"Science Stays True Here"
Journal of Mathematics and Statistical Science, 104-109 | Science Signpost Publishing

Identifying Hard-to-Survey Provinces in Burkina Faso

Baguinébié Bazongo[*]

Department of statistical coordination and research
Institut national de la statistique et de la démographie, Ouagadougou, Burkina Faso

Abstract

Probabilistic sampling is theoretical the best method to select representative sample from target population. However, this result does not always hold in practice because some selected sampling units are missed during survey. The aim of this paper is to categorize the 45 provinces of Burkina Faso in terms of household surveys complexity in order to adapt specific methodology to each group of similar provinces.

We applied hierarchical clustering based on principal components analysis on provinces data and found five clusters. According to these results, two provinces (Kouritenga and Loroum) were the most hard-to-survey. Therefore, adapted methodology should be applied to each group of provinces during survey implementation to maximize data quality.

Keywords: hard-to-survey, hierarchical clustering, FactoMineR

Introduction

Household surveys are widely used to assess individuals living conditions, particularly poverty in developing countries. These surveys are based on probability sampling methods because “reliable estimates can be produced along with estimates of the sampling error and inferences can be made about the population” (Statistics Canada, 2010). The effectiveness of probability sampling assumes that variables of interest are measured for all units in the sample. Unfortunately, this assumption is not always true in practice because some units may be hard to survey. The hardness results from unavailability of certain units, high mobility (nomadic), language barriers, low cognitive abilities, refusals, etc. (Roger Tourangeau et al. 2014). Therefore, a sample drawn with probability method may be unrepresentative as some categories of units are under-represented in the sample. Low participation in the survey will decline the quality of data given increase of non-response error, coverage error and decrease of sample size (Roger Behrens et al. 2008).

The National Institute of Statistics and Demography conducts regularly household and businesses surveys to monitor and evaluate national development policies in Burkina Faso. Thus, seven household surveys have been conducted since 1994 with probability sampling to assess individuals' poverty. However, few studies have used data from these surveys to improve the quality of future surveys. Indeed, we do nothave clear answers to the following questions: (1) does interviewers

[*] Head of Research and Methodology Unit, bazongo15@yahoo.fr, 00226 74 33 90 08

experience the same difficulties when conducting survey in the provinces of Burkina Faso? (2) Are there groups of provinces with similarity in terms of complexity? The answers to these questions will help survey statisticians to design and adapt specific methodology for each group of provinces. This methodology includes communication strategy, workload distribution, recruitment of interviewers, mode of data collection, etc. The aim of this paper is to categorize the 45 provinces of Burkina Fasoin terms of household survey complexity to adapt specific methodology to each group of similar provinces.

Data and methods

Data come from household survey conducted in 2014 by the National Institute of Statistic and Demography. A random sample of 10,800 households was drawn from non-institutional households using stratified probability sampling. Interviewers were trained to conduct face-to-face interviews with paper questionnaire, and data were collected from January to Mars, 2014. The aims of this survey were to assess households living conditions, particularly monetary poverty. The measurements included individual background, education, health, employment, cash transfer and household consumption.

We selected five variables to assess survey complexity at household level and aggregated data at province level using average function. Each variable is described in the table 1 below. After aggregating data, we applied hierarchical clustering based on principal components analysis to categorize provinces (François Husson et al., 2010). FactoMineR package from R.3.3 was used to implement this method and clusters map was created with QGIS 2.14.3.

Table 1. Description of variables of interest

Variable name	Variable description
replacement	Whether the household is drawn firstly (value = 0) to be interview or secondly (value = 1) to replace another household who misses the interview INT: 12 households were drawn in each primary Unit (enumeration area). If the selected household misses the interview, he is replaced by another household until 12 households are successfully interviewed.
language	Whether interview language is French (value = 0) or local language (value = 1) INT: The questionnaire was designed in French and each interviewer should translate question in respondent language. Interviewers were trained to do so.
missing	Number of missing values per household INT: Each observation contains 14 individuals’ background variables at household level.
duration	The duration of interview (minutes)
interview	Whether respondent was reluctant to participate to the interview after interviewer introducing the aims of the survey
hhsize	Number of household members

Results

Sample description

A total of 10,800 non-institutional households were included in the dataset. The duration of interview ranged from 20 and 800 minutes, with a median of 80 minutes; 0.86[†]% of households expressed reluctance to interviews; 81.3% of them were interviewed in local languages, and 4.1% were missed during interviews and replaced by other households. The number of missing values per household (14 individual background variables) ranged from 0 to 43 with median equal to 4 missing values. Households' size was right skewed with median equal 6 members and a maximum of 63 members. When aggregating data at provinces level, the minimum number of households per province was 30 (Komandjoari) and the maximum was 1038 households (Kadiogo).

Provinces clustering

The dendrogram was cut into 5 clusters using visual criteria such as distance between clusters. Each cluster is characterized by the means and proportions of active variables. The significance of association between clusters and variables was tested using v.test (A. Morineau, 1992) and marked with star symbol (Table 2).

Table 2. Summaries of variables by cluster

Cluster Number	Cluster size	Percentage of missed households	average household size	Percentage of interview in local languages	average number of missing values	Interview duration (minutes)	Percentage of reluctances
1	18	4.2	6.5*	75.3*	4.1*	82.9	0.9
2	22	2.6*	8.0	91.2*	5.3	87.5	0.4
3	3	9.0*	8.6	81.3	6.9*	100.2	1.5
4	1	2.1	8.9	85.5	5.9	115.4	5.4*
5	1	0.0	12.2*	93.3	8.4	126.5*	0.0
Global	45	3.3	7.5	84.1	5.0	88.0	0.8

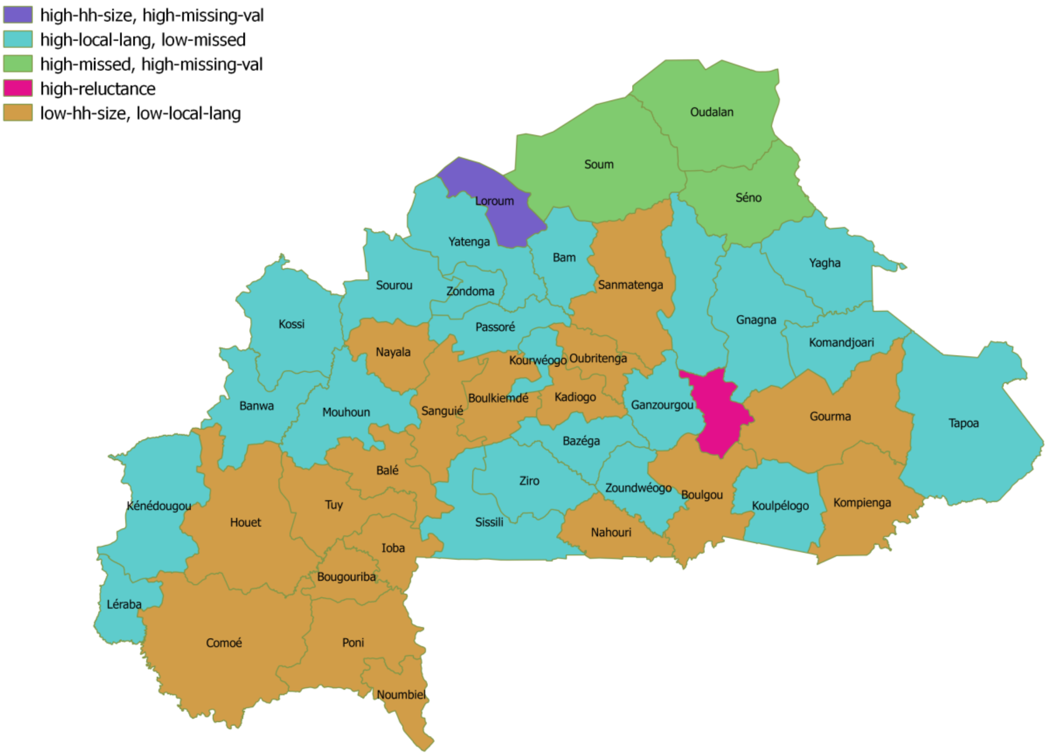
* = significant association between clusters and variable (|v.test|>3, pvalue<0.01)

Clusters characterization

The figure 1 below shows the five clusters of provinces in the map and the characterization of each cluster.

[†] This does not take into account households who missed interviews

Figure 1. Provinces characterization



Discussion

Hierarchical clustering suggests five groupsof homogenous provinces based on active variables. Group 1 consists of 18 provinces characterized by lower household size (6.5 individuals), lower percentage of interviews in local languages (75.3%) and lower number of missing values per household (4.1). Kadiogo and Houet provinces which contain the two biggest cities (Ouagadougou and Bobo Dioulasso) of the country are included in group 1. Moreover, provinces in group 1 have higher urbanization rate, higher education rate and lower percentage of polygamous. For group 2, we have 22 provinces characterized by higher percentage of interviews in locale languages (91.2%) and lower percentage of missed households (2.6%). These provinces have lower urbanization rate and lower education rate compared to those in group 1. The third group consists of 3 provinces characterized by higher percentage of missed households (9%) and higher number of missing values per household (6.9). One potential explanation is that most of respondents in these provinces are nomads who move from place to place to search for food and water for their livestock. Also, these provinces are bordering with Mali and Niger, and breeders move from one country to another.Given enumeration and interviews were not simultaneous, surveyors were more likely to miss respondents who were selected during enumeration for interviews. Group 4 with only one province (Kouritenga) is characterized by highest percentage of reluctant households (5.4%) and higher interview duration (115.4 minutes). This province contains one of the biggest commercial cities of Burkina Faso (Pouytenga) so that most of respondents are traders who are reluctant to provide information to interviewers. They fear that the latter provide information to tax authorities. The last group (group 5) have also one province (Loroum) characterized by higher household size (12.2) and higher interview duration (126.5 minutes). However, explanation about larger household size for this province is unclear.

Conclusion

The aim of our study was to identify similar provinces in terms of household survey complexity. Our results suggest fives clusters with different sizes.The first group contains provinces with higher urbanization rate and higher educated households. This group may require interviewers with higher education during household survey. In contrast, the second cluster consists of provinces with lower urbanization rate and lower educated households. So, this cluster requires interviewers who speak local languages. The third cluster however contains provinces with mobile households, and the fourth cluster contains province with reluctant households. As results, we should not apply the same survey methodology to the 45 provinces during survey design. Adapted methodology should be applied to each group of provinces based on their characteristics.

One limitation of our study is the small sample size in some provinces because sample was drawn to be representative at region level (province is a sublevel). Thus, some provinces may have unrepresentative sample. Moreover, some relevant variables such as the number of visits required to conduct an interview were not included in our study. Future household surveys should include questions to measure these variables that are more relevant to assess hard-to-survey population.

References

[1]. A Morineau (1992). *Tests et valeurs-Tests: application à l'étude de mastic utilisés dans la fabrication des vitraux*.Revue de statistique appliquée, tome 40, n4.

[2]. François HUSSON et al. (2009). *Analyse des données avec R*. Presses Universitaires de Rennes. Collection Pratique de la statistique.

[3]. Roger Behrens. Mark Freedman et Nancy McGuckin (2008). *The challenge of surveying 'hard to reach' groups*. 8th International Conference on Survey Methods in Transport: Harmonisation and Data Quality

[4]. Roger Tourangeau et al. (2014). *Hard-to-Survey Populations*. Cambridge University presse. ISBN: 9781107628717

[5]. Sharon L. Lohr (2010). *Sampling: Design and Analysis.* Second Edition.

Annexe

R scripts for data analysis

```
#Setting the working directory
setwd("---")
#Load library to read data
library(foreign)
#Read data
emc<-read.spss("emc_passage1.sav",to.data.frame = TRUE)
#Summary of variable
summary(emc)
#Create replacement variable
emc$replacement<-ifelse(emc$A5>12,1,0)
#Select variables of interest
emc<-emc[,c("A2","replacement","hhsize","language","missing","duration","interview")]
#Aggregate data at province level
prov<-aggregate(emc,by=list(emc$A2),FUN=mean, na.rm=TRUE)
#Drop province variable
row.names(prov)<-prov$Group.1
prov<-prov[,-2]
#Principle components analysis
#Loading package FactoMineR
library(FactoMineR)
#Implémentation de l'ACP
pca=PCA(prov[,-1])
summary(pca)
#Clusteringprovinces
clas=HCPC(pca,nb.clust=5)
clas$desc.var
#Clusters size
table(clas$data.clust$clust)
#Characteristics of clusters
aggregate(prov,by=list(clas$data.clust$clust),FUN=mean)
#Clusters characterization
clas$data.clust$charac=clas$data.clust$clust
levels(clas$data.clust$charac)=c("low-hh-size, low-local-lang","high-local-lang, low-missed","high-missed, high-missing-val","high-reluctance","high-hh-size, high-missing-val")
write.table(clas$data.clust,"clustering.csv")
```