Tapping into existing household survey data for research or policy use: hands-on exercise on water access in Kinshasa

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

2016-10-20

Abstract: In most countries, national statistical offices periodically run large surveys that provide outstanding insights on several subjects: social, economic, health, cultural, political. In many cases, this data is only used to produce nationally aggregated indicators that feed international statistical portals (WDGs, SDGs, World Bank, WHO, United Nations…). However, it is possible to do much more with the raw data collected during these surveys: calculate other indicators, cross different variables, run analysis for subnational areas, as well as more sophisticated analysis. This survey data is accessible to scholars, students, policy makers or practitioners and is most useful for the appraisal, monitoring and evaluation of development projects, programs and public policies. We first describe the typology of data generally available in developing countries, its possible uses and how to obtain it. Then we provide detailed guidelines on how to analyze it, using the *survey* package, available for *R*, a free and open source statistical software. We illustrate the step by step procedure for survey data processing using Democratic Republic of the Congo as an example: several surveys of different type and from different dates (MICS 2010, 1-2-3 2012 and DHS 2014) are analyzed to understand the levels and trends of access to drinking water, in particular in the Kinshasa megacity.

Introduction

We have entered an Open Data era and witnessed a clear improvement of the diversity, quantity and quality of freely available information. This trend also occurred in developing countries. Agence française de développement (AFD) evaluation unit wants to help practitioners, researchers and development stakeholders to better harness the potential of existing administrative databases, operational information systems, remote sensing imagery, big data and national surveys. Regarding national surveys in particular, we can draw upon a wealth of reliable, harmonized and recurring data collections regarding household social, economic and health situation, individual perceptions of public institutions, basic services and governance, firms, informal sector or agriculture. Analyzing such existing data provides, with almost no cost and delays, relevant and accurate information to more precisely describe and understand socio-economic contexts. It also provides baseline, follow-up and end-line information that is useful to evaluate projects, programs or policies.

Many policy makers and practitioners cite and use statistics on developing countries' situation that are disseminated by the World Bank, OECD, the World Health Organization and other UN agencies, without knowing where this data comes from. At the same time, they often think the data available at national levels is lacking or untrustworthy. In reality, the vast majority of this data is produced at national level by the domestic statistical systems – and predominantly through surveys. International institutions only curate, compile and disseminate this information. Governments and donors spend millions in financing local data collection. Still, it is almost only used to produce very general and aggregated statistics. However, we will show that much more can be done with the raw data produced during these surveys, and in particular that it can be used to fill the gap of knowledge we have on the context and outcomes of the interventions we finance.

The survey landscape is quite rich in developing countries:

- Multiple indicators cluster surveys (MICS) and Demographic and health surveys (DHS) focus on social, demographic and health aspects. They are typically based on a 10 to 40 thousand household sample and repeated every 3 to 5 years.
- Living standard measurement surveys (LSMS) are supported and supervised by the World Bank. They focus on social and economic aspects, including the household budget appraisal that is necessary to calculate country poverty ratios, national incomes and inflation rates. LSMS are typically applied on samples made of 5 to 15 thousand household and they repeated every 4 to 6 years. An increasing number of these surveys include now panel components: smaller samples of 2 to 4 thousand households are re-interviewed every year or every two years between larger surveys, in order to monitor trends. In many countries, living standard measurement survey have been completed through a "1-2-3" approach, that is adding prior to budget analysis ("3") substantial components on employment ("1") and informal enterprises ("2"). Both aspects are characterized by a high degree of informality in developing economies and are thus poorly reflected by conventional methods relying on administrative registries and must be approached at household level. Some of the LSMS surveys also adapted their methods to gather specific data on agriculture.
- Less demanding survey gathering only key well-being indicators. They have been/become widespread since 2000, to facilitate the reporting on the World Development Goals and other basic socio-economic and health variables, while going into less depth than specialized surveys.
- A wealth of information on socio-political and governance aspects is also produced through surveys. The aforementioned health, economic or wellbeing surveys often include additional modules to gather citizen's experience and perceptions on public institutions (police, justice, local authorities, etc.), conflicts, security, corruption and social cohesion. Standalone annual or biannual surveys like Afrobarometer or Gallup World Poll also specialized on such socio-political aspects. Other specific issues are also tackled by ad hoc surveys, like Global Findex 2011 and 2014 for financial inclusion.

Most of this data is made accessible through online repositories, such as ihsn.org, dhsprogram.com, mics.unicef.org, microdata.worldbank.org, etc. This way, it can be freely downloaded, often simply after filling a short information form and committing to respect some essential ethical principles, stating in particular not to disseminate microdata without authorization, not to use it for commercial or politician means and not to try to breach respondent anonymity.

These surveys are subject to a close quality control by several instances. The national statistical offices follow standards and guidelines at several stages: recruitment of skilled surveyors and training on each specific survey, questionnaire pilot testing, field supervision, audit on a sample of filled questionnaires, double data entry, etc. In almost all cases, they also benefit from support and supervision from international entity: World Bank LSMS team, UNICEF MICS team, ICF International DHS team, AfriStat, etc. Moreover, it would be technically very hard to falsify some of the respondents' information while maintaining the coherence of detailed and intertwined answers of several household individuals on hundreds, respondent answers on thousands of records, as well as the plausibility of data comparison between subsequent surveys. Open access is therefore a factor of reliability: potential inconsistencies can quite easily be identified by any user when processing the data for statistical analysis.

The principles of complex survey analysis

Household surveys are based on a stratified two-stage cluster sampling scheme. In other words, they are based on the zoning of the entire national territory into several thousand area units ("clusters"), each of which, according to the previous census, comprises an equivalent number of households. Several hundred clusters are randomly selected and in each of these, an exhaustive population census is conducted. Ten to twenty-five households are randomly drawn from each cluster census list to form the sample.
The representativeness of a survey stems from the fact that all households in the country have the same chance of being drawn. Due to some lack of frequency or reliability of the national census, some disparities

in population numbers between the areas may however introduce an equiprobability bias. A comparison between the number of randomly drawn households in each area and the total number of households living in the same area is used to calculate a weight that translates its probability of being drawn and thus corrects the bias in all the calculations that are taken into account to compute all the indicators and error margins estimated from the survey.
Estimates and confidence intervals are calculated using the Horvitz-Thomson method (Lumley, 2010, p. 3-12). To be able to calculate valid population estimates, every individual in the population must have a non-zero probability of ending in the sample, which must be known for every individual who does end up in the sample. To calculate the accuracy of the estimate, every pair of individuals in the sample must have a non-zero probability of both ending up in the sample which must be known for every pair that does end up in the sample.
These probabilities enable to compute the Horvitz-Thompson estimator of the variables for the population total and the variance estimate.

This computation is quite straightforward and largely facilitated by specific functionalities offered by statistical package, such as svy for Stata or survey for R. It however implies to obtain a precise description of the overall sampling methodology and to get, for each surveyed unit in the dataset, its sample weight and the strata it belongs to. The subsequent data analysis procedure always begins with the reproduction of some estimates published in the official report to make sure the survey parameters have been correctly specified.

Applications

The primary purpose of the method we've described is to produce useful **descriptive statistics** for development project design and monitoring. We roughly estimate that about two thirds of the variables collected at household level through DHS or MICS questionnaires are actually analyzed in the published reports. Yet, many of the social, demographic and health indicators are only calculated as national aggregates, without being cross-tabulated by region, rural and urban areas or household wealth status. The proportion of gathered information that is rendered in the official reports is clearly lower for most of socio-economic surveys like LSMS, where a wealth of data on household budgets is collected at household level, only to be computed at national level to monitor consumer price index, gross incomes or poverty rates. We even found several cases where reports still hadn't been published 2 to 4 years after a socio-economic survey was implemented and its data made available.

> **Caution**: While narrowing the scope of calculation to local areas and subpopulation, margins of error increase and it is crucial to take it into account not to overstate the significance of the results. This is why standard errors and/or confidence intervals are always calculated and disclosed within every data analysis report. Except in rare occasions where it might seem relevant to mention it (for instance to say that the data does not allow for relevant comparisons) it is not recommended to produce estimate visualization or comments when confidence interval overlaps impede relevant interpretation.

A second purpose of this approach is to **follow up trends**, ie.the evolution of outcomes of interest over time (which is not impact evaluation). It can provide policy makers and practitioners with crucial insights that are otherwise lacking, like knowing what is the rate of access to electricity before and after a grid extension program in a certain area, or similar details regarding assisted birth levels for maternal health policies or improved water and sanitation use for water access programs.

Proving a causal effect for **impact evaluation** purpose requires a more demanding statistical analysis. Except for panel surveys, national household survey samples are drawn independently across iterations. It is therefore unlikely for a household sampled in a survey to be sampled again in the next round a few years later. Even if a same household is sampled in two subsequent surveys, observations are anonymized in a way that impedes to identify it across both datasets. Therefore, most survey data provide pooled cross sections rather than panel data. The compared benefits of panel data and pooled crossed sections for impact evaluation

have been extensively commented in the literature (A. S. Deaton, 1997; Wooldridge, 2010, 2011, p. 432-484): panel data are in principle preferable to disentangle intertwined structural factors, but samples are generally small and hard to maintain over time and over time, they tend to lose their proprieties of representativeness of the overall population. Pooled cross sections also have strengths: they rely on significantly larger samples, providing more precise estimators and tests statistics with more power are designed to maximize population representativeness and are repeated with harmonized methods over decades. Major policy impact evaluations that are today considered as textbook classics have been implemented using pooled cross sections from household surveys (A. Deaton, 1985; Kiel & McClain, 1995; Sander, 1992). A recent impact evaluation supported by AFD also consisted in pooling cross sectional data to assess the impact of a maternal health financial scheme in Mauritania (Philibert et al., s.d.).

Fostering open and reproducible analysis

At AFD evaluation unit, we want to foster the use of such data and methods by developing countries specialized Ministries, local stakeholders, researchers and consulting first. We implement most of this analysis in R (open source statistical package and language) and publish our reports in RMarkdown, ie. the document includes all calculation scripts. This allows to automatically update all the results if some data or parameter is modified. More importantly, it facilitates results verification and reproduction by others.

Hands-on with an example on access to water in Kinshasa

During the last 10 years, the Democratic Republic of the Congo (DRC) National Statistics Institute implemented the main types of household surveys: MICS, DHS and LSMS/1-2-3. We will guide you through the process of gathering as analyzing this data as a practical example on how to use survey data for your own research.

Data access and download

Finding the available data of interest

To identify all the available surveys for your country of interest, it is usually preferable to go through each of the websites mentioned at the beginning of this post. Regarding water and sanitation issues however, the WHO-UNICEF Joint Monitoring Program mainains an updated list of all available information in order to organize the corresponding World Development Goals monitoring. This makes it much easier to find the most recent surveys addressing Water Supply and Sanitation issues in a country. Simply download the country file for you aera of interest (for this example the DRC).
In this excel file, look at the sheet **"Tables_W"**. The rightmost survey is the last completed, for DRC, the DHS 2014 survey.

By using the already processed information contained in this sheet, we can build the following graphic on the use of improved water sources in urban DRC's population.

Figure 1 shows the estimates published in the end-of-survey reports. As these data are calculated for all urban areas in the DRC, it is impossible to separate out water access information specifically for the city of Kinshasa. We attempted to do this by using the raw data from these surveys. Each survey was in fact based on a sample constructed to be statistically representative of certain regions, and, each time, Kinshasa was selected as one of the regions whose drawn sample produced satisfactory estimates (see Table 1).

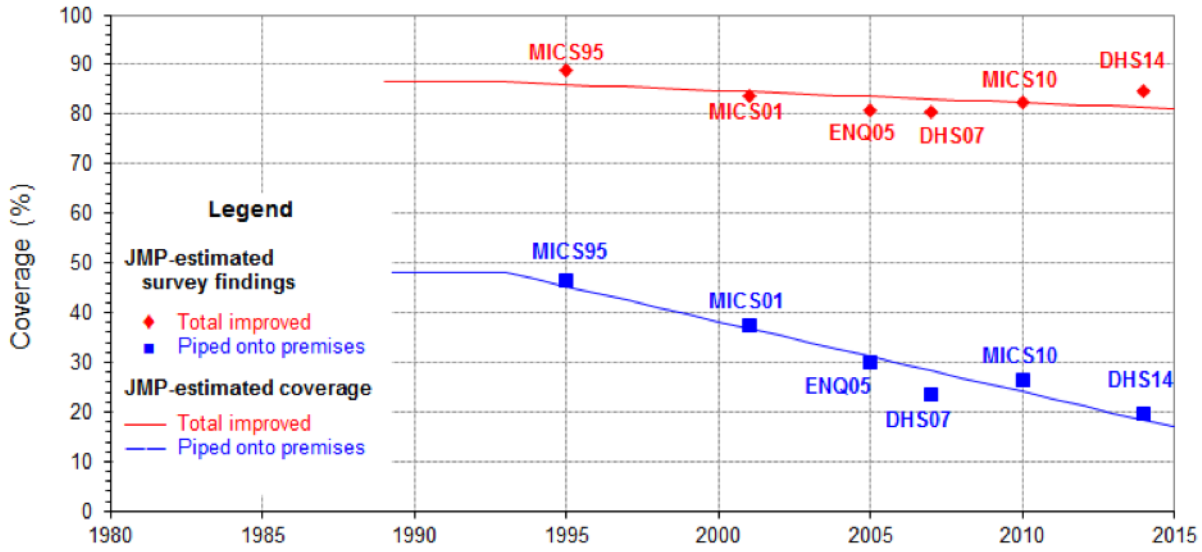

Figure 1: Estimated proportion of the urban population in the DRC using improved drinking water sources, from the Joint Monitoring Program country files

Survey	Supervised by	Collection period for Kinshasa	No. of surveyed households	No. of households surveyed in Kinshasa
DHS 2007	Macro International	February–August 2007	8,886	2,665
MICS 2010	UNICEF	February–April 2010	11,393	1,004
1-2-3 2012	Afristat and DIAL	October–December 2012	21,454	1,969
DHS 2013-2014	ICF Macro	August–September 2013	18,360	1,224

Table 1: Representation of Kinshasa in recent surveys carried out in DRC

DHS and MICS surveys gather detailed information on how water is accessed, distance to a water source, type of sanitation, hygiene practices, water storage, diarrheal diseases. Their surveys also include summary information on household living standards (approximated on the basis of assets) and the main economic activity of household members. The 1-2-3 surveys collect only some of this WASH (Water, Sanitation and Hygiene) information, using slightly different criteria: user access to water and sanitation, occurrence of diarrheal diseases. On the other hand, the information contains much more detailed socioeconomic data on household activity, income source and spending. The DHS surveys provide GPS coordinates for groups of households. The 1-2-3 surveys provide the exact address of each surveyed household. These surveys enabled us to describe the water access situation in Kinshasa and its outskirts.

Download DHS and MICS data bases

To realize a small statistical analysis, we will use this four surveys. DHS and MICS are directly available online. 1-2-3 survey must be requested contacting directly the National Statistics Institute (NSI).

Download DHS surveys

To download the two DHS surveys, you will need to register on the DHS website, filling out the form in Figure 2.

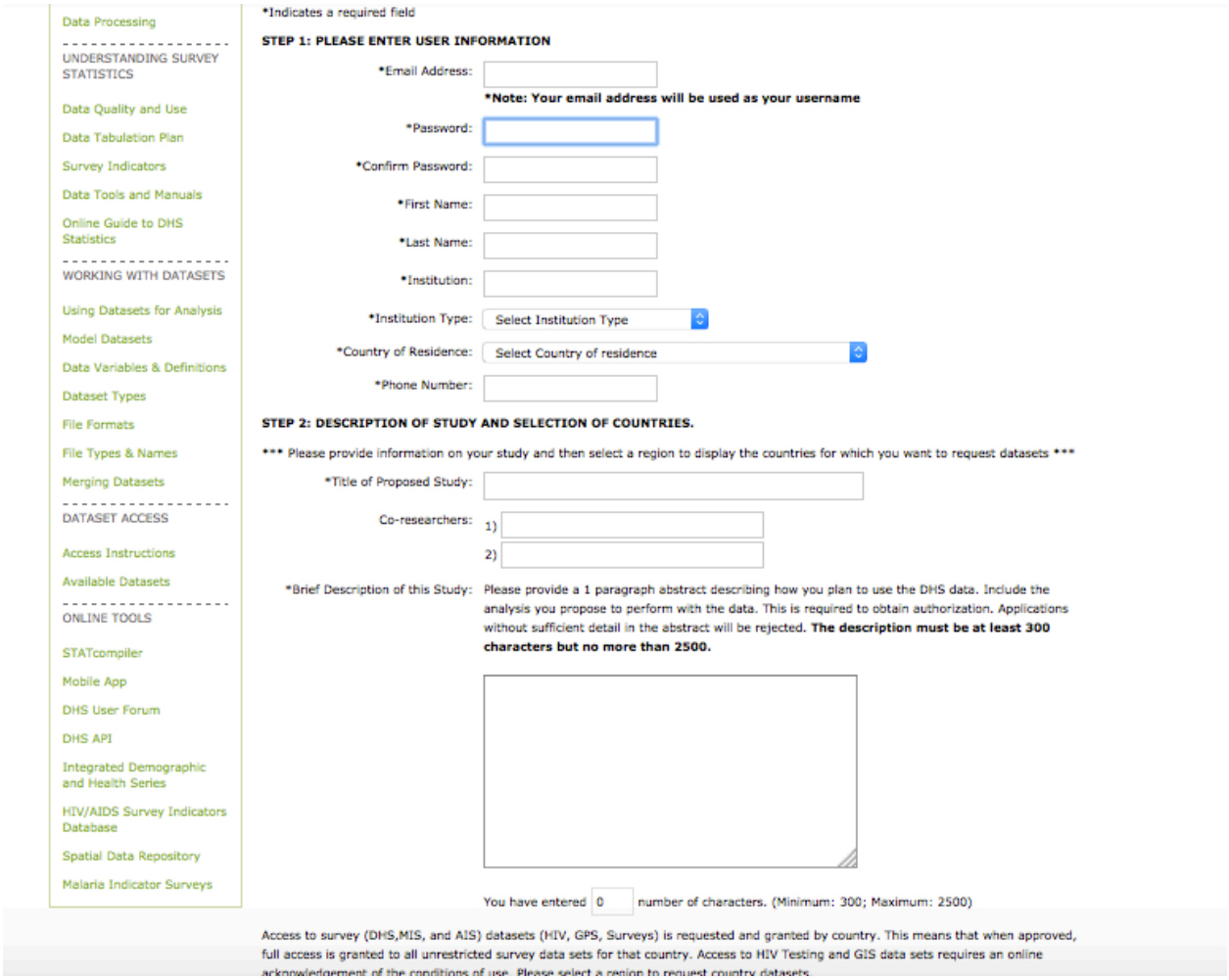

Figure 2: Screenshot of the DHS inscription and suvery query form

Once you've provided the required information, link it with the corresponding datasets you're requesting. Here we select "Sub-Saharan Africa" in **Select Region**, then select "DRC" in **Select country** and tick **Survey** and **GPS** checkboxes.

Survey questionnaires gather information on several units of analysis: households, female and male adults, children and births for social and health surveys (households, individuals, farms, informal production unit, expenses, etc. for LSMS). Here our interest focus on household as it is at household level that access to water is assessed.

You will receive an email a few hours later to allow you to download the data. In our case, Download the surveys for 2007 and 2014.

The files downloaded can be unzipped and copied directly in your working directory, so you can load them directly with their name, without having to specify "the/full/path/to/the/directory/you/stored/them/in".

Download MICS survey data

The procedure is similar for MICS data. Register on the corresponding section of UNICEF website, then click on **SURVEYS** and select in **Any Country** "Congo, Democratic Republic of Congo". Here you will be immediately allowed to download the 2010 dataset.

Data analysis with DHS and MICS surveys

A prerequisite is of course to install R statistical software. We strongly recommend to install also RStudio as it provides a very ergonomic integrated development interface (ie. windows, menus, and a lot of convenient functionalities) that will make your life with R easier.

Install packages

R basic package contains a large number functions that are required to perform the most common statistical tasks. The installation of packages provides additional functions that enable to perform more specific tasks.We will use the packages *foreign* to read the datasets that are provided by MICS and DHS in SPSS format. The package *survey*, developped by Thomas Lumley, is crucial to facilitate estimations for complex survey designs (here two-stage cluster samples). The *ggplot2* is very convenient and flexible to produce graphics.

The following commands install the required packages:

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

And then tell R that we will use it during this session:

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

Open and prepare data

We will now open the data:

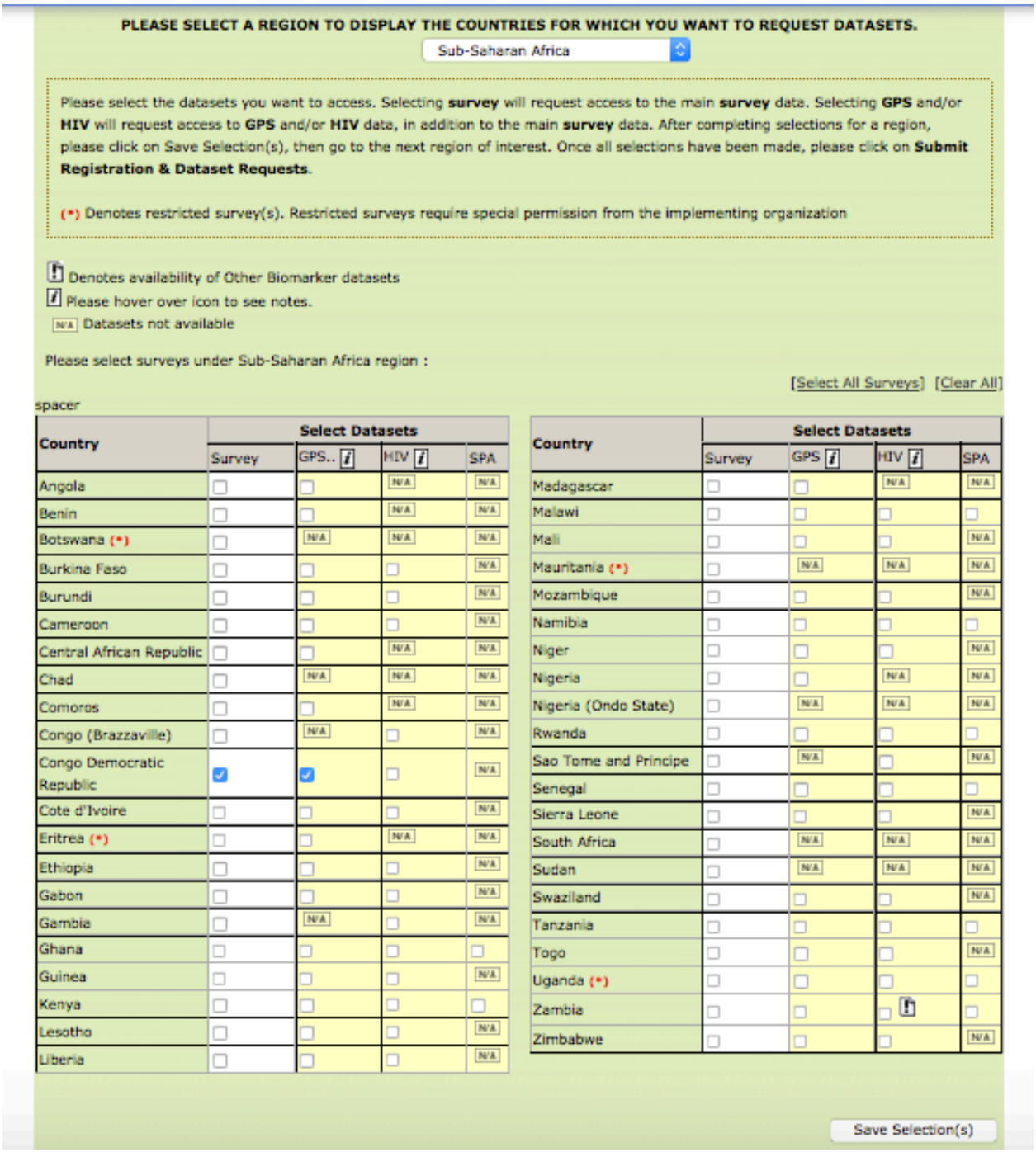

Figure 3: Screenshot of the different files that can be downloaded for a survey

```
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)
```

The DHS suvey codebook is available on DHS website, together with a great number of useful documents, and similar resources can be found on MICS website. The name and codification of the variables can also be explored in the environment window on RStudio.

It is generally very important to also collect and review the survey questionnaire (always published on survey pages and/or at the end of official survey reports), to make sure the variable codification is correctly interpreted.

For our example, we restrict the analysis to Kinshasa province:

```
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")
```

In DHS the sample weights are in base 1000000, while literature suggests weight average should be 1 (Lumley, 2010 p. 10). In our tests we obtain the same results for estimates and confidence interval with base 1 and 1000000 sample weights. But in any case, we will set the DHS sample weights in base 1.

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

Evolution of private connections in Kinshasa

Our first production is a graphic presenting the evolution of private connections in Kinshasa between 2007 and 2014 with the DHS and MICS survey.

We have to create a new variable to separate the households with a private connection from the other. We will use as basis the variables **HV201** for DHS and **WS1** for MICS, which code the household's main water source in 15 categories. To build the private connection variable, we are only interested in two of them, the *Piped into dwelling* and *Piped to yard/plot*.

We need first to categorize all the household, for which the water source is known, in the "Other water sources" category.

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

The part `{r, eval=FALSE}is.na(DHS_2014$HV201)==FALSE` gives the condition that the answer to the HV201 variable was not missing.

Then we categorize all the households, which are *Piped into dwelling* and *Piped to yard/plot* as having a private connection:

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

We do the same for DHS 2007 and MICS 2010. For MICS 2010, the WS1 categories are in french and we are interested in *Robinet dans le logement* and *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"
```

Then, we have to specify the sample's designs with the function *svydesign* from survey package. The design enable to take into account the survey's statistical structure to compute robust standard errors and confidence intervals for our statistic. We enter the following information and the package will use them for statistical computation:

- ids = The sample units, first the cluster and second the household
- strata = the stratification level (region and environment)
- weights = the sample weight
- data = data

```
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)
```

The svydesign function re-creates a database with the original data, but adding attributes that are necessary for the calculations taking into account the complex sample design.

Now, we prepare the data for the graphic by building a common database with the computed proportions.

First, we compute the proportion of each type of connections in Kinshasa using the *svymean* function. We enter the following information:

- ~source: the variable for which we want the proportion ("~" indicates that it a variable included in the database used: design_2014)
- design: the design we specify for the wanted survey
- na.rm=TRUE: the NA should not been counted in the mean
- vartype = c('se'): we want to obtain the proportion standard error

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

Second, we select only the *private connection* category, it is on the new dataframe's second line. But we need all the dataframe columns, so we tell it to conserve the two columns but only for the second line.

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

We do the same for DHS 2007 and 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) ]
```

We now want to construct a single dataframe with this three separate dataframe. We will binding them by rows using the *rbind* function:

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

And we create a new variable called *annee*, containing the survey's names and year:

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

Finally, we construct the graphic with **ggplot2** package. The dataframe has three columns that we will use for the graphic: *annee*, *mean* and *SE*. First, we use the ggplot function and specify the basic elements of the graphic:

- the dataframe *source_evo*
- the axis in aes: in x-axis the variable *annee* and in y-axis the variable *mean*

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

Then we add the different elements:

```
graph_source_evo +
  # the errorbar, computed by adding and substrating the standard error to the mean
  geom_errorbar(aes(ymin=mean-SE, ymax=mean + SE), width=.2) +
  geom_line() +   # the line between the two limits
  geom_point(size=2)+   # the point at the level of the mean
  xlab("")+   # no title on the x-axis
  ylab("Proportion") + # a title on the y-axis
  ylim(0, 0.6) + # scale indication for the y-axis
  # text placement on the x-axis
  theme (axis.text.x = element_text(angle=90, vjust=0.5, size=11))
  ggtitle("Evolution of the rate of private connections in Kinshasa according
          to the social and health surveys conducted between 2007 and 2013")
```

Sharp disparities between the different residential area

Sharp disparities are observed between the different residential area.

We have used the cluster location to categorize the households between two main areas: the ASUREP and the rest of the city. The ASUREP districts are area chosen by the AFD and other funders for the PILAEP and WASH water supply projects in Kinshasa. As shown in Figure 5, these districts are in the city suburbs and should be poorer with a worse water connection than the other districts. We will try to confirm it with the survey data.

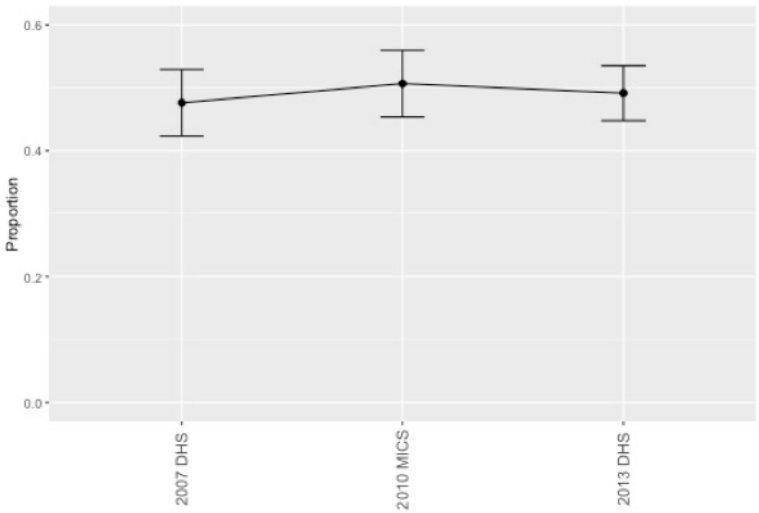

Figure 4: Evolution of the rate of private connections in Kinshasa according to the social and health surveys conducted between 2007 and 2013

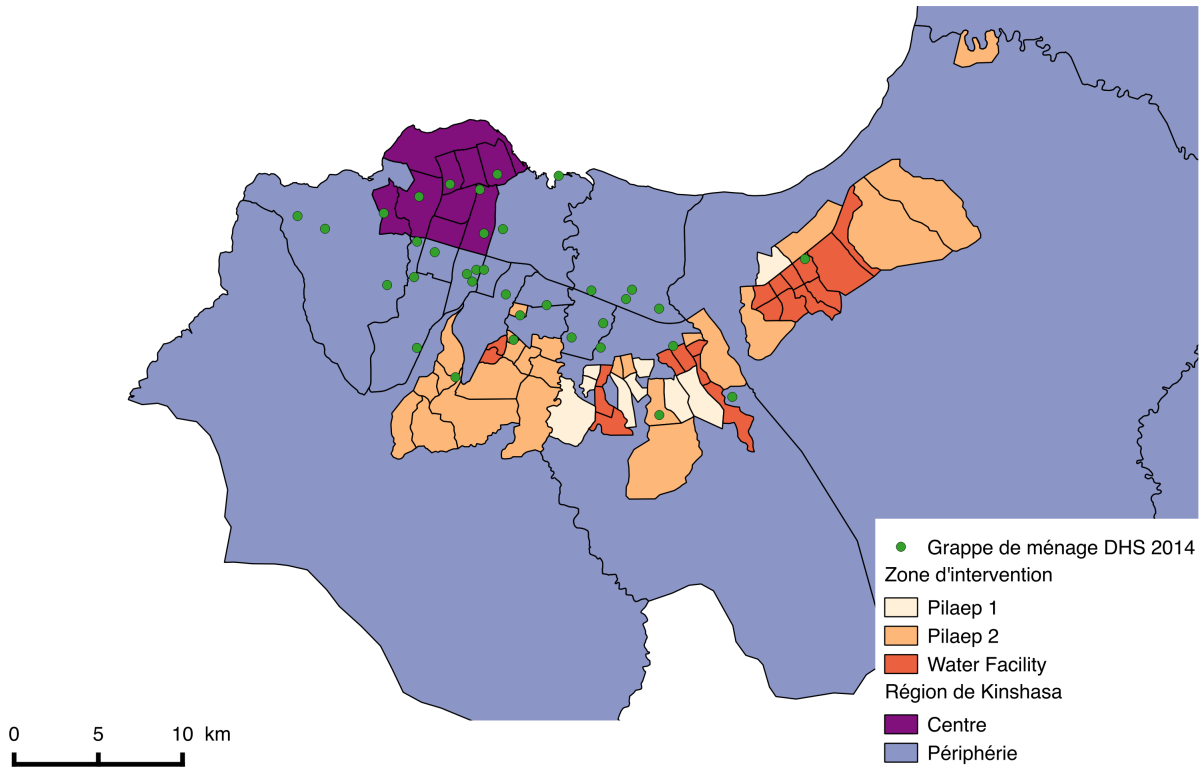

Figure 5: Location of sampled clusters for DHS 2014

In a future post, we will explain how to handle spatial data and create maps with R. But for now we will skip this part. From the spatial analysis, we use a csv file in which we have allocated an area to each cluster, as illustrated in Figure 6.

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: Geographic classification of Kinshasa sampled clusters

We open the csv file with the *read.csv2* function (specially adapted to csv files generated by MS Office).

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

We merge the DHS 2014 and csv file to incorporate the new information to our DHS_2014 database. To do so, we create a new variable called *matching*, which is similar in the two files. We will use the *HV001* variable which contains the cluster identification in DHS survey and the *DHSCLUST* variable which contains the same information in the csv file.

```
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")
```

As we modify the database, we need to specify the design again:

```
design_2014 <- svydesign (ids=~HV021+HV002, strata= ~HV023, weights=~weight, data=DHS_2014)
```

We compute the proportion of private connection by area and their standard errors with the survey's function *svyby*. It is very similar to svymean but there is two new elements to ad in the information:

- by: the category for which the proportion should be computed
- FUN: the function that we want to use (for proportion: svymean)
- vartype = c('ci'): we want to obtain the confidence interval boundaries (and not standard error)

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

And we prepare the data frame to be used in the graphic.

We generate a new variable with the exact name of the area:

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

This time, we conserve all the line but selct only the column with the private connection information, the confidence intervals boundaries and the area names

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

Finally, we build the graphic:

```
### specify the x and y axes
graph_source_zone <- ggplot(source_zone, aes(x=Zone_1, y=Raccordement))
graph_source_zone +
  # specify the graphic type and characteristics:
  #here a bar graph with choosen filling color
  geom_bar(position=position_dodge(), stat="identity", color="black",
           fill=c("#003399", "#0099FF")) +
  geom_errorbar(aes(ymin=source_zone$ci.l, ymax=source_zone$ci.u),
                width=.2,
  # specify the position of the error bar with the confidence interval's boudaries
                position=position_dodge(.9))+
  geom_line(position=position_dodge(0.9)) + # add line between the boudaries
  geom_point(position=position_dodge(0.9)) + # add a point in the middle
  xlab("Areas")+ # add a title to x-axis
  ylab("Proportion") + # add a title to y-axis
  # x-axis text characteristics
  theme (axis.text.x = element_text(angle=90, vjust=0.5, size=11))
```

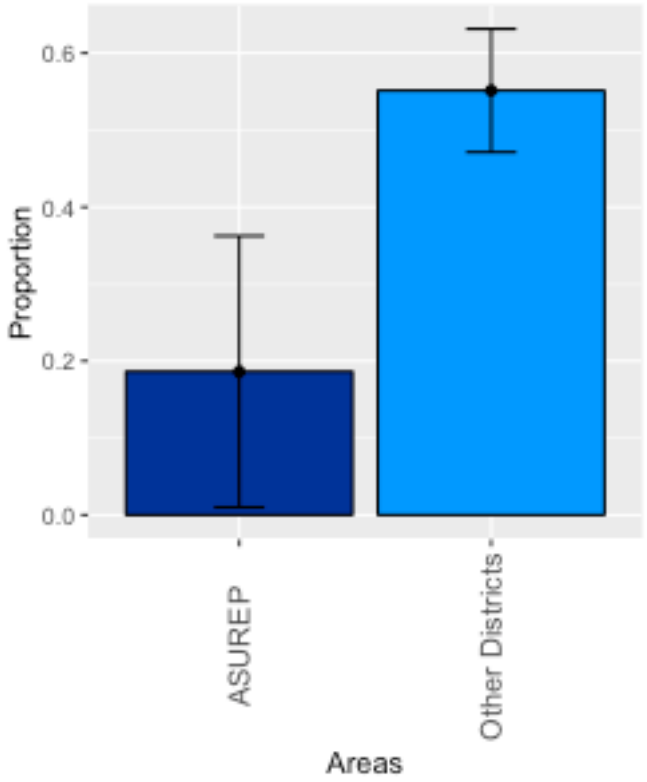

Figure 7: Difference in the rate of private connections estimated by the 2013 DHS Survey

Data Analysis with 1-2-3 survey

The 1-2-3 survey bring more information on household's economic situation. It is not freely available as DHS and MICS survey and we thanks the DRC's national statistical institute and DIAL laboratory to authorize us to use it.

Comparison between DHS-2014 and 1-2-3 2012 private connection rate in ASUREP and the rest of the city

First we open 1-2-3 household's database and merge it with a csv file where the site are allocated to the ASUREP and the rest of the city

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

And we merge the two file thanks to the *SITE* variable:

```
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")
```

We modify the weight variable:

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

And we generate a new variable to identify the households having a private connection, based on the answer to the *H10* variable:

```
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"
```

We specify the design as for the other database:

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

We compute the proportion by area with the *svyby* function:

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

We create the Zone variable and select only the column that interet us for the graph:

```
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")
```

We combine this dataframe with the dataframe from *source_zone* dataframe issued from 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")
```

And we construct the graphic:

```
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))
```

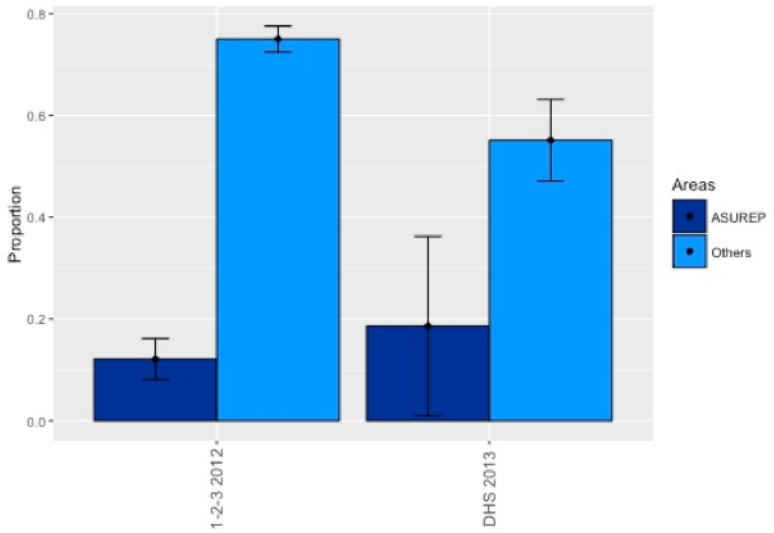

Figure 8: Difference in the rate of private connections estimated by the 2012 1-2-3 Survey and the 2013 EDS Survey

There is significant differences between the two surveys' results. There is three possible explanations to this differences. First, the two surveys have not been realised the same year so it can be some small changes in the water access situation. It can also be small differences in the water sources definition.

Second, the sample are drawn from the last RDC census. But the last one has been realized in 1984 and is not representative of the DRC's population. The teams which draw the sample try to actualize it but it is very difficult for a city like Kinshasa, which have grow very quickly in the past 10 years. Especially, the ASUREP districts are recent and so worse capted than richer districts from the city center. The 1-2-3 survey has done a particular effort to capt the poorest households and city area. There is many more households surveyed in this districts, that is why the estimate is more precise. For all the city, the 1-2-3 survey focuses more on the Kinshasa population and on wealth disparities, so the sample was bigger and the final statistics more precise.

Third, the sample has not been drawn to be representative at a smaller scale than the entire city. The statistics for smaller area can be biased and not perfectly representative of the total area. It gives us a confirmation of disparities and an approximation of the possible rate but are not the true real rate.

Analyzing water spendings with 1-2-3 survey

We have to open a new database with information on households expenses:

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

We merge it with the household database by creating a similar individual identification variable:

```
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")
```

We restrict the data base to the households living in Kinshasa:

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

Then, we construct the variable that regroup the household's water expenditures and their proportion in total budget by setting as missing values the 0 and 9999999 answers and adding bills *H18* and the other expenditures *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
```

And we build the data frame that will be used for the graphic:

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

And finally the graphic:

```
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)
```

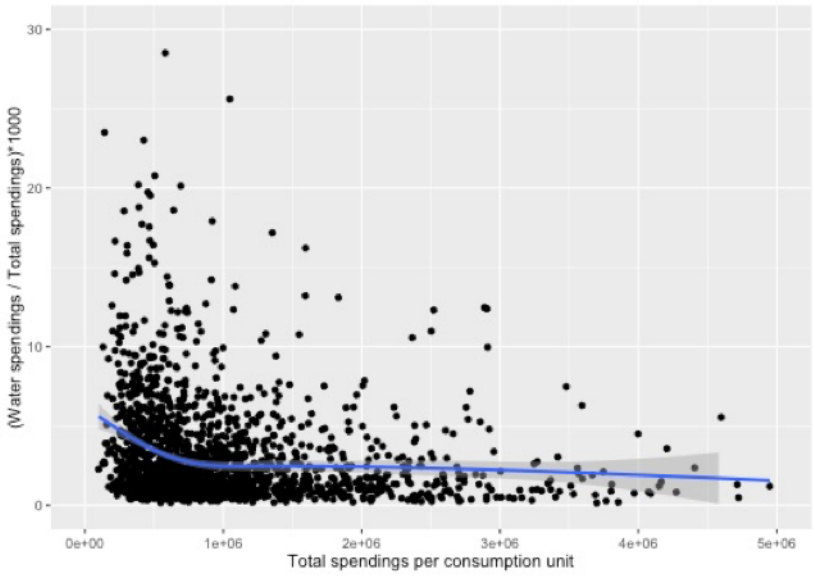

Figure 9: Difference in the rate of private connections estimated by the 2013 DHS Survey

This figure clearly shows that households generally spend less than 1% of their budget on water, which is consistent with the trends observed for the rest of the continent (Banerjee et al., 2008). Constently with the literature, we observe that expense amounts increase with rising income levels, mainly because middle to high-income households are disproportionately connected to formal water networks. But they decrease in percentage of household budget, as the poor spend a higher part of their budget to buy water because higher prices are paid by unserved consumers of alternative suppliers. Beyond the financial aspects, the time spent in fetching free water represents a significant opportunity cost (Hutton, Haller, & Bartram, 2007). This hidden cost is clearly shown by the time that households devote to water collection chores.

Average time per trip according to DHS 2014

For this table, we will use again the DHS 2014 survey. We begin by recoding the time to water source variable:

```
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
```

We analyze the data and construct a table with the number of interviewed households:

```
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

This table summarizes the average time needed by household members for a round trip to their main source to fetch water. In the City center, this time is about 3 minutes, versus 28 minutes in the neighborhoods targeted by the projects supporting ASUREP creation. The "min" and "max" column provide the lower and upper limits of the confidence interval.

Conclusion

We prepared this guide to provide a didactic explanation on why and how retrieving raw data from national statistical surveys and how to analyse them with *R*, thanks to the *survey* package. To illustrate our tutorial, we developed an example, i.e. the state and trends of access to drinking water in Kinshasa, DRC. With this document, we hope to have contributed to draw scholars', students', policy makers' and practitioners' attention to the potential of these sources to improve appraisal, monitoring and evaluation of development projects, programs and public policies.

References

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. 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, available online

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

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 (Édition : 1). 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.