

Présentation de R

Sophie Baillargeon

2021-01-14

Qu'est-ce que R ?

- un **environnement** et un **langage de programmation** pour effectuer des *calculs statistiques* et créer des *graphiques*;
- un **logiciel libre** : logiciel *gratuit*, distribué avec son *code source*, il peut être *étudié, modifié et partagé librement*;
- un logiciel formé des composantes suivantes :
 - **R de base** (fonctions statistiques et graphiques standards),
 - **extensions appelées packages** (collections de nouvelles fonctions créées par des utilisateurs).

Que fait-on avec R ?

Typiquement, on utilise d'abord R pour faire de l'**analyse statistique de données**. On réalise donc en R les tâches suivantes :

- manipuler des **données**;
- appeler des fonctions préexistantes de **calculs statistiques** (estimation de statistiques, ajustement de modèle, etc.);
- produire des **graphiques**;
- rédiger des **rapports d'analyse** de données.

Plusieurs utilisateurs de R doivent un jour **développer leurs propres fonctions** de calcul statistique. On peut devenir un développeur lorsque :

- on a besoin de faire la même analyse à plusieurs reprises (**automatiser des calculs**);
- on souhaite **améliorer les implantations existantes** d'une méthode de calcul;
- on cherche à **développer de nouvelles méthodes** de calcul.

On passe d'utilisateur de R à développeur R souvent sans même s'en rendre compte.

1990 - Au Département de statistique de l'**Université d'Auckland** en Nouvelle-Zélande, **Ross Ihaka** et **Robert Gentleman** (alors en année sabbatique de l'Université de Waterloo au Canada) ont l'idée de créer un nouveau logiciel statistique pour tester quelques idées dans leurs **travaux de recherche**.

1992 - Le langage est nommé **R**, car il s'agit de la première lettre des prénoms des deux créateurs et parce qu'il est décidé que le langage utilisera la syntaxe du **langage S** développé dans les Bell Laboratories par **John Chambers** et collaborateurs.

1994 - Une version initiale du logiciel est utilisée pour **donner des cours** d'introduction à la statistique et elle est **distribuée sur internet**. Il est décidé que R sera un **logiciel libre**.

Robert Gentleman et Ross Ihaka dans les années 90 :



Source : <https://www.stat.auckland.ac.nz/~ihaka/downloads/Otago.pdf>

1996 - **Martin Mächler** de l'École polytechnique fédérale de Zurich en Suisse se joint à l'équipe des R & R pour développer R, qui connaît de plus en plus de succès. Malgré tout, bien vite trois personnes ne suffisent plus pour gérer le flot constant de courriels reçus d'utilisateurs.

1997 - Le CRAN est créé par **Kurt Hornik** et **Friedrich Leisch** à l'Université technique de Vienne en Autriche. Il s'agit d'un dépôt informatique pour les contributions des utilisateurs (packages). Peu après, le « **R core team** », soit le noyau de développeurs de R, est formé.

2000 - La première version officielle de R est publiée à une date particulière : le **29 février 2000!**

Les membres du *R core team* sont typiquement des chercheurs en calcul statistique. Ils travaillent pour différentes organisations, souvent des universités, aux quatre coins du monde.

- Depuis 2003, la **fondation R** amasse des dons et les utilisent pour maintenir des infrastructures et pour commanditer quelques travaux de développement.

<https://www.r-project.org/foundation/>

- Des **conférences** sont organisées et réunissent les développeurs de R ainsi que des utilisateurs de R.

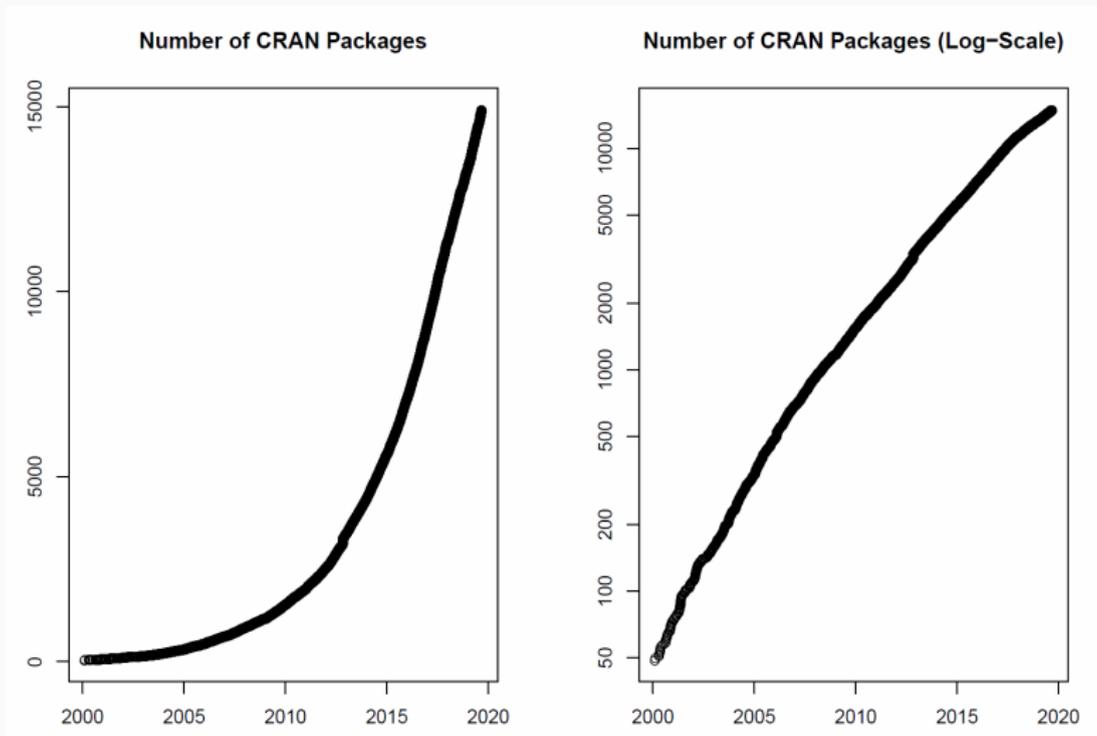
<https://www.r-project.org/conferences/>

- R a aussi sa **revue scientifique**, *The R journal* (successeur des *R News*). **<https://journal.r-project.org>**

- R est en **constant développement**. Il est mis à jour environ 4 à 6 fois par année. La version courante de R (4.0.3) est téléchargeable sur le site web du projet R : **<https://www.r-project.org>**.
- Il y a maintenant :
 - **presque 17 000 packages** sur le CRAN (**<https://cran.r-project.org/web/packages/>**);
 - **presque 2000 packages** sur Bioconductor, un dépôt informatique pour des contributions R spécialisées en bio-informatique (**<http://www.bioconductor.org/>**);
 - plusieurs packages R uniquement disponibles sur GitHub ou ailleurs sur le web.

Note : Informations relevées le 14 janvier 2021.

Évolution exponentielle du nombre de packages sur le CRAN



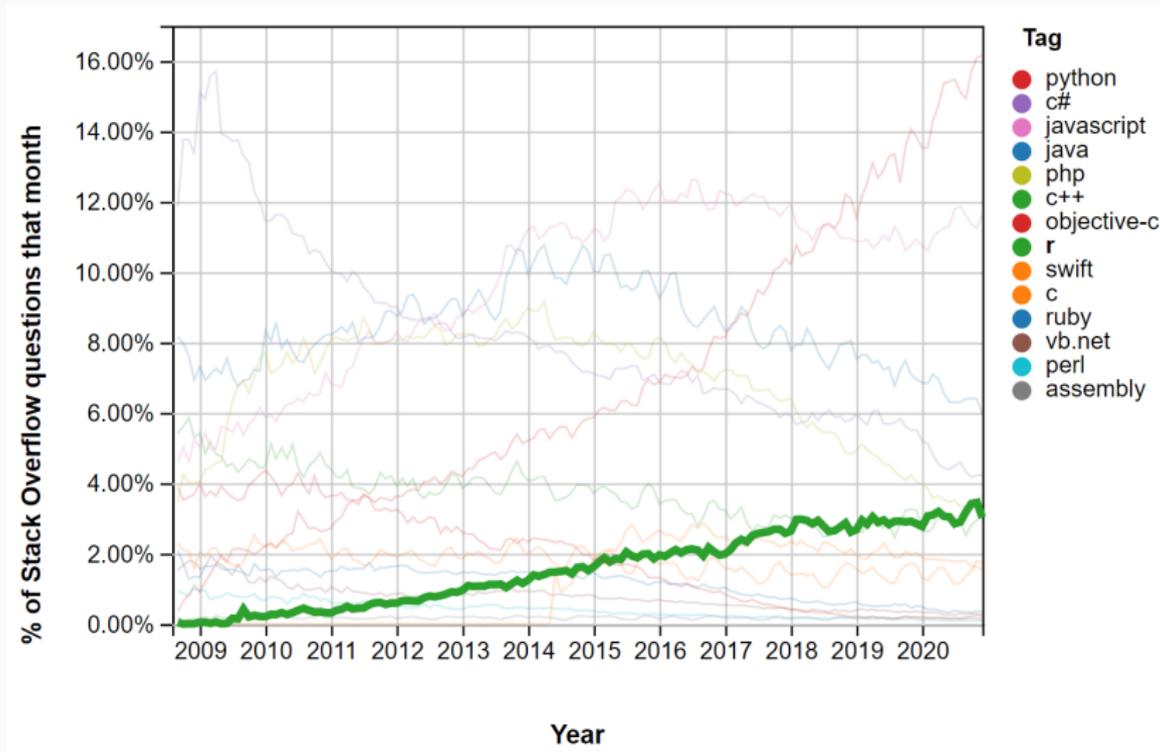
Kurt Hornik, Uwe Ligges, Achim Zeileis. (2019). Changes on CRAN. The R Journal. Vol. 11 (1), pp. 438-441. <http://journal.r-project.org/archive/2019-1/cran.pdf>

- Beaucoup de gens l'utilisent et en parlent sur internet.
- L'environnement intégré de développement RStudio (<https://rstudio.com/>) et les packages du tidyverse (<https://www.tidyverse.org/>) contribuent à la popularité de R.
- Savoir programmer en R est une compétence fréquemment demandée pour des emplois en statistique / science des données. (<https://github.com/ThinkR-open/companies-using-r>)

Quelques statistiques sur la popularité de R :

<http://r4stats.com/articles/popularity>

Augmentation du nombre de questions relatives à R sur Stack Overflow



Source : <https://insights.stackoverflow.com/trends>, 14 janvier 2021.

- logiciel libre : a l'avantage d'être **gratuit** et de favoriser la **recherche reproductible**;
- **langage interprété** : langage plus proche de notre langage que du langage machine, donc plus simple et direct que, par exemple, C ou C++;
- **partage et réutilisation de code facilité** grâce au système des packages et au CRAN;
- communauté active de développeurs et d'utilisateurs :
 - R **évolue vite**, ses bogues sont identifiés et corrigés rapidement;
 - il y a **beaucoup d'information** concernant la programmation en R sur internet;
 - le **nombre de packages R est toujours grandissant**, ainsi de nouvelles fonctionnalités sont fréquemment ajoutées à R.

Une communauté pour supporter les utilisateurs



Illustration de @allison_horst

<https://github.com/allisonhorst/stats-illustrations>

Les caractéristiques de R sont particulièrement adaptées au monde de la recherche et de l'enseignement. Il est donc très utilisé dans les universités. Par contre, il comporte les limites suivantes :

- logiciel libre : certaines organisations **préfèrent utiliser un logiciel commercial**;
- langage interprété : **R est parfois lent** pour réaliser certains calculs;
- gestion de la mémoire : **le R de base est plus ou moins adapté à la manipulation de données volumineuses** (mais de plus en plus de packages sont créés pour contrer ce problème).

Certains préfèrent utiliser un logiciel commercial, car :

- Un logiciel libre n'offre **aucune garantie officielle quant à la validité de ses résultats**.
 - Cependant, Keeling et Pavur (2007) arrivent à la conclusion que les résultats obtenus en R pour les analyses statistiques les plus courantes sont tout aussi exacts que ceux obtenus d'autres logiciels statistiques.
- Un logiciel libre n'offre **pas de soutien technique**.
 - Cependant, plusieurs sites de questions/réponses (comme **<https://stackoverflow.com/>**) ainsi que des listes courriel existent. On obtient rapidement une réponse d'autres utilisateurs lorsque l'on y soumet une question.

Est-ce que ça vaut la peine de se perfectionner en R ?

Oui!

Ça pourrait vous permettre de **décrocher un emploi**.

Sur certains aspects, R surpasse ses concurrents (SAS, Python, etc.) :

- plus **grande offre d'implémentations de méthodes statistiques**,
- un nouvel utilisateur peut **rapidement être capable** de mener une **analyse exploratoire de données**,
- **communication de résultats** facilitée par ses outils de production de **graphiques, rapports et applications web interactives**.

<https://github.com/matloff/R-vs.-Python-for-Data-Science>

<https://blog.rstudio.com/2019/12/17/r-vs-python-what-s-the-best-for-language-for-data-science/>

Références intéressantes pour en savoir plus

- Fox, J. (2009). Aspects of the Social Organization and Trajectory of the R Project. *The R Journal*. Vol. 1 (2), pp. 5-13. http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Fox.pdf
- Ihaka, R. (2011, April 20). The R Project: A Brief History and Thoughts About the Future [Présentation].
<https://www.stat.auckland.ac.nz/~ihaka/downloads/Otago.pdf>
- Keeling, K. B. et Pavur, R. J. (2007). A comparative study of the reliability of nine statistical software packages. *Computational Statistics & Data Analysis*. Vol. 51, pp. 3811-3831.
- Peng, R. (2018, 12 juillet). Teaching R to New Users - From tapply to the Tidyverse [Billet de blogue].
<https://simplystatistics.org/2018/07/12/use-r-keynote-2018/>
- Thieme, N. (2018). R generation. *Significance magazine*, Vol. 15 (4), pp. 14-19.
<https://doi.org/10.1111/j.1740-9713.2018.01169.x>