R, RStudio et l'enseignement de la quantification

Enseigner R en SHS — Séminaire RUSS

Anton Perdoncin

ENS — Département de Sciences Sociales

5 avril 2018 – EHESS

Section 1

Introduction

Contexte

Un point de vue d'enseignant-chercheur en sciences sociales :

- **Formation aux méthodes quantitatives** à l'ENS Lyon lorsque j'étais étudiant (Modalisa).
- Mais **auto-formation à R et RStudio** pour l'enseigner...
- ... grâce à une **pratique soutenue** dans le cadre de mes activités de recherche.

Contexte

Un **éco-système particulier**

- Etudiants du département de Sciences Sociales de l'ENS, et des master SocStat et PDI (ENS–EHESS)...
- ... motivés et bien équipés de micro-ordinateurs récents;
- Petits effectifs ;
- Volumes horaires confortables ;
- Partage des tâches et charges d'enseignement du quanti avec d'autres collègues (J. Deauvieau, F. Maillochon, C. Zalc).

Une **réflexion en cours sur l'enseignement du quanti**, élaborée avec mes collègues et amis Pierre Blavier et Samuel Coavoux.

Contexte

Accueillir des étudiants peu, ou plus fréquemment, **pas formés en quanti.**

Rassurer les étudiants... si si... le quanti peut être intéressant... et coder ce n'est pas (forcément) traumatisant !

Faire sauter la distinction et la division du travail pédagogique entre stats-maths, pratique et réflexion sur la quantification, et logiciel.

Un enjeu **pédagogique et intellectuel** majeur pour le développement de nos disciplines.

Objectifs

- **Pourquoi** enseigner R et RStudio ? Et **pour quoi faire** ?
- Dans quel **cadre pédagogique** et avec quels **objectifs** ?
- Aperçu de **contenus** d'enseignement.

Section 2

Enseigner la quantification par la pratique

Subsection 1

Trois contextes d'enseignement

CPES 2e année (Paris Sciences & Lettres)

Pratique de la recherche :

- 60h de CM et 15h de TD sur l'année ;
- Licence 2, parcours histoire et sociologie ;
- Initier les étudiant.e.s à la recherche en sciences sociales ;
- Construire et exploiter collectivement des données (prosopographie des ingénieurs des mines) ;
- Fournir les bases pour une critique des chiffres, des nomenclatures et de leurs usages scientifiques et politiques ;
- Développer des compétences techniques : production, lecture et interprétation de sorties statistiques.

Master Pratique de l'Interdisciplinarité (ENS – EHESS)

Quantifier en sciences sociales (1, 2 et 3) :

- Trois modules de 24h, divisés en CM et en TD ;
- Un master pluridisciplinaire à coloration ethnographique ;
- Convaincre les réticent.e.s et nourrir les mordu.e.s ;
- Enjeux pratiques et épistémologiques de la quantification ;
- QSS 1 (M1) : construction des données, statistique descriptive, inférence, initiation à R ;
- QSS 2 (M1) : analyse des données et modélisation : fondements, enjeux et pratique ;
- QSS 3 (M2) : approfondissements.

Master Sociologie et Statistique (SocStat, ENS – EHESS)

Introduction à R (M1)

- Cours de 24h consacré exclusivement à R (en plus d'un cours équivalent SAS/Excel) ;
- Master orienté quanti ;
- Des étudiants qui, pour la plupart, n'ont jamais fait de programmation, ni eu de cours de quanti ;
- Enjeux premier de maîtrise technique du logiciel ;
- Présentation d'une large palette d'outils : statistique descriptive, analyse géométrique des données, modélisation, analyse longitudinale, réseaux, etc.

Subsection 2

Quelques principes généraux

Les trois piliers de l'enseignement du quanti

Il s'agit de tenir ensemble, dans un même cours, trois dimensions :

- réflexions sur le **raisonnement quantitatif** ;
- formation aux **techniques statistiques** sans formalisation excessive ;
- formation pratique à un **logiciel**.

A l'exception de SocStat : master spécialisé où les logiciels sont enseignés à part.

Objectifs pédagogiques

L'objectif principal (modulé selon les niveaux) : développer l'**autonomie des étudiants face aux données quantitatives et aux chiffres**, et leurs **compétences techniques**.

Déconstruire et maîtriser l'**ensemble de la chaîne de production des chiffres** : de la construction des données à leurs usages, en passant par le codage et le traitement.

Confronter les étudiants le plus **rapidement** et le plus **simplement** possible à l'**écriture de scripts**, donc privilégier les **logiciels à langage de programmation explicite**.

Pourquoi R et RStudio ?

Dissocier **saisie** (Calc... ou tout autre avatar non-libre) et **manipulation/traitement** des données (R, SAS, Stata...).

R : un couteau-suisse, libre, (plus ou moins) user-friendly, multi-plateformes, qui ne nécessite pas forcément de salle info.

RStudio : aide à l'écriture du code, interface graphique sympathique, fonctionnalités d'import/export, intégration d'outils pour usage avancé (RMarkdown, Git).

Section 3

Comment enseigner R et RStudio ?

Subsection 1

Généralités

Format des cours

Je privilégie le **recopiage du code par les étudiants**, à partir d'une présentation (réalisée sous RMarkdown), **sauf pour les méthodes plus complexes** nécessitant d'écrire beaucoup de code (typiquement l'AGD : distribution et commentaire en cours de scripts "clé en main").

Une séance = une partie **cours** + une partie **pratique**.

Des **exercices d'application** à faire à la maison.

Installer R, RStudio et les packages

Une **première étape**... qui est souvent **un petit challenge**...

Il a fallu apprendre progressivement à **déminer les obstacles les plus fréquents** : OS ancien ou non mis à jour, erreurs de téléchargement, accents dans le nom du répertoire racine, etc.

Quand les soi-disant "digital natives" sont des "**digital captives**" : dès cette première étape, il s'agit de faire sortir les étudiants de la **réconfortante léthargie** de l'utilisateur conditionné à un usage captif des applis.

Organiser son espace de travail : un préalable

Bien maîtriser le logiciel suppose de **comprendre comment sont rangés et stockés des documents sur un disque dur**... ce qui est tout sauf évident !

Les **projets RStudio** sont ici très utiles : un cours = un projet.

A la racine du projet, distinction entre **Editeurs**, **Sorties**, **Data** (les bases de l'exigence de réplicabilité...).

Quels packages ?

- `questionr` : la bibliothèque de fonctions la plus utile pour les opérations ordinaires de traitement quantitatif en sciences sociales ;
- `tidyverse` (notamment `dplyr`, `tidyr`, `stringr`, `ggplot2`) : une conversion (personnelle) récente... mais sans retour en arrière possible ;
- `foreign` ou `haven` : pour certaines fonction d'import/export.

Quels packages ?

- `FactoMineR`, `explor`, `factoextra`, `cluster` : analyse géométrique des données ;
- `lmtest` : modélisation ;
- `TraMineR` : analyse de séquences par appariement optimal ;
- `tm`, `wordcloud`, `topicmodels` : lexicométrie ;
- `survival` : analyse démographique des biographies (modèles de durée / event history analysis).
- etc.

Subsection 2

Objectifs pédagogiques

Objectifs de niveau 1

Premier semestre L2 CPES, QSS 1 (PDI), premières séances SocStat

- Initiation à la **syntaxe** : objets, fonctions, vecteurs ;
- **Comprendre** et **résoudre** les erreurs les plus usuelles ;
- **Importer** des données ;
- **Inspecter** et **décrire** la **structure** d'un jeu de données ;
- **Décrire** des **variables** numériques et catégorielles ;
- **Découper** une variable numérique, **renommer** et **regrouper** des modalités ;
- **Croiser** deux variables et **tester** la significativité d'une liaison statistique (chi2).

Objectifs de niveau 2

Deuxième semestre L2 CPES, QSS 2 (PDI), suite des séances SocStat ;

- **Manipuler** des **chaînes de caractère** : rechercher, remplacer et substituer des pattern (`stringr`) ;
- **Mettre en forme** graphiques (rudiments de `ggplot2`) et tableaux ;
- Manipuler et **transformer** des **jeux de données** : sous-populations, fusions de données ;
- Explicitation de ce que la **réplicabilité** signifie en sciences sociales.

Objectifs de niveau 3

QSS2 et QSS3 (PDI), fin des séances SocStat

- **Recoder** et **sélectionner** des variables afin de réaliser une analyse statistique complexe ;
- **Réaliser** et **commenter** les sorties graphiques et statistiques d'une **AGD** ;
- **Réaliser** et **commenter** les sorties statistiques de modèles de **régression** ;
- . . . en fonction des **années** et des **envies** des étudiants : analyse longitudinale, réseaux, etc. ;
- Présentation de **RMarkdown**.

Section 4

Conclusion

Synthèse

Enseigner R et RStudio s'inscrit dans un **enseignement général des méthodes quantitatives et de la quantification** en sciences sociales :

- Tenir ensemble formation aux statistiques, acquisition d'une attitude réflexive et informée à l'égard des chiffres et de leurs usages scientifiques et politiques, et production de résultats quantitatifs par le traitement de données ;
- R et RStudio : la meilleure des solutions (actuellement disponible) pour permettre aux étudiants de progresser de façon autonome.

Une affaiR de temps

Les **objectifs de premier niveau sont les mêmes**, quels que soient les étudiants et les formations : ce qui change est **le temps que l'on y consacre**.

L'apprentissage des techniques avancées ne peut être réalisée sans un **cours de méthodes quantitatives préalable** : cela n'a pas de sens d'apprendre à se servir de FactoMineR... si l'on ne sait pas ce qu'est une ACM ! Ou de manipuler `chisq.test` sans savoir ce qu'est un test d'inférence et quelles sont ses limites !

Les étudiants doivent accepter de **consacrer (un peu) de temps à l'apprentissage d'un nouveau langage** et de nouvelles habitudes intellectuelles.

Dédramatiser les stats... par l'apprentissage du code !

Un gros enjeu pour des étudiants de licence et pour des mastérant à sensibilité ethnographique : **dédramatiser** et montrer que les stats peuvent être non seulement **utiles**... mais aussi **amusantes** !

Entrer pas à pas, sans se presser, dans un **langage de programmation** est un des moyens de cette dédramatisation :

- totalement contre-intuitif... les étudiants eux-mêmes sont supris !
- ils comprennent vite l'avantage comparatif à maîtriser ce type d'outils...

Quelques ressources utiles

- *Introduction à R et au Tidyverse* de Julien Barnier : https://juba.github.io/tidyverse/
- Le site du projet Analyse-R : http://larmarange.github.io/analyse-R/
- Diverses pages sur le site STHDA : http://www.sthda.com/french
- La liste R-Soc : https://groupes.renater.fr/sympa/info/r-soc : inciter les étudiants à oser y poser leurs questions !
- Le groupe slack Grrr : https://frama.link/r-grrr (pour rejoindre le groupe)
- R Stackoverflow (en anglais. . . ce qui peut rebuter certains étudiants) : https://stackoverflow.com/questions/tagged/r