Une interface graphique pour analyser des données distantes sous R

Raphaël Coudret, Gilles Durrieu, Jerôme Saracco

Une interface graphique pour analyser des données distantes sous R

R. Coudret[a] and **G. Durrieu**[b] and **J. Saracco**[a]

[a]Équipe CQFD et Institut de Mathématiques de Bordeaux UMR CNRS 5251
INRIA et Université de Bordeaux
351 cours de la Libération, 33405 Talence
{rcoudret, Jerome.Saracco}@math.u-bordeaux1.fr

[b]Laboratoire de Mathématiques de Bretagne Atlantique UMR CNRS 6205
Université de Bretagne Sud
Campus de Tohannic, 56017 Vannes
gilles.durrieu@univ-ubs.fr

Mots clefs : Analyse de données, Base de données, Interface graphique.

Utiliser ou faire utiliser des méthodes statistiques récentes pour analyser un jeu de données peut représenter une tâche ardue. Lorsqu'il s'agit de s'intéresser à des données dont la diffusion doit être contrôlée ou qui représentent un volume important, la question de leur accès et de leur stockage peut également se poser. Dans ce cadre, grâce à l'utilisation combinée des *packages* `RMySQL` et `RGtk2`, qui permettent de relier R à MySQL et à Gtk+, nous proposons une méthode pour concevoir une interface graphique pour des algorithmes d'analyse statistique lorsque les données ne sont pas situées sur la machine de l'utilisateur. Dans une première partie, nous décrirons le système de gestion de base de données (SGBD) MySQL ainsi que le *package* `RMySQL` qui lui est associé. Nous nous intéresserons ensuite à la librairie GTK+, relative à la création d'interface graphique et au *package* `RGtk2` qui permet d'y accéder sous R. Nous terminerons par un exemple d'analyse de données qui tire profit de ces deux logiciels.

1 Accéder aux données

MySQL est un SGBD utilisable gratuitement hors d'un contexte commercial. Il possède de nombreux concurrents dont PostgreSQL et Ingres, sous licence libre. Il permet de stocker des données sur un serveur tout en les rendant disponibles via Internet. L'accès à un serveur MySQL est sécurisé de telle sorte que l'administrateur peut définir, pour chaque utilisateur, les éléments de la base de données sur lesquels il a la permission de travailler. Pour se connecter au serveur MySQL il est nécessaire d'avoir installé un client sur son ordinateur. La principale manière d'interagir avec MySQL réside en l'utilisation de scripts au format SQL (Structured Query Language). L'exécution de ces scripts, aussi appelées requêtes, peut se faire à l'aide d'une interface textuelle, disponible après exécution dans un terminal de la commande `mysql`. Le *package* `RMySQL` offre l'opportunité d'envoyer des scripts SQL au serveur MySQL, et d'en récupérer les résultats, avec R. Ceci est réalisé avec la fonction `dbSendQuery()`, sur laquelle nous nous focaliserons. Nous nous en servirons d'une part pour montrer comment il est possible de récupérer des données depuis le serveur MySQL afin de les traiter, et d'autre part pour expliquer comment faire pour enregistrer les résultats de ces analyses sur le serveur MySQL. Cette dernière fonctionnalité est avantageuse dans le cadre de travaux en équipe, par exemple.

2 Des méthodes d'analyse dans une interface graphique

GTK+ est un ensemble de librairies sous licence libre permettant la création d'interfaces graphiques. L'environnement de bureau GNOME et le logiciel de traitement d'images GIMP sont des exemples de logiciels basés sur GTK+. Les objets graphiques disponibles avec ce dernier sont très variés et comprennent entre autres des boutons, des barres de progressions, des zones d'affichage et des zones de texte. Ces objets peuvent être disposés dans une fenêtre selon la volonté du créateur de l'interface grâce à un logiciel approprié, comme par exemple Glade. Le fichier XML généré par ce logiciel peut alors être lu depuis R grâce au *package* `RGtk2` et à la fonction `gtkBuilder()`. Le comportement des objets graphiques doit alors être géré par des programmes en R grâce à la fonction `gSignalConnect()`. `RGtk2` permet également d'afficher des graphiques R dans des zones d'affichage de l'interface graphique en utilisant le *package* `cairoDevice` et la fonction `asCairoDevice()`.

3 Application à des mesures biologiques

MySQL et GTK+ sont tous deux utilisés dans l'étude de mesures, au cours du temps, de distances entre les deux parties de coquilles d'huîtres. Ces animaux sont analysés dans la baie d'Arcachon en France et sur plusieurs autres sites comme Santander en Espagne, Locmariaquer en France ou encore Tromsø en Norvège, par le laboratoire EPOC (Environnements et Paléoenvironnements Océaniques et Continentaux) UMR CNRS 5805. Les données sont acquises à une fréquence de 0.625 Hz pour chaque animal. Des liens entre la qualité de l'eau et ces signaux ont été mis en évidence (voir par exemple Tran et al. [4]). Afin de déterminer si une huître est en bonne santé, nous estimons la densité de probabilité f des distances entre les deux parties de sa coquille.
L'estimation est réalisée grâce à un estimateur à noyau $\hat{f}_{K,h}$ (voir Parzen [1]), défini quel que soit $t \in \mathbb{R}$ par :

$$\hat{f}_{K,h}(t) := \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right).$$

De tels estimateurs ont déjà été étudiés dans ce contexte (voir par exemple Sow et al. [3]). Ceux-ci requièrent le choix d'un noyau K et d'une fenêtre de lissage h. Pour K, nous prenons le noyau gaussien, défini par $K(t) := \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. Pour h, nous choisissons la fenêtre de lissage critique de Silverman [2] donnée par $h_{crit} := \min_{N(\hat{f}_{K,h})=N(f)}(h)$, où la fonction N associe son nombre de modes à une densité de probabilité. Ainsi, l'utilisation de h_{crit} nécessite une hypothèse concernant $N(f)$. Nous supposons donc que $N(f) = 2$. Sur des exemples d'huîtres mourantes ou saines, nous détaillerons les caractéristiques couramment observées de $\hat{f}_{K,h_{crit}}$.

References

[1] Parzen, E. (1962). On estimation of probability density function and mode, *The Annals of Mathematical Statistics*, **91**, 115–132.
[2] Silverman, B. W. (1986). Using kernel density estimates to investigate multimodality, *Journal of the Royal Statistical Society. Series B (Methodological)*, **43**(1), 97–99.
[3] Sow, M., Durrieu, G., Briollais, L. (2011). Water quality assessment by means of HFNI valvometry and high-frequency data modeling. *Environmental Monitoring and Assessment*, **182**, 155–170.
[4] Tran, D., Fournier, E., Durrieu, G., Massabuau, J.-C. (2003), Copper detection in the Asiatic clam Corbicula fluminea: Optimum valve closure response. *Aquatic Toxicology*, **65**, 317–327.