Pourquoi R devient incontournable en recherche, enseignement et développement

E. Matzner-Løber

Rencontre R, BoRdeaux 2012

Plan

- Introduction
- Recherche
- Enseignement
- Développement (entreprise)
- Conclusions

Les logiciels de Statistique

- SAS
  intégration totale de la donnée au reporting, leader dans les grandes entreprises, logiciel de référence, coût élevé
- SPSS, produit IBM
  GUI, possibilité de menus ou de programmation, intégrations externes (R, Pyhton,...) facilitées, coût moins élevé
- R
  Open Source, évolutif, La réference scientifique, en augmentation constante
- STATA, **Excel**.....

Popularités respectives

Une belle étude de R. Muenchen

http ://r4stats.com/popularity

Beaucoup de classements

- TIOBE community Programming (R 24th, SAS 31th) en 2012
- trafic mensuel sur les mailings listes (R premier)
- nombre de blogs dédiés
- compétition data analysis contests au début 2012 ; 25000 analystes, 72 000 problèmes, R premier et 50 % parmi les vainqueurs
- questionnaire de Rexer Analytics, quels outils utilisez-vous pour le data mining : R 40 %, SAS et SPSS 30 % (plus utilisés que leurs versions dédiées !)
- questionnaire de KDnuggets R devant Excel

Historique : S

Au début S (« Statistics »)

- Bell Laboratories, 1976
- 1980 première version publique
- 1988 « blue book » : Fortran → 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.

- 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

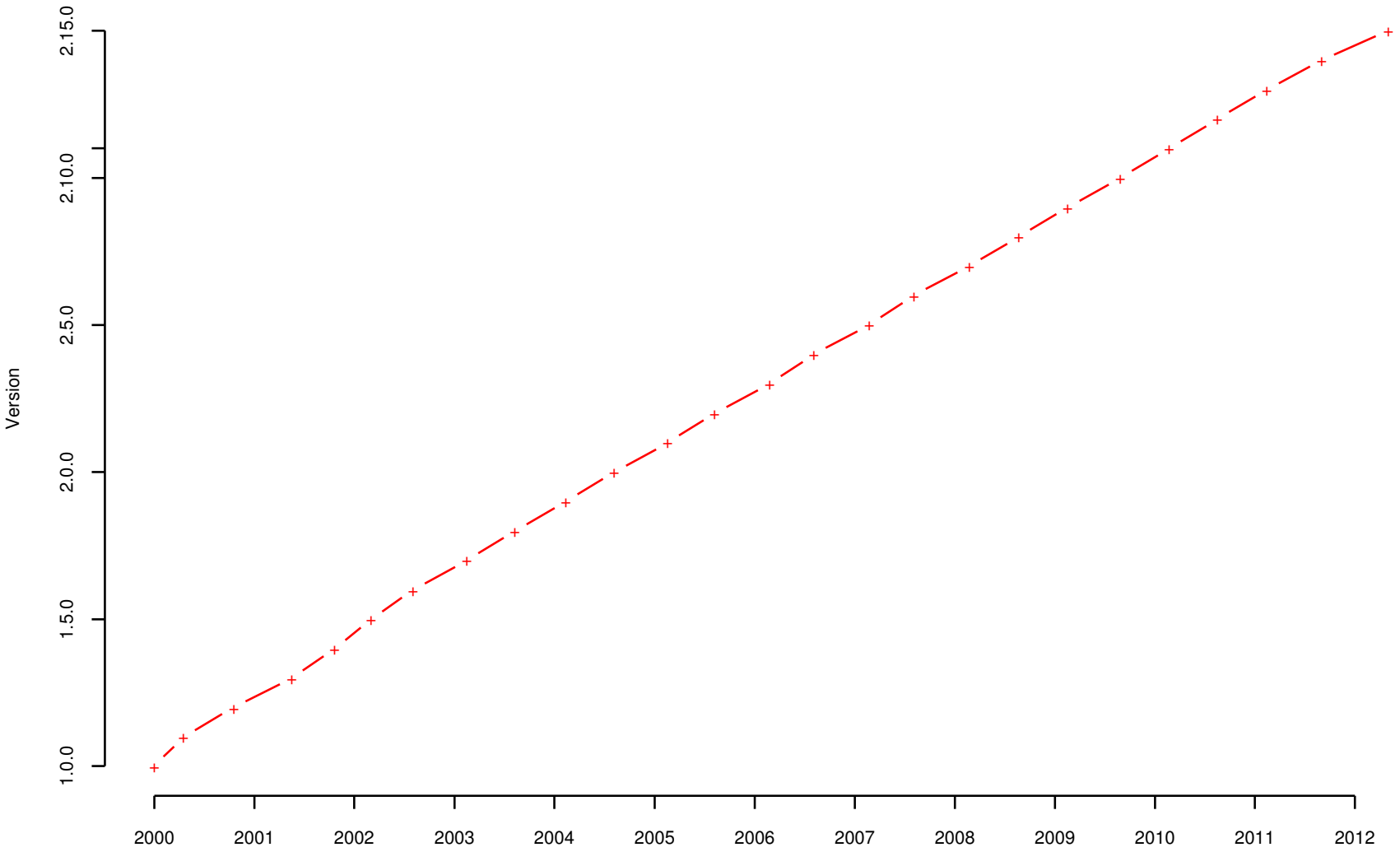
Caractéristiques

- 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)

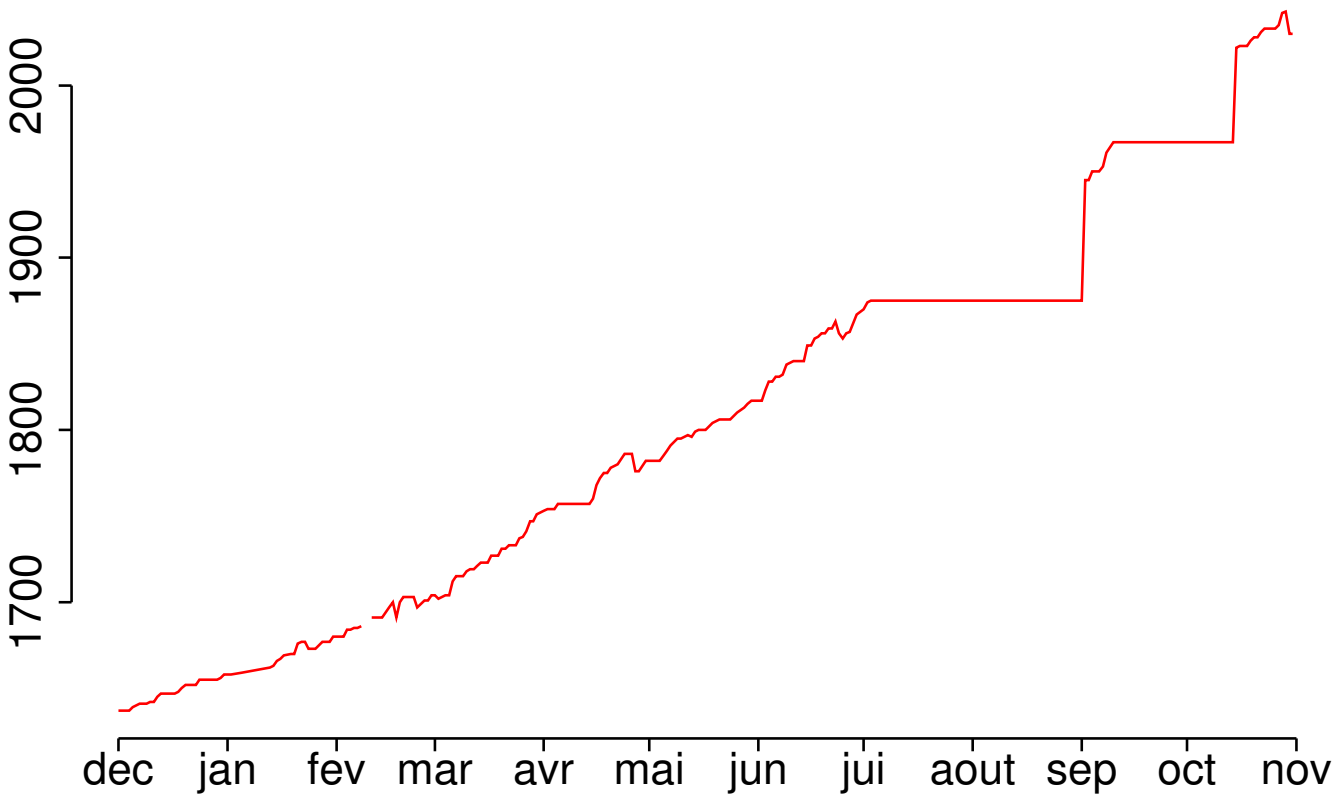
Installation, page(s) web, aide

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

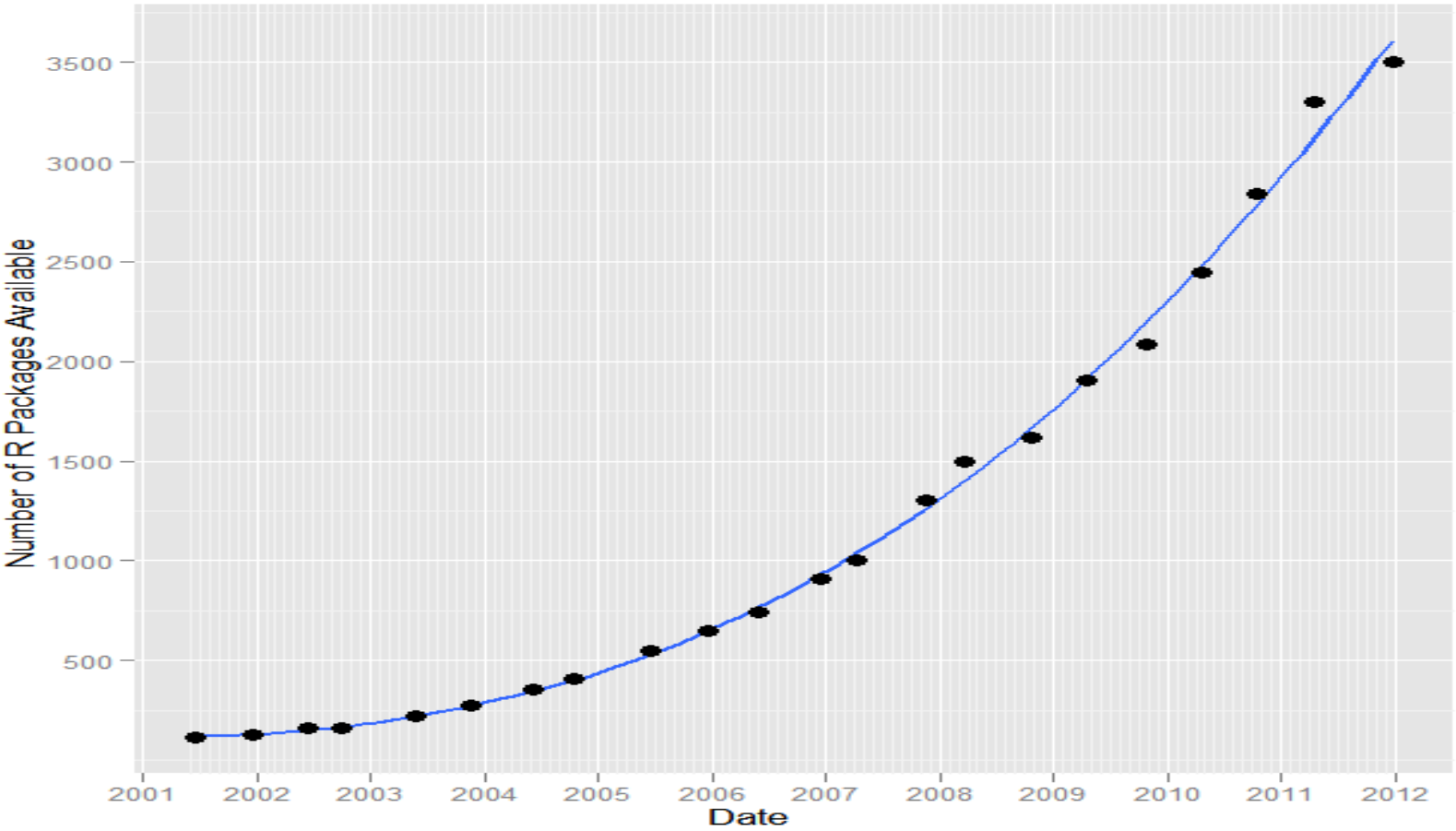
Versions


Version
1.0.0
1.5.0
2.0.0
2.5.0
2.10.0
2.15.0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012

Evolution des packages en 2009



Packages hors bioconductor



Facteur important de la croissance de R, la capacité à traiter tous types de problèmes, à comparer des méthodes...

Livres

Documentation en plein essor :

- Plus de 110 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
  - Séries temporelles avec R PratiqueR, Springer
- Collections spécifiques : UseR ! PratiqueR (Springer)

Livres en français

1. *Statistiques avec R*, Presses Universitaires de Rennes, France (2008,2010,2012)
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. *Le logiciel R*, Springer
5. *PratiqueR* une collection française chez Springer, **projets bienvenus**

Positionnement

- Statistique, branche des mathématiques appliquées
  - possibilité de simulations
  - possibilité de traiter des données réelles

    $\rightarrow$ besoin d'outils
- Effet de masse critique au contraire des autres branches : Matlab, Scilab, Octave,
  calcul symbolique : Maxima, géométrie algébrique : Cocoa, Singular, Macaulay, groupes : GAP, arithmétique : Pari, Axiom ou Sage/Python scientifique.

Points forts

- Effet de masse critique
- Simulation/Programmation
  - Langage de statistiques
  - Command line interface (CLI) : script et programmes → simulations
- Cluster de calcul (hétérogène) : packages `snow`, `Rmpi`, interfaces web etc. voir `http://epub.ub.uni-muenchen.de/8991/`

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/`

Piloter des programmes externes

- C : fonction .C
- Fortran : .Fortran
- C avec expression R : .Call
- 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

- Difficulté de créer des nouveaux cours (LRU)
- Prendre le temps sur les cours de Statistique
- Difficulté d'apprentissage & Autonomie

mais de plus en plus d'enseignants potentiels !

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

Graphical User Interface

1. `Rcmdr` : R commander
2. `pmg` : poor man gui

Packages et boites noires

Dans un package

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

On ne peut pas faire plus simple...

Tableurs et statistiques

Pour le moment gratuit...

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

Graphiques et carto

1. Pour l'aspect spatial voir `http://geodacenter.asu.edu/r-spatial-projects`. Un exemple : package `maps` ou `RgoogleMaps`
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 $\rightarrow$ problème de mémoire (cf les différentes versions de R, de l'OS, 32 ou 64bt)
- Positionnement par rapport à SAS, SPSS, Excel

Points forts

- prix + mailing list très active
- possibilité d'installation locale régulières et non pas par le DRI (CRI)
- CLI (scripts)
- interaction avec base de données
- GUI (avec rappel de commandes) ?
- graphiques complets

Etats des lieux

Les entreprises ont déjà des compétences dans d'autres logiciels SAS, SPSS et il faut donc envisager le coût d'une double (triple) compétence

ou n'ont pas de logiciel spécifique et font déjà tout avec un tableur

Points faibles

- Interlocuteur en cas de problème
- Les données sont stockées en mémoire $\rightarrow$ problème de mémoire
- Agrément FDA
- compétences dans d'autres logiciels

Points forts

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

Entreprise

Deux environnements :

- Exploratoire
- Production

Développement

Exploratoire, outil idéal

- faible volume
- nombreuses méthodes implémentées
- graphiques développés

Production

Peut-être compliqué si volume de données important mais des packages existent

- biglm
- bigmemory
- biganalytics
- ...

Production suite

On peut utiliser R directement depuis

- SPSS : tous les outils de modélisations ne sont pas implémentés
- SAS
- Excel

Oracle (Oracle analytics est basé sur R) Oracle R Enterprise Integrating Open Source R with Oracle Database 11g
Revolution Analytics ...

Conclusions

Développements futurs, notre rôle est de positionner R dans

- enseignements
- maquettes
- stages
- conseils lors des stages
- écriture de livres
- traiter des interfaçages de R avec les bases de données (RODBC, RMySQL)
- ...