ADDIS ABABA UNIVERSITY
OFFICE OF GRADUATE PROGRAMS
FACULTY OF SCIENCE
DEPARTMENT OF STATISTICS

WOREDA LEVEL ESTIMATES OF AGRICULTURAL AREA AND PRODUCTION: SMALL AREA ESTIMATION USING AUXILIARY DATA

By

Seid Jemal

A Thesis submitted to the Office of Graduate Programs of Addis Ababa University in Partial fulfillment of the requirement for the Degree of Master of Science in Statistics

AUGUST 2009
ADDIS ABABA

ADDIS ABABA UNIVERSITY
OFFICE OF GRADUATE PROGRAMS
FACULTY OF SCIENCE
DEPARTMENT OF STATISTICS

WOREDA LEVEL ESTIMATES OF AGRICULTURAL AREA AND PRODUCTION: SMALL AREA ESTIMATION USING AUXILIARY DATA

By
Seid Jemal

Approved by the Board of Examiners:

Sileshi Fenta, Asst. Prof. . . .
Department Head **Signature**

Eshetu Wencheko, Prof. . . .
Examiner **Signature**

Emmanuel G/Yohannes, Ph.D. . 28/08/09
Examiner **Signature**

Tel. 251-1-239465 **P.O.Box 1176**
AddisAbaba, Ethiopia

ACKNOWLEDGMENTS

Profound thanks are due to my advisor Dr. Fentaw Abegaz for his unfailing encouragement, guidance, constructive comments and useful suggestions.

I also owe my gratitude to my mother W/o Aysha Ahmed and my sisters and brothers for their familial love, affection and encouragement.

I would also like to extend my heartfelt thanks to those friends who have been friends indeed. I have to mention at least Joshua, Wonde, Minilik, Jemil, Zele, Ayu, Yarede, Lege, Zemen, Henny, Dawd, Amex, Zerish, Mekue, Mamen, Tedo, Mube, Getcho for their brotherly support, advice and encouragement.

I would like also to thank the Central Statistical Agency, for financial and material support of my postgraduate study and this thesis. Specially, I would like to offer my thanks to Ato Yaekob Mudesir, Ato Dawit Dinku, Ato Zenaselases Sium, Ato Kassu Gebeyehu and Ato Alemayehu Teferi.

Finally, I would like to forward my thanks to my teachers who gave me their invaluable assistance, encouragement and useful advice that enable me to pursue not only this thesis but also my academic life too!

Table of Contents

ABSTRACT

This study applies small area estimation technique to analyze available agricultural data to estimate the total of basic agricultural variables such as total cultivated area in hectares and total production in quintals of Teff for woredas in the Tigray Region using cultivated area and production data from 2001/02 National Agriculture Sample Census Enumeration and from Ministry of Agriculture and Rural Development as an auxiliary variable. An area level model for small area analysis is used to produce estimate on cultivated area and production for 33 woreda in the Tigray Region. The estimation process shows significant woreda-specific random effect which leads to acquire a better estimate for cultivated area and production.

CHAPTER ONE

INTRODUCTION

1.1. Background of the Study

Sample surveys have long been recognized as cost-effective means of obtaining information on wide-ranging topics of interest at frequent intervals over time. The data generated from these sample surveys are extensively used to provide reliable direct estimates of totals and means for the whole population and large areas or domains. Over time, the range of topics investigated using survey methods has broadened enormously as policy makers and researchers have learned to appreciate the value of quantitative data. In response to policy makers' demand, survey researchers have tackled topics previously considered unsuitable for study using survey methods. The range of analyses of survey data has also expanded, as users of survey data have become more sophisticated and as major developments in computing power and software have simplified the computation involved. In the early days, users were mostly satisfied with national estimates and estimates for major geographical regions and other large domains.

The situation is very different today: more and more, policy makers are demanding estimates for small domains for use of making policy decisions. In recent years, the demand for small area statistics has greatly increased worldwide. This is due, among other things, to their growing use in formulating policies and programs, in the allocation of government funds and in regional and local planning.

The need to provide estimates for small domains has led to developments in different directions. One direction is the use of sample designs that can produce domain estimates of adequate precision within the standard design-based mode of inference used in survey analysis (i.e., direct estimates). Many sample surveys are now designed to yield sufficient sample sizes for the key domains to satisfy the precision requirements for those domains.

This approach is generally used for socio-economic domains and for some larger geographic domains. However, the increase in sample sizes may limit the survey's resources and capabilities.

In the other direction, small area estimation technique was developed and used nowadays by many researchers. Here, the term small area can be defined as a small geographical area of country, region or census division. It can also be defined as a small domain of population or as a small sub-population of the total population such as age, sex, race in a large geographical area. For instance "woreda" or "kebele" could be considered as small area for Ethiopian context and in similar fashion domain is regarded as large if the domain size is large enough to yield direct estimates of adequate precision; otherwise, if any subpopulation, for which direct estimates of adequate precision cannot be produced, is regarded as small (Rao, 2003).

Accordingly, small area estimation is defined as a statistical technique of producing reliable estimates for local areas (small domain) with a certain precision based on survey data, which was launched aiming to produce the necessary information at some higher level. Usually, small area estimates are obtained by fitting statistical models to survey data and applying these models to available information for the small area population of interest.

Currently, many small area estimation techniques have been developed which make use of information from other data sources. They also borrow strength from related or similar areas through explicit and implicit models that connect the small area via supplementary data for example census and administrative records. The current emphasis and recent advances in this area also take advantage of the significant advances in statistical data processing. The advances in computing facilities have also provided convenient tools for many theoretical developments in this area (Rao, 2005).

In line with the development of small area estimation techniques, the demand for small area statistics has greatly increased in many countries worldwide. This is due to small

area statistics importance in formulating and execution of policies and programs in local and regional administrative units. Accordingly, many countries take viable action to practice the small area estimation technique based on the sample surveys and auxiliary information available within the country. The situation has not been different in the case of Ethiopia.

Ethiopia is a country which situated in the Horn of Africa between 3 and 15 degrees north latitude and 33 and 48 degrees east longitude. As the country is located within the tropics, its physical conditions and variations in altitude have resulted in great diversity of terrain, climate, soil, flora and fauna. Ethiopia has a population 73.9 million with the majority of the population living in the highland areas of the country. Agriculture remains to be the base of Ethiopian economy. In 1960's agriculture accounted for about 65 percent of gross domestic product (GDP). Recently it accounts nearly for 45.9 percent of GDP and about 60 percent of exports. Agriculture provides raw materials for 70 percent of the country's large and medium sized agro-industries. About 80 percent of the population depends on agriculture for their livelihood and the main occupation of most of the settled population is farming (NBE, 2007). The Ethiopian economy has shown mixed performance in which the variability of growth was mostly a result of the variability in the output of the agricultural sector. The Tigray Region has also almost similar socio-economic condition with a population of 3.47 million.

By and large, agriculture in Ethiopia is subsistence. This is particularly true to the major food crops grown in the country. The major food crops are produced in almost all regions of the country in spite of the variation in volume of production across the regions. The variation may be attributed to the extent of area devoted to each crop type, weather change and a shift in preference for the crops grown. These crops have been categorized into eight groups: cereals, pulses, oilseeds, vegetables, root crops, fruit crops, stimulant crops and sugar cane. Stimulant crops consist of chat, coffee, etc.

In Ethiopia, information has been collected on various socio-economic dynamics through censuses and sample surveys since 1961. In particular, data on agriculture were collected

using annual sample surveys. According to the 2007/08 Ethiopia Agricultural Sample Survey cereals are produced in larger volume in the country compared to other crops because they are the principal staple crops. Cereals are grown in all the regions with varying quantity. The finding of the survey shows that, out of the total grain crop area, 79.69% (8.7million hectares) was under cereals. Out of this Teff, maize, wheat and sorghum took up 23.42% (about 2.6 million hectares), 16.12% (about 1.8 million hectares), 13.01% (1.4 million hectares) and 14.01% (1.5 million hectares) of the grain crop area, respectively. As to production, the finding gives a similar picture as that of the area. Cereals contributed 85.11% (about 137.1 million quintals) of the grain production. Maize, wheat, Teff and sorghum made up 23.24% (37.5 million quintals), 14.36% (23.1 million quintals), 18.57% (29.9 million quintals) and 16.52% (26.6 million quintals) of the grain production, respectively.

Similar figures were observed in Tigray Region. Out of the total grain crop area, 80.82% (0.71million hectares) was under cereals. Out of this Teff, maize, wheat and sorghum took up 25.18% (about 0.18 million hectares), 8.91% (about 0.06 million hectares), 14.65% (0.10 million hectares) and 24.02% (0.17 million hectares) of the grain crop area, respectively. As to production, the finding gives a similar picture as that of the area. Cereals contributed 85.04% (about 10.0 million quintals) of the grain production. Teff, wheat, sorghum and maize made up 22.77% (2.28 million quintals), 14.74% (1.48 million quintals), 10.39% (1.04 million quintals) and 9.74% (0.94 million quintals) of the grain production, respectively (CSA, 2008).

1.2. Statement of the Problem

Agriculture in Ethiopia is one of the major sectors to the development of the country. Especially, its contribution to food sufficiency makes the sector's operation compulsory and decisive. The prime role that agriculture plays in the country's political, economic and social stability makes the measures taken on the sector extremely sensitive. Therefore, basic information on the sector such as information on area of cultivated land and production is needed for socio-economic planning and policy formulation. Beside

most of the administrative, social and economic planning is undertaken at smaller administrative unit, i.e., at woreda level. But the question is, is information on cultivated area and production available in Ethiopia? Is this information also provided at smaller administrative unit like woreda? In light of this, the study attempts to give reliable information on cultivated area and production at woreda level which is useful for planning and policy formulation.

1.3. Objective of the Study

1.3.1. General Objective

The main objective of the study is to introduce the different small area estimation techniques and also to provide better and reliable estimates at small area, in our country context at woreda level.

1.3.2. Specific Objective

The specific objective of the study is:-

- to provide estimates on total area of cultivated land in hectares and total production in quintal of Teff at woreda level for Tigray Region in 2007/08 Meher Season.
- to compare the efficiency of an estimate of small area model with that of direct estimate at small area level.

1.4. Significance of the Study

Nowadays, most countries around the world conduct censuses. These censuses can give reliable estimates at a very lower level or geographic area. But, the censuses content is inherently severely restricted to very few numbers of most important demographic variables. So, censuses cannot provide estimates for much of the characteristics of interest at small area level. In addition, in most countries censuses are conducted once in

a decade and that also censuses cannot provide satisfactory estimates for intermediate time points.

In other situations, most of sample surveys are conducted with the aim to get reliable estimates at a large domain such as national, regional or zonal level. To provide estimates for small domain using sample design that can give adequate precision, we need to increase the sample size which in turn can lead a cost burden or restriction. For instance, the Ethiopian Agricultural Sample Enumeration conducted in 2001/02 by CSA gave reliable estimate at small area level i.e., woreda level for most agricultural variables. But, this enumeration was done in a total of 15,670 enumeration areas (EAs) which is 7 times more than that of the annual agriculture sample surveys conducted yearly by CSA to give reliable estimates at large domain (national, regional and zonal level) based on around 2200 EAs (CSA, 2003 and CSA, 2008).

However, currently information at lower level is needed to assist in socio-economic planning such as policy formulation, budget allocation and for delivering better services to people. In Ethiopia, most of administrative, social and economic planning is undertaken at woreda level. Therefore, the importance of statistical indicators at woreda level is increasing from time to time. Since the agriculture sector is the backbone of the country's economy, basic information on the sector is useful (1) to formulate and implement timely food security measures and to alert policy makers about the food situation of the country, (2) to develop and monitor farm program, (3) to design and allocate funding for extension service project, and (4) in order to execute effective production and distribution system, specially for private sector. On this ground, the significance of this study relies on filling the data gap at lower level by providing small area estimates at woreda level for basic agricultural variables.

1.5. Limitations of the Study

Small area estimation technique developed into the estimation process through either implicit or explicit model that provides a link to related areas through the use of

supplementary information related to the variable such as recent census counts and current administrative records. Therefore, the availability of good auxiliary data and determination of suitable linking models are crucial for the increment of precision of the estimate (Rao, 2003). However, in Ethiopia, there is substantial problem on the availability of small area (woreda) level data on predictor variables like amount of fertilizer supplied, annual average temperature/ rainfall and so on. Moreover, the quality of data obtained from administrative records is questionable. Thus, it makes the efficiency of the technique to be limited. In addition, the software used in the analysis could not manipulate all of data analysis specially model diagnostics in fitting small area models, since the data analysis technique in small area estimation is under development.

1.6. Definition of Terms

Enumeration Area (EA):- an enumeration area in the rural parts of the country is a locality which is part of farmers' association and usually consists of 150-200 households.

Household: - a household consists of one or more persons who live together and make common provisions for food and other essentials of living. These persons may pool their incomes and have a common budget to a greater or lesser extent. They may be related or unrelated persons or a combination of both. These persons are taken as members of the household.

Agriculture: - is the growing of crops and/or raising of animals for own consumption and /or sale.

Agricultural Household: - a household is considered as agricultural household when at least one member of the household is engaged in growing crops and/or raising livestock in private or in combination with others.

Crop: - includes cereals, pulses, oilseeds, vegetables, root crops, fruits, coffee, Enset, Chat, hops, sugarcane, cotton, tobacco, etc which are produced for food, making drinks, stimulation and making fabrics or clothing.

Cultivated Area:- is the process of clearing and cultivating a piece of land in order to grow the above crops.

Crop production: - is the process of growing and harvesting of the above crops for own consumption and/or sale.

Woreda: - It is an administrative division of Ethiopia, equivalent to a district. Woredas are composed of a number of Kebeles or neighborhood farmers' associations, which are the smallest unit of administrative unit in Ethiopia. Woredas are typically collected together into zones, which form a kilil (Regional administration); some woredas are not part of a zone, and are called Special Woredas, which function as autonomous entities.

1.7. Organization of the Thesis

This thesis is organized into five chapters. Chapter one deals with the introductory part which includes background, objectives, significance and limitation of the study and definition of basic terms used in the thesis. Chapter two deals with review of related literature on small area estimation either through referring studies made in Ethiopia or outside. Chapter three discusses the data and methodology of the study such as sources of data and variables to be included in the study with their description and methods of data analysis used in the study. Chapter four presents statistical data analysis and summary of the main findings. Finally, discussion and conclusions of the study are dealt with in chapter five.

CHAPTER TWO

LITERATURE REVIEW

2.1. Direct Domain Estimation

In the development of survey methods in the last decades, the rapid increase in the number and types of sample surveys is observed. Sample surveys have been designed to estimate the parameter of interest of the study population. Cochran (1977) has introduced different applications of sampling theory in the estimation of the parameter which is under consideration in sample surveys. Furthermore, Sardnal et al. (1992) gave detailed descriptions on variance estimation of the estimators. Most sample survey methodologies/ techniques have the purpose to obtain information on crop area and production and land use, size of labor force, industrial production, wholesale and retail prices, health status of the people and family income and expenditure. In particular, collection of information on the area covered by different crops and the quantity produced was conducted in most countries around the world using sample surveys since this information is considered as an important measure for the overall performance of the agriculture sector.

Agriculture sample survey data are broadly used to provide direct estimate of agricultural parameter for the whole population and large areas or domains. For instance, annual agricultural sample surveys in Ethiopia have been conducted with the aim to get reliable direct estimates at larger domain level such as national, regional as well as zonal level for the agricultural variables into consideration (CSA, 2008).

In other studies, in order to compare the precision of estimates of cropland area based on unbiased and ratio estimators, Biratu and Eskinder (2005) computed and compared unbiased direct estimate and ratio estimates using total number of holders as an auxiliary variable for the five major crops (maize, teff, sorghum, barley and wheat). They

concluded that the ratio estimator could be taken as a better alternative in providing more precise estimate of area for each crop than the design-based unbiased estimate. However, they remarked that availability of timely and accurate auxiliary information such as total number of holders from recent census or administrative records are essential in order to provide precise estimates of cropland areas.

Whereas, if the demand of estimates at sub-domain level such as woreda occurs, the domain-specific sample cannot give a reliable estimate of the parameter of interest (Gosh and Rao, 1994). This is because one of the main criteria usually used to determine sample size of nationwide surveys have the objective to give a specified level of precision for a given (large) domain. Thus, it leads the domain specific sample is to be not large enough to support direct estimate of adequate precision which in turn likely produce large standard errors due to the unduly small size of the sample. Therefore, as a rule, a domain will be regarded as *small* if the domain specific sample is not large enough to support direct estimate of adequate precision.

In practical situations, it is seldom possible to have a large overall sample size to support reliable direct estimates for all the domains of interest due to limitation in resources and scope of surveys. Therefore, it is often necessary to generate indirect estimates by using values of the variable of interest from related areas. Often there is auxiliary information that can be used to define estimators for small areas. In some cases there are values of the variable of interest in other, similar areas, or past values of the same area. These values are brought into the estimation process through either implicit or explicit model that provides a link to related areas through the use of this supplementary information which are related to the variable. The use of auxiliary information has been characterized in the statistical literature as "borrowing strength" from the relationship between the values of the response variables and the auxiliary information. Therefore, availability of good auxiliary data and determination of suitable linking models are crucial to the formation of indirect estimates. This indirect estimate offers several advantages, most importantly, it increase precision of the estimate made at small area level. The technique involved in this estimation is commonly known as small area estimation (Rao, 2003).

2.2. Indirect Domain Estimation

Various studies have been made and conclusions were given by different scholars in the past regarding the investigation of different types of methods and models for the estimation and prediction of small area parameters. Early applications of small area estimation methods employed only simple methods. However, the situation has changed enormously in recent years, and particularly in the last decades. Now, there exist a wide range of different, often complex models that can be used depending on the nature of measurement of small area estimates and the auxiliary data available. One key distinction in the methods arises from the assumption which is set on the model. Survey of literatures categorized the method used and the model applied in these studies as indirect domain estimation. Indirect domain estimation is mainly subdivided into two. One of the categories is model-assisted method which involves an estimator such as synthetic estimator, composite estimator and alike. The other is model-based method or small area model which involves mainly area level model and unit level model.

Stasny et al. (1991) used a regression-synthetic estimator to produce county estimates of wheat production in the state of Kansas. The study used a non-probability sample of farms, assuming a linear regression model relating wheat production of each farm in a specific county to a vector of predictors. The predictor variables chosen for this application consists of acres planted with wheat and district indicators on condition that these variables have known county totals. In the study, it was not necessary to know the individual values of each predictor variables for all farms in the county. The regression-synthetic estimator of a specific county total is obtained from the least square estimator of the sample data. From the result of the study, they reported the direct and synthetic estimates with their mean square error of each county. The result of the study shows that the synthetic estimates performed better than direct estimates in terms of estimated mean square error (MSE).

In general, synthetic estimates are based on implicit model that provide a link to related small areas through supplementary data based on certain assumptions and it is formed in a sense of the estimates developed for the larger areas are to be scaled down to the smaller areas on the basis of certain model assumptions. It is assumed that the relation between the larger area and small areas remain same (homogenous) for the characteristics under study as well as for the auxiliary variable. If this model assumption is true, synthetic estimation has its strength in borrowing information from larger groups for use in small domain areas. But, if the implicit model assumption that the small domains resemble each other fails, the synthetic estimator may be badly design-biased; hence it seems artificial to the probability sampler. Nevertheless, one may gamble on the synthetic estimator because strength is borrowed and the design variance is often low. Even a small departure from the assumption, however, puts the whole method in question and the mean squared error may be very high (Sardnal, 1984).

Several authors have noted that the bias potentially produced by purely synthetic methods can be reduced through procedures that essentially combine a synthetic component with another component. Gonzalez (1973) suggested that a choice between direct and synthetic estimators need not be made but a combination of the two is better than either one. A natural way to balance the potential bias of a synthetic estimator against the instability of a direct estimator is to take a weighted average of both. Under suitable choice of the weight, such type of estimators is known as composite estimator.

2.3. Indirect Domain Estimation: Small Area Model

Rao (2003) described that the model-assisted indirect domain estimator could give a biased estimator if the implicit model assumption failed. Therefore, it leads this estimate to be limited as a consequence of the method's inability to account properly for local factors such as area variation. Many researchers shifted to *explicit* small area models that make specific allowance for between area variations. In particular, they introduced mixed models involving random area-specific effects that account for between areas variation beyond that explained by auxiliary variables included in the model.

Fay and Harriot (1979) first studied improved estimation in small areas using regression model for the domain sample means of the dependent variable on the vector of domain sample means of the independent variables. The result of the study shows that the success of any model-based small area method depends on the availability of good auxiliary data. Thus, more attention should be given to the compilation of auxiliary variables that are good predictors of the study variables. Hence, this method may be classified into two broad types as that of aggregate level (or area level) models that relate small area means to area-specific auxiliary variables and unit level models that relate the unit values of the study variable to unit-specific variables. Here, it is necessary to note that area level model is essential if unit (element) level data are not available.

A review of the literature, including web sites and on-line journals, was conducted to determine the extent to which small area estimation techniques have been used in the prediction of cultivated area and production estimate at small area level. Rao (2002) describes a particular application of small area estimation in agricultural sector in detail. On estimation crop area, production and yield at district level in India, Singh and Goel (2000) used remote sensing satellite data and crop surveys. As a result, they remarked that the use of supplementary information could give a better estimate at the district level. Bartosinka (2006) used small area estimation technique in agricultural sample surveys and agricultural census data on estimation of some agricultural characteristics by region in Poland.

Russo et al. (2002) made an assessment to get small area estimates on major agricultural variable based on the 1999 Italian Farm Structure Survey data. In the paper, they described the growing demand for reliable small area statistics in the agricultural discipline in order to assess or to put into effect agricultural policies and programs. The survey direct estimates in specific areas were not reliable, because the smallness of sample sizes in the areas can drive to unacceptably large standard errors. With the goal to estimate proportions of farms falling in some qualitative classes and in certain small

areas, they used small area models to estimate the farm characteristics and they got small area statistics are powerful methods in estimating it.

In another study, with availability of auxiliary data on each farm, a particular application of unit level model on prediction of crop area of small domain using survey and satellite data was discussed by Bettese et al. (1988). In the study, an estimate crop area of corn and soybeans at county level (small domain) using the data of 1978 June Enumerative Survey of the U.S. Department of Agriculture and data obtained from land observatory satellites (LANDSAT) during the 1978 growing season. A nested-error regression model is specified and it defined a correlation structure among reported hectares within the counties. In this model, the mean hectares of the crop per segment is the sum of a fixed component, involving unknown parameter to be estimated and a random component to be predicted. Subsequently, variance-component estimators were defined and the generalized least-square estimators of the parameters of the linear model were obtained. The result of the study suggested predictor for the county mean crop per segment has a standard error that is considerably less than that of the traditional survey regression predictor.

Prasad and Rao (1990) regarded small area models particularly unit-level models as special cases of a general linear mixed model. Accordingly, Minilik (2004) used general linear mixed model to the estimation of mean harvested area per holder of barley for woredas of two zones namely North and South Gondar using survey data collected in annual agriculture sample survey. Using the weight for each woreda provided by CSA, small area estimates on total cultivated area of barley for each woreda of North and South Gondar were obtained and the result was compared with the 2001/02 Ethiopian Agricultural Sample Census data in order to assess the reliability of the estimate. The result of the study shows that the difference between the two estimates was found to be insignificant and suggests the usefulness of small area estimation technique and their application in various aspects to get estimates at woreda level (small area level).

2.4. Application of Small Area Estimation in Various Discipline

In poverty count study by National Research Council (1999) in the United States, basic area level model had been used to produce model-based county estimates of poor school-age children. The predictor variables used in the study were previous census and administrative records such as number of poor school-age children in last census, 3-year weighed average of poor school-age children, estimated population under age 18, number of child exemptions reported by families in poverty on tax returns, number of people receiving food stamps of counties and number of child exemptions on tax returns. Since the information was available at county level, basic area level model was applied for estimation.

Recently, Claudio et al. (2007) conducted a similar study to estimate income poverty measures in the Italian provinces using area-level model. In the study, they used head count ratio as the dependent variable and a number of indicators such as activity rate, unemployment rate, population density, resident population, crude birth rate, crude death rate, infant mortality rate, marriage rate and so on as independent variables. Stepwise procedure for the selection of covariates is used to determine some of the factors that contribute to the poverty level and to investigate in depth the territorial perspective in the poverty analysis at a provisional level.

In Ghana, Amoako et al. (2003) conducted a study which aims to derive district-level estimates of home deliveries and assess spatial variations between districts using data from the 2003 Ghana Demographic and Health Survey and auxiliary information from 2000 Population and Housing Census. The empirical best linear unbiased prediction (EBLUP) extension of the Fay-Harriot model for small area analysis is used to produce estimates for the 110 districts of Ghana. The small area estimates show significant clustering effect in delivery care uptake across districts that warrant policy and programmer attention. Valid measures of error are also used to access the validity of the estimates. They suggested that the importance of the analysis in supporting central

government allocation of health funds and local government and health practitioner's implementation, monitoring and evaluation of maternal health activities.

In Philippines, the 2000 Census of Population and Housing was conducted simultaneously with the Family Income and Expenditure Survey. This combined data set used to build models that can predict provincial (small area) poverty statistics which were eventually used to produce the EBLUP estimates (Albacea, 2003).

Torelli and Trevisani (2008) described the problem behind using auxiliary data. Small area estimation model depends crucially on the availability of auxiliary information. For larger geographical domains corresponding to administrative regions, it could be easier to find good auxiliary covariates, while this cannot be true when areas are small domains. In many application of small area estimation in Italy the census has been the major source of auxiliary information. Census data have many advantages since they allow to build appropriate information for any area level worth of consideration. Moreover, the census collects data on many aspects which are related to different variables. Unfortunately census data become outdated very quickly. A much more important source of auxiliary information are administrative archives. This information is often strictly related to phenomena under study and its explanatory power is very high compared to potential auxiliary variables collected in the survey. It is important to note however that in many cases, and especially for geographical domains that do not have administrative relevance, it could be the case that area level auxiliary information is not defined over the same geographical grid.

In most studies, after fitting the data with small area models especially unit level model it is found that the EBLUP estimates could not be design-consistent with the survey based estimate of larger domain. This means, the sum of EBLUP estimates of small areas within some definite larger area could not be same with the reliable design-based estimate of that larger area. In the study of per capita income for small places, Fay and Harriot (1979) obtained an estimate for small places. But, the sum of estimates at some county was found to be inconsistent with the county estimates. Thus, to insure

consistency with aggregated sample estimates, they adjusted EBLUP estimate to be (i) the total estimated income for all places equals the direct estimate at state level (ii) the total estimated income for all places in a county equals the direct estimate of total income.

Similarly, in the study of Canadian census under coverage Dick (1995) used basic area level model to estimate undercount in the decennial census of the United States and in Canadian census. In the study, it is found that the EBLUP estimates are inconsistent with design estimates. Then, the estimate is subjected to two-step ranking to insure consistency with the reliable direct estimate. The ranked EBLUP estimates were used as final estimates. Battese et al. (1988) also reported the adjusted EBLUP estimates as a final estimate in their study of county crop area estimation study.

CHAPTER THREE

MATERIALS AND METHODOLOGY

3.1. Source of Data

The source of the data on the area of cultivated land and production of Teff will be the raw data of 2007/08 Ethiopia Agricultural Sample Survey which is obtained from the Central Statistical Agency (CSA). The survey is part of the annual successive agriculture survey designed to provide estimates for basic agricultural variables such as land use, cultivated area and production/ yield for the following domains: national level, regional and zonal level for all 9 regions namely: Tigray, Affar, Amhara, Oromiya, Somali, Benishangul-Gumuz, Southern Nations, Nationalities and Peoples (SNNP), Gambela and Harari.

A stratified two-stage cluster sample design was implemented to select the sample. Thus, enumeration areas (EAs) were taken to be the primary sampling units (PSUs) and the secondary sampling units (SSUs) were agricultural households. EAs from each stratum were selected systematically using the probability proportional to size sampling technique; size being the number of agricultural households. The survey was intended to cover 44,200 agricultural households in 2,200 enumeration areas (EAs) and succeeded to cover 42,523 (96.21%) of agricultural households and 2,125 EAs (96.59%) throughout all regions. Specific to Tigray Region it was planned to cover 3300 agricultural households in 165 EAs and it was possible to cover 3299 of them was succeeded to cover in the survey.

In addition, data were used from the 2001/02 Ethiopia Agriculture Sample Enumeration, which was launched with the aims to provide woreda level direct survey estimates for different agriculture variables by CSA, and administrative record from Ministry of

Agriculture and Rural Development on area of cultivated land and production of Teff at woreda level.

3.2 Variables Included in the Study

a. The Response Variable

In this study, since the target is to improve the efficiency of the direct estimate by borrowing strength from the auxiliary variables through small area models, direct estimate of cultivated area and production is used as response variable. This direct survey estimate of 2007/08 at woreda level can be estimated using the survey weight provided by CSA for the annual agriculture sample survey (See Appendix I).

b. Explanatory Variables

The predictor variables which are taken as fixed effects and considered as the auxiliary variables in the small area estimation model are: -

- direct survey estimate of area cultivated and production at woreda level of 2001/02 Ethiopia Agriculture Sample Enumeration.
- administrative record data of cultivated area and production 2007/08 at woreda level from Ministry of Agriculture and Rural Development.

3.3. Methodology

In the development of small area estimation technique, there were challenges about how to formulate the methods and models which can give reliable or better estimates at small area level and how to test and practice them with appropriate empirical data. The terminology used in small area estimation can be confusing. This term small area is frequently used because in most applications the domains of interest have been relatively small geographic areas. However, it is the small number of sample observations and

resulting large variance of standard direct estimators, that are of concern, rather than the size of the population in the area or the size of the area itself.

There are different types of small area estimation techniques that can be used to estimate small area parameters through available supplementary information. Amongst them, direct domain estimation and indirect domain estimation are well known and widely used. The indirect domain estimation method is classified as model-assisted and model-based (small area model). The following explanations for direct and indirect estimators are given to distinguish them.

A **direct estimator** uses values of the variable of interest only from the time period of interest and only from units in the domain of interest.

An **indirect estimator** uses values of the variable of interest from a domain and/or from time period other than the domain and time period of interest.

However, each of these techniques has its own advantages and disadvantages. In general, availability of good auxiliary data and determination of suitable linking models are crucial to increase the precision of the estimate generated by the technique. To have a good picture of the small area estimation methods, it is better to present the methodology of each types of technique in detail.

3.3.1. Direct Domain Estimation Method: Design-Based Method

The direct domain estimation technique mainly focuses on the estimation of the small domain based on the sample design estimate. Thus, the estimate can be found using classical design-based method that are obtained by applying survey weights and complex survey design to the sample units in each small area. The method of estimation is also known as design-based approach. Sardnal and Hidroglou (1989) gave a description of this technique and its mathematical model in their study which is presented as follows.

Let us consider a population size U with N distinct elements identified through the labels $j = 1,2,\ldots,N$ and a sampling design to select sample size of S from U with probability of selection $p(s)$. Assume the characteristics of interest y, associated with element j,

(y_j) can be measured exactly by observing element j (thus, measurement errors are assumed to be absent). And also, consider design-weight, $w_j(s)$ which is, in most cases, a reciprocal of inclusion probability π_j, i.e., $\pi_j = \sum_{\{s: j \varepsilon s\}} p(s)$, $j = 1,2,\ldots,N$ where $\{s : j \varepsilon s\}$ denote summation over all samples s containing the element j and $w_j(s)$ may be interpreted as the number of elements in the population represented by the sample element.

So, in estimation of the population total Y of the characteristics of interest, we use the expansion estimator

$$\hat{Y} = \sum_{\{s: j \in s\}} w_j(s) y_j$$

Here, the choice of $w_j(s) = 1/\pi_j$ satisfies the unbiased condition and leads to the Horvity-Thompson (H-V) estimator (Cochran, 1977 and Kish, 1965)

Note that, the above expansion estimator is given considering the larger domain. But, we can also have an estimate for a small domain in a similar fashion. Consider q $(q = 1,2,\ldots,Q)$ small domains and U_q denotes the size of small domain (sub-population) with N_q distinct elements, Y_q is the total of characteristics of interest of small domain. Then, define

$$y_{qj} = \begin{cases} y_q, & j \in U_q \\ 0, & otherwise \end{cases}$$

Thus, in the population it will have

$$\sum_{\{j \in U\}} y_{qj} = \sum_{\{j \in U_q\}} y_q .$$

So, to obtain an estimate of Y_q, $\hat{Y}_q$ we use the expansion estimator

$$\hat{Y}_q = \sum_{\{j \in s\}} w_j(s) y_j = \sum_{\{j \in s_q\}} w_j(s) y_{qj}$$

where s_q denotes the sample of estimate belonging to U_q.

The direct domain estimation technique is simple and straightforward to apply. But, this technique cannot be applied in practice. The main problem in this technique lies on size of sample in the areas of interest is usually too small to obtain accurate and/or precise estimates. This problem becomes much worse for areas in which no sample at all was collected.

3.3.2. Indirect Domain Estimation: Model-Assisted Method

The model-assisted method is one of the indirect domain estimation techniques which is brought into the estimation process through an *implicit* model that provides a link to related areas through the use of supplementary information related to the variable. The technique involves estimators which include synthetic estimators, composite estimators and James-Stein (Shrinkage) estimators. Gonzalez (1973) pointed out that a reliable direct estimator for a large area, covering several small area, is used to derive an indirect estimator for a small area under the assumption that the small areas have the same characteristics as the large area. Such type of an estimator is known as synthetic estimator. Synthetic estimates formed in a sense of the estimates developed for the larger areas are to be scaled down to the smaller areas on the basis of certain model assumption like the relation between the larger area and small areas remain same for the characteristics under study as well as for the auxiliary variable. If this model assumption is true, synthetic estimation has its strength in borrowing information from larger groups for use in small domain areas. On the other hand if the model assumption is not true, the techniques of synthetic estimation gives biased estimates (Sardnal, 1984). Synthetic estimator is described as follows.

Consider a linear regression of the characteristics of interest, y on $X^T = (X_1, X_2, \cdots, X_p)$, a p-vector of auxiliary variables, $X_P^T = (x_{1p}, x_{2p}, \cdots, x_{Np})$. Assume y_j : j = 1, ..., N are independent and let v_j be the variance of y_j. Then the model postulates that $E(y_j) = X_j^T \beta$; where $X_j^T = (x_{j1}, x_{j2}, \cdots, x_{jp})$, the j^{th} unit of the auxiliary

variables and $\beta^T = (\beta_1, \beta_2, \cdots, \beta_p)$ is a (px1) vector of regression coefficient of the model. If all N points (y_j, x_j) are observed, with population size U, β can be estimated by weighted least squares estimates, B which is given by

$$B = \left(\sum_U X_j X_j^T / v_j\right)^{-1} \left(\sum_U X_j y_j / v_j\right)$$

However, in practice, y_j is observed for the sampled observation. Therefore, B in turn is estimated by the sampled weighted least squares, $\hat{B}$ which is given by:-

$$\hat{B} = \left(\sum_s w_j(s) X_j X_j^T / c_j\right)^{-1} \left(\sum_s w_j(s) X_j y_j / c_j\right)$$

S and $w_j(s)$ are the sample and design weight (as defined previously)

c_j is a specified constant (for instance if $c_j = x_j$, the estimator becomes the ratio synthetic estimator, or it may be estimated from the sample.)

Therefore, the regression synthetic estimator has a form:-

$$\hat{Y}_q = X_q^T \hat{B}$$

where $\hat{Y}_q$ is an estimate of Y_q, q^{th} small domain total.

$X_q^T = (X_{q1}, X_{q2}, \cdots, X_{qp})$ are p vector of total for q^{th} domain auxiliary variables.

Note that $\hat{B}$ is estimated from all samples in the larger domain considering the regression coefficient of each small domain is same as to the larger domain (Sardnal and Higroglou, 1989).

A special case of the above synthetic-regression estimator in the case of single auxiliary variable x is the synthetic-ratio estimator. It is obtained by letting $c_j = x_j$. For an auxiliary variable of small area, x_q with the synthetic-ratio estimator is given by:-

$$\hat{Y}_q = x_q \frac{\hat{Y}}{\hat{X}}$$

where $\hat{Y}_q$ - synthetic-ratio estimate of q^{th} small area

$\hat{Y}$ - an estimate of total of q^{th} for larger area

$\hat{X}$ - an estimate of total of an auxiliary variable, X for larger domain

Note that the mean square error will be small if the area-specific ratio $R_q = Y_q / X_q$ is close to the overall ratio $R = Y/X$.

Synthetic estimators have been very widely used. However, the bias has been a matter of concern and attempts have made to mitigate it through the application of an estimator called composite estimator.

3.3.3. General Linear Mixed Model

Several small area models may be regarded as a special case of a general linear mixed model. It is better here to discuss the general linear mixed model in order to better understand small area models. The general linear mixed model represents a class of fixed and random effects regression models for dependent variables where some of the regression parameters are population-specific, i.e., the same for all subjects whereas other parameters are subject-specific. For instance, in cluster design, subjects are observed nested within larger units, for example, students nested in schools, patients nested in hospitals, inhabitants nested in localities, workers nested in work-places and so on. In the general linear model, it is assumed that the systematic variability in the population values of response variable is explained by the variation in the values of predictor variable. This validates the assumption that the error term is "noise" and observations are independent of each other. However, in practice the observed sample residual do not look like noise. Often it contains significant between-subject variation. This implies that the model is mis-specified and there are missing covariates in the model, whose values vary from one area to another.

Whereas, in the general linear mixed models there is no assumption that all observations are independent of each other which makes these models different from the general linear model. In this case the covariates are considered as fixed effects and the subject-specific variation can be handled by random effects which are taking into account the subject-specific covariate information. Therefore, it is appropriate for the analysis of several types of correlated data structures.

Mixed models have a long history, but received special interest only in the last few decades. This is partly due to the heavy computational burden of estimation methods used with such models. Recent developments in computing hardware, software and estimation methods have, however, led to increased attention being paid to the use of mixed models for the data analysis. Specially, linear mixed models have a wide range of applications. In particular, their ability to predict a linear combination of fixed and random effects is one of the more attractive properties of such models. In the mixed models, the mean or total of the response variable can be expressed as a linear combination of fixed and random effects.

3.3.3.1. Linear Mixed Model Structure

The general linear mixed model can be written as:-

$$y = X\beta + Z\nu + e$$

Here y is an $n\times 1$ vector of sample observations of response variable, X is a known $n\times p$ matrix of fixed effects with unknown regression coefficient β and Z is a known $n\times h$ matrix of random effects, and ν and e are non observable random vectors which are normally and independently distributed with mean 0 and with covariance matrices G and R depending on some known variance parameters $\delta=(\delta_1,\delta_2,\cdots,\delta_q)^T$, respectively. Thus, the variance-covariance matrix of the response variable y is given by $\text{var}(y) = V = R + ZGZ^T$.

3.3.3.2. Parameters and Mean Square Error Estimation of the Model

In the general linear mixed model, the parameter of the model β and ν can be estimated using an estimator given by:-

$\tilde{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ which is the BLUE for β

$\tilde{\nu} = GZ^T V^{-1}(y - X\tilde{\beta})$.

But, in the general linear mixed model it is important to estimate the means or totals of the dependent variable. The means or totals of the model can be expressed as the linear combination of fixed and random effects. Thus, estimators of such parameters can be obtained in the classical frequentist-framework and they are known as Best Linear Unbiased Predictor (BLUP). Henderson (1975) developed the best unbiased prediction (BLUP) estimators of such parameters. In this case "best" stands for minimum mean square error among all linear unbiased predictors, "linear" means that the predictor is a linear combination of the response variable values and "unbiased" means that the expected value of the prediction error (predicted value of variable – actual value of variable) is zero. The BLUP estimator has become a powerful and widely used procedure for fitting mixed models.

The derivation of BLUP estimator of the mean, μ involves the linear combination of fixed and random effects and it is given by

$$\tilde{\mu} = l^T \tilde{\beta} + m^T \tilde{v}$$

where $\tilde{\beta}$ and $\tilde{v}$ can be estimated as given above and l and m are any specified vector of constants.

The mean square error (MSE) of the BLUP, defined as the expected value of its squared deviation from the true value, consists of two parts and it can express as:-

$$MSE(\tilde{\mu}) = g_1(\delta) + g_2(\delta)$$

where

$$g_1(\delta) = m^T(G - GZ^T V^{-1} ZG)m$$

$$g_2(\delta) = d^T(X^T V^{-1} X)^{-1} d \text{ with } d^T = l^T - m^T GZ^T V^{-1} X \text{ and}$$

$\delta = (\delta_1, \delta_2, \cdots, \delta_q)^T$ is a variance parameter in G and R.

BLUP estimators minimize the MSE among the class of linear unbiased estimators and do not depend on normality of the random effects. However, they depend on the variance-covariance of random effects. The method described in Henderson (1975)

assumes that the variances associated with random effects in the mixed model (the variance components) are known. In practice, such variance components are unknown and have to be estimated from the data. There are several methods for estimating variance components. Harville (1977) reviews maximum likelihood (ML) and restricted maximum likelihood (REML). Different researchers showed that substituting estimated values of variance components in the BLUP led to biased predictions. However, they used a two-stage estimator approach such as first estimating variance components, then using these to estimate and predict fixed parameters and random components. A further description on estimation of random effect has also given by Liard and Ware (1982).

The predictor obtained from the BLUP using this two-stage estimator approach, i.e., when unknown variance components are replaced by associated estimators is called the empirical best linear unbiased predictor (EBLUP) and it is described in Robinson (1990) in detail. A naïve mean square error (MSE) estimate of the EBLUP was suggested by Henderson (1975), whereas, this estimate underestimates the true MSE of the estimators. However, Kacker and Harville (1984) introduced the MSE of an estimator of total or mean based on an approximation to its true MSE under mixed models. Prasad and Rao (1990) put many of the estimators in a unified framework and used second-order Taylor approximation to derive an estimator of the MSE of the EBLUP under the assumption of normality of the mean. Thus, they included a third component in the MSE estimate of EBLUP which accounts for uncertainty due to estimation of the variance component of the random effects.

3.3.3.3. Block Diagonal Covariance Structure

A special case of the general linear mixed model which covers many small area models is a model with block diagonal covariance structure. For m small areas, the general linear mixed model can be decomposed into m sub models and can be written as:-

$$y_q = X_q\beta + Z_q v_q + e_q, \qquad q = 1, \cdots, m$$

Here y_q is an $n_q \times 1$ vector of sample observations of response variable, X_q is a known $n_q \times p$ matrix of fixed effects with unknown regression coefficient β and Z_q is known $n_q \times h$ matrix of random effects, and v_q and e_q are non observable random vectors which are normally and independently distributed with mean 0 and with covariance matrices G_q and R_q depending on some known variance parameters $\delta = (\delta_1, \delta_2, \cdots, \delta_q)^T$, respectively. Thus, the variance-covariance matrix of the response variable y_q is given by $\text{var}(y_q) = V_q = R_q + Z_q G_q Z_q^T$. Thus, the parameter of the model β and v_q can be estimated using an estimator given by:-

$$\tilde{\beta} = (\sum_q X_q^T V_q^{-1} X_q)^{-1} (\sum_q X_q^T V_q^{-1} y_q) \quad \text{which is the BLUE of } \beta$$

$$\tilde{v}_q = G_q Z_q^T V_q^{-1} (y_q - X_q \tilde{\beta})$$

Similarly, the BLUP is estimated as

$$\tilde{\mu}_q = l_q^T \tilde{\beta} + m_q^T \tilde{v}_q$$

where $\tilde{\beta}$ and $\tilde{v}_q$ can be estimated as given above and l_q and m_q are any specified vector of constants at small area and also the MSE estimate can express as:-

$$MSE(\tilde{\mu}_q) = g_{1q}(\delta) + g_{2q}(\delta)$$

with

$$g_{1q}(\delta) = m_q^T (G_q - G_q Z_q^T V_q^{-1} Z_q G_q) m_q$$

$$g_{2q}(\delta) = d_q^T (X_q^T V_q^{-1} X_q)^{-1} d_q$$

$$d_q^T = l_q^T - m_q^T G_q Z_q^T V_q^{-1} X_q \text{ and}$$

$\delta = (\delta_1, \delta_2, \cdots, \delta_q)^T$ is a variance parameter in G_q and R_q.

The model described above is known as linear mixed model with block diagonal covariance structure.

3.3.4. Indirect Domain Estimation: Small Area Model

Small area models are models which consider some explicit model that makes specific allowance for between area variation which are essentially mixed models and are used in specific situations based on data availability on the response variables of interest. The use of explicit models offers several advantages such as: (1) model diagnostics can be used to find suitable model(s) that fit the data well. Such model diagnostics include residual analysis to detect departures form the assumed model, selection of auxiliary variables for the model, and case-deletion diagnostics to detect influential observations, (2) area-specific measures of precision can be associated with each small area estimate, (3) linear mixed models as well as nonlinear models with random area effects can be entertained, (4) recent methodological developments of random effects models can be utilized to achieve accurate small area inferences.

Saei and Chambers (2003) discussed the different small area models in relation to mixed effect model. They explained the difference in the synthetic estimators and small area model estimators based on two main ideas. The first idea bases on the assumption that the inter-domain variability in the response variable can be explained entirely in terms of corresponding variability in the auxiliary information, leading to so-called fixed effect model, but the other idea require the assumption that "unexplained" domain specific variability remains even after accounting for the auxiliary information, leading to so-called mixed models incorporating domain specific random effects. Fixed effect models explain inter-domain variation in the response variable of interest entirely in terms of known factors. Such models have been the mainstay of statistical analysis. Estimates of small area characteristics based on fixed effect models are referred to as synthetic estimators, composite estimator and so on (Sardnal, 1984).

Small area models are mostly known as model-based indirect estimation method. The estimator in small area models is based on some auxiliary information that is incorporated into a model and used to calculate the estimated criterion of interest. There are two categories in small area model. These are a) area level models where information

on response variable is available only at the small area level; and b) unit level models where information on the response variable is available at the unit level. The unit level model mostly involves a one-folded nested error linear regression model in the estimation of the parameter of interest.

3.3.5. Area Level Model

The area level models are aggregate level models which relate the small area totals or means to area-specific auxiliary variables in order to estimate the variable of interest. The models are widely used in many practical applications of small area estimation and especially they are essential if unit (element) level data are not available.

3.3.5.1. The Model Structure

The derivation of area level model involves two components. These are: -

1. Suppose that direct survey estimate of the parameter θ_q, (which is a function of finite population total i.e., $\theta_q = g(Y_q)$ for some specified $g(.)$ which is assumed related to area-specific auxiliary data and for some small area totals, Y_q) based on the sampling design.

Thus, $\hat{\theta}_q$ can be expressed as follows:-

$$\hat{\theta}_q = \theta_q + e_q, \qquad q = 1,2,\ldots,Q$$

where $\hat{\theta}_q$ is a direct survey estimate of the q^{th} small domain of θ, θ_q.

e_q is a design based random error term which is assumed to be normally and independently distributed with mean zero and known variances ψ_q.

This model is called a sampling model and ψ_q is customarily assumed as design-based sampling variance.

2. A linking model: For area-specific auxiliary data X_q^T, θ_q can be related through a linear model as

$$\theta_q = X_q^T \beta + v_q, \qquad q = 1,2,\ldots,Q$$

where $X_q^T = (X_{q1}, X_{q2}, \cdots, X_{qp})$ is a transpose of $(px1)$ vector of the q^{th} small domain auxiliary variables.

$\beta^T = (\beta_1, \beta_2, \cdots, \beta_p)$ is a $(px1)$ vector of regression coefficient of the model.

v_q is a model-based random error term which is assumed to be normally and independently distributed across small areas with mean zero and known variances σ_v^2.

The model variance σ_v^2 is a measure of homogeneity of the areas after accounting for the covariates X_q. Combining these two models, the resultant mixed linear model is

$$\hat{\theta}_q = X_q^T \beta + v_q + e_q, \qquad q = 1,2,\ldots,Q.$$

From the model structure of area level model, it can be noted that it has a similar structure with block diagonal covariance structure model. Thus, area level model can be considered as special case of the general linear mixed model.

The corresponding notation between block diagonal covariance structure model and area level model is given as follows:-

$y_q = \hat{\theta}_q$, $X_q^T = X_q^T$ and

Z_q is the $q \times q$ identity matrix

$v_q = v_q, e_q = e_q$, $G_q = \sigma_v^2$, $R_q = \psi_q$ and $V_q = \psi_q + \sigma_v^2$

Further,

$l_q = X_q^T$ and m_q is the $q \times q$ identity matrix.

Using these notations, the estimation of parameters and their mean square error of the area level model are derived in the next section.

3.3.5.2. Parameters and Mean Square Error Estimation of the Model

In a model which doesn't involve random effect like the general linear regression model, the total of a response variable is estimated only from the auxiliary variable (fixed part). But, in the general linear mixed model the response variable is estimated by the linear combination of fixed and random component of the model.

Henderson (1975) developed the method called best linear unbiased prediction (BLUP) estimator, which can be used to predict the linear combination of fixed and random effects. Thus, the BLUP estimator for θ_q is given by:-

$$\tilde{\theta}_q = X_q^T \tilde{\beta} + \gamma_q (\hat{\theta}_q - X_q^T \tilde{\beta}),$$

where $\tilde{\beta} = [\sum_{i=1}^{Q} x_q x_q^T / (\psi_q + \sigma_v^2)]^{-1} [\sum_{i=1}^{Q} x_q \hat{\theta}_q / (\psi_q + \sigma_v^2)]$

$\gamma_q = \sigma_v^2 / (\sigma_v^2 + \psi_q)$ known as shrinkage factor and it is the ratio between model variance and total variance.

Note that $X_q^T \tilde{\beta}$ is an estimator for the fixed part, whereas $(\hat{\theta}_q - X_q^T \tilde{\beta})$ is an estimator for the random part.

Similarly, the BLUP of θ_q can also be re-written as:-

$$\tilde{\theta}_q = \gamma_q \hat{\theta}_q + (1 - \gamma_q) X_q^T \tilde{\beta}.$$

This is the weighted average of direct estimator $\hat{\theta}_q$ and regression synthetic estimator $X_q^T \tilde{\beta}$, where the weight $\gamma_q (0 < \gamma_q < 1)$ measures the uncertainty in modeling $\tilde{\theta}_q$, namely, the model variance σ_v^2 relative to the total variance $\sigma_v^2 + \psi_q$.

Prasad and Rao (1990) provided the measure of the mean square error of the BLUP estimator. It depends on the unknown variance parameter σ_v^2 and it is:

$$MSE[\tilde{\theta}_q(\sigma_v^2)] = g_{1q}(\sigma_v^2) + g_{2q}(\sigma_v^2)$$

with $g_{1q}(\sigma_v^2) = \sigma_v^2 z_q^2 \psi_q (\sigma_v^2 z_q^2 + \psi_q) = \gamma_q \psi_q$ and

$$g_{2q}(\sigma_v^2) = (1-\gamma_q)^2 X_q^T \left[\sum_q \left(\frac{X_q X_q^T}{\psi_q + \sigma_v^2} \right) \right] X_q$$

In practice, with the availability of the data $\{(\hat{\theta}_q, X_q), q = 1,2,\ldots,Q\}$ it is unrealistic to have a known variance of random effect, σ_v^2. To estimate the parameter of interest in the above model, it is necessary to replace the variance parameter σ_v^2 by a suitable estimator, i.e., $\hat{\sigma}_v^2$. Therefore, an estimator of θ_q^* with $\hat{\sigma}_v^2$ can be expressed as:-

$$\theta_q^* = X_q^T \hat{\beta} + \hat{\gamma}_q (\hat{\theta}_q - X_q^T \hat{\beta}),$$

$$= \hat{\gamma}_q \hat{\theta}_q + (1 - \hat{\gamma}_q) X_q^T \hat{\beta}$$

where $\hat{\theta}_q$ and X_q^T are as defined above and $\hat{\gamma}_q$ and $\hat{\beta}$ can be estimated by

$\hat{\gamma}_q = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \psi_q)$

$\hat{\beta} = (\sum_q \hat{\gamma}_q X_q X_q^T)^{-1} (\sum_q \hat{\gamma}_q X_q \hat{\theta}_q)$ is the weighted least square estimate of β obtained by regressing $\hat{\theta}_q$ on X_q using $\hat{\gamma}_q$ as a weights.

The above estimator of θ_q, θ_q^* is a two stage estimator since it involves first estimation of σ_v^2 then estimation of θ_q. θ_q^* is commonly known as Empirical Best Linear Unbiased Prediction (EBLUP). It may be noted that θ_q^* is a linear combination of direct estimate $\hat{\theta}_q$ and the model based regression synthetic estimate $X_q^T \hat{\beta}$, with weights inversely proportional to their respective variances. The EBLUP estimate can lead to large gains in efficiency over the direct estimate with variance ψ_q, when γ_q is small i.e. the model variance σ_v^2 is small relative to the sampling variance ψ_q. Therefore, choice of good auxiliary data to provide a good model fit is the key to successful application of the small area technique. Here, it can be noted that estimation of small area parameter involves the estimation of σ_v^2. The suitable estimator of this variance of can be estimated using the Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) methods.

The mean square error of EBLUP appears to be insensitive to the choice of the estimator $\hat{\sigma}_v^2$. Prasad and Rao (1990) put many of the estimators of small area mean in a unified framework and used second-order Taylor approximation to derive an estimator of the MSE of the EBLUP under an assumption of normality of small area means. Under normality of the random effects, the approximated form of the mean square error of EBLUP is obtained as:-

$$MSE\left[\tilde{\theta}_q(\sigma_v^2)\right] = g_{1q}(\sigma_v^2) + g_{2q}(\sigma_v^2) + g_{3q}(\sigma_v^2)$$

where σ_v^2 is estimated by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML); so an unbiased estimator of this mean square error has been computed using the following expression:

$$MSE[\tilde{\theta}_q(\hat{\sigma}_v^2)] = g_{1q}(\hat{\sigma}_v^2) - b_{\hat{\sigma}_v^2}(\hat{\sigma}_v^2)\nabla_{g_{1q}}(\hat{\sigma}_v^2) + g_{2q}(\hat{\sigma}_v^2) + 2g_{3q}(\hat{\sigma}_v^2)$$

where

$$g_{1q}(\hat{\sigma}_v^2) = \hat{\sigma}_v^2 z_q^2 \psi_q (\hat{\sigma}_v^2 z_q^2 + \psi_q) = \hat{\gamma}_q \psi_q$$

$$g_{2q}(\hat{\sigma}_v^2) = (1-\hat{\gamma}_q)^2 X_q^T \left[\sum_q \left(\frac{X_q X_q^T}{\psi_q + \hat{\sigma}_v^2}\right)\right]^{-1} X_q$$

$$g_{3q}(\hat{\sigma}_v^2) = \psi_q^2 (\psi_q + \hat{\sigma}_v^2)^{-3} \overline{V}(\hat{\sigma}_v^2) \quad \text{where} \quad \overline{V}(\hat{\sigma}_v^2) = 2m\left[\sum_q \frac{1}{(\psi_q + \hat{\sigma}_v^2)}\right]^{-2}$$

$$b_{\hat{\sigma}_v^2}(\hat{\sigma}_v^2) = \frac{2\left[m\sum_q (\psi_q + \hat{\sigma}_v^2)^{-2} - \left\{\sum_q (\psi_q + \hat{\sigma}_v^2)^{-1}\right\}^2\right]}{\left[\sum_q (\psi_q + \hat{\sigma}_v^2)^{-1}\right]^3} \quad \text{and}$$

$$\nabla_{g_{1q}}(\hat{\sigma}_v^2) = (1-\hat{\gamma}_q)^2$$

The derivation and detail explanation of the estimation of parameters and mean square error of EBLUP is well discussed in Rao(2003), Prasad and Rao (1990), Saei and Chambers (2003) and Robinson (1991).

In addition to EBLUP method, empirical Bayes (EB) and hierarchical Bayes (HB) estimation and inference methods have been applied to small area estimation. These

methods have been used quite extensively and effectively in recent years for small area estimation problems (Ghosh et al., 1998). Under the EB approach, Bayes estimation and inferential approaches are used in which posterior distributions are estimated from the data. Under the HB approach, unknown model parameters (including variance component) are treated as random, with values drawn from specified prior distribution. Posterior distributions for the small area characteristics of interest are then obtained by integrating over these priors, with inferences based on these posterior distributions. Yogendra et al. (2003) explain about the application of these methods under specific distributional assumptions. They also suggest that under normality assumptions EB is identical with EBLUP method.

3.3.6. Model Selection and Validation in Small Area Model

3.3.6.1. Model Selection

Verbeke and Molenberghs (2000) discussed the application of log likelihood ratio test as a classical statistical test for the comparison of linear mixed model with different mean and covariance structure. It is applied under the null hypothesis that the coefficients for all the terms in the fitted model, except the constant term, are zero. Log likelihood ratio test is given by –2 times the difference between the log-likelihood of fitted model and the null model. It then follows from classical likelihood theory under some regularity conditions, log likelihood ratio follows asymptotically, under the null hypothesis, a chi-squared distribution with degrees of freedom equal to the difference between the dimension parameter in the fitted model and the dimension of null model. Similarly, testing a reduced model against a full model is based on -2LL, where LL is the natural log of the maximized likelihood. It is based on the fact that $(-2LL)_{\text{full}} - (-2LL)_{\text{reduced}}$ is distributed approximately as chi-square with the number of d.f. equal to k, where k is the numbers of explanatory variables which is found in both models.

Verbeke and Molenberghs (2000) also discussed some other frequently used functions which are all leading to different discriminating rules, called information criteria. The main idea behind information criteria is to compare models based on their maximized log-likelihood value, but to penalize for the use of too many parameters. Among them, the Akaike Information Criteria (AIC) and Schwarz Bayessian Criteria (SBC) are well known and widely used in model selection. Generic function calculating the Akaike information criterion for one or several fitted model objects for which a log-likelihood value can be obtained, according to the formula -2*log-likelihood + k*npar, where npar represents the number of parameters in the fitted model, and k = 2 for the usual AIC. When comparing fitted models, the smallest the AIC has given a better fit.

3.3.6.2. Model Validation in Small Area Model

Model diagnostics and validation in small area models as well as linear mixed models has not been developed appropriately and it is very difficult to make a clear distinction between different fitted models. Research suggests that rather than a single measure of quality, there are a range of diagnostic measures available for assessing the performance and reliability of various small area modeled estimates. In reality not all diagnostic measures will point to a single best model or small area estimator. It is not at all uncommon for a small area estimator to perform well against several of the diagnostics but not so well against others. Statistical judgment is therefore required to choose the best small area estimator from a group of candidate small area estimators (ABS, 2006). In general, mainly three diagnostic procedures are discussed by many of the researchers on the area which are used to validate the reliability of the estimates generated from the estimation model (Brown et al., 2001). The diagnostic methods discussed are: (a) Relative standard error diagnostics (b) Model diagnostics and (c) Bias plot (the goodness of fit diagnostic).

a. Relative standard error (CV) diagnostics

The relative standard error (RSE) is one of the measures of reliability of an estimate. The RSE is obtained by dividing the standard error of the EBLUP estimate by the standard error of the direct estimate and expressed as a percentage.

RSE = [S.E. (EBLUP estimator)] / [S.E. (Direct Estimate] x 100%

Researchers suggested that estimates with large RSE's are considered unreliable (ONS, 2001).

b. Model Diagnostics

As with all statistical models, a number of assumptions are made about small area models. In most of these models the model error term are assumed to have a normal distribution with mean zero and constant variance. The model diagnostics are used to verify that the model assumptions are satisfied using residual analysis. If the model assumptions are satisfied the relationship between the area level residuals and the model estimates is expected to lie around the line residual=0. In the literature it is indicated that the model evaluation in small area model especially the model diagnostic using the plot of EBLUP estimate versus residual signifies whether the assumption of independence of observations, homogeneity of the variance of the residuals, the presence of outliers and the linearity between the dependent and the independent variables satisfied or not. Any systematic movement of the scatter points shows the violation in one or more of the assumption of the model.

That is, similar to the general linear model, diagnostics must also made on linearity of the auxiliary variable, normality of the error term and diagnostic on outliers. Linearity of the model can be diagnosed by inspecting the plot of the dependent variable and independent variable. Partial correlation coefficient is also used as measures of the linear relationship between two variables after adjusting for a group of variables. The Shapiro-Wilks test is used to test normality of residuals. The histogram of residual and pp-plot can also helpful to inspect the assumption of normality. Similarly, the residual analysis is applied in order to inspect the unusual data points, i.e., outliers.

c. Bias test

Because of the design of the sample survey and the weighting procedure, it is common to assume the response variable i.e., in this study direct estimate, is unbiased. Thus, to test if the model-based estimates are unbiased, plot the direct estimates (y-axis) against the

model-based, EBLUP estimates (x-axis). If the regression line is inconsistent with the Y=X line, there is evidence for bias in the model-based estimates. This test also used as goodness of fit diagnostics tests if the EBLUP estimates are close to the direct estimates.

3.3.5. Adjustment on EBLUP Estimate

After model formulation and validation, the final selected model is fitted to get the EBLUP estimates for the means or totals of small areas. Since the EBLUP estimate is the linear combination of direct estimate and model-based synthetic estimate, EBLUP estimates will be a better estimate than either of the two. In addition, the EBLUP estimates is superior over direct estimate because it uses an explicit model which incorporates area-specific random variation. Further, the EBLUP estimate would give an estimate with minimum variance.

But in the process of estimating mean or total of small area, the EBLUP estimate could not be consistent with reliable direct estimate of larger domain. This means, the sum of the EBLUP estimates of small areas in a definite larger area could not be equal to the design-based reliable estimate of that larger area. Rao (2003) remarked the EBLUP estimate must be subjected to adjustment in order to report the final estimate of small area level. There are various methods of adjustment of EBLUP estimate. The method of interpolation is the simplest and commonly used in the adjustment of EBLUP estimate and it is applied by many researchers (Fay and Harriot, 1979). This method involves first the EBLUP estimate of small areas of some larger domain or sub-domain aggregated and each small area EBLUP estimates are interpolated with the aggregated sum and the reliable direct estimate of the larger domain.

CHAPTER FOUR

STATISTICAL DATA ANALYSIS

4.1. Introduction

The purpose of this chapter is to give empirical analysis about woreda (small area) level estimates of cultivated area in hectare and production in quintal of Teff in Tigray Region in 2007/08 Meher Season using appropriate small area estimation techniques based on auxiliary variables: cultivated area and production from 2001/02 National Agricultural Sample Census Enumeration and 2007/08 Ministry of Agriculture reports.

4.1.1. Data Nature

In most studies of small area estimation, the data on auxiliary information could be acquired basically from sources such as previous censuses and administrative records. In this study two auxiliary variables are used from census and administrative data in order to improve the efficiency of direct estimate about small area level.

Census Data

The census data were acquired from the 2001/02 National Agricultural Sample Census Enumeration which was launched with the aim to provide wereda level direct survey estimates for different agricultural variables. The Enumeration was conducted by the Central Statistical Agency (CSA). Data on cultivated area in hectare and production in quintal of Teff in Tigray Region in 2001/02 is obtained from the Ethiopia Agricultural Sample Enumeration Report on Area and Production of Crops and Crop Utilization for Tigray Region.

Administrative data

The administrative data accessed from Ministry of Agriculture and Rural Development. The data were collected by woreda offices of the ministry on cultivated area and production on Teff in the 2007/08.

Direct estimate

The response variable in this study is the direct estimate of cultivated area/ production at

woreda level which is obtained from the 2007/08 National Agricultural Sample Survey. For this study, woreda level direct estimates on cultivated area and production were obtained using the survey weight available from Central Statistical Agency with direct domain estimation technique. (The estimation procedure for totals and standard error for direct estimate is presented in Appendix I.)

4.1.2. Data Problem

At the time of conducting the 2007/08 National Agricultural Sample Survey the Tigray Region was demarcated into 34 woreda. Data for response and auxiliary variable of cultivated area/ production were available for 33 woreda. Because for one woreda, Erob, there was no report on cultivated area and production of Teff. The data for the 33 woredas are presented in Appendix II.a & b.

4.2. Predicting Small Area Estimates of Cultivated Area Using Direct Domain Estimation

As it is indicated in the methodology part, direct domain estimate of small area can be obtained based on sample survey data. The estimate can be found using classical design-based estimators that are obtained by applying survey weights to the sample units taken from each woreda. The agriculture sample survey conducted by CSA was designed as a stratified two-stage cluster sample. Thus, enumeration areas (EAs) were taken as the primary sampling units (PSUs) and the secondary sampling units (SSUs) were agricultural households. EAs from each stratum were selected systematically using the probability proportional to size sampling technique; size being number of agricultural households. The survey covered 42,523 agricultural households and 2,125 EAs throughout all nine regions. Specific to Tigray Region the survey covered 3299 agricultural households and 165 EAs.

The estimate of cultivated area at woreda level, i.e., direct estimate and its variance can be found by applying the estimation technique used for stratified two-stage cluster sample

design (Cochran, 1977). Direct estimates, their variance and coefficient of variation (CV) of cultivated area in hectares of the 33 woredas in Tigray Region are given in Appendix II.a. The result of direct estimate has shown large sampling CV of the design-based direct estimates of cultivated area has a range of 74.49 of Atsbi Wenberta woreda and with the lowest value of 3.54 for Tahtay Adiyabo woreda. In the literature the conventionally accepted CV of an estimate must be below 30. As can be seen from the table the average CV of the direct estimates of woreda for Tigray Region is equal to 32.08 which is higher than the conventional remarked minimum CV.

4.3. Predicting Small Area Estimates of Cultivated Area Using Small Area Model

4.3.1. Preliminary Analysis

In order to assess the characteristics of the auxiliary variables and the response variable a preliminary analysis such as descriptive statistics, correlations and scatter plots and fitting the general linear model is useful. The preliminary analyses on the data of the 33 woreda are presented in Table 4.1.

Table 4.1: Descriptive statistics on cultivated area

Variable	No of Observation	Mean	St. Dev.	Min	Max
Census Estimate	33	4754.091	3724.523	132	16075
MOARD Data	33	3778.656	2796.489	127.42	9739.03
Direct Estimate	33	4870.347	3817.35	125	14665.04

The descriptive analysis shows high variation of area in hectare between woreda levels. For instance, as shown in Table 4.1 direct estimate has mean of 4870.347 hectare with standard deviation 3817.35. In addition to this, the scatter plot of direct estimates with census estimates and data from Ministry of Agriculture and Rural Development (MOARD) indicate approximate linear relationship. The value of partial correlations of direct estimate versus census estimate and MOARD data are 0.908 and 0.729,

respectively (see Appendix IV.a.). The result shows that a significant linear relationship between the dependent variable and independent variables.

Before incorporating the small area model which involves two error components such as area-specific random error and model error to the model, first it is better to fit the data using general linear model. Thus fitting the general linear model to the data can give us some idea about the inter-relationship between the response variable and the predictors. In addition, it can shed light about on the validity of the model assumption and goodness of fit of the small area model which is going to be fitted in next subsection.

The general linear model is given by:-

$$y_q = \beta_0 + \beta_1 x_{1q} + \beta_2 x_{2q} + e_q \quad q = 1,2,\ldots,33$$

where

y_q = 33x1 vector of observation of direct estimate of area cultivated in hectares of Teff for woreda in Tigray region.

$\beta^T = (\beta_0, \beta_1, \beta_2)$ is a 3x1 vector of unknown coefficient of auxiliary variables.

x_{1q} = 33x1 vector of observation of the 2001/02 census estimate of cultivated area in hectares of Teff for woreda in Tigray Region

x_{2q} = 33x1 vector of observation of the administrative data from MOARD of cultivated area in hectares of Teff for woreda in Tigray Region

e_q = is a random error term which is assumed to be independently distributed with mean zero and known variances σ_e^2

σ_e^2 = is a measure of variance across the error term.

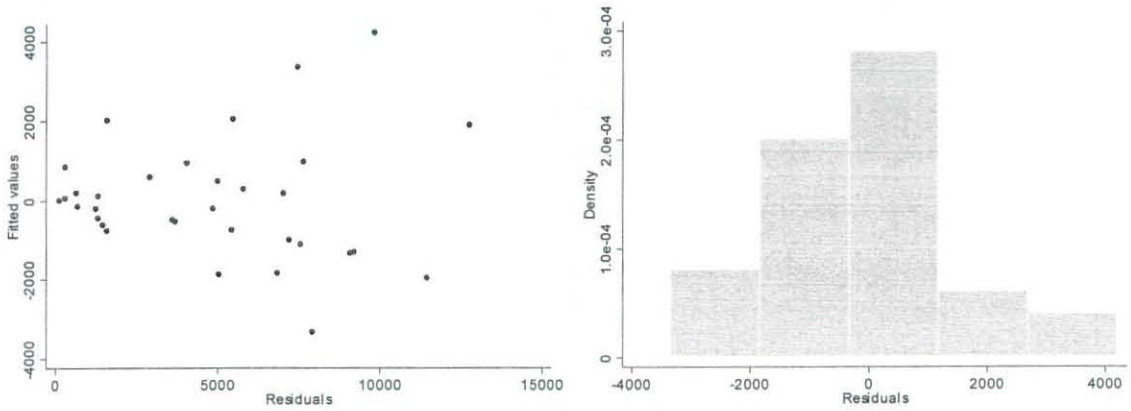
The empirical result of fitting the general linear model shows that the computed F-statistic is 79.06 with p-value 0.0000, adjusted R-Squared is 0.8299. In addition, variance inflation factor (VIF) of census estimate is 1.982 and VIF of MOARD data is 1.982 in which both VIF values show no problem of multicollinearity (Appendix IV.a.). All these imply that the fitted regression equation may have given a better fit.

Table 4.2: Estimates of variables in fitting general linear model of the 33 woreda

Variable	Coefficients	Standard Error	t	P>|t|	[95% Confidence Interval]	
Census Estimate	1.067822	0.1401255	7.62	0.000	0.7816479	1.353997
MOARD Data	0.183054	0.1052106	1.74	0.092	-0.0318144	0.397923
_Constant	-34.84367	482.475	-0.07	0.943	-1020.189	950.5017

However, from the plot of residual versus fitted values, we observe a funnel opening of the scatter points which shows the problem of hetroscedasticity and also the negatively skewed histogram shows the failure of the assumption of normality of error term (See Appendix IV.a for full output of general linear model analysis).

Figure 4.1: Plot of residual versus fitted values and histogram of residual of the model



In the context of regression analysis, these problems arise because of various reasons. The problems may be solved by taking measures such as applying proper transformation or incorporating woreda specific random factor in the model which will lead the above model to mixed model representation.

4.3.2. Small Area Models

One of the breakthroughs in small area estimation technique is the use of explicit models such as small area models which accounts for the area-specific random factor (between area random variations) in addition to fixed factors (auxiliary information). As it was

indicated, there are two types of small area models: area level models and unit level models. The choice of appropriate model is determined by the condition of available data on the covariates. The unit level model assumes that the data are available at unit (individual) level. Whereas, if the data on the covariates available in aggregate form at small area, then area level model can only be applicable to fit the data. Since the covariates considered in this study, i.e., the data from census and data from MOARD are available at woreda level, *area level model* is considered as an appropriate choice.

4.3.3. Area Level Model

The area level model which is going to be fitted for the area in hectares of Teff in Tigray Region is presented as:-

$$y_q = \beta_0 + \beta_1 x_{1q} + \beta_2 x_{2q} + v_q + e_q \qquad q = 1,2,\text{K},33$$

Description of model parameters and assumptions is:-

y_q = 33x1 vector of observations of direct estimate of area cultivated in hectares of Teff for woreda in Tigray Region.

$\beta = (\beta_0, \beta_1, \beta_2)$ is a vector of unknown coefficient of auxiliary variables.

x_{1q} = 33x1 vector of observations of the 2001/02 census estimate of cultivated area in hectares of Teff for woreda in Tigray Region.

x_{2q} = 33x1 vector of observation of the administrative data from MOARD of cultivated area in hectares of Teff for woreda in Tigray Region

v_q = is a model-based random error term which is assumed to be normally and independently distributed across small areas with mean zero and known variance σ_v^2.

e_q = is a design-based random error term which is assumed to be normally and independently distributed with mean zero and known variance ψ_q.

σ_v^2 = is a measure of homogeneity of small areas after accounting for the covariates.

ψ_q = is a variance of the error term e_q and customarily assumed as design-based sampling variance.

4.3.4. Model Formulation

In order to make a model formulation a null model (Model 0) was initially fitted. Then, three models were used to check for the explanatory power of the covariates. Model 1 accounted for effect of the census estimate on the response variable. Model 2 accounted for effect of the MOARD data on the response variable. Finally, Model 3 is fitted in order access the effect of both covariates on the response variable. Below is an illustration of the model building process.

Model 0: $X_0 = 1$

Model 1: $X_0 = 1$, X_1 = census

Model 2: $X_0 = 1$, X_2 = MOARD

Model 3: $X_0 = 1$, X_1 = census and X_2 = MOARD

Nowadays, small area models are fitted using standard software. For this particular study the R package version 2.9.0 is used. The method and application used to fit the area level model in the R package is prepared using the training material of Rubio (2008). The specific procedure used to fit the small area models in this study using R package is presented in Appendix III.

In the estimation of model parameters, the method known as Restricted Maximum Likelihood (REML) estimation is applied. Restricted Maximum likelihood estimation is a method of estimation which is based on the likelihood principle which leads to useful properties such as consistency, asymptotic normality and efficiency. Verbeke and Molenberghs (2000) show that REML can be used in estimating the parameters and their variance in linear mixed model analysis and usually REML follows the Newton-Raphson procedures to estimate all parameters in the model. Similarly, for the estimation of EBLUP in the area level model and its mean square error, the REML method is applied. As it was indicated in Chapter 3, log likelihood ratio and Akaike Information Criteria (AIC) can be used for model selection. Based on the log likelihood ratio test, the

performance of the models was checked. Table 4.3 shows the log likelihood ratio and chi-square value of Model 0–3 and Table 4.4 presents estimates of regression coefficients.

Table 4.3: The value of the log likelihood and log likelihood ratio test and Akaike Information Criteria of 33 woredas

	Log Llikelihood (LL)	df	Log Likelihood Ratio Test (-2LL)	Akaike Information Criteria (AIC)
Model 0	-7365.991	-	-	14735.980
Model 1	-7365.510	1	-0.962	14737.020
Model 2	-7365.555	1	-0.872	14737.110
Model 3	-7370.976	2	9.970	14736.810

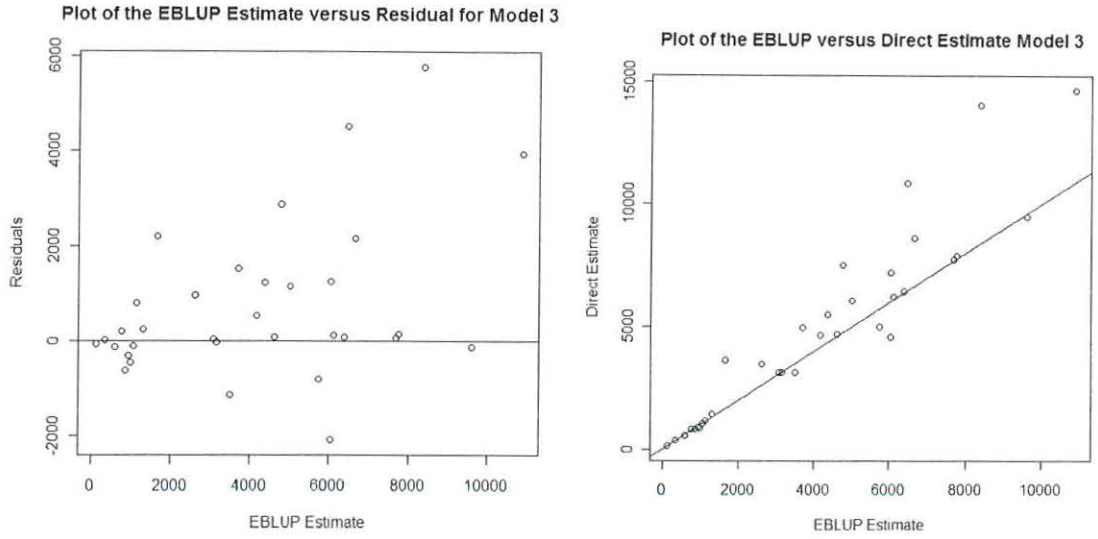
Table 4.4: Estimates of coefficient of independent variables in the area level Model fitted for the 33 woreda: Models 0-3

Variable	d.f.	Coefficients	Standard Error
Model 3			
Census Estimate	30	0.88	0.10
MOARD Data	30	0.16	0.08
_Constant	30	58.38	223.07
Model 2			
MOARD Data	31	0.12	0.16
_Constant	31	3565.53	960.91
Model 1			
Census Estimate	31	-0.11	0.21
_Constant	31	4549.21	1015.83
Model 0			
_Constant	30	4108.15	574.56

In order to get theoretically meaningful and statistically reliable estimates, each model must be properly validated. Thus, the model diagnostic using the plot of EBLUP estimate versus residuals and the plot of EBLUP estimate versus direct estimate of the fitted model indicates some sort of failure in model assumption (See Figure 4.2 below).

In both plots we observe some outlier observations. This is also happen in the plots of the other three models (See Appendix IV.b). To overcome the problem with outliers, Christensen (1996) suggested that variable selection and the elimination of outliers would lead to an improved model. Using residual diagnosis the outliers are inspected. Then, the problem would be illuminated by deleting these outliers from the observation.

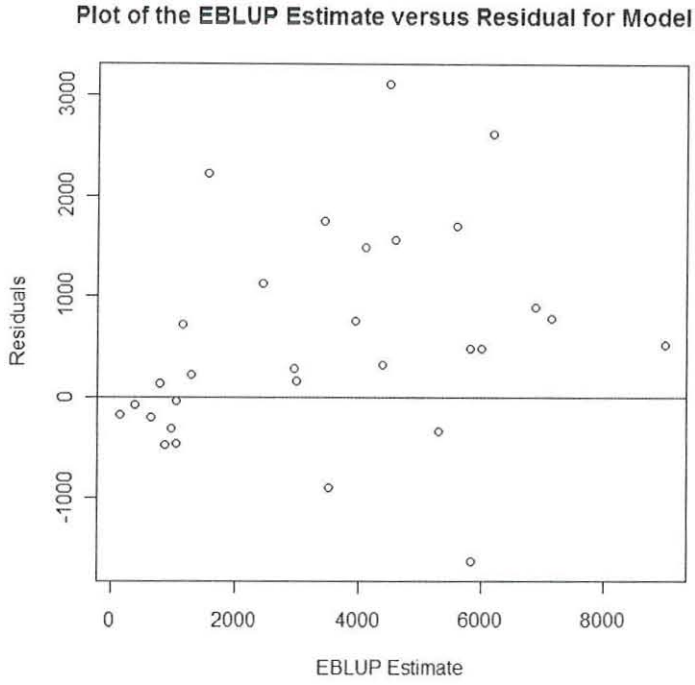
Figure 4.2: Plot of EBLUP estimate versus residuals and the EBLUP estimate versus direct estimate for Model 3



4.3.5. Fitting Area Level Model for Cultivated Area after Deleting Outliers Observations

By inspecting the residual of the fitted area level model, the observations of three woreda are deleted and analysis is made on the remaining 30 woreda in the region. The woredas whose observations were deleted are Medebay Zana, Worie Lehe and Rayaazebo woreda. A similar model formation, validation and selection is made on the data of the rest 30 woreda. To verify whether the model assumption are satisfied or not, the relationship between the EBLUP estimate and residuals is useful. Under a good fitted model scatter points of the plot of EBLUP estimate versus residuals is expected to lie around the line residual=0. Figure 4.3 below shows the plot of the fitted model for Model 3. From the plot in Figure 4.3, it can see that most of the points lie just below and above the line of residual is equal to zero.

Figure 4.3: Plots of the EBLUP estimate versus residual of the fitted Model 3



The result of log likelihood and information criteria of the four models is presented in Table 4.5. The log likelihood ratio test result shows fitting the Model 1 and Model 3 has given a better model. In order to determine the better model we compare the log likelihood ratio of the full model versus reduced model.

Table 4.5: The value of the log likelihood and log likelihood ratio test and Akaike Information Criteria of 30 woredas

	Log Llikelihood (LL)	df	Log Likelihood Ratio Test (-2LL)	Akaike Information Criteria (AIC)
Model 0	-5977.467	-	-	11958.93
Model 1	-5983.413	1	11.892	11972.83
Model 2	-5978.515	1	2.096	11963.03
Model 3	-5981.537	2	8.14	11971.07

*At 5 percent, tabulated value of chi-square is 3.84 and 5.99 with 1and 2 d.f., respectively.

Thus, comparing Model 3 (full model) with the Model 1 (reduced model), –2 times log likelihood ratio of the full minus the reduced is -3.752 which is insignificant for chi-square at 0.05 with 1 d.f. But using the information criteria, the AIC of Model 3 has

smallest value. Assessment on the estimated coefficients of the models shows that (See Table 4.6 below), the coefficient of census estimate is significant at 0.05 level.

Table 4.6: Estimates of variables in the area level model fitted for the 30 woredas: Model 0-3

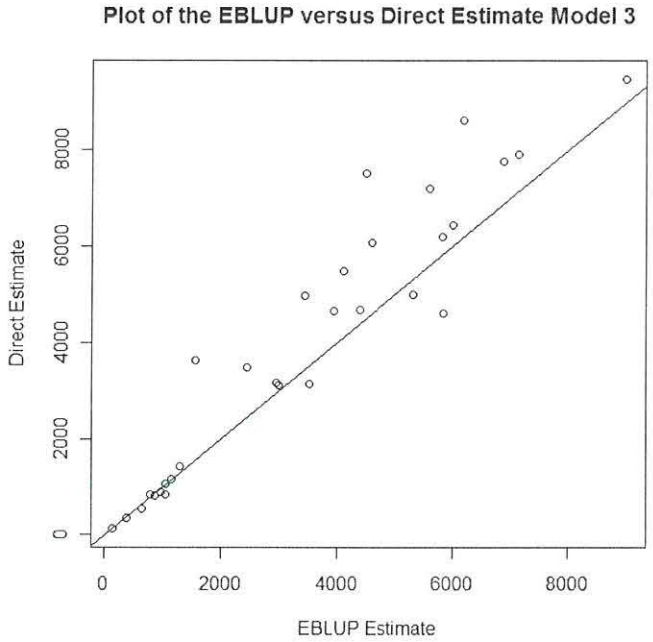
Variable	Coefficients	Standard Error	t	Significance At 0.05 level
Model 3				
Census Estimate	0.83	0.09	8.96	Signf.
MOARD Data	0.12	0.07	1.62	Not signf.
_Constant	177.88	199.50	0.89	Not signf.
Model 2				
MOARD Data	0.62	0.12	5.10	Signf.
_Constant	933.46	524.46	1.78	Not signf.
Model 1				
Census Estimate	0.91	0.08	12.12	Signf.
_Constant	323.89	168.45	1.92	Not signf.
Model 0				
_Constant	3372.39	478.18	7.05	Signf.

*At 5 percent, t-value is given by 2.0518, 2.0484 and 2.0452 with 27, 28 and 29 d.f., respectively.

Furthermore, validation of the selected model is made. Validation of Model 3 involves all diagnostic procedures discussed in the methodology part. It has mean relative standard error (RSE) of 56.41. Even though the RSE with value less than 100 shows that the fitted small area model gives a better estimate, the value is far from the RSE recommended in the literature, in which the RSE must be as low as 20 if good auxiliary variables are available. However, Model 3with mean RSE of 56.41 seems a reasonable fit using the census data as auxiliary variable. The RSE for all woreda are given in Table 4.7.

In addition to these, normality diagnostic is also performed on the residual. The statistic W from the Shapiro-Wilks test (W=0.9689) and its corresponding p-value (0.5108) indicate the normality of the residuals. Therefore, Model 3: using the auxiliary variable from census and MOARD data is taken as a *final model* in order to make estimation on cultivated area at small area level. (The full diagnostic result of selected model, Model 3 is given in Appendix IV.d).

Figure 4.4: Plots of the EBLUP estimate versus direct estimate of the fitted model for Model 3



Bias plot is used to test whether the EBLUP estimates are unbiased or not. If the regression line is inconsistent with the Y = X line, there is evidence for bias in the model-based estimates. This test is used as a goodness of fit diagnostics test. The fitted model is good if the EBLUP estimates are close to the direct estimates. The plot of EBLUP estimate versus direct estimate and linear regression for Model 3 is made using the R package. Figure 4.4 shows the plot of the EBLUP estimate versus direct estimate for Model 3. From the plots in Figure 4.4, even if some points are a little further apart, it can see that most of the points lie somehow around the line of Y=X. This is verified by the linear regression fit of direct estimate over EBLUP estimate which gives a slope of 1.11067.

4.3.6. Estimates of Final Small Area Model for Cultivated Area

In this section, results of the final model estimates of cultivated area in hectares of Teff in Tigray Region are presented in Table 4.7. The coefficient of variation (CV) of the design-based direct estimates (See Appendix II.a) is high 74.49 in Astbi Wonberta woreda and

with the lowest value 3.54 of Tahtay Adiyabo woreda. The average CV of the direct estimates of 30 woreda is equal to 33.02. From Table 4.7, the coefficient of variation (CV) of the EBLUP estimates using area level model is high 62.77 of Endamehoni woreda and with the lowest value 3.56 of Tahtay Adiyabo woreda. The average value of the CV of EBLUP estimates 30 woreda is equal to 19.30. From the comparison of the CV between the EBLUP estimate and direct estimate, we observe the gain in efficiency of using small area model with a highest gain (that fall at most) in Laelay Adyabo with 34.82 reduction in the CV and minimum gain in Tsegede woreda with only 0.13. Tahtay Adiyabo woreda has a negative value with -0.02 which shows the direct estimate gave a better estimate than the EBLUP estimate to this woreda (See Table 4.7).

The average of the CV of direct estimate is the 30 woreda is 33.02 but the average of the CV of EBLUP estimate of these 30 woreda is 19.30. This shows that on average the EBLUP estimate gave 13.72 gains in efficiency on the CV. Thus, it can be suggested that small area estimation technique has an advantage in improving the efficiency of the woreda (small area) level estimates.

Table 4.7: The EBLUP estimates of cultivated area in hectares of the 30 Woreda in Tigray Region and their MSE, CV and RSE using area level model

Sr. No	Zone Name	Woreda Name	EBLUP Estimate	MSE of EBLUP Estimate	CV of EBLUP Estimate	Relative Standard Error	Gain in Efficiency
1	North	Tahtay Adiyabo	1144.41	1662.75	3.56	99.99	-0.02
2	Western	Laelay Adiyabo	5319.46	313537.72	10.53	24.73	34.82
3		Tahtay Koraro	4581.72	298247.55	11.92	29.81	18.28
4		Asegede Tsimbela	4483.93	267654.36	11.54	21.45	20.57
5		Tselemti	5815.72	281792.03	9.13	58.49	5.49
6	Central	Mereb Lehe	3435.29	208682.53	13.30	39.19	10.14
7		Ahiferom	6008.72	315472.59	9.35	38.93	13.01
8		Adwa	3937.18	232867.51	12.26	26.02	27.60
9		Laelay Maychew	6180.68	309603.71	9.00	35.22	9.33
10		Tahtay Maychew	4106.22	227737.27	11.62	32.18	15.46
11		Naeder Adet	3517.78	100983.70	9.03	89.59	2.23
12		Kola Temben	7133.63	433791.91	9.23	19.42	33.66
13		Degua Temben	2998.19	182859.75	14.26	56.18	10.12
14		Tanqu Abergele	5847.18	258848.38	8.70	66.04	8.06
15	Eastern	Gulumahada	130.52	5973.15	59.21	99.93	2.66
16		Saesi Tsaedamba	1042.87	93768.10	29.36	85.88	12.54

17		Ganta Afeshum	780.86	98349.09	40.16	85.61	3.57
18		Hawzen	1561.18	196099.63	28.37	29.07	13.65
19		Wukro (kelete awlalo)	2451.83	207655.44	18.59	32.06	22.12
20		Atsbi Wenberta	644.29	113115.27	52.20	82.39	22.29
21	Southern	Seharte Samre	8976.61	596836.54	8.61	25.32	23.56
22		Enderta	4378.65	266830.97	11.80	23.72	34.68
23		Hintalo Wajirat	5607.58	291847.72	9.63	31.79	13.95
24		Ambalage	975.02	62163.02	25.57	92.64	4.83
25		Endamehoni	375.66	55596.78	62.77	94.71	7.36
26		Alamata	6868.48	856956.81	13.48	29.60	26.89
27		Wofla	1299.38	144632.39	29.27	69.80	8.68
28	Western	Kafta humera	871.07	23617.31	17.64	100.57	1.07
29		Welkait	2956.21	227787.65	16.14	72.74	4.67
30		Tsegede	1054.91	18418.35	12.87	99.35	0.13
	Total		104485.23	6693389.99	-	-	-
	Mean		3482.84	223113.00	19.30	56.41	13.71

4.4. Predicting Small Area Estimates of Production Using Direct Domain Estimation

A similar analysis is also made in the small area estimation of production in quintal of Teff for Tigray Region in 2007/08. The estimate of production at woreda level, i.e., direct estimate and its variance can be found by applying the estimation technique used for stratified two-stage cluster sample design in the sampling methodology. The full sampling methodology used in the estimation of a direct estimate of production in quintal of the 33 woredas in Tigray Region is given in Appendix I. The result of direct estimate has shown large sampling errors. The coefficient of variation (CV) of the design-based direct estimates of production of Teff has a range of 98.73 of Atsbi Wenberta woreda and with the lowest value of 8.14 Neader Adet woreda. In the literature, it is remarked that the conventionally accepted CV to report an estimate has to be below 30. The average CV of the direct estimates woredas for Tigray Region is equal to 42.99 which is much higher than the conventional minimum CV.

4.5. Predicting Small Area Estimates of Production Using Small Area Model

4.5.1. Preliminary Analysis

The preliminary analyses on the production data of the 33 woredas are presented as follows. The descriptive analysis shows that high variation of production between woreda levels. For instance, as shown in Table 4.8 the direct estimate has a mean of 69170.18 with standard deviation 69210.53.

Table 4.8: Descriptive statistics of the variable used in the analysis

Variable	No of Observation	Mean	St. Dev.	Min	Max
Census Estimate	33	31165.64	26426.75	896.45	128570.4
MOARD Data	33	52694.61	51074.77	883.35	226512.0
Direct Estimate	33	69170.18	69210.53	1008.42	307174.5

In addition to this, the value of partial correlation between the dependent variable and each of the independent variables is calculated. Thus, the value of partial correlation of direct estimate versus census estimate and MOARD data is 0.5093 and -0.1650, respectively. The scatter plot of direct estimate versus census estimate and MOARD data suggest an approximate linear relationship between the direct estimate versus census estimate and MOARD data.

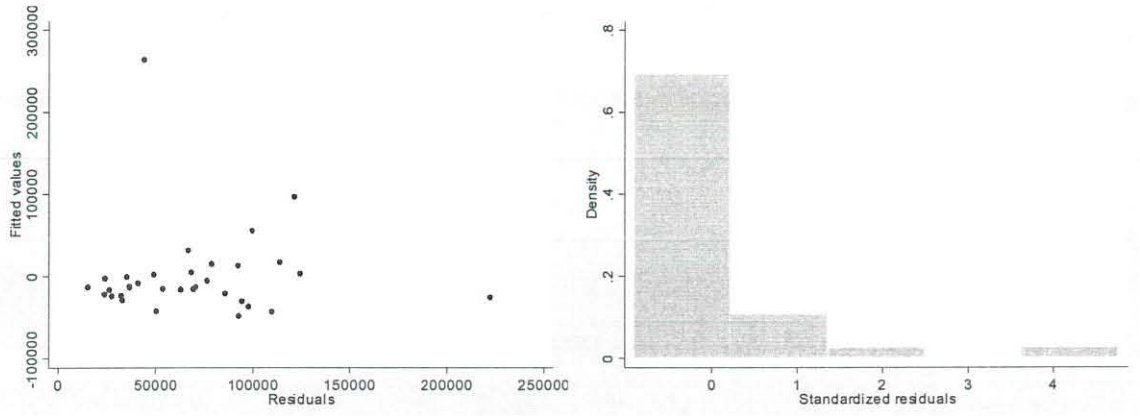
Before incorporating the small area model, it is better to fit the data using the general linear model. The empirical results of fitting general linear model show that the computed F-statistic is 8.64 with p-value 0.0011 adjusted R-squared is 0.3231. In addition, variance inflation factor (VIF) of census estimate is 2.578 and VIF of MOARD data is 2.578 in which both VIF values show no problem of multicollinearity (Appendix V.a.).

Table 4.9: Estimates of variables in fitting general linear model of the 33 woreda

Variable	Coefficients	Standard Error	t	P>|t|
Census Estimate	1.982554	0.61159	3.24	0.003
MOARD Data	-0.28989	0.316445	-0.92	0.367
_Constant	22658.26	15565.39	1.46	0.156

The plot of residuals versus fitted values in Figure 4.5 shows, except some outlier observations, that the general linear model seems to provide good fit. But the histogram shows assumption of normality of error term is not satisfied (See Appendix V.a, for full output of general linear model analysis). In the context of regression analysis, these problems arise for various reasons. Similar to the cultivated area model the problems may be solved by incorporating woreda specific random factor in the model; this will lead to a mixed model.

Figure 4.5: Plot of residual versus fitted values and histogram of residual of the model



4.5.2. Area Level Model Formulation

Similar to the cultivated area analysis, the appropriate small area model for production data is also area level model. Four models are formulated in the modeling process of area level model.

Model 0: $X_0 = 1$

Model 1: $X_0 = 1$, $X_1 =$ census

Model 2: $X_0 = 1$, $X_2 =$ MOARD

Model 3: $X_0 = 1$, $X_1 =$ census and $X_2 =$ MOARD

The log likelihood ratio test and AIC can be used for model selection. Based on the log likelihood ratio test, the performance of the models was checked. Table 4.10 shows log

likelihood ratio and AIC for Models 0–3. The regression coefficient estimates are also presented in Table 4.1.

Table 4.10: The value of the log likelihood and log likelihood ratio test and Akaike Information Criteria of 33 woredas

	Log Llikelihood (LL)	df	Log Likelihood Ratio Test (-2LL)	Akaike Information Criteria (AIC)
Model 0	-10326.78	-	-	20657.56
Model 1	-10330.70	1	7.840	20667.40
Model 2	-10326.28	1	-1.000	20658.57
Model 3	-10330.40	2	7.240	20668.80

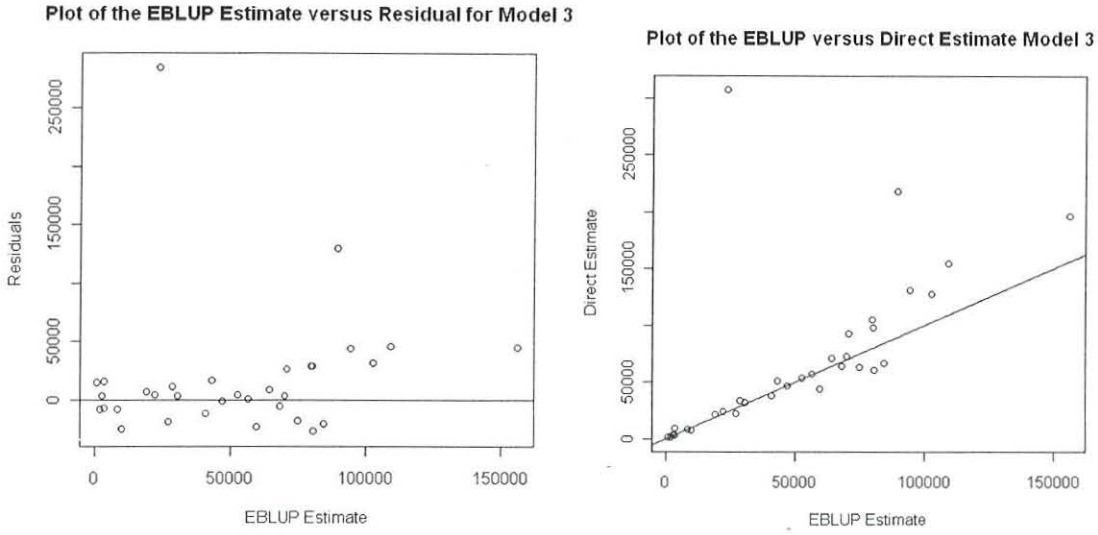
In order to get theoretically meaningful and statistically reliable estimate appropriate measures or adjustment must be taken/ made for each model to be properly validated.

Table 4.11: Estimates of coefficient of independent variables in the area level Model fitted for the 33 woredas: Model 0-3

Variable	d.f.	Coefficients	Standard Error
Model 3			
Census Estimate	30	1.811	0.200
MOARD Data	30	-0.049	0.100
_Constant	30	-219.533	3685.966
Model 2			
MOARD Data	31	0.712	0.141
_Constant	31	13313.525	7992.562
Model 1			
Census Estimate	31	1.751	0.152
_Constant	31	-927.063	3422.187
Model 0			
_Constant	30	47836.673	7158.004

The model diagnostic using the plot of EBLUP estimate versus residuals and the plot of EBLUP estimate versus direct estimate of the fitted model indicate some sort of failure in model assumption. (See Figure 4.6 below)

Figure 4.6: Plot of EBLUP estimate versus residuals and the EBLUP estimate versus direct estimate for Model 3



In both of the above plots, we observe some outlier observations. This also occurs in the plots of the other three models (See Appendix V.b). To overcome the problem with outliers, we deleted the outliers after residual diagnostic.

4.5.3. Fitting Area Level Model for Production after Deleting Outliers Observations

By inspecting the residual of the fitted area level model, the observations from five woreda are removed and analysis will be made on the remaining 28 woredas in the region. Those woredas are Atsbi Wonberta, Asgede Tsimblela, Seharte Samre, Rayaazebo woreda and Alamata. A similar model formation, validation and selection are made based on the data of the 28 woredas. The result of log likelihood and information criteria of the four models is presented on Table 4.12. The log likelihood ratio test result shows that Model 1 and Model 3 are better. But from the AIC, model 1 has smallest value.

Table 4.12: The value of the log likelihood and log likelihood ratio test and Akaike Information Criteria of 28 woredas

	Log Llikelihood (LL)	df	Log Likelihood Ratio Test (-2LL)	Akaike Information Criteria (AIC)
Model 0	-7179.453	-	-	14362.91
Model 1	-7183.897	1	8.888	14373.79
Model 2	-7179.321	1	-0.264	14364.64
Model 3	-7183.576	2	8.246	14375.15

*At 5 percent, tabulated value of chi-square is 3.84 and 5.99 with 1and 2 d.f., respectively.

Thus, it can be concluded that Model 1 is the superior. Assessment of the estimated coefficients of the model shows the coefficient of census is significant at 0.05 level (See Table 4.13 below) but not the coefficient of MOARD data.

Table 4.13: Estimates of coefficients in the area level model fitted for the 28 woredas: Model 0-3

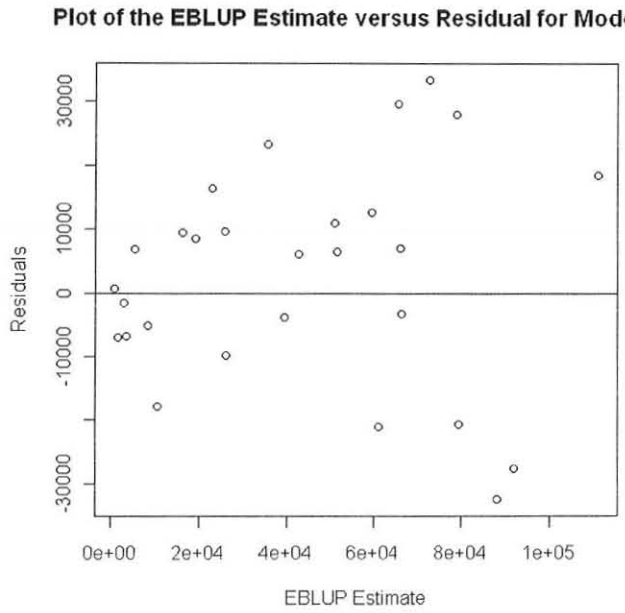
Variable	df	Coefficients	Standard Error	t	Significance At 0.05 level
Model 3					
census Estimate	25	1.796	0.221	8.128	Signf.
MOARD Data	25	-0.035	0.102	-0.349	Not Signf.
Constant	25	-652.021	3920.712	-0.166	Not Signf.
Model 2					
MOARD Data	26	0.580	0.158	3.677	Signf.
Constant	26	15128.237	7873.138	1.922	Not signf.
Model 1					
census Estimate	26	1.756	0.177	9.896	Signf.
Constant	26	-1237.056	3624.322	-0.341	Not Signf.
Model 0					
Constant	27	39449.067	6025.249	6.547	Signf.

*At 5 percent, t-value is given by 2.060, 2.056 and 2.052 with 25, 26 and 27 d.f., respectively.

Furthermore, validation of the selected model is made. Validation of Model 1 involves all diagnostic procedures discussed in the methodology part. It has a mean relative standard error (RSE) of 66.91. The RSE for all woredas is given in Table 4.14. The RSE with value less than 100 show that the fitted area level model gives a better estimate.

To verify whether the model assumptions are satisfied or not, it is useful to look at the relationship between the EBLUP estimate and residuals. For a good fitted model the scatter points of the plot of EBLUP estimate versus residuals is expected to lie around the line residual=0. Figure 4.7 below shows the plot of the fitted model for Model 1. From the plot in Figure 4.7 it can see that most of the points lie just below and above the line of residual=0.

Figure 4.7: Plots of the EBLUP estimate versus residual of the fitted Model 1

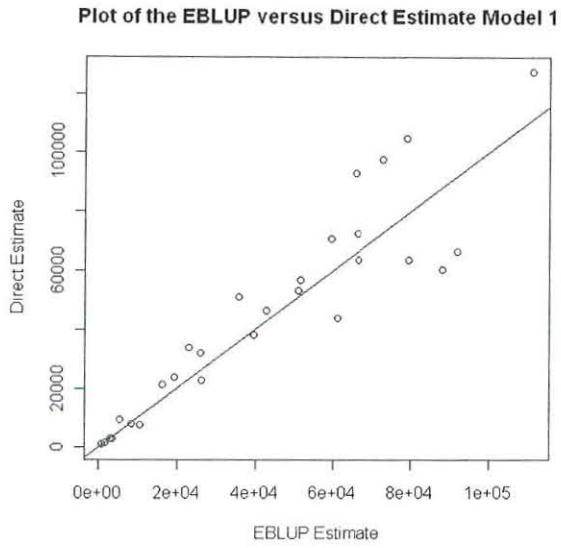


Bias plot is used to test whether the EBLUP estimates are unbiased or not. If the regression line is inconsistent with the Y = X line, there is evidence for bias in the model-based estimates. This test uses as goodness of fit diagnostics tests if the EBLUP estimates are close to the direct estimates. Figure 4.8 shows the plot of the EBLUP estimate versus direct estimate for Model 1.

Even if some points are a little further apart, it can be seen that most of the points lie around the line of Y=X. In addition to this, normality diagnostic is also performed on the residual. The statistic W from the Shapiro-Wilks test (W=0.9786) and its corresponding p-value (0.8169) indicate the normality of the residuals. Therefore, Model 1: using the auxiliary variable from census is chosen as a *final model* in order to make estimation on

cultivated area at small area level. (The full diagnostic result of selected model, Model 3 is given as Appendix V.d)

Figure 4.8: Plots of the EBLUP estimate versus direct estimate of the fitted model for Model 1



4.5.4. Estimates of Final Small Area Model for Production

In this section, results of the final model of production of Teff in Tigray Region are presented in Table 4.14 below.

The average of the CV of direct estimate of 28 woreda is 41.58 while the average of the CV of EBLUP estimate of these 28 woreda is 31.93. This show on average the EBLUP estimate gave a gain of 13.71 in efficiency on the CV. Therefore, it suggests that the small area estimation technique has an advantage in improving the efficiency of the woreda (small area) level estimates.

Table 4.14: The EBLUP estimates of production in quintal of the 28 Woredas in Tigray Region and their MSE, CV and RSE using area level model

Sr. No	Zone Name	Woreda Name	EBLUP Estimate	MSE of EBLUP Estimate	CV of EBLUP Estimate	Relative Standard Error	Gain in Efficiency
1	North	Tahtay Adiyabo	5535.64	46561173.42	123.27	74.85	-25.14
2	Western	Laelay Adiyabo	59289.17	78346570.07	14.93	37.44	18.57
3		Medebay Zana	72489.99	69638896.85	11.51	57.59	3.36
4		Tahtay Koraro	66081.88	85422304.1	13.99	32.56	25.15
5		Tselemti	60917.18	74985917.39	14.22	51.20	24.20
6	Central	Mereb Lehe	35820.29	53233446.7	20.37	64.74	1.73
7		Ahiferom	111031.80	123583021.6	10.01	49.77	7.53
8		Werie Lehe	78739.18	95596065.08	12.42	31.44	17.21
9		Adwa	65598.80	84736894.91	14.03	30.48	18.39
10		Laelay Maychew	88098.65	101316290.6	11.43	50.27	21.77
11		Tahtay Maychew	39490.61	30907941.62	14.08	85.72	2.98
12		Naeder Adet	50964.26	16592234.24	7.99	93.95	0.15
13		Kola Temben	79348.91	82941930.41	11.48	59.92	12.48
14		Degua Temben	42920.87	48342819.33	16.20	71.43	4.77
15		Tanqu Abergele	51428.94	71964095.78	16.49	43.23	18.05
16	Eastern	Gulumahada	999.02	985738.9099	99.38	99.81	-0.73
17		Saesi Tsaedamba	10688.53	12884408.83	33.58	95.32	15.53
18		Ganta Afeshum	3576.24	6885089.894	73.37	97.99	17.53
19		Hawzen	8516.62	7008881.96	31.09	97.88	2.57
20		Wukro(kelete awlalo)	25968.70	50934268.48	27.48	67.44	5.75
21	Southern	Enderta	66455.22	81548283.72	13.59	42.76	19.59
22		Hintalo Wajirat	91936.13	112097286.9	11.52	37.60	30.75
23		Ambalage	23062.82	54730312.02	32.08	63.44	2.24
24		Endamehoni	3055.30	7753516.297	91.14	97.82	7.40
25		Wofla	19379.99	41282085.25	33.15	77.34	2.04
26	Western	Kafta humera	1806.42	1783558.378	73.93	99.57	8.44
27		Welkait	26248.13	30183201.99	20.93	85.56	7.50
28		Tsegede	16200.47	42986494.84	40.47	76.28	0.26
	Total		1205649.74	1515232729.58	-	-	-
	Mean		43058.92	54115454.63	31.93	56.41	13.71

4.6. Final Estimate of Cultivated Area and Production

As described in Chapter 3, in order to report the final estimate at woreda level an adjustment must be made in the EBLUP estimates to make them consistent with zonal and regional reliable direct estimates. Therefore, the final estimate of cultivated area and production at woreda level for Tigray Region is reported after making adjustments on

EBLUP estimates using the reliable design estimate at zonal level. The adjustment is made as follows:-

$$\text{Adjusted EBLUP estimate of specific woreda} = \frac{\text{EBLUP estimate of the woreda}}{\text{Sum of EBLUP estimates of specific zone}} \times \text{Reliable Zonal Estimate}$$

The estimate of adjusted EBLUP estimates for cultivated area and production of each woredas is presented in Table 4.15. Note that the woreda with outliers observations are approximated by their direct estimate in reporting of adjusted EBLUP estimate.

Table 4.15: Direct estimate and adjusted EBLUP estimates of cultivated area and production for woreda of Tigray Region

			Cultivated Area		Production	
Sr. No	Zone Name	Woreda Name	2007/08 Direct Estimate	Adjusted EBLUP Estimate	2007/08 Direct Estimate	Adjusted EBLUP Estimate
1	North	Tahtay Adiyabo	1423.924	1604.57	9290.42	5876.3
2	Western	Laelay Adiyabo	6618.698	6955.60	70610.62	62937.79
3		Tahtay Koraro	5700.771	5797.29	97570.80	76950.98
4		Medebay Zana	13474.42	10829.42	72519.47	70148.52
5		Asegede Tsimbela	5579.096	6478.14	217799.73	231203
6		Tselemti	7236.166	8368.05	43991.53	64665.99
7	Central	Mereb Lehe	4158.055	4198.05	50921.23	38668.64
8		Ahiferom	7272.919	7855.85	127305.10	119860.8
9		Werie Lehe	17015.53	14057.84	104945.78	85000.34
10		Adwa	4765.539	5134.76	93149.04	70815.07
11		Laelay Maychew	7481.059	7834.37	60310.97	95104.05
12		Tahtay Maychew	4970.144	5263.18	38019.76	42630.81
13		Naeder Adet	4257.9	4657.21	53280.68	55016.82
14		Kola Temben	8634.504	8671.37	63422.92	85658.55
15		Degua Temben	3628.991	3987.81	46423.18	46333.84
16		Tanqu Abergele	7077.392	7601.60	56828.71	55518.45
17	Eastern	Gulumahada	119.6556	125.52	1008.42	4394.902
18		Saesi Tsaedamba	956.0624	1015.19	7667.57	47021.13
19		Ganta Afeshum	715.8619	769.46	2945.95	15732.64
20		Hawzen	1431.229	1396.03	8034.31	37466.43
21		Wukro (kelete awlalo)	2247.742	2069.05	31886.51	140275.6
22		Atsbi Wenberta	590.6598	685.96	307174.51	114241.9

23	Southern	Seharte Samre	12419.63	14945.73	131035.00	128506.3
24		Enderta	6058.102	7443.45	63609.93	65172.77
25		Hintalo Wajirat	7758.393	8849.80	66602.44	90161.96
26		Ambalage	1348.993	1635.73	33954.51	22617.76
27		Endamehoni	519.7461	680.11	2888.67	2996.339
28		Rayaazebo	20289.88	14665.04	196124.12	192339.3
29		Alamata	9502.917	9235.17	154371.34	151392.3
30		Wofla	1797.763	2240.40	23606.74	19006
31	Western	Kafta humera	658.428	696.51	1628.38	1849.724
32		Welkait	2234.552	2118.21	22586.91	26877.36
33		Tsegede	797.3897	875.65	21100.62	16588.83
	Total		160721.5	178742.12	2282615.87	2283031.14
	Mean		4870.347	5416.43	69170.18	69182.76

CHAPTER FIVE

DISCUSSION AND CONCLUSION

The principal objective of this study is to discuss the recently developed methodologies in small area estimation technique that may be applicable to generate the woreda level estimate on the cultivated area and production of Teff in 2007/08 for Tigray Region. The study mainly uses small area model specifically area level model which is applied to get woreda level estimates.

Beside the area level model, woreda level estimates can also be generated using direct domain estimation method. Although direct domain estimation method is the easiest to calculate, it has some disadvantages. One of the major disadvantages is they would have used small or no sample size in the small area under consideration which will lead to generate an estimate with a high standard error. For instance, in cultivated area estimation, the direct domain estimate gives an estimate with coefficient of variation as much as 74.49. Therefore, reliable estimate cannot be generated using the direct domain estimation method. But, using the estimate generated from this direct estimation with combining multiple sources of data, each with different facets of the necessary information, small-area estimation method could be fitted to get better small area estimates.

In this study, using the available census data and administrative data from Ministry of Agriculture and Rural Development which contained information corresponding to the direct estimate, a small area models is fitted and woreda level estimates are generated. The major advantage of the model based estimators such as small area model is the use of auxiliary data to predict the variable of interest. All of the small area estimators presented in this report "borrow strength" for surrounding small areas in the estimation. As a result, in the analysis of cultivated area estimation and production, highest gain in efficiency with 34.82 and 30.75 reduction in the coefficient of variation are observed, respectively.

The finding of the study is consistent with other studies and suggestions given in the literature. Thus, to get more reliable estimate in woreda, i.e, small area the availability and the quality of auxiliary data is essential. In this study, estimates from previous census give supplementary information in order to improve the efficiency of direct estimate. This can be justified by the final model for cultivated area estimation as well as production. Therefore, we can say that the contribution of the census estimate is invaluable for the fitted small area model in this analysis. Besides, the study clearly shows there are woreda specific random variations. The area level model which considers the woreda specific random factor could give a better fit.

Appropriate assessment was also made while comparing the total of the direct estimate of Tigray Region and the EBLUP estimate. For all the weredas the difference between direct and EBLUP estimate is found to be small but EBLUP estimate leads with a remarkable reduction in mean square error. This strongly supports the usefulness of small area estimation techniques and their application in various aspects to get estimates at wereda level (small area level).

In Ethiopia as well as many developing countries, the application of small area estimation has been almost non-existent. Recently, some endeavors are being undertaken in applying small area estimation technique on agricultural variables as well as in poverty count studies. For instance, the Central Statistical Agency is undertaking preliminary works on small area estimation of cultivated area for some common crop. Therefore, this study has shown that with the availability of auxiliary data, small area estimation can be practical, in particular, on agricultural variables. The successful continuation of studies to derive reliable small area statistics such as woreda level estimates as well as Kebele (Peasant Association) level estimates highly depend on the availability of the needed information from government, statistical offices and related organizations. Using small area estimation method, the data generated from the 2007 Population and Housing Census can be used to the maximum as auxiliary information to get reliable estimates at woreda and kebele level.

REFERENCES

Albacea Z.V. (2003) Estimating Sub-national Poverty Incidence Using Small Area Estimation Technique. University of the Philippines, Los Baños. Laguna, Philippines.

Amoako F., Brown J. and Padmadas S. (2003). District Estimates of Home Deliveries in Ghana: A Small Area Analysis Using DHS and Census.

Australian Bureau of Statistics (ABS)(2006). A Guide to Small Area Estimation. Version1.1. available online at

http://www.nss.gov.au/nss/home.NSF/pages/Small+Areas+Estimates?Open Docu ment.

Bartosinka D. (2006) Attempts at Applying of Small Area Estimation Methods in Agricultural Sample Surveys. Statistics in Transition, Vol.7, No. 6, pp.1203-1218.

Battese G.E., Harter R.M., and Fuller W.A. (1988) An Error component Model for Prediction of County Crop Areas Using Survey and Satellite Data. Journal of American Statistical Association, 83, 28-36.

Biratu Y. and Eskinder T. (2005) Review of the 2001/2 Ethiopian agricultural sample enumeration methodology. The Ethiopian Agricultural Sample Enumeration In-depth Analysis. Proceedings of a Workshop 6th June 2005, Addis Ababa, Ethiopia.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001). Evaluation of Small Area Estimation Methods – An Application to Unemployment Estimates from the UK LFS. Proceedings of Statistics Canada Symposium 2001 – Achieving Data Quality in a Statistical Agency: a Methodological Perspective, Ottawa, Canada, October 17-19, 2001.

Central Statistical Agency (CSA) (2003). The 1994 Ethiopia Agricultural Sample Enumeration. Report on Area and Production of Crops and Crop Utilization for Tigray Region. Addis Ababa, Ethiopia.

Central Statistical Agency (CSA) (2008) The 2000 Agricultural Sample Survey. Report on Land Utilization. Addis Ababa, Ethiopia.

Christensen R. (1996) Analysis of Variance, Design and Regression. *Applied statistical methods*. Chapman & Hall, New Mexico, USA.

Claudio Q., Rosalia C., and Gennaro P. (2007) Estimating Poverty in the Italian Provinces Using Small Area Estimation Models. Metodoški zveski, 4(1):37-70.

Cochran W.G. (1977) Sampling Techniques, 3rd ed., New York: Wiley.

Dick P. (1995) Modeling Net Under coverage in the 1991 Canadian Census. Survey Methodology, 21, 45-54.

Fay R. E. and Harriot R.A. (1979) Estimation Of Income From Small Places: An Application Of James- Stein Procedures To Census Data. Journal of the American Statistical Association, 74, 269-277.

Gonzalez M.E. (1973) Use and Evaluation of Synthetic Estimates. Proceedings of Social Statistics Section, America Statistical Association, 33-36.

Ghosh M., Natrajan K., Stroud T.W.F. and Carlin B.P. (1998) Generalized linear models for small area estimation. Journal of the American Statistical Association, 93, 273-282.

Ghosh M. and Rao J.N.K. (1994) Small Area Estimation: An Appraisal; Statistical Science, 9, No.1, 55-93.

Harville D.A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association, 72, 320-340.

Henderson C. R. (1975) Best Linear Unbiased Estimation and Prediction under a selection Model. Biometrics, 31, 423-447.

Kacker R.N. and Harville D.A. (1984) Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models. Journal of the American Statistical Association, 79, 853-862.

Kish L. (1965) *Survey Sampling*. John Wiley & Sons, New York.

Liard N. M and Ware J. H. (1982) Random-Effects Models for Longitudinal Data. Biometrics, 38, 963-974.

Minilik T. (2004) Small area estimation using mixed model with special emphesis on Best Linear Unbiased Predictor (BLUP). Statistics Department. MSc. Project Paper, Addis Ababa University.

National Bank of Ethiopia (NBE) (2007) The 2006/2007 Annual Report of National Bank of Ethiopia. Addis Ababa, Ethiopia.

National Research Council (1999) Small-area estimates of school-age children in poverty: Interim report 3. C. F. Citro and G. Kalton, eds. Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics. Washington, D.C.: National Academy Press.

Office for National Statistics (ONS) (2001). Small Area Estimation in ONS, National Statistics Methodology Advisory Committee. United Kingdom.

Prasad N.G.N and Rao J.N.K. (1990) The estimation of the mean squared error of small area estimators. Journal of the American Statistical Association, 85, 163-171.

Rao J. N. K. (2003) Small Area Estimation. Wiley Series in Survey Methodology, New York.

Rao J. N. K. (2005) Inferential issues in small area estimation: Some new developments. Statistics in Transition, 7, 523 -526.

Rao J. N. K. (2002) *Small area estimation with applications to Agriculture*. Proceedings of the Conference on agricultural and environmental statistical appications in Rome, Vol. III, 555-564.

Robinson G.K. (1991) That BLUP is a good thing: The estimation of random effects. Statistical Science, 6, 15-51.

Rubio V.G. (2008) Small Area Estimation with R. BIAS Project UseR! 2008. Epidemiology and Public Health Imperial College, London, England

Russo C., Sabbatini M. and Salvatore R. (2002) General Linear Models in Small Area Estimation: An Assessment in Agricultural Surveys. Department Economia e Territorio, University of Cassino, Italy.

Saei A. and Chambers R. (2003) Small area estimation: A review of methods based on the application of mixed models, Southampton Statistical Sciences Research Institute Methodology Working Paper M03/16. University of Southampton, United Kingdom.

Sardnal C.E. (1984) Design Consistent Versus Model Dependent Estimation For Small Domains. Journal of the American Statistical Association, 79: 624-631.

Sardnal C.E. and Hidroglou M.A. (1989) Small domain estimation: A conditional analysis. Journal of American Statistical Association, 84, 266-275.

Sardnal C.E., Swensson B. and Wretman J.H. (1992) Model assisted survey sampling. New York: Springer-Verlag.

Singh R. and Goel R.C. (2000) Use of Remote Sensing Satellite Data in Crop Surveys. Technical Report, India Agricultural Statistics Research Institute, New Delhi, India.

Stasny E., Goel, P.K. and Rumsey D.J. (1991) County estimates of wheat production. Survey methodology, 17, 211-225.

Torelli N. and Trevisani M (2008) Labour force estimates for small geographical domains in Italy: problems, data and models. Working Paper n. 118, Dipartimento di Scienze Economiche e Statistiche, Universit` di Trieste, Italy.

Verbeke, G. and Molenberghs, G. (1997) Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.

Yogendra P.C., Tak Mak and Fasil N. (2003) A review of some recent methods in small area estimation with special emphesis on Monte-Carlo based techniques. Journal of the Ethiopian Statistical Association, XIII, 1-15.

Appendix I: Estimation Procedures of Totals and Sampling Errors

The following formulas were used to estimate total area of land under specific crop and production of specific crop in a stratum such as woreda.

1. For estimating Total Area of Land under Specific Crop such as Teff:

$$\hat{y}_q = \sum_{i=1}^{n_q} w_{qi} \sum_{j=1}^{q_{qi}} y_{qij} = \sum_{i=1}^{n_q} w_{qi} y_{qi}$$

in which, $w_{qi} = \frac{M_q H_{qi}}{n_q m_{qi} q_{qi}}$ is the basic weight as given in methodology part

Where:

q represents woreda

n_q is the total number of sample EAs successfully covered in the q^{th} woreda.

M_q is the measure of size of the q^{th} woreda as obtained from the sampling frame.

m_{qi} is the measure of size of the i^{th} sample EA in the q^{th} woreda obtained from the sampling frame.

H_{qi} is the total number of agricultural households of the i^{th} sample EA in the q^{th} woreda.

q_{qi} is the number of sample agricultural households successfully covered in the i^{th} sample EA in the q^{th} woreda.

y_{qij} is the value of area for agricultural households j, in the i^{th} EA in the q^{th} woreda under a specific crop.

y_{qi} is the sample total area under specific crop for EA i in woreda q

$\hat{y}_q$ estimate of total area under specific crop in woreda q

2. For estimating Total Production under Specific Crop such as Teff:

$$\hat{P}_q = \sum_{i=1}^{n_q} w_{qi} P_{qi}$$

in which, $P_{qi} = y_{qi} * \overline{Y}_{qi}$

where, $\overline{Y}_{qi} = \dfrac{Y_{qi}}{16C_{qi}}$ is average yield per square meter of a specific crop in the i^{th} EA in the q^{th} woreda.

$\hat{P}_q$ is estimate of total quantity of production of a specific crop in the q^{th} woreda.

Y_{qi} is sample total quantity of production of a specific crop from defined area of land for crop cutting of a crop in the i^{th} EA in the q^{th} woreda.

P_{qi} is estimate of total quantity of production under specific crop for EA i in woreda q.

C_{qi} is the number of crop cutting of a specific crop in the i^{th} EA in the q^{th} woreda.

3. Sampling Variance of Estimates:

Sampling variance for the estimate of woreda total of area and production for a specific crop are estimated by the following formulas.

$$Var(\hat{y}_q) = (1-f_q)\frac{n_q}{n_q-1}\sum_{i=1}^{n_q}\left(\hat{y}_{qi} - \frac{\hat{y}_q}{n_q}\right)^2 + f_q\sum_{i=1}^{n_q}(1-f_{qi})\left(\frac{q_{qi}}{q_{qi}-1}\right)\sum_{j=1}^{q_{qi}}\left(\hat{y}_{qij} - \frac{\hat{y}_{qi}}{q_{qi}}\right)^2$$

$$Var(\hat{P}_q) = (1-f_q)\frac{n_q}{n_q-1}\sum_{i=1}^{n_q}\left(\hat{P}_{qi} - \frac{\hat{P}_q}{n_q}\right)^2 + f_q\sum_{i=1}^{n_q}(1-f_{qi})\left(\frac{q_{qi}}{q_{qi}-1}\right)\sum_{j=1}^{q_{qi}}\left(\hat{P}_{qij} - \frac{\hat{P}_{qi}}{q_{qi}}\right)^2$$

f_q = average first stage probability of selection of EAs within woreda q.

$f_{qi} = \dfrac{q_{qi}}{H_{qi}}$ = average for the proportion of systematic selection of agriculture household within the i^{th} sample EA in woreda q.

$\hat{y}_{qi}, \hat{P}_{qi}$ are weighted total area and production, respectively, of a specific crop in the i^{th} EA and q^{th} woreda.

$\hat{y}_{qij}, \hat{P}_{qij}$ are weighted values of area and production, respectively, from j^{th} agricultural household in the i^{th} EA and q^{th} woreda under a specific crop.

Note: - The strata are independent. In estimating the sampling variance by the above formula, selection of EAs within a woreda is assumed to be with replacement. By so doing the variance estimate may be slightly over estimated but it greatly simplifies the estimation procedure.

4. Coefficient of Variation (CV) of Estimates:

Coefficient of Variation (CV) in percentage of estimate of woreda total of area and production for a specific crop are given by:

$$CV(\hat{y}_q) = \frac{\sqrt{Var(\hat{y}_q)}}{\hat{y}_q} \times 100\%$$

$$CV(\hat{P}_q) = \frac{\sqrt{Var(\hat{P}_q)}}{\hat{P}_q} \times 100\%$$

Appendix II.a: The Census estimate, MOARD data and Direct estimate with its MSE and CV of cultivated area in hectares of the 33 Woredas in Tigray Region.

Sr. No	Zone Name	Woreda Name	2000/02 Census Estimate	2007/08 MOARD Data	2007/08 Direct Estimate	2007/08 Variance of Direct Estimate	CV of Direct Estimate
1	North	Tahtay Adiyabo	300.00	271.55	1150.81	1663.00	3.54
2	Western	Laelay Adiyabo	7908.00	5093.69	4992.88	5127914.96	45.35
3		Tahtay Koraro	7968.00	4087.85	6066.32	3356443.84	30.20
4		Medebay Zana	7763.00	5697.48	10829.42	6980322.52	24.40
5		Asegede Tsimbela	3050.00	4634.90	7511.97	5818853.57	32.11
6		Tselemti	3719.00	6152.09	6206.02	823628.85	14.62
7	Central	Mereb Lehe	4855.00	2990.38	4974.26	1358903.00	23.44
8		Ahiferom	4846.00	6285.10	6452.27	2081556.42	22.36
9		Werie Lehe	7337.00	7991.48	14057.84	14110090.20	26.72
10		Adwa	3597.00	3971.06	4653.59	3440208.85	39.86
11		Laelay Maychew	6306.00	6121.23	8617.34	2495294.12	18.33
12		Tahtay Maychew	4165.00	4004.03	5476.41	2198755.15	27.08
13		Naeder Adet	3153.00	4218.86	3149.63	125805.00	11.26
14		Kola Temben	9733.00	6991.10	7901.30	11485049.88	42.89
15		Degua Temben	2315.00	3019.90	3121.46	579349.32	24.38
16		Tanqu Abergele	5776.00	6472.73	4596.95	593562.38	16.76
17	Eastern	Gulumahada	137.00	127.42	125.00	5981.00	61.87
18		Saesi Tsaedamba	1227.00	1201.22	850.93	127149.30	41.90
19		Ganta Afeshum	491.00	562.77	837.82	134205.00	43.73
20		Hawzen	2236.00	1155.83	3624.69	2320107.78	42.02
21		Wukro (kelete awlalo)	4382.00	2010.69	3491.42	2020264.00	40.71
22		Atsbi Wenberta	132.00	669.14	548.00	166643.00	74.49
23	Southern	Seharte Samre	6604.00	9641.23	9483.71	9307441.66	32.17
24		Enderta	2961.00	4621.08	4685.82	4744119.61	46.48
25		Hintalo Wajirat	6785.00	5459.62	7205.22	2887756.44	23.58
26		Ambalage	1626.00	1004.36	885.43	72436.34	30.40
27		Endamehoni	332.00	264.93	355.00	61986.00	70.13
28		Rayaazebo	13188.00	9739.03	14665.04	6270717.14	17.08
29		Alamata	16075.00	5788.42	7748.89	9783633.29	40.37
30		Wofla	1032.00	1102.95	1435.69	296828.83	37.95
31	Western	Kafta humera	5451.00	589.29	816.73	23350.95	18.71
32		Welkait	7983.00	2130.10	3152.86	430480.55	20.81
33		Tsegede	3452.00	624.15	1050.74	18658.52	13.00
	Total		156885	124695.7	160721.5	99249160.5	-
	Mean		4754.091	3778.656	4870.347	3007550.32	32.08

Appendix II.b: The Census estimate, MOARD data and Direct estimate with its MSE and CV of production in quintal of the 33 Woredas in Tigray Region.

Sr. No	Zone Name	Woreda Name	2000/02 Census Estimate	2007/08 MOARD Data	2007/08 Direct Estimate	2007/08 Variance of Direct Estimate	CV of Direct Estimate
1	North	Tahtay Adiyabo	2063.81	1067.41	9290.42	83108384.30	98.13
2	Western	Laelay Adiyabo	33667.44	44263.00	70610.62	559402912.89	33.50
3		Tahtay Koraro	37248.67	102306.00	97570.80	210446956.10	14.87
4		Medebay Zana	38022.99	101985.00	72519.47	805661106.59	39.14
5		Asegede Tsimbela	55847.94	39643.50	217799.73	13491668085.25	53.33
6		Tselemti	37747.82	15558.00	43991.53	285729326.46	38.42
7	Central	Mereb Lehe	16385.17	19941.06	50921.23	126618756.25	22.10
8		Ahiferom	62645.74	75627.58	127305.10	498620861.23	17.54
9		Werie Lehe	44472.94	62367.00	104945.78	967012214.11	29.63
10		Adwa	36865.76	58493.00	93149.04	911826800.46	32.42
11		Laelay Maychew	53629.98	107085.36	60310.97	400938149.43	33.20
12		Tahtay Maychew	24583.77	60315.00	38019.76	42058986.38	17.06
13		Naeder Adet	24838.14	8144.00	53280.68	18796473.54	8.14
14		Kola Temben	48632.77	84021.59	63422.92	230853685.58	23.96
15		Degua Temben	23681.99	22555.00	46423.18	94756791.18	20.97
16		Tanqu Abergele	29438.03	36218.00	56828.71	385333366.61	34.54
17	Eastern	Gulumahada	896.45	2233.10	1008.42	989567.35	98.65
18		Saesi Tsaedamba	15247.80	7604.40	7667.57	14181852.17	49.11
19		Ganta Afeshum	6230.64	6938.80	2945.95	7170398.62	90.90
20		Hawzen	8172.07	22179.00	8034.31	7315618.47	33.66
21		Wukro (kelete awlalo)	13397.02	28201.05	31886.51	112290474.76	33.23
22		Atsbi Wenberta	10995.61	883.35	307174.51	91966275818.74	98.73
23	Southern	Seharte Samre	54642.86	56951.50	131035.00	746464362.25	20.85
24		Enderta	38805.91	48226.00	63609.93	445587347.92	33.18
25		Hintalo Wajirat	54332.92	69140.78	66602.44	792590282.88	42.27
26		Ambalage	10660.39	29243.50	33954.51	135801964.63	34.32
27		Endamehoni	3260.96	4853.25	2888.67	8102448.39	98.54
28		Rayaazebo	128570.35	189502.80	196124.12	1254058613.04	18.06
29		Alamata	72030.63	226512.00	154371.34	9606238450.42	63.49
30		Wofla	9310.72	15717.38	23606.74	69011067.14	35.19
31	Western	Kafta humera	5661.83	63850.00	1628.38	1799005.21	82.37
32		Welkait	19178.88	82549.00	22586.91	41226544.22	28.43
33		Tsegede	7298.18	44745.60	21100.62	73876431.62	40.73
	Total		1028466.18	1738922.01	2282615.87	124395813104.19	-
	Mean		31165.64	52694.61	69170.18	3769570094.07	42.99

Appendix III: Procedure used to fit the area level model using the R package (used to fit cultivated area data of 33 woredas)

```
library(nlme)
library(foreign)
library(spam)
library(SAE2)
spam.options(eps = 1e-11)
area33=read.dta("C:\\area33.dta")
sr<-area33$sr
dir<-area33$dir
cen <-  as.vector(by(area33$cen,area33$sr,mean))
min <-  as.vector(by(area33$min,area33$sr,mean))
aux <- data.frame(cen,min)
a33<-cbind(data.frame(sr=area33$sr,dir=area33$dir,vard=area33$vard),aux)
a33eblup <- EBLUP(dir ~ cen+min, ~vard, data = a33,method="REML")
a33eblup
AIC(a33eblup)
deviance(a33eblup)
f<-a33eblup$fitted.values
r<-a33eblup$residuals
b<-a33eblup$coef
varb<-a33eblup$varcoeff
v<-a33eblup$randeff
u<-a33eblup$sigma2u
varu<-a33eblup$varsigma2u
vard<-a33eblup$desvar
y<-a33eblup$eblup
mse<-a33eblup$mse
b
varu
```

```
varb
cvd<-sqrt(vard)/dir*100
length(cvd)
cvy<-sqrt(mse)/y*100
mean(cvd)
mean(cvy)
esta<-data.frame(sr,cen,min,dir,vard,cvd,y,mse,cvy,f,r,v)
write.table(esta, file = "E:\\esta.txt", sep = ",")
plot(y,dir,xlab="EBLUP Estimate",ylab="Direct Estimate")
abline(a=0,b=1)
title("Plot of the EBLUP versus Direct Estimate Model 3",sub)
fm1<-lm(dir~y)
summary(fm1)
plot(y,r,xlab="EBLUP Estimate",ylab="Residuals")
abline(a=0,b=0)
title("Plot of the EBLUP Estimate versus Residual for Model 3",sub)
shapiro.test(r)
hist(r)
qqnorm(r)
qqline(r)
```

Appendix IV.a.: Output of the descriptive analysis and general linear model in cultivated area data analysis

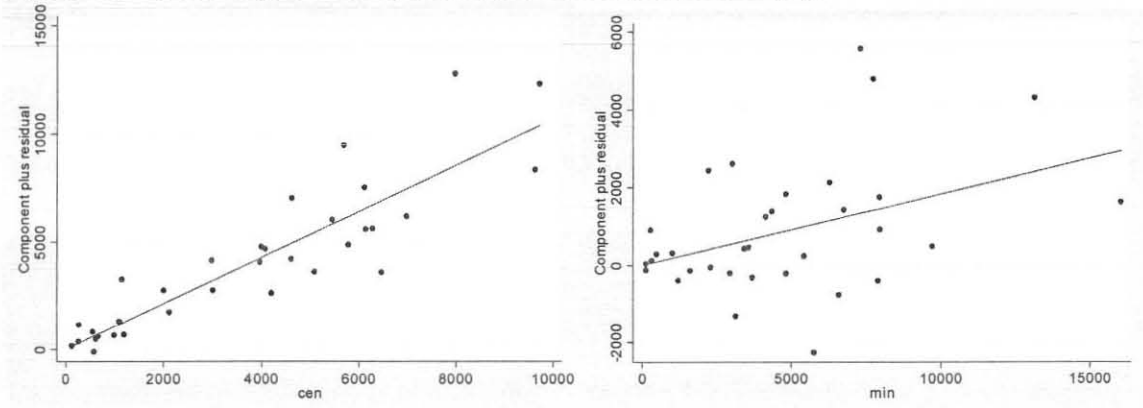
Correlation between direct with census and MOARD data(obs=33)

	dir	cen	min
dir	1.0000		
cen	0.9080	1.0000	
min	0.7293	0.7040	1.0000

Partial correlation of direct with census and MOARD data

Variable	Corr.	Sig.
cen	0.8120	0.000
min	0.3028	0.092

Plot of direct estimate versus census estimate and MOARD data



Source	SS	df	MS	Number of obs = 33
Model	391947101	2	195973551	F(2, 30) = 79.06
Residual	74362088	30	2478736.27	Prob > F = 0.0000
Total	466309189	32	14572162.2	R-squared = 0.8405
				Adj R-squared = 0.8299
				Root MSE = 1574.4

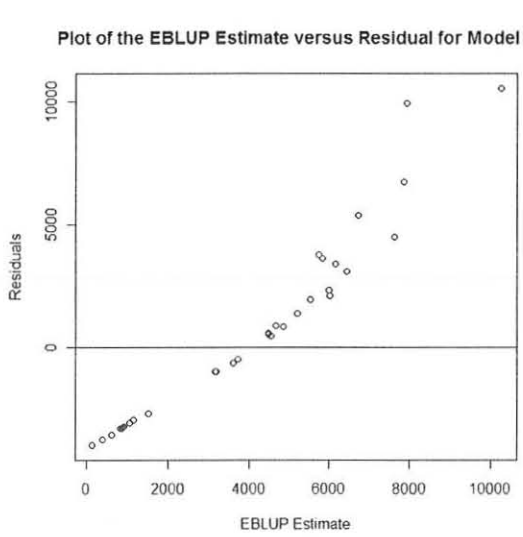
dir	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cen	1.067822	.1401255	7.62	0.000	.7816479	1.353997
min	.1830544	.1052106	1.74	0.092	-.0318144	.3979232
_cons	-34.84367	482.475	-0.07	0.943	-1020.189	950.5017

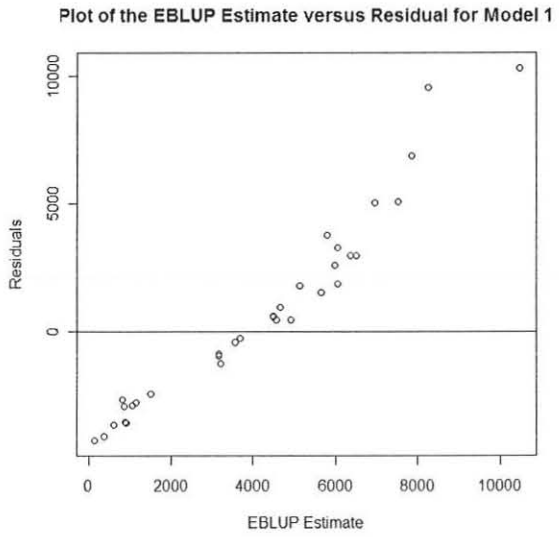
Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	-34.844	482.475	-.072	.943		
	cen	1.068	.140	7.620	.000	.504	1.982
	min	.183	.105	1.740	.092	.504	1.982

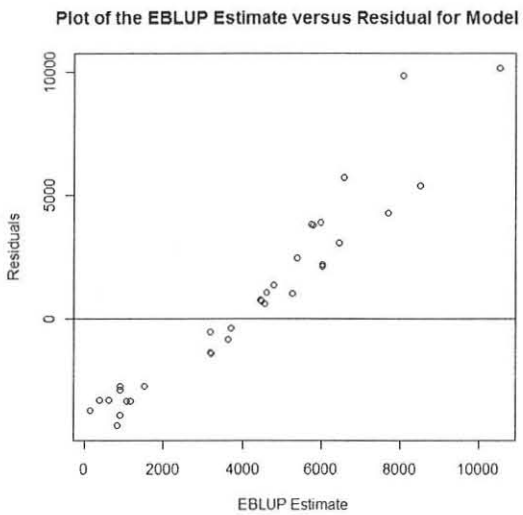
a Dependent Variable: dir

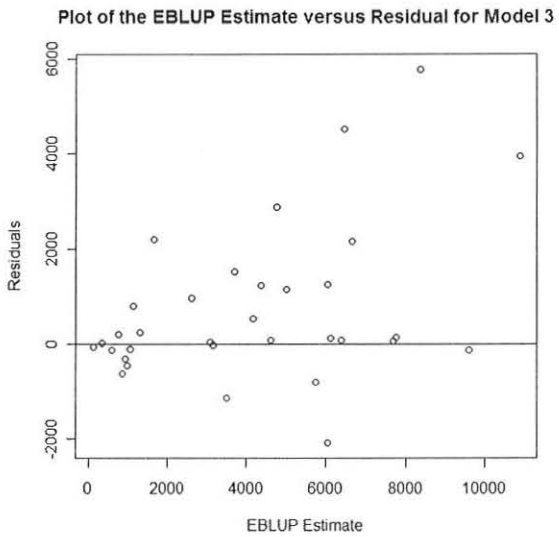
Appendix IV.b.: The result of model diagnostics plots of the 33 woredas in fitting cultivated area model.

i) The plot of the EBLUP estimate versus Residual

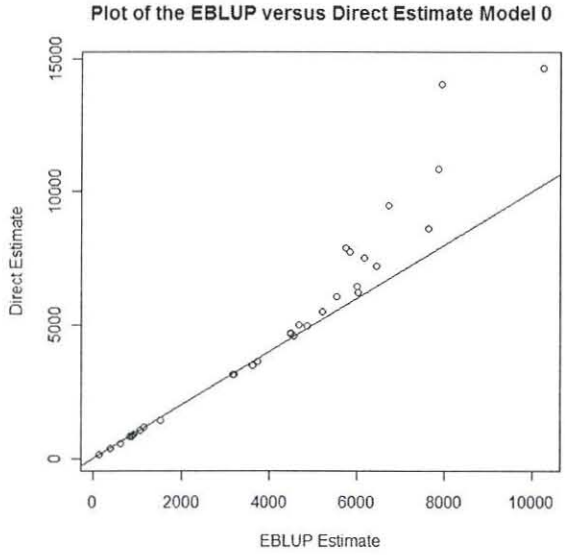


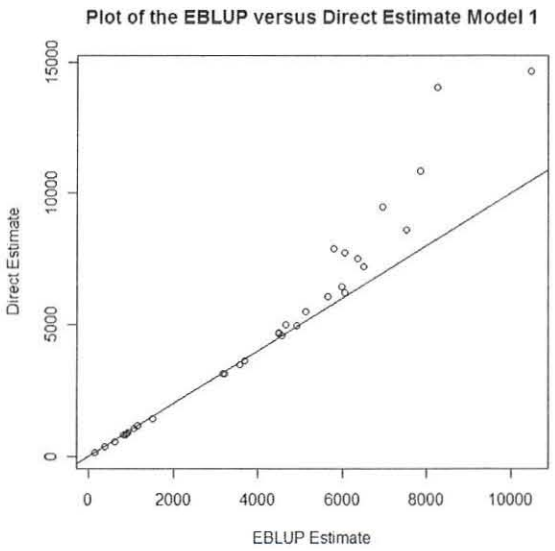


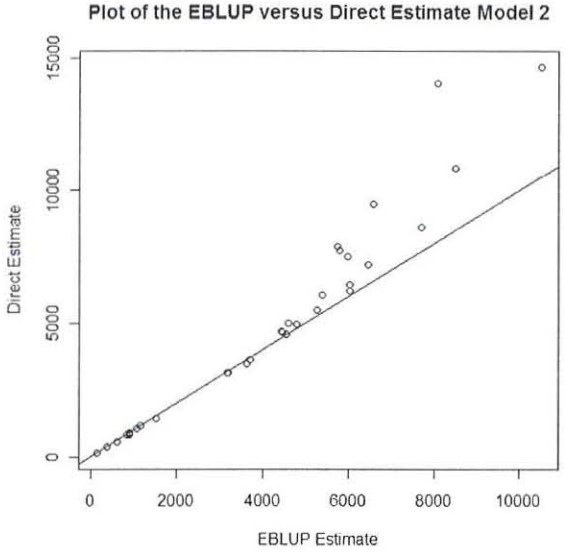


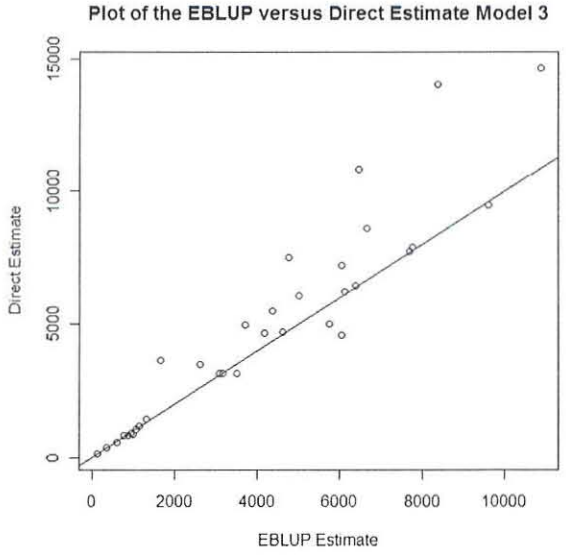


ii) The plot of the EBLUP estimate versus Direct Estimate



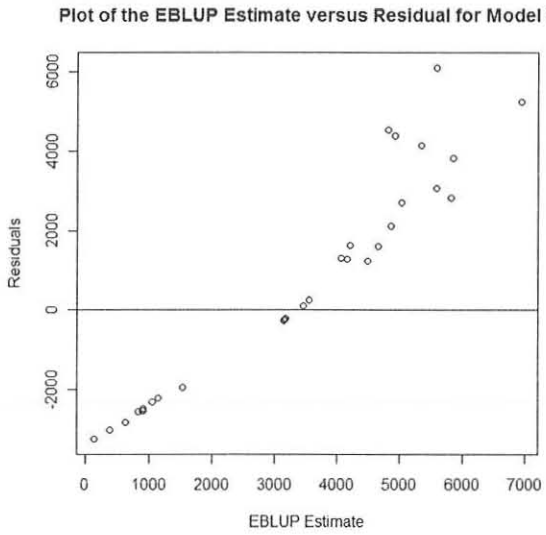


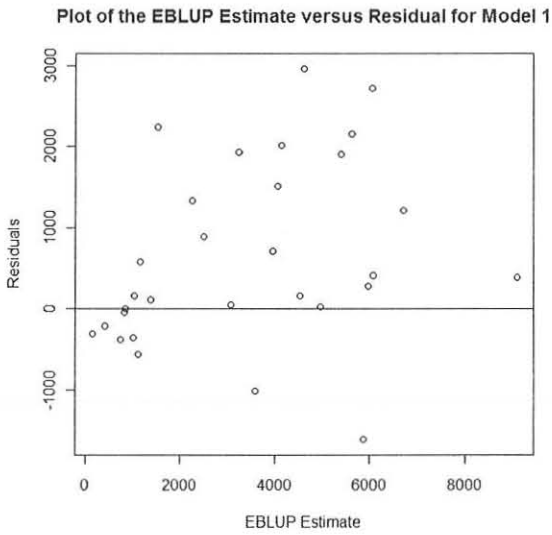


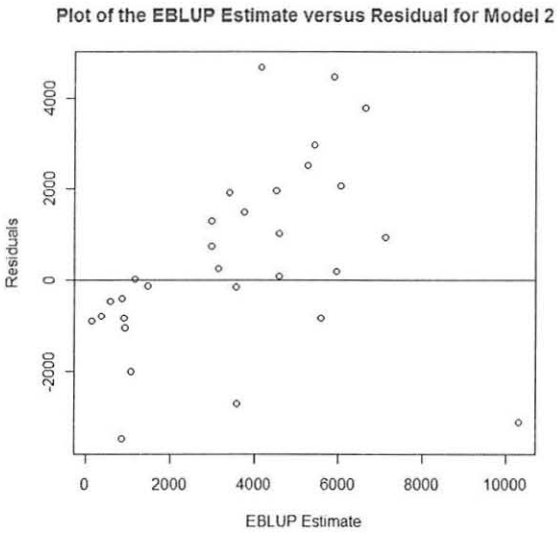


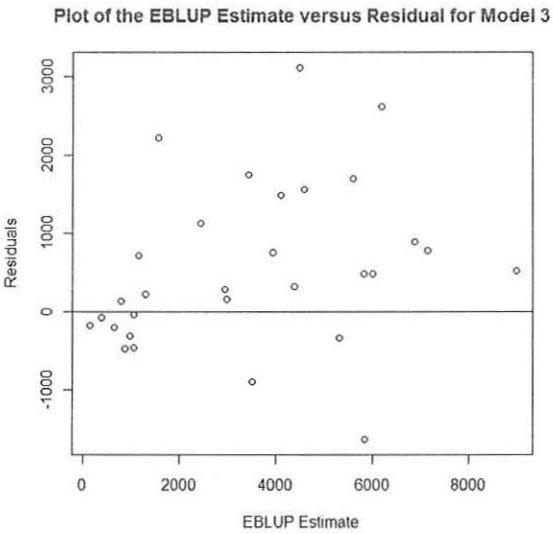
Appendix IV.c.: The result of model diagnostics plots of the 30 woredas in fitting cultivated area model.

i) The plot of the EBLUP estimate versus Residual

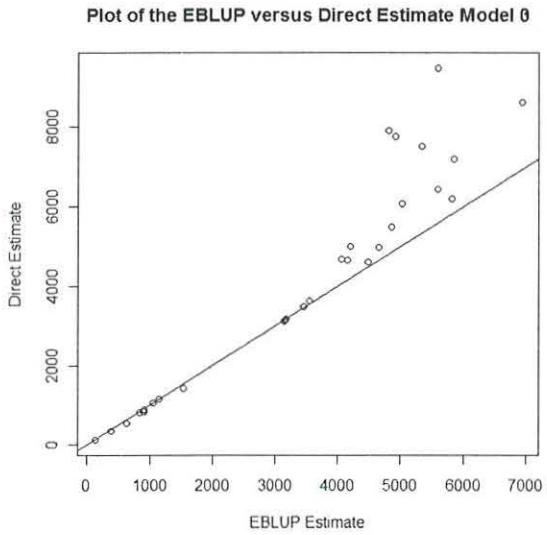


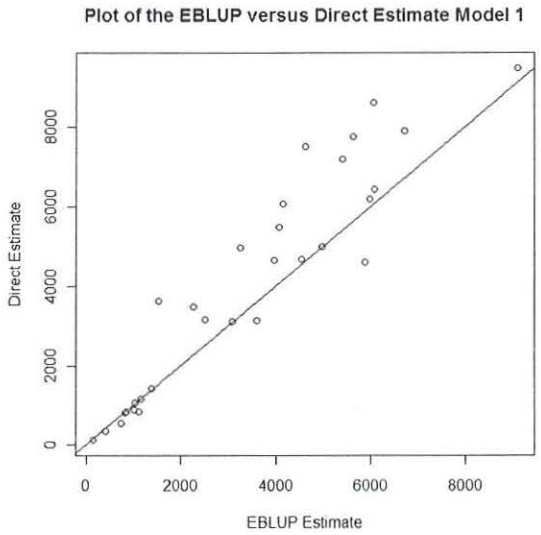


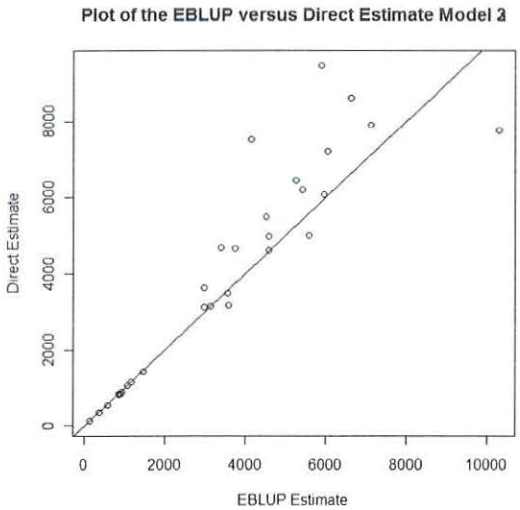


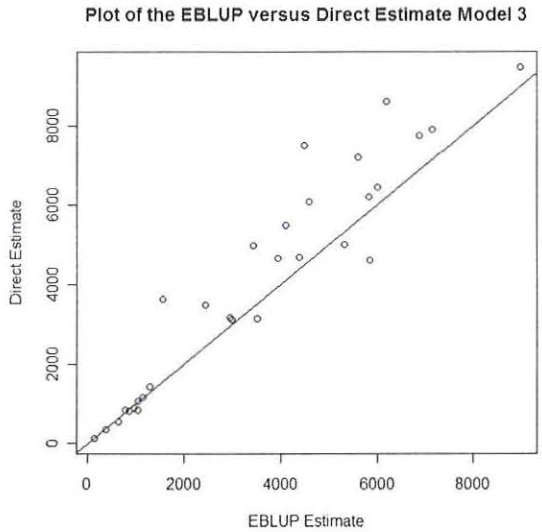


ii) The plot of the EBLUP estimate versus Direct Estimate





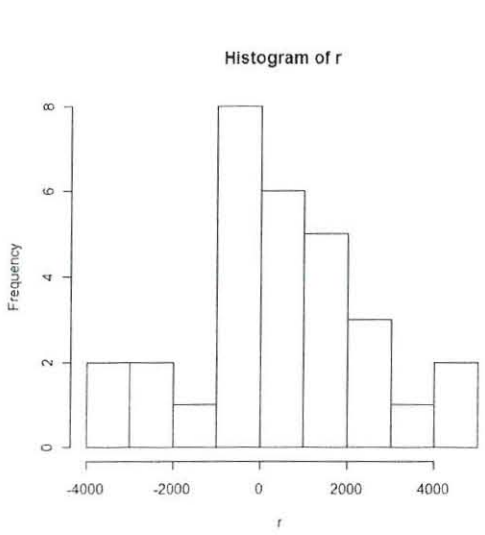


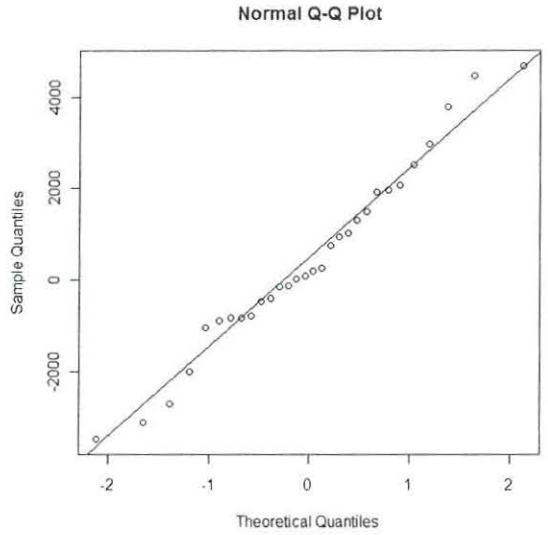


Appendix IV.d.: Output of diagnostic on the selected model, Model 1

```
Call:
EBLUP(formula = dir ~ cen + min, varformula = ~vard, data = a32,method =
"REML")
Coefficients:
          [,1]
[1,] 177.879524762916
[2,] 0.832533440736301
[3,] 0.115400196773738
Class 'spam'

Variance of the random effects: 183265.0
Log likelihood: -5981.537
AIC(a32eblup)
[1] 11971.07


        Shapiro-Wilk normality test
data:  residual
W = 0.9689, p-value = 0.5108
```

Histogram and Normal Q-Q Plot for selected model, Model 1



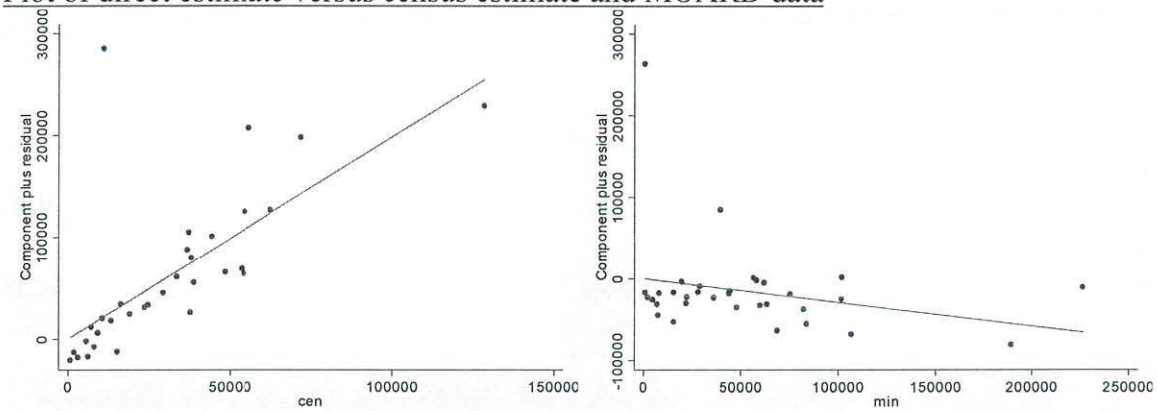


Appendix V.a.: Output of the descriptive analysis and general linear model in cultivated area data analysis

Result of 33 woreda

```
. correlation between dir versus census and MOARD(obs=33)
        |   dir     cen     min
------------+---------------------------
        dir |  1.0000
        cen |  0.5896  1.0000
        min |  0.3783  0.7824  1.0000
Partial correlation of direct with census and MOARD data
   Variable |  Corr.    Sig.
------------+------------------
        cen |  0.5093   0.003
        min | -0.1650   0.367
```

Plot of direct estimate versus census estimate and MOARD data



Source	SS	df	MS	Number of obs = 33
				F(2, 30) = 8.64
Model	5.6012e+10	2	2.8006e+10	Prob > F = 0.0011
Residual	9.7271e+10	30	3.2424e+09	R-squared = 0.3654
				Adj R-squared = 0.3231
Total	1.5328e+11	32	4.7901e+09	Root MSE = 56942

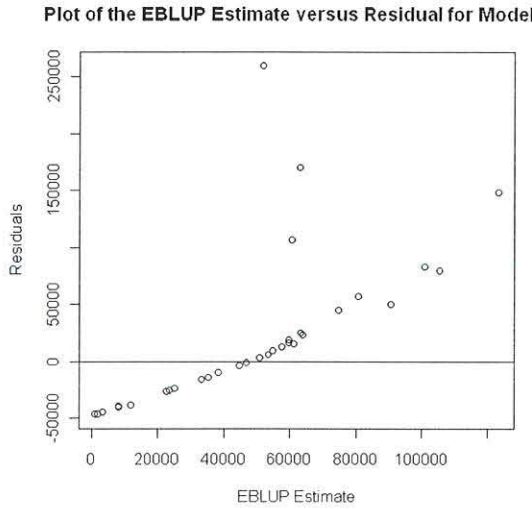
dir	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cen	1.982554	.61159	3.24	0.003	.7335202	3.231587
min	-.28989	.3164447	-0.92	0.367	-.9361563	.3563763
_cons	22658.26	15565.39	1.46	0.156	-9130.515	54447.03

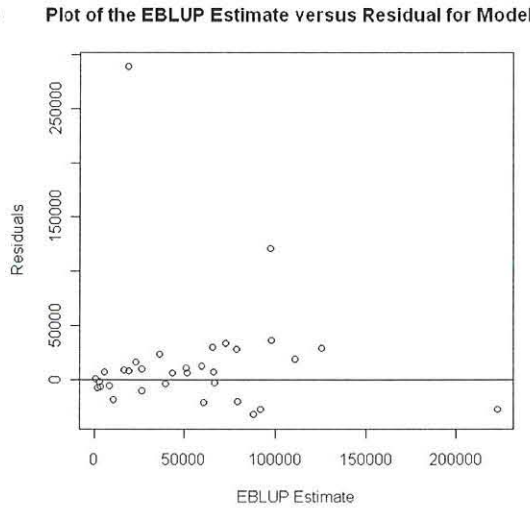
Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	22658.259	15565.394	1.456	.156		
	cen	1.983	.612	3.242	.003	.388	2.578
	min	-.290	.316	-.916	.367	.388	2.578

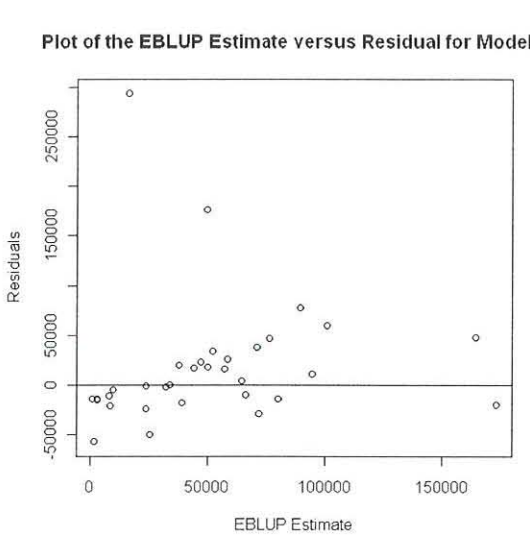
a Dependent Variable: dir

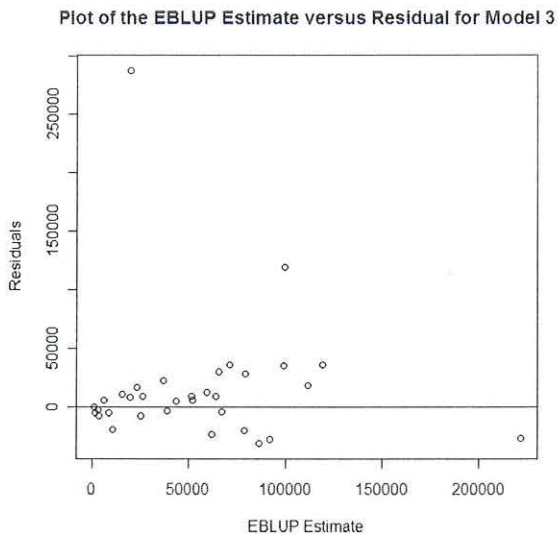
Appendix V.b.: The result of model diagnostics plots of the 33 woreda in fitting production model.

i) The plot of the EBLUP estimate versus Residual

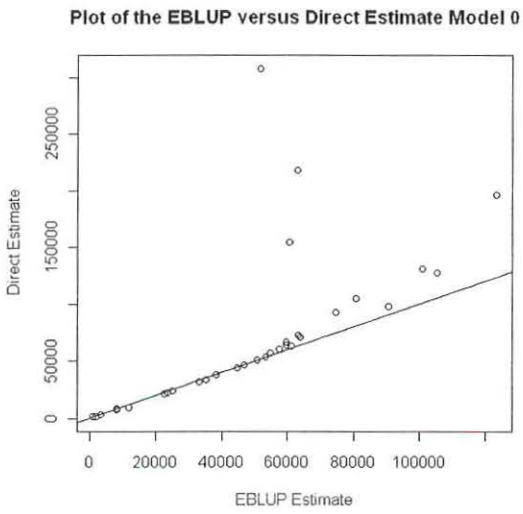


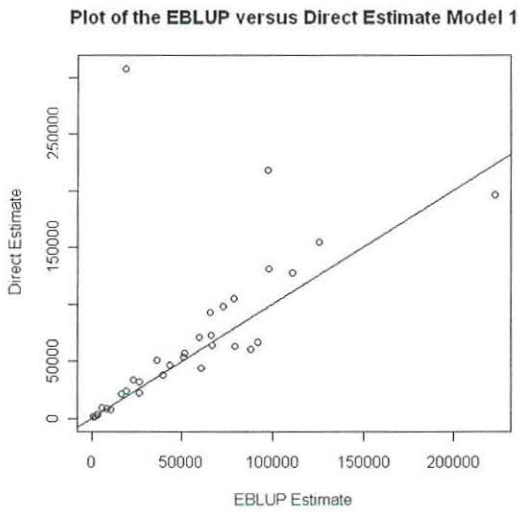


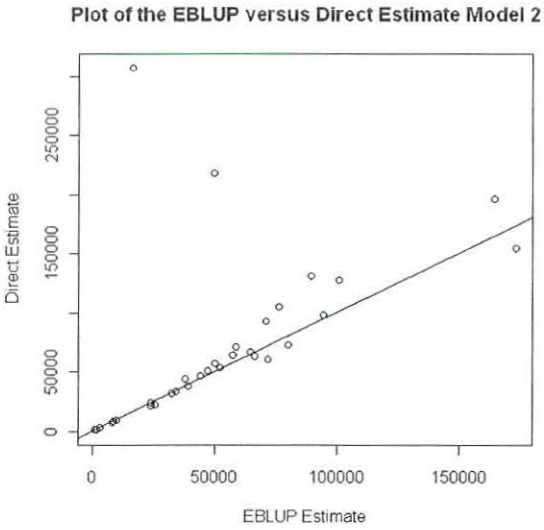


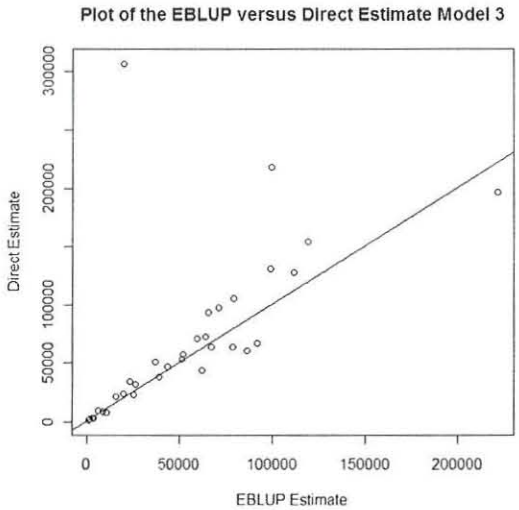


ii) The plot of the EBLUP estimate versus Direct Estimate



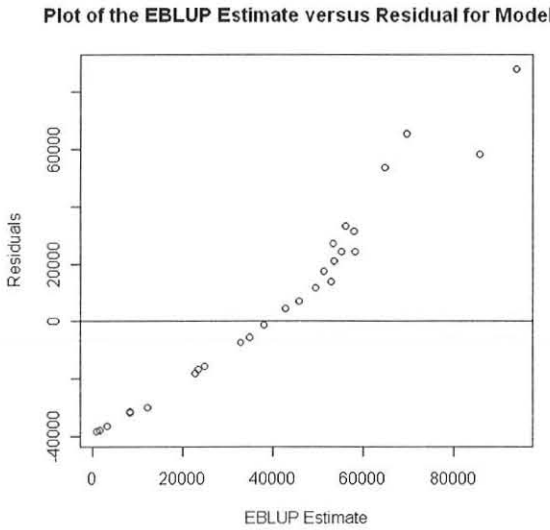


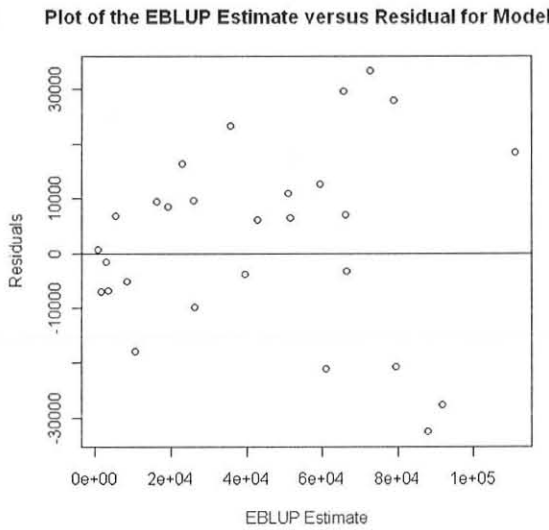


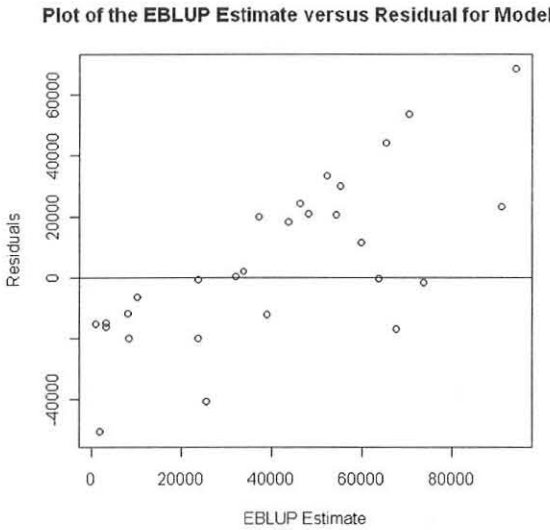


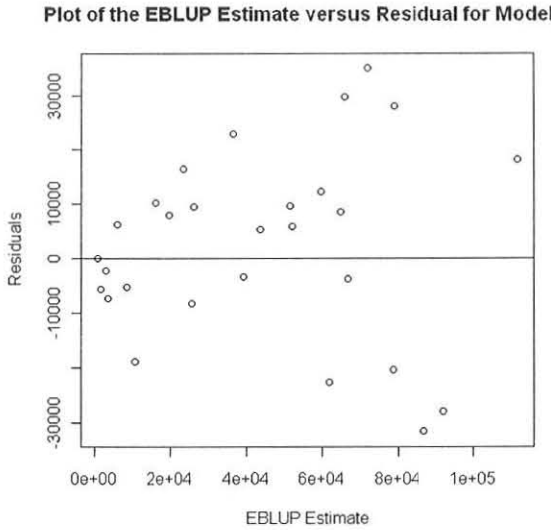
Appendix V.c.: The result of model diagnostics plots of the 28 woredas in fitting production model.

ii) The plot of the EBLUP estimate versus Residual

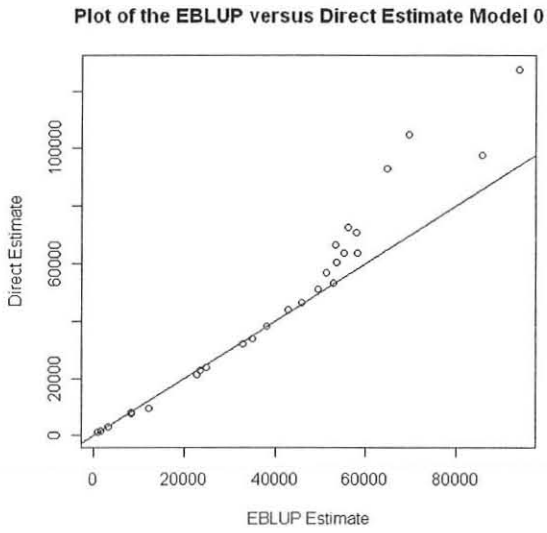


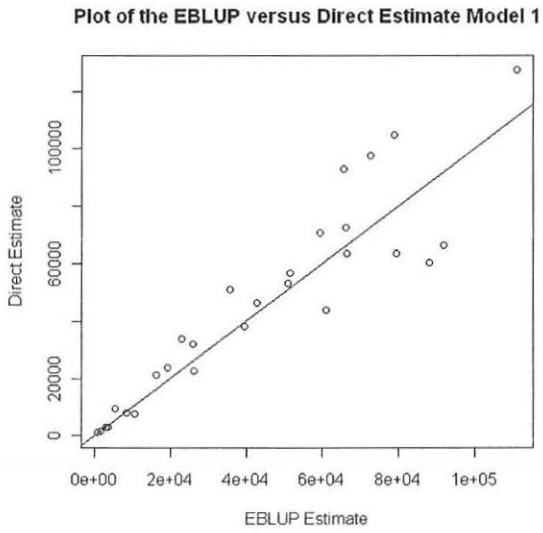


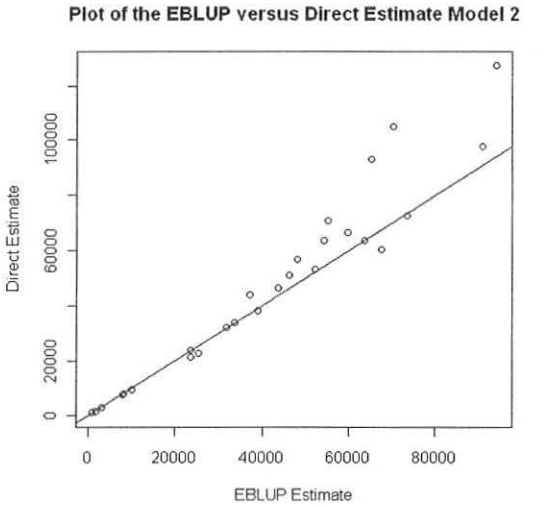


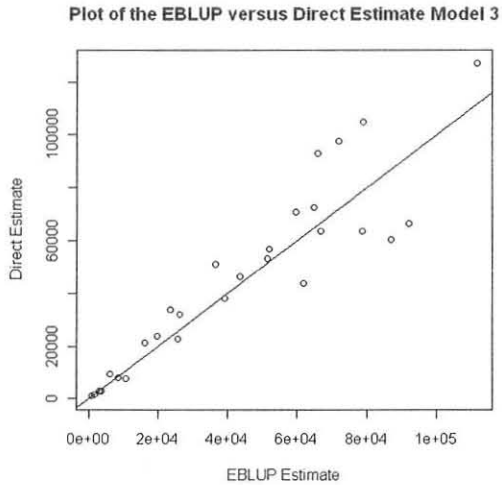


ii) The plot of the EBLUP estimate versus Direct Estimate





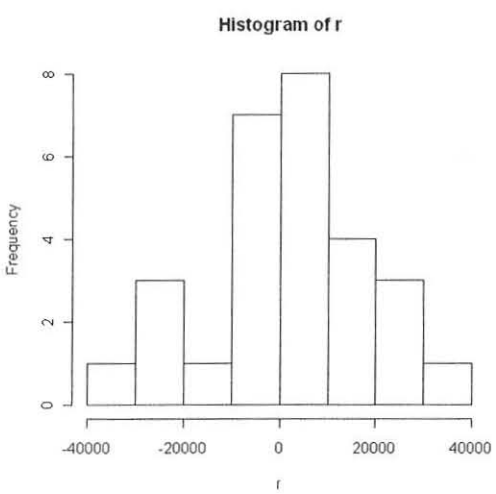


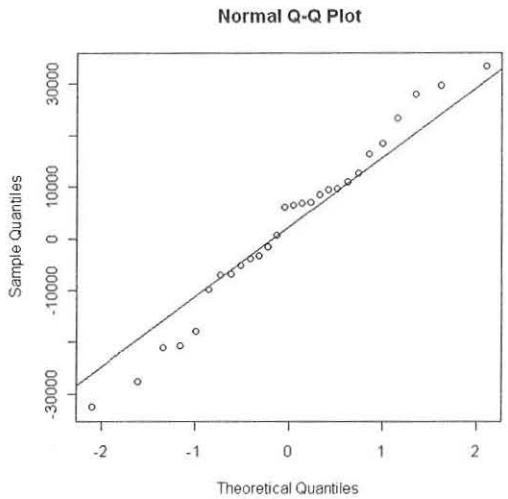


Appendix V.d.: Output of diagnostic on the selected model, Model 3

```
Call:
EBLUP(formula = dir ~ cen, varformula = ~vard,
data = a32,
   method = "REML")

Coefficients:
            [,1]
[1,] -1237.056127
[2,]     1.755839
Class 'spam'

Variance of the random effects: 69699504
Log likelihood: -7183.897
> AIC(a32eblup)
[1] 14373.79


shapiro.test(r)
      Shapiro-Wilk normality test
data:  r
W = 0.9786, p-value = 0.8169
```

Histogram and Normal Q-Q Plot for selected model, Model 3





DECLARATION

I, the undersigned, declare that the thesis is my original work, has not been presented for degrees in any other university and all sources of material used for the thesis have been duly acknowledged.

Name: Seid Jemal

Signature: .....

Place: Faculty of Science, Addis Ababa University

Date: August 2009

This thesis has been submitted for examination with my approval as a University Advisor.

27/08/09

Dr. Fentaw Abegaz