Master I APE – MECI

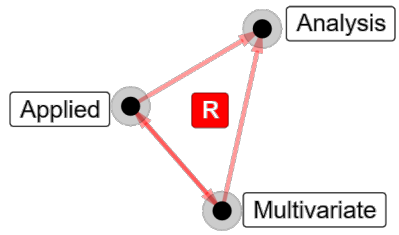
U.F.R. GHES

UNIVERSITÉ DE PARIS

ME3AY020 – AMA-R

SYLLABUS

ANALYSE MULTIVARIÉE APPLIQUÉE



Année universitaire

2020-2021

Thibaud DEGUILHEM

Maître de Conférences en Économie

Département d'Économie

LADYSS UMR 7533

www.tdeguilhem.com

thibaud.deguilhem@u-paris.fr

Présentation et objectifs pédagogiques

Destiné aux étudiant-e-s de première année du Master MECI (PISE et CCESE option Data), cet enseignement offre un large panorama sur l'analyse multivariée appliquée. Utilisant Rstudio, certains packages spécifiques et l'environnement `Markdown`, ce cours propose aux étudiant-e-s de devenir autonomes dans l'utilisation des techniques usuelles d'exploration, de prédiction et d'identification des relations causales en sciences sociales (socio-économie en particulier) en leur permettant de développer différentes compétences : (*i.*) connaître les outils d'analyse multivariée afin d'apporter des réponses quantitatives à différents types de questions du même ordre, (*ii.*) maîtriser les packages essentiels pour conduire une analyse multivariée à partir d'un jeu de données, (*iii.*) interpréter et mettre en valeur des résultats d'analyse multivariée à l'aide de visualisations adaptées et de tableaux pertinents, (*iv.*) travailler en équipe lors d'un challenge sur `Markdown` à partir de jeux de données spécifiques. Avoir suivi le cours d'ASA-R au premier semestre est un prérequis pour ce cours.

Organisation et déroulement

Trois parties structurent ce cours abordant les trois approches essentielles de l'analyse multivariée appliquée : (*i.*) l'exploration de jeux de données par la réduction des dimensions et les algorithmes non-supervisés de regroupement (*data mining*), (*ii.*) la prédiction par l'utilisation de modèles et d'algorithmes supervisés de classification ou l'extension du modèle linéaire (*predictive modelling*) et (*iii.*) l'identification de relations causales à partir de données non-expérimentales à travers les méthodes d'évaluation d'impact (*causal modelling*). Les étudiant-e-s retrouveront toutes les informations en ligne sur la page dédiée au cours : `ama-r-me3ay020.html`. A la fin du semestre, les étudiant-e-s participeront à un "challenge" en groupe sous l'environnement `Markdown`. A partir d'applications pédagogiques, les étudiant-e-s seront accompagné-e-s dans leur pratique de Rstudio tout au long du semestre par la mise à disposition de jeux de données, une communauté (AMA-R sur le forum DATALAB) dédiée aux questions de méthodes, de "syntaxe" (script), de "tronçons" (chunk) ou de "tricotage" (knitr/-compilation) (`https://helpstudents.tribe.so`) et leur apprentissage sera guidé par la réalisation de deux fiches d'exercices. Les étudiant-e-s peuvent enfin prendre rendez-vous durant les heures de permanence (mardi 9h-10h30).

Plan du cours

Introduction à l'analyse multivariée appliquée (AMA) avec Rstudio

(25-01) *Présentation du syllabus, questions d'AMA et multivariate data*

- ☆ Définitions, questions et stratégies multivariées, exemples du cours
  - ⋆ Lectures conseillées : (Denis, 2015) ⟶ Chap. 1 (introduction : 33-46) ; (Zelterman, 2015) ⟶ Chap. 1 (introduction 1–13)

Partie I **Approches exploratoires "non-supervisées" : explorer, résumer et regrouper**

(1-02) *Réduction des dimensions et analyse factorielle : Analyse en Composantes Principales (ACP)* ⟶ `US Arrests dataset (1973)`

- ☆ Étapes méthodologiques de l'ACP, optimisation et validation avec le package `factoextra`
  - ⋆ Lectures conseillées : (Husson et al., 2017) ⟶ Chap. 1 (1–44)
  - ⋆ Références complémentaires : (Denis, 2020) ⟶ Chap. 10

(8-02) *Clustering et méthodes de regroupement algorithmiques non-supervisées : proximité et K-means (KM)*
⟶ `US Arrests dataset (1973)`

☆ Similarités et distances, optimisation, nombre de clusters et validation

⋆ Lectures conseillées : (Husson et al., 2017) ⟶ Chap. 4 (169–204)

⋆ Références complémentaires : (Denis, 2020) ⟶ Chap. 12

(15-02) *Typologies et algorithmes hiérarchiques : Classification Ascendante Hiérarchique (CAH)*
⟶ `US Arrests dataset (1973)`

☆ Méthodes agglomératives et "dissimilarité", complémentarité CAH et ACP

⋆ Lectures conseillées : (Husson et al., 2017) ⟶ Chap. 4 (169–204)

⋆ Références complémentaires : (Denis, 2020) ⟶ Chap. 12

(8-03) Séance de Q&R à propos de la fiche d'exercices 1 (en ligne sur Zoom)

Partie II **Approches prédictives "supervisées" : classer et prédire l'assignation**

(15-03) *Tree Methods : Regression et Classification Trees*
⟶ `Ames Iowa Housing dataset`

☆ Modèles de mélange gaussien, algorithme EM et convergence, critère d'informations (AIC/BIC), qualité de la partition, package `mixtools`

⋆ Lectures conseillées : (Deguilhem and Seetahul, 2016)

⋆ Références complémentaires : (Stahl and Sallis, 2012)

(22-03) *Prédictions supervisées et modèles à variable dépendante qualitative*
⟶ `Affairs dataset`

☆ Maximum vraisemblance, modèles Probit et Logit, effets marginaux, qualité de l'ajustement et tests avec le package `margins`

⋆ Lectures conseillées : (Stock and Watson, 2014) ⟶ Chap. 7 (235–258)

⋆ Références complémentaires : (Denis, 2020) ⟶ Chap. 8

(29-03) *Modèles de médiation et de modération (SEM)*
⟶ `Mediation Hallquist dataset`

☆ Identifier un effet direct et indirect, modération et médiation avec le package `lavaan`

⋆ Lectures conseillées : (MacKinnon et al., 2007)

⋆ Références complémentaires : (Stock and Watson, 2014) ⟶ Chap. 9

Partie III **Approches causales : identifier et évaluer un impact**

(5-04) *Stratégie d'identification par l'appariement : Propensity Score Matching*
⟶ `STAR dataset`

☆ "Traitement" et "contrôle", spécification du modèle, algorithmes d'appariement et tests avec le package package `MatchIt`

⋆ Lectures conseillées : (Li, 2013)

(12-04) *Introduction du temps pour identifier l'effet causal ("before"-"after" models) : Difference-In-Difference*
⟶ `Krueger & Card dataset`

☆ "Traitement" et "contrôle", spécification du modèle, validation et test

⋆ Lectures obligatoires : (Wing et al., 2018)

(3-05) Séance de Q&R à propos de la fiche d'exercices 2 (en ligne sur Zoom)

Modalités d'évaluation

Les étudiant-e-s sont évalué-e-s individuellement et collectivement en 100% CC.

☆ **Assiduité et implication : 15%**

- ⋆ Présence durant le semestre
- ⋆ Participation en séance et durant les séances de Q&R
- ⋆ Dépôt des rendus avant la deadline

☆ **Fiches d'exercices : 50%** ⟶ rendues les 14-03 et 9-05

- ⋆ Les étudiant-e-s devront déposer en ligne les réponses aux différentes questions posées dans les deux fiches d'exercices avant l'heure limite sur un seul et unique document, réalisé dans l'environnement Markdown et compilé au format html

☆ **Challenge en groupe : 35%** ⟶ organisé le 17-05

- ⋆ Différents groupes seront formés pour réaliser chacune des trois épreuves d'une heure (pause de 15 minutes entre chaque épreuve)

Contrat pédagogique et règles de fonctionnement du semestre

☆ **Les absences durant les séances ou les sessions d'évaluation**

- ⋆ La présence est obligatoire tout au long du semestre. Toute absence devra être dument justifiée dans les plus brefs délais et dans une limite de 3 jours. Un-e étudiant-e qui présentera une absence injustifiée obtiendra 0 pour la composante "assiduité" de son évaluation.
- ⋆ Les justificatifs fournis devront correspondre explicitement aux cas définis par l'Université de Paris et sous réserve d'acceptation de la part de l'enseignant. Aucun justificatif en dehors des cas listés par l'administration ne pourra être accepté.

☆ **Règles concernant les différentes évaluations tout au long du semestre**

- ⋆ Les problèmes matériels devront être signalés le plus tôt possible, et resteront à l'appréciation de l'enseignant. Aucune exemption de dernière minute ne sera accordée en dehors des cas listés par l'UFR GHES. Toute demande d'exemption devra être précisément justifiée et son acceptation restera à l'appréciation de l'enseignant.
- ⋆ A la suite d'une évaluation, aucune réponse ne vous sera fournie par mail. Une correction sera disponible en ligne au maximum dans les deux semaines suivant l'évaluation (à l'exception du "challenge").
- ⋆ Le plagiat de ressources en ligne, d'ouvrages ou de travaux de camarades est formellement interdit et demeure soumis à la charte "anti-plagiat" de l'université de Paris, disponible en suivant ce lien. Une procédure disciplinaire sera engagée dès lors qu'un cas de plagiat sera constaté.
- ⋆ Les évaluations individuelles devront être réalisées individuellement. Si un doute réel et sérieux se présente à propos de deux ou plusieurs étudiant-e-s, la note définitive pour chaque étudiant au test concerné sera de 0.

- Les différents rendus des fiches d'exercices devront se faire dans le strict respect des règles indiquées lors de la séance introductive : dépôt en ligne et au format html après compilation sous Markdown. En l'absence de respect de ces règles, le devoir rendu par un autre moyen et sous une autre forme ne sera pas corrigé et obtiendra la note de 0.

☆ Les interactions avec l'enseignant

- Vous êtes invité-e-s à poser toutes vos questions et à participer à la vie de la communauté "AMA-R" sur le forum à votre disposition. Vous pourrez alors poser toutes vos questions : cours, exercices, nouvelles ressources...
- Vous pouvez également prendre rendez-vous afin que nous puissions échanger durant les heures de permanence : mardi 9h-10h30.
- Vous pouvez également m'adresser vos mails si nécessaire (thibaud.deguilhem@u-paris.fr), en respectant bien entendu les recommandations de la fiche disponible en ligne. Les réponses à vos questions par mail vous parviendront seulement durant les heures de permanence.

Informations pratiques

☆ Page dédiée, communauté AMA-R et ressources en ligne

- Retrouvez toutes les informations et les ressources en ligne sur la page du cours : ama-r-me3ay020
- Participez à la communauté AMA-R sur le forum dédié : ama-r-master-pise-ccese-data

☆ Types de séance et nombre d'heures affectées

- Séances de cours et d'applications sur site en présentiel (ou à distance sur Zoom) (4h/séance)
  - ⟶ du 25-01 au 15-02
  - ⟶ du 15-03 au 12-04
- Séances de Q&R sur les fiches d'exercices en ligne sur Zoom (2h/séance)
  - ⟶ le 8-03
  - ⟶ le 3-05

☆ Horaire et salle du cours

- Lundi matin
  - ⟶ 10h-12h puis 13h-15h pour les séances de cours (sur site ou sur Zoom)
  - ⟶ 13h-15h pour les séances de Q&R (en ligne sur Zoom)
- salle 375, bâtiment Olympe de Gouges campus PRG (localisation : maps)

☆ Dates des rendus (évaluations)

- 14-03 : rendu de la fiche d'exercices 1 en ligne (distribution de la fiche le 15-02)
- 09-05 : rendu de la fiche d'exercices 2 en ligne (distribution de la fiche le 12-04)
- 17-05 : challenge en groupe organisé en salle 375 (en fonction des conditions sanitaires)

☆ Contact et permanence

- ⋆ mail : thibaud.deguilhem@u-paris.fr
- ⋆ Permanence le mardi de 9h à 10h30 (bureau 820, bâtiment Olympe de Gouges ou par mail)

Références

Deguilhem, T. and Seetahul, S. (2016). Modèles de mélange fini (FMM) appliqués à la segmentation du marché du travail à Bogota. *Bulletin of Sociological Methodology*, 132(1) :26–43.

Denis, D. J. (2015). *Applied Univariate, Bivariate, and Multivariate Statistics*. Wiley-Blackwell, Hoboken.

Denis, D. J. (2020). *Univariate, Bivariate, and Multivariate Statistics Using R : Quantitative Tools for Data Analysis and Data Science*. Wiley-Blackwell, Hoboken.

Husson, F., Le, S., and Pagès, J. (2017). *Exploratory Multivariate Analysis by Example Using R*. CRC Press, Boca Raton, 2nd edition.

Li, M. (2013). Using the Propensity Score Method to Estimate Causal Effects : A Review and Practical Guide. *Organizational Research Methods*, 16(2) :188–226.

MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation Analysis. *Annual Review of Psychology*, 58(1) :593.

Stahl, D. and Sallis, H. (2012). Model-based cluster analysis. *WIREs Computational Statistics*, 4(4) :341–358.

Stock, J. H. and Watson, M. (2014). *Principes d'économétrie*. Pearson, Paris.

Wing, C., Simon, K., and Bello-Gomez, R. A. (2018). Designing Difference in Difference Studies : Best Practices for Public Health Policy Research. *Annual Review of Public Health*, 39(1) :453–469.

Zelterman, D. (2015). *Applied Multivariate Statistics with R*. Statistics for Biology and Health. Springer International Publishing, Geneva.