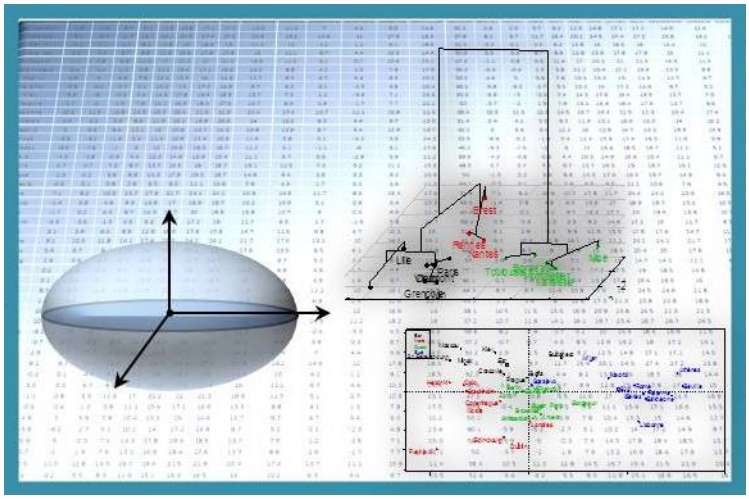
Note de consultation rédigée par Nathalie VILLA-VIALANEIX[1]

ANALYSE DES DONNÉES MULTIDIMENSIONNELLES

François HUSSON, Jérôme PAGÈS et Magalie HOUÉE-BIGOT

MOOC publié sur France Université Numérique (FUN)
Numéro 40001 – 2 mars au 6 avril 2015



Avant-propos

Cette note de consultation est un peu particulière dans le sens où elle fait le compte-rendu d'un support de cours non imprimé, un MOOC[2]. Ce type de moyen de formation se développant, il semble utile de faire une recension des expériences pédagogiques innovantes et réussies d'enseignement de la statistique au travers de ce type de média afin de les faire connaître et de permettre au plus grand nombre de profiter d'une session ultérieure du cours[3] : en particulier, celles-ci peuvent servir à des étudiants souhaitant compléter leur compréhension d'un sujet, à des professionnels souhaitant approfondir leurs connaissances dans ce domaine ou des enseignants souhaitant s'appuyer sur un support de cours complet pour alimenter leurs propres enseignements.

Cette note est organisée en deux parties : la première, classique, décrit le cours, son contenu et contient des retours d'expérience directs de professionnels non statisticiens ayant suivi le cours en auto-formation pour leurs besoins professionnels. La seconde partie présente

[1] Chargée de Recherche, INRA de Toulouse, UR 0875 MIAT, nathalie.villa@toulouse.inra.fr

[2] MOOC : Massive Online Open Courses (Cours en Ligne Ouverts et Massifs), voir le volume 5(1) de la revue *Statistique et Enseignement*, ou bien Villa-Vialaneix, N. (2013), J'ai testé pour vous… un MOOC, *Statistique et Enseignement*, **4**(2), 3-17.

[3] En particulier, une prochaine session de ce MOOC est programmée en mars 2016 : https://www.france-universite-numerique-mooc.fr/courses/agrocampusouest/40001S02/session02/about

une analyse plus macroscopique de la satisfaction des apprenants inscrits au cours, analyse qui a été rendue possible par les réponses collectées lors d'un sondage réalisé à la fin du cours. Les données et l'analyse ont été fournies par François Husson, l'un des créateurs du MOOC présenté ici.

1re partie – Note de consultation du MOOC « Analyse des données multidimensionnelles »

Nathalie Villa-Vialaneix, à partir des retours d'expérience de Claire Hoede, Erika Sallet et Clément Delestre

Le cours « Analyse des données multidimensionnelles » a été proposé sur la plateforme FUN[4]. Il est décrit à la page :

https://www.fun-mooc.fr/courses/agrocampusouest/40001/Trimestre_1_2015/about

Ce cours a commencé le lundi 2 mars 2015 pour une durée de 5 semaines (il s'est donc achevé le 6 avril 2015). Il a été créé et animé par François Husson (l'unique enseignant présent sur les vidéos du cours), Jérôme Pagès et Magalie Houée-Bigot, et il était porté par l'école d'ingénieurs Agrocampus Ouest[5].

L'objectif affiché du cours était de présenter des méthodes permettant « *d'analyser, d'explorer, de visualiser des tableaux de données afin d'en extraire l'essentiel de l'information* ». De manière concrète, le cours a été divisé en 4 grandes parties, plus une partie de synthèse. Chacune de ces parties était traitée sur une semaine entière et correspondait à une méthode particulière qui était décrite de manière détaillée et illustrée sur des exemples concrets. Les méthodes présentées lors du cours étaient :

- Semaine 1 : l'Analyse en Composantes Principales (ACP) ;
- Semaine 2 : l'Analyse Factorielle des Correspondances (AFC) ;
- Semaine 3 : l'Analyse Factorielle des Correspondances Multiples (AFCM) ;
- Semaine 4 : la classification.

Chacune des 4 premières semaines est organisée de manière similaire : le support de cours est composé d'un diaporama global, qui est commenté au travers de courtes séquences vidéos d'une dizaine de minutes (3 séquences vidéos pour la semaine 1, 5 pour la semaine 2, 4 pour les semaines 3 et 4). Toutes les vidéos sont téléchargeables en divers niveaux de résolution et la transcription de l'audio est également disponible (ceci peut s'avérer très pratique pour les étrangers ne bénéficiant pas d'une connexion suffisante). Chaque vidéo de cours est suivie d'un court quizz récapitulatif de 5 questions. Chaque semaine, un didacticiel pour la mise en pratique des notions vues dans le cours conclut la semaine. La mise en pratique est effectuée par le logiciel libre R, avec le package **`FactoMineR`**[6] dont le principal instigateur du cours, François Husson, est le mainteneur. Elle est illustrée dans une vidéo qui analyse deux jeux de données par semaine : les jeux de données ainsi que le script R au format texte et PDF (produit par RMarkdown[7], ce fichier contient les commandes R, les

[4] FUN : France Université Numérique ; http://www.france-universite-numerique.fr/sciences.html.

[5] http://www.agrocampus-ouest.fr

[6] http://factominer.free.fr

[7] http://rmarkdown.rstudio.com : RMarkdown est un format de fichier qui permet la création de rapports avec R en combinant une syntaxe de mise en forme en simple texte très facile (langage Markdown) avec des morceaux de code R qui sont exécutés. Les résultats de l'exécution du code R sont inclus dans le document

sorties produites et de très brefs commentaires) est fourni aux apprenants. Chaque semaine se termine par deux exercices : un qui se présente sous la forme d'un quizz plus long que les quizz qui correspondent aux diverses séquences vidéos de la semaine et dont le but est de récapituler les notions principales abordées durant le cours. Un second, toujours sous la forme d'un quizz, demande à l'apprenant d'analyser avec la méthode décrite dans le cours de la semaine, un jeu de données réel. Des questions en relation avec les résultats de l'analyse et l'interprétation de ceux-ci sont posées. Enfin, les trois premières semaines se terminent par une séquence vidéo d'approfondissement : la première sur la gestion des données manquantes, la seconde sur une étude de cas d'application de l'AFC à l'analyse de données textuelles et la troisième sur l'imputation de données manquantes.

La dernière semaine est une semaine récapitulative qui commence par une vidéo décrivant la démarche générale en analyse de données multidimensionnelles et qui se solde par un exercice récapitulatif global posé sous la forme d'une étude de cas concret (à réaliser avec R) et d'un quizz permettant de vérifier les résultats obtenus et leur bonne interprétation.

Tous les exercices des différentes séances sont notés et il est possible de suivre sa progression sous la forme d'un score de réussite aux divers types d'exercices. Tous les exercices peuvent être refaits autant de fois que le souhaite l'apprenant et seule la dernière tentative est comptabilisée. Ainsi, contrairement au parti pris de beaucoup de cours en ligne de type MOOC, il n'y a pas de validation de type « scolaire » dans ce MOOC, avec un exercice à essai unique ou limité ou des devoirs à rendre à date fixe. Après quelques questions sur les jeux de données étudiés, l'interprétation des résultats est laissée à l'apprenant et un Wiki permet de proposer son interprétation. Les autres apprenants peuvent alors corriger et compléter l'analyse. L'interprétation des résultats est ainsi co-construite avec tous les apprenants. Enfin, un forum de discussion, organisé par thématiques et aspects logiciels, permettait les échanges plus directs entre apprenants : son organisation permettait de retrouver facilement les discussions sur chaque sujet.

De mon point de vue, l'organisation et la structuration du MOOC est remarquablement claire : les semaines contiennent la juste dose d'information et le cours ne commet pas l'erreur de présenter un trop grand nombre de méthodes au détriment de la compréhension de celles-ci par des apprenants qui n'en ont jamais entendu parler au préalable. Les aspects théoriques des méthodes ne sont pas passés sous silence mais la mise en œuvre pratique sur des jeux de données réels et parlants apportent systématiquement une illustration complémentaire claire des concepts développés. Les supports de cours sont tous d'une grande qualité et les quizz et exercices sont suffisamment nombreux et intelligemment pensés pour permettre l'acquisition des notions abordées. Il semble que le profil idéal d'un apprenant pour ce MOOC est celui d'une personne ayant déjà une formation de base en mathématiques (comme elle peut être enseignée dans la plupart des cursus universitaires de premier cycle en France). Comme la mise en œuvre pratique se fait sous R, il est également préférable que l'apprenant soit un minimum familiarisé avec ce langage de programmation : même si les méthodes peuvent être mises en œuvre via l'interface graphique RCommander, il paraît assez peu réaliste de pouvoir assimiler à la fois les aspects méthodologiques et la mise en œuvre pratique en seulement 5 semaines. Ces caractéristiques correspondent effectivement à la description du cours et des pré-requis qui est faite sur sa page de présentation.

Les trois participants qui ont accepté de faire un retour d'expérience sur le MOOC ont tous les trois un profil et des aspirations similaires : de formation scientifique en bio-

---

produit lors de la compilation.

informatique, ils ont suivi durant leur cursus scolaire quelques cours de statistique mais ceux-ci ne constituaient pas le sujet central de leurs formations d'origine. Actuellement insérés dans le secteur professionnel comme ingénieurs d'études ou de recherche en bio-informatique, leur motivation était de découvrir des outils d'analyse de données et particulièrement de savoir comment les mettre en œuvre. Pour tous, cela correspondait à un besoin non crucial mais utile dans leur vie professionnelle. Les trois ont noté que le MOOC avait correspondu à leurs attentes, qu'ils avaient réussi à le suivre jusqu'au bout et à compléter entièrement tous les exercices. Ils ont tous noté que le cours était très dense avec beaucoup d'informations à assimiler en peu de temps. Les principales difficultés didactiques notées sont relatives aux notions de géométrie utilisées pour présenter les analyses factorielles. Des suggestions ont aussi été apportées pour limiter les difficultés d'installation de R et des packages, ainsi que les problèmes dus aux différences de plateforme d'installation (en particulier, R est sensible à l'encodage des fichiers importés lorsque ceux-ci contiennent des caractères accentués : ce problème, rencontré sous Linux et Mac OS X, n'était pas abordé dans le premier cours mais a été corrigé dès la deuxième semaine).

En conclusion, globalement, les trois apprenants se sont déclarés très satisfaits d'avoir participé et disent tous avoir appris des choses utiles pour leur pratique professionnelle.

2e partie – Courte analyse du questionnaire de satisfaction final du MOOC « Analyse des données multidimensionnelles »

Cette partie présente une brève analyse des informations données par les participants sur leur profil et leur satisfaction vis-à-vis du MOOC, telles que collectées au travers du questionnaire final. 5 053 personnes se sont inscrites à la première session du MOOC « Analyse des données multidimensionnelles », provenant de 93 pays. Le nombre de participants actifs est, comme c'est habituel sur ce type de cours, très inférieur puisque 1 450 personnes ont participé au premier quizz et seulement 410 au dernier (certaines personnes ont suivi le cours sans faire les quizz, donc le nombre réel de participants est difficile à estimer). Enfin, 315 personnes ont accepté de donner leur avis sur le cours lors du questionnaire final de satisfaction. Ce sont ces réponses qui sont brièvement décrites dans cette partie.

Les apprenants ayant répondu au questionnaire semblent bien correspondre à la cible visée par le cours : la moyenne d'âge est de 36,4 ans (bien que sa distribution soit très asymétrique avec une forte proportion d'apprenants de moins de trente ans comme le montre la FIGURE *1*). Ceux-ci ont un niveau d'études avancé, master ou ingénieur (50,8%) ou doctorat (22,5%).

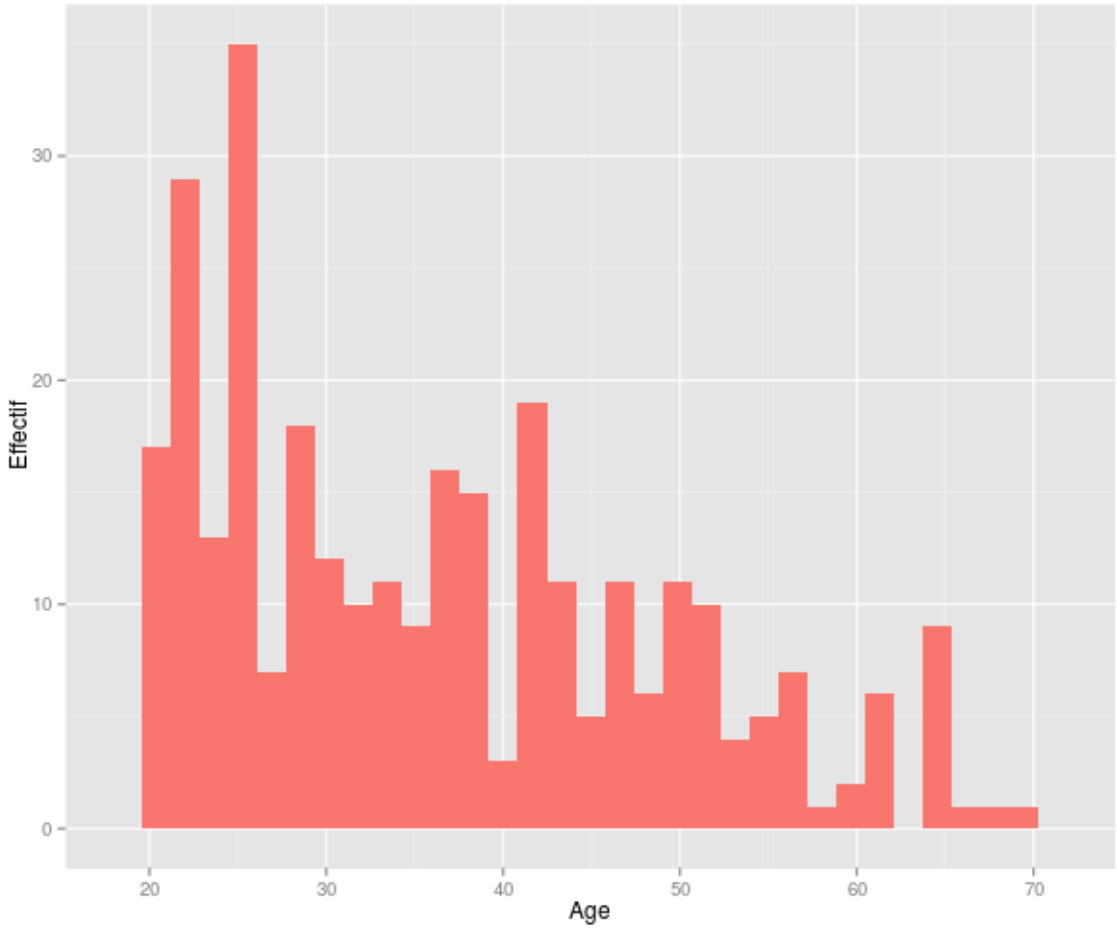


FIGURE 1 – *Distribution de l'âge des apprenants*

Pour les répondants, l'utilisation des différentes facettes du MOOC a été assez complète : 58,7% des apprenants ont utilisé le forum et 41,6% le wiki. Sur les 17 vidéos, les répondants en ont, en moyenne, regardé plus de 15 (avec près de 71% d'entre eux qui ont regardé les 17). Ils ont en moyenne fait 14 des 16 quizz (et près des trois-quarts ont fait les 16) et réalisé 5 des 8 exercices proposés (plus de 39% ont fait les 8, ce qui démontre une bonne participation des apprenants à l'ensemble des activités du cours). En termes de temps consacré au MOOC, ces activités correspondent en moyenne à presque 5 heures de travail hebdomadaire, ce qui correspond au descriptif du cours, avec de fortes disparités selon les sujets : l'ACP a requis environ 6,3 heures de travail alors que les autres thématiques ont requis entre 4,4 et 4,6 heures de travail par semaine.

Le niveau de satisfaction des apprenants a été évalué par les répondants sur une échelle allant de 0 (pas satisfait du tout) à 10 (parfaitement satisfait). La distribution de la satisfaction est donnée à la FIGURE *2*, ce qui correspond à une note moyenne d'environ 9,6/10. De manière plus précise, près de 50% des répondants ont jugé le niveau du cours « adapté » et 46% l'ont jugé « élevé mais j'ai pu suivre », ce qui illustre le bon dosage de la difficulté du cours : des notions jugées complexes ont été abordées par le cours sans que celles-ci découragent les apprenants. Du point de vue de l'équilibre du contenu et des choix pédagogiques, 45% des répondants l'ont trouvé bon et 46% très bon. L'utilisation de **`FactoMineR`** a été

majoritairement jugée très facile (13%), facile (37%) ou acceptable (35%). Seuls 6% l'ont trouvée difficile et moins de 1% trop difficile (les autres ne l'ont pas utilisé).

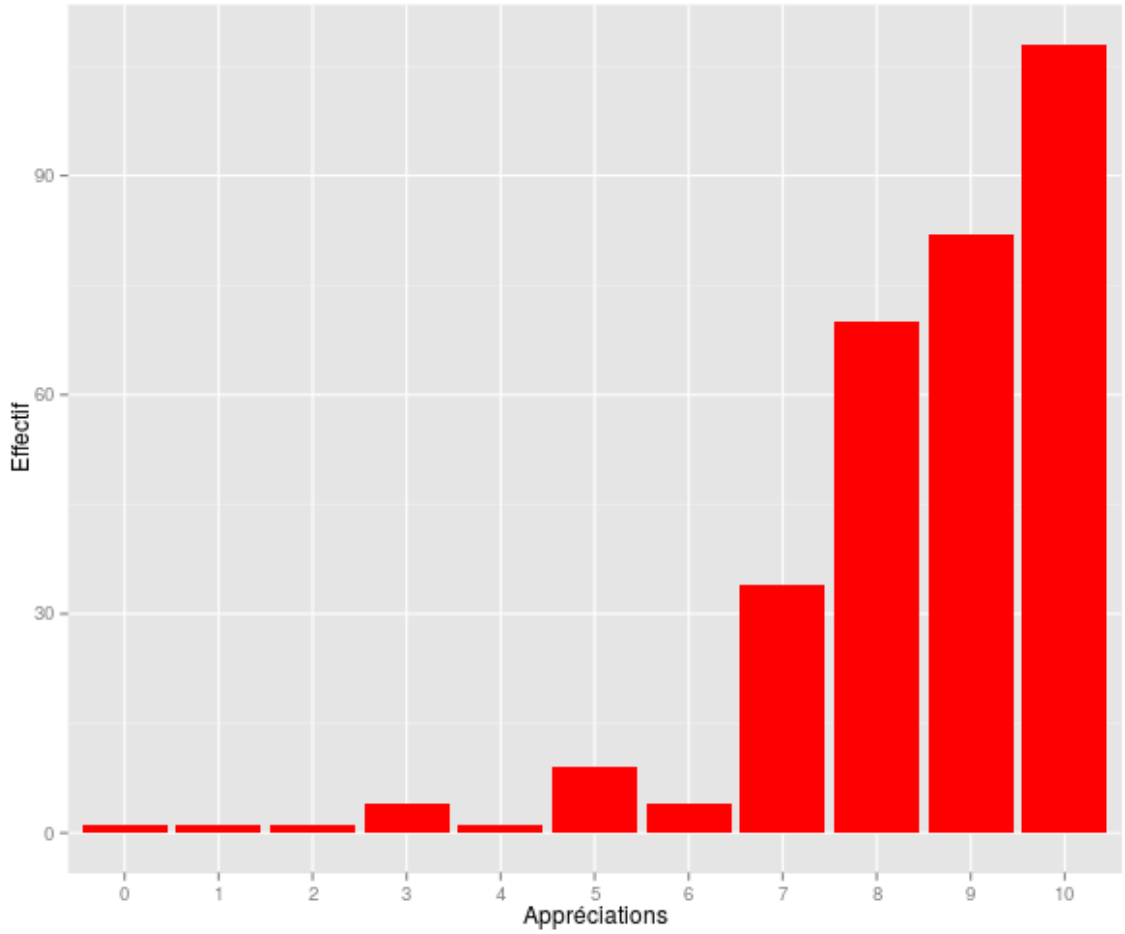


FIGURE 2 – *Distribution des notes d'appréciation données par les répondants (10 correspond à la satisfaction maximale)*

Notons encore que 96% des répondants seraient prêts à recommander ce MOOC à d'autres. Les commentaires libres sont pour la plupart très élogieux sur la structuration du MOOC, la complémentarité cours-exercices-logiciel, mais aussi sur les approches pédagogiques. Certains étudiants étrangers (en particulier africains) ont souligné le bénéfice de la mise à disposition de tels cours en ligne. Parmi les quelques personnes ayant répondu à la question « Si vous avez abandonné le MOOC (vous pensiez le faire en entier mais n'avez pas tout fait et ne comptez pas le finir), quelles en sont les raisons ? », la plupart soulignent le manque de temps et la difficulté à comprendre le formalisme mathématique du cours.

En conclusion, l'analyse du questionnaire de satisfaction est concordante avec les retours directs des utilisateurs. Pour l'apprenant cible, qui est un scientifique ayant des bases en mathématiques et statistique ainsi que dans le langage de programmation R, le cours est jugé difficile, mais il est apprécié et permet effectivement l'acquisition de connaissances et compétences pratiques dans le domaine de l'analyse de données multidimensionnelles. Le temps moyen de travail estimé pour l'apprenant type est d'environ 5 à 6 heures d'investissement par semaine.