Reproductibilité, transparence et à ouverture des données

Séminaire Administrer la preuve statistique en sciences sociales

Anton Perdoncin

Lundi 4 novembre 2019

Section 1

Reproductibilité, réplicabilité et transparence

Enjeux notionnels

Réplication : tentatives de vérification des résultats d'une étude scientifique en reproduisant le protocole de l'enquête originelle.

Reproductibilité : ensemble de règles pratiques permettant la réplication et la vérification des résultats ; facilitant aussi le travail collaboratif.

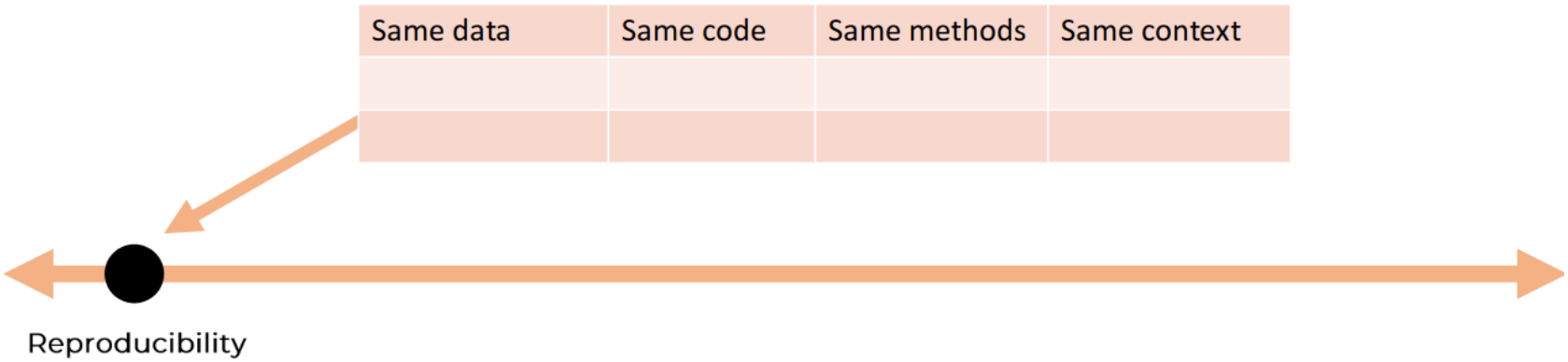
Transparence : ensemble de règles pratiques permettant l'accès à l'ensemble de la chaîne de production de la preuve statistique : + Données brutes + Données nettoyée et recodées + Scripts + Résultats intermédiaires + Documentation de l'enquête et de son traitement

Enjeux pratiques

Organiser de façon reproductible (et efficace !) le travail de quantification n'est pas chose aisée :

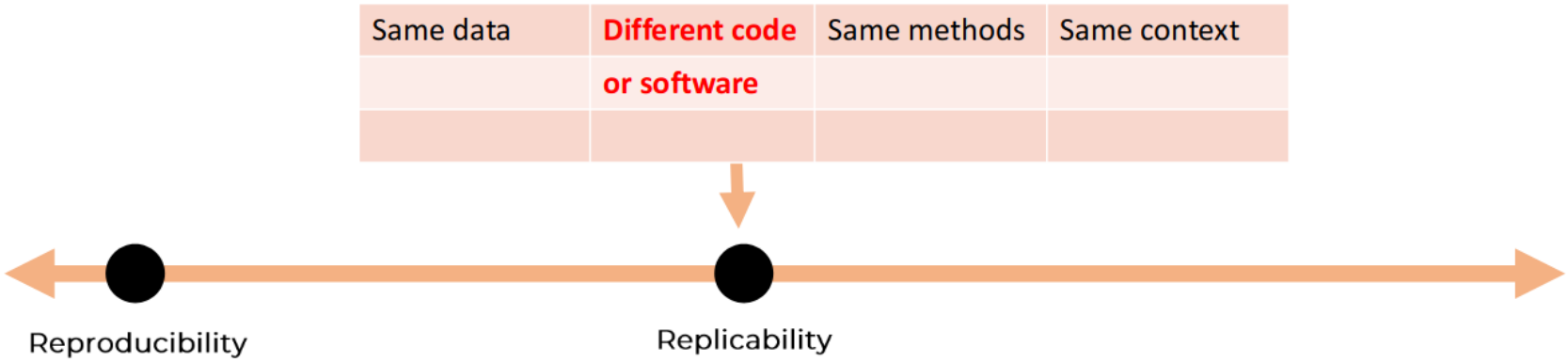
- des fichiers de nature différentes : jeux de données, scripts, tableaux, images, notes, documentation diverse ...
- des va et vient entre opérations diverses : importer et nettoyer des données, recoder, décrire, modéliser ...
- on oublie assez vite ce qu'on a fait si l'on n'y touche pas pendant... un certain temps !
- lire le code des autres peut être compliqué ...

Reproductibility continuum



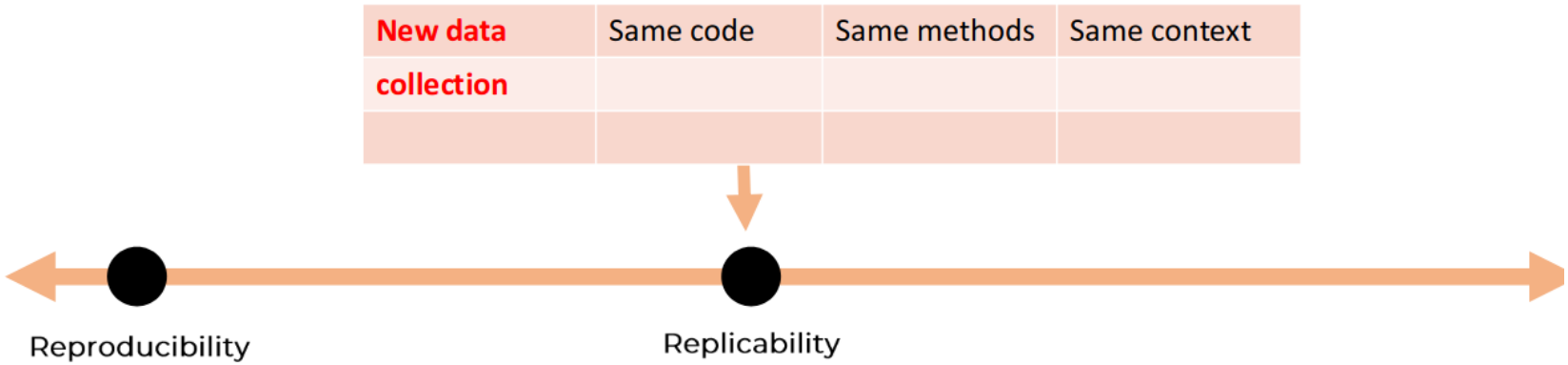
Lars Vilhuber : https://www.ciqss.org/sites/default/files/documents/Lars%20Vilhuber_%C3%A9thiques.pdf

Reproductibility continuum



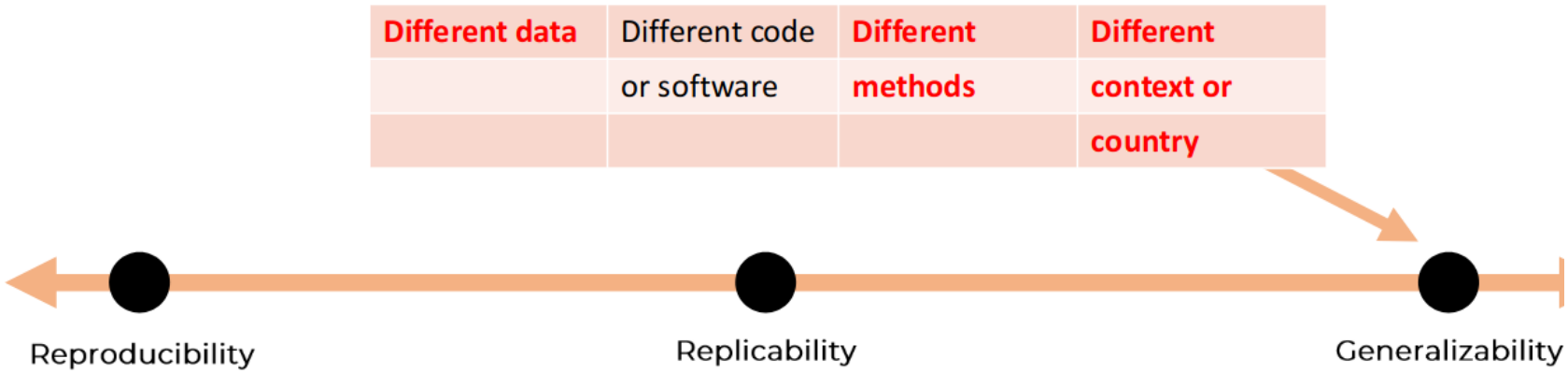
Lars Vilhuber : https://www.ciqss.org/sites/default/files/documents/Lars%20Vilhuber_%C3%A9thiques.pdf

Reproductibility continuum



Lars Vilhuber : https://www.ciqss.org/sites/default/files/documents/Lars%20Vilhuber_%C3%A9thiques.pdf

Reproductibility continuum



Lars Vilhuber : https://www.ciqss.org/sites/default/files/documents/Lars%20Vilhuber_%C3%A9thiques.pdf

Les principes FAIR et la FAIRification de la recherche

Principes FAIR publiés en 2016 dans *Scientific Data* une revue associée à *Nature*.

"Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments. Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments."

Les principes FAIR et la FAIRification de la recherche

Findable : afin d'utiliser ou réutiliser des données, il faut pouvoir les trouver :

- F1 : LEs (méta)données identifiées de façon unique et persistante
- F2 : Les données sont décrites à l'aide de métadonnées riches (voir R1)
- F3 : Les métadonnées incluent l'identifiant des données qu'elles décrivent
- F4 : Les (méta)données sont enregistrées ou indexées dans une support cherchable

Les principes FAIR et la FAIRification de la recherche

Accessible : une fois qu'un utlisateur a trouvé des données, il doit pouvoir savoir rapidement à quelles conditions celles-ci sont accessibles :

- A1 : Les (méta)données peuvent être retrouvées à partir de leur identifiant, selon un protocole standardisé de mise à disposition
  - A1.1 : ce protocole doit être ouvert, gratuit et inter-opérable
  - A1.2 : ce protocole permet une authentification ou une procédure d'autorisation, si nécessaire.
- A2 : Les métadonnées sont accessibles, même lorsque les données ne sont plus disponibles.

Enjeux institutionnels : les principes FAIR et la FAIRification de la recherche

Interoperable : les données doivent pouvoir être intégrées à d'autre données, et être intégréees à des workflow et des supports de stockage divers

- I1 : Les métadonnées utilisent un langage formel, accessible, partagé et applicable largement
- I2 : Les métadonnées utilisent un vocabulaire qui suit les règles du FAIR
- I3 : Les métadonnées incluent d'éventuelles références explicites à d'autres métadonnées

Les principes FAIR et la FAIRification de la recherche

Reusable : afin d'être réutilisées, les données et les métadonnées doivent être bien décrites, de façon à ce qu'elles puissent être répliquées et/ou combinées dans différents contextes.

- R1: Les (méta)données dont décrites de façon riche, avec une pluralité d'attributs pertinents et précis.
  - R1.1 : Les méta(données) sont diffusées avec une licence d'usage claire et accessible
  - R1.2 : La provenance des méta(données) est décrite de façon détaillée
  - R1.3 : Leur description répond aux standards en vigueur dans l'univers scientifique ou professionnel

Section 2

Le problème de la reproductibilité

Au principe de pratiques scientifiques collaboratives

La reproductibilité est au principe de l'**échange d'expériences**, de la **résolution de problèmes de codage** et du **travail scientifique collaboratif** :

- Apprenez à rédiger des MWE (Minimal Working Examples) ou EMR (Exemples minimaux reproductibles) : cf. stackoverflow
- Organiser correctement ses scripts est une condition pour pouvoir travailler avec d'autres...
- ... mais c'est aussi une condition pour pouvoir bien travailler avec soi-même !

Trois objectifs principaux

1. Reproduire **soi-même** ses propres résultats, même après un certain temps passé sans y avoir touché...
2. Permettre à d'**autres** de les reproduire
3. Permettre le travail collaboratif

Un script reproductible permet donc, à condition de disposer des données, de **reproduire exactement les résultats** d'une publication.

Principes généraux

1. **Conserver la trace de la manière dont sont produits les résultats**: à chaque document que l'on produit (article, rapport, mémoire...) correspond un (ou un série de) script(s) R. L'ensemble des résultats statistiques présentés dans ce document doit être produit par l'exécution de ce(s) script(s).
2. **Conserver la mémoire des modifications successives du code**: des outils spécialisés et très efficaces permettent de faire du "contrôle de version" (*version control* : GitHub ou GitLab par exemple). De façon plus simple (mais aussi moins fiable), numérotez vos scripts et vos documents pour chaque nouvelle version (V1, V2.... Vdef).

Principes généraux

3. **Sauvegarder les résultats intermédiaires** : vous nettoyez les données... recodez un grand nombre de variables... définissez une sous-population, etc. : exportez le résultat dans un fichier (.csv par exemple) afin de pouvoir le réutiliser sans difficulté et de façon sûre.
4. **Structurer et expliciter votre script de façon logique et non ambiguë** : on fait les choses dans un certain ordre (importer, nettoyer, recoder, analyser) et on explicite tout ce que l'on fait !

Ce qu'on vous pourrait vous demander en annexes d'une thèse ou d'un article

1. Fournir (format .csv) **deux jeux de données** : données "brutes" (avant manipulation/nettoyage) et "propres" (après manipulations/nettoyages).
2. Fournir (format .R) un **script reproductible** qui permette de vérifier que toutes vos manipulations statistiques sont correctes.
3. Ces aspects sont malheureusement trop souvent négligés... ils font (ou devraient faire) pourtant partie de l'**évaluation** d'un mémoire ou d'un article de recherche.

Les règles d'un script reproductible

1. Structure : titre, sectionnage, enchaînement logique des opérations (chargement des packages avant celui des données ; recodage avant l'analyse, etc.) ;
2. Propreté du code
3. Absence d'erreurs de compilation
4. Commentaires abondants

A propos des commentaires

Pourquoi commenter ?

1. pour vous-mêmes et pour les autres
2. pour créer des sections et structurer le code
3. pour expliciter des passages complexes
4. pour expliquer les étapes d'un travail
5. pour noter des idées telles qu'elles vous viennent... des premières ébauches de lecture statistique d'un tableau ou d'un graphique, etc.
6. De façon générales, afin de documenter les étapes de la chaîne de production de l'argument statistique.

Section 3

Organiser le workflow

Structurer des projets

- Utiliser des **projets RStudio**
  - une publication = un projet
  - un mémoire = un projet
  - un projet... = un projet !
- Un **projet** RStudio correspond à un **dossier** dans votre ordinateur
  - Bien travailler, c'est d'abord bien organiser l'arborescence de son ordinateur
  - Distinguer : types de productions et types de documents.

Exemple de structure d'un projet

Supposons que vous ayiez à rédiger une thèse . . . Un exemple d'arborescence du dossier *these* dans votre ordinateur :

- Les sources et matériaux empiriques
  - *data*
  - *entretiens*
  - *observations*
  - attention à bien sauvegarder tout cela sur des support physiques et jamais sur des clouds !
- Les documents intermédiaires permettant de produire des résultats
  - *editeurs* : tous les fichiers .R et .Rmd
  - *analyses* : reprise de vos matériaux quanti et quali avant rédaction
  - un fichier *todo* est aussi utile. . .

Exemple de structure d'un projet

- Les résultats de vos analyses
  - *graphiques* : .png, .jpg, .pdf
  - *tableaux* : sorties numériques de vos traitements stats
  - . . . et autres : *portraits*, *recits_observations* . . .
- Les autres documents utiles
  - *illustrations* : les images et autres documents que vous utiliserez pour produire vos documents finaux
  - *documentation* : dictionnaires de codes, inventaires d'archives, cartes, plans, etc.

Exemple de structure d'un projet

- Enfin, l'ensemble des documents permettant de produire le mémoire
  - dans un dossier *redaction*
  - partitionnez votre theses en chapitres que vous ne réunirez qu'à la toute fin
- Cette structure est indicative et à adapter en fonction des projets : exemple : mon projet *lubartworld*.

Versionner son travail

- Le **package *checkpoint*** : https://cran.r-project.org/web/packages/checkpoint/vignettes/checkpoint.html
- **Deux modes de versionnage**
  - de façon manuelle (document_date)
  - de façon automatisée : Git (GitHub, GitClone . . .)
- **Sauvegardez régulièrement** : sur support physique (disque dur) et en synchronisation de type “cloud” (mais pas les données personnelles !!!).

Six principes d'organisation d'un espace de travail quantitatif

1. **Transparence** : les éléments doivent être organisés de façon logique et lisible par n'importe qui ouvrirait le dossier (y.c vous. . .)
2. **Adaptabilité** : un bon workflow doit être aisément modifiable et adaptable ; d'où : opter pour des noms de fichiers standardisés et adopter de bonnes pratiques de commentaire.
3. **Modularité** : chaque tâche distincte doit être affectée à un fichier (e.g. un script) distinct ; il est ainsi toujours facile de savoir où faire des modifications, et quelles parties reprendre pour d'autres projets.

Six principes d'organisation d'un espace de travail quantitatif

4. **Portabilité** : un bon workflow doit être transposable sur une autre machine, un autre système informatique. D'où : opter pour des liens relatifs vers les données, les illustrations, les sorties, etc.
5. **Reproductibilité** : cf. *supra*.
6. **Efficacité** : la vôtre, pas celle du processeur de votre ordinateur ! Un bon workflow vous fait gagner du temps et vous simplifie la vie en automatisant tout ce qui peut l'être !

David Smith : https://blog.revolutionanalytics.com/2010/10/a-workflow-for-r.html