L'apport et les enjeux de l'analyse multiniveau en démographie

Valérie Golaz

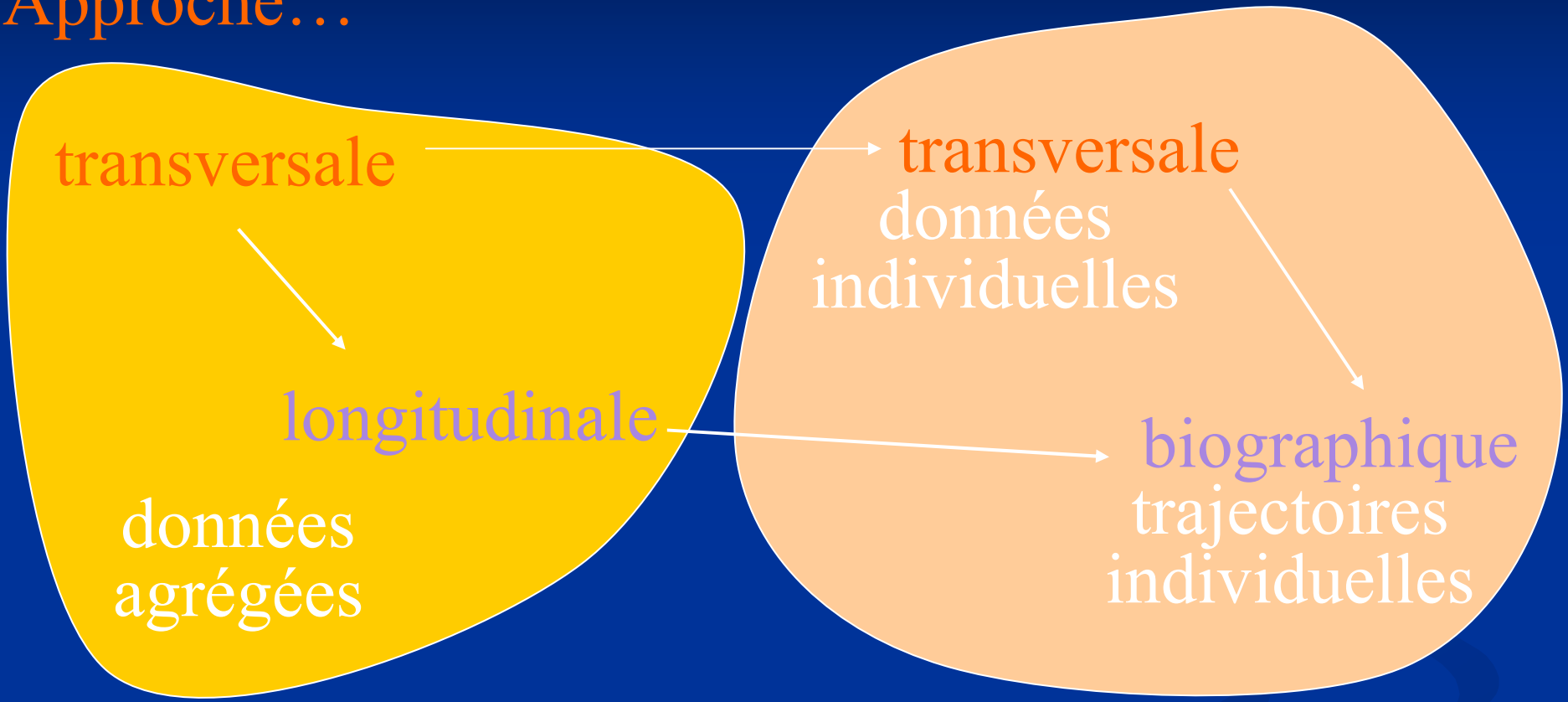
INED-CEPED

Plan de la présentation

- Qu'est-ce que l'analyse multiniveau ?
- 3 exemples :
  - Les migrations norvégiennes
  - La non-scolarisation des enfants au Kenya
  - L'évaluation de la vie en grandes périodes
- Apports et enjeux pour la démographie

Différentes approches se succèdent…

Approche…



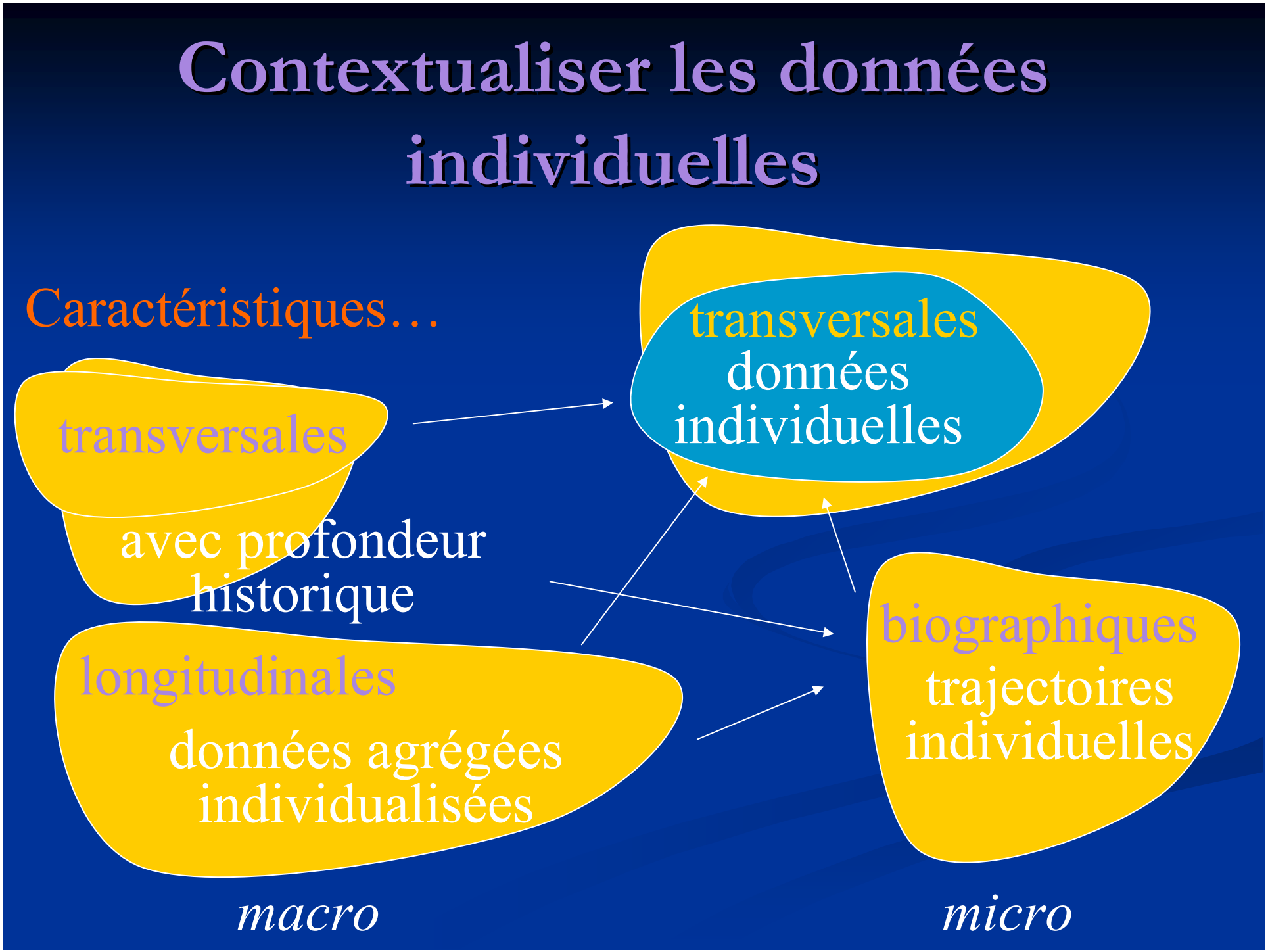
macro

micro

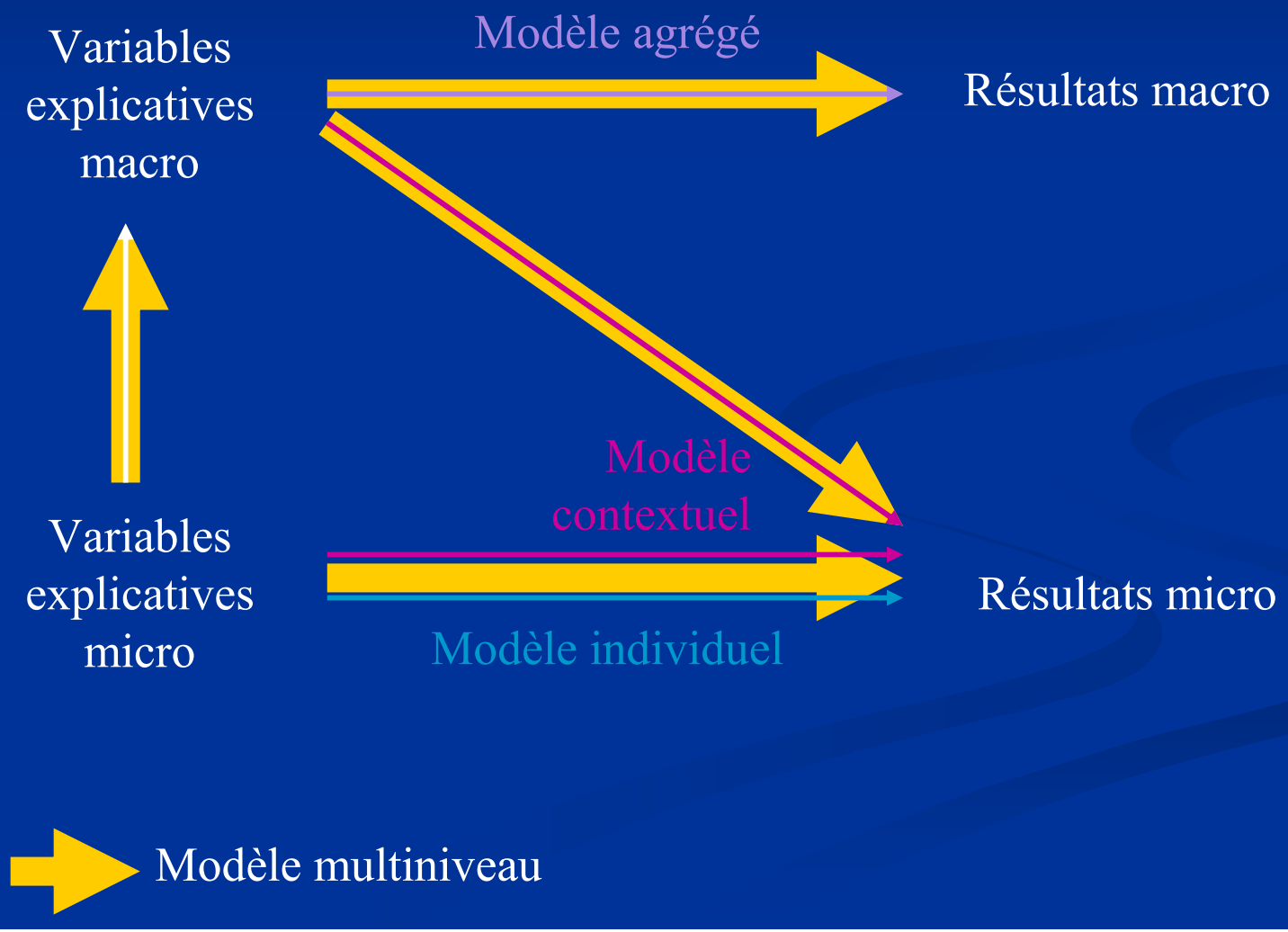
! Erreur écologique !

! Erreur atomiste !

Contextualiser les données individuelles



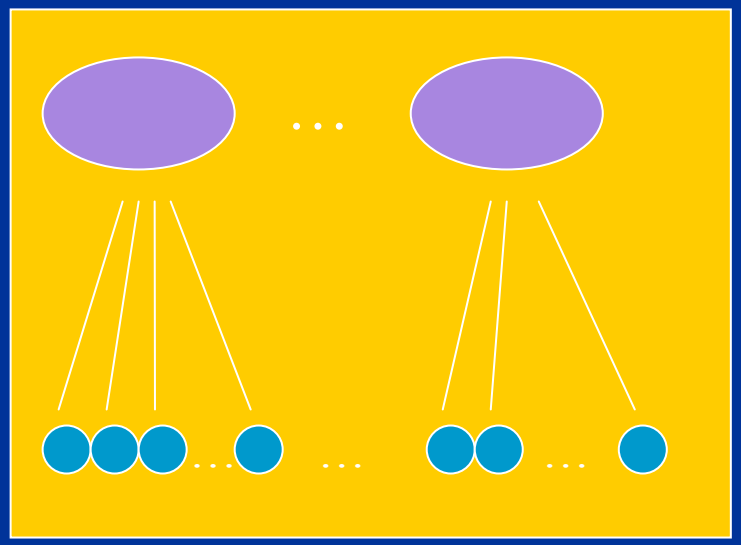
Modèles à différents niveaux



Définir le niveau d'agrégation à utiliser…

Quelques exemples courants

Zones administratives



Individus

<u>Micro / Macro</u>

Patients / Hôpital

Enfants / École

Électeurs / Circonscription

…

Individu / Famille

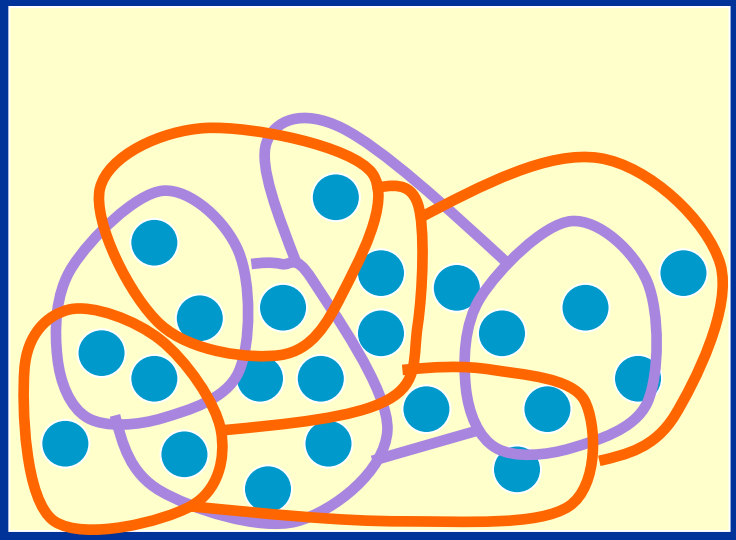
Evénement / Trajectoire

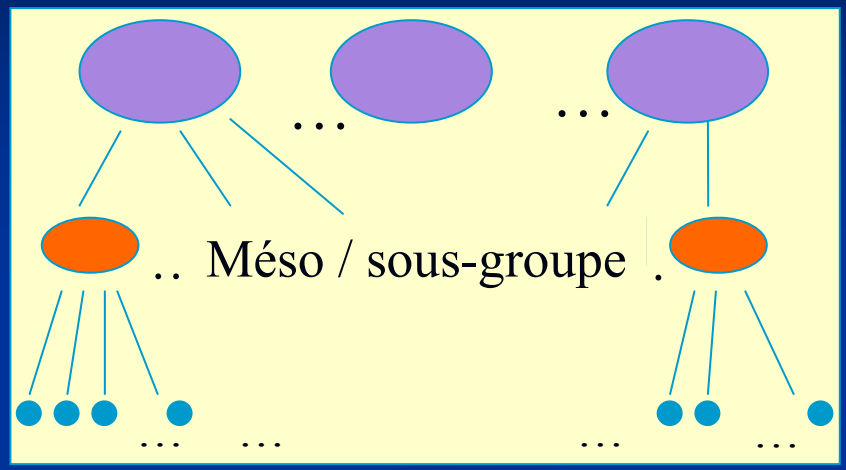
Employé/Entreprise

…

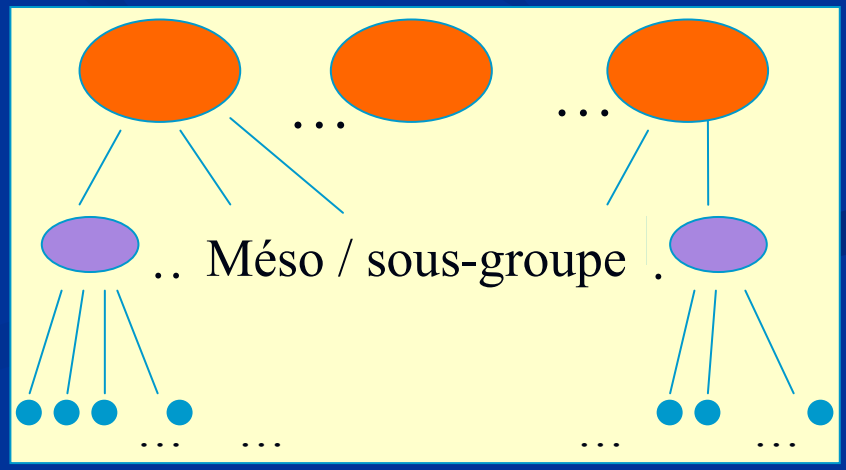
Définir les niveaux d'agrégation à utiliser…

Différents types de contextes peuvent être importants





Entrelacement…



L'apport du multiniveau

Structurer les données sur différents niveaux permet la décomposition en effets fixes et aléatoires

Dans un modèle statistique, il y a toujours une part d'inobservé, une partie de la réalité qui n'est pas explicitée

En dissociant dans un modèle les caractéristiques à différents niveaux d'observation, on peut percevoir de manière plus fine cette « hétérogénéité non observée » : on obtient une mesure de la variance par niveau.

Spécificités de l'analyse multiniveau

Modèle logit

Variables au niveau 1 (V) et au niveau 2 (C)

$$P(Yij = 1) = (1 + \exp-[\underbrace{\alpha_0 + \alpha_1 V_{1ij} + \alpha_2 V_{2ij} + \ldots + \alpha_n V_{nij}}]) ^{-1}$$

$$\alpha_0 + u_{0j} + \alpha_1 V_{1ij} + \alpha_2 V_{2ij} + \ldots + \alpha_n V_{nij}$$

avec $Var(u_{0j}) = \sigma^2 u_0$

A quoi cela sert-il ?

Quelle part de l'hétérogénéité se place au niveau 1 et au niveau 2 ?

- entre individus au sein de chaque groupe
- ou entre groupes ?

⇒ Identifier et réduire la variance entre groupes

Toutes choses égales par ailleurs, quelles caractéristiques contextuelles vont expliquer les différences entre groupes ?

Quels sont les types de contextes qui entrent en jeu ?

Un objectif et trois modes d'action…

Objectif : (toujours le même) réduire la part d'hétérogénéité non observée dans le modèle

- En jouant sur la définition des groupes
- En jouant sur la modélisation statistique des effets fixes et aléatoires
- En cherchant les variables explicatives pertinentes

Un exemple classique: Les migrations norvégiennes

Données de registres / recensements

28462 hommes nés en 1948 résidant en Norvège en 1991, qui n'ont jamais vécu hors du pays

19 régions

Objet : étude des changements de région dans les 2 années qui suivent le recensement de 1970,

Avec une variable : être agriculteur ou non

source : Courgeau, 2002

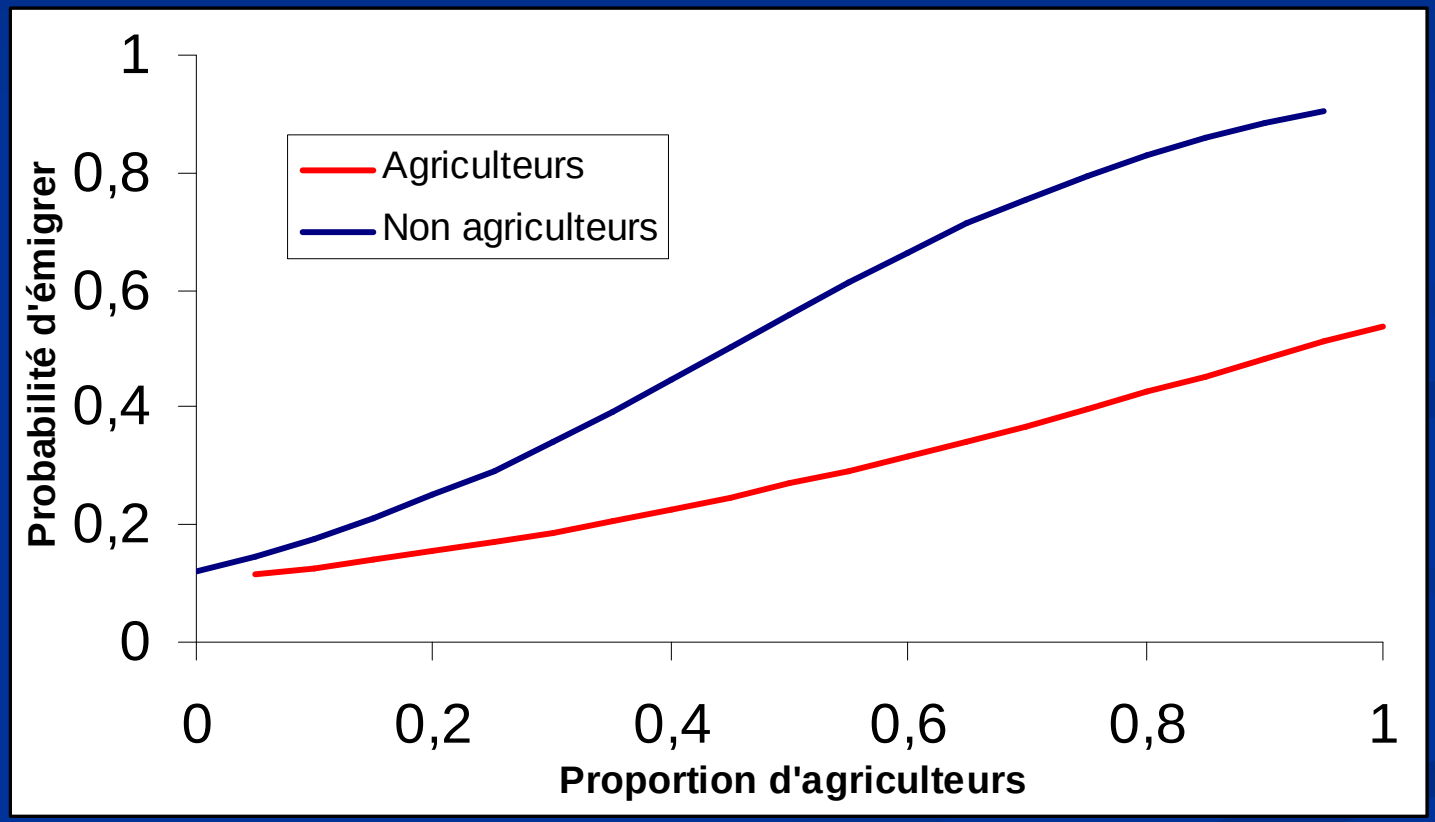
Modèle contextuel avec interaction

- la proportion d'agriculteurs par région.
- effet sur les agriculteurs et les non agriculteurs

$$P(\mu_{ij} = 1 \mid a_{ij}) = (1 + \exp-[\ \alpha_0 (1-a_{ij}) + \alpha_1 a_{ij} + \alpha_2 a_{.j} + \alpha_3 a_{ij} a_{.j}\ ])^{-1}$$

Paramètres	Modèle c	Modèle ci
non agric.	-1,984 (0,032)	-1,996 (0,033)
agric.	-2,614 (0,093)	-2,155 (0,209)
% agric	4,266 (0,453)	4,469 (0,461)
agric x %agric		-5,774 (2,447)

Les migrations norvégiennes: Résultats de la régression logistique, niveau individuel, modèle contextuel



Source : D.Courgeau, à partir du registre de population norvégien

Modèle multiniveau

$$P(\mu_{ij} = 1 \mid a_{ij}) = (1 + \exp-[\ (\alpha_0 + u_{0j})\ (1-a_{ij}\ ) + (\alpha_1 + u_{1j})\ a_{ij}\ ])^{-1}$$

$$\text{var}\ (u_{0j}\ ) = \sigma^{2}_{u0}$$

$$\text{var}\ (u_{1j}\ ) = \sigma^{2}_{u1}$$

$$\text{cov}\ (u_{0j}\ , u_{1j}\ ) = \sigma_{u01}$$

Résultats

	simple	c	ci
non agric.	-1,710 (0,070)	-2,150 (0,110)	-2,067 (0,119)
agric.	-2,306 (0,133)	-2,786 (0,200)	-2,067 (0,340)
% agric		6,654 (0,989)	5,420 (1,209)
agric x %agric			-8,691 (3,238)
σ^2_{u0}	0,088 (0,032)	0,049 (0,025)	0,047 (0,024)
σ_{u01}	0,054 (0,044)	0,104 (0,068)	0,085 (0,042)
σ^2_{u1}	0,167 (0,135)	0,312 (0,238)	0,181 (0,119)

Le contexte régional a-t-il un effet sur la scolarisation des enfants en milieu rural kenyan ?

Données IPUMS-Int du recensement de 1999

160 531 enfants de 10 à 14 ans dont 18 % ne sont pas scolarisés

67 districts ayant une partie rurale

Variables explicatives niveau individu, ménage, district

La non-scolarisation - 0

		Modèle 0		
	Fixes	**Coeff**	**SE**	**Pr > |t|**
Individu	constante	-1,58	0,15	<,0001
	Aléatoires	**Coeff**	**SE**	**Pr > |t|**
	distke	1,58	0,28	<,0001

La non-scolarisation - 1

		Modèle 1		
	Fixes	**Coeff**	**SE**	**Pr > |t|**
Individu	constante	-1,58	0,15	<,0001
	genre=garçon	0,04	0,02	0,0058
	relation=enf du CM	0,00	0,00	<,0001
Ménage	taille ménage petite	-0,08	0,00	<,0001
	niveau d'instr CM < primaire	-0,01	0,00	<,0001
	Aléatoires	**Coeff**	**SE**	**Pr > |t|**
	distke	1,41	0,25	<,0001

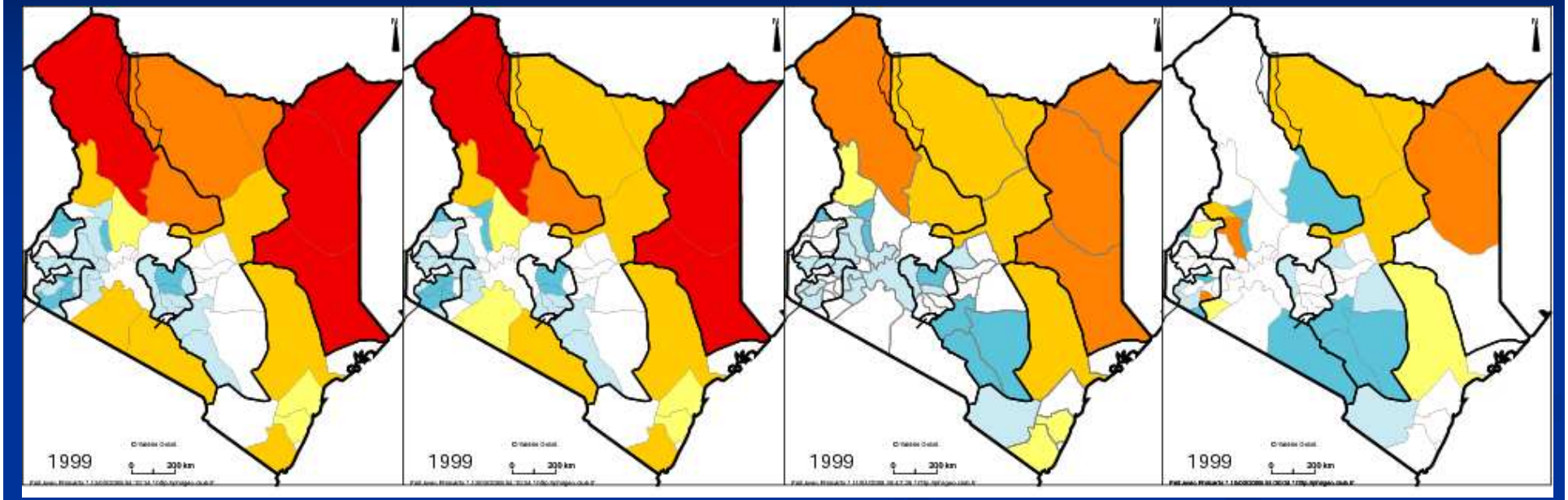
La non-scolarisation - 2

		Modèle 2		
	Fixes	**Coeff**	**SE**	**Pr > |t|**
Individu	constante	-0,32	0,18	0,0895
	genre=garçon	0,05	0,02	0,0018
	relation=enf du CM	0,00	0,00	<,0001
Ménage	taille ménage petite	-0,08	0,00	<,0001
	niveau d'instr CM < primaire	-0,01	0,00	<,0001
	genre du CM = homme	-0,23	0,02	<,0001
District	Densité de population du district	-0,04	0,01	<,0001
	Indice d'emploi (formel)	-0,04	0,01	<,0001
	Aléatoires	**Coeff**	**SE**	**Pr > |t|**
	distke	1,07	0,19	<,0001

La non-scolarisation - 3

		Modèle 3		
	Fixes	**Coeff**	**SE**	**Pr > |t|**
Individu	constante	-0,23	0,27	0,3932
	genre=garçon	0,05	0,02	0,0024
	relation=enf du CM	0,00	0,00	<,0001
Ménage	taille ménage petite	-0,08	0,00	<,0001
	niveau d'instr CM < primaire	-0,01	0,00	<,0001
	genre du CM = homme	-0,23	0,02	<,0001
District	Densité de population du district	-0,04	0,01	<,0001
	Indice d'emploi (formel)	-1,47	0,53	0,0052
Mén*Dist	Densité* niv inst CM	0,37	0,03	<,0001
	Aléatoires	**Coeff**	**SE**	**Pr > |t|**
	distke	0,90	0,16	<,0001

Représentation des résidus



Modèle vide

Modèle individuel

Modèle contextuel

Modèle contextuel avec interaction

L'évaluation de la vie passée en grandes périodes

D'après l'enquête Biographies et entourage (INED, 2001)

⇒ les biographies de 2830 individus résidents en Ile de France

⇒ qui ont découpé leur vie en 1 à 12 périodes, qu'ils ont qualifiées (TD, D, SP, B, TB)

Les périodes difficiles

		Coeff.	S.
	constante	-0,63	***
Sexe (ref = femme)	Hommes	-0,24	***
Année de naissance (ref = 1930-1935)	1946-1950	-0,18	**
	1941-1945	-0,13	*
	1936-1940	0,01	
Lieu de naissance (ref= Etranger)	Ile de France	-0,20	**
	Province	-0,24	***
Début de période… (ref=à partir de 50 ans)	Pendant l'enfance ou l'adolescence (0-20 ans)	0,37	***
	Pendant l'âge actif (21 - 49 ans)	0,46	***
Durée de la période (ref = moyenne)	courte	0,61	***
	longue	-0,83	***
Variance de la taille du ménage au cours de la période (ref = nulle)	forte	-0,06	
	moyenne	-0,19	*
Evolution de la taille du ménage (ref = non monotone)	croissante	-0,19	**
	décroissante	0,12	
Zone d'enquête (ref = Paris)	Grande couronne	-0,17	**
	Petite couronne	0,05	

Construire un modèle multiniveau

- Quels individus ?

Ici, l'individu statistique est la période

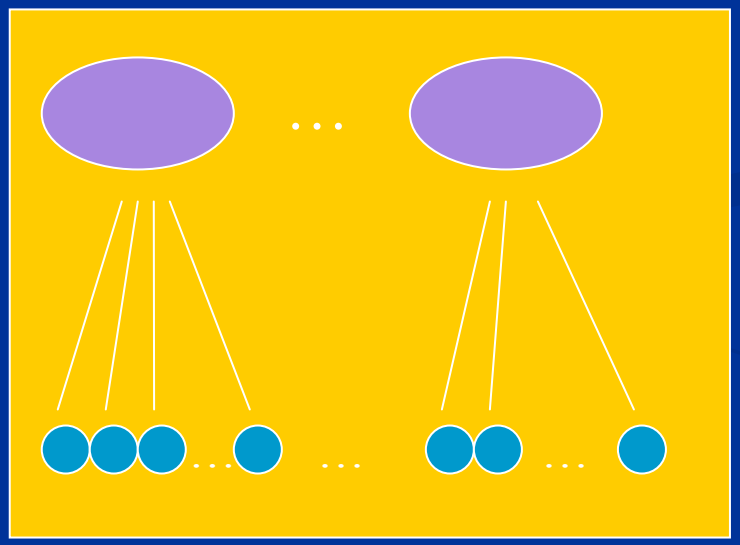
- Quels groupes ?

Les individus ? Les zones géographiques ? Des cohortes ? Croiser ces informations ?

- Quelle modélisation des résidus ?

Des groupes particuliers

Age au début de la période (0-20 / 21-49 / 50+)
* Département d'enquête (75 / 78 / 92 / 93 / 94 / 95)



24 groupes,

Avec 104 à 1167 périodes

Périodes (11905 périodes)

Un modèle logistique à 2 niveaux

avec effet aléatoire sur la constante

$$P(Eval<0) = (1 + \exp-[a + b\,X + \mathbf{u_j}])^{-1}$$

$$\text{var}(u_j) = \sigma^2_{u0}$$

Des effets fixes relativement stables…

		Modèle simple		2 niveaux Pér*Dep	
		Coeff.	S.	Coeff.	S.
	constante	-0,63	***	-0,43	***
Sexe (ref = femme)	Hommes	-0,24	***	-0,24	***
Cohorte (ref = 1930-1935)	1946-1950	-0,18	**	-0,15	*
	1941-1945	-0,13	*	-0,11	
	1936-1940	0,01		0,03	
Lieu de naiss. (ref= Etranger)	Ile de France	-0,20	**	-0,19	**
	Province	-0,24	***	-0,24	***
Début de période… (ref=à	Enf. ou l'adoles.	0,37	***		
	Age actif	0,46	***		
Durée de la période (ref =	courte	0,61	***	0,59	***
	longue	-0,83	***	-0,82	***
Variance de la taille du ménage	forte	-0,06		-0,03	
	moyenne	-0,19	*	-0,17	*
Allure (ref = non monotone)	croissante	-0,19	**	-0,18	**
	décroissante	0,12		0,11	
Zone d'enquête (ref = Paris)	Grande	-0,17	**		
	Petite couronne	0,05			

Des effets aléatoires significatifs

	2 niveaux Pér*Dep		
Variance inter-groupes ($\sigma^2 u_0$) (vide=0,95***)		0,04	**
Groupes aux résultats significativement différents de la moyenne (u_{0j} significatifs)	0-20 ans / 75	-0,19	*
	0-20 ans / 93	-0,18	*
	21-49 ans / 92	-0,24	**
	21-49 ans / 93	-0,36	***
	50 ans et + / 78	0,33	*

Conclusion des exemples

Se rapprocher le plus possible de la réalité

Modéliser la complexité

Guider la modélisation

=>Identifier les niveaux d'observations pertinents et les caractéristiques importantes à chaque niveau

=>Travailler sur des données moins nombreuses que si l'on considérait les groupes séparément les uns des autres

Mais… faire face à encore beaucoup d'écueils

Avantages

Analyse => identifier les niveaux, modéliser, réduire la variance

MAIS… COMPLEXITE MAL PRISE EN CHARGE

Collecte => envisager la collecte de données moins nombreuses

MAIS… HARMONISATION DES COLLECTES INTERNATIONALES

Valoriser des sources de données différentes

MAIS… TRES DIFFICILE DE TROUVER LES SOURCES ADEQUATES

Conclusion

Un outil de valorisation / d'analyse de données existantes qui peut nécessiter un travail de recueil de données important

Envisager des collectes de données contextuelles dans le projet d'enquête individuelle / ménage ?