Sommaire

Remerciements i
Avant-propos iii

1 Introduction 1
  1.1 Préambule ................................................. 1
  1.2 Notations .................................................. 2
  1.3 Objectifs statistiques ...................................... 3
  1.4 Packages .................................................. 4
  1.5 Jeux de données ............................................ 5
    1.5.1 Exemple fil rouge : détection de spams ............... 5
    1.5.2 Pollution par l’ozone ................................ 6
    1.5.3 Données génomiques pour une étude vaccinale ......... 7
    1.5.4 Pollution par les poussières .......................... 7

2 Arbres CART 9
  2.1 Le principe ................................................ 9
  2.2 Construction de l’arbre maximal ......................... 10
  2.3 Élagage .................................................... 13
  2.4 Le package rpart ......................................... 16
  2.5 Découpes concurrentes et de substitution ............ 24
    2.5.1 Découpes compétitives ou concurrentes .......... 24
    2.5.2 Découpes de substitution ........................... 24
    2.5.3 Interprétabilité ..................................... 26
  2.6 Exemples ................................................. 28
    2.6.1 Prédire la concentration d’ozone ................. 28
    2.6.2 Analyser des données génomiques ................. 31

3 Forêts aléatoires 35
  3.1 Principes général ......................................... 35
    3.1.1 Instabilité d’un arbre ............................... 36
    3.1.2 D’un arbre à un ensemble : le Bagging .......... 39
3.2 Forêts aléatoires « Random Inputs » .......................... 41
3.3 Le package randomForest ........................................ 43
3.4 Erreur Out-Of-Bag ............................................. 45
3.5 Réglage des paramètres pour la prédiction ..................... 46
   3.5.1 Le nombre d’arbres ntree .................................. 46
   3.5.2 Le nombre de variables choisies à chaque nœud mtry ...... 47
3.6 Exemples .......................................................... 50
   3.6.1 Prédire la concentration d’ozone ............................ 50
   3.6.2 Analyser des données génomiques ........................... 53
   3.6.3 Analyser la pollution par les poussières .................. 56

4 Importance des variables ........................................... 61
4.1 Notions d’importance ........................................... 61
4.2 Comportement de l’importance ................................ 65
   4.2.1 Comportement vis-à-vis de n et p .......................... 66
   4.2.2 Comportement vis-à-vis de groupes de variables corrélées . 67
4.3 Diversité des arbres et importance .............................. 68
4.4 Influence des paramètres et importance ........................ 70
4.5 Exemples .......................................................... 72
   4.5.1 Une illustration par simulation en régression ............... 72
   4.5.2 Prédire la concentration d’ozone ............................ 74
   4.5.3 Analyser des données génomiques ........................... 77
   4.5.4 Pollution de l’air par les poussières : quelle part locale? . 79

5 Sélection de variables ............................................... 83
5.1 Généralités ....................................................... 83
5.2 Le principe ....................................................... 85
5.3 La procédure ..................................................... 86
5.4 Le package VSURF .............................................. 87
5.5 Réglage des paramètres pour la sélection ...................... 94
5.6 Exemples .......................................................... 96
   5.6.1 Prédire la concentration d’ozone ............................ 96
   5.6.2 Analyser des données génomiques ........................... 97

Bibliographie .......................................................... 101

Index des fonctions ................................................... 107

Index ................................................................. 108