Articuler données individuelles et données exogènes grâce aux méthodes multiniveaux:

De nouvelles possibilités mais toujours des écueils

Valérie Golaz (Ined / LPED)
Fabrice Boyam Yameogo (Aix-Marseille Université / LPED)

Plan

1/ Ce qu'est un modèle multiniveau, enjeux et limites

2/ La disponibilité des données : une révolution internationale qui permet d'envisager de nouvelles analyses

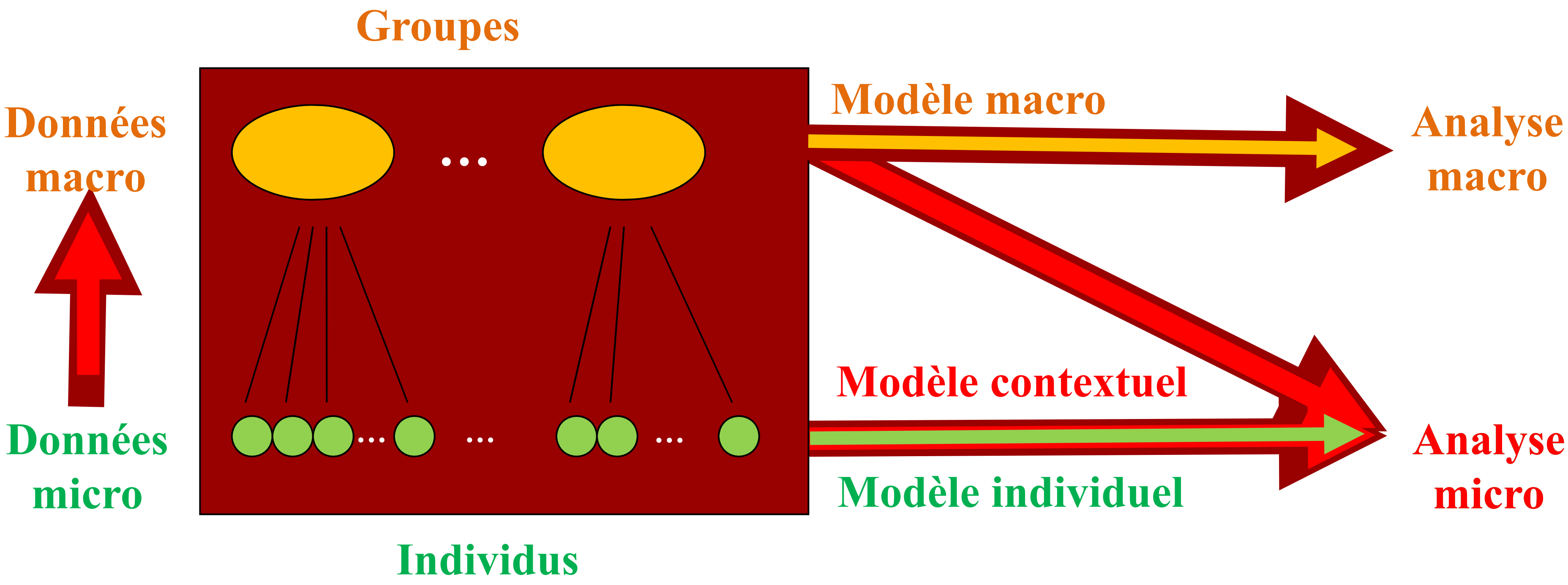
- données administratives
- données spatiales / issues de l'imagerie satellite

3/ Les écueils toujours présents:

- la difficulté d'accéder à des découpages administratifs/spatiaux fins
- la question de la temporalité des découpages et des informations disponibles (décalage / sources disparaissent au fil du temps, se périment…)

4/ Mais des résultats qui justifient l'effort: analyse multiniveau de la déscolarisation au Kenya

L'approche multiniveau, au croisement des sources statistiques



=>Un modèle unique évalue les effets fixes et aléatoires liés à différents niveaux d'analyse, en utilisant éventuellement des sources différentes

Spécificité

Introduction d'un effet aléatoire entre unités agrégées : on autorise un effet différencié de ces unités, ou de certaines variables entre chaque unité.

Par exemple:

$$p_{ij} = \frac{1}{1+\exp^{-[\beta_{00}+\beta_1 X_{ij}+(u_{0_j})]}}$$

Le terme aléatoire u_{0j} suit une loi normale d'espérance nulle et de variance $\sigma^2_{u_0}$

⇒ C'est un modèle à constante aléatoire

Avantages et inconvénients

Evaluer effets fixes et aléatoires à différents niveaux dans un seul modèle …

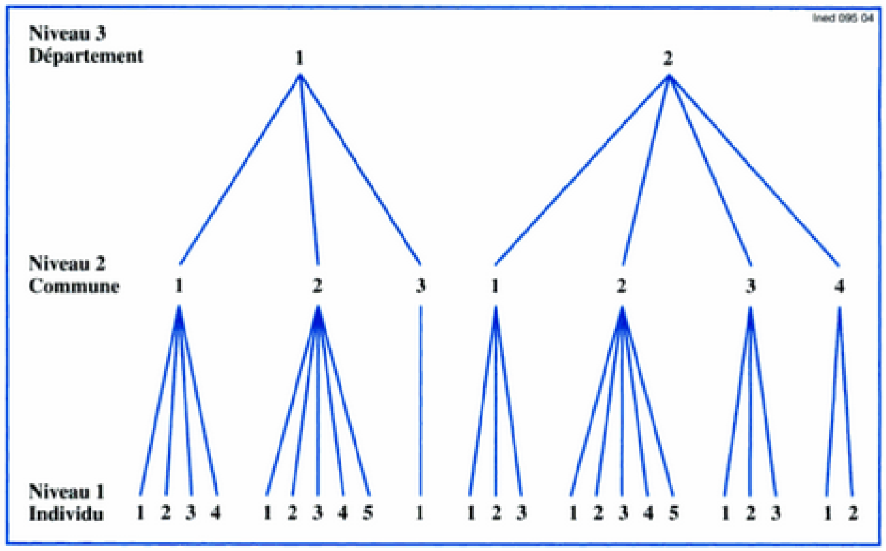
Avantages

- … ne surestime pas l'effet des variables contextuelles et permet de mieux mesurer leurs rôles respectifs dans le phénomène étudié,
- La répartition de la variance sur différents niveaux permet de savoir à quel niveau le modèle peut être complété

Limites

- Il est parfois difficile d'affiner l'analyse autant que l'on aimerait:
  - → niveaux pertinents
  - → variables
  - → le modèle lui-même
- Effets de frontière / mobilité
- Trouver les données pertinentes reste difficile

Des modèles particulièrement pertinents pour analyser des données structurées par zones administratives



Source: Courgeau, 2004

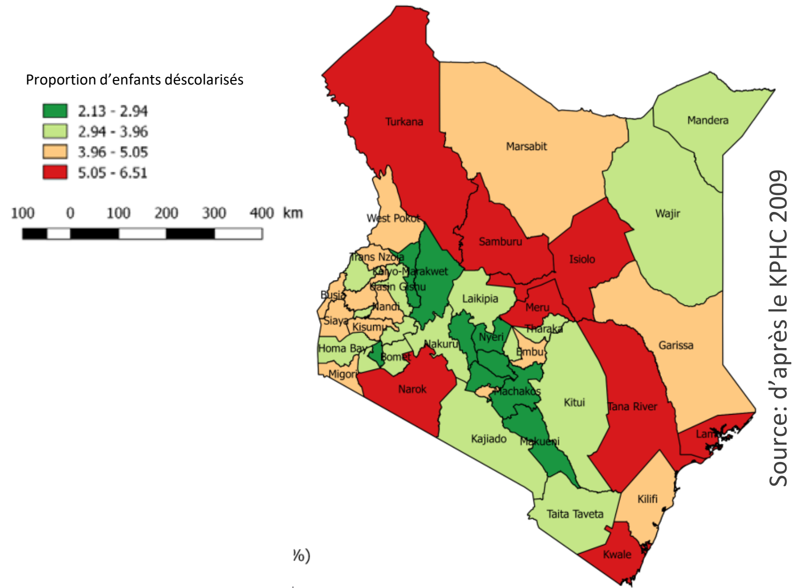
L'approche multiniveau permet alors de mobiliser des données individuelles et des données agrégées issues d'autres sources (données administratives, données d'enquêtes agrégées, données issues de l'analyse spatiale, …)

=> C'est particulièrement utile dans des pays où l'on dispose de moins de données d'enquêtes (mais aussi dans les autres!)

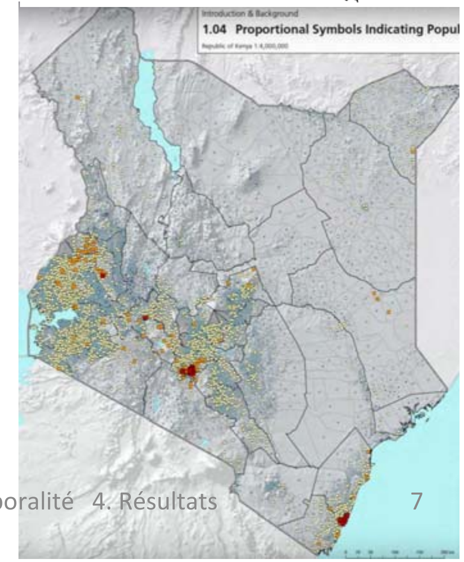
Etudier la déscolarisation des enfants au Kenya et en Ouganda

- Un problème encore prégnant à l'école primaire malgré les politiques de scolarisation universelle en vigueur (Ouganda, 1997; Kenya, 1973)
- Des différences régionales marquées et encore mal comprises
- Un sujet peu traité (enfants hors l'école = non scolarisation + dé-scolarisation) adapté à une analyse multiniveaux
- Des données abondantes

=> Est-ce lié aux caractéristiques des individus et de leur famille, ou bien à l'environnement plus général dans lequel ils se trouvent (qualité de l'offre scolaire, distance des écoles, pression sociale, …)?



Source: d'après le KPHC 2009



Source : Atlas socio-économique 2016

Des données de recensement de plus en plus accessibles…

ined

IPUMS INTERNATIONAL

HOME | SELECT DATA | MY DATA | FAQ | HELP

DATA CART
YOUR DATA EXTRACT
0 VARIABLES
0 SAMPLES

SAMPLE INFORMATION

Argentina	1970 · 1980 · 1991 · 2001 · 2010	Hungary	1970 · 1980 · 1990 · 2001 · 2011	Peru	1993 · 2007
Armenia	2001 · 2011	Iceland	1703 · 1729 · 1801 · 1901 · 1910	Philippines	1990 · 1995 · 2000
Austria	1971 · 1981 · 1991 · 2001 · 2011	India	1983 · 1987 · 1993 · 1999 · 2004 · 2009	Poland	1978 · 1988 · 2002 · 2011
Bangladesh	1991 · 2001 · 2011	Indonesia	1971 · 1976 · 1980 · 1985 · 1990 · 1995 · 2000 · 2005 · 2010	Portugal	1981 · 1991 · 2001 · 2011
Belarus	1999 · 2009	Iran	2006 · 2011	Puerto Rico	1970 · 1980 · 1990 · 2000 · 2005 · 2010
Bolivia	1976 · 1992 · 2001	Iraq	1997	Romania	1977 · 1992 · 2002 · 2011
Botswana	1981 · 1991 · 2001 · 2011	Ireland	1971 · 1979 · 1981 · 1986 · 1991 · 1996 · 2002 · 2006 · 2011	Rwanda	1991 · 2002
Brazil	1960 · 1970 · 1980 · 1991 · 2000 · 2010	Israel	1972 · 1983 · 1995	Saint Lucia	1980 · 1991
Burkina Faso	1985 · 1996 · 2006	Italy	2001	Senegal	1988 · 2002
Cambodia	1998 · 2008	Jamaica	1982 · 1991 · 2001	Sierra Leone	2004
Cameroon	1976 · 1987 · 2005	Jordan	2004	Slovenia	2002
Canada	1852 · 1871 · 1881 · 1891 · 1901 · 1911 · 1971 · 1981 · 1991 · 2001 · 2011	Kenya	1969 · 1979 · 1989 · 1999 · 2009	South Africa	1996 · 2001 · 2007 · 2011
Chile	1960 · 1970 · 1982 · 1992 · 2002	Kyrgyz Republic	1999 · 2009	South Sudan	2008
China	1982 · 1990 · 2000	Liberia	1974 · 2008	Spain	1981 · 1991 · 2001 · 2011
El Salvador	1992 · 2007	Nepal		Uganda	1991 · 2002
Ethiopia	1984 · 1994 · 2007	Netherlands	1960 · 1971 · 2001	Ukraine	2001
Fiji	1966 · 1976 · 1986 · 1996 · 2007	Nicaragua	1971 · 1995 · 2005	United Kingdom	1851a · 1851b · 1851c · 1861a · 1861b · 1871b 1891a · 1891b · 1901a · 1901b · 1911 · 1991 · 2
France	1962 · 1968 · 1975 · 1982 · 1990 · 1999 · 2006 · 2011	Nigeria	2006 · 2007 · 2008 · 2009 · 2010	United States	1850a · 1850b · 1860 · 1870 · 1880a · 188 1960 · 1970 · 1980 · 1990 · 2000 · 2005 ·

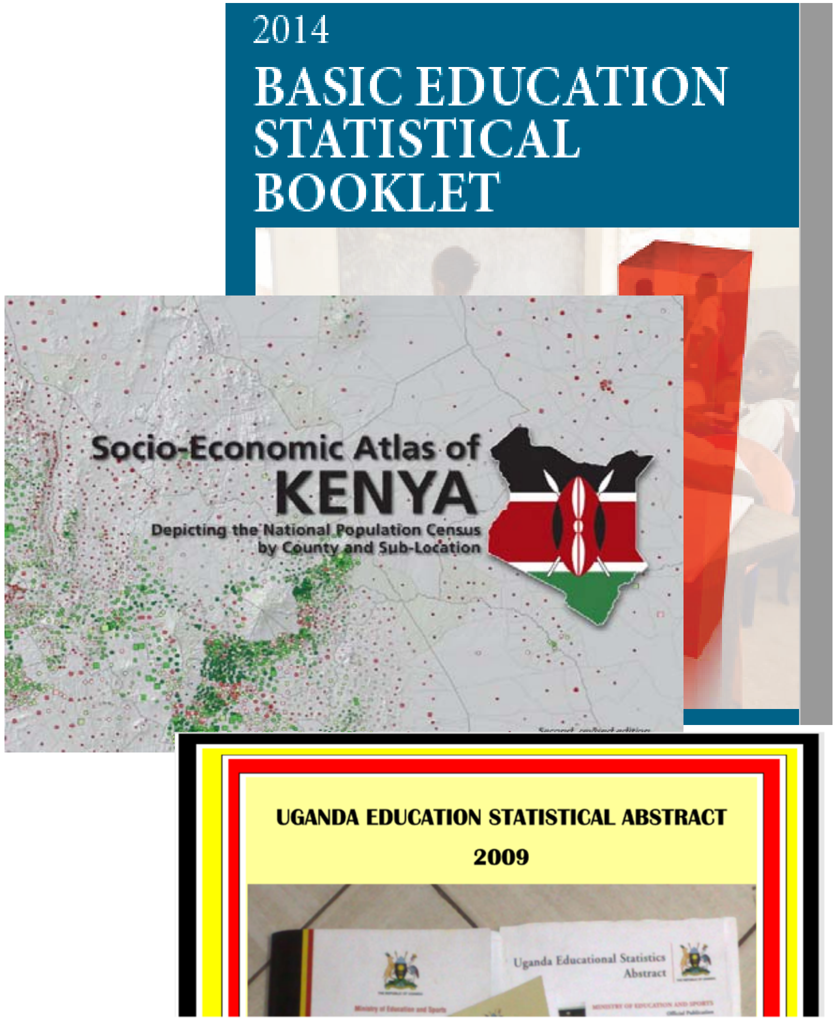
Des données administratives de plus en plus accessibles également

Kenya

- Le [recensement des écoles et leurs caractéristiques](http://www.opendata.go.ke/)de 2007 sur http://www.opendata.go.ke/
- Le livret statistique sur l'éducation de base en 2014 (2015)
- L'Atlas Socio-économique du Kenya (2016)

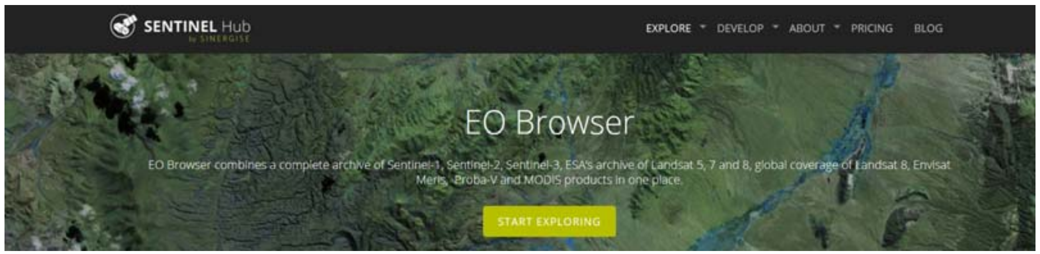
Ouganda

- Le recensement scolaire annuel (2014)
- Résultats aux examens (PLE, UCE)
- *Uganda Education Statistical Abstract* 2005, 2009, 2010…



+ Les *statistical abstracts* annuels

D’autres sources de données qui s’ouvrent



- Les images Sentinel 1 et 2 apportent des visuels précis à 10 m près, pris tous les 3 à 5 jours, de l’ensemble de la planète

⇒ possibilité d’en tirer des indicateurs précieux, en particulier dans le domaine de l’environnement

- Des images régulières depuis les années 1970

Des difficultés toujours présentes

Internet et l'archivage

- la non pérennité de l'accès à certaines données

Une question de temporalité

- La question de la date dans l'appariement

Une question d'échelle spatiale

- la difficulté d'accéder à des découpages administratifs/spatiaux plus fins / plus pertinents pour l'analyse dans l'ensemble des données
- Le difficile appariement des données dans un contexte où de nouvelles entités administratives sont créées régulièrement

Kenya

- Le recensement des écoles et leurs caractéristiques de 2007 sur http://www.opendata.go.ke/
- Le livret statistique sur l'éducation de base en 2014 (2015)
- L'Atlas Socio-économique du Kenya (2016)

Ouganda

- Le recensement scolaire annuel (2014)
- Résultats aux examens (PLE, UCE)
- *Uganda Education Statistical Abstract* 2005, 2009, 2010…

+ Les *statistical abstracts* annuels

La temporalité du zonage administratif

Kenya: du district au county

- 1969: 41 districts
- 1979: 41 districts
- 1989: 41 districts
- 1999: 69 districts
  - → 2007: 71 districts
- 2009: 158 districts

2010: Nouvelle constitution, introduction des *counties*

→ 47 *counties*

(mais 7 149 *sub-locations*)

Ouganda: *county* et districts

- 1969: 21 districts, 111 *counties*
- 1980: 33 districts, 140 *counties*
- 1991: 38 districts, 163 *counties*
  - → 1994: 39 districts, 163 counties
- 2002: 56 districts, 163 *counties*
  - → 2006: 80 districts
  - → 2011: 112 districts
- 2014: 112 districts, 181 *counties*

Les enfants déscolarisés au Kenya: modélisation

Population: Enfants de 6 à 17 ans, qui ont déjà été scolarisés en cycle primaire mais n'ont pas dépassé ce niveau

Nous nous intéressons à ceux de ces enfants qui ne sont pas scolarisés au moment du recensement. Ils n'ont donc pas encore achevé le cycle primaire (qui comprend 7 années).

Modèles logistiques / modèles multiniveaux

=> La déscolarisation des enfants en cours de cycle primaire est-elle liée aux caractéristiques des individus et de leur famille, ou bien à l'environnement plus général dans lequel ils se trouvent (qualité de l'offre scolaire, distance des écoles, pression des pairs, …)?

Dans un premier temps, modèles logistiques, puis modèles multiniveaux à 2 niveaux (individu / county) avec un aléa sur la constante.

Les enfants déscolarisés au Kenya: variables individuelles

D'après la littérature (UEPA, 1999):

- les caractéristiques de l'enfant : sexe, âge, statut d'orphelin, lien de parenté avec le chef de ménage;
- les caractéristiques du ménage : nombre de personnes habitant dans le ménage, milieu de résidence (urbain-rural), la zone de résidence (*county*) ;
- les caractéristiques du chef de ménage: sexe, âge, statut matrimonial, niveau d'éducation, activité professionnelle

Les enfants déscolarisés au Kenya: variables contextuelles endogènes

- Les variables contextuelles issues du recensement: la proportion de chefs de ménage ayant au moins achevé le primaire dans le *county*, la proportion de chefs de ménage en emploi salarial dans le *county*
- Méthode de construction:

  regroupement des chefs de ménage de 20-59 ans

  dichotomisation des variables (niveau d’instruction, activité professionnelle)

  calcul de la moyenne de ces variables par *county*

  variables centrées et réduites

  fusion avec les données individuelles

Les enfants déscolarisés au Kenya: variables contextuelles exogènes

- les variables contextuelles issues d'autres sources de données:

* **caractéristiques économiques du *county* (source: Atlas de 2016, croisement de données de recensement de 2009 et de l'enquête budget des ménages de 2005-2006)** : dépenses moyennes mensuelles par personne, incidence de la pauvreté, intensité de la pauvreté, proportion de personnes qui travaillent dans le secteur informel, dans l'agriculture et dans le secteur formel ;

* **caractéristiques de l'offre scolaire (source: recensement des écoles de 2014)** : taille moyenne des écoles, taille moyenne des classes, ratio maitre/élèves et densité d'écoles primaires

Résultats (régression logistique)

ined

	M0 Modèle simple	M1 Modèle avec car. ménage et districts	M2 Modèle avec car. ménage et car. contextuelles
Intercept	- - -***	- - -***	- - -***
Sexe : Femme	+***	+***	+***
Age : 13-17 ans	+ +***	+ +***	+ +***
Orphelins : au moins 1 parent décédé	+***	+***	+***
Lien : pas l'enfant du chef de ménage	+ +***	+ +***	+ +***
Taille de ménage : ≥6		-***	-***
Age du CM : <60 ans		+***	+***
Sexe du CM : Femme		-***	-***
Niv. Inst. du CM : <fin du secondaire		+***	+***
Profession du CM : pas salarié		-***	-***
Stat. Mat du CM : pas marié monogame		-***	-***
Milieu : rural		-***	-***
County : pas Nairobi		+/- ***	
% de CM ayant achevé le primaire			-***
Incidence de la pauvreté			- -***
Densité d'écoles			+ ***
Taille moyenne des écoles			+***

Note : seuils de significativité : *p<0,05 ; **p<0,01 ; ***p<0,001 ; ns=non significatif.
Source : KPHC 2009, MOES 2014, KNBS 2016 (estimation des auteurs, proc logistic et proc glimmix de SAS®)

Passage au multiniveau

- Seul changement: le statut matrimonial du CM n'est plus significatif
- Pas de changement majeur des rapports de côtes pour toutes les autres caractéristiques de l'individu, du ménage et le milieu de résidence,
- Constance de ces rapports de côtes dans la modélisation multiniveau
- Réduction de la variance contextuelle: de 0,084***(modèle vide) à 0,055*** (avec car ind et mén et milieu de résidence), puis 0,048*** (avec car ind et mén et var cont endo) et 0, 034*** (avec car ind et mén et var cont endo et exo)

	M2 Modèle logistique avec car. ménage et car. contextuelles	M3 Modèle multiniveau avec car. ind, mén et milieu de résidence	M5 Modèle multiniveau avec car. ind, ménage et car. contextuelles
Intercept	0,018***	0,009***	0,015***
Sexe : Femme	1,099***	1,1***	1,1***
Age : 13-17 ans	3,727***	3,7***	3,7***
…			
Taille de ménage : ≥6	0,888***	0,887***	0,886***
…			
Milieu : rural	0,919***	0,932***	0,933***

Note : seuils de significativité : *p<0,05 ; **p<0,01 ; ***p<0,001 ; ns=non significatif.

Source : KPHC 2009, MOES 2014, KNBS 2016 (estimation des auteurs, proc logistic et proc glimmix de SAS®)

Résultats (logistique et multiniveau)

	M2’ Meilleur Modèle logistique	M6 Modèle multiniveau équivalent	M5 Modèle multiniveau simplifié
% de CM ayant achevé le primaire	0,689***	0,715***	0,771***
Incidence de la pauvreté	0,264***	0,352**	0,361**
Densité d’écoles	1,063***	1,073	1,103***
Taille moyenne des écoles	1,109***	1 ,065	1,05
% des CM en emploi salarial	0,882***	0,987	
% des individus en emploi dans le secteur formel	6,735***	2,859	
Ratio maitre/élève	0,928***	0,979	
Effet aléatoire (Intercept)		0,3438***	0,3623***

Note : seuils de significativité : *p<0,05 ; **p<0,01 ; ***p<0,001 ; ns=non significatif.
Source : KPHC 2009, MOES 2014, KNBS 2016 (estimation des auteurs, proc logistic et proc glimmix de SAS®)

⇒ le modèle logistique surestime les effets des var. contextuelles: Les variables relatives à la qualité de l’offre scolaire ne sont pas significatives, seule l’accessibilité de l’école (distance / densité) compte. L’explication de la déscolarisation se situe entre les variables liées à l’individu et à son entourage immédiat et à la répartition des écoles sur le territoire.

- Un tri possible entre les variables contextuelles

Limites et perspectives

- Question des dates dans les données transversales et les appariements

⇒ Besoin de caractéristiques contextuelles du même moment que les données individuelles

⇒ Besoin de données longitudinales

- La modélisation n'est toujours pas parfaite, poursuivre le travail?
  - → Ajouter un aléa sur une caractéristique individuelle
  - → Introduire un niveau ménage?

Conclusion

- Le multiniveau apporte une bien meilleure modélisation des effets contextuels

- L'ouverture des données permet des analyses inédites et beaucoup mieux contextualisées qu'avant

- Mais l'accès à des données définies au niveau qui nous intéresse, à la bonne date, demeure difficile

Références bibliographiques

7E Réseau thématique de recherche de l'UEPA (dir.), (1999), Guide d'exploitation et d'analyse des données de recensements et d'enquêtes en matière de scolarisation, Paris, Ceped/UEPA/Unesco, Documents et manuels du Ceped.

Bressoux, P. (2010). Modélisation statistique appliquée aux sciences sociales, De boeck Brussels,464p.

Bringé, A., **Golaz**, V., (2017). Manuel pratique d'analyse multiniveau, Paris, France, INED éditions, 117 p.

Courgeau, D. (2004). Du groupe à l'individu, Synthèse multiniveau, Paris, Ined Editions, coll. « Les manuels»

Ministry of Education Science and Technology, **UNICEF** (2015). 2014 Basic Education Statistical Booklet.

Wiesmann, U., **Kiteme,** B., **Mwangi**, Z. (2016). Socio-Economic Atlas of Kenya: Depicting the National Population Census by County and Sub-Location. Second, revised edition. KNBS, Nairobi. CETRAD, Nanyuki. CDE, Bern.