COLLEGE OF SCIENCE AND TECHNOLOGY
SCHOOL OF SCIENCES
DEPARTMENT OF MATHEMATICS

SMALL AREA ESTIMATION OF POVERTY INDICATORS IN RWANDA SECTORS: A case study in KARONGI district.

MUKANDAYISENGA Beatrice

Master of Science

in

Applied Mathematics

Kigali, May 12, 2019

COLLEGE OF SCIENCE AND TECHNOLOGY
SCHOOL OF SCIENCES
DEPARTMENT OF MATHEMATICS

SMALL AREA ESTIMATION OF POVERTY INDICATORS IN RWANDA SECTORS: A case study in KARONGI district.

By
MUKANDAYISENGA Beatrice
Student number: 217130704

Thesis Submitted in Partial Fulfillment of the Academic Degree of
Master of Science
in
Applied Mathematics
Option of Mathematical Modeling and Scientific Computing

Supervisor: Dr. Innocent NGARUYE

Kigali-Rwanda

May 12, 2019

Declaration

I, MUKANDAYISENGA Beatrice, hereby declare that this masters thesis titled "*Small Area Estimation of Poverty Indicators in Rwanda Sectors: A case study in Karongi district.*"is my original work and has never been submitted or presented in any University or Institution of higher learning for academic purposes or otherwise. This was done under supervision of Dr. Innocent NGARUYE, University of Rwanda-College of Science and Technology, Rwanda.

May 12, 2019

MUKANDAYISENGA Beatrice

Dedication

This thesis is dedicated to my lovely family who were always there for me; my beloved Husband Dr. HATUNGIMANA Fabien, my daughter ICYEZA Lina Bernice, my parents BATSEMBISANO Evariste and MUKANGANGO Victoire, brothers NDAYISHIMIYE Deogratias and TUYAMBAZE Felix, sisters NIRERE Beata, TUYISABE Ancille and NIYONSABA Marie Chantal, for their invaluable moral and financial support. You deserve my deepest gratitude for the unfailing patience and sacrifice for my education.

I am also grateful to my generous relatives, colleagues, friends, classmates and well wishers who shared and supported me throughout my longstanding journey .

You taught me to be resilient.

Abbreviation

BLUP: Best Linear Unbiased Prediction

EBLUP: Empirical Best Linear Unbiased Predictor

EDPRS: Economic Development and Poverty Reduction Strategy

EICV: Integrated Household Living Conditions Survey(Enquete Integrale sur les Condition de Vie des menages)

FGT: Foster-Greer-Thorbecke

GIS: Geographic Information System

GLS: Generalized Least Squares

HCR: Head Count Ratio

MANOVA: Multivariate Analysis of Variance

NCD: Non Communicable Diseases

NISR: National Institute of Statistics of Rwanda

ODI: Oversees Development Institute

PG: Poverty Gap

PovMap: Poverty Mapping

Rwf: Rwandan Franc

SAE: Small Area Estimation

UR: University of Rwanda

VUP: Vision Umurenge Program

Abstract

This research reflects on the small area estimation (SAE) with the principal objective of presenting the status of the poverty and extreme poverty at sector level. To accomplish this objective, we first present the theory of the small area estimation (SAE) technique. The SAE is concerned on the generation of believable estimations of characteristic of interest for small demesnes, starting on small or non samples coming from these demesnes; and the assessment of the estimate or error prediction. To improve direct estimationsfor a small demesne, SAE technique tries to "borrow strength" (covariates) from other related data sets, either from similar areas, or relevant/auxiliary information obtained from a recent census or some other administrative records. The covariates used that are related to poverty indicators were the rates of household with no electricity, production, roof with local tiles, roof with other materials, and unimproved sanitation facilities. These five covariates were chosen as they were found to be of great impact on poverty status as published in the fourth integrated household living conditions (EICV4) report 2013-14 and in Poverty mapping report 2013-14.

This study showed that, there is inequality in sectors of Karongi District. Bwishyura and Rubengera sectors were less suffering from poverty, with poverty incidence of 43.047% and 49.9902% respectively and with the extreme poverty incidence 27.7159% and 32.3044% respectively. Mutuntu sector was the most suffering from the poverty in Karongi district with the poverty incidence of 73.2498% and the extreme poverty of 49.2456% followed by Rwankuba sector with the poverty incidence of 70.2774% and the extreme poverty incidence of 46.6706%.

Acknowledgment

My inner gratitude is for the government of Rwanda through the Ministry of Education for sponsoring my studies at the University of Rwanda.

My special thanks go to Dr NGARUYE Innocent, who despite many engagements, withtook the supervision of this research. His simpleness, comments and specially his scientific rigour have been of great significance on the realization of this research.

I also express my acknowledgement toward KARONGI District for permitting me to carry this work .

My thanks are also highly addressed to my family and my classmates for their support and motivation in my study. Im very thankful to the University of Rwanda, especially NYARUGENGE Campus and its devoted staff, which have made our training possible.

Last but not the least; my special acknowledgement goes to all my lecturers at University of Rwanda for their commitment to teaching, mentoring and encouragement. The fruit of your work is this step reached so far.

May God bless you all!

CONTENTS

1. INTRODUCTION

1.1 General introduction

1.1.1 Background

Following Hidiroglou, 2007, the domain or area is said to be small when the statistic of interest of its subpopulation cannot be estimated accurately because of same constraints of existing data. Example of small domain: geographic region such as country, province, district, and so on. Due to the useful of such data in regional planning, distribution of different funds, government policy and the development programs; their neediness has very gone up in the past period.

The programs which had the goal of producing estimates for the small areas so that the new demand may be satisfied, have been presented by many statisticians. However existing data to present these kind of estimates are founded on surveys which are not indicated for theses small areas. But many method are usable to produce different parameters statistic of interest for the small area levels, when the data at the higher level are available, and they have a great correlation with the variables of interest at the corresponding level. The main purpose of estimate the variables parameters at the small area level is to plan and to allocate the different founds. So to produce these parameters is very important for the government. We can say same small area estimates such as poverty indicators, stunting indicators, education indicators and development indicators. The number of used factors influence the generation of the small area estimates. A number of questions are raised up to unhance this mater. What is the requirement of these statistics? Which is the engagement and desire the management support methodology, arrangement, and subject matter staff? What is a number of methodology and subject matter ability which are in the management? What is the intensive of the relationship between the variables of interest and the available auxiliary information? There is the sufficient sample size of survey so that the estimates have the accuracy by utilizing the data of survey and available auxiliary information? How much bias are the management and customers wishing to allow the estimates? What is the implication of making the wrong decision? The size of the units of

the small area a significant consideration. It is possible that the small area estimates differ absolutely from the statistics based on local knowledge.

The Government of Rwanda wishes that all population live in the enjoyable life. The poverty is defined to be the state or personal or family income lower than the one which is supposed to be standard.
In Rwanda, a household is considered to be poor if it is not able to fulfill its basic needs such as food, clothing, children's school fees, medical and insurance fees, etc. To determine the poverty level, one use the poverty line. By definition, the poverty line is the lowest income level utilized as an official standard in order to determine the number of people which are in poverty.

1.1.2 Definition of some key concepts

- Poverty is defined as the condition in which there is lack of basic needs to live like food, house, etc.
- Extreme poverty is part of people who have the food and non food consumption in term of money is less than which is assumed to be food poverty line (Rwf 105,064 in June 2014 prices).
- Poverty line is the lowest income level which is used to be standard in order to determine the share of people which are in poverty (for example in Rwanda, poverty line is Rwf 159,375).
- Headcount index is defined as the proportion of people who have all consumption (food and non food) is less than the total poverty line (Rwf 159,375 in January 2014 prices). Headcount index is also called incidence of poverty.
- Total poverty line is determined by considering both food and non food items cost.
- Food basket is defined as a mixture of basic products which make up the usual diet of population in sufficient amounts to cover adequately the energy requirements of each member of family (NISR, 2015).

1.2 Problem statement

Despite a considerable progress for sustainable development and poverty reduction that the Government of Rwanda has made for the last decades, there is still a long way to go about fighting against poverty.

In order to put up some programs and planning for poverty reduction, the first step is to know the current status across all the parts of the country. The Integrated Household Living Conditions survey (EICV) conducted in Rwanda every 5 years tries to show the poverty mapping and poverty indicators in Rwanda up to district level. However, the statistics about poverty indicators are also needed at sector level by stakeholders in order to know where to put more emphasis since the unit of planning in Rwanda is a sector.

1.3 Objectives

The goal of this study is to estimate the status of poverty at the sector level based on existing micro data and other relevant auxiliary data; so that policy-makers may not be misled mainly when allocating funds. Based on the status of the poverty indicators in a district, we want to present the status of the poverty and extreme poverty at sector level to identify which factor contributes to poverty the most in a given sector. We want to find the poverty and extreme poverty status by sector based on the key poverty indicators as identified by EICV 2014. Even though, Rwanda counts 416 sectors, this work was only limited on sectors of Karongi district, because of the time and the budget constraints. This study uses Small Area Estimation approach by borrowing strength from other related data sets, either from similar areas, or relevant/auxiliary information obtained from a recent census or some other administrative records.

2. SMALL AREA ESTIMATION METHODOLOGY

2.1 Small Area Estimation

Small area estimation is referred to several statistical and mathematical methods aiming on the estimation of quantities such as means, total, proportions... at sub-population level, utilized when the sub-population of interest is included in a large survey and the direct estimates can not estimated accurately because of the small sample size corrected to the area. The accuracy is attained by borrowing strength for the estimation for a spacial small area by making use of facts from areas to which it is similar. There are small area estimation methods which join the data coming from many sources. By using an example, the census and the new survey information may be joined to modernize estimations from the initial census.

Historically, small area estimation took place in eleventh century England and seventeenth century in Canada (Brackstone, 1987; Marchall, 1991). Extremely long time, demographers have been applying a variation of indirect methods for small area estimation of population and other characteristics of interest. Recently, the request for definitive small area estimates has remarkably expanded far-reaching; because of their many application in formulating strategies and programs, the allocation of government investments and in regional planning. Application of SAE has also expanded in the private sector due to the business decisions, particularly the small businesses, depend deeply on the regional conditions.

Many authors have investigated some new developments and directions in small area estimation. Cite some authors SAE has been used in many applications such as poverty mapping, unemployment rates, demography, disease mapping, etc. The model-based method to small area estimation provides different benefits, and the most important of them are: increased accuracy, the production of optimal estimations and related measurements of variableness for a supposed model, and the validation of models from the sample data.

"Small area" or "small domain" is directed to a subpopulation where the sample of specific domain is not sufficient to generate direct estimations accurately. Some examples of this

subpopulation are geographical area (province, district, sector etc.), a group of people who are in the same region (for instance the group formulated by referring to age and sex, etc.) or any group of people (Pfeffermann, 2002; Rao, 2003; Rao and Isabel, 2015).

Small area estimation is a technique which uses data collected from one or more data sources, to generate estimations whose the accuracy, for the small area level, instead of using only the data from each small area level. In order to achieve that accuracy, it is necessary to borrow strength from the related areas. A few small area estimation methods Some merge the data coming from different sources. For instance, survey and census can be merged in order to upgrade estimations obtained in the previous census. For poverty estimations, the suitable model of statistic for the data of survey is gathered at the same time like the census, and this model is utilized for forecasting a variable not gathered in the census, based on the variables which are gathered in survey and census. Generally, small area estimation mentions the gathering of statistical methods planned in order to upgrade the sample survey estimations by using the auxiliary information. By starting with the targeted variable, say Y, of which we need estimations over a range of small subpopulations, sometimes is similar to small geographical areas. For every subpopulation, the direct estimates of Y are accessible due to the data of the sample survey, where Y is immediately determined by measurement on the sample units. Due to the smallness of the sample size in subpopulation, these direct estimates will not be reliable because it will have the large standard errors. In fact, certain subpopulations can not be in the sample in the least he survey. Let X be the auxiliary information which can be utilized for some situation to upgrade estimates, offering the lowest standard errors, X stands for the auxiliary variables that have been determined by measurement of the hole population.
The simple linking model relating X and Y is the following:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{2.1}$$

This relationship may be estimated utilizing information from the survey, where the accessibility of target variable and the auxiliary variables is possible. $\boldsymbol{\beta}$ stands the estimated regression coefficients providing the effect of the X variables on Y, and e represents a random error showing that portion of Y may not be described utilizing only the auxiliary information. By supposing that the above relationship still the same in entire population, it can be used to forecast Y for those units where X have been measured but not Y. Small-area estimates by considering these forecasted Y values will usually possess the tolerated standard errors comparing with the standard errors of the direct values, because of that these values increase the sample size. So it is necessary to borrow the strength from the other related area in order to obtain the large samples.

2.2 Estimation of poverty indicators.

In the study of poverty, the majority common variable forecast is income or cost by using a model that contains level of education, the age of the members who are in the family, Household representative and size of the family etc... This method is called "poverty mapping" as on it the poverty estimates are frequently mapped in detail. In this context, the maps can make interpretation as simple as possible, unless the central point is not the map in itself, nevertheless this poverty may be evaluated at a great thinner level and cheaper than by augmenting the data adequately or repeating the census. It is understandable that there is a cost for the statistical modeling, even though this may not be expensive and possess the best information at small area(Stephen, 2008).

The World Bank recorded, reinforced and advocated the technique for small area estimation of expenditure or income poverty which is progressed by Elbers et al (2001); Elbers et al, (2003); Leite et al, (2001). This is accessible as unlimited software (PovMap) from the website of the World Bank (Zhao, 2006). The data were sampled in order to fit a statistical model for income (or expenditure) on the scale of logarithm and apply the model to the census data to make an estimate of income (or expenditure) of all households, and put together the transformations of the predictions, small area estimates utilizing the World Bank approach for headcount index, intensity and gap may be mapped, and deduced at small area level. The variations of Elbers, Lanjouw and Lanjouw technique have been utilized for the World Bank in other countries such as Thailand (Healy et al, 2003), the Philippines (Stephen and Geoffery, 2005), South Africa (Alderman et al, 2002), Brazil (Elbers et al 2001; Leite et al, 2001) and for the World Food Programme in Bangladesh (Stephen and Geoffery, 2003), etc.

2.2.1 Definition of the principal poverty indicators

By the aim of monitoring the process of social inclusion, an enumeration of 18 indicators which monitor the poverty and social exclusion was forthput in 2001 (Atkinson et al. 2002). the enumeration is permanently modified and completed. Among these indicators, there are some which are based on the incomes of the household (monetary indicators and others are based on the symptoms which are not related to the money (non monetary indicators).

By taking the reference on the monetary poverty indicators and beginning from the distribution of the income, the average mean of the equalized income, the Head Count Ratio (HCR) and the Poverty Gap are the most frequently utilized indicators. The

Headcount index is measured by the HCR and this HCR shows the percentage of each household under the poverty line, which may be determined at national or regional level. The PG index shows the severity of poverty, whose the deepness of poverty by reflecting on farness, on average, the poor are determined by that poverty line. Officially, the headcount index or the PG may be produced by the generalized measures presented by 1984. Let t represents the poverty line, the Foster-Greer-Thorbecke (FGT) poverty measures is given by:

$$F(\alpha, t) = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{t - y_j}{t} \right)^{\alpha} I(y_j \leq t). \tag{2.2}$$

In this equation, y_j represent the amount of income for individual (household) j, N stands for the number of individuals (household) and α represents a sensitivity parameter. If $\alpha = 0$, we have the HCR, $F(0, t)$, but, when $\alpha = 1$, we have the PG, $F(1, t)$. The HCR indicator is a greatly utilized measure of poverty because of its ability of interpretation and construction, although that it has some constraints. By assumptions, all poor households live in the same condition, the fastest way of decreasing its valence is to put in practice the actions which favor the people who area under the poverty line so that some of these people may be motivated toward the line easily. Consequently, the policies could not be totally effective because of that they based on the poverty incidence but not based on the exam of the entire income distribution. That is why the estimates of the PG indicator have the significant. The PG may be considered as the mean shortfall of the people who are poor. It demonstrates the intensity of transferring the whole poor to raise their disbursement until to the poverty line. Collectively with these indicators, the mean valence of the income distribution of the household has also the significant. This is specially correct when the income level possesses a high tail. In this situation the value of median on which the poverty line is accomplished is hoped to be small and the HCR is likely small as possible. Moreover the relevance of PG may be lost, providing a false indication of the deprivation of the population on which the research is conducted. In a lot of situations, these policies are taken as a beginning point in the seriously research of poverty and living conditions. As a matter of fact, analysis are performed utilizing too indicators which are not related to the money so that the entire feature of poverty and the deprivation can be achieved (Cheli and Lemmi, 1995). Also, the fixed indicators are commonly expanded by the indicators which are in the powerless groups, in which it may be probably to transfer toward the poverty status

(Pratesi, 2016). The unusual distribution of these indicators of poverty is the structure of benefit which is on the high level. It can be provided by the construction of the maps showing the poverty status. To construct the maps of poverty, one may usa the different

data such as the data from the survey, census and the data from the administrative. In this case, the poverty mapping is taken as a reference to make visible the unusual distribution of indicators of the poverty. This one is very important to track the areas which are the most powerless and poverty localization (Pratesi, 2016).
The following are Poverty indicators:

1. Natural resource indicators: measure how the natural resources is very important on the people who are poor by showing the dependance of these people on them and the impact of their reduction on these people. According to ODI (Oversee Development Institute), natural resource indicators is one "which changes when better management of natural resources leads to decline in poverty" (Twesigye and Ntabana, 2007). There is the relationship between these indicators and the food security.

2. Environmental health indicators: show how the people who are poor are exposed to the diseases because of that they use the air and the water which are not proper.

3. poverty vulnerability to natural disasters: show how the disasters affect the people who are poor because of that the most of them are unable to handle the problem caused by these disaster and they have not the material which facilitate them to avoid these disasters (for example they have not the houses which are able to resist on the rain). Some disasters are natural and others are man made disasters. The droughts, landslides, floods and volcanic eruptions are examples of disasters.

4. Housing indicators: these indicators show the physical situation of habitations, kind of habitations and the area where the habitations are localized (Twesigye and Ntabana, 2007). The appearance of the house of any family can gives some information about poverty status. Thus these indicators are very important poverty indicators.

2.2.2 Indirect and direct estimate of poverty indicators at small area level

To estimate the some poverty indicators at small area level may be performed with the design based (Hansen et al. 1953; Kish, 1965), model assisted (Sarndal et al. 1992) and model based method (Valliant et al.2000; Rao , 2003), such as indirect or direct small area estimates. The disign based method influences the production of the direct estimates by utilizing the data of one survey only. For the indirect estimates, the auxiliary variables are used in order to ameliorate the correctness of estimates produced by the survey. Let U be a population whose N size in which there are D non intersecting subsets U_d whose size $N_d, d = 1, ..., D$. Let j and d be the units of population and small areas respectively, the

interest variable is y_{jd}, x_{jd} is a vector composed by p auxiliary variables. It is assumed that the first component of x_{ij} is 1.

Assume that s is a sample which is taken by considering some, perhaps not simple, designed sample so that the unit j of area d has the probability π_{jd} of being included, samples of specific area $s_d \subset U_d$ having the size $n_d \geq 0$ are profitable for every area. For non-sample areas $n_d = 0$, means that s_d is the empty set. Let $r_d \subset U_d$ be the set of the $N_d - n_d$ non-sampled units of small area$d$. The valences of y_{jd} are recognized but for the vector of auxiliary variable (vector p), we suppose that the totals area levels X_d or means area levels$\bar{X}_d$ or individual values x_{jd} are exactly recognized from exterior resources. The simple method of calculating FGT indicators of poverty by considering the interested areas, is to calculate the direct estimates. The direct estimators of each areas, utilize only the data relating to the households which are sampled because of that the information about income of a household for these households is profitable.

The form of the direct estimators for the FGT indicators of poverty are given by:

$$F_d^{dir}(\alpha, t) = \frac{1}{\sum_{i \in s_d} w_{id}} \sum_{j \in s_d} w_{jd} \Big(\frac{t - y_{jd}}{t} \Big)^{\alpha} I(y_{jd} \leq t), \quad d = 1, ..., D, \tag{2.3}$$

where w_{jd} represents the sampling weight (with $w_{jd} = \frac{1}{\pi_{jd}}$)
of household j belonging to area d. In the similar case, the average of the household equivalized income for every small area may be calculated as:

$$\mu_i^{dir} = \frac{1}{\sum_{i \in s_d} w_{id}} \sum_{j \in s_d} w_{ij} y_{ij}, \quad i = 1, ..., m \tag{2.4}$$

If we have the limited sample size of the interested ares, estimators like 2.2 and 2.4 may not be utilized. As a matter of fact, the size is very small such that it cannot produce the important statistical direct estimates got from the designed sample. So really designed based response and the using of the direct estimates frequently means the augmentation of the sample size, using additional sample coming from the different domains which are studied before. If the additional sample is used, acceptable estimates may be got with suitable direct estimators and the difficulty concern to the Small Area Estimation is resolved. However in different practical conditions, to use the additional sample is not very easy because of its costbenefit, time consuming and non affordable response. Assisted and based model SAE methods are required to be used for these conditions. Thus, to compute the targeted parameters at non global level by using indirect approaches, the auxiliary information are used. These information are often obtained in the data of administrative which are too profitable at non global area level. An appropriate model describes the connection between the targeted parameters and the auxiliary variables. Referring to

Sarndal et al. (1992), in this case we identify a model which having few assumptions of connection, non verifiable but not completely away of situation, to keep sources of survey or to avoid other specific problems. With these methods, it is very important that the average and the FGT indicators for the small area d can be expressed.
For the small area, the population average mean may be expressed as follow:

$$m_d = N_d^{-1}\Big(\sum_{j\in s_d} y_{jd} + \sum_{j\in r_d} y_{jd}\Big). \tag{2.5}$$

Due to that y valences of r_d non-sampled units are not known, we use their prediction. For small area d, the FGT poverty inicators may be expressed as follow:

$$F_d(\alpha, t) = N_d^{-1}\Big(\sum_{j\in s_d} z_{jd}(\alpha) + \sum_{j\in r_d} z_{jd}(\alpha)\Big), \tag{2.6}$$

where

$$z_{jd}(\alpha, t) = \Big(\frac{t - y_{jd}}{t}\Big)^{\alpha} I(y_{jd} \leq t). \tag{2.7}$$

In addition, the z valences of r_d unsampled units are not known, and we use their prediction for the basis of prediction of y valences. In general, the y prediction is founded on the set of auxiliary variables by considering the regression model. In this case, the based model methods permit the achieving of an appropriate estimators because of borrowing due to the use of a fit model. The procedure of predicting can meet insufficiency, difficulties and the issues because of the data which are available and the fitness of the model. The extent and the amount of the information about the study and the auxiliary variables are the sources of these issues.

2.3 Clustering

The measurements are made for the units which are often dependent, although are clustered normally in groups of the same unity. Households which are similar, tend to be grouped collectively in small administrative or geographic units. The households which are closed must be put in the same group because they have same similarity than those are far apart, and households could also be supposed to have similar characteristics. The regression model which governing this structure in population can be given by:

$$Y_{ij} = X_{ij}\beta + c_i + h_{ij}. \tag{2.8}$$

Here Y_{ij} is the measurement of the j^{th} household into the i^{th} group, c_i the error term held in common by the i^{th} group, h_{ij} the household level error into the group. These two sources of error may be shown by their variances respectively σ_c^2 and σ_h^2. We see that the auxiliary

variables X_{ij} can be helpful mainly in describing the level of variation in the cluster, or the level of variation in the household. The measurement for a specific small area will be the mean of the predicted within that area. As the sample size becomes bigger as the standard error of an average becomes smaller, size of the sample at each level gives the contribution to the whole standard error of the variation at that level, household and group.

In a small area, we could have a great number of households comparing with the number of groups, in order to obtain appropriate standard errors for estimates of the small area it is necessary that, the non explained group$_l evel variance \sigma_c^2$, at the very elevated level could be small.

2.4 Design-based methods

Design-based method is also called the randomization method. This method is founded on standard probability sampling theory (see Fuller 2009). The factor that influences the randomness in the estimation method, is the sampling design utilized to choose the sample. The population measurements of this sample are considered as fixed (e.g., Lehtonen and Veijanen 2009; Pfeffermann 2002). Due to the source of the related data utilized, there is two types of estimators: direct estimators (when the data used are only coming from the specific domain) and indirect estimators (when the estimation method "borrows strength" from related areas).

2.4.1 Direct estimators

The direct estimator is performed by utilizing lone the sample data coming from the area in which the research is conducted. In the design-based method, the direct estimators of small area quantities concerns to estimate the characteristics of interest, for instance population mean or area mean (Sarndal et al., 1992).
In this subsection U represents a whole population of size N, whose m areas. Each area is denoted by U_i and it has population size $N_i; (i = 1, ..., m)$. $U = \bigcup_{i=1}^{m} U_i$ and $N = \sum_{i=1}^{m} N_i$; (the areas are disjoint), s of size n, represents the selected sample from the whole population U; $\pi_j = \frac{n}{N}$ of unit j, represents its probabilities of selection and $w_j = \frac{1}{\pi_j}, j \in s$, represents its sampling weights.
Y represents our characteristic of interest, y_{ij} represents the observation of the j^{th} subpopulation taken from the small area i (with $i = 1; ...; m; j = 1; ...; N_i$) and s_i represents the corresponding sub-sample of size n_i coming from the small area i such that $s = s_1 \bigcup ... \bigcup s_m$ and $n = \sum_{i=1}^{m} n_i$. The total area mean $\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$

By considering simple random sampling without replacement, $\pi_{ij} = \frac{n_i}{N_i}$; $w_{ij} = \frac{N_i}{n_i}$ the direct estimator of means is given by

$$\widehat{\bar{Y}}_i = \frac{1}{N_i}\sum_{j=1}^{n_i} w_{ij}y_{ij} = \frac{1}{N_i}\sum_{j=1}^{n_i} y_{ij} = y_{i.} \tag{2.9}$$

$\bar{y}_i$ is the sample area mean.

The variance of this estimator is unbiased and it is calculated as follow:

$$Var(\widehat{\bar{Y}}_i) = (1-\pi_{ij})\frac{S_i^2}{n_i}, \tag{2.10}$$

where $(1-\pi_{ij})$ represents the finite population correction factor and

$$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{Y}_{i.})^2, \quad N_i > 2. \tag{2.11}$$

If N_i is unknown

$$Var(\widehat{\bar{Y}}_i) = (1-\pi_j)\frac{s_i^2}{n_i}. \tag{2.12}$$

with

$$s_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{Y}_{i.})^2, \quad N_i > 2. \tag{2.13}$$

s_i^2 is unbiased estimator of S_i^2. Danny Pfeffermann (2013). Important Developments in Small Area Estimation

It is clear that for small n_i, the variance will be bigger except that the variation of the y values is very small.

In order to have an efficient estimator we assume that we have the auxiliary variables for every sample unit and $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$; $_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$ which are the corresponding sample mean and population mean respectively are known. Here x_{ij} is the vector of auxiliary variables for j^{th} observation in area i. Therefore, the efficient estimator which is also called *regression estimator* and is obtained by

$$\bar{Y}_i^{reg} = \bar{y}_i + (\bar{X}_i - \bar{x}_i)'\beta_i \tag{2.14}$$

its variance equals $Var(\bar{Y}_i^{reg}) = (1-\rho_i^2)(1-\pi_{ij})\frac{S_i^2}{n_i} = (1-\rho_i^2)Var(\bar{Y}_i)$,

where β_i is the vector of regression coefficients and it is calculated using the data taken from the area i and ρ_i is the multiple correlation between the survey variable Y and auxiliary variables x_{ij} in area i.

It is obvious that the variance is reduced because of the factor $(1-\rho_i^2)$. So, the auxiliary information increase the level of precision in SAE.

2.4.2 Indirect estimators

Indirect estimators concern to "borrows strength" from related areas and/or from the other time periods is used, in order to have accuracy of small area estimates. This indirect estimators increase the sample size. As the sample size increases, as the estimate is more efficient due to the decreasing of variance as it is seen in the above subsection. The auxiliary data and the sample data are combined to produce an efficient precision. In this model the auxiliary data can used alone but they may produce inefficient precision. So it is better to use both auxiliary data and sample data. the efficient of indirect estimators is not only based on the appropriate auxiliary data, but also is based on the choice of the appropriate linking model. The estimate may be inefficient due to errors coming from using of the linking model which is not appropriate. To use appropriate liking model reduces the errors depending on mischoosing of the model.
As it is said above, indirect estimates are used to be more precise (when the direct estimates are not precise) in order to make a conclusion of the population. The direct estimator of the broad small area, can be used to obtain indirect estimate of the small areas covered by that broad area (Gonzalez, 1973). This estimator is called *regression synthetic estimator* and it is denoted by $\bar{Y}^{reg}_{syn,i}$. It is called synthetic because it uses the estimate of broad domain to estimate the quantity of the small domains, which are covered by this broad domain. All these domains are supposed to be homogeneous and having the same quantity to be estimated.

$$\bar{Y}^{reg}_{syn,i} = \bar{y}_i + (\bar{X}_i - \bar{x}_i)'\hat{\beta}, \tag{2.15}$$

where $\hat{\beta}$ is calculated using the data coming from the whole areas.

2.5 Model-based methods

Model-based methods are also indirect estimators because of the use of covariates (the auxiliary data) in order to increase the level of precision. These auxiliary data are gotten from the large survey or/and administrative records such as census or registers. As it is seen before the sufficient accurate prediction can not be achieved when the sample size is small and the model is not appropriate. So the model-based methods are wonderful due to their utility of achieving on the sufficient accurate prediction of small areas. In this section, $\tilde{\theta}_i$ represents direct estimate of true area mean μ_i; y_i denotes the observed variable for area i and x_i denote the corresponding values of the covariates, y_i is a scalar, when x_i is a vector, and it is a vector when x_i is a matrix; μ_i represent the true value of mean in area i; u_i represent the random area effect; e_i are sampling errors of direct estimator.

2.5.1 Basic Area Level Model

The area level model uses the auxiliary information which are only coming from the area level, so x_i is a vector. Firstly, the area level model is studied Fay and Herriot (1979). This model is defined by:

$$\tilde{\theta}_i = \mu_i + e_i, \quad \mu_i = x_i'\beta + u_i, \quad i = 1, ..., m. \tag{2.16}$$

From the above area level model (2.16), it is obvious that the model is divided into two models: the first is the sampling model (composed by direct estimate and the sampling error) and the second is the linking model (which relate the population value to some covariates with unknown random effect).
If the sampling model and the linking model are combined, one obtain the area level linear mixed model

$$\tilde{\theta}_i = x_i'\beta + u_i + e_i, i = 1, ..., m. \tag{2.17}$$

In the area level model, we assume that u_i and e_i are independent; the sampling errors e_i are also independent and normal distributed with $\mathbb{E}(e_i|\mu_i) = 0$ and $Var(e_i|\mu_i) = \sigma_{ei}^2$. Moreover, the random effects u_i's are assumed to be independent and identically normally distributed with $\mathbb{E}[u_i] = 0$ and $Var(u_i) = \sigma_u^2$.

2.5.2 Nested error unit level model.

In SAE, Nested error unit level model is known as the unit level model. The suggestion of this model was firstly given by Battese, Harter and Fuller (1988). In SAE, the unit level model is used if and only if the true population area means and population mean $\bar{x}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$ and auxiliary variables x_{ij} are known. It uses individual variables of interest y_{ij} such that y_i is a vector, and x_i is generally a matrix. The unit area level model is formulated as follow:
$y_{ij} = x_{ij}'\beta + u_i + e_{ij}, i = 1, ..., m, j = 1, ..., N_i$, with y_{ij} represent the values of variable of interest Y; the random effect u_i are the effect of area characteristic that are not considered for the auxiliary variables and e_{ij} are the sampling errors. The random effects u_i's and the sampling errors e_{ij}'s are supposed to be mutually independent with means zero and variances σ_u^2 and σ_e^2, respectively. The true area means are given by $\bar{y}_i = \bar{x_i'}\beta + u_i + \bar{e}_i$, For the large sample size N_i, the mean of sampling error $\bar{e}_i = \dfrac{1}{N_i}\sum_{j=1}^{N_i} e_{ij} = 0$
thus, the true area means become $\bar{y}_i = \bar{x_i'}\beta + u_i$, which are the target means. If the variances σ_e^2 and σ_u^2 are known, the BLUP of the small area means $\bar{Y}_i$ is defined as:

$$\hat{\mu}_i^{BLUP} = \bar{x}_i'\hat{\beta} + \hat{u}_i$$

$$= \bar{x}_i'\hat{\beta} + \gamma_i(\bar{y}_i - \hat{x}_i'\hat{\beta}$$
$$= (1-\gamma_i)\bar{x}_i'\hat{\beta} + \gamma_i\left[\bar{y}_i + (\hat{X}_i - \hat{x}_i)'\hat{\beta}\right],$$

with $\hat{\beta}$, the GLS estimator of $\beta, \gamma_i = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/n_i}$; represents the shrinkage factor and $\bar{y}_i$; $\bar{x}_i$, stand the sample means of y_{ij} and x_{ij} for area i, respectively and n_i stands the sample size taken from N_i units in the i^{th} area. For the unknown variances (σ_e^2 and σ_u^2), we use their estimators ($\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$). The use of these estimators, produce the EBLUP of small area means $\bar{Y}_i$ which calculated as follows:
$\hat{\mu}_i^{EBLUP} = \bar{x}_i'\hat{\beta} + \gamma_i(\bar{y}_i - \hat{x}_i'\hat{\beta})$,
with $\hat{\gamma}_i = \dfrac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i}$. For $n_i = 0$ (if there is no samples in small areas),$\hat{\gamma}_i = 0$, so the EBLUP of small area means $\hat{\mu}_i = \bar{x}_i'\hat{\beta}$.

3. MULTIVARIATE LINEAR MODEL UNDER SMALL AREA ESTIMATION SETTINGS

3.1 Univariate and multivariate normal distribution

Definition 3.1 (**Univariate normal distribution):** A standard normal random variable x with mean 0 and variance 1, denoted simply by $x \sim \mathcal{N}(0,1)$ is a random variable whose density is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \quad -\infty < x < \infty.$$

If we put $x = \frac{y-\mu}{\sigma}$, that is $y = \sigma x + \mu$, then the random variable y is normally distributed with mean μ and variance σ^2, or simply $y \sim \mathcal{N}(\mu, \sigma^2)$ and its density is given by

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < \mu, y < \infty, \ \sigma > 0.$$

Definition 3.2 (**Multivariate normal distribution):** Let $\boldsymbol{x} = (x_1, \ldots, x_p)$ be a vector which consists of p independent identically standard normally distributed elements (i.e., $\mu_i = 0$, $\sigma_i^2 = 1$ for all i and $\sigma_{ij}^2 = 0$ for $i \neq j$. Then $\boldsymbol{x}$ has standard multivariate normal distribution, or simply $\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I})$.

If we put $\boldsymbol{y} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{x} + \boldsymbol{\mu}$, then $\boldsymbol{y}$ has multivariate normal distribution $\boldsymbol{y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and its density is given by
f($\boldsymbol{y}$)=$(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\}}$.

Definition 3.3 (**Matrix normal distribution):** Let $\boldsymbol{\Sigma} : p \times p$ and $\boldsymbol{\Psi} : n \times n$ be such that $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$ and $\boldsymbol{\Psi} = \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}'$. A random matrix $\boldsymbol{Y} : p \times n$ is said to be matrix normally distributed with parameters $\boldsymbol{M} : p \times n$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, we write $\boldsymbol{Y} \sim \mathcal{N}_{p,n}(\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$ if

$$\boldsymbol{Y} \overset{d}{=} \boldsymbol{M} + \boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{\Upsilon}',$$

where $\boldsymbol{X} : p \times n$ consists of pn i.i.d. standard normal distributions x_{ij}, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, n$. If $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are positive definite, then the corresponding density function is given by

$$f(\boldsymbol{Y}) = (2\pi)^{-\frac{pn}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} |\boldsymbol{\Psi}|^{-\frac{p}{2}} e^{-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{M})\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{M})'\right\}}.$$

It follows that $\boldsymbol{Y} \sim \mathcal{N}_{p,n}(\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$ if and only if

$$\text{vec}\boldsymbol{Y} \sim \mathcal{N}_{pn}(\text{vec}\boldsymbol{M}, \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}).$$

As result

$$\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{M}, \quad \mathbb{D}[\boldsymbol{Y}] = \mathbb{D}[\text{vec}\boldsymbol{Y}] = \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma},$$

where vec$(\cdot)$ stands for is the usual column-wise vectorization operator and the symboland $\otimes$ is the usual Kronecker product. The notation " $\stackrel{d}{=}$ " means "*equality in distribution*", $|.|$ and tr denote the determinant and the trace of a matrix, respectively.

Definition 3.4 (**MANOVA model):** The Multivariate Analysis of Variance (MANOVA) model is defined by

$$\boldsymbol{Y} = \boldsymbol{BC} + \boldsymbol{E},$$

where $\boldsymbol{Y} : p \times n$, $\boldsymbol{B} : p \times k$, $\boldsymbol{C} : k \times n$ are the observation, unknown parameter and design matrices, respectively. It is assumed that $\boldsymbol{E} \sim \mathcal{N}_{p,n}(\boldsymbol{0}, \boldsymbol{\Sigma}_e, \boldsymbol{I})$ with $\boldsymbol{\Sigma}_e$ supposed to be positive definite and $\mathcal{N}_{p,n}(\boldsymbol{0}, \boldsymbol{\Sigma}_e, \boldsymbol{I})$ denotes the matrix normal distribution where the columns of $\boldsymbol{E}$ are independent and the dispersion of each column equals $\boldsymbol{\Sigma}_e$. When $n \geq \text{rank}(\boldsymbol{C}) + p$ and $\boldsymbol{C}$ assumed to have full rank, the MLEs for $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_e$ are given by

$$\begin{aligned}\widehat{\boldsymbol{B}} =& \boldsymbol{Y}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1},\\ \widehat{\boldsymbol{\Sigma}}_e =& \frac{1}{n}\boldsymbol{Y}\big(\boldsymbol{I} - \boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1}\boldsymbol{C}\big)\boldsymbol{Y}'.\end{aligned}$$

3.2 Model formulation

The working model is a MANOVA model with random effect and the unit of study is the household. In this study we suppose that the data are available at small area level. Consider a finite population U into m disjoint subpopulations or domains $U_1, ..., U_m$ called *small areas* of sizes $N_i, i = 1, ..., m$ such that $\sum_{i=1}^{m} N_i = N$, where N is the total size of the target population. Suppose that there are $\boldsymbol{y}_i : p \times 1$ characteristic vectors of interest and $\boldsymbol{x}_i : r \times 1$ vectors of fixed covariates associated to the characteristic of interest in the i-th small area, $i = 1, ..., m$. Consider the following multivariate model in the i-th small area

$$\boldsymbol{y}_i = \boldsymbol{B}\boldsymbol{x}_i + \boldsymbol{1}_p u_i + \boldsymbol{e}_i, \tag{3.1}$$

where

$$\boldsymbol{e}_i \sim \mathcal{N}_p\big(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_p\big)$$

is independently distributed with

$$u_i \sim \mathcal{N}\big(0, \sigma_u^2\big),$$

for $i = 1, ..., m$, where $\boldsymbol{B} : p \times r$ is a matrix of regression coefficients, $\boldsymbol{u}_i$ is a vector of specific random effects, σ_e^2 is the sampling error variance assumed known since stems from a survey study. and σ_u^2 is a model error variance corresponding to the specific area effects. The notation $\mathbf{1}_p$ stands for a vector of ones of order p and the variances σ_e^2 and σ_u^2 are assumed to be common for all small areas.

Put

$$\begin{aligned}
\boldsymbol{Y} =& [\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_m] : p \times m, \\
\boldsymbol{X} =& [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m] : r \times m, \\
\boldsymbol{E} =& [\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_m] : p \times m, \\
\boldsymbol{u} =& [u_1, u_2, ..., u_m]' : m \times 1.
\end{aligned}$$

Collecting all m small areas together, then the model at area level can be expressed as

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X} + \mathbf{1}_p \boldsymbol{u}' + \boldsymbol{E}, \tag{3.2}$$

$$\tag{3.3}$$

$$\boldsymbol{E} \sim \mathcal{N}_{p,m}(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_p, \boldsymbol{I}_m), \quad \boldsymbol{u} \sim \mathcal{N}_p(\mathbf{0}, \sigma_u^2 \boldsymbol{I}_p).$$

That is

$$\boldsymbol{Y} \sim \mathcal{N}_{p,m}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{I}_m), \quad \text{where} \;\; \boldsymbol{\Sigma} = \sigma_u^2 \mathbf{1}_p \mathbf{1}_p' + \sigma_e^2 \boldsymbol{I}_p.$$

The proposed model (3.3) is a multivariate version that belongs to the class of the extension of the univariate Fay-Herriot model, originally proposed by Fay and Herriot Fay1979.

3.3 Estimation of the mean parameter and variance for random effects

We assume that the model defined in (3.3) holds for both sampled and nonsampled population units, i.e., the sampling method is non-informative. For the purpose of

estimation, we consider the corresponding model for the sampled data and reduce it to a Multivariate Analysis of Variance (MANOVA) model by

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X} + \widetilde{\boldsymbol{E}},$$

with $\widetilde{\boldsymbol{E}} = \mathbf{1}_p\boldsymbol{u}' + \boldsymbol{E}$; where $\boldsymbol{Y} : p \times m$, $\boldsymbol{B} : p \times r$, $\boldsymbol{X} : r \times m$ are the observation, unknown parameter and design matrices, respectively. It is assumed that $\widetilde{\boldsymbol{E}} \sim \mathcal{N}_{p,m}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{I})$ with $\boldsymbol{\Sigma}$ supposed to be positive definite and $\mathcal{N}_{p,m}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{I})$ denotes the matrix normal distribution where the columns of $\boldsymbol{E}$ are independent and the dispersion of each column equals $\boldsymbol{\Sigma}$. We assume that $m \geq \text{rank}(\boldsymbol{X}) + p$ and $\boldsymbol{X}$ to have full rank, the MLEs for $\boldsymbol{B}$ and $\boldsymbol{\sigma}_u^2$ are given by

$$\begin{aligned}
\widehat{\boldsymbol{B}} =& \boldsymbol{Y}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}, \\
\widehat{\sigma}_u^2 =& \frac{1}{pm}\text{tr}\Big\{\big(\boldsymbol{Y} - \widehat{\boldsymbol{B}}\boldsymbol{X}\big)\big(\boldsymbol{Y} - \widehat{\boldsymbol{B}}\boldsymbol{X}\big)'\Big\} - \sigma_e^2 \\
=& \frac{1}{pm}\text{tr}\Big\{\boldsymbol{Y}\big(\boldsymbol{I} - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\big)\boldsymbol{Y}'\Big\} - \sigma_e^2.
\end{aligned}$$

3.4 Prediction of random effects and small area proportions

In this section, we use the approach developed by Henderson1973 for prediction of random effects which consists of maximizing the joint density of the observable and non-observable random variable.
Consider the model in (3.3) given by

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X} + \mathbf{1}_p\boldsymbol{u}' + \boldsymbol{E},$$

and according to Henderson1973 we shall maximize the joint density $f(\boldsymbol{Y}, \boldsymbol{u})$ with respect to $\boldsymbol{u}$ assuming the other parameters $\boldsymbol{B}, \sigma_u^2$ and σ_e^2 to be known. We get

$$\begin{aligned}
f(\boldsymbol{Y}, \boldsymbol{u}) =& f(\boldsymbol{u})f(\boldsymbol{Y}|\boldsymbol{u}) = \lambda \exp\Big\{-\frac{1}{2}\{\sigma_u^{-2}\boldsymbol{u}'\boldsymbol{u}\}\Big\} \\
& \times \exp\Big\{-\frac{\sigma_e^{-2}}{2}\text{tr}\big\{(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{X} - \mathbf{1}_p\boldsymbol{u}')(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{X} - \mathbf{1}_p\boldsymbol{u}')'\big\}\Big\},
\end{aligned}$$

where λ is a known constant. Then, the predicting equation for $\boldsymbol{u}$ equals

$$\sigma_e^{-2}\mathbf{1}_p'(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{X} - \mathbf{1}_p\boldsymbol{u}') - \sigma_u^{-2}\boldsymbol{u}' = \mathbf{0}.$$

which gives

$$\widehat{\boldsymbol{u}} = \frac{\widehat{\sigma}_u^2}{p\widehat{\sigma}_u^2 - \sigma_e^2}(\boldsymbol{Y} - \widehat{\boldsymbol{B}}\boldsymbol{X})'\mathbf{1}_p,$$

after some calculations.

The prediction of small area means is performed under the superpopulation approach to finite population in the sense that estimating the small area means is equivalent to predicting small area means of non sampled values, given the sample data and auxiliary data. The model in (3.1) is partitioned into two parts $\boldsymbol{Y}^{(s)}$ and $\boldsymbol{Y}^{(r)}$ of sampled n and $N-n$ non sampled units and similar partitions apply for vector of covariates $\boldsymbol{x}_i^{(s)}$ and $\boldsymbol{x}_i^{(r)}$. One can refer to Battese1988 about estimation of characteristics of interest under the superpopulation approach for small area estimation statistics.

Therefore, the small area proportions of the characteristics of interest are obtained by

$$\widehat{\boldsymbol{\mu}}_i = f_i\Big(\widehat{\boldsymbol{B}}\boldsymbol{x}_i^{(s)} + \mathbf{1}_p\widehat{u}_i\Big) + (1-f_i)\Big(\widehat{\boldsymbol{B}}\boldsymbol{x}_i^{(r)} + \mathbf{1}_p\widehat{u}_i\Big),$$

where $f_i = \dfrac{n_i}{N_i}$ is the fraction of sampled units in the i-th small area.

4. EMPIRICAL STUDY OF OF POVERTY INDICATORS IN RWANDA

4.1 Background

- Background and status of poverty in Rwanda

In Rwanda, the poverty has a relation with the chain of the different problems, in specific land demography, degradation of an environment, unfavorable administration and in addition nether restricted sources of growth. These problems and constraints played a big role to continue the reduction of the income support and cause the extensive of the poverty. The incidence of poverty is assessed over 65 percent and rising. This increasing of poverty runs especially in the prematurely 1990s due to the friable social capital and the genocide of1994 which destroyed many infrastructure, and rose the portion of the defenseless groups (NISR, 2015).

In order to fight against poverty, Rwanda has made a strong effort where Economic Development and Poverty Reduction Strategy (EDPRS) has been introduced. This one (EDPRS), adjusts the objectives of country, the policies which are major and priorities. EDPRS is very important because it gives the guide lines which are necessary in the different domain. It gives a mediumly restriction stricture in order to achieve the sustainable development of the the country and aspirations as presented in Rwanda Vision 2020.

Viable expansion of exportation and and jobs, VUP and good process of governing are the main programs which present the importance of strategy. The advanced importance is assigned by EDPRS in order to advance the creation of jobs and production of exportation. This one will be achieved due to an ambition, level of excellence in general program of to invest in order to reduce the costs of commercial enterprise or establishment. This force will generate powerful interest on non governmental organization and on the businessperson so that they can augment their investment (NIRS, 2015).

In order to put up these above programs and planning for poverty reduction, the first step is to know the current status across all the parts of the country.

4.2 Description of the data

4.2.1 Source of the data

This research has the target to produce the estimates of poverty and extreme poverty at sector level in Karongi district. So, in this study the small area is sector. To perform it, we have used the data from the fourth Integrated Household Living Conditions Survey (EICV2013-2014) containing the poverty status in Rwanda at the district level and Rwanda poverty mapping report 2013-2014, District profile Karongi 2012-15 containing the different information which show how Karongi district stands in different domains (aspects).
Table 4.1, shows the status of poverty at district level

Tab. 4.1: Population identified as poor and extremely poor by district, 2013/14

Districts	**Poverty Incidence(%)**	**Extreme Poverty Incidence(%)**
Kicukiro	16.3	6.5
Nyarugenge	19.9	8.4
Gasabo	23.4	11.3
Rwamagana	25.4	6.0
Kamonyi	25.9	6.0
Kayonza	26.4	9.5
Muhanga	30.5	7.8
Huye	32.5	5.7
Bugesera	34.3	13.4
Musanze	34.9	16.8
Rusizi	35.1	15.8
Rubavu	35.5	14.2
Ruhango	37.8	12.8
Nyanza	38.0	17.6
Nyabihu	39.6	12.6
Nyamagabe	41.5	13.0
Kirehe	41.8	17.8
Gakenke	42.0	16.2
Gatsibo	43.8	18.5
Nyagatare	44.1	19.5
Karongi	45.3	21.3
Ngoma	46.8	19.5
Nyaruguru	47.9	20.0
Rulindo	48.1	20.2
Ngororero	49.6	23.5
Burera	50.4	23.0
Rutsiro	51.4	23.6
Gisagara	53.3	20.6
Gicumbi	55.3	24.7
Nyamasheke	62.0	39.2

Source:Rwanda poverty profile report EICV 2013-2014

As this research is only conducted in Karongi district, the data used in this thesis are found in District profile Karongi 2012-15. Karongi is one of the 30 districts of Rwanda and it is located in Western Provence. Rwanda has 4 provinces which are Western, Eastern, Northern and Southern Province. This research is limited on Karongi district due to the some limitations such as time, budget... Karongi district has 13 sectors, 88 cells, 537 villages and 73,326 households.

In Rwanda, every district has its own profile. This profile shows the status of district in different aspect such as the number of its citizens, households, cells and sectors. It also shows how the district's citizen stand according to their satisfaction of their basic needs such as food security, clothing, access to health insurance, the capacity to attend the school, the nature of house in which they live, their occupation and so on. This profile is updated after every 5 years, and it is done in the following way:

The chief of village collects the information showing how every household in its village stands. This report is sent at cell level and the executive secretary of cell verifies if the given information are correct; and after verification, he/she sends it at sector level. The sector also does the deep verification on this report and after, it send it at district level and then they make this district profile. Table 4.2, shows how the households of Karongi district involved in this research.

Tab. 4.2: Distribution of private households in Karongi district

Sectors	**Private households(PH)**	**Sampled households(SH)**	**Unsampled households(USH)**	**Ratio f=SH/PH**	**1-f**
Bwishyura	7919	736	7183	0.093	0.907
Gishari	4693	512	4181	0.109	0.891
Gishyita	4783	560	4223	0.117	0.883
Gitesi	5633	720	4913	0.128	0.872
Mubuga	4369	512	3857	0.117	0.883
Murambi	5106	576	4530	0.113	0.887
Murundi	6142	624	5518	0.102	0.898
Mutuntu	5075	688	4387	0.136	0.864
Rubengera	7869	880	6989	0.112	0.888
Rugabano	7283	944	6339	0.130	0.870
Ruganda	4007	528	3479	0.132	0.868
Rwankuba	4870	656	4214	0.135	0.865
Twumba	5577	656	49217	0.118	0.882

Source:District profile Karongi 2012-15

4.2.2 Statistics of covariates

As in this research the target parameters are the poverty and extreme poverty, the covariates are chosen based on the factors that influence the poverty in Rwanda which are also the factors which influencing the poverty in Karongi district. Among the many factors influencing the poverty, we chose the five factors which are the following:

1. The production of agriculture: The majority of the population of Rwanda is engaged in agriculture sector. That is why the crop production is an essential factor that can be based on, to determine the status of poverty and extreme poverty in certain areas in Rwanda.

2. Roof with local tiles: Actually, if you look at the living house of someone, you can get some inspiration about of her/his social status. That is why the roof with local tiles was chosen as a factor that can be used in this research.

3. Roof with other materials: In general, the building materials of a house is an important factor that can give some information about the poverty.

4. Unimproved sanitation facilities: In the poor household, it is difficulty to obtain the facilities about the improved sanitation. That is why this factor also was chosen in this study.

5. Number of households have access to electricity: This factor was chosen due that the poor households can not get the capacity of using the electricity.

Tab. 4.3: Poverty indictors in Karongi district

Sectors	**No access to electricity(%)**	**Production(%)**	**Roof with local tiles(%)**	**Roof with other materials(%)**	**Unimproved sanitation facilities(%)**
Bwishyura	71.90	14.74	20.10	1.20	2.30
Gishari	94.70	5.67	92.40	1.20	2.30
Gishyita	91.00	5.84	31.40	2.10	3.00
Gitesi	97.70	5.99	83.20	0.80	2.20
Mubuga	93.90	8.44	43.50	1.30	3.70
Murambi	93.00	4.07	92.20	0.60	2.40
Murundi	99.00	8.44	94.40	0.70	1.90
Mutuntu	98.60	5.02	93.70	0.70	3.30
Rubengera	82.80	19.18	68.20	1.10	4.90
Rugabano	98.80	9.86	96.10	0.90	1.40
Ruganda	98.40	5.42	96.10	0.90	1.80
Rwankuba	98.50	3.38	77.10	1.30	2.10
Twumba	96.10	3.93	79.40	1.00	1.30

Source:District profile Karongi 2012-15

4.3 Data analysis and discussion of the results

The working model used in this study to find estimates is MANOVA model

$$\boldsymbol{Y} = \boldsymbol{BX} + \mathbf{1}_p \boldsymbol{u}' + \boldsymbol{E}, \tag{4.1}$$

Where

$$\begin{aligned}
\boldsymbol{Y} &= [\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{13}] : 2 \times 13, \\
\boldsymbol{X} &= [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{13}] : 5 \times 13, \\
\boldsymbol{E} &= [\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_{13}] : 2 \times 13, \\
\boldsymbol{u} &= [u_1, u_2, ..., u_{13}]' : 13 \times 1.
\end{aligned}$$

In the above model, Y is the matrix of order 2×13 of observations, this 2 means that there is two target parameters which are poverty and extreme poverty; 13, means there are 13

sectors in Karongi district. X represents the covariate matrix of order 5×13 means we have 5 chosen factors which influence the poverty in 13 sectors of Karongi district. E is the random error matrix. $\boldsymbol{B}$ is a matrix of regression coefficients and $\boldsymbol{u}$ is a vector of random effects.

By using MATLAB software, version 9.0.0.341360, The MathWorks, In USA, we obtain the following results presented in table 4.4.

Tab. 4.4: Summary table of the obtained results: Estimates of poverty and extreme poverty at sector level.

Sectors	**Poverty Incidence(%)**	**Extreme Poverty Incidence(%)**
Bwishyura	43.047	27.7159
Gishari	60.7178	38.3568
Gishyita	57.4989	36.7102
Gitesi	65.5478	42.0906
Mubuga	61.1681	39.2641
Murambi	57.2281	34.5107
Murundi	64.9943	41.5354
Mutuntu	73.2498	49.2456
Rubengera	49.9902	32.3044
Rugabano	66.9409	43.9775
Ruganda	66.0656	42.5042
Rwankuba	70.2774	46.6706
Twumba	65.4524	42.3555

According to the Table 4.4, Bwishyura and Rubengera sectors are less suffering from poverty, with poverty incidence of 43.047%, and 49.9902% respectively; and with the extreme poverty incidence of 27.7159% and 32.3044% respectively. Mutuntu sector is the most suffering from poverty with the poverty incidence of 73.2498% and the extreme poverty incidence of 49.2456% followed by Rwankuba with the poverty incidence of 70.2774% and the extreme poverty incidence of 46.6706%.

5. CONCLUSION AND RECOMMENDATION

5.1 Conclusion

During this study, we identified the poverty indicators in Rwanda, in order to provide the reliable estimates of poverty mapping and poverty status at sector level. These estimates was found by using existing micro-data and other relevant auxiliary data. The result obtained in the Table 4.3 shows that the sectors of Karongi district are not poor at the same level. So, there is inequality in sectors of Karongi District. This is to mean that all sectors are not likely equally poor. In case of using funds for aid, it is not wise to think on countrywide level but focus on some more suffering sectors. The poverty is found to be the main challenge in all sectors of Rwanda and this shows how much attention is needed. Among all the sectors in Karongi District, Bwishyura sector is less suffering from poverty and Mutuntu sector is the most suffering from poverty as it is seen on the estimated proportions of each of 13 sectors in Karongi District.

5.2 Recommendation

- It will be better if the poverty status is conducted in the all sectors of the whole country but because of time, it was impossible with this work; that is why we recommend to the other researchers to work for other remaining sectors of Rwanda
- This work can not be limited to poverty. We recommend to other researchers to conduct a similar research in other field like malnutrition, malaria, HIV, non communicable diseases (NCD) like hypertension and diabetes mellitus ,...
- An other recommendation also goes to Karongi District Executive committee to prioritize the poorest sectors in order to help them in the progress of sustainable development and poverty reduction.

BIBLIOGRAPHY

[1] Alderman, H., Babita, M., Demombynes, G., Makhatha, N., & zler, B. (2002). How low can you go? Combining census and survey data for mapping poverty in South Africa. Journal of African Economies, 11(2), 169-200.

[2] Atkinson, P. M. & Lloyd, C. D. (2002). Deriving DSMs from LiDAR data with kriging. International Journal of Remote Sensing, 23(12), 2519-2524.

[3] Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association, 83(401), 28-36.

[4] Brackstone, G. J. (1987). Small area data: policy issues and technical challenges. Small Area Statistics, 3, 20.

[5] Cheli, B., & Lemmi, A. (1995). Atotallyfuzzy and relative approach to the multidimensional analysis of poverty. Economic notes, 24, 115-134.

[6] Elbers, C., Lanjouw, J. and Lanjouw, P. (2003), Micro-level estimation of poverty and inequality, Econometrica, 71, 355-364.

[7] Elbers, C., & Lanjouw, P. (2001). Intersectoral transfer, growth, and inequality in rural Ecuador. World Development, 29(3), 481-496.

[8] Fay III, R. E., & Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366a), 269-277.

[9] Fuller, W. A. (2009). Introduction to statistical time series (Vol. 428). John Wiley & Sons.

[10] Gonzalez, M. E. (1973). Use and evaluation of synthetic estimators. Proceedings of the social statistics section, American Statistical Association, pages 33-36.

[11] Hansen, M. H., & Hurwitz, W. N. (1953). Sample survey methods and theory. Vol. I. John Wiley And Sons, Inc.; New York.

[12] Healy, A. J., Jitsuchon, S., & Vajaragupta, Y. (2003). Spatially disaggregated estimation of poverty and inequality in Thailand. preprint.

[13] Henderson, C. R. (1973), Sire evaluation and genetic trends. Journal of Animal Science, Symposium, 10-14.

[14] Hidiroglou, M. (2007). Small-area estimation: Theory and practice. Proceedings of the Survey Research Methods Section.

[15] Kish, L. (1965). Sampling organizations and groups of unequal sizes. American sociological review, 564-572.

[16] Lehtonen, R., & Veijanen, A. (2009). Design-based methods of estimation for domains and small areas. In Handbook of statistics (Vol. 29, pp. 219-249). Elsevier.

[17] Leite, P. G., Elbers, C., Lanjouw, J., & Lanjouw, P.,(2001). Poverty and inequality in Brazil: new estimates from combined PPV-PNAD data, World Bank. DECRG, Mimeo.

[18] Marshall, R. J. (1991). A review of methods for the statistical analysis of spatial patterns of disease. Journal of the Royal Statistical Society: Series A (Statistics in Society), 154(3), 421-441.

[19] NISR. (2015). Rwanda poverty profile report: integrated household living conditions survey 2013/14.

[20] Pfeffermann, D. (2002). Small area estimation new developments and directions. International Statistical Review, 70(1), 125-143.

[21] Pfeffermann, D. (2013). New important developments in small area estimation. Statistical Science, 28(1), 40-68.

[22] Pratesi, M. (Ed.). (2016). Analysis of poverty data by small area estimation. John Wiley & Sons.

[23] Rao, J. N. K. (2003). Some new developments in small area estimation. Journal of the Iranian Statistical Society, 2(2), 145-169.

[24] Rao, J. N. K. and Isabel, M. (2015), Small area estimation. John Wiley and Sons, New York,2edition.

[25] Sarndal, C. E., Swensson, B., & Wretman, J. (1992). Model assisted survey sampling. Springer Science & Business Media.

[26] Stephen, H., & Geoffery, J. (2003). Local Estimation of Poverty and Malnutrition in Bangladesh.

[27] Stephen, H., & Geoffery, J. (2005). Estimation of Local Poverty in the Philippines. Manila: National Statistical Coordination Board.

[28] Stephen, H., & Geoffery, J. (2008). Potential for small area estimation and poverty mapping at constituency and at gewog/town level in Bhutan. World Food Programme, New Zealand.

[29] Twesigye, C., & Ntabana, I. (2007). Poverty-environment Indicators & Strategies for Monitoring Them Within the Framework of the EDPRS. Rwanda Environment Management Authority.

[30] Valliant, R., Brick, J. M., & Morganstein, D. (2000). Analysis of complex sample data using replication.

[31] Zhao, Q. (2006). User manual for povmap. World Bank.