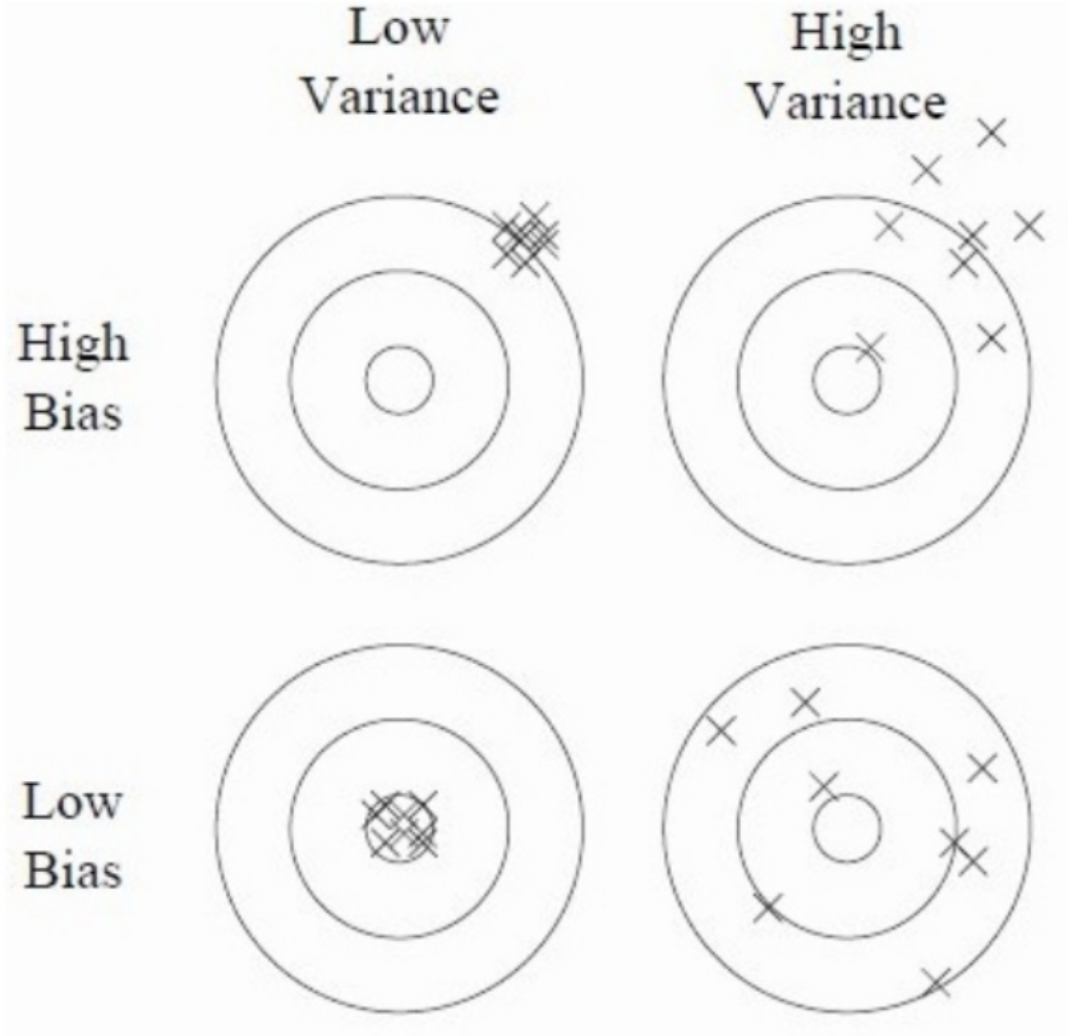
= critère de précision

b) On atteint un seuil minimal quant au nombre d’observations dans un nœud

= critère de support

c) On atteint un seuil quant au changement minimal dans la mesure d’impureté

Bias-variance decomposition

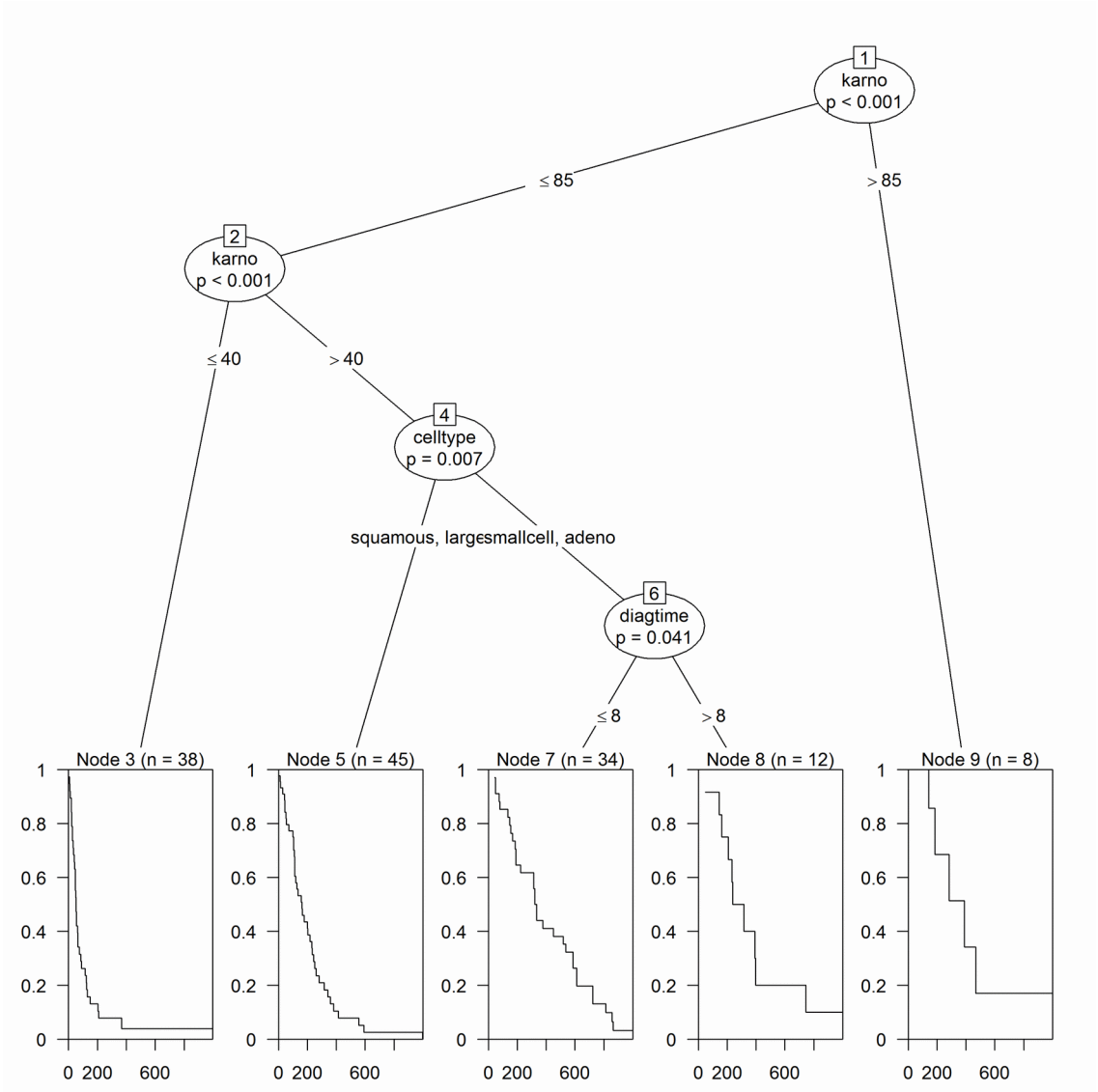


Arbres de survie: le cas « vétéran »

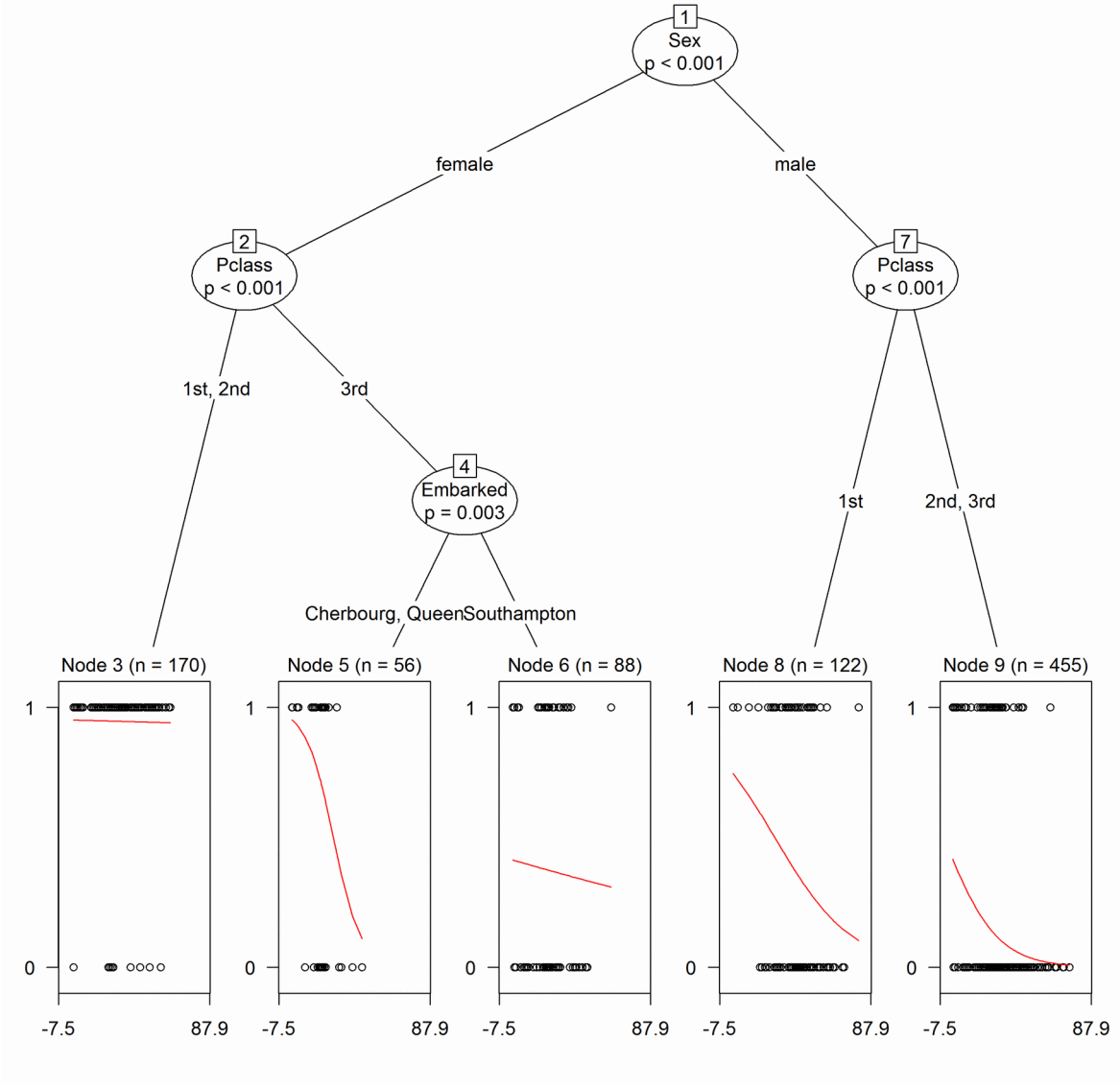
Randomised trial of two treatment regimens for lung cancer (standard survival analysis data set)

- trt: 1=standard 2=test
- celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
- time: survival time
- status: censoring status
- karno: Karnofsky performance score (100=good)
- diagtime: months from diagnosis to randomisation
- age: in years
- prior: prior therapy 0=no, 1=yes

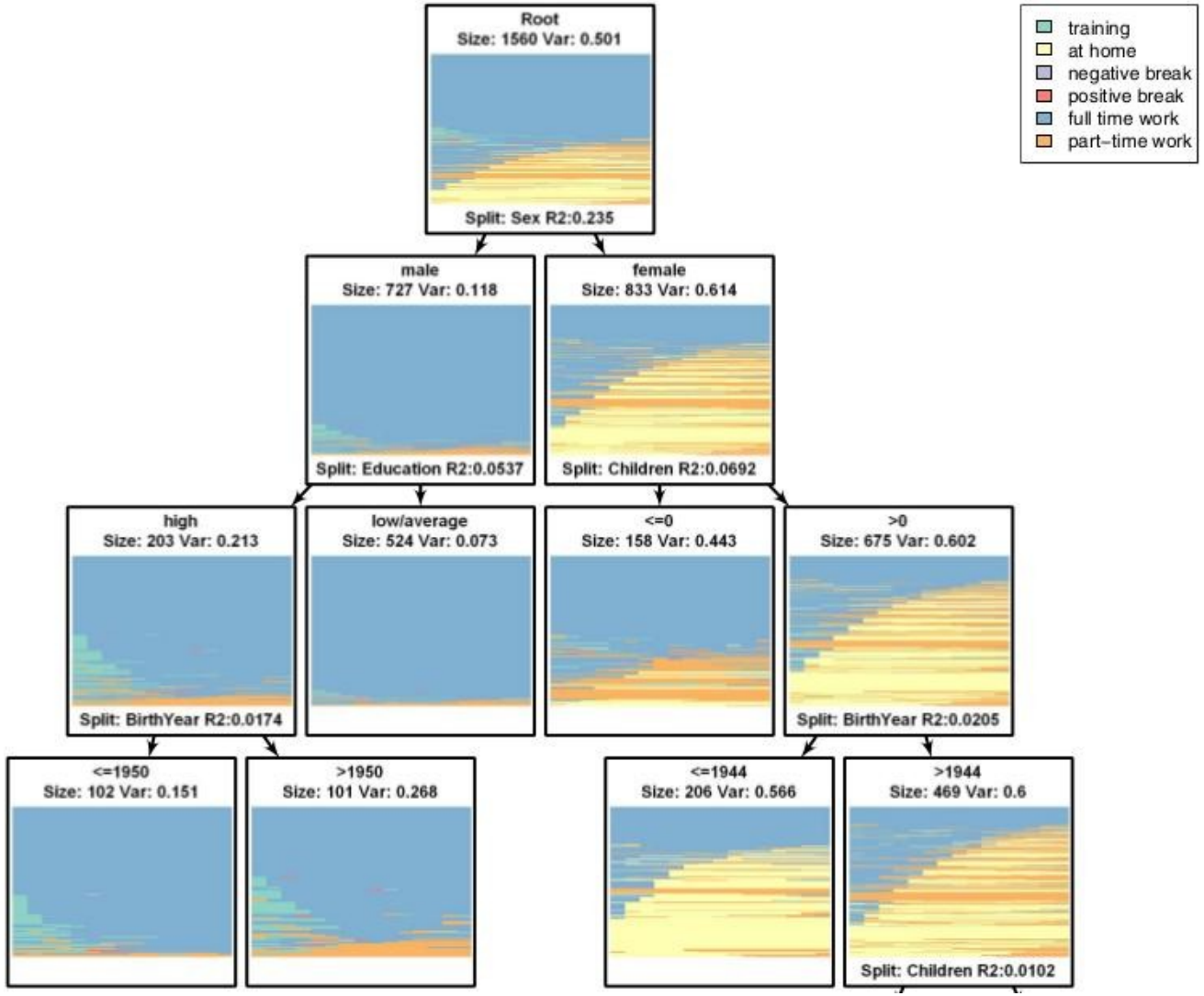
Arbres de survie: le cas « vétéran »



Model-based recursive partitioning



Matrices de distances et séquences



Avantages

• Applicables aux régressions et aux classifications, autrement dit la variable à expliquer peut être continue ou catégorielle. Possibilité également de traiter des données censurées (modèles de durée, Cox, etc.).

• Traitement indifférencié selon le type des variables explicatives, ie prennent en compte variables continues et catégorielles sans aucun souci.

• Pas d'hypothèses sur les distributions statistiques (normalité, etc.) = non-paramétriques.

• La sélection des variables est automatique.

• Robuste face aux données aberrantes ; solutions pour les données manquantes (cf *surrogate variables*), et non suppression comme dans les régressions

• Robuste face aux variables redondantes (cf multicolinéarité)

• Rapidité et capacité à traiter des très grandes bases

• Très faciles à analyser, lorsque l'arbre n'est pas trop grand.

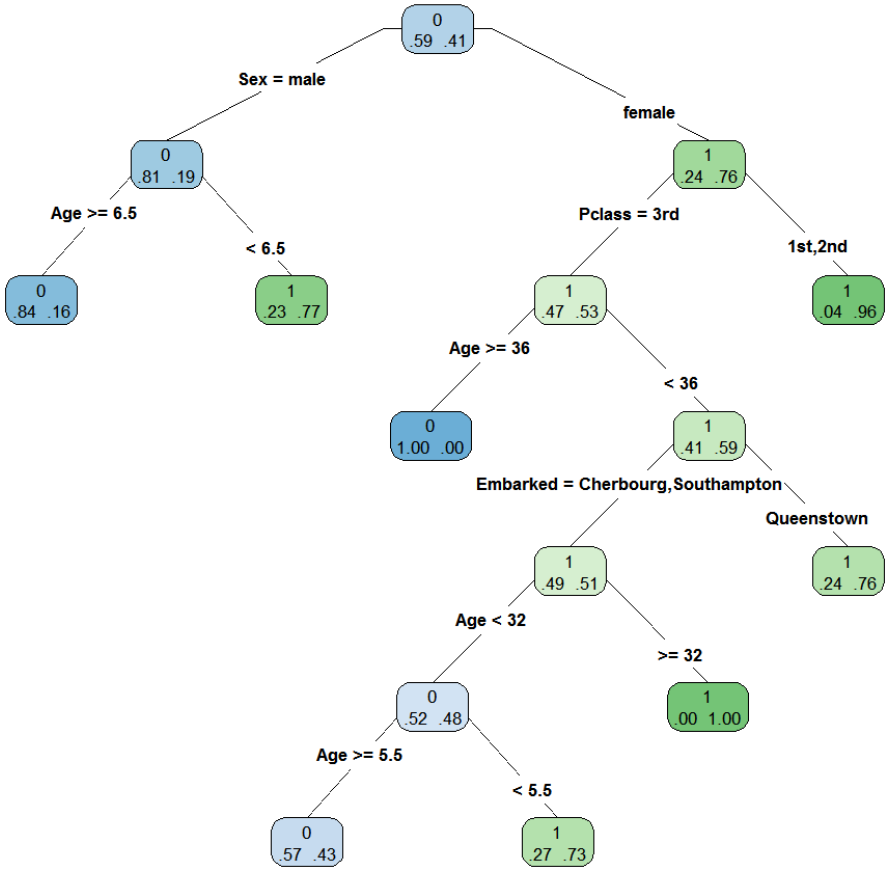
  - La représentation de l'arbre permet d'analyser quelles variables sont importantes, et où elles le sont (cf interactions).
  - Les nœuds finaux (*terminal nodes*) suggèrent une partition naturelle des observations en groupes homogènes.

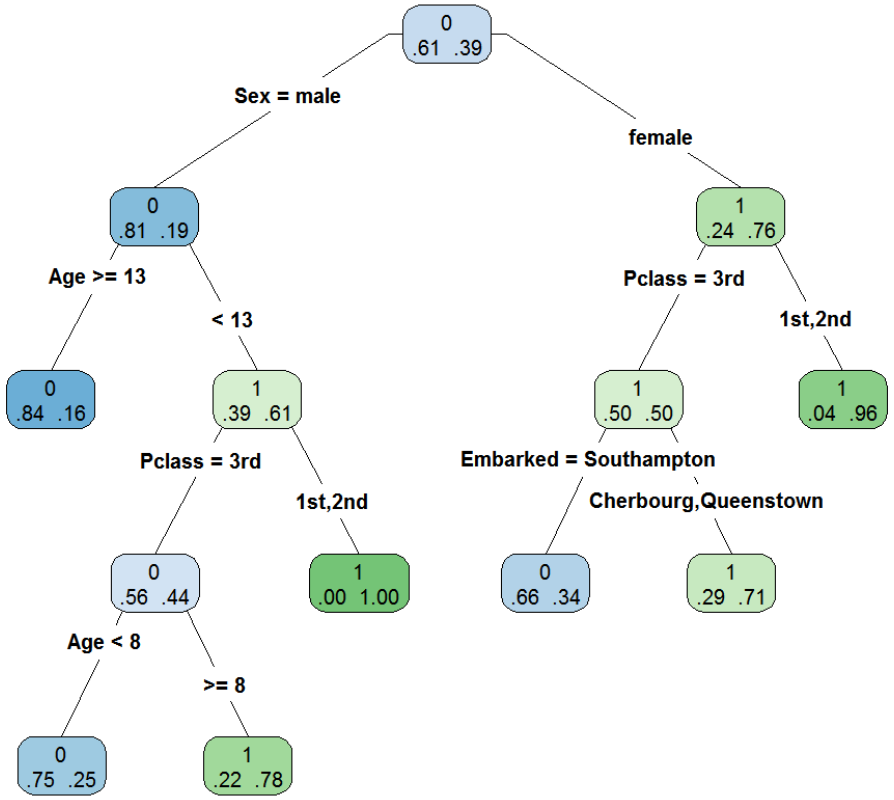
• Peut analyser des interactions non-linéaires (*highly non-linear interactions*), d'ordre élevé (*high order*) et les frontières de classification (*classification boundaries*).

Limites

- Précision
- Stabilité
- Binary splits

La stabilité des arbres





La stabilité des arbres

```
Node number 2: 577 observations,    complexity param=0.02339181
  predicted class=0  expected loss=0.1889081  P(node) =0.647587
    class counts:   468   109
   probabilities: 0.811 0.189
  left son=4 (553 obs) right son=5 (24 obs)
  Primary splits:
      Age      < 6.5  to the right, improve=10.788930, (124 missing)
      Pclass   splits as  RLL,      improve=10.019140, (0 missing)
      Embarked splits as  RLL,      improve= 3.079304, (0 missing)
```

Bagging = Boostrap AGGregatING

i. On construit un échantillon "*bootstrap*" à partir des données = tirage au sort de n observations, avec remise.

ii. On construit un arbre à partir de cet échantillon.

iii. On répète l'opération un grand nombre de fois, souvent plusieurs centaines : on obtient donc un ensemble d'arbres.

iv. On combine / agrège enfin ces arbres, par le vote (pour la classification) ou la moyenne (pour la régression).

Erreur « out-of-bag » (OOB)

i	M_1	M_2	M_3	M_4	M_5	Y^	Y	ERR
1	.	+	.	+	−	+	+	0
2	.	.	.	−	−	−	−	0
3	+	.	−	−	−	−	+	1
4	.	+	+	.	.	+	+	0
5	+	−	−	.	.	−	+	1
6	.	.	.	.	+	+	−	1
7	−	.	.	−	.	−	−	0
8	−	+	.	.	+	+	+	0
9	.	.	+	.	.	+	+	0
10	+	.	.	+	.	+	−	1
…								

Prédiction OOB par vote à la majorité des modèles (dans la ligne).

ERR_{OOB} = Proportion(ERR)

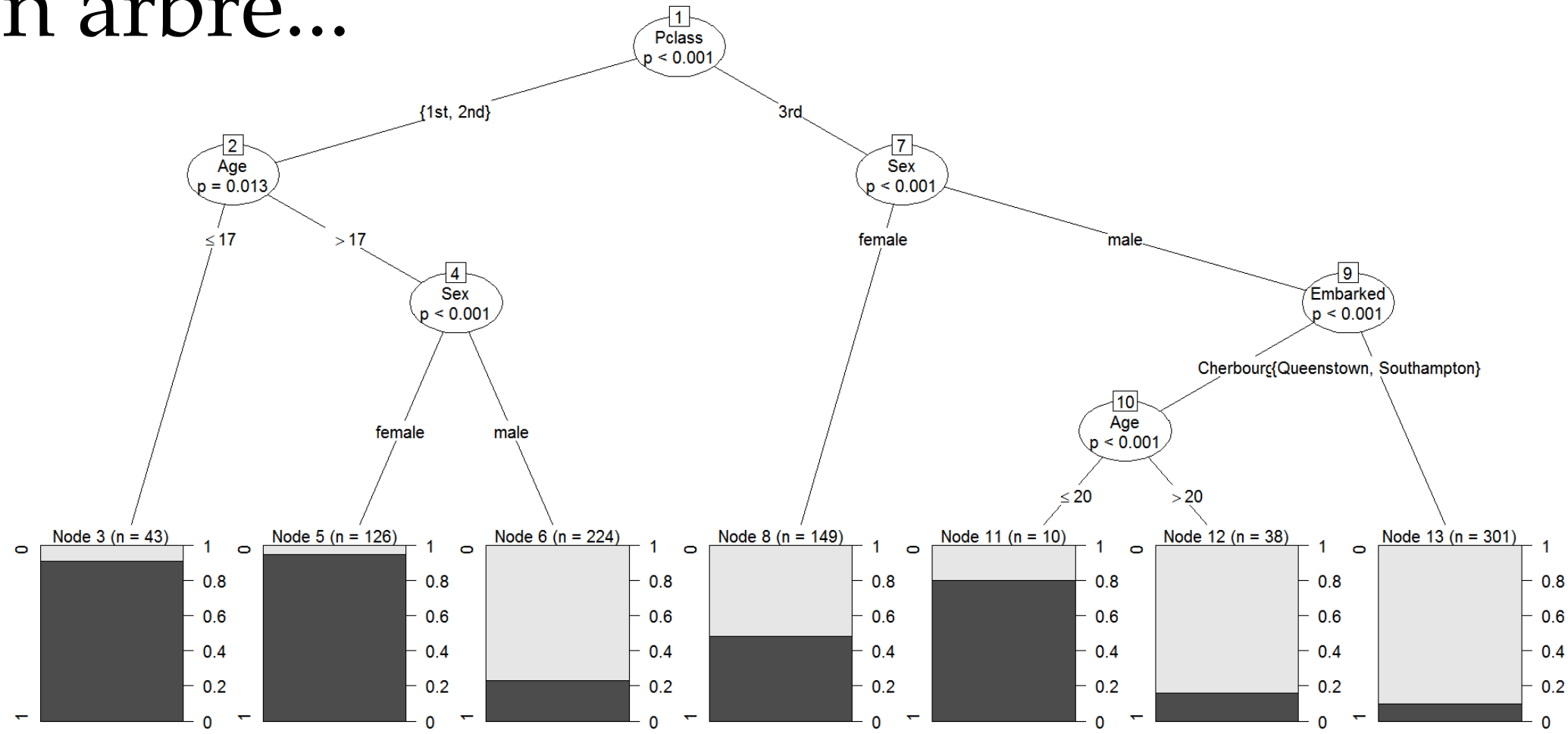
Err_{OOB} est une estimation viable de l'erreur en prédiction du modèle bagging.

« . » signifie que l'individu a été utilisé pour la construction de l'arbre M_b
« + » ou « − » est la prédiction de l'arbre M_b pour l'individu « i » qui est OOB

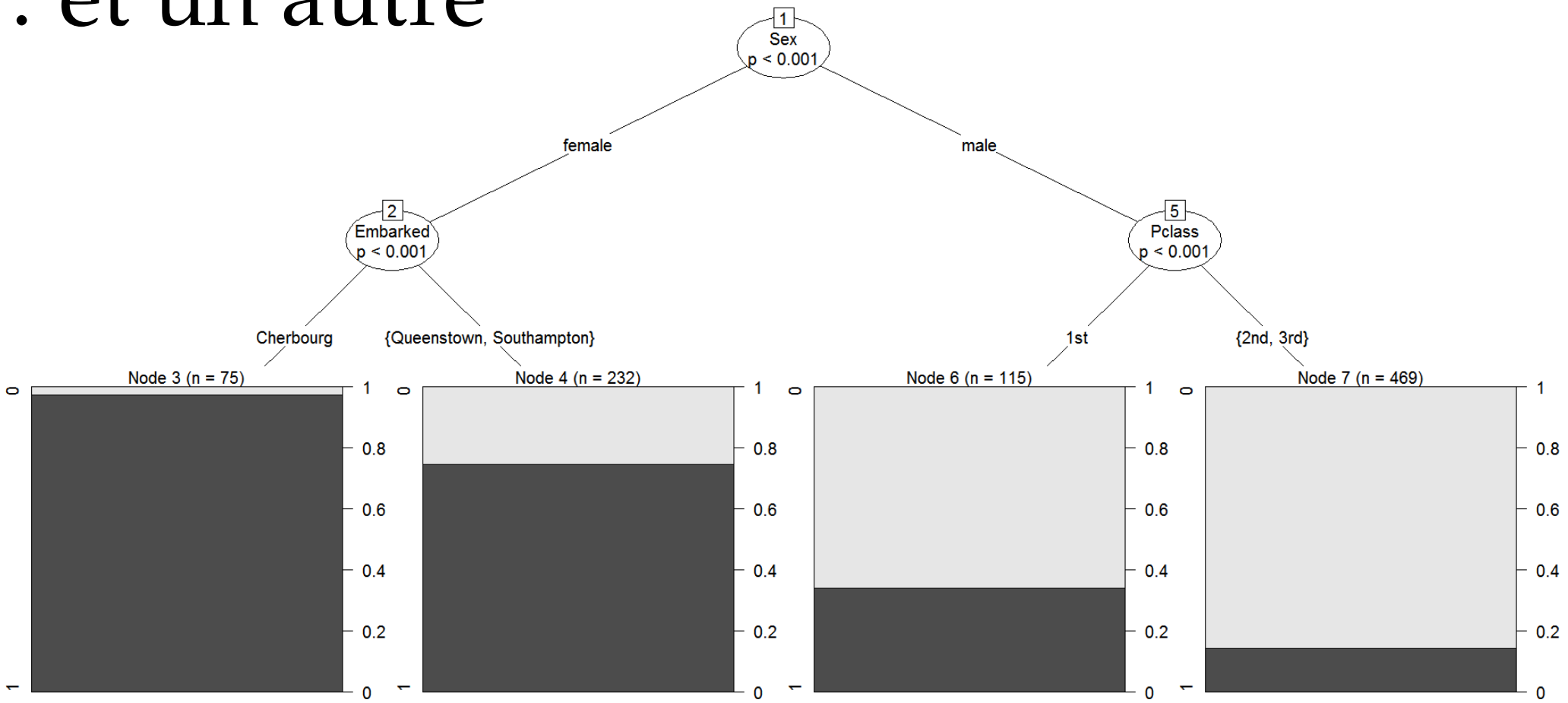
Random Forest

- Faire pousser un arbre à partir d'un échantillon *bootstrap* des données de départ (ie d'apprentissage).
- A chaque noeud:
  1. Sélectionner mtry variables au hasard parmi les M variables possibles (tirage au sort indépendant à chaque nœud).
  2. Trouver la meilleure segmentation (split) à partir de ces mtry variables.
- Faire pousser l'arbre à sa profondeur maximale (classification). Pas d'élagage.
- Reproduire ces étapes un grand nombre de fois (500 par défaut dans R, par exemple)
- Combiner les arbres par vote/moyenne pour obtenir les valeurs prédites de chaque observation.

Un arbre...


1
Pclass
p < 0.001
{1st, 2nd}
3rd
2
Age
p = 0.013
7
Sex
p < 0.001
≤ 17
> 17
female
male
4
Sex
p < 0.001
9
Embarked
p < 0.001
female
male
Cherbourg
{Queenstown, Southampton}
10
Age
p < 0.001
≤ 20
> 20
Node 3 (n = 43)
Node 5 (n = 126)
Node 6 (n = 224)
Node 8 (n = 149)
Node 11 (n = 10)
Node 12 (n = 38)
Node 13 (n = 301)
0
0.2
0.4
0.6
0.8
1

… et un autre


1
Sex
p < 0.001
female
male
2
Embarked
p < 0.001
5
Pclass
p < 0.001
Cherbourg
{Queenstown, Southampton}
1st
{2nd, 3rd}
Node 3 (n = 75)
Node 4 (n = 232)
Node 6 (n = 115)
Node 7 (n = 469)
0
1
0.8
0.6
0.4
0.2

Qualité du modèle

Matrice de confusion

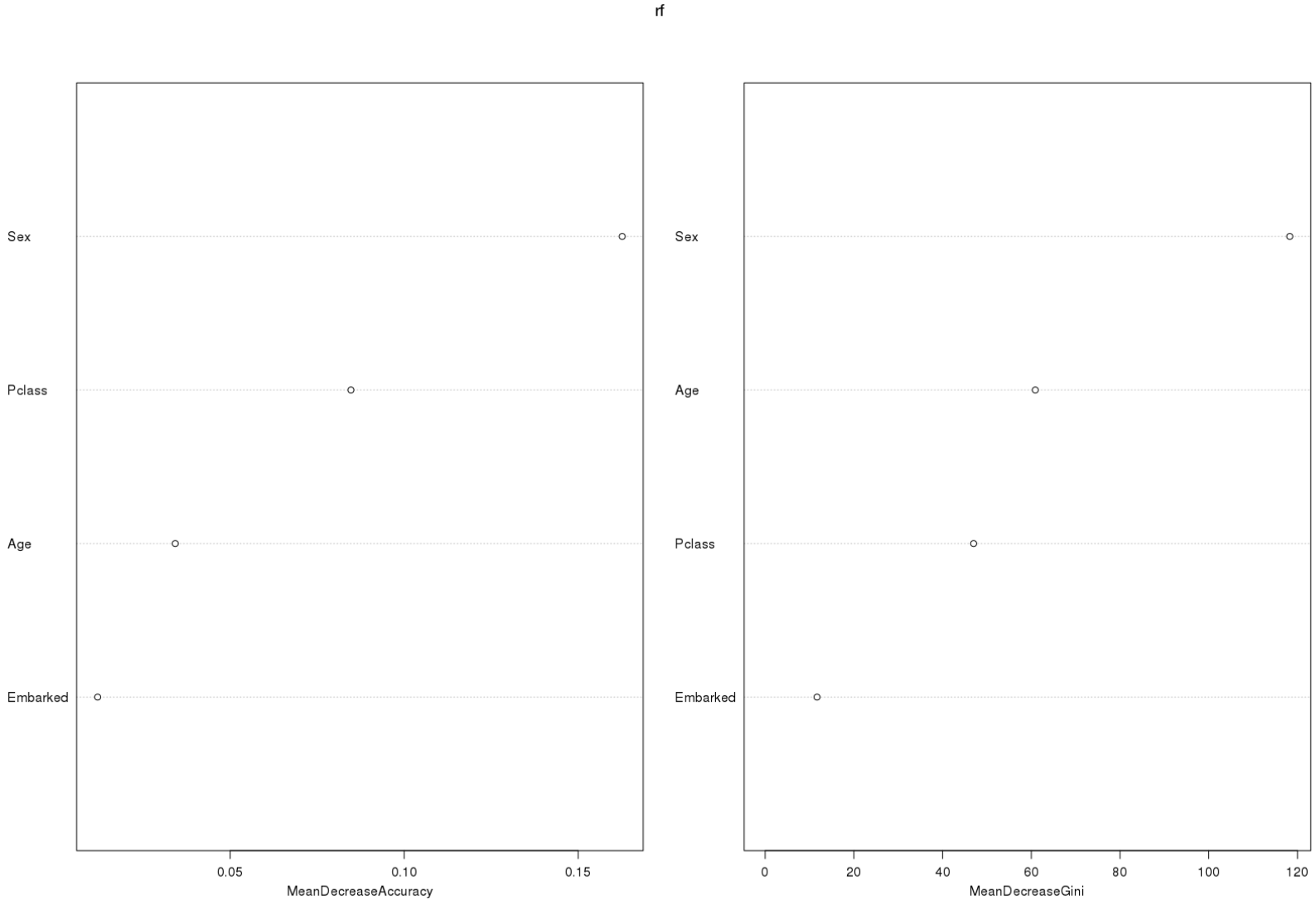
		actual		
		0	1	class.error
predict	0	524	25	0.04553734
	1	132	210	0.38596491

Taux d'erreur OOB = 0.1762065

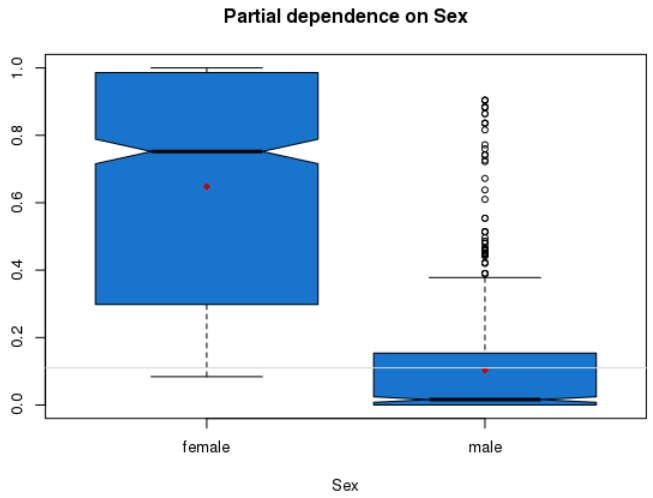
Importance des variables

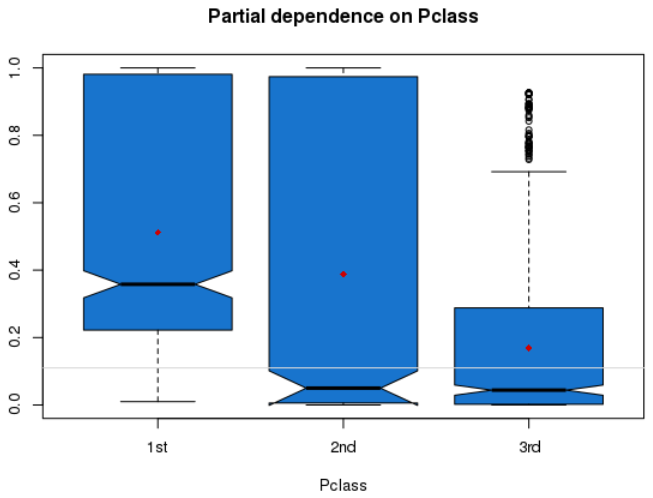
a) Nombre d’utilisations des variables
b) Importance de Gini
c) Importance par permutation

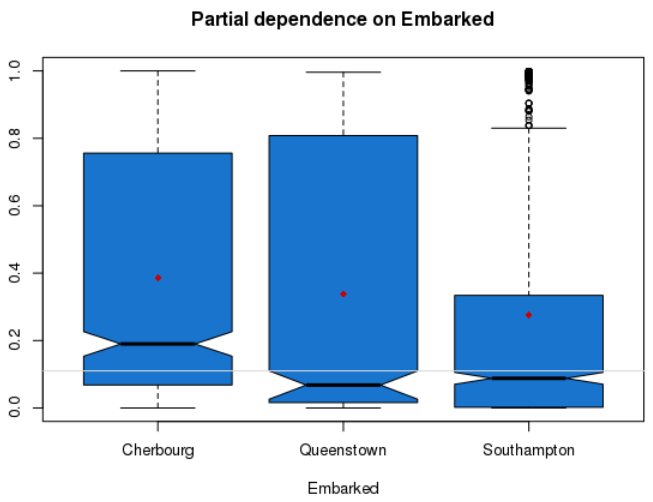
Importance des variables



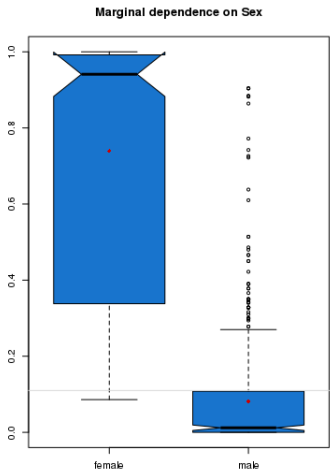
Dépendances partielles

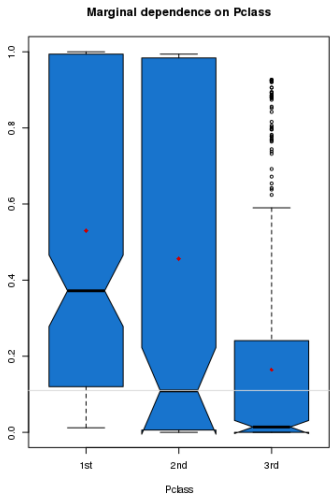


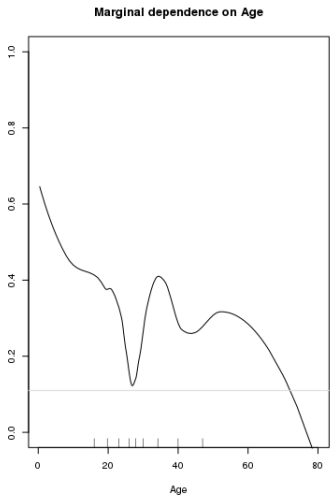


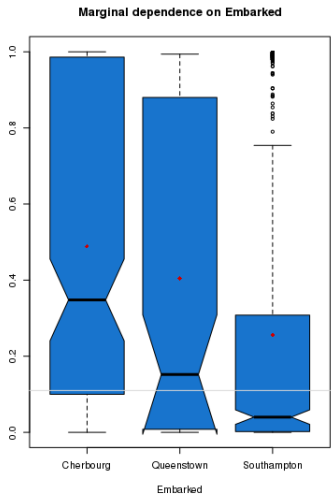


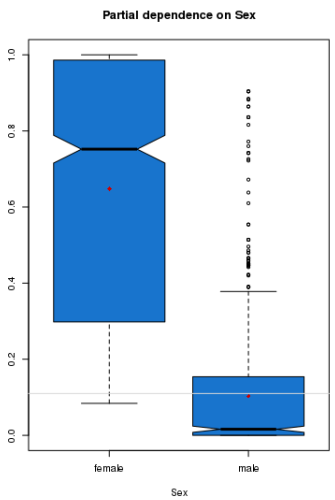
Dépendances partielles *vs* marginales

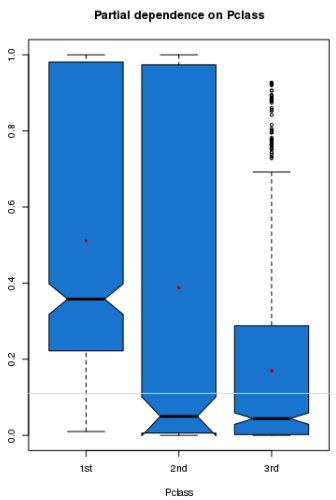


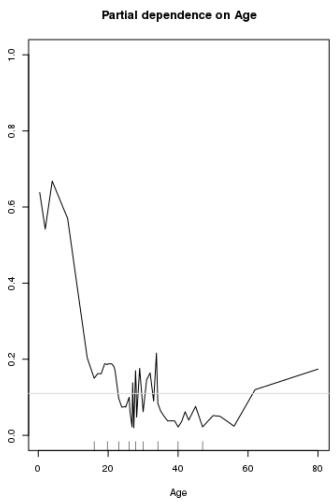


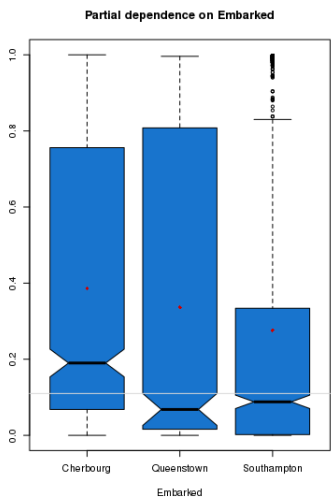








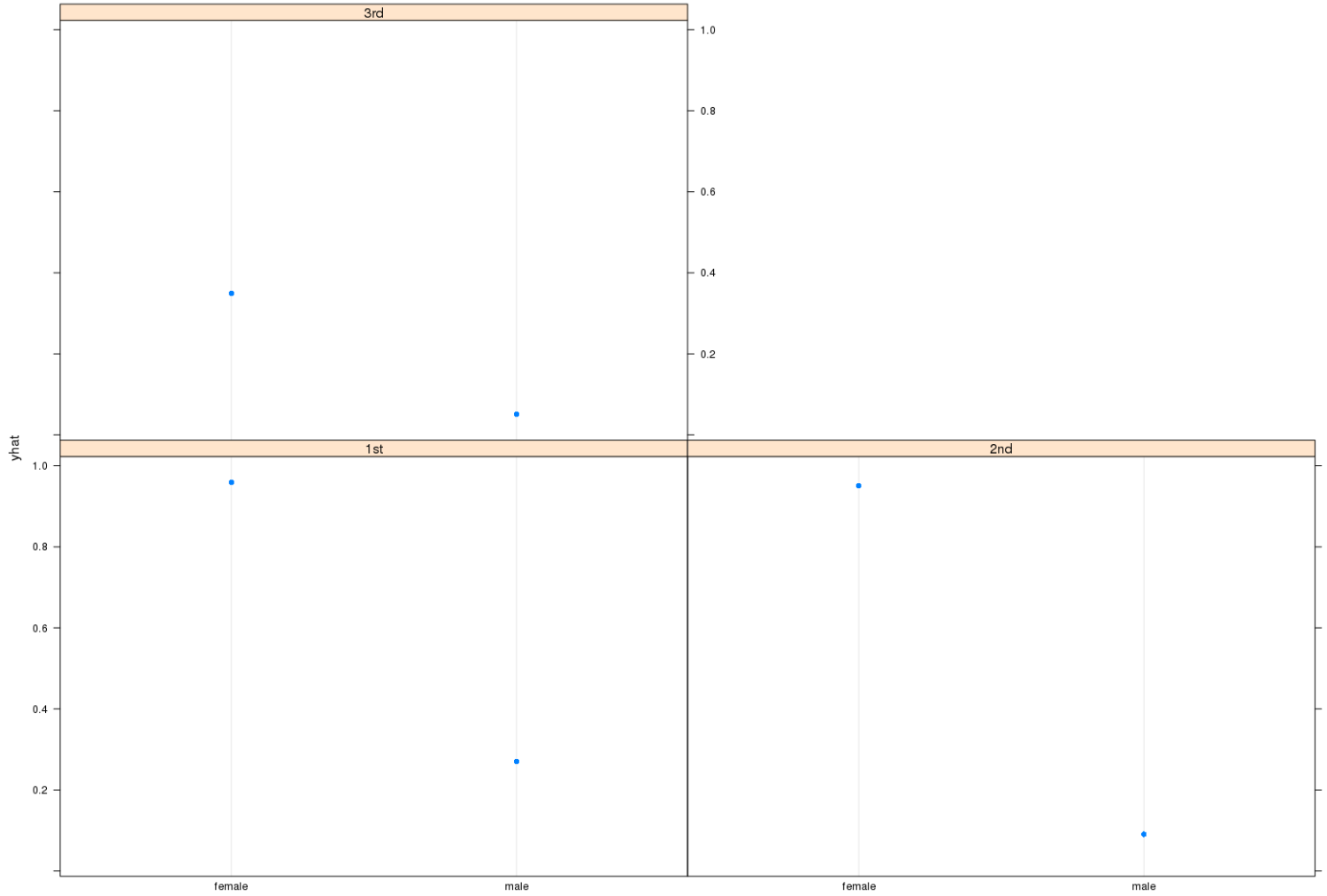




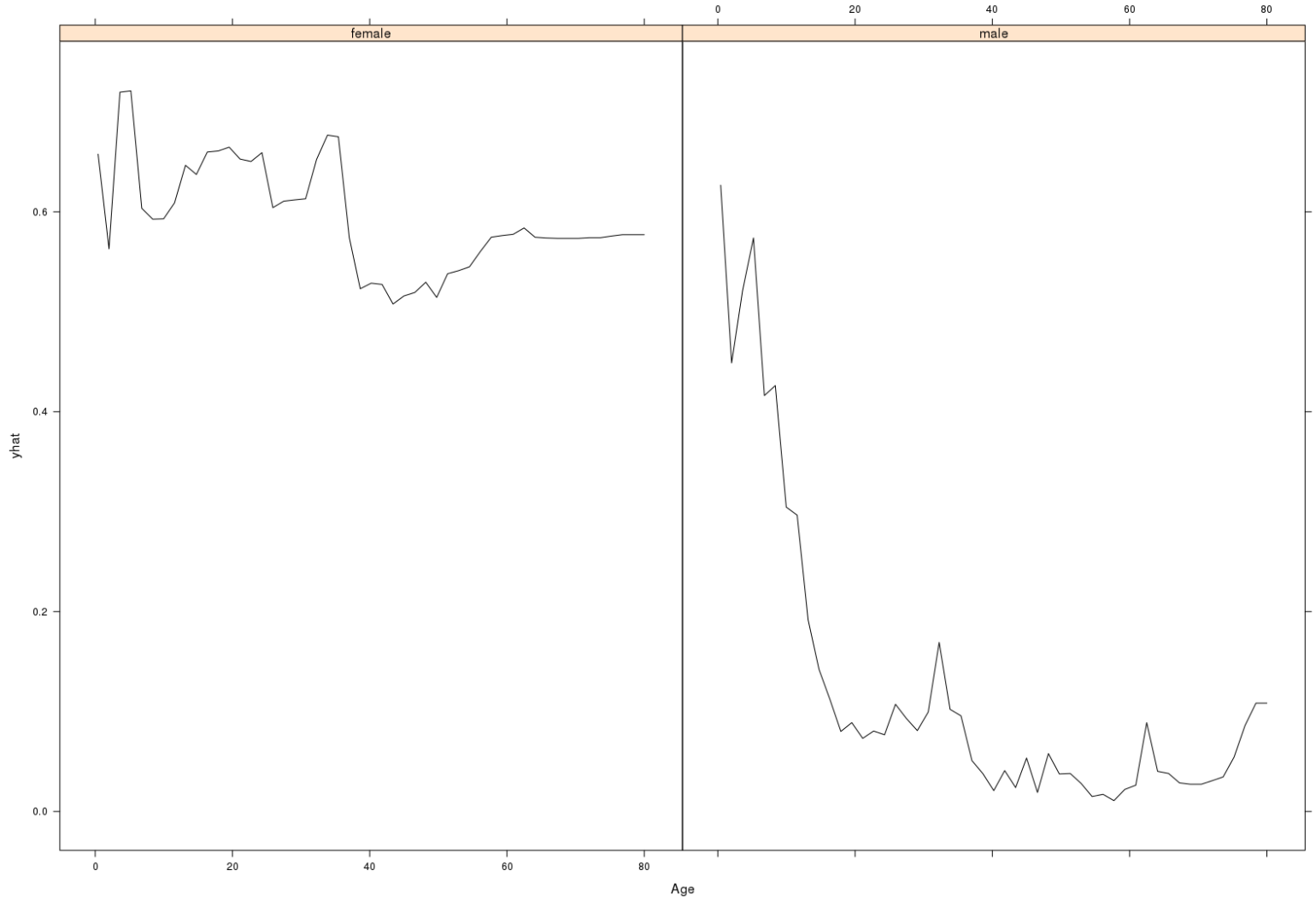
Interactions

	Var 1	Var 2	Paired	Additive	Difference
Pclass:Sex	0.0871	0.1505	0.2155	0.2376	-0.0221
Pclass:Age	0.0871	0.0440	0.1112	0.1311	-0.0199
Pclass:Embarked	0.0871	0.0117	0.0945	0.0988	-0.0043
Sex:Age	0.1511	0.0440	0.1783	0.1951	-0.0168
Sex:Embarked	0.1511	0.0117	0.1603	0.1628	-0.0025
Age:Embarked	0.0441	0.0117	0.0480	0.0558	-0.0079

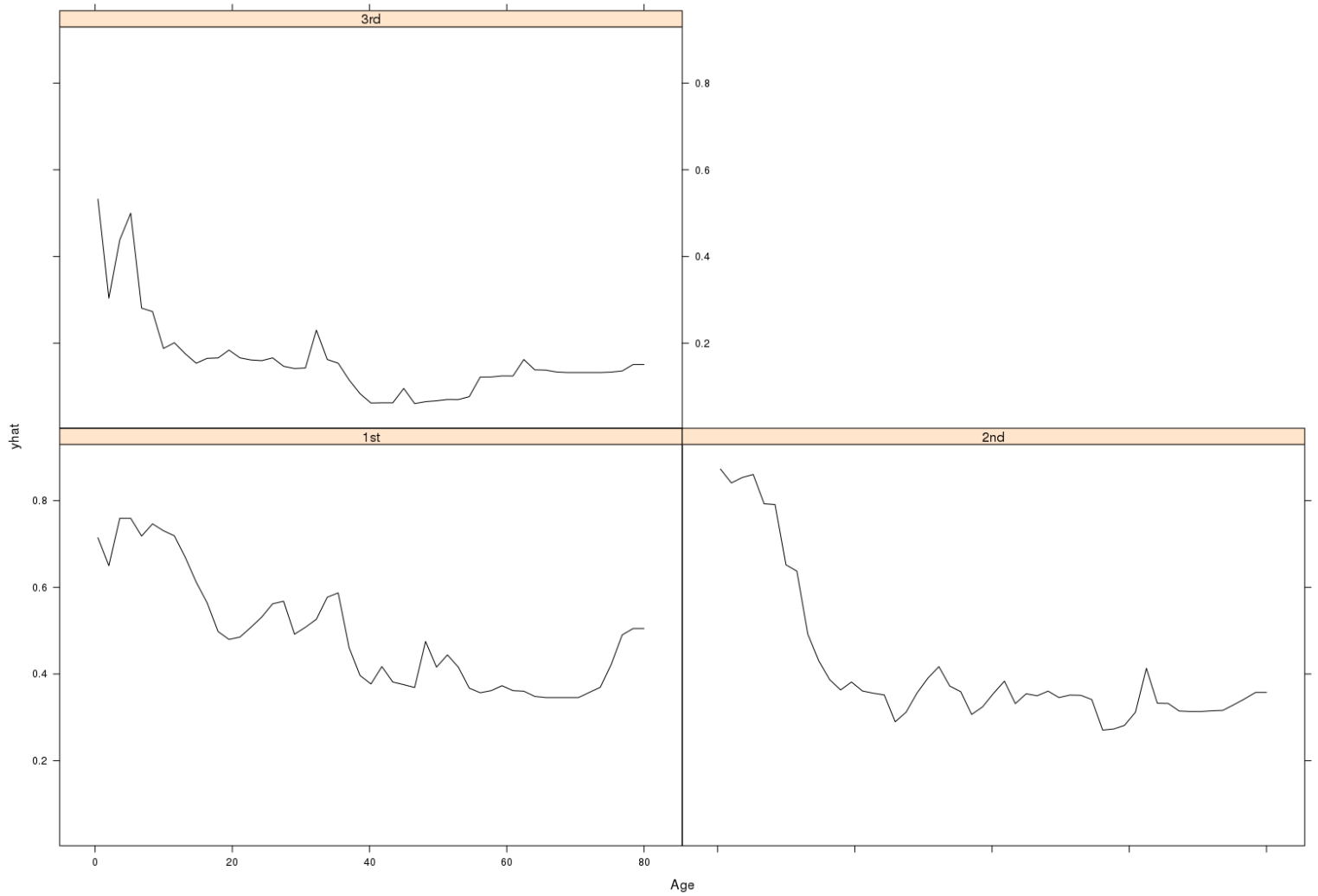
Interaction sexe*classe



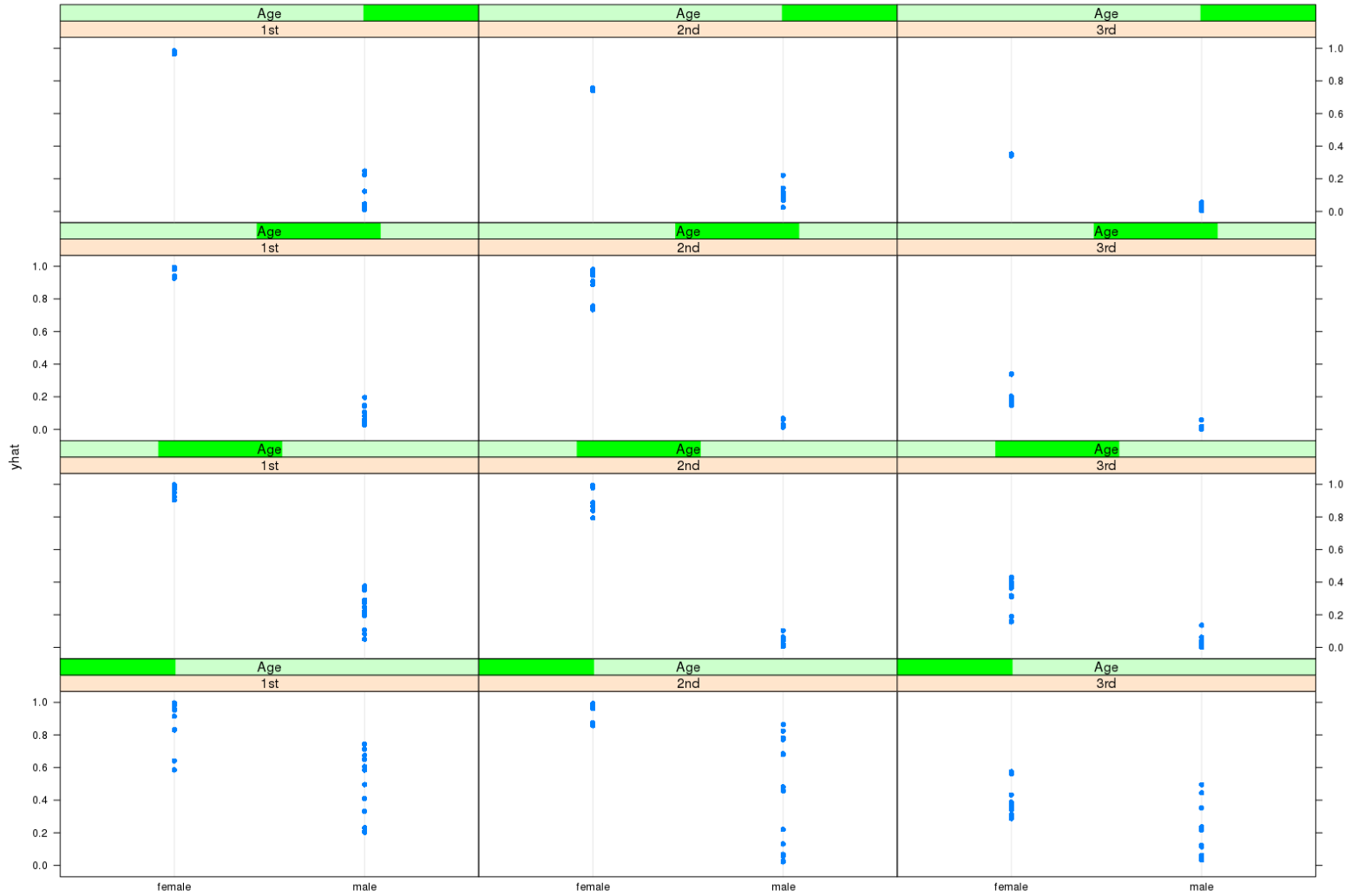
Interaction sexe*âge



Interaction classe*âge



Interaction sexe*classe*âge



Proximités

- Définition simple = la proximité entre deux observations est le nombre de fois (ie d'arbres) dans lesquels elles se trouvent dans le même nœud terminal. Prennent donc en compte l'importance des variables !
- Des proximités aux distances, MDS, clustering

Prototypes

```
$`0`
     [,1]  [,2]   [,3]                [,4]
[1,] "3rd" "male" "13.5"              "Queenstown"
[2,] "3rd" "male" "30"                "Southampton"
[3,] "3rd" "male" "29"                "Southampton"
[4,] "3rd" "male" "29"                "Southampton"
[5,] "3rd" "male" "24.1143000126459" "Southampton"

$`1`
     [,1]  [,2]     [,3]                 [,4]
[1,] "1st" "female" "52.5"               "Southampton"
[2,] "2nd" "female" "30"                 "Queenstown"
[3,] "1st" "female" "21.5"               "Southampton"
[4,] "3rd" "female" "19.9928248101107" "Southampton"
[5,] "2nd" "female" "20.96058058365"    "Southampton"
```

Arbres représentatifs

Mesures de similarité entre les arbres:

- Arbres similaires s'ils utilisent les mêmes variables pour les splits
- Arbres similaires si les mêmes individus sont ensembles / séparés dans les nœuds terminaux
- Arbres similaires si les prédictions sont les mêmes

Imputation de valeurs manquantes

- Méthode rapide = via médiane / mode
- Méthode raffinée =
    1. Imputation par la manière rapide.
    2. Calcul des proximités.
    3. Imputation des valeurs manquantes, pour l'observation I, par la moyenne pondérée des valeurs non-manquantes, avec des poids proportionnels aux proximités entre l'observation i et les observations aux valeurs non-manquantes.
    4. Répétition des étapes 2 et 3 à plusieurs reprises (4 à 6 suffisent)

Avantages

- Hérite de beaucoup des avantages de CART
- Avec en plus:
  - Qualité de prédiction
  - Stabilité
  - Évaluation de l'erreur intégrée (OOB)
  - « Frontières » plus douces

Limites

- Interprétation moins facile: pas de représentation graphique, ni d'arbre « moyen »
- Overfitting ?
- Aléa et stabilité ?

Usages

- Explorer
- Expliquer
- Prédire
- autres: imputation de valeurs manquantes, analyse de survie…

Boosting

Pondération des individus. A l'étape (b+1), l'idée est de donner une pondération plus élevée aux individus mal classés par Mb. La construction des modèles est séquentielle.

➔Alors qu'on parle de « stratégies aléatoires » pour les ensembles construits par *bagging*, on parle de « stratégies adaptatives » pour les ensembles construits par *boosting*.

Références

• Leo Breiman, Jerome Friedman, Richard Olshen, Charles Stone (1984). *Classification and Regression Trees*. Wadsworth.

• Leo Breiman (1996). “Bagging Predictors”. *Machine Learning*, 24, 123-140.

• Leo Breiman (2001). “Random Forests”. *Machine Learning*, 45, 5-32.

• Site web de Breiman: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

• Présentation de Adele Cutler : http://www.math.usu.edu/adele/RandomForests/UofU2013.pdf

• Carolin Strobl, James Malley, Gerhard Tutz (2009). “An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests”. *Psychological Methods*, 14(4), 323-348. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2927982/

• Trevor Hastie, Rob Tibshirani, Jerome Friedman (2009). *Statistical Learning*. Springer.

Packages R

- **`rpart`** : arbres CART
- **`rpart.plot`** : représentations plus jolies des arbres
- **`randomForest`** : RF « à la Breiman »
- **`randomForestSRC`** : RF « à la Breiman » élargies aux modèles de durée + qq outils utiles (interactions, partial dependence plots...)
- **`party`** : RF avec « inférence conditionnelle »
- **`adabag`** : bagging, boosting