AJAS / SPAS

ISSN 2316-0861

**African Journal of Applied Statistics**
Vol. 6 (1), 2019, pages 589 - 598.
DOI : http ://dx.doi.org/10.16929/ajas/2019.589.232

ISSN 2316-0861

# Developing an area-based sampling system in the urban district of Bobo Dioulasso, Burkina Faso

**Serge Manituo Aymar Somda [1,2,*], Do Edmond Sanou[2], Armel Soubeiga[2] and John Emmanuel Marie Sawadogo[1]**

[1] Centre MURAZ, 2054 Av. Mamadou Konaté, Bobo Dioulasso, Burkina Faso
[2]Université Nazi Boni, 01 BP 1091 Bobo-Dioulasso 01, Burkina Faso

Received on January 15, 2019. Accepted on April 04, 2019. Published online on May 2019



**Résumé Abstract.** We proposed a simple and comprehensive approach to build a probabilistic sampling system in low income settings. We conducted this study in Bobo-Dioulasso, Burkina Faso. Our method consists in first selecting commonly known spots in the study area. Key information is collected in these spots and the population density around the spots is estimated. We then predicted the population density in all the study area by kriging, controlling for the variance. We can then compose, for any selected point in the study area a virtual cluster with an estimated population density. We then described a two-stage sampling design with equal probability for each stage. We described the unbiased Horvitz-Thompson estimators which will be obtained.

**Key words:** sampling ; kriging ; low income countries ; Burkina Faso.
**AMS 2010 Mathematics Subject Classification Objects :** 62D05 ; 62H11 ; 62P99.

---

Presented by Dr. Olusegun Ewmooje Federal University of Technology, Akure, Nigeria
Correspondent Member of the Editorial Board.

---

*Corresponding author Serge M.A. Somda : manituo@gmail.com
Do Edmond Sanou : doedmondsanou@gmail.com
Armel Soubeiga : armelsoubeiga93@gmail.com
John E.M. Sawadogo : johnemmanuel@gmail.com

**Résumé** (French) Nous avons proposé une approche simple pour construire un système d'échantillonnage dans des situations à revenus limités. L'étude a été menée à Bobo-Dioulasso, Burkina Faso. La méthode a consisté à sélectionner des lieux-dits bien connus dans la zone d'étude. Des informations clés étaient collectées sur ces lieux et la densité de la population était estimée. Nous avons ensuite prédit la densité de la population sur tout le territoire par krigeage, en controlant la variance. Nous pouvons ensuite proposer, pour un point désigné dans la zone d'étude, un groupe virtuel avec une densité de population estimée. Nous avons ensuite décrit une méthode de sondage à deux degrés à probabilités égales et l'estimateur sans biais de Horvitz-Thompson associé.

## 1. Introduction

Households surveys are the principal source for national, regional and local statistics in low income countries. In fact, complete information sources are generally nonexistent or of poor quality. Household surveys need a comprehensive household list to prepare a probabilistic sampling design. A probabilistic representative sampling design is one where all the individuals of the population have a known, non-null, probability of being selected.

In developed countries, comprehensive lists for population surveys can be obtained through vital registry systems, addressing and mailing records, voting lists, etc. These registries are usually made complete and up-to-date. In developing countries, the authorities can barely manage the population dynamics. Births and deaths as well as internal and external migrations are not recorded in a systematic way. The streets and houses indications are often missing or unknown by the population. Finally, it is very hard to build a comprehensive list (Leete , 2001). Non-probabilistic approaches can be used to drive fast information with very poor quality.

The procedure used for population surveys includes several stages. The first stage consists in the selection of clusters. These clusters can be administrative areas such as villages or city sections. The clusters are selected using a first probabilistic approach. In the selected clusters, data collection agents are sent to record all the individuals, in order to prepare sampling lists. However, the administration areas are not suitable for sampling. They can be too big for instance. For large scale studies, the national statistics offices use census enumeration areas EA. EAs are areas where live around a thousand individuals but this varied where population density or environmental features required larger or smaller boundaries to facilitate enumeration (INSD , 2006). These areas are set up during census map production of the most recent national census and then used since the next one. Operations of updating of the census EA maps can also be done. All the research institutes and organizations do not have access to these census area lists. However, setting them up is very expensive and demands a lot of logistics.

The World Health Organization (WHO) has proposed an Expanded Programme on Immunization (EPI) cluster survey design (Henderson and Sundaresan , 1982; Bennet *et al.* , 1991). This method is widely used as the standard sampling method for health surveys in developing countries when EAs are not available. The standard method proposes to select 30 clusters in the study area and then to select 7 children in each cluster to observe.

Clusters may be villages or the like. This method has been discussed, tested and validated in the literature (Bennet *et al.* , 1994) and several variations have been proposed (Turner *et al.* , 1996). However, this method lies in pre-defined cluster. The cities and the villages are fast increasing in developing countries and it is now more and more difficult to consider an entire village as a cluster.

In many research areas such as epidemiology, ecology and sociology, geographical approaches are being more and more used (Keating *et al.* , 2003; Heeringa *et al.* , 2004). Considering the populations in terms of geographical location could simplify the sample selection (Troyo *et al.* , 2008; Vanden Eng *et al.* , 2007). A good knowledge of the distribution of the households and the people in a city can help to define a consistent sampling strategy for population surveys. Once the places people leave are known, one can go and select them. Spatial methods help to provide estimation on characteristics according to geographical distance with known points. Their use has been extended to several domains. These can contribute to reinforcing surveying systems. We propose an estimation of the density of the population in each geographical point as a database for preparing a sampling method for population surveys in an urban statement in Africa, the city of Bobo-Dioulasso.

## 2. Material and methods

### *2.1. Introducing the city of Bobo-Dioulasso*

Burkina Faso is a landlocked country in Western Africa. It's 2006 population was estimated at 14,017,262 (INSD , 2006). The country is listed among the low income countries. The Human Development Index was estimated 0.402 in 2015, ranking the country 185 on 188 (INSD , 2017). Bobo-Dioulasso is the second biggest city of the country. It is located at 365 kilometers from Ouagadougou, the capital city and covers an area of 136.78 square kilometers (figure 1). It's 2006 population was 554,042 residents living in 106,660 households (Zida-Bangré , 2009).

The city of Bobo-Dioulasso is organized in seven districts (arrondissements) containing thirty-three (33) neighborhood areas (sectors).

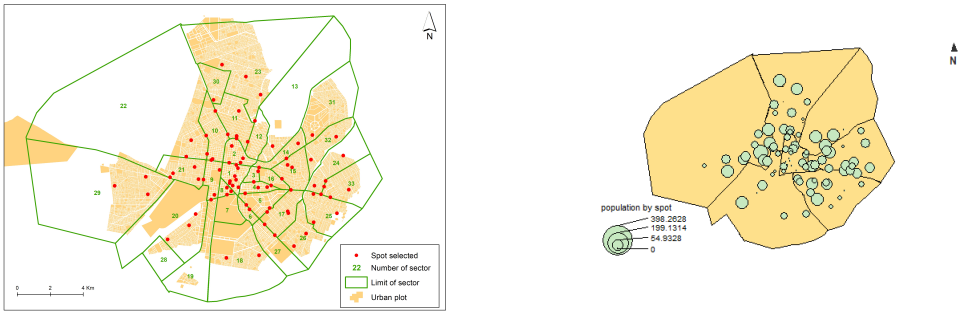
### *2.2. Data collection*

The first task consisted in the determination of significant spots in the urban area of Bobo-Dioulasso. The spots were defined as known and remarkable places, small extent, which were associated a proper name. These places may be markets, cemeteries, fountains, pubs, monuments, intersections (roundabouts, crossroads, ...), etc. We selected the spots according to their spatial spread. These spots were identified only in residential areas of the city.

A primary data collection and localization were performed in the eighty five (85) selected spots (figure 2a). The blank areas represent nonresident areas such as factory area or protected forest. The military zone was not included in the sample. Data were collected using CSPro version 6.3.2 application on Windows (US Census Bureau , 2015) and CSENTRY on


FIGURE 1: Bobo-Dioulasso in Burkina Faso

Android (US Census Bureau , 2015). Data were collected from 30 May 2016 to 1 June 2016. Analysis and estimations were performed using R software (R core team , 2015).




(a) (b)

FIGURE 2: Spatial distribution of the observed spots

For each spot, the geographical coordinates were collected, with each nature and place, including a picture. Another important was an estimation of the population density in terms of households around the spot. The density around the spot was estimated using web

utilities, including *Openstreetmap* (Openstreetmap Contributors, 2016) and *Google Earth* (Google Inc. , 2016). The number of habitable plots were counted in an area of a hundred (100) meters radius around the identified spot. We were supported in this task by the specialists of the ministry of housing and town planning in Bobo-Dioulasso. The images were considered for the 06th of June 2016.

The number of households living in the plots was estimated using the city statistics (Zida-Bangré , 2009). We mapped the spots according the estimated densities (figure 2b).

## 3. Spatial analysis

### *3.1. Definitions and notations*

The variogram is a function playing a key role in geostatistics and spatial analysis. It quantifies the interdependence of sampling locations ; i.e. two samples from nearby locations tend to be more alike than two taken from widely separate locations (Warrick and Myers , 1987).

Let's use the following notations :

- $Z(s)$ : the estimated density at a given spot $s$,
- $h$ : the distance in meters between two spots,
- $\gamma(h)$ : the variogram function
- $C(h)$ : the covariance function

and assume that the density function $Z(s)$ is defined by $\{Z(s), \quad s \in D\}$ where $D$ represents the entire study area. We then have the following intrinsic hypotheses (Cressie , 1985) :

$$\forall s \in D \quad E[Z(s) - Z(s+h)] = 0 \tag{1}$$

$$\forall s \in D \quad E[Z(s) - Z(s+h)]^2 = 2\gamma(h) \tag{2}$$

$$\forall s \in D \quad Cov[Z(s), Z(s+h)] = C(h) \tag{3}$$

### *3.2. Spatial estimation*

The purpose is to estimate the density in a given point $Z^*$ in the study area. This estimate is obtained using an ordinary kriging system described by Olivier and Webster (2014) :

$$Z^* = \sum_{s \in D} \lambda_s Z(s) \tag{4}$$

where the $\lambda_s$ are the parameters to be estimated. The estimation error will be defined as

$$e = (Z - Z^*) \quad \text{with} \quad V(e) = E[Z - Z^*]^2$$

This variance $\sigma_e^2$ can be given an equivalent formula :

$$\sigma_e^2 = V(Z) + V(Z^*) - 2COV(Z, Z^*) \tag{5}$$

$$\sigma_e^2 = V(Z) + \sum_{s \in D} \sum_{t \in D} \lambda_s \lambda_t Cov(Z(s), Z(t)) - 2 \sum_{s \in D} Cov(Z(s), Z) \tag{6}$$

Then the $\lambda_s$ are the solutions of the following equation :

$$\sum_{t \in D} \lambda_t Cov(Z(s), Z(t)) + \mu = Cov(Z, Z(s)) \ \forall s \in D \tag{7}$$

with $\sum_{s\in D} \lambda_s = 1$, where $\mu$ is a Lagrange multiplier introduced for the minimization of the error variance.

## 4. The sampling system

The exercise described above permitted to get a new sampling basis. In fact we could define a regular grid, covering the urban surface. The density around any given point of the area can be predicted using this model, as shown in figure 3.

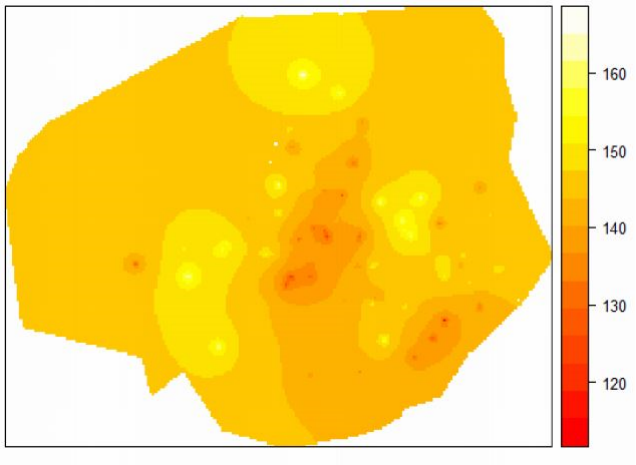



FIGURE 3: Density predictions in the urban area of Bobo-Dioulasso

We now proposed a probabilistic sampling method for a given survey. This is a two-stage sampling method. The primary units are circular areas with 100-meters radius in which the number of households is identified. The secondary units are these households. We propose the inclusion probabilities for a random sampling with equal probabilities. The parameters estimations are then described according to the Horvitz and Thompson (1952) approach.

### *4.1. Description*

For a two-stage sampling, the following hypotheses are observed : :

*invariance :* the mode of selection of secondary units is the same for each selected primary unit ;

*independence :* the mode of selection of secondary units is independent from the mode of selection of primary units.

In the selected area ($U_C$) we can consider $M$ primary units (clusters). Let's consider for the cluster $h; h = 1, \ldots, M$, named $u_h$, $N_h$ households. We aim to select $m$ clusters in $S_m$ and in any selected cluster $u_h$, we will select $s_h$ households.

### *4.2. Inclusion probabilities*

Let us define :

$\pi_h$ the probability of selecting the cluster $u_h$ ,

$\pi_{hk}$ the probability of selecting both clusters $u_h$ and $u_g$,

$\pi_{i|h}$ the probability of selecting the household $i$ in the selected cluster $u_h$,

$\pi_{ij|h}$ the probability of selecting both the households $i$ and $j$ in the selected cluster $u_h$,

$\Delta_{ij|h} = \pi_{ij|h} - \pi_{i|h}\pi_{j|h}$

For a sampling with equal probabilities for the clusters, we have :

$$\pi_h = \frac{m}{M} \tag{8}$$

$$\pi_{hk} = \frac{m}{M} * \frac{m-1}{M-1} \tag{9}$$

For a sampling with equal probabilities for the households into the clusters, we have :

$$\pi_{i|h} = \frac{n_h}{z_h} \tag{10}$$

$$\pi_{ij|h} = \frac{n_h}{z_h} * \frac{n_h-1}{z_h-1} \tag{11}$$

Where $z_h$ results to the kriging interpolation described above.

### *4.3. Horvitz-Thompson estimator*

The objective of sampling is to provide the most accurate estimation of a function of the key variable $\theta = \theta(Y_i), i \in U$ and its variance. Here we propose the estimation of the total (sum). The other characteristics such as the mean can be easily obtained using the method. Considering $Y_{i|h}$ the value of the character in the household $i$ located in the cluster $h$, we have :

$$\theta(Y_i) = T(Y_i) = \sum_{h=1}^{M}\sum_{i=1}^{N_h} Y_{i|h} \tag{12}$$

As all our inclusion probabilities are more than zero, the Horvitz and Thompson (1952) estimator is given by :

$$\hat{\theta} = \sum_{h\in S_m}\sum_{i\in s_h} \frac{y_{i|h}}{\pi_h \pi_{i|h}} \tag{13}$$

where $y_{i|h}$ is the measured value of the parameter in the sampled household $i$ in the cluster $h$. This estimator is unbiased.

In fact we can consider estimators of all clusters with their corresponding variance and even the estimators of the variance (Thompson , 1991) (table 1).

TABLE 1: Estimation formulas for a cluster $h$

| Value | Notation | Formula |
|---|---|---|
| Total | $\theta_h$ | $\theta_h = \sum_{i=1}^{N_h} Y_{i\|h}$ |
| Estimator of the total | $\hat{\theta}_h$ | $\hat{\theta}_h = \sum_{i=1}^{n_h} y_{i\|h}$ |
| Variance of the estimator | $V(\hat{\theta}_h)$ | $V(\hat{\theta}_h) = \frac{N_h(N_h - n_h)}{n_h(N_h - 1)} \sum_{i\in 1}^{N_h} (Y_{i\|h} - \frac{\theta_h}{N_h})^2$ |
| Estimator of the Variance | $\hat{V}(\hat{\theta}_h)$ | $\hat{V}(\hat{\theta}_h) = \frac{N_h(N_h - n_h)}{n_h(n_h - 1)} \sum_{i\in 1}^{n_h} (y_{i\|h} - \frac{\hat{\theta}_h}{n_h})^2$ |

The variance of $\theta$ and its estimator are then calculated using the formula of the analysis of variance :

$$V(\hat{\theta}) = \frac{M(M-m)}{m(M-1)} \sum_{h=1}^{M} (\theta_h - \frac{\theta}{M})^2 + \frac{M}{m} \sum_{h=1}^{M} \frac{N_h(N_h - n_h)}{n_h} V(\hat{\theta}_h) \quad (14)$$

$$\hat{V}(\hat{\theta}) = \frac{M(M-m)}{m(M-1)} \sum_{h=1}^{m} (\hat{\theta}_h - \frac{\hat{\theta}}{M})^2 + \frac{M}{m} \sum_{h=1}^{m} \frac{N_h(N_h - n_h)}{n_h} \hat{V}(\hat{\theta}_h) \quad (15)$$

## 5. Discussion and conclusion

We proposed a comprehensive method to conduct a random sample with estimated inclusion probabilities in case of population based surveys where no list is available. This is always a challenge when conducting surveys in low income countries where vital registries, voting lists and addressing is very poorly organized. This approach will permit robust and accurate estimations of different information one could be trying to measure in households.

We used an analytic approach for the variance estimation. Usually no sample correction is performed to estimate variances in the most of the small surveys and the final results could be wrong. Other methods can be used to compute that variance, including jackknife and bootstrap estimation proposed by Wu (1986). Similar approaches have been described in literature like Siri *et al.* (2008) but these are applied to specific fields. We are somehow proposing a generalization of these approaches.

One limit of this approach is that the estimations will need updates. The cities like Bobo-Dioulasso experience constant changes and new high density areas are created. So it will be important to conduct these surveys like every five years to update the estimates. This experience can be reproduced in any situation, including urban and rural areas.

**Acknowledgement**

The study was conducted during the internship of DES and AS in Centre MURAZ and we will like to acknowledge both Centre MURAZ and the Université Nazi Boni for making this possible. We will also like to acknowledge the staff from the ministry of urban affairs, the city office and all the kind persons who helped us on the field.

We acknowledge the reviewers for their valuable contributions to our work.

**Références**

Bennett, S., Radalowicz, A., Vella, V. and Tomkins, A. (1994). A Computer Simulation of Household Sampling Schemes for Health Surveys in Developing Countries *International Journal of Epidemiology*, doi :10.1093/ije/23.6.1282, ISSN :0300-5771, 23(6), 1282-1291.

Bennett, S., Woods, T., Liyanage, W.M. and Smith, D.L. (1991). A simplified general method for cluster-sample surveys of health in developing countries. *World Health Statistics Quaterly, WHO*, 44, 98-105, url : *http ://apps.who.int/iris/bitstream/handle/10665/47585/WHSQ_1991_44%283%29_98-106_eng.pdf ?sequence=1&isAllowed=y.*

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17(5), 563-586.

Google Inc. (2016). Google Maps / Google Earth. url : *http ://www.earth.google.com*

Heeringa, S.G., Wagner, J., Torres, M., Duan, N., Adams, T. and Berglund, P. (2004). Sample designs and sampling methods for the Collaborative Psychiatric Epidemiology Studies (CPES) *International journal of methods in psychiatric research*, 13(4), 221-240.

Henderson, R. H. and Sundaresan, T. (1982). Cluster sampling to assess immunization coverage : a review of experience with a simplified sampling method. *Bullettin of the World Health Organization*, ISSN :0042-9686, 60 (2), 253-260.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260), 663-685.

Institut National de la Statistique et de la Démographie (Burkina Faso) (2007). Recencement Général de la Population et de l'Habitation 2006. url : *http ://www.insd.bf/n/index.php/publications ?id=141.*

Institut National de la Statistique et de la Démographie (Burkina Faso) (2017). Annuaire Statistique 2016. url : *http ://www.insd.bf/n/contenu/pub_periodiques/annuaires_stat/Annuaires_stat_nationaux_BF/Annuaire_stat_2016.pdf.*

Keating, J., Macintyre, K., Mbogo, C., Githeko, A., Regens, J.L. Swalm, C., *et al.* (2003). A geographic sampling strategy for studying relationships between human activity and malaria vectors in urban Africa *The American Journal of Tropical Medicine and Hygiene*, doi :10.4269/ajtmh.2003.68.357, ISSN :0002-9637, 1476-1645, 68(3), 357-365.

Leete,R (2001). Population and housing censuses : A funding crisis. *Population and housing censuses : A funding crisis* url : *http ://www.ldsa.gr/european-census/Files/general-data/Symposium/crisis.pdf.*

Openstreetmap Contributors (2016). Openstreetmap. url : *https ://www.openstreetmap.org.*

Oliver, M. A. and Webster, R. (2014). A tutorial guide to geostatistics. *Catena*, 113, 56-69.

R Core Team (2015). R : A Language and Environment for Statistical Computing. url : *http ://www.R-project.org/.*

Siri, J. G., Lindblade, K. A., Rosen, D. H., Onyango, B., Vulule, J. M., Slutsker, L. and Wilson, M. L. (2008). A census-weighted, spatially-stratified household sampling strategy for urban malaria epidemiology. *Malaria Journal*, doi :10.1186/1475-2875-7-39, ISSN :1475-2875 7(1), 1475-2875.

Thompson, S. K. (1991). Adaptive cluster sampling : designs with primary and secondary units. *Biometrics*, 47(3), 1103-1115.

Troyo, A., Fuller, D.O., Calderón-Arguedas, O. and Beier, J.C. (2008). A geographical sampling method for surveys of mosquito larvae in an urban area using high-resolution satellite imagery *Journal of vector ecology : journal of the Society for Vector Ecology*, ISSN :1081-1710, 33(1), 1-7.

Turner, A.G., Magnani, R.J. and Shuaib, M. (1996). A Not Quite as Quick but Much Cleaner Alternative to the Expanded Programme on Immunization (EPI) Cluster Survey Design *International Journal of Epidemiology*, doi :10.1093/ije/25.1.198, ISSN :0300-5771, 25(1), 198-203.

US Census Bureau and Macro International and Serpro S.A. (2015). Census and Survey Processing system (CSPro). url : *http ://www.census.gov/population/international/software/cspro/*.

US Census Bureau and Macro International and Serpro S.A. (2015). CS Entry. url : *https ://play.google.com/store/apps/details ?id=gov.census.cspro.csentry*.

Vanden Eng, J.L., Wolkon, A., Frolov, A.S., Terlouw, D.J., Eliades, M.J, Morgah, K. *et al.* (2007). Use of Handheld Computers with Global Positioning Systems for Probability Sampling and Data Entry in Household Surveys *The American Journal of Tropical Medicine and Hygiene*, doi :10.4269/ajtmh.2007.77.393, ISSN :0002-9637, 1476-1645, 77(2), 393-399.

Warrick, A. W. and Myers, D. E. (1987). Optimization of sampling locations for variogram calculations. *Water Ressources Research*, 23(3), 496-500.

Wu, C. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4), 1261-1295.

Zida-Bangré, H. (2009). Monographie de la commune urbaine de Bobo-Dioulasso. Institut National de la Statistique et de la Démographie (Burkina Faso), url : *http ://www.issp.bf/index.php/fr/droits-des-enfants-au-bf/2-documents-recenses/5-autres/ouvrages/389-monographie-bobo/file*.