Une bouffée d'R pour les actuaires

La technicité croissante des modèles utilisés en actuariat nécessite de recourir à des logiciels performants. Dans ce domaine, SAS et ses surcouches (telles que Emblem, Pretium, Earnix,...) se positionnent comme des logiciels de statistiques leader en entreprise. Mais après seulement quinze ans d'existence, le logiciel R, développé par deux néo-zélandais, cf. [1], et initialement utilisé dans le milieu universitaire, est devenu un outil incontournable de statistiques et de visualisation de données.

Avec ses 35 listes de discussion et plus de 450 blogs actifs (avec au moins une mise à jour par semaine), le logiciel R est largement utilisé et il est porté par une communauté active. Il est toujours largement utilisé dans le domaine universitaire, mais aussi dans des entreprises telles que Google ou Oracle (deux des sponsors de la dernière conférence UseR) ou encore par le marché de l'assurance britannique Lloyds of London et par la plupart des assureurs français. C'est aussi le logiciel privilégié dans le cadre de compétitions comme celles proposées par le site du Kaggle, [2].

On peut se demander pourquoi le logiciel R semble si attractif.

Premier atout, c'est un logiciel libre. "*Libre"* signifie qu'il est gratuit, ce qui constitue en soit un avantage non négligeable surtout compte tenu du prix des licences d'autres logiciels. Outre la gratuité, "*Libre"* sous-entend aussi que le code source est accessible à n'importe quel utilisateur ([3]) : c'est la fin des boîtes noires ! On peut enfin visualiser le code préprogrammé, et ainsi connaître la méthode implémentée. Pouvoir disposer très facilement du code sous-jacent permet à la fois de mieux comprendre les fonctionnalités et aussi de se les approprier en les modifiant. Cette flexibilité ouvre de multiples perspectives.

De plus, R n'est pas un simple logiciel de statistiques, mais aussi un langage pour les calculs statistiques et pour les graphiques. Il contient près de 32 000 fonctions préprogrammées, contre 1 200 implémentées dans la dernière version de SAS. Mais comme les fonctions sont emboîtables et personnalisables, le nombre de possibilités s'avère en réalité infini. En tant que langage, le logiciel R permet donc d'implémenter n'importe quelle idée.

Il n'est pas nécessaire d'être informaticien pour manipuler le logiciel R. On peut même créer très facilement ses propres fonctions. Et si nécessaire, le logiciel R peut aussi interagir avec des langages plus avancés, en intégrant par exemple des parties de programme en C ou C++. Il permet aussi de piloter ou d'interagir des programmes externes tels que (ACCESS, SAS, ou les systèmes de base de donnés Oracle, MySQL,...)

R est donc un logiciel très complet qui permet de traiter tous types de données et de générer des graphiques complexes. Mais son atout essentiel réside dans la communauté qui le porte. En effet, outre les analyses classiques chargées par défaut lors de son installation, de nombreux **packages** sont disponibles depuis le site officiel http://cran.r-project.org. Un package est un ensemble de fonctionnalités documentées, développées (généralement par des

enseignants-chercheurs) dans le cadre de projets de recherche, et rendues libres. L'accroissement du nombre de packages est fulgurant : on en compte déjà plus de 5 000, cf. [4]. Ils couvrent de très nombreux domaines (écologie, épidémiologie, optimisation, théorie mathématique, finance, assurance...), certains parfois inattendus (sons, musique, création de pages web interactives, aide à la création d'un package, interaction avec la wii...), et les fonctionnalités proposées sont de plus en plus spécialisées, cf. [5]. R est donc un logiciel en essor permanent, qui propose les développements les plus novateurs (méthode ABC en statistique bayésienne, machine learning, étude des valeurs extrêmes...) qui lui permet de garder un temps d'avance.

En ce qui concerne l'actuariat, les outils actuariels les plus classiques et les plus novateurs y sont développés (calibration de loi de sinistres, modèle de Lee-Carter et autres modèles à mortalité stochastique, modèles de provisionnement non-vie, valorisation d'options, théorie des extrêmes,...). Parmi les nombreux packages utiles, les packages actuar et ActuDistns fournissent la majorité (pour ne pas dire l'intégralité) des lois de probabilités utiles en actuariat. Le premier propose aussi des fonctions pour les modèles de crédibilité et la théorie du risque. De plus, le package lifecontingencies implémente les principales méthodes de calcul d'annuités et de capital décès en assurance vie. La grande flexibilité du logiciel permet en outre de créer des analyses sur mesure. Un ouvrage à venir édité par Arthur Charpentier regroupe une grande quantité d'applications actuarielles et financières sous R au travers de 16 chapitres, cf. [6]. Notons que cette année s'est tenue la première conférence R in insurance, cf. [7], à Londres, qui a montré à quel point R est utile en actuariat.

On a longtemps reproché à R une incapacité à traiter de gros volumes de données. En effet, R, comme d'autres logiciels (matlab, scilab, octave,...), travaille en mémoire vive. Par conséquent dans le passé, on atteignait vite la limite de 2Go ou de 4Go suivant le système. Comme la plupart des systèmes d'exploitation utilisent de nos jours un adressage 64 bit, cette limite théorique est devenue une limite matérielle. Ainsi, R permet de gérer d'assez gros volumes de données du moment que l'ordinateur dispose d'une grande quantité de mémoire vive. Dans le cas où on attendrait la limite matérielle, le package ff permet de travailler sur le disque dur plutôt qu'en mémoire vive et réplique ainsi les "external-memory" algorithmes utilisés par SAS.

Les arguments classiques avancés contre les logiciels libres sont leur manque de fiabilité et de sécurité. Qu'en est-il pour le logiciel R ? D'un point de vue sécurité, il faut souligner que R ne contient pas de virus car seuls les membres de R Core Team (qui sont les premiers utilisateurs de R) ont accès en écriture sur le dépôt svn. De toute façon, rien n'empêche quiconque de vérifier le code source [3]. Enfin, du fait de la grande communauté de R, les éventuels bugs sont trouvés et corrigés très rapidement voire plus rapidement que sur les logiciels propriétaires. A ce propos, le numéro de version de R est déterminé selon la règle suivante 'major.minor.patchlevel', où patchlevel correspond au numéro de correction de bugs.

Libre et très complet, le logiciel R connaît un essor particulièrement important ces dernières années, cf. [8,9,10]. Il se pose comme une alternative particulièrement intéressante dans le monde de l'entreprise. D'ailleurs, beaucoup d'actuaires ont déjà adopté le logiciel R, [11, 12, 13].

Alors, pas d'hésitation, prenez une bouffée d'R.

Manuela Royer-Carenzi et Christophe Dutang

Références :

[1] http://cran.r-project.org/doc/html/interface98-paper/paper.html

[2] http://www.kaggle.com/

[3] http://cran.r-project.org/src/base/R-3/

[4] http://r.789695.n4.nabble.com/Milestone-5000-packages-on-CRAN-td4680090.html

[5] http://cran.r-project.org/web/views/

[6] http://www.crcpress.com/product/isbn/9781466592599

[7] http://www.cass.city.ac.uk/news-and-events/conferences/r-in-insurance

[8] http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?_r=0

[9] http://blog.revolutionanalytics.com/2011/03/how-the-new-york-times-uses-r-for-data-visualization.html

[10] http://www.theactuary.com/archive/old-articles/part-6/general-insurance-3A-r-you-ready-3F/

[11] http://opensourcesoftware.casact.org/

[12] http://blog.casact.org/2013/10/02/using-r-for-actuarial-work/

[13] http://www.institutdesactuaires.com/gene/main.php?base=364&action=details&id_news=3305

Biographies

Manuela Royer-Carenzi est agrégée de mathématiques (1998) et diplômée de l'ENS. Après avoir soutenu son doctorat de probabilités (2003), elle a eu un poste d'enseignant-chercheur à l'ISFA (2004 à 2007). Depuis septembre 2007, elle est maître de conférences à l'Université de Provence, cf. http://www.cmi.univ-mrs.fr/~carenzi/

Christophe Dutang a été diplômé de l'Ensimag (2007) et l'ISFA (2008). Il a effectué une thèse Cifre au Group Risk Management d'AXA entre 2008 et 2012. Une fois le diplôme de doctorat en poche, il a été maitre de conférence à l'Université de Strasbourg enseignant notamment dans la filière d'actuariat. Depuis septembre 2013, il est maitre de conférence à l'Université du Maine (Le Mans) et l'Institut du Risque et de l'Assurance (IRA). Enfin, il est membre certifié de l'Institut des Actuaires, cf. http://dutangc.free.fr