LA PART DU LOGICIEL R DANS L'ENSEIGNEMENT DE LA STATISTIQUE EN BIOLOGIE LE SITE WEB DE LYON

Anne-Béatrice DUFOUR[1]

TITLE

Teaching statistics in biology: the role of the R software – The website of Lyon

RÉSUMÉ

L'enseignement de la statistique en biologie à l'Université Lyon 1, dont le site pédagogique en est la vitrine, s'inscrit dans une relation tripartite. La face visible de cette relation est le logiciel R articulant méthodes statistiques et données biologiques. De la présentation algébrique à la visualisation des données et des méthodes, il autorise un apprentissage auprès d'étudiants possédant des bagages mathématiques variés.

Mots-clés : *biologie, biométrie, biostatistique, analyse de données, logiciel R.*

ABSTRACT

At Université Lyon I, introductory statistics courses for biologists emphasize the connection between statistical methods and biological data. Online course material was developed for this purpose based on the statistical software R. Through this material, sophisticated modeling tools and data visualization techniques can be made accessible to students with widely varying mathematical background.

Keywords: *biology, biometry, biostatistics, data analysis, R software.*

1 Introduction

L'enseignement de la statistique en biologie à l'université Lyon 1 s'inscrit dans l'histoire du Laboratoire de Biométrie fondé par J.-M. Legay qui chercha à rassembler mathématiciens et biologistes. Pour ne s'attacher ici qu'à la statistique, l'objectif était et reste de susciter le dialogue entre ceux qui en ont besoin sans en avoir forcément la maîtrise et ceux qui la conçoivent et qui n'en ont pas l'usage.

L'expérience acquise lors des enseignements et/ou des consultations statistiques est rassemblée sur le site web [1]. Ce dernier s'adresse aussi bien aux étudiants en formation initiale qu'aux chercheurs. Créé par D. Chessel, A.-B. Dufour et J.-R. Lobry, le site est aujourd'hui alimenté par des membres du Laboratoire de Biométrie et Biologie Evolutive. Il est, par choix, disponible à l'ensemble de la communauté scientifique de langue française.

L'enseignement de la statistique en biologie et la démarche biométrique sont indissociables. C'est pourquoi l'enseignement s'articule autour de trois axes : les données, les méthodes et les procédures informatiques, ceci quel que soit le public auquel les enseignants s'adressent. Le logiciel R, libre et gratuit, s'est imposé rapidement [2]. Ainsi, documents et

[1] Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive, anne-beatrice.dufour@univ-lyon1.fr

logiciel s'associent dans un même principe : celui de la communauté d'échange. L'analyse en composantes principales a été retenue pour illustrer notre point de vue.

2 Contexte historique

L'enseignement de statistique en biologie et par conséquent le site qui lui est dédié [1] s'inscrivent dans l'histoire même du Laboratoire de Biométrie et Biologie Evolutive (LBBE) créé par J.-M. Legay au début des années 1960. Ce dernier réunit des mathématiciens et des biologistes. Comme le souligne D. Chessel en 1992 [3], chacun arrivait au laboratoire avec ses connaissances sur la seule base qu'il faudrait générer un dialogue avec un autre dont le langage était inconnu. La biométrie n'est pas une discipline nouvelle mais une « méthodologie nouvelle à la disposition des sciences expérimentales, c'est-à-dire dans les sciences expérimentales ». Et c'est en cela que la biométrie se distingue de la statistique qui ne se préoccupe pas des objets qu'elle manipule (J.-M. Legay, 1976 [4]).

Le site pédagogique a été fondé par trois enseignants-chercheurs ayant réalisé leur thèse au LBBE, dans cette démarche et cette volonté de lier données et méthodes : Daniel Chessel, Anne-Béatrice Dufour et Jean Lobry. Il est enrichi par tous les membres de l'équipe pédagogique, enseignants-chercheurs et chercheurs enseignant, quelle que soit leur discipline d'origine. Il s'adresse en premier lieu aux étudiants de biologie de l'Université Lyon 1. Les documents sont volontairement accessibles à l'ensemble de la communauté scientifique : étudiants, enseignants et chercheurs. Le site, s'il est porté par quelques-uns, est la propriété de tous.

Le site d'enseignement de la statistique en biologie se décompose en cinq menus principaux : cours (env. 85 fiches), annales, fiches de travaux dirigés (env. 120), consultations (env. 116) et données (env. 370). Chaque menu contient des sous-menus permettant d'affiner la recherche selon le niveau universitaire et/ou le niveau de compétence acquis en statistique. Il contient également des documents rédigés lors de consultations statistiques permettant ainsi de valoriser certaines réflexions portées par la recherche au sens strict ou portées par la recherche liée à l'enseignement lui-même. Tout document s'appuie sur les données issues de la recherche, mises à disposition de tous, avec l'accord des chercheurs les ayant récoltées. C'est une des richesses reconnues par les utilisateurs à laquelle nous sommes très attachés.

3 Entre méthodes et données, le logiciel R

Le contexte historique décrit ci-dessus explique la démarche intégrative adoptée entre les données et les méthodes statistiques. L'enseignement de statistique assuré dès la première année de biologie, ceci à l'Université Lyon 1, est un enseignement général en licence et devient plus thématique en master. Il comprend les bases de la statistique descriptive univariée et bivariée, de la statistique inférentielle (échantillonnage, estimations et tests). Mais le site d'enseignement de la statistique en biologie possède une marque de fabrique initiée par ses fondateurs : l'analyse exploratoire des données. Il contient les méthodes classiques telles que l'analyse en composantes principales, les analyses des correspondances simple et multiple, les méthodes de couplage et les principales méthodes K-tableaux.

Les méthodes exploratoires les plus complexes ont été développées à partir de problématiques issues de l'écologie et de la biométrie humaine. Ces disciplines constituent le

socle de nombreux jeux de données. D'autres contributions ont été portées par la génomique, la génétique des populations, la biodiversité, la criminalistique, etc. Les méthodes résultent du dialogue entre le statisticien et le biologiste. Plus que les formules, les graphiques permettent les échanges. Cette idée a été défendue par D. Chessel dès 1993 avec la création du logiciel ADE (Analyse des Données Environnementales) : un outil orienté-objet d'aide à l'analyse et la représentation graphique des données environnementales. ADE a évolué jusqu'à sa forme actuelle, une librairie du logiciel R [2] nommée `ade4` ([5], [6], [7]). Elle contient des fonctions d'**A**nalyse de **D**onnées destinées d'abord à la manipulation des données **E**cologiques et **E**nvironnementales avec des procédures **E**xploratoires d'essence **E**uclidienne.

En enseignement comme en recherche, le logiciel R a été retenu car c'est un langage orienté-objet. Il est libre, gratuit et multi plate-forme. Il peut donc être installé dans toutes les salles informatiques du campus de l'université et les étudiants comme les chercheurs peuvent le télécharger sur leurs propres ordinateurs. Il contient l'ensemble des méthodes qu'un étudiant de licence en biologie doit acquérir lors de son cursus. Son langage vectoriel le rend proche de la formulation mathématique et permet ainsi d'assembler méthodes et procédures. La variance descriptive, par exemple, est la moyenne des carrés des écarts à la moyenne que l'on pourra écrire simplement `mean`((*x*-`mean`(*x*))^2) où *x* est le vecteur contenant les valeurs de la variable quantitative étudiée et la fonction `mean`, la moyenne.

Le logiciel R permet de visualiser les données tant dans la phase initiale de compréhension du tableau que dans la phase liée à l'analyse. J.-R. Lobry visite dans le document intitulé lang04.pdf les principales représentations graphiques que l'on peut obtenir avec les fonctions graphiques de haut niveau de R. Construire un graphique avec les étudiants, en séance, entraîne une discussion sur le sens de la donnée et le choix de la méthode. Le temps de calcul étant réduit, il n'est plus besoin de s'appuyer sur des jeux de données tronqués. Nous pouvons montrer la complexité réelle de l'analyse statistique à travers un contexte lié à la recherche ou à l'histoire. La figure 1 illustre les données : stature et longueur du majeur de 3000 criminels – `crimtab` – ayant servi à W. S. Gossett *alias* Student pour la construction de la théorie liée aux petits échantillons (Dufour et Lobry, 2008 [8]).

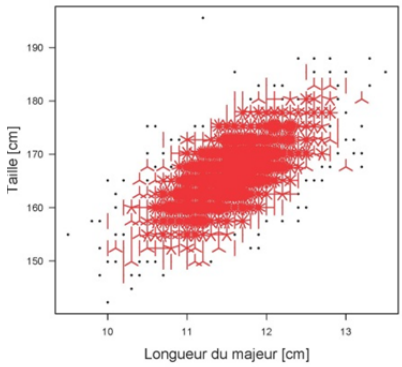

FIGURE 1 – *Un exemple de représentation graphique sous R : les données de Student (1908)*

L'enseignement de la statistique en biologie s'appuie sur la relation entre trois disciplines : la biologie au sens large du terme pour les données, la mathématique pour les méthodes et l'informatique pour les procédures. De plus, le public concerné par cet enseignement est de plus en plus hétérogène : des étudiants en formation initiale ou continue, aux chercheurs débutants ou confirmés quelle que soit leur origine, possédant ou non un bagage

mathématique. Le rôle de l'enseignant, tel que nous le percevons, est d'assurer l'apprentissage des méthodes statistiques selon l'un des deux points d'entrée : la donnée ou la méthode s'appuyant sur le logiciel R pour ouvrir le dialogue entre enseignant et enseigné.

4 Un exemple d'enseignement à travers la relation donnée, méthode et procédures

L'exemple proposé se veut un témoignage de la démarche adoptée depuis une dizaine d'années. Il est illustré à travers l'analyse en composantes principales. Celle-ci est enseignée au niveau du master (première ou seconde année) à des étudiants provenant des filières biologie des populations, écologie, activités physiques et sportives, mathématique et informatique du vivant. La forme privilégiée est le cours-TD intégré. Quel que soit le public, nous avons choisi de toujours présenter un cas concret à travers les données, le questionnement et la mise en œuvre assurée en direct avec R. Les graphiques sont au centre de la discussion, qu'ils soient simples pour s'approprier les variables et les individus du cas étudié, qu'ils soient complexes pour asseoir le vocabulaire lié à la méthode (valeurs propres, axes principaux, composantes principales, projection orthogonale, etc.).

L'exposé d'un cas concret, la mise en œuvre de représentations graphiques permettent aux étudiants de s'exprimer plus facilement. S'ils sont familiers du domaine, ils vont tenter des interprétations. S'ils sont plus proches des concepts mathématiques, ils vont découvrir le sens de la donnée. Trente stations ont été échantillonnées le long du Doubs et des prélèvements d'eau ont été effectués afin d'en étudier ses caractéristiques physico-chimiques [9]. La figure 2[2] représente la mesure du nitrate (en mg/l). La rivière est représentée en bleu avec l'indication de sa source pour assurer le repère géographique. La donnée a été centrée. La taille du carré représente l'écart à la moyenne et la couleur le sens de cet écart. En résumé, la présence de nitrate augmente avec l'activité humaine.

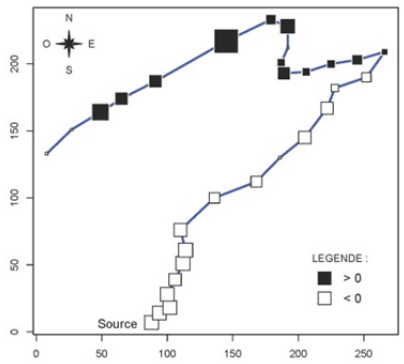

FIGURE 2 – *Présence de nitrate dans 30 stations réparties le long du Doubs – variable centrée*

La seconde étape consiste à faire expérimenter le sens de l'analyse en composantes principales à travers un jeu de données simple extrait de R (data(survey), [10]). Trois mesures (toutes exprimées en cm) ont été relevées chez 84 hommes et 84 femmes : la taille ou stature, les empans de la main dominante et non dominante, c'est-à-dire les distances entre le pouce et

[2] La programmation de la rose des vents en langage R provient d'un article de Tanimura *et al.* [11].

l'auriculaire, poignet posé à plat, les doigts écartés au maximum. Les données sont visualisées en dimension 3 à l'aide de la librairie `rgl` [12]. Le nuage de points (figure 3) peut être tourné dans tous les sens. Les étudiants recherchent ainsi la position donnant la plus grande dispersion, c'est-à-dire l'axe majeur (tdr601.pdf).

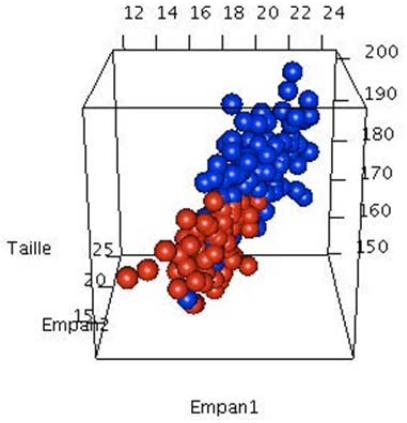

FIGURE 3 – *Représentation graphique en dimension 3 – librairie `rgl`*

Les notions d'espace, de sous-espaces, de critère à maximiser, de projection orthogonale peuvent être énoncées. L'analyse en composantes principales (ACP) peut alors être conceptualisée à l'aide des éléments de l'algèbre linéaire (add1.pdf, bs8.pdf), expérimentée grâce au logiciel R et son approche vectorielle et de quelques fonctions : `%*%` (produit de matrices), `eigen` (valeurs et vecteurs propres), `solve` (inversion de matrice), etc. Enfin, l'ACP peut être mise en œuvre et résolue – calculs et représentations graphiques – à l'aide des fonctions de la librairie `ade4`. Dans tous les cas, aux exercices type papier-crayon d'autrefois peuvent se substituer des petits exemples où chaque étape est énoncée (tdr80.pdf, tdr601.pdf). Enfin, de nombreux exemples d'analyse en composantes principales sont présentés dans la fiche tdr61.pdf.

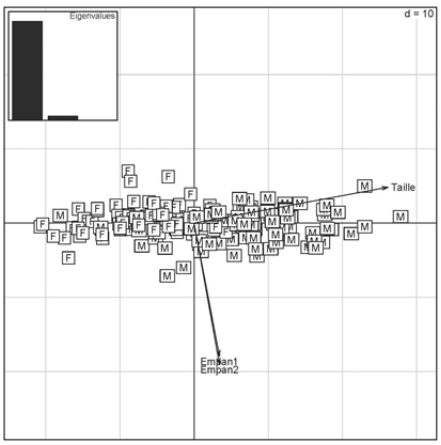

FIGURE 4 – *Représentation simultanée des individus et des variables sur le premier plan factoriel de l'ACP centrée - librairie `ade4`*

Pour achever l'exemple présenté (figure 3), la solution optimum est apportée par l'analyse en composantes principales sur les données centrées. Toute la variance est portée par le premier axe (95,93%) représentant l'effet-taille lié à la proportionnalité des mesures de taille d'un être humain. Un individu de grande stature aura globalement de plus grands empans qu'un individu de plus petite stature. Les deux empans, fortement corrélés entre eux,

définissent le deuxième axe (3,94%). La superposition du sexe (F pour Femelle et M pour Mâle), comme variable illustrative, souligne que, pour l'espèce humaine, l'homme est généralement plus grand que la femme.

Cette illustration ne serait pas complète sans discussion de la représentation des résultats sur le premier plan factoriel. En effet, le pourcentage de variance expliquée montre que seul le premier axe a un sens. Il ne faut donc pas céder aux habitudes. Mais peut-on ne représenter qu'un seul axe ? Cette question, récurrente lors de consultations statistiques, a fait l'objet d'une réponse détaillée par D. Chessel en 2004 (qr4.pdf). Elle est illustrée dans la figure 5. La relation entre une mesure (axe vertical) et le premier facteur (axe horizontal) est représentée par un nuage de points. La droite est le résultat de la régression linéaire simple de la mesure en fonction des scores des individus sur le premier axe. Le sens de la pente indique ainsi le sens de la corrélation, ici positive (effet-taille). Plus les points sont proches de la droite, plus la mesure intervient dans l'interprétation du premier facteur.

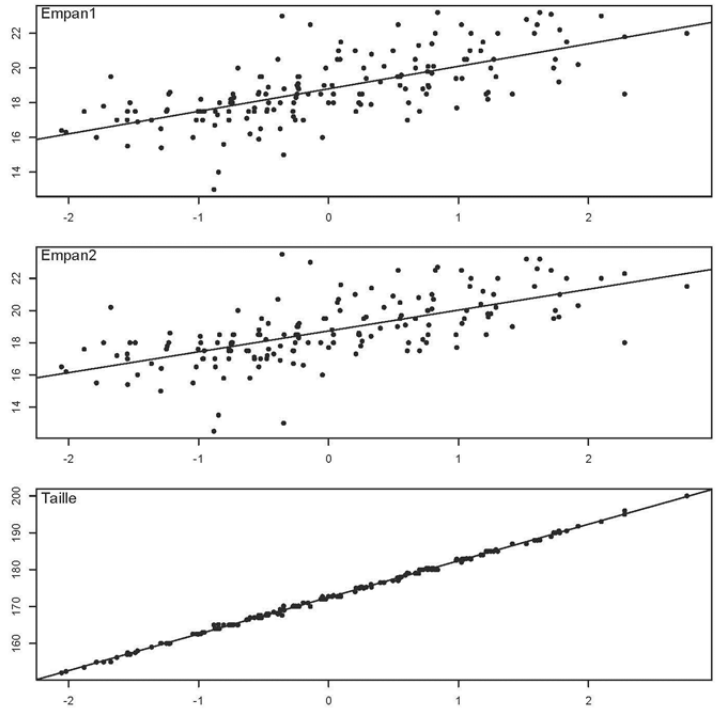

FIGURE 5 – *Représentation simultanée des individus et des variables sur le premier axe de l'ACP centrée - librairie* `ade4`

5 Conclusion

L'enseignement de la statistique en biologie, concrétisé par les documents du site WEB [1], est lié à la démarche biométrique et tente de rassembler données, méthodes et procédures informatiques. Il n'est aisé ni pour l'enseignant, ni pour l'enseigné, qui se retrouvent au croisement de trois disciplines. Mais il en est d'une richesse extrême car il permet, selon l'origine des étudiants, de naviguer d'un bord à l'autre et de créer un dialogue.

Celui-ci s'appuie principalement sur la visualisation des données, des méthodes et ne pourrait se faire sans le logiciel R, sa gratuité et son accès au code source. La rapidité de l'exécution des fonctions permet de dynamiser l'enseignement dans la présentation de

problèmes biologiques issus de la recherche. Cette vision, quelque peu idyllique, doit s'accompagner cependant d'une mise en garde sur l'utilisation de fonctions associées à des méthodes non maîtrisées.

En guise de conclusion, plus personnelle, le contenu du site WEB de Lyon d'enseignement de la statistique en biologie se transforme et évolue grâce au dynamisme de certains enseignants-chercheurs et chercheurs du Laboratoire de Biométrie et Biologie Evolutive, aux questions et consultations statistiques suscitées par les étudiants et les chercheurs, de Lyon et d'ailleurs, de biologie ou autres disciplines, tous à la recherche d'une statistique impliquée. Qu'ils en soient ici remerciés.

Références

[1] http://pbil.univ-lyon1.fr/R/enseignement.html, Enseignements de statistique en biologie.

[2] R Development Core Team (2010), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

[3] Chessel, D. (1992), *Echanges interdisciplinaires en analyse des données écologiques*, Mémoire d'Habilitation à diriger des recherches, Université Claude Bernard - Lyon 1.

[4] Legay, J.-M. (1976), Pour une Biométrie, *Statistique et Analyse des Données*, **1**, 5-11.

[5] Chessel, D., A.-B. Dufour, and J. Thioulouse (2004), The ade4 package - I: One-table methods, *R-News*, **4**, 5-10.

[6] Dray, S., A.-B. Dufour, and D. Chessel (2007), The ade4 package - II: Two-table and K-table methods, *R-News*, **7**, 47-52.

[7] Dray, S. and A.-B. Dufour (2007), The ade4 package: implementing the duality diagram for ecologists, *Journal of Statistical Software*, **22**(4), 1-20.

[8] Dufour, A.-B. and J.-R. Lobry (2008), From W. S. Gosset's pieces of card samples (1908) to the R software: analyzing Macdonell dataset on 3000 criminals, *Anthropologie et Biométrie Humaine*, **26**, 33-39.

[9] Verneaux, J. (1973), *Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie*, Thèse d'état, Besançon.

[10] Venables, W. N. and B. D. Ripley (1999), *Modern Applied Statistics with S-Plus*, Springer, New-York.

[11] Tanimura, S., C. Kuroiwa, and T. Mizota (2007), Auxiliary cartographic functions in R: North arrow, scale bar, and label with a leader arrow, *Journal of Statistical Software*, **19** (Code Snippet 1), 1-20.

[12] Adler, D. and D. Murdoch (2010), *rgl: 3D visualization device system (OpenGL)*, R package version 0.91.