Exploiter les données d'enquêtes ménages pour la recherche ou la décision publique : guide et étude de cas sur l'accès à l'eau à Kinshasa

Florent Bedecarrats, Oriane Lafuente-Sampietro, Martin Lemenager, Thimothée Makabu

Exploiter les donnée d'enquêtes ménages pour la recherche ou la décision publique : guide et étude de cas sur l'accès à l'eau à Kinshasa

F. Bedecarrats, O. Lafuente, M. Lemenager, T. Makabu

26 septembre 2016

Résumé : Les enquêtes menées périodiquement par les instituts nationaux de la statistique dans la plupart des pays du monde couvrent des thèmes très divers : sociaux, économiques, sanitaires, culturels et politiques. Sous-exploitées, elles ne débouchent souvent que sur le calcul de quelques grands agrégats nationaux qui alimentent les portails statistiques internationaux (Banque mondiale, Organisation mondiale de la santé, nations unies…). Cette utilisation reste très limitée, en comparaison aux multiples calculs d'indicateurs, croisements de variables et périmètres d'analyse que permettent les informations détaillées collectées lors de ces enquêtes. Ces bases de données brutes sont pourtant accessibles aux chercheurs, étudiants, décideurs ou praticiens et elles permettent par exemple d'éclairer la conception, le suivi et l'évaluation de projets, programmes et politiques publique. Après avoir décrit ces applications, ce document décrit où et comment obtenir ces données et la manière de les analyser, grâce au package survey, disponible sur R, un logiciel statistique libre et gratuit. La démarche d'analyse est illustrée par l'exemple de la République Démocratique du Congo où des enquêtes de nature différentes (MICS 2010, 1-2-3 2012 et DHS 2014) permettent d'étudier les caractéristiques et l'évolution de l'accès à l'eau à Kinshasa.

Introduction

Avec l'avènement de l'« Open Data », on assiste à un accroissement impressionnant de la diversité, de la quantité et de la qualité des informations librement accessibles. Cette tendance concerne tous les continents, et notamment l'Afrique sub-saharienne. La division d'évaluation de l'Agence française de développement accompagne ses homologues des services opérationnels et les partenaires pour mieux exploiter le potentiel des bases de données administratives existantes, systèmes d'information opérationnelle, données satellitaires et enquêtes ménages. Concernant ces dernières, on dispose de données collectées régulièrement, de manière fiable et harmonisée sur les conditions sociales, économiques et sanitaires des ménages, leur appréciation des institutions, des services essentiels et de la gouvernance, les entreprises, le secteur informel ou encore l'agriculture. L'analyse de ces données fournit, à moindre coûts et rapidement des informations pertinentes et précises pour comprendre les contextes d'interventions au moment de concevoir projets, programmes et politiques publiques. Cela permet aussi de documenter une situation initiale et son évolution, ce qui est essentiel pour évaluer des projets, programmes ou politiques publiques.

De nombreux professionnels de l'aide au développement citent ou utilisent des statistiques sur leurs pays d'interventions qui sont produites par la Banque mondiale, l'Organisation pour la coopération et le développement économique, l'Organisation mondiale pour la santé, ou d'autres agences des Nations Unies. Mais souvent, ces professionnels sont en même temps persuadés qu'au sein des pays en question, les données sont rares ou peu fiables. En réalité, la plupart des données relayées par les institutions internationales sont produites par les systèmes nationaux de statistique – principalement au moyen d'enquêtes – et les institutions internationales ne font qu'harmoniser, consolider et relayer ces informations. Ainsi, gouvernements et bailleurs consacrent des millions à une collecte locale d'information qui n'est exploitée presque que pour produire des indicateurs agrégés. On peut pourtant faire bien plus en exploitant les données brutes de ces enquêtes. Elles peuvent en particulier être mobilisées pour renseigner le contexte et les résultats de projets, programmes et politiques de développement.

Le panorama des enquêtes menées périodiquement dans les pays en développement est très riche :

- Les enquêtes en grappes à indicateurs multiples (MICS en anglais) et enquêtes démographiques et de santé (DHS) s'intéressent principalement à des aspects sociaux, démographiques et sanitaires. Elles reposent typiquement sur des échantillons de 10 000 à 40 000 ménages et ont lieu tous les trois à cinq ans.
- Les enquêtes de mesure du niveau de vie des ménages (LSMS) analysent le budget des ménages et permettent d'établir les taux de pauvreté et servent au calcul de plusieurs agrégats macroéconomiques. Les enquêtes LSMS sont généralement mises en œuvre tous les quatre à six ans. Elles sont de plus en plus nombreuses à inclure une composante de panel, c'est-à-dire qu'un sous-échantillon de ménage est réinterrogé tous les ans ou tous les deux ans, afin de suivre les tendances entre deux enquêtes de plus grande ampleur. Dans plusieurs pays, les enquêtes LSMS ont été complétées au travers d'une approche « 1-2-3 », c'est-à-dire qu'en amont d'une analyse du budget des ménages (« 3 »), des collectes de données sont effectuées sur l'emploi (« 1 ») et les entreprises informelles (« 2 »). Dans les économies des pays les moins avancés ou à revenus intermédiaires, une part significative des emplois et des entreprises n'est déclarée auprès des autorités. En l'absence de registre administratif fiable, ces activités informelles sont identifiées en interrogeant les membres des ménages sur leur occupation. Certaines enquêtes LSMS ont aussi été adaptées pour mieux appréhender les activités agricoles.
- Des enquêtes moins exigeantes visent uniquement à collecter quelques indicateurs clés de bien-être. Elles se sont multipliées depuis les années 2000, afin de suivre l'évolution des Objectifs du Millénaire pour le Développement et de renseigner quelques variables socio-économiques et sanitaires essentielles. Elles n'approfondissent cependant pas ces sujets autant que les enquêtes MICS, DHS ou LSMS.
- Nombre d'autres informations sur des aspects socio-politiques ou de gouvernance sont aussi produites avec des enquêtes. Les enquêtes sanitaires, économiques ou de bien-être mentionnées aux points précédents incluent souvent des modules additionnels sur les expériences et la perception des citoyens à l'égard des institutions publiques (police, justice, autorités locales, etc.), conflits, sécurité, corruption et cohésion sociale. Des enquêtes spécifiques sur ces sujets sont aussi menées annuellement ou bisannuellement comme Afrobaromètre ou Gallup World Poll. D'autres enquêtes sont aussi plus exceptionnellement menées dans de nombreux pays sur des sujets spécifiques, comme Global Findex 2011 et 2014 sur l'inclusion financière.

Les données brutes de la plupart de ces enquêtes sont rendues disponibles au travers de portails internet tels que ihsn.org, dhsprogram.com, mics.unicef.org, microdata.worldbank.org, etc. Ils peuvent ainsi être téléchargés librement, souvent après avoir rempli un court formulaire et s'être engagé à respecter quelques principes éthiques essentiels, comme ne pas diffuser les données détaillées sans autorisation, ne pas les utiliser à des fins commerciales ou politiciennes ou ne pas mettre en péril l'anonymat des personnes enquêtées.

Ces enquêtes font l'objet d'un contrôle de qualité étroit de la part de différentes instances. Les instituts nationaux de la statistiques, qui produisent la plupart de ces données, suivent des standards concernant plusieurs aspects : recrutement d'enquêteurs qualifiés, formation pour chaque enquête, tests en conditions réelles des questionnaires, supervision de terrain, audit d'un échantillon de registres, double saisie des données, etc. Dans la plupart des cas, ils bénéficient également de l'assistance technique et de la supervision de partenaires internationaux : la Banque mondiale pour LSMS, UNICEF pour MICS, ICF international pour DHS, AfriStat, etc. De plus, il faut signaler que le libre accès aux données brutes est en soi un facteur de fiabilité de l'information, car les anomalies peuvent aisément être repérées lors de la phase de préparation et de description des données. Il semblerait en effet techniquement ardu de falsifier certaines réponses des enquêtés tout en assurant la cohérence de renseignements détaillés et interdépendants concernant des centaines de variables et des milliers de registres, mais aussi en assurant la plausibilité des comparaisons avec les résultats d'enquêtes passées et futures.

Les principes de l'analyse d'enquêtes complexes

On parle d'enquêtes « complexes » car, en l'absence d'inventaire complet et à jour de l'ensemble des individus (recensement général de la population, registre des entreprises), elles ne reposent pas sur des tirages aléatoires simples qui permettraient des calculs élémentaires de probabilité. Elles utilisent en revanche une imbrication de tirages successifs. Les estimations doivent donc recourir à une arithmétique un peu plus sophistiquée, heureusement prise en charge automatiquement par les fonctions préprogrammées fournies par les logiciels statistiques.

Ces enquêtes sont fondées sur un échantillonnage aréolaire stratifié à deux degrés. C'est-à-dire qu'elles se basent en première instance sur un zonage du territoire national comprenant plusieurs milliers d'aires. Un tirage aléatoire est réalisé pour sélectionner plusieurs centaines d'aires, si possible en faisant en sorte que la probabilité de tirage de chaque aire soit proportionnelle au nombre de tirage qu'elle comportait lors du dernier recensement disponible. Ce tirage peut s'effectuer indépendamment pour des sous-ensembles, par exemple par région administrative ou par zones rurales et urbaines, afin de garantir que l'échantillonnage sera représentatif pour chacun de ces sous-ensembles. On effectue ensuite un dénombrement exhaustif des ménages résidant dans chacune des aires sélectionnées lors de la première phase. On tire enfin au sort un nombre fixe de ménages (généralement entre 10 et 25) dans chaque aire. Ce sont ces ménages qui seront interrogés lors de l'enquête.

Cette procédure d'échantillonnage comporte avantages et inconvénients. Elle réduit la dispersion géographique des individus à interroger – et donc le temps et le coût de collecte – en favorisant la sélection de ménages localisés en grappes. Elle a en revanche un effet négatif sur les marges d'erreur, car des voisins tendent à se ressembler. Chaque estimation doit donc inclure un calcul de corrélation intra-grappe, afin de tenir compte du biais qu'introduit ce phénomène pour la représentativité de l'échantillon (une variance moins grande dans l'échantillon que dans la population dans son ensemble). La stratification du tirage améliore en revanche la précision des estimations. Par exemple pour une stratification rural/urbain, ménages ruraux partagent des traits communs et sont souvent très différents des ménages urbains. La variance au sein de chacune de ces deux catégories est donc plus faible que la variance de ces deux catégories prises ensemble. Réaliser les calculs d'estimations et de marge d'erreurs séparément pour urbains et ruraux, pour ensuite en faire la moyenne produit les mêmes indicateurs, mais avec des marges d'erreur plus petites.

La représentativité de l'enquête repose sur le principe selon lequel tous les ménages du territoire doivent avoir la même chance d'être tirés au sort. Des disparités d'effectifs de populations entre les aires sont de nature à biaiser cette équiprobabilité. Pour rectifier ce biais, on compare le nombre de ménages tirés au sort dans chaque aire au nombre total de ménages résidant dans l'aire. Le quotient traduit la probabilité de tirage pour chaque ménage sélectionné et sera employé comme pondération pour le calcul de l'ensemble des indicateurs estimés à partir de l'enquête. A partir du moment où les probabilités de tirage sont connues pour chaque individu et chaque paire d'individus tirés au sort, on peut calculer des estimations applicables à l'ensemble de la population et calculer la fiabilité de ces estimations au moyen de la méthode Horvitz-Thomson (Lumley, 2010).

On peut recommander deux ouvrages consacrés à l'analyse des enquêtes auprès des ménages. Celui d'Angus Deaton (1997, en accès libre) reste probablement la meilleure référence sur la théorie et la pratique des enquêtes auprès des ménages dans les pays en développement. On y trouve aussi des méthodes permettant à partir de ces sources d'étudier les sujets majeurs qui ont valu à l'auteur son Prix Nobel d'économie en 2015 : consommation, pauvreté, inégalités, nutrition, épargne et formation des prix. Angus Deaton a en outre apporté une contribution à la communauté des analystes qui s'intéressent à ces sources en partageant et commentant les scripts d'analyse qu'il avait conçu pour l'analyse de ces enquêtes avec le logiciel propriétaire Stata®. Ceux-ci ont été complétés et actualisés, notamment par des chercheurs associés au programme LSMS, qui ont développé un programme gratuit intitulé ADePT qui contient une version allégée de Stata® et fournit une interface simplifiée permettant de produire à partir d'une enquête sur le budget des ménages une série de tableaux et graphiques standardisés, sur plusieurs grands thèmes : pauvreté, emploi, genre, protection sociale, inégalités, éducation et alimentation. Ce logiciel oblige toutefois à rester dans le cadre des traitements prédéfinis et une version complète de Stata® coûte dans un pays Africain entre 900 USD pour un étudiant et 1800 USD pour les autres utilisateurs.

Le second ouvrage que nous recommandons est celui de Thomas Lumley (2011) consacré à l'analyse des enquêtes complexes. Voir le site créé par l'auteur pour obtenir les fichiers d'exercice et la référence du livre, ainsi que des publications en accès libre pour ceux qui ne voudraient pas l'acheter. Cet ouvrage synthétise l'état de l'art sur le sujet et explique de manière très didactique et détaillée la manière de les mettre en œuvre avec *survey*, le paquet de fonctions préprogrammées rendu librement accessible par l'auteur pour le logiciel statistique libre et gratuit R. Le paquet *survey* simplifie grandement l'analyse des données d'enquête mais, comme pour Stata®, il implique de connaître et de spécifier adéquatement la méthode d'échantillonnage sur laquelle l'enquête a été construite. Une bonne pratique consiste dans un premier temps à reproduire les résultats déjà publiés dans le rapport officiel de l'enquête et de s'assurer que cela concorde, afin de s'assurer que tout a été correctement paramétré.

Applications pour concevoir, suivre et évaluer projets et politiques publiques

La fonction première de l'analyse des données d'enquêtes est de produire des statistiques descriptives qui soient utiles à la conception et au suivi de politiques publiques. Nous estimons approximativement qu'environ les deux tiers des variables collectées auprès des ménages lors des enquêtes DHS ou MICS sont finalement analysées dans les rapports officiels. Même ainsi, nombre de ces indicateurs sociaux, démographiques et sanitaires sont seulement calculés comme des agrégats nationaux, sans être déclinés par région, zones rurales et urbaines, ou niveau de vie des ménages. La part des données collectées qui sont effectivement restituées dans les rapports publics est bien inférieure pour les enquêtes LSMS, où une part importante des informations sur la consommation des ménages sont simplement traitées pour établir des indices de prix, de revenus ou de pauvreté au niveau national. Nous avons même rencontré de nombreux cas où les rapports d'enquêtes n'avaient toujours pas été publiés 2 à 4 ans après que les données socio-économiques ont été collectées, et parfois rendues publiques.

> **Attention** : en réduisant le périmètre des calculs pour se concentrer sur des régions moins vastes ou des catégories de population, les marges d'erreur s'accroissent et il devient indispensable de les examiner soigneusement afin de ne pas sur-interpréter la portée des résultats. Pour ce faire, les erreurs-type et/ou les intervalles de confiance doivent être systématiquement calculés et inclus dans les rapports d'analyse. Sauf de rares cas où cela peut sembler pertinent (par exemple pour dire que les données ne permettent pas de comparer) et les implications en sont discutées clairement, il convient de ne pas produire de graphique ou de commentaire lorsque les intervalles de confiance empêchent toute interprétation.

Une deuxième fonction de ces analyses est de suivre des tendances, c'est-à-dire l'évolution de variables d'intérêt au cours du temps. Cela fournit des informations importantes aux acteurs qui participent à la prise de décision ou à la mise en œuvre de projets et politiques publiques, par exemple pour savoir si le taux d'accès à l'électricité a substantiellement augmenté après un programme d'extension du réseau ciblant certaines régions, ou pour obtenir des informations concernant l'assistance des parturientes lors de l'accouchement pour des programmes de santé maternelle, ou encore pour connaître l'usage de sources d'eau ou de dispositifs d'assainissement amélioré suite à des investissements dans ce domaine.

Evaluer l'impact des programmes est une troisième fonction possible pour les données d'enquêtes auprès des ménages. Ce type d'évaluation implique de prouver l'existence d'un rapport de cause à effet entre l'intervention et le résultat, ce qui requiert une analyse plus exigeante. Hormis pour les enquêtes en panel, les échantillons des enquêtes nationales auprès des ménages sont tirées indépendamment à chaque fois. Il est donc très peu probable qu'un ménage sélectionné lors d'une enquête le soit à nouveau quelques années plus tard lors de l'enquête suivante. Même si c'était le cas, les observations sont anonymisées de sorte que les recoupements entre les bases de données sont impossibles. Par conséquent, sauf pour les enquêtes en panel, les enquêtes auprès des ménages fournissent des observations transversales (on dit aussi « en coupe ») et non des données longitudinales. La supériorité des observations longitudinales pour les évaluations d'impact a été

abondamment commentée (A. S. Deaton, 1997; Wooldridge, 2010, 2011, p. 432-484). Les données de panel sont en principe préférables pour distinguer le lacis de facteurs structurels qui déterminent l'évolution des variables au cours du temps, mais les échantillons sont généralement plus petits et difficiles à conserver au cours du temps et elles tendent à perdre leurs propriétés de représentativité de l'ensemble de la population. Mais les données en coupe ont aussi des avantages : elles s'appuient généralement sur des échantillons sensiblement plus grands, fournissent des estimateurs plus précis et des tests statistiques plus puissants, sont conçues pour maximiser la représentation de l'ensemble de la population et sont réitérées à quelques années d'intervalles avec des méthodes harmonisées. Plusieurs évaluations d'impact utilisées aujourd'hui dans les manuels d'économétrie ont été réalisées à partir de données en coupes issues d'enquêtes nationales auprès des ménages (A. Deaton, 1985; Kiel & McClain, 1995; Sander, 1992). Une évaluation d'impact récemment commanditée par l'AFD a aussi consisté à rassembler cinq enquêtes DHS et MICS menées pendant plus de dix ans en Mauritanie, pour évaluer l'impact d'un dispositif de financement de l'accès aux soins qui avait progressivement été étendu dans le pays pendant cette période (Philibert et al., s.d.).

Encourager des analyses ouvertes et reproductibles

L'AFD essaye de promouvoir l'usage de ces données et méthode par les ministères concernés dans nos pays en développement partenaires, par des parties prenantes locales, des chercheurs, étudiants et des bureaux d'études. La plupart des analyses de données d'enquêtes produites par la division évaluation sont produites avec R, car ce logiciel est devenu en quelques années le plus populaire auprès des statisticiens et que sa gratuité le rend accessible à tous les partenaires. Les rapports sont élaborés dans des fichiers de type RMarkdown, c'est-à-dire contenant à la fois le texte rédigé et les scripts de calcul. Cela permet d'actualiser automatiquement l'ensemble des résultats si des données ou des paramètres sont modifiés. Plus important encore, cela permet à toute personne recevant le rapport de vérifier la validité des résultats et de s'inspirer de ces travaux pour mener des études similaires.

Exemple guidé avec l'analyse de l'accès à l'eau à Kinshasa.

Au cours de la dernière décennie, l'Institut national de la statistique (INS) de la République démocratique du Congo (RDC) a mené des enquêtes auprès des ménages qui reflètent la diversité de la typologie présentée plus haut : enquêtes MICS, DHS et 1-2-3. Nous allons ci-dessous décrire pas-à-pas le processus de récupération, d'analyse et de visualisation de ces données pour décrire l'accès à l'eau à Kinshasa, avec l'espoir que cet exemple pourra vous être utile pour vos propres recherches.

Accès aux données et téléchargement

Trouver les données d'intérêt

Pour identifier les données d'enquêtes disponibles pour votre pays d'étude, il est en général recommandé de vous rendre sur chacun des sites mentionnés dans la première partie de ce post. Pour notre sujet en particulier, le site du Joint Monitoring Program de l'OMS et de l'UNICEF tient à jour une liste de toutes les sources existantes permettant de suivre les indicateurs des Objectifs du Millénaire pour le Développement relatif à l'eau et à l'assainissement. Cela facilite grandement l'identification de la source de données la plus récente quand on s'intéresse spécifiquement à ce sujet. Il faut simplement télécharger le fichier pays de la zone d'étude (ici la RDC).

Dans ce fichier Excel, il faut se reporter à la feuille intitulée **"Tables_W"**. La colonne la plus à droite correspond à l'enquête la plus récente qui ait été menée. En l'occurrence pour la RDC, il s'agit de l'enquête 2014.

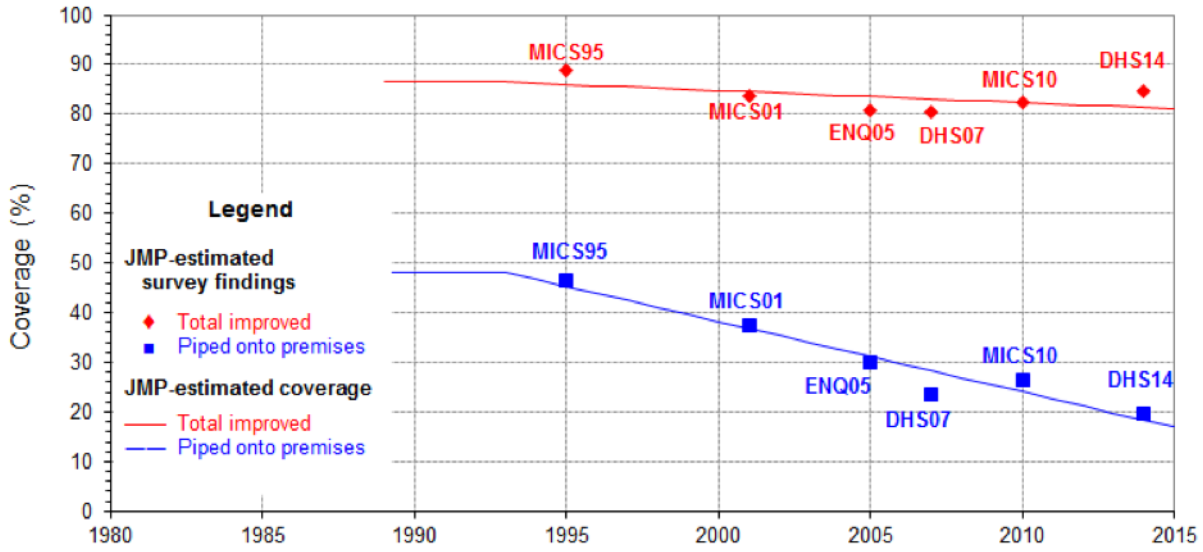


Figure 1: Part estimée de la population urbaine en RDC qui utilise une source d'eau améliorée (Source : Joint Monitoring Program)

Les informations déjà analysées dans ce fichier composent le fichier suivant sur l'accès à des sources d'eau améliorées pour la population urbaine en RDC.

Dans le graphique ci-dessus, les points correspondent aux taux d'accès à une source d'eau potable améliorée (en rouge) et en particulier à un branchement privé pour l'ensemble de la population urbaine, tels que publiés dans les rapports d'enquêtes ménages par l'institut national de la statistique congolais. Ces statistiques sont calculés pour l'ensemble des régions urbaines de RDC, et ne renseignent pas sur l'accès à l'eau spécifiquement pour la ville de Kinshasa. Nous avons à produire ces mêmes indicateurs pour Kinshasa en utilisant les données brutes issues des enquêtes. En effet, ces enquêtes ont été construites de façon à être statistiquement représentatives au niveau de certaines régions géographiques. A chaque fois, Kinshasa a été choisie pour être l'une de ces régions et donc avoir un échantillon représentatif de sa population.

Enquête	Supervisée par	Période de collecte	Nb. total de ménages interrogés	Nb. de ménages interrogés à Kinshasa
DHS 2007	Macro International	Février–août 2007	8 886	2 665
MICS 2010	UNICEF	Février–avril 2010	11 393	1 004
1-2-3 2012	Afristat et DIAL	Octobre-décembre 2012	21 454	1 969
DHS 2013-2014	ICF Macro	Août–septembre 2013	18 360	1 224

Les enquêtes DHS et MICS donnent des informations sur l'accès à l'eau, la distance au point d'eau, le type de sanitaires, les pratiques d'hygiène, le stockage de l'eau, ou encore les maladies diarrhéiques. Ces enquêtes contiennent également des informations sur le niveau de vie des ménages et les principales activités économiques des membres du ménage. L'enquête 1-2-3 ne collecte qu'une partie de ces informations et utilise des critères légèrement différents : l'accès à l'eau et aux sanitaires, l'occurrence des maladies diarrhéiques. En contrepartie, ces données contiennent des informations socio-économiques bien plus détaillées sur les activités du ménage, les sources de revenus et les dépenses. Les enquêtes DHS fournissent des coordonnées GPS pour des groupes de ménages et les enquêtes 1-2-3 l'adresse exacte de chaque ménage interrogé. Ces enquêtes nous permettent de décrire la situation de l'accès à l'eau à Kinshasa et dans ses quartiers.

Télécharger les bases de données DHS et MICS

Pour réaliser une petite analyse statistique, nous allons utiliser ces quatre enquêtes. Les enquêtes DHS et MICS sont disponibles en ligne tandis que l'enquête 1-2-3 doit être demandée à l'Institut National de la Statistique (INS) de RDC.

Télécharger les enquêtes DHS

Pour télécharger les deux enquêtes DHS, il faut s'enregistrer sur le site de DHS et remplir le formulaire suivant :

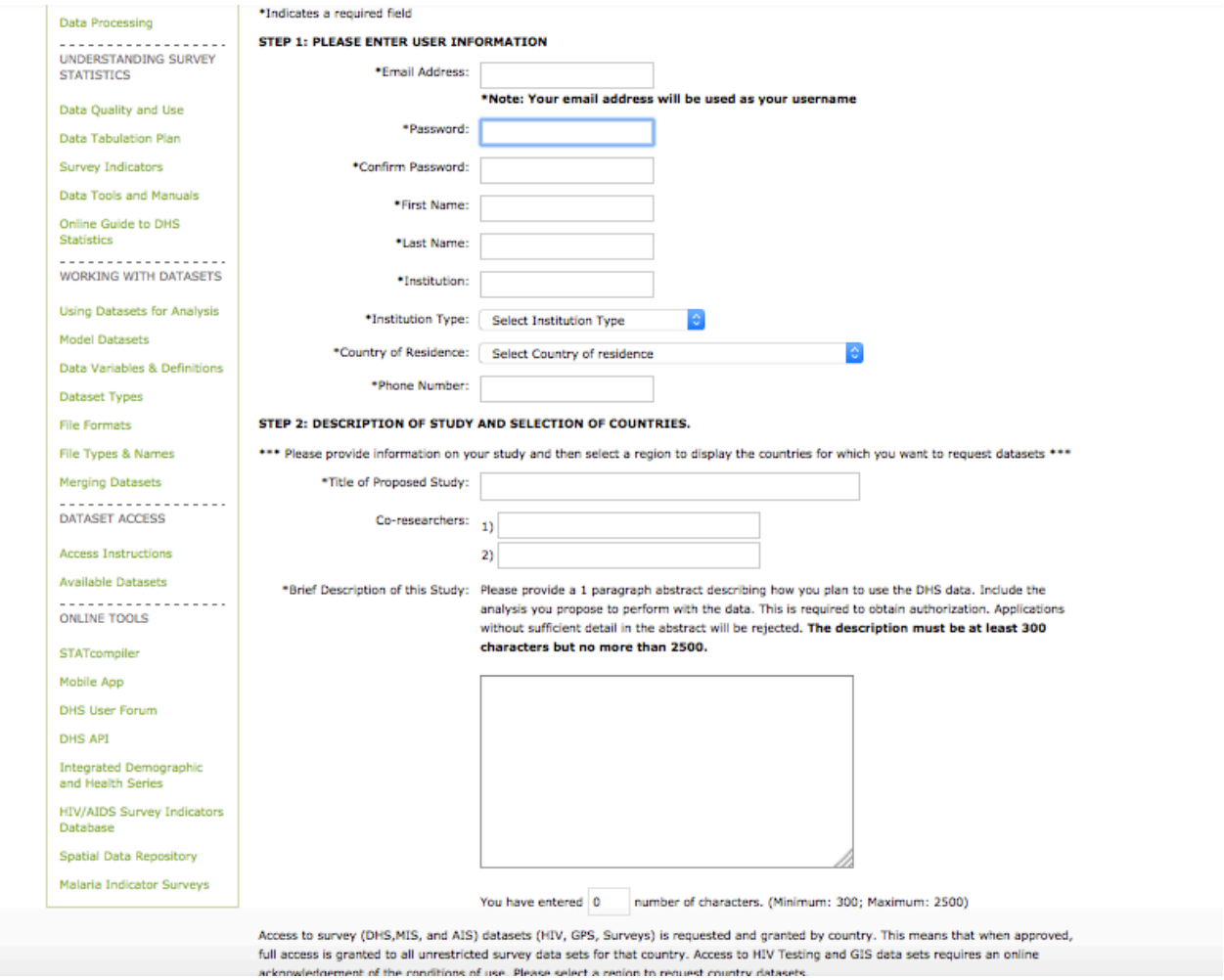
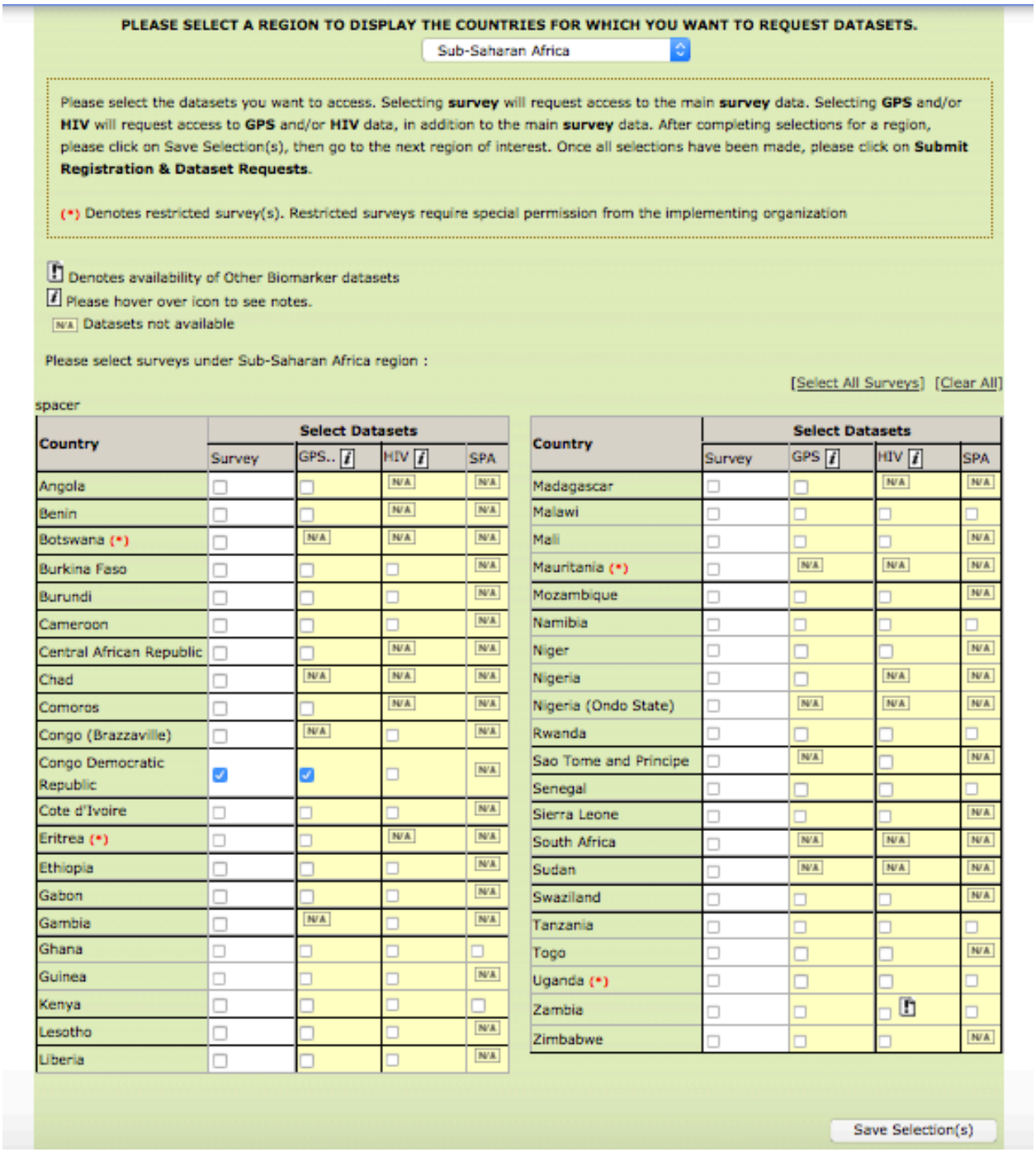


Figure 2: Capture d'écran du formulaire d'inscription et de téléchargement des données DHS

Une fois que vous avez fourni les informations précédentes, il faut effectuer une demande pour les bases de données souhaitées. Ici, nous sélectionnons « Sub-Saharan Africa » dans **Select Region** puis « Congo Democratic Republic » dans **Select country** en cochant les cases **Survey** et **GPS**.

Les questionnaires des enquêtes de santé contiennent des informations pour plusieurs niveau d'analyse : les ménages, les femmes et les hommes adultes et les naissances. Ici, notre intérêt se porte sur les ménages car c'est au niveau du ménage que l'accès à l'eau est déterminé.

Vous recevrez un email quelques heures plus tard pour vous autoriser à télécharger les données. Pour notre étude, il faut télécharger les enquêtes de 2007 et 2014.



PLEASE SELECT A REGION TO DISPLAY THE COUNTRIES FOR WHICH YOU WANT TO REQUEST DATASETS.

Sub-Saharan Africa

Please select the datasets you want to access. Selecting **survey** will request access to the main **survey** data. Selecting **GPS** and/or **HIV** will request access to **GPS** and/or **HIV** data, in addition to the main **survey** data. After completing selections for a region, please click on Save Selection(s), then go to the next region of interest. Once all selections have been made, please click on **Submit Registration & Dataset Requests**.

(*) Denotes restricted survey(s). Restricted surveys require special permission from the implementing organization

Denotes availability of Other Biomarker datasets

Please hover over icon to see notes.

N/A Datasets not available

Please select surveys under Sub-Saharan Africa region :

[Select All Surveys] [Clear All]

spacer

Country	Select Datasets			
	Survey	GPS..	HIV	SPA
Angola	☐	☐	N/A	N/A
Benin	☐	☐	N/A	N/A
Botswana (*)	☐	N/A	N/A	N/A
Burkina Faso	☐	☐	☐	N/A
Burundi	☐	☐	☐	N/A
Cameroon	☐	☐	☐	N/A
Central African Republic	☐	☐	N/A	N/A
Chad	☐	N/A	N/A	N/A
Comoros	☐	☐	N/A	N/A
Congo (Brazzaville)	☐	N/A	☐	N/A
Congo Democratic Republic	☑	☑	☐	N/A
Cote d'Ivoire	☐	☐	☐	N/A
Eritrea (*)	☐	☐	N/A	N/A
Ethiopia	☐	☐	☐	N/A
Gabon	☐	☐	☐	N/A
Gambia	☐	N/A	☐	N/A
Ghana	☐	☐	☐	☐
Guinea	☐	☐	☐	N/A
Kenya	☐	☐	☐	☐
Lesotho	☐	☐	☐	N/A
Liberia	☐	☐	☐	N/A

Country	Select Datasets			
	Survey	GPS	HIV	SPA
Madagascar	☐	☐	N/A	N/A
Malawi	☐	☐	☐	☐
Mali	☐	☐	☐	N/A
Mauritania (*)	☐	N/A	N/A	N/A
Mozambique	☐	☐	☐	N/A
Namibia	☐	☐	☐	☐
Niger	☐	☐	☐	N/A
Nigeria	☐	☐	N/A	N/A
Nigeria (Ondo State)	☐	N/A	N/A	N/A
Rwanda	☐	☐	☐	☐
Sao Tome and Principe	☐	N/A	☐	N/A
Senegal	☐	☐	☐	☐
Sierra Leone	☐	☐	☐	N/A
South Africa	☐	N/A	N/A	N/A
Sudan	☐	N/A	N/A	N/A
Swaziland	☐	☐	☐	N/A
Tanzania	☐	☐	☐	☐
Togo	☐	☐	☐	N/A
Uganda (*)	☐	☐	☐	☐
Zambia	☐	☐	☐	☐
Zimbabwe	☐	☐	☐	N/A

Save Selection(s)

Figure 3: Capture d'écran des différents types fichiers pouvant être téléchargés pour une enquête

Les fichiers téléchargés peuvent être copiés directement sans être dézippés dans votre dossier de travail. Ainsi, vous pourrez les utiliser directement avec leur nom sans avoir à spécifier « le/chemin/complet/jusqu'au/dossier/ où/vous/le/rangez ».

Télécharger les enquêtes MICS

La procédure est similaire pour les données MICS. Inscrivez-vous sur la section correspondante du site de l'UNICEF. Cliquez ensuite sur **SURVEYS** et sélectionnez dans **Any Country** « Congo, Democratic Republic of Congo ». Vous serez immédiatement autorisé à télécharger les bases de données de 2010.

Analyse de données MICS et DHS avec R

Un prérequis à ce tutoriel est d'installer le logiciel de traitement statistique R. Nous recommandons d'installer également RStudio, car il procure une interface très ergonomique qui facilite l'utilisation de R.

Installer des packages

L'installation basique de R contient un grand nombre de fonctions nécessaires pour réaliser les analyses statistiques les plus courantes. L'installation de package procure des fonctions additionnelles permettant de réaliser des tâches plus spécifiques. Nous allons utiliser le package *foreign* pour lire les bases de données fournies par MICS et DHS en format SPSS. Le package *survey*, développé par Thomas Lumley est crucial pour faciliter le travail sur des échantillons complexes (ici un échantillon stratifié par grappe). Le package *ggplot2* est très simple et flexible pour créer des graphiques.

Si les packages en question n'ont encore jamais été installés sur l'ordinateur, les commandes suivantes permettent de le faire :

```
install.packages("foreign")
install.packages("survey")
install.packages("ggplot2")
```

On indique ensuite à R que nous allons les utiliser durant la session :

```
library(foreign)
library(survey)
library(ggplot2)
```

Charger et préparer les données

Nous ouvrons ensuite les fichiers données DHS et MICS que nous avons téléchargés :

```
DHS_2014 <- read.spss("CDHR61FL.SAV", to.data.frame=TRUE)
DHS_2007 <- read.spss ("cdhr50fl.sav", to.data.frame=TRUE)
MICS_2010 <- read.spss("hh.sav", to.data.frame=TRUE)
```

Le dictionnaire des variables des enquêtes DHS est disponible sur le site de DHS avec de nombreux documents utiles. Des ressources similaires peuvent être trouvées sur le site internet des enquêtes MICS. Le nom et la codification des variables peuvent aussi être retrouvés via la fenêtre environnement de Rstudio.

Il est généralement très important de télécharger et lire les questionnaires d'enquêtes (toujours publiés sur les pages web des enquêtes et/ou à la fin des rapports officiels), pour être sûr que la variable est correctement interprétée.

Pour notre exemple, nous restreignons l'analyse à la province de Kinshasa :

```
DHS_2014 <- subset(DHS_2014, DHS_2014$HV024=="Kinshasa")
DHS_2007 <- subset(DHS_2007, DHS_2007$HV024=="Kinshasa")
MICS_2010 <- subset(MICS_2010, MICS_2010$HH7 == "Kinshasa")
```

Dans la base de données DHS, la pondération est en base 1000000 alors que la littérature suggère que la moyenne de la pondération doit être 1 (Lumley, 2010 p. 10). Dans nos tests, nous obtenons les même résultats pour les estimations et les intervalles de confiance avec les pondérations en base 1 et 1000000. Dans cette étude, nous utiliserons la pondération en base 1.

```
DHS_2014$weight <- DHS_2014$HV005/1000000
DHS_2007$weight <- DHS_2007$HV005/1000000
```

Évolution des raccordements privés à Kinshasa

Notre première production est un graphique présentant l'évolution des raccordements privés à Kinshasa entre 2007 et 2014 avec les enquêtes MICS et DHS.

Nous devons créer une nouvelle variable qui sépare les ménages avec une connexion privée et les autres. Nous allons utiliser comme base les variables **HV201** pour DHS et **WS1** pour MICS, qui codent la principale source d'eau des ménages en 15 catégories. Pour construire la variable de connexion privée, nous sommes intéressés par seulement deux catégories : *Piped into dwelling* et *Piped to yard/plot*.

Tout d'abord, nous catégorisons tous les ménages dont nous connaissons la source d'eau dans une catégorie « Other water sources ».

```
DHS_2014$source[is.na(DHS_2014$HV201) == FALSE] <- "Other water sources"
```

La partie *is.na(DHS_2014$HV201)==FALSE* donne comme condition que les ménages aient répondu à la question correspondant à la variable HV201.

Ensuite, nous catégorisons tous les ménages présents dans les catégories *Piped into dwelling* et *Piped to yard/plot* comme ayant une connexion privée :

```
DHS_2014$source[DHS_2014$HV201 =="Piped into dwelling" |
                  DHS_2014$HV201 == "Piped to yard/plot"  ] <- "Private connection"
```

Nous répétons l'opération pour DHS 2007 et MICS 2010. Pour MICS 2010, les catégories de la variable WS1 sont en français et nous sélectionnons donc *Robinet dans le logement* et *Robinet dans quartier, cour ou parcelle*.

```
### DHS 2007 ###
DHS_2007$source[is.na(DHS_2007$HV201) == FALSE] <- "Other water sources"
DHS_2007$source[DHS_2007$HV201 =="Piped into dwelling" |
                  DHS_2007$HV201 == "Piped to yard/plot" ] <- "Private connection"

### MICS 2010 ###
MICS_2010$source[is.na(MICS_2010$WS1)== FALSE] <- "Other water sources"
MICS_2010$source [MICS_2010$WS1=="Robinet dans le logement" |
                    MICS_2010$WS1=="Robinet dans quartier,
                  cour ou parcelle" ] <- "Private connection"
```

Ensuite, nous devons spécifier le design de l'échantillon grâce à la fonction svydesign du package survey. Le design permet de tenir compte de la structure statistique de l'enquête lors du calcul des erreurs standards et des intervalles de confiance. Nous intégrons les informations suivantes et le package va les utiliser pour les calculs statistiques :

Ids : l'unité d'échantillonnage : en premier la grappe et en second le ménage Strata : le niveau de stratification (region et environnement) *Weights : la pondération* Data : la base de données

```
design_2014 <- svydesign (ids=~HV021+HV002, strata= ~HV023,
                          weights=~weight, data=DHS_2014)
design_2007 <- svydesign (ids=~HV021+HV002, strata= ~HV023,
                          weights=~weight, data=DHS_2007)
design_MICS <- svydesign(ids=~HH1+HH2, strata=~HH6+HH7,
                         weights=~hhweight, data=MICS_2010)
```

La fonction *svydesign* reproduit la base de données avec les données originales mais en ajoutant des informations nécessaires aux calculs pour tenir compte de la complexité de l'échantillonnage.

Ensuite, nous préparons les données pour le graphique en rassemblant les proportions et les intervalles de confiance dans une petite base de données commune.

Tout d'abord, nous calculons les proportions de chaque type de connexions à Kinshasa en utilisant la fonction *svymean*. On entre les informations suivantes :

~source : la variable pour laquelle nous voulons les proportions ("~" indique que la variable se trouve dans la base de données utilisée) design : le design spécifié pour l'enquête choisie *na.rm=TRUE : les valeurs manquantes ne doivent pas être prises en compte pour le calcul de la moyenne* vartype=c('se') : on veut obtenir également les erreurs standards associées aux moyennes

```
source_eau_kin_2014 <- data.frame(svymean(~source, design=design_2014,
                                          na.rm=TRUE, vartype = c('se')))
```

Ensuite, on sélectionne uniquement la catégorie private connection qui se trouve sur la deuxième ligne du dataframe. Par contre, nous avons besoin des deux colonnes du dataframe, ainsi nous indiquons que nous conservons les deux colonnes mais seulement la deuxième ligne.

```
source_eau_kin_2014 <-  source_eau_kin_2014[c(2), c(1, 2)]
```

On reproduit les mêmes commandes pour DHS 2007 et MICS 2010 :

```
source_eau_kin_2007 <- data.frame(svymean(~source, design=design_2007,
                                          na.rm=TRUE, vartype = c('se')))
source_eau_kin_2007 <-  source_eau_kin_2007[c(2), c(1, 2)]

source_eau_kin_MICS <- data.frame(svymean(~source, design=design_MICS,
                                          na.rm=TRUE, vartype = c('se')))
source_eau_kin_MICS <-  source_eau_kin_MICS[c(2), c(1, 2)]
```

Nous construisons maintenant un seul tableau (dataframe) avec les trois que nous venons de créer. Nous les combinons avec la fonction *rbind*.

```
source_evo <- rbind.data.frame (source_eau_kin_2007, source_eau_kin_MICS, source_eau_kin_2014)
```

Et on ajoute une nouvelle variable appelée *annee* contenant le nom de enquêtes et les années :

```
source_evo$annee <- c("2007 DHS", "2010 MICS", "2014 DHS")
```

Enfin, on construit le graphique avec le package *ggplot2*. Le dataframe a trois colonnes que nous allons utiliser pour le graphique : *annee*, *mean* et *SE*. Premièrement, on utilise la fonction *ggplot* et on spécifie les éléments de base de graphique : *le dataframe : source_evo* les axes dans aes : en abscisse la variable annee et en ordonnée la variable mean

```
graph_source_evo <- ggplot(source_evo, aes(x=annee, y=mean, group=1))
```

On ajoute ensuite les différents éléments pour obtenir le graphique :

```
graph_source_evo +
  geom_errorbar(aes(ymin=mean-SE, ymax=mean + SE), width=.2) +
  ### les barres d'erreurs calculés en soustrayant et ajoutant
  ### l'erreur standard à la moyenne
  geom_line() +   ### la ligne entre les deux limites
  geom_point(size=2)+   ### le point au niveau de la ligne
  xlab("") +   ### pas de titre sur l'axe des abscisses
  ylab("Proportion") + ### un titre sur l'axe des ordonnées
  ylim(0, 0.6) + ### indication d'échelle pour les ordonnées
  theme (axis.text.x = element_text(angle=90, vjust=0.5, size=11)) +
  ### positionnement du texte sur l'axe des abscisses
  ggtitle("Evolution du taux de connexions privées à Kinshasa selon
          les enquêtes sociales et de santé menées entre 2007 et 2013")
```

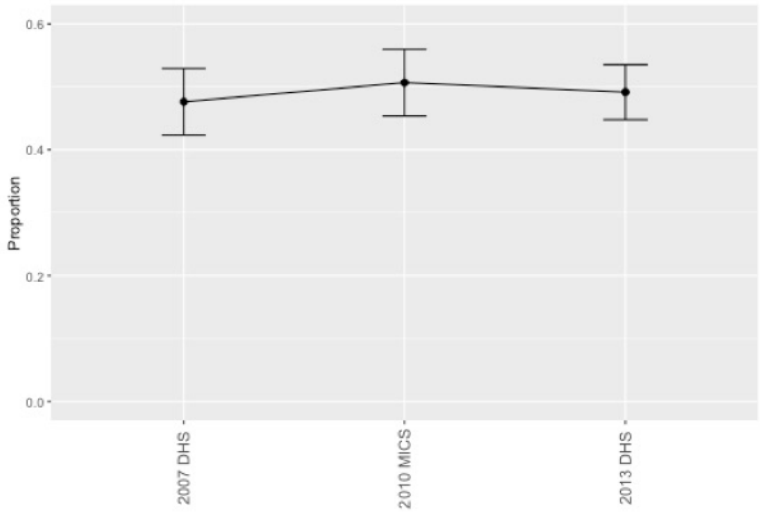


Figure 4: Evolution du taux de connexions privées à Kinshasa selon les enquêtes sociales et de santé menées entre 2007 et 2013

D'importantes disparités entre les différentes zones résidentielles

Nous avons utilisé la géolocalisation des grappes pour catégoriser les ménages entre deux type de quartiers : les quartiers ciblés par des projets visant à créer des mini-réseaux d'adduction gérés par des associations

d'usagers d'eau potable (ASUREP) et le reste de la ville. Ces quartiers ASUREP se trouvent en périphérie de la ville et devraient être plus pauvres et avec une plus mauvaise connexion que les autres quartiers. Nous allons essayer de confirmer ces hypothèses avec les données des enquêtes.

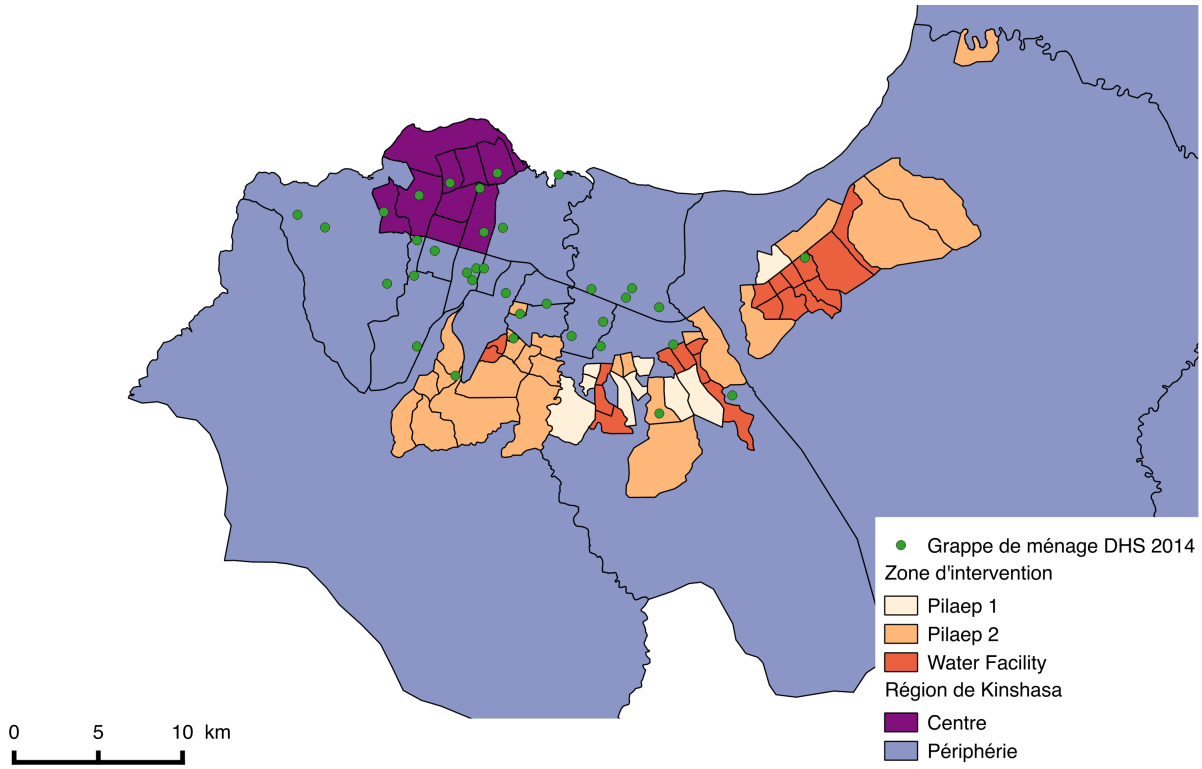


Figure 5: Localisation des grappes de ménages de l'enquête DHS 2014

Dans un futur article, nous expliquerons comment utiliser les données spatiales pour créer des cartes avec R. Pour le moment, nous allons éluder cette partie. Pour l'analyse spatiale, nous allons utiliser un fichier csv (coma separated values : un fichier de texte où les valeurs sont séparées par des virgules ou des sauts de ligne) dans lequel nous avons associé chaque grappe à une zone.

DHSCLUST	ADM1NAME	DHSREGCO	URBAN_RUR	ALT_DEM	Zone_bis	Zone
3	Kinshasa	1	U	285	Peripherie	Centre
7	Kinshasa	1	U	278	Centre	Centre
34	Kinshasa	1	U	297	Peripherie	Centre
50	Kinshasa	1	U	356	Programme	Asurep
65	Kinshasa	1	U	272	Peripherie	Centre
103	Kinshasa	1	U	312	Programme	Asurep

Figure 6: Classification géographique des grappes de ménage de l'enquête DHS 2014

On ouvre le fichier csv avec la fonction *read.csv2* (spécialement adaptée pour ouvrir les csv générés par MS Office).

```
zone_2014 <- read.csv2("Zones Kinshasa DHS2014.csv")
```

On combine l'enquête DHS 2014 et le fichier csv pour incorporer les nouvelles informations à la base DHS_2014. Pour ce faire, on crée une nouvelle variable appelée *matching* et qui sera identique dans les deux fichiers. On utilise la variable *HV001* qui contient l'identification des clusters dans l'enquête DHS et la variable *DHSCLUST* qui contient les mêmes informations dans le fichier csv.

```
DHS_2014$matching <- paste(DHS_2014$HV001, sep="_")
zone_2014$matching <- paste(zone_2014$DHSCLUST, sep="_")
DHS_2014 <- merge(DHS_2014, zone_2014, by="matching")
```

On calcule la proportion de connexion privées par zone et leurs erreurs standards avec la fonction *svyby*. Cette fonction est très proche de svymean mais avec deux nouveaux éléments :

by : la catégorie pour laquelle la proportion doit être calculée FUN : la fonction que nous voulons utiliser (ici svymean) *vartype=c('ci') : on veut les limites de l'intervalle de confiance et pas l'erreur standard

```
source_zone <- svyby(~source, by=~Zone, design=design_2014,
                     FUN=svymean, na.rm=TRUE, vartype = c('ci'))
```

On prépare le dataframe à utiliser dans le graphique et on génère une nouvelle variable avec le nom exact des zones :

```
source_zone$Zone_1<- c( "ASUREP", "Other Districts")
```

Cette fois-ci, on conserve toutes les lignes mais seulement les colonnes avec l'information sur les connexions privées, les limites des intervalles de confiance et les noms des zones :

```
source_zone <- source_zone[, c(3, 5, 7, 8)]
colnames(source_zone) <- c("Raccordement", "ci.l", "ci.u", "Zone_1")
```

Enfin, on construit le graphique :

```
### on définit les axe
graph_source_zone <- ggplot(source_zone, aes(x=Zone_1, y=Raccordement)) s
graph_source_zone +
  ### le type de graphique et ses caractéristiques
  ##ici un graphique à barre avec des couleurs de remplissage précise)
  geom_bar(position=position_dodge(), stat="identity", color="black",
           fill=c("#003399", "#0099FF")) +
  ### la position des barres d'erreur avec les limites de l'intervalle de confiance
  geom_errorbar(aes(ymin=source_zone$ci.l, ymax=source_zone$ci.u),
                width=.2,
                position=position_dodge(.9))+
 geom_line(position=position_dodge(0.9)) + ### les lignes entre les limites
  geom_point(position=position_dodge(0.9)) + ### le point à la moyenne
  xlab("Areas")+ ### le titre de l'axe des abscisses
  ylab("Proportion") + ### le titre de l'axe des ordonnées
  theme (axis.text.x = element_text(angle=90, vjust=0.5, size=11)) +
  ### les caractéristiques de l'axe des abscisses
  ggtitle("Différence entre les taux de connexions privées dans l'enquête DHS 2014 ")
```

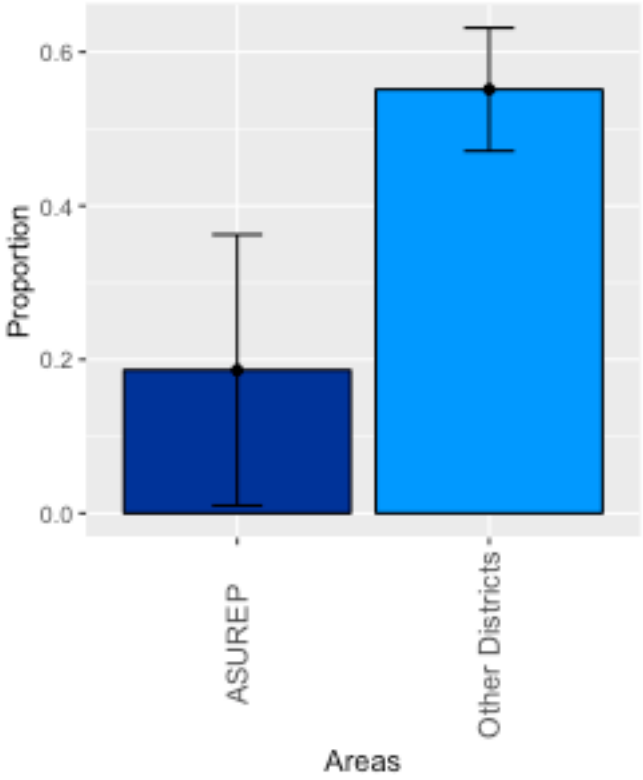


Figure 7: Différence entre les taux de connexions privées dans l'enquête DHS 2014

Analyse des données de l'enquête 1-2-3

L'enquête 1-2-3 contient plus d'informations sur la situation économique des ménages. Elle n'est pas disponible librement comme les enquêtes DHS et MICS et nous remercions l'institut national de la statistique de RDC et le laboratoire DIAL de nous autoriser à les utiliser.

Comparaison des connexions privées dans les ASUREP et le reste de la ville dans les enquêtes DHS 2014 et 1-2-3 2012

Tout d'abord, nous ouvrons la base de données des ménages de l'enquête 1-2-3 et on la combine avec le fichier csv où les sites sont répartis entre les ASUREP et le reste de la ville.

```
data_123 <- read.csv("menages.csv")
merge <- read.csv2("merge kin.csv")
```

Et on combine les deux fichiers grâce à la variable *SITE* :

```
data_123$matching <- paste(data_123$SITE, sep="_")
merge$matching <- paste(merge$site, sep="_")
data_123 <- merge(data_123, merge, by="matching")
data_123 <- subset(data_123, data_123$q03=="Kinshasa")
```

On modifie la variable de pondération pour la mettre en base 1 :

```
data_123$weight <- data_123$Coefext/1000
```

Et on crée une nouvelle variable pour identifier les ménages ayant une connexion privée à partir de la réponse à la variable *H10* :

```
data_123$source[is.na(data_123$H10) == FALSE] <- "Other water sources"
data_123$source [data_123$H10 == 1 | data_123$H10 == 2 ] <- "Raccordement privé"
```

On génère le design pour cette nouvelle base de données :

```
design_123 <- svydesign(ids=~HHID+SITE, data=data_123, weights=~weight)
```

On calcule la proportion de raccordements privés par zone avec la fonction *svyby* :

```
source_zone_123 <- svyby(~source, by=~Zone, design=design_123,
                         FUN=svymean, na.rm=TRUE, vartype = c('ci'))
```

On crée la variable de zone et on sélectionne seulement les colonnes qui nous intéressent pour le graphique :

```
source_zone_123$Zone_1<- c("ASUREP", "Other Districts")
source_zone_123 <- source_zone_123[, c(3, 5, 7, 8)]
colnames(source_zone_123) <- c("Raccordement", "ci.l", "ci.u", "Zone_1")
```

On combine ce dataframe avec le dataframe source_zone issu de l'enquête DHS 2014 :

```
source_zone <- rbind.data.frame(source_zone_123, source_zone)
source_zone$enquete <- c("1-2-3 2012", "1-2-3 2012", "DHS 2014", "DHS 2014")
```

Et on construit le graphique :

```
graph_source_zone <- ggplot(source_zone, aes(x=enquete, y=Raccordement, fill=Zone_1))
graph_source_zone + geom_bar(position=position_dodge(), stat="identity", color="black") +
  scale_fill_manual(values=c("#003399", "#0099FF"), name="Areas",
                    breaks=c("ASUREP", "Other Districts"),
                    labels=c("ASUREP", "Other Districts") ) +
  geom_errorbar(aes(ymin=source_zone$ci.l, ymax=source_zone$ci.u),
                width=.2,
                position=position_dodge(.9))+
  geom_line(position=position_dodge(0.9)) +
  geom_point(position=position_dodge(0.9)) +
  xlab("")+
  ylab("Proportion") +
  theme (axis.text.x = element_text(angle=90, vjust=0.5, size=11))+
  ggtitle("Différence entre les taux de connexions privées estimés
          par les enquêtes 1-2-3 2012 et DHS 2014")
```

Il y a des différences significatives entre les résultats des deux enquêtes. On peut expliquer ces différences de trois façons. Premièrement, les deux enquêtes n'ont pas été réalisées la même année donc il peut y avoir eu de petits changements dans la distribution d'eau. Les deux enquêtes peuvent également utiliser des définitions légèrement différentes pour la source d'eau.

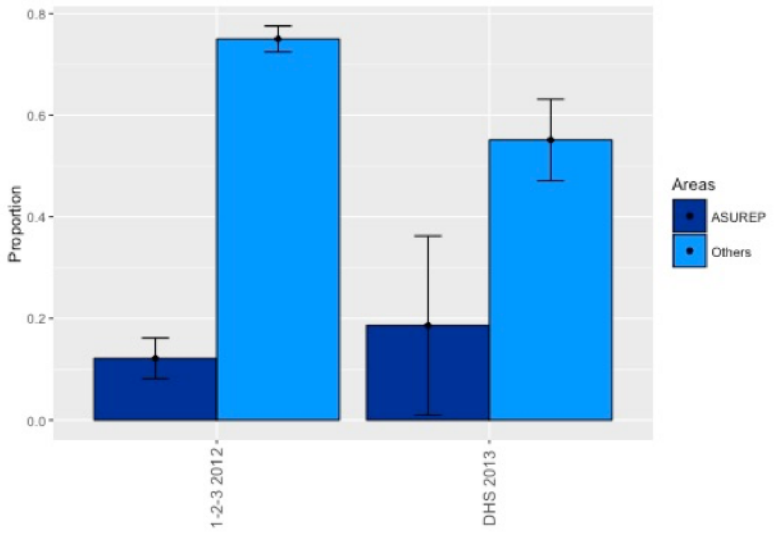


Figure 8: Différence entre les taux de connexions privées estimés par les enquêtes 1-2-3 2012 et DHS 2014

Deuxièmement, l'échantillon est tiré du dernier recensement effectué en 1984 et qui n'est donc plus représentatif de la population congolaise en 2014. Les équipes ayant construit l'échantillon ont essayé de l'actualiser mais c'est très difficile pour une ville comme Kinshasa qui a grossi très rapidement les dix dernières années. Les quartiers des ASUREP sont relativement récents et donc moins bien représentés que les quartiers plus riches du centre-ville. L'enquête 1-2-3 a effectué un effort particulier pour saisir les ménages les plus pauvres et les quartiers périphériques. Dans l'enquête 1-2-3, il y'a donc d'avantage de ménages interrogés dans ces quartiers. C'est pour ça que les estimations les concernant sont plus précises. L'enquête 1-2-3 se concentre plus sur Kinshasa et sur les inégalités de richesse. Ainsi, l'échantillon choisi est plus grand et les statistiques plus précises.

Troisièmement, l'échantillon n'a pas été construit pour être représentatif de plus petites zones que la ville entière. Les statistiques pour de plus petites zones peuvent être biaisées et non représentatives de l'ensemble de la zone. Les calculs nous donnent donc une confirmation des disparités entre les zones et une approximation des taux réels.

Analyser les dépenses en eau avec l'enquête 1-2-3

Nous devons ouvrir un autre fichier contenant les informations sur les dépenses des ménages :

```
depense <- read.dta("/way/to/my/1-2-3 data/Fonctions dépenses.dta")
```

On le combine avec la base de données en créant une variable d'identification commune :

```
data_123$id = paste(data_123$SITE, data_123$MENAGE, sep="_")
depense$id = paste(depense$site, depense$menage, sep="_")
data_123 <- merge(data_123, depense, by="id")
```

On restreint la base de données aux ménages vivant à Kinshasa :

```
data_123 <- subset(data_123, data_123$q03=="Kinshasa")
```

Et on construit les variables comprenant les dépenses en eau du ménage et leur proportion dans le budget total en considérant comme valeur manquante les réponses 0 et 9999999 et en additionnant les factures d'eau *H18* avec les autres dépenses *H20* :

```
data_123$H20_bis <- data_123$H20
data_123$H20_bis [ is.na(data_123$H20)] <- 0
data_123$H20_bis [data_123$H20 == 9999999] <- 0
data_123$H18_bis <- data_123$H18
data_123$H18_bis [ is.na(data_123$H18)] <- 0
data_123$H18_bis [data_123$H18 == 9999999] <- NA
data_123$eau <- NA
data_123$eau = data_123$H18_bis + data_123$H20_bis
data_123$eau [data_123$eau == 0] <- NA

data_123$budget_eau = (data_123$eau/data_123$deptot)*1000
data_123$budget_eau [data_123$budget_eau >= 1000] <- NA
```

Enfin on construit le dataframe à utiliser pour le graphique :

```
budget_eau <- data.frame(data_123$budget_eau, data_123$deptotuc)
colnames(budget_eau) <- c("Eau", "Depenses")
```

Et le graphique :

```
budget_eau_graph <- ggplot(budget_eau, aes(x=Depenses, y=Eau))
budget_eau_graph + geom_point() +
  geom_smooth() +
  xlab("Total spendings per consumption unit")+
  ylab("(Water spendings / Total spendings)*1000") +
  xlim(0, 5000000) +
  ylim(0, 30) +
  ggtile("Poids des dépenses en eau dans le budget des ménages
         selon leur niveau de vie estimé par l'enquête 1-2-3")
```

Ce graphique montre clairement que la plupart des ménages dépense moins de 1% de leur budget en eau, ce qui est cohérent avec les tendances observées sur le reste du continent (Banerjee et al., 2008). En accord avec la littérature, on voit également que le montant des dépenses augmente avec le niveau de revenu, particulièrement parce que les ménages des classes moyennes et les ménages les plus riches sont disproportionnellement plus connectés à des réseaux de distribution formels. La diminution du poids dans le budget du ménage montre que les plus pauvres dépenses une plus grande part de leurs revenus pour acheter de l'eau, notamment à cause de prix plus élevés payés à des fournisseurs alternatifs. En plus de ces aspects financiers, le temps passé pour chercher de l'eau représente une coup d'opportunité important (Hutton, Haller, & Bartram, 2007). Ce coup caché est clairement montré par le temps passé à la collecte de l'eau.

Temps moyen pour aller chercher de l'eau selon DHS 2014

Pour ce tableau, nous allons utiliser l'enquête DHS 2014. On commence par modifier la variable représentant le temps pour la source d'eau :

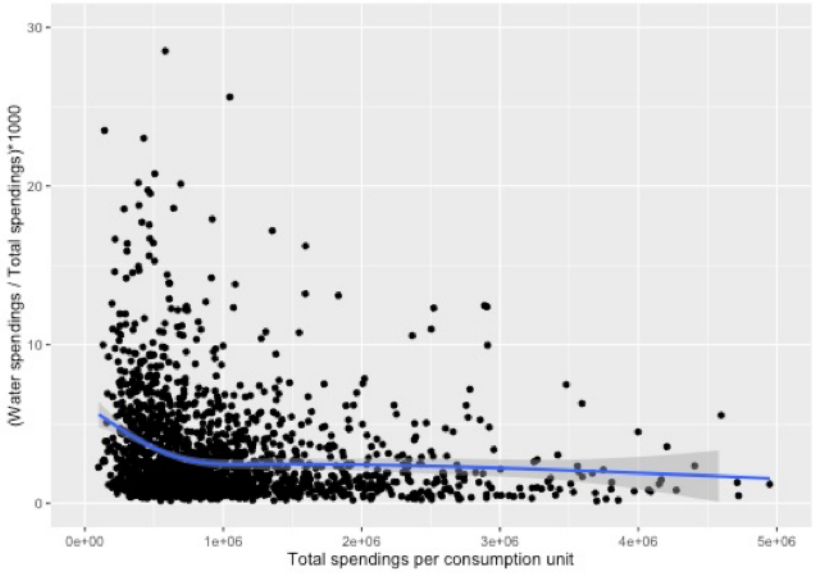


Figure 9: Poids des dépenses en eau dans le budget des ménages selon leur niveau de vie estimé par l'enquête 1-2-3

```
DHS_2014$time = DHS_2014$HV204
DHS_2014$time [DHS_2014$HV204 == "996"] <- 0
DHS_2014$time [DHS_2014$HV204 =="998"] <- NA
DHS_2014$time_rep <- NA
DHS_2014$time_rep [DHS_2014$time != "NA"] <- 1
```

On analyse les données et on construit un tableau avec le nombre de ménages interrogés :

```
temps_eau <- svyby(~time, by=~Zone, design=design_2014,
                   FUN=svymean, na.rm=TRUE, vartype = c('ci'))
temps_eau <- temps_eau[, c(2,3,4)]
total_temps <- data.frame(table(data$rep_time))
temps_eau <- data.frame(temps_eau, total_temps$Freq)
colnames(temps_eau) <- c("Estimate", "min", "max", "No. of households interviewed ")
rownames(temps_eau) <- c("Center", "Suburbs", "Neighborhoods targeted by ASUREPs")
```

Area	Estimate	min	max	No. of households interviewed
Center	3.07	0.77	5.37	192
Suburbs	6.76	3.44	10.09	802
Neighborhoods targeted by ASUREPs	27.72	13.91	41.53	201
All Kinshasa	9.71	0	0	1,195

Ce graphique indique que le temps moyen requis par les membres du ménage pour aller chercher de l'eau (aller-retour à chaque voyage). Dans le centre-ville, ce temps est estimé à 3 minutes environ, contre 28 minutes dans les quartiers ciblés par les projets de soutien aux ASUREP. Les colonnes « min » et « max » indiquent les bornes inférieure et supérieure de l'intervalle de confiance.

Conclusion

Nous élaboré ce guide afin d'expliquer de manière didactique comment obtenir les données des enquêtes menées par les INS et comment les analyser avec R, grâce au package survey. Pour illustrer notre propos, nous avons développé un exemple d'analyse, en l'occurrence les caractéristiques de l'accès à l'eau à Kinshasa, en République Démocratique du Congo. Nous espérons avoir contribué avec ce document à attirer l'attention des chercheurs, étudiants, décideurs et praticiens sur la richesse que recèlent ces sources pour améliorer la conception, le suivi et l'évaluation de projets, programmes et politiques publiques.

Bibliographie

Banerjee, S., Wodon, Q., Diallo, A., Pushak, T., Uddin, H., Tsimpo, C., & Foster, V. (2008). *Access, affordability, and alternatives: Modern infrastructure services in Africa.*

Deaton, A. (1985). "Panel data from time series of cross-sections". *Journal of econometrics*, 30(1), 109–126, disponible en ligne.

Deaton, A. S. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy.* Baltimore, MD: World Bank Publications, disponible en ligne.

Hutton, G., Haller, L., & Bartram, J. (2007). "Global cost-benefit analysis of water supply and sanitation interventions". *Journal of water and health*, 5(4), 481–502.

Kiel, K. A., & McClain, K. T. (1995). "House prices during siting decision stages: the case of an incinerator from rumor through operation". *Journal of Environmental Economics and Management*, 28(2), 241–255.

Lumley, T. (2010). Complex Surveys: *A Guide to Analysis Using R.* Hoboken, N.J: Wiley-Blackwell.

Philibert, A., Ravit, M., Ridde, V., Dossa, I., Bonnet, E., Bédécarrats, F., & Dumont, A. (s. d.). "Maternal and neonatal health impact of Obstetrical Risk Insurance scheme in Mauritania: a controlled before-and-after study". *Health Policy and Planning*, in press.

Sander, W. (1992). "The effect of women's schooling on fertility". *Economics Letters*, 40(2), 229–233.

Wooldridge, J. (2010). *Econometric analysis of cross section and panel data.* MIT press.

Wooldridge, J. (2011). *Introductory econometrics.* South-Western.