

Données

Structures fondamentales : tables
et bases de données

Qu'est ce que c'est qu'une base
de données ?

Structure des données issues de méthodes quantitatives (= données statistiques)

1) caractérisées par un format récurrent (=multi-occurent)

2) peuvent être mis en tableau avec:

- en lignes = de nombreux objets ou unités observés (= 'individus' ou items), repérés par un identifiant

- en colonnes: des variables ou 'descripteurs' (ou 'paramètres') qui décrivent les objets/unités observés

Unités observées

Variables ou descripteurs

Iden	Sexe	Age	Educ at	Qu'est ce le plus important pour vous dans la vie?
391	F	71	bas	la santé.
313	F	59	bas	C'est de faire ce qu'on veut.
314	F	50	bas	la santé puisqu'il faut toujours travailler quand on est commerçant.
251	F	80	moyen	la santé.
252	F	35	élevé	le calme intérieur.
253	M	60	bas	etre bien dans sa peau et heureux dans son ménage.
254	M	22	moyen	etre heureux.
255	M	39	élevé	l'équilibre entre toutes les choses importantes de la vie.
256	F	45	moyen	la santé.
771	M	30	moyen	bonheur.
772	M	46	élevé	s'épanouir pleinement.
773	F	41	élevé	bonne santé.
774	F	84	bas	etre avec les enfants.
775	M	35	moyen	vivre bien.
776	M	18	moyen	etre bien dans sa peau.
777	F	56	moyen	travail.
778	F	25	élevé	amitié.
779	M	66	bas	santé.
780	F	22	élevé	bonheur
781	M	27	moyen	etre heureux.
165	F	57	bas	réussir sa vie sur un plan humain.
166	M	73	bas	la santé cela a toujours été notre fortune.

(Objets =) Unités observées ou unités d'observation

- Une unité d'observation est une entité ou « chose » que l'on décide d'observer. Cette entité peut avoir une nature de différents types:

Ex. d'entités
naturelles
'palpables',
'tangibles' ou
quasi-tangibles

- personnes, êtres vivants (animaux, plantes)
- objets physiques tangibles et pérennes (ex.: embarcations, habitations, parcelles foncières, débarcadères, villages, bassins versants etc..)
-

Ex.: d'entités
abstraites,
entités
artificielles
« ad hoc »

- groupes de personnes (ex.: ménages), institutions (ex.: coopératives, entreprises, associations), Etats
- événements (accidents, catastrophes naturelles, guerres, crises)
- chansons, romans (textes de)
- actions simples (ex.: transactions de vente), des séquences d'actions (ex.: campagnes agricoles)
- morceaux de l'espace (unités d'espace, 'quadrats'), des périodes du temps (semaine, mois),
-

Pour une étude donnée: les unités d'observation → les unités statistiques 'de base' ou 'originelles'

Les « unités observées » sont pleinement définies lorsqu'on a précisé la façon dont on les observe/décrit (→ liste de *descripteurs/ variables*)

Une unité d'observation = un objet + la façon dont voit/décrit cet objet

Descripteurs d'une pirogue (= variables)

Unité
d'observation
« pirogue »

N° immatriculation	Date de construction	longueur	Nom du propriétaire
-----------------------	-------------------------	----------	------------------------

Descripteurs d'une sortie de pêche (= variables)

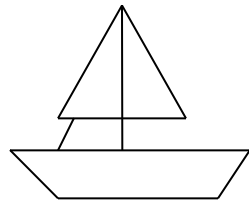
Unité d'observation
« sortie de pêche »

Date	Heure de départ	Heure de retour	Techniques de pêche	Nombre de participants	Quantité de poisson capturée
------	--------------------	--------------------	------------------------	---------------------------	------------------------------------

Il peut y avoir des relations structurelles entre les unités d'observation ex.: l'u.o. « pirogue » et l'u.o. « sortie de pêche »

*Unité concrète,
tangible, assez
Pérenne: l'unité est
bien une
« chose physique »*

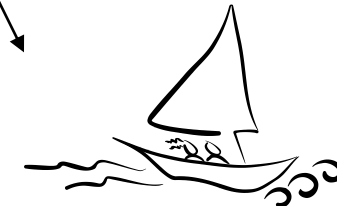
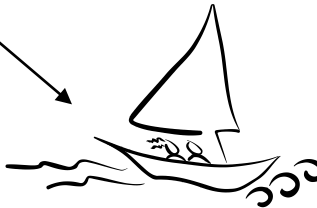
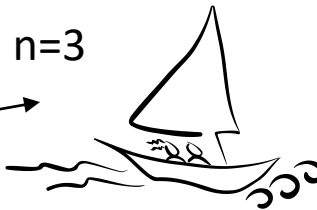
une pirogue ,



- n° d'immatriculation
- longueur
- nom du propriétaire
- année de construction
- type de motricité (voile, moteur..)

1

une sortie de pêche



- Date,
- heure de départ,
- heure de retour,
- techniques de pêche utilisées,
- nbre de participants,
- quantité capturée

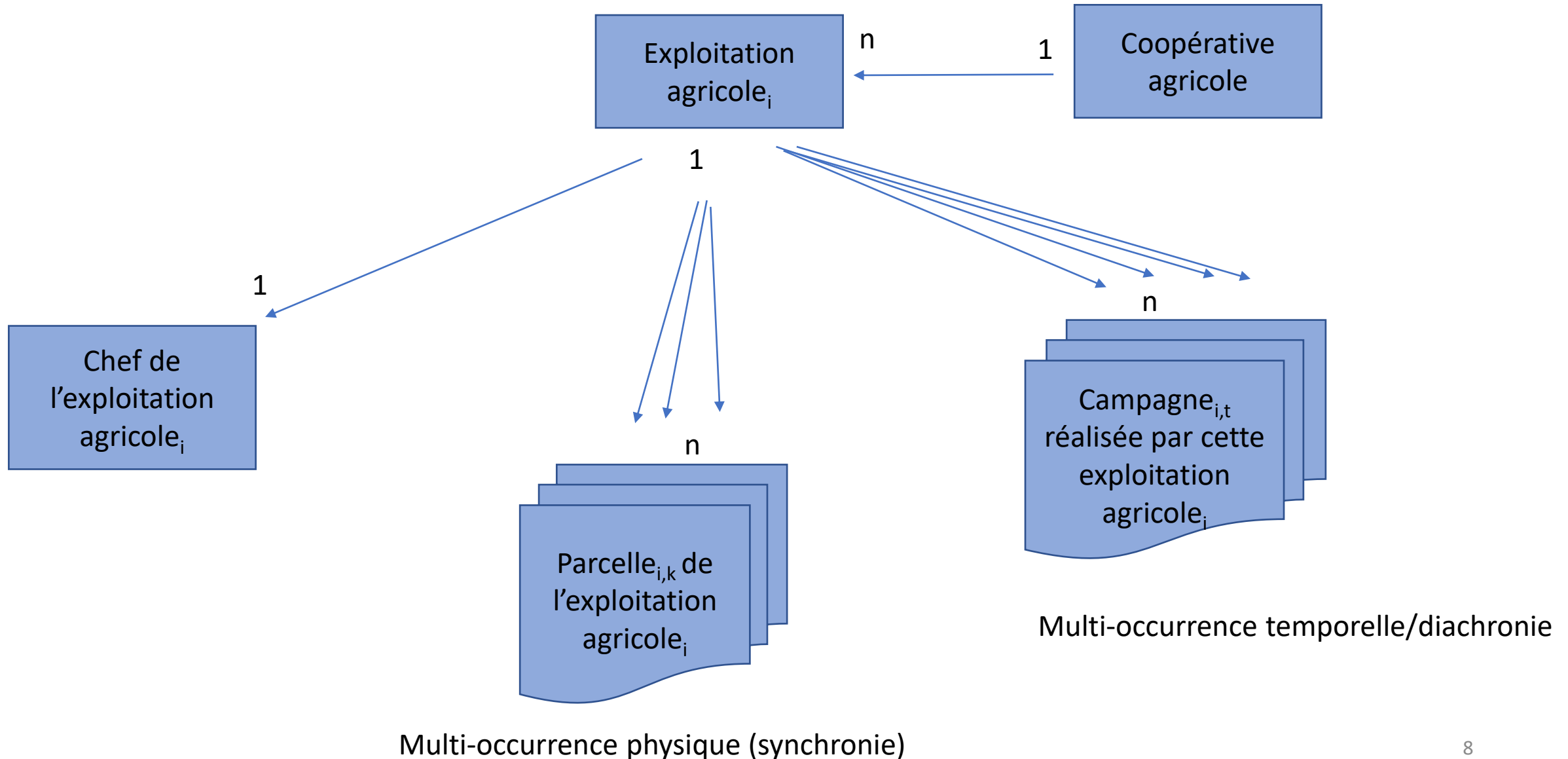
- Date,
- heure de départ,
- heure de retour,
- techniques de pêche utilisées,
- nbre de participants,
- quantité capturée

- Date,
- heure de départ,
- heure de retour,
-

*Unité non tangible:
séquence d'actions, relative à
une échelle de temps assez
brève: qq heures à qq jours*

Types de relation:
possession, inclusion
subordination, appartenance, emboîtement

Autre exemple: diverses unités d'observation dans le cadre d'une étude quantitative en agronomie sur les exploitations céréalières d'un département, avec leurs relations



Descripteurs d'une exploitation

Unité d'observation
« exploitation agricole »

Localisation administrative des bâtiments principaux (ferme) (nom de la commune)	Nombre d'adultes actifs impliqués dans l'exploitation	Nbre de parcelles rattachées de façon à l'exploitation	Surface totale des parcelles rattachées à l'exploitation
--	---	--	--

Descripteurs d'un(e) chef(fe) d'exploitation

Unité d'observation
« chef d'exploitation »

Nom, prénom	Age	Sexe	Situation maritale	Niveau de formation atteint	Nbre d'enfants
-------------	-----	------	--------------------	-----------------------------	----------------

Descripteurs d'une parcelle agricole

Unité d'observation
« parcelle »

Localisation GPS de la parcelle (centre de gravité)	Surface de la parcelle	Type de sol (pH)	Types de droit d'usage (propriété, location)
---	------------------------	------------------	--

Descripteurs d'une campagne agricole année t d'une exploitation

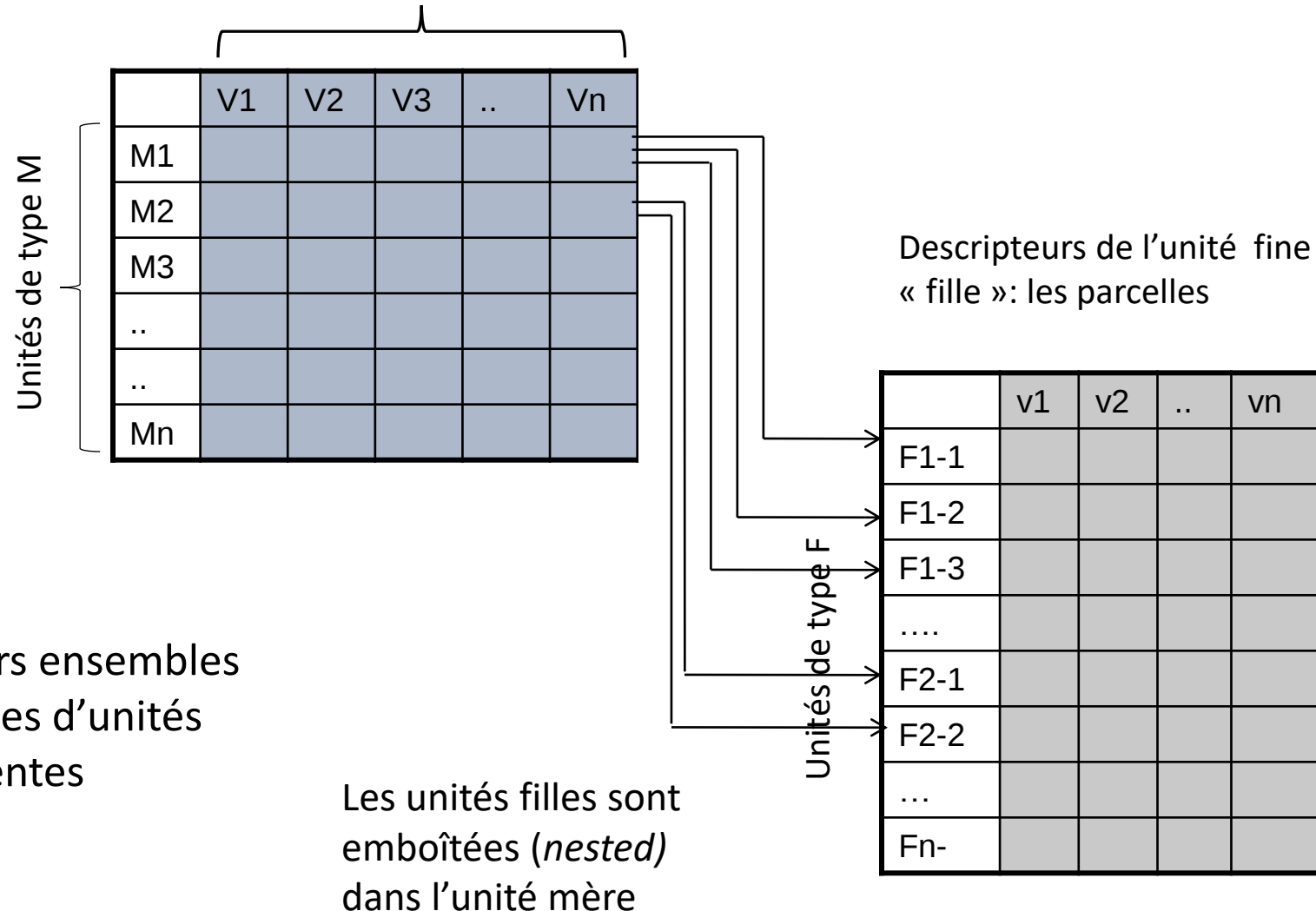
Unité d'observation
« Campagne agricole d'une exploitation »

Année t	Nbre totale de parcelles cultivées	Date début des travaux de labours	Coûts engagés dans l'achat semences	Date début récolte	Production en céréales (tonnes) lors de l'année t	Recettes (C.A.) tirées de la vente de la récolte
---------	------------------------------------	-----------------------------------	-------------------------------------	--------------------	---	--



La structure des données est 'naturellement branchue' (= non plate) du fait des liens existants entre types d'unités lorsqu'il y a des relations de 1 à n

Descripteurs de l'unité supérieure « mère »: l'exploitation agricole



Enjeux d'accessibilité

Est-ce que vous utilisez souvent
des données ouvertes ?

Si oui, où les trouvez-vous ?

Open data policy principles

Les principes d'une stratégie d'ouverture des données

1. ***Rendre publiques les données*** : par défaut, sauf contradiction avec des restrictions dues au respect de la vie privée, de la confidentialité, de la sécurité
2. ***Accessibilité des données***: présentées sous des formats techniquement reconnus, sur des supports appropriés, non sélectifs, lisibles par informatique pour autoriser le traitement automatique.
3. ***Documentation***: Les données doivent être décrites de façon à ce que les utilisateurs puissent en comprendre le sens et les limites et les utiliser à bon escient (meta-information). La méthodologie d'échantillonnage et les référents temporels et spatiaux doivent accompagner les données.

Open data policy principles

4. **Réutilisabilité:** Les données doivent être disponibles sans restriction d'utilisation.
5. **Complétude:** Les données doivent être publiées de façon complète, dans leur forme fine (telles que collectées à la source = microdata).
6. **Sans délai ou sous peu délai:** Les données doivent être rendues disponibles aussi vite que possible.
7. **Gestion et accompagnement après diffusion:** Un point de contact doit être fourni par le fournisseur de données pour répondre aux plaintes et aux questions et pour assister les utilisateurs dans leur demande par rapport aux données.

Ouverture des données

- Données ouvertes

vs.

- Données FAIR

"Aussi ouvert que possible, aussi fermé que nécessaire"

Typologies de données

Source

Méthode

Dimensions temporelles et spatiales

A votre avis, qui produit les données qu'on utilise habituellement ?

- Institutions internationales (rôle souvent surestimé) :
 - enquêtes d'expert, p. ex. Banque mondiale sur les institutions WGI, CPIA, Doing Business...
 - surtout, compilations de données issues des systèmes nationaux.
- Systèmes statistiques nationaux, les sources plus sous-estimés ;
- Autres sources “contributives”, à croiser (voir après).

Système statistique national

- Quels sont selon vous les grands types d'instances qui composent un système statistique national ?
- Instituts nationaux de statistique : INSTAT
- Instituts géographiques nationaux : FTM
- Systèmes d'information :
 - Ministères, collectivités locales;
 - Opérateurs privés de services publics;
 - Thématiques (Museums, météo...).

Principales méthodes de collecte

- Enquêtes
 - Ménages : niveau et conditions de vie, emploi, informel
 - Entreprises
 - Etablissements publics (santé, écoles...)
- Recensements
 - Population et habitat
 - Agricoles
- Données administratives
- Capteurs

Où trouver ces données ?

- Catalogues par famille d'enquête
 - Demographic & Health Surveys : www.dhsprogram.com
 - Multiple indicator cluster surveys : <https://mics.unicef.org>
 - Living standard measurement survey : www.worldbank.org/LSMS
- Enquêtes internationales
 - Banque mondiale : [Global Findex](http://GlobalFindex), Doing business...
 - Enquêtes de tiers : Afrobarometer/[Latinobarómetro](http://Latinobarometro)/etc.
- Catalogues internationaux
 - International Household Survey Network : <http://ihsn.org>
 - Integrated Public Use Microdata Series : www.ipums.org
 - World Bank microdata catalog : <https://microdata.worldbank.org>
- Autres sources
 - Administrations des collectivités territoriales, aux associations (ex. data.gouv.fr)
 - Projets de recherche qui veulent bien fournir leurs fichiers
 - Aux institutions statistiques officielles: INS, DNS, INSEE
 - Opérateurs privés générant des données comme « *by product* » (caisses de supermarchés)

Où trouver ces données ? Sources nationales

- **Catalogues internationaux**

- International Household Survey Network : <http://ihsn.org>
- Integrated Public Use Microdata Series : www.ipums.org
- World Bank microdata catalog : <https://microdata.worldbank.org>

- **Catalogues par famille d'enquête**

- Demographic & Health Surveys : www.dhsprogram.com
- Multiple indicator cluster surveys : <https://mics.unicef.org>
- Living standard measurement survey : www.worldbank.org/LSMS

- **Sites nationaux**

- INS : propres à chaque pays
- Ministères : *idem*

Où trouver ces données ? Sources internationales

- Indicateurs issus des systèmes nationaux
 - World Bank: WDI, SCI, Dette... <https://databank.worldbank.org/databases>
 - Portails de données des agences des Nations Unies : [PNUD](#), [FAO](#), [UNICEF](#), [BIT](#)...
- Enquêtes internationales
 - Banque mondiale : [Global Findex](#), Doing business...
 - Enquêtes de tiers : [Afrobarometer](#)/[Latinobarómetro](#)/etc.
- Plateformes internationales d'échanges
 - Humanitarian Data Exchange : <https://data.humdata.org/>
- Autres sources
 - Projets de recherche qui veulent bien fournir leurs fichiers
 - Opérateurs privés générant des données comme « *by-products* » (ex. téléphonie)

Où trouver ces données ? Sources satellitaires

Plus compliqué car elles sont généralement classées par source et pas par indicateur, voir :

- **Open Geospatial Solutions** : <https://github.com/opengeos>

Exercice

Cherchez des données pour votre propre analyse.

Dimensions temporelles et
spatiales

Dimension temporelle

- Données en coupes transversales
 - Coupe transversale simple
 - Coupes répétées
- Données en panel
 - Panels cylindrés
 - Panels non cylindrés
- Pseudo-panels

Cross section

individual	Y	X	Z
1	Y1	X1	Z1
2	Y2	X2	Z2
3	Y3	X3	Z3
...			
N	YN	XN	ZN

Time series

year	Y	X	Z
1	Y1	X1	Z1
2	Y2	X2	Z2
3	Y3	X3	Z3
...			
T	YT	XT	ZT

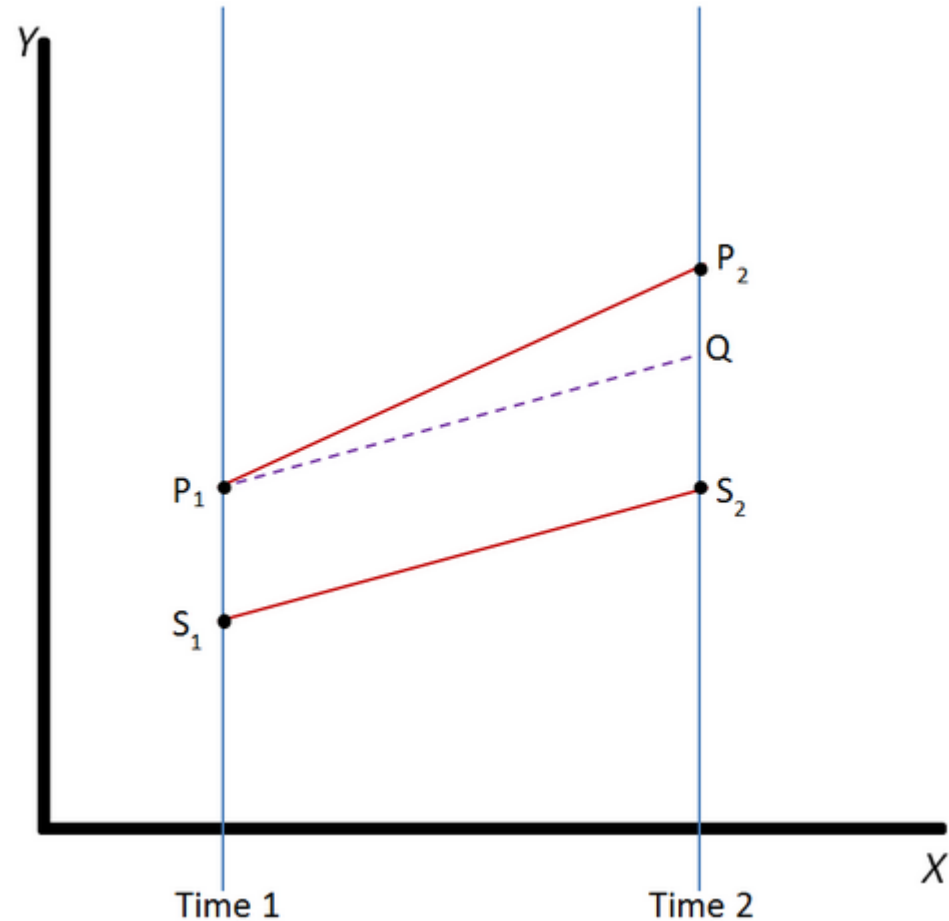
==>

Panel data

individual	year	Y	X	Z
1	1	Y11	X11	Z11
1	2	Y12	X12	Z12
...	...	...	...	...
1	T	Y1T	X1T	Z1T
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
N	1	YN1	XN1	ZN1
N	2	YN2	XN2	ZN2
...	...	...	...	...
N	T	YNT	XNT	ZNT

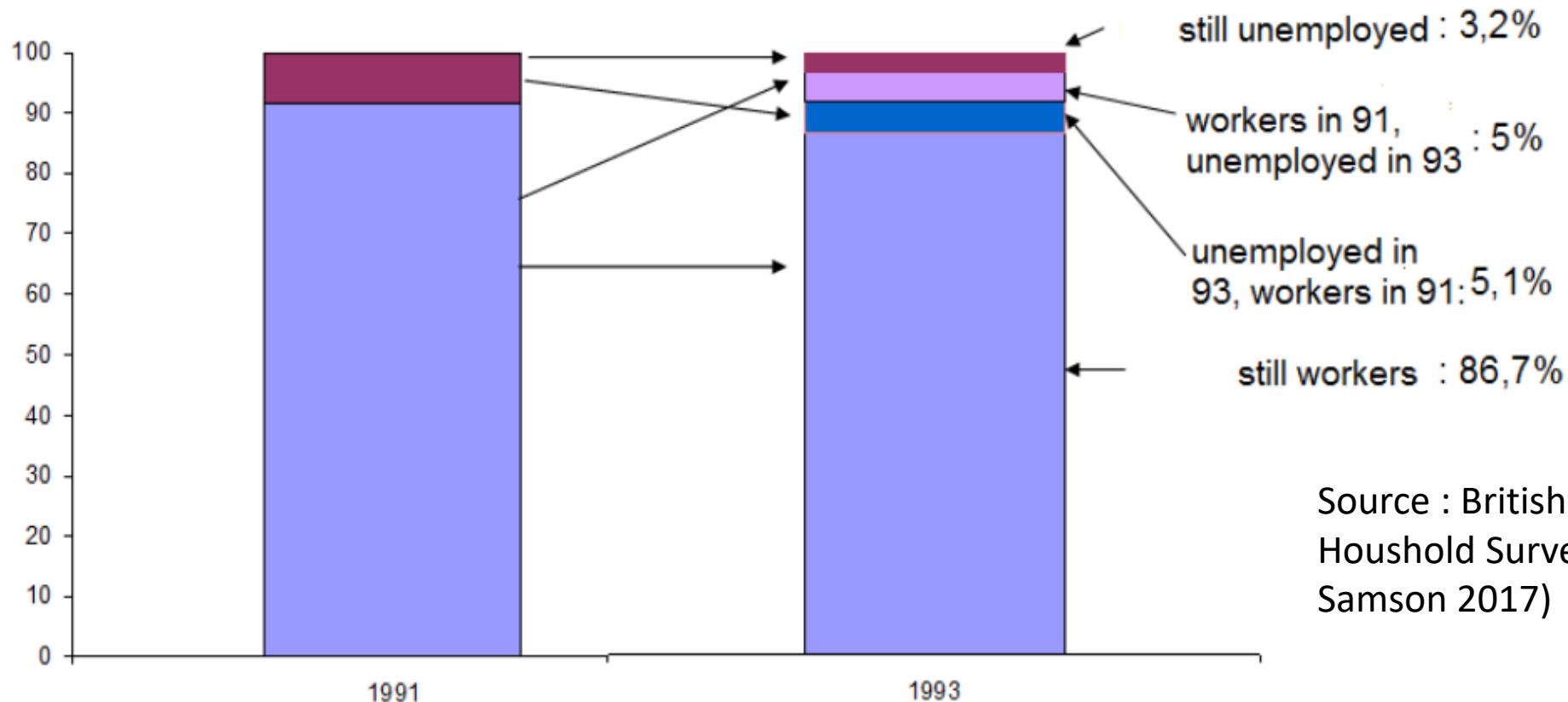
Avantages des données de panel (1/3)

- Contrôler l'hétérogénéité spatiale et temporelle



Avantages des données de panel (2/3)

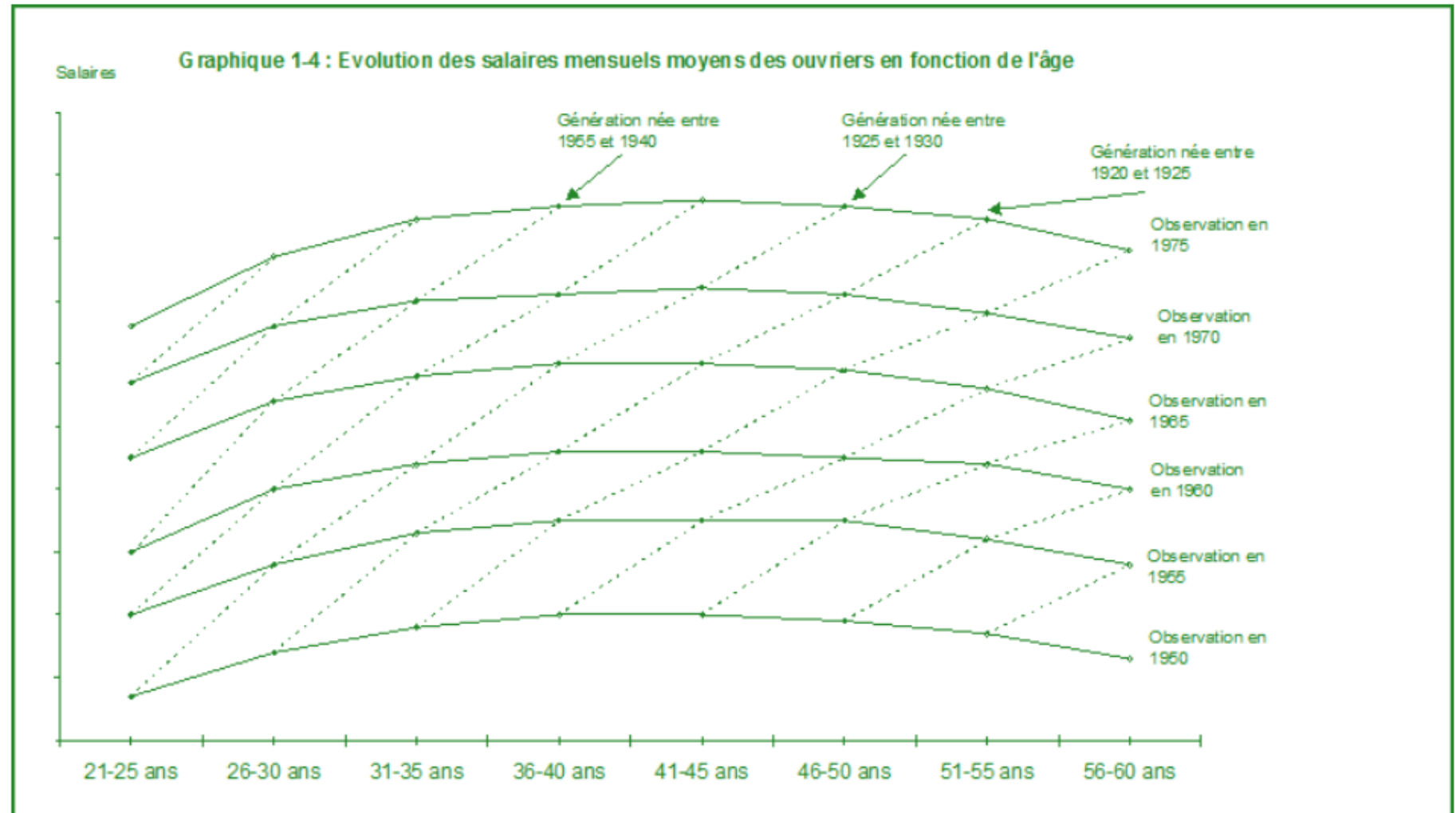
- Distinguer l'évolution et l'hétérogénéité des comportements individuels



Avantages des données de panel (3/3)

Distinguer les effets:

- âge,
- date
- cohorte



Diffusion des usages des données spatiales satellitaires

- **Versant méthodologique**
 - Des sciences naturelles aux sciences sociales
 - Du fondamental aux applications
- **Versant de la mise en oeuvre**
 - Evaluation ex post
 - Suivi-évaluation
 - Conception
 - Opérationnel (ex. [projet Talaky à Beampigaratsy](#))

Champ toujours plus large

Quelques synthèses de la littérature en économie ([Donaldson and Storeygard 2016](#))

Inventaire en cours : www.3ieimpact.org/resources/remote-sensing-inventory

Sujets “classiques”

- Environnement/biodiversité
- Gestion forestière
- Qualité de l’air
- Climat, météorologie
- Topographie : accidenté
- Agriculture, sécurité alimentaire
- Hydrologie, risques naturels

“Nouveaux” sujets

- Développement économique : luminosité nocturne, infrastructures
- Urbanisation, bâti, localisation et mouvements de population, transports
- Déterminants de santé
- Raffinement des sujets classiques (ex. dégradation des forêts)
- Interfaces entre ces sujets (“nexus”)

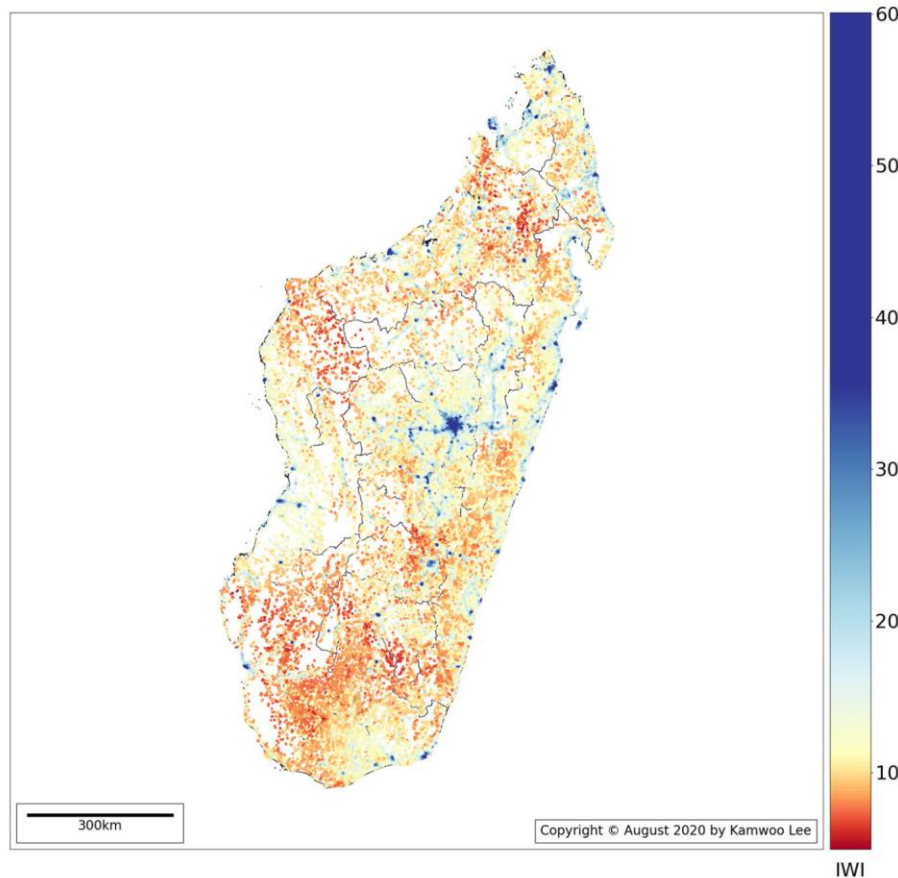
- Les données satellitaires sont mobilisées pour calculer 30/231 indicateurs ODD et elles renseignent utilisation 71 des 169 cibles ODD ([GEO 2019](#); [Estoque 2020](#)).

Relation aux autres sources

Besoins de croisement

- Relevés sur le terrain
- Données administratives
- Données contributives
- Fusion de données : pauvreté, biodiversité, densité de population

Exemple de fusion de données



Cartes de pauvreté ([Lee and Braithwaite 2022](#))

Combinent des données :

- Enquêtes DHS
- Satellitaire (VIIRS, Google SM)
- Population (HRLS)
- OpenStreetMap

Intelligence artificielle (XGBoost et CNN), assez fiable ($R^2 = 89\%$)

Enorme potentiel d'utilisation

Risques toutefois : quantification des incertitudes, (més)usages

Facteurs favorisant l'essor

- Disponibilité gratuite
- Couverture mondiale, même là où peu d'autres données
- Accès au terrain (COVID, conflits) ([Yokoi, Vaessen, and Vandecasteele 2020](#))
- Fréquences de mise à jour
- Accessibilité des infrastructures et systèmes de traitement
- Disponibilité de données pré-processées
- Matériel didactique
- Initiatives de soutien (ex. Space for International Development Assistance)

Risques et dérives

- Risque méthodologique : mésusages des données, traitements ou interprétation
- Risque épistémologique : métaphore du lampadaire
- Risque ontologique : quelle est la nature même des enjeux homme-nature qu'on étudie (conservation "moderne" vs. savoirs vernaculaires) (Culas and Veriza 2022)
- Risque sociopolitique : cybertariat, solutionnisme (Fourcade and Gordon 2020)
- Risque techno-économique : place centrale de Google Earth Engine dans ce phénomène
- Economie politique : le terrain à distance