Atouts et faiblesses de R

P.A. Cornillon & E. Matzner-Løber

Journée PLUME Paris 2010

Exposés (auxquels vous avez échappé)

Définir l'objectif didactique

En fonction du public, l'objectif est différent.

Exposés (auxquels vous avez échappé)

Définir l'objectif didactique

En fonction du public, l'objectif est différent.

- statisticien 1 : compiler et déboguer un programme C utilisant des objets R

▸ debug

Exposés (auxquels vous avez échappé)

Définir l'objectif didactique

En fonction du public, l'objectif est différent.

- statisticien 1 : compiler et déboguer un programme C utilisant des objets R ▸ debug
- statisticien 2 : créer une interface graphique avec R ▸ interface

Exposés (auxquels vous avez échappé)

Définir l'objectif didactique

En fonction du public, l'objectif est différent.

- statisticien 1 : compiler et déboguer un programme C utilisant des objets R ▸ debug
- statisticien 2 : créer une interface graphique avec R ▸ interface
- statisticien 3 : comprendre l'aide du package `lattice`

Exposés (auxquels vous avez échappé)

Définir l'objectif didactique

En fonction du public, l'objectif est différent.

- statisticien 1 : compiler et déboguer un programme C utilisant des objets R ▸ debug
- statisticien 2 : créer une interface graphique avec R ▸ interface
- statisticien 3 : comprendre l'aide du package `lattice`
- statisticien débutant : débuter en R (et repartir avec les bases après 45 mn d'exposé)

Exposés (auxquels vous avez échappé)

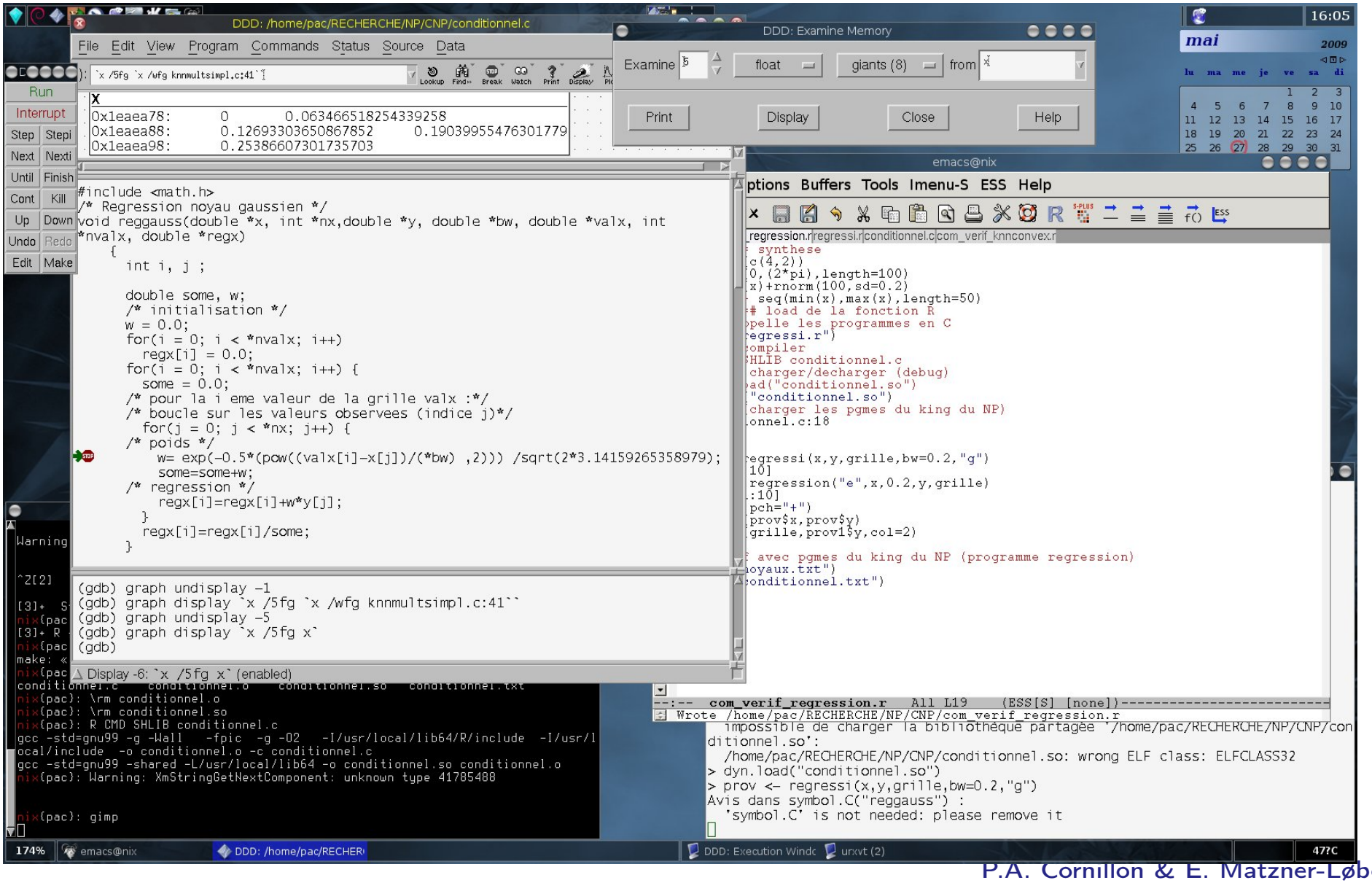
Définir l'objectif didactique

En fonction du public, l'objectif est différent.

- statisticien 1 : compiler et déboguer un programme C utilisant des objets R ▸ debug
- statisticien 2 : créer une interface graphique avec R ▸ interface
- statisticien 3 : comprendre l'aide du package `lattice`
- statisticien débutant : débuter en R (et repartir avec les bases après 45 mn d'exposé)
- public mixte : exposer quelques potentialités du logiciel ▸ suite

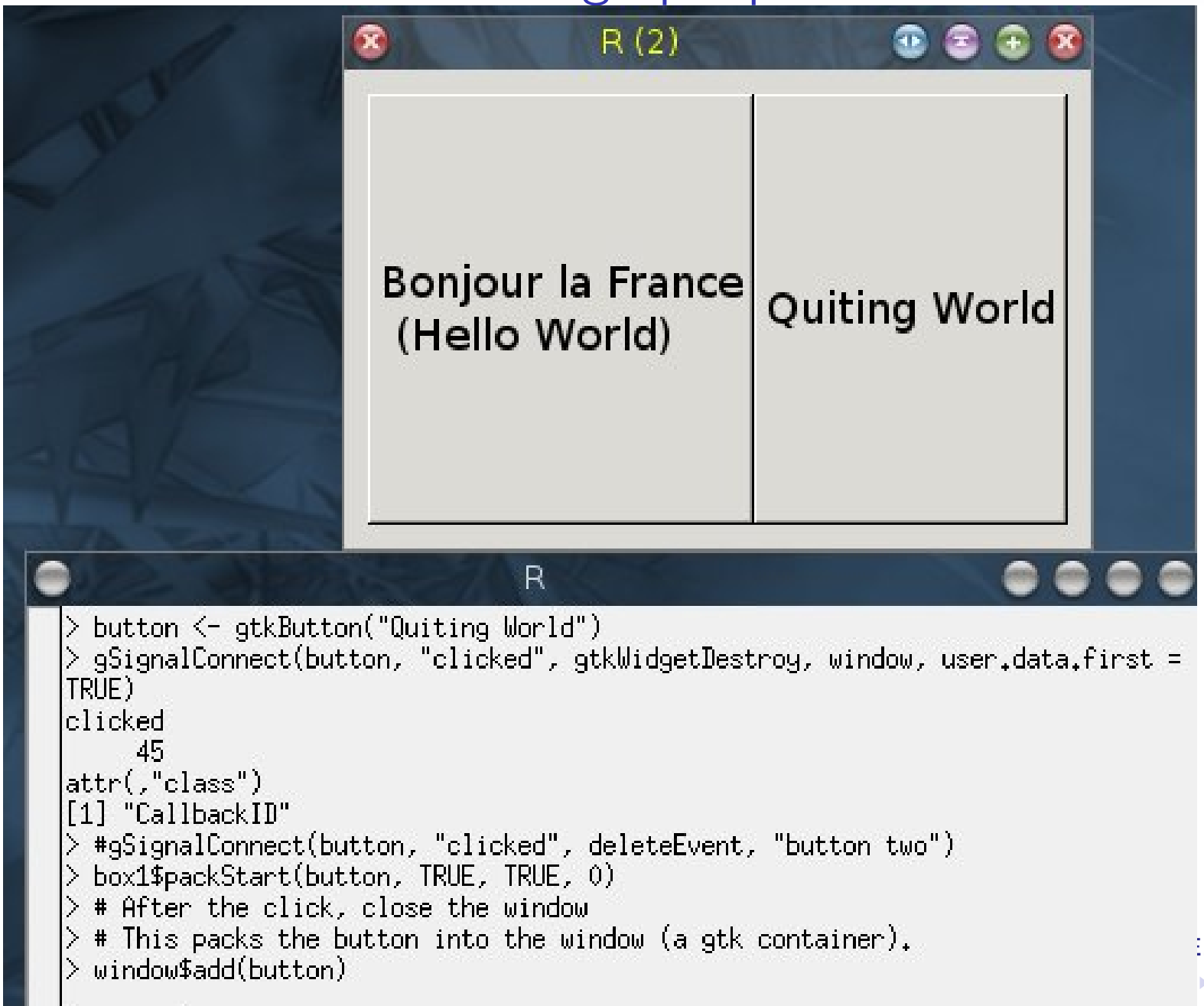
R et C (débogage avec ddd)

exposé



Créer une interface graphique avec R

exposé



Historique : S

Au début S (« Statistics »)

- Bell Laboratories, 1976
- 1980 première version publique
- 1988 « blue book » : Fortran $\longrightarrow$ C, fonction, devices (X11, postscript)
- 1991 Statistical Models in S « white book » : formules, méthodes, classes

Historique : R

R crée par Robert Gentleman & Ross Ihaka (lettre R).
R avant S.
Implémentation gratuite de S avec portée lexicale (lexical scoping)

- Version 0.16 début de la « mailing list » : 1er avril 1997
- Version 1.0.0 - 29 février 2000
- Version 2.0.0 – 4 octobre 2004 : lazy loading (fast loading of data with minimal expense of system memory)
- Version 2.9.0 - 17 avril 2009 : Package 'Matrix' recommandé dans la distribution basic

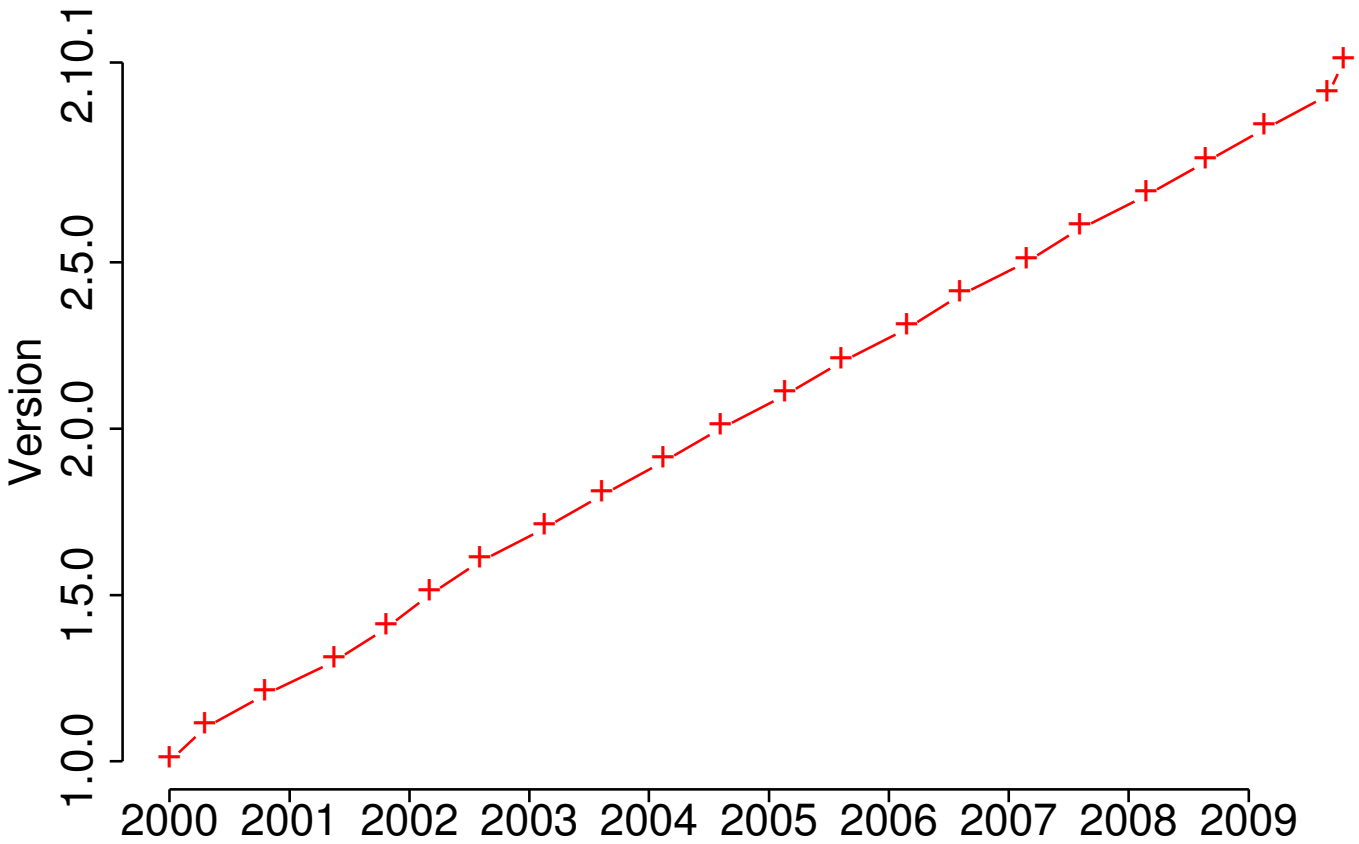
Caractéristiques

- Langage pour les statistiques
- Gratuit, partie du projet GNU depuis le 5 décembre 1997.
- Multiplateforme & Multi OS depuis 1997
- Modulaire : possibilités de base extensibles par des « packages » (équivalent module scilab)

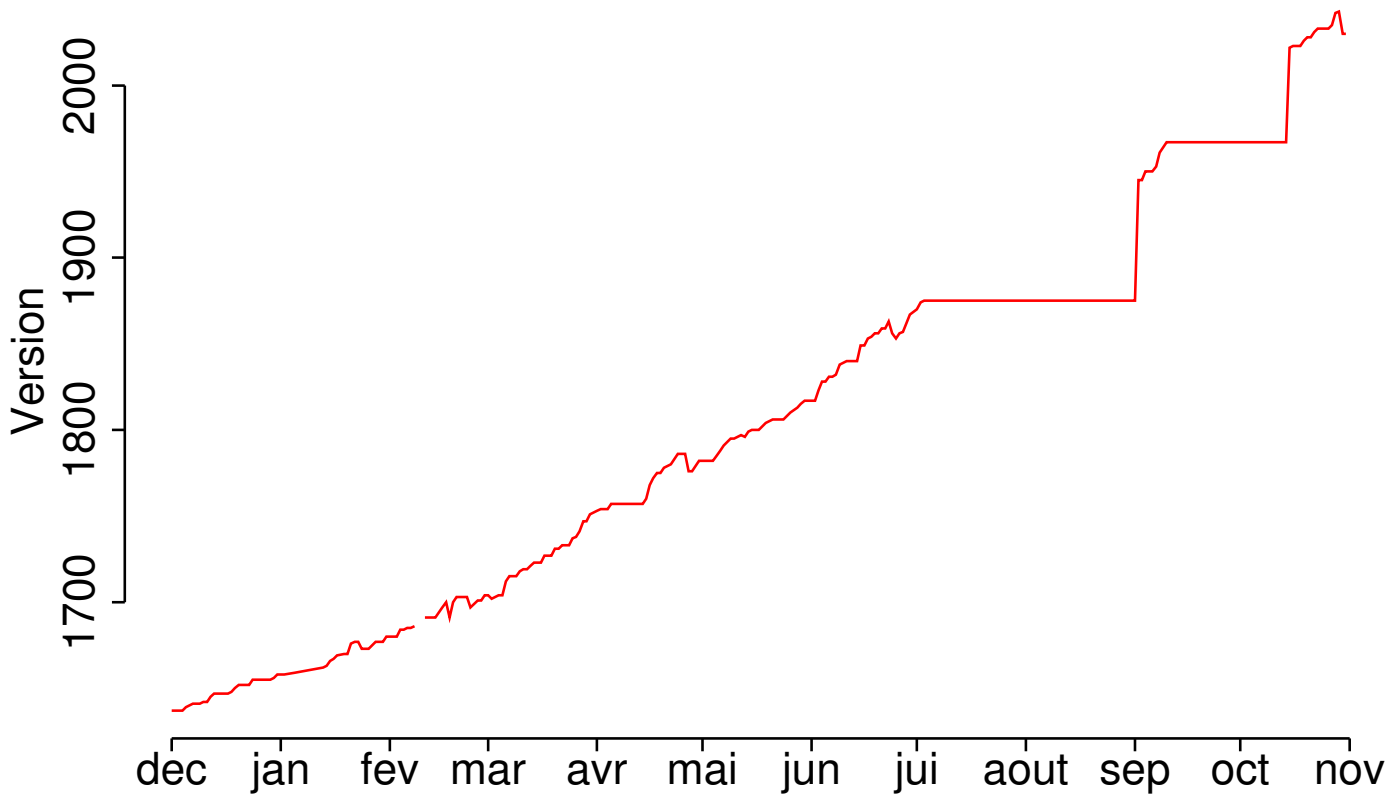
Installation, page(s) web, aide

1. Page du projet : `http://www.r-project.org/`
2. CRAN : `http://cran.univ-lyon1.fr/`

Versions


Version
1.0.0
1.5.0
2.0.0
2.5.0
2.10.1
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009

Packages


Version
1700
1800
1900
2000
dec
jan
fev
mar
avr
mai
jun
jui
aout
sep
oct
nov

Livres

Documentation en plein essor :

- Plus de 80 livres
  `http://www.r-project.org/doc/bib/R-books.html`
- Illustration de méthodes avec R :
  - Hidden Markov Models for Time Series : An Introduction Using R (Chapmann & Hall)
  - Statistical Data Analysis Explained : Applied Environmental Statistics with R (Wiley)
- Collection spécifique : UseR ! (Springer)

Livres

Des livres en français

1. *Statistiques avec R*, Presses Universitaires de Rennes, France (2008)
2. *Comprendre et réaliser les tests statistiques à l'aide de R*, de Boeck université, Louvain-la-Neuve, Belgique (2009)
3. *Analyse de données avec R*, Presses Universitaires de Rennes, France (2009)
4. *PratiqueR* une collection française chez Springer

Points forts

- Simulation/Programmation
  - langage de statistiques
  - command line interface (CLI) : script et programmes $\rightarrow$ simulations
- cluster de calcul (hétérogène) : packages `snow`, `Rmpi`, interfaces web etc.
  voir `http://epub.ub.uni-muenchen.de/8991/`
- Effet de masse critique

Comparaison de méthodes

Grâce aux packages
par exemple la discrimination :

- CART (`rpart`),
- Random Forest (`randomForest`),
- Mixture and Flexible Discriminant Analysis (`mda`)
- Boosting (`ada`, `mboost`)
- SVM (`e1071`)

Populariser sa méthodologie

1. (Proposer une méthode et exposer dans un article ses propriétés)
2. Ecrire et déposer un package sur CRAN
3. (Publier dans « journal of statistical software » `http://www.jstatsoft.org/`)

Reproductibilité de la recherche

principe de Claerbout (Géophysicien, Stanford)

> *An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.*

1. Ecrire et déposer un package sur CRAN
2. Décrire la méthode, ses propriétés et ses résultats (avec l'implémentation)
   comme les livres de statistiques...

Simplicité de la création d'un package

Objectif

Fonction pour la représentation d'une variable quantitative discrète : diagrammes en bâtons.

Organisation

1. Fichier DESCRIPTION
2. Répertoire `R` : fonctions R (fonction `batons`)
3. Répertoire `man` : documentation des fonctions
4. Répertoire `data` : données
5. (Répertoire `src` : pour les fichiers à compiler, header etc.)

Simplicité de la création d'un package

- Aide complète dans le Manuel *Writing R extensions*
- Liste des mots clefs

  `> file.show(file.path(R.home("doc"), "KEYWORDS"))`
- Création des packages sous windows
  - « Windows toolset » : `http://cran.univ-lyon1.fr/doc/manuals/R-admin.html#The-Windows-toolset`
  - Création en 1/2 heure automatiquement : `http://win-builder.r-project.org/`

Interfaces

Widgets (fenêtres, menus etc. dans R)

Les standards

- Tcl/Tk
- Gtk
- Qt (voir conférence UseR ! 2009 à Rennes)

Vers un effort d'uniformisation

- `iwidgets`
- `SciViews`

Piloter des programmes externes

- C : fonction .C
- Fortran : .Fortran
- C avec expression R : .Call
- C++ (voir conférence UseR ! 2009 à Rennes)
- java : `rJava`

Piloter des programmes R

- en mode batch à partir du shell

  `R --vanilla < commandesbatch.r > sorties.r`

- en python `http://rpy.sourceforge.net/`

Enseignement

Constatation

Outil naturel de l'enseignant-chercheur

Question

Raisonnable pour l'enseignement ?

Problèmes ?

Langage à apprendre

- Temps non disponible pour les stats
- Difficulté d'apprentissage & Autonomie

Réponses ?

Temps non disponible pour les stats

- proposer les commandes
- proposer une fonction « boite noire »

Difficulté d'apprentissage & Autonomie

consolide les aptitudes en informatique/capacités d'abstraction

- fichier de commandes $\longrightarrow$ question de l'éditeur (sous windows tinn-R ?)
- couper/coller à partir du pdf
- aide partielle
- tous documents
- fiches
- aide partielle
- commandes, boites noires

Graphical User Interface

1. `Rcmdr` : R commander
   Exemple d'une régression linéaire
2. `pmg` : poor man gui

Tableurs et statistiques

Pour le moment gratuit...

1. `Rexcel` : R et Excel
2. `ROoo` : R et OpenOffice

Packages et boites noires

Dans un package

1. Inclure des fonctions « boite noire »
2. Inclure des données

On ne peut pas faire plus simple...

Graphiques et carto

1. Pour l'aspect spatial voir `http://geodacenter.asu.edu/r-spatial-projects`.
   Un exemple : package `maps`
2. Pour tracer des graphiques en 3D : package `rgl`
3. Exploration des données `rggobi` (et ggobi)
4. Voir aussi les packages `ggplot` ou `lattice`

Base de données

- packages spécifiques pour une base de données : `RMySQL`, `RPostgreSQL`, `RSQLite`, `ROracle`
- package de driver ODBC `RODBC`

Points faibles

- Langage interprété (lent sur les boucles)
- Les données sont stockées en mémoire $\longrightarrow$ problème de mémoire
- Agrément FDA (et pas de SS de type III)
- Interlocuteur en cas de problème

Points forts

- CLI (scripts)
- prix + mailing list très active
- interaction avec base de données
- GUI (avec rappel de commandes) ?
- graphiques complets

Entreprise

Deux environnements :

- Exploratoire
- Production

R en entreprise

Exploratoire

Outil idéal : faible volume, nombreuses méthodes, graphiques

Production

Selon le volume de données et les méthodes (`biglm`)
Problème de l'agrément FDA.