ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES
DEPARTMENT OF STATISTICS
GRADUATE REGULAR PROGRAM

MSc Thesis

Small Area Estimation of Maize Yield of Wereda-level Using Mixed Effect Linear Model with Spatial Auxiliary Information

Damtew Berhanu

Advisor:

Professor M.K. Sharma

Jun 17, 2016

ADDIS ABABA

Small Area Estimation of Maize Yield of Wereda-level Using Mixed Effect Linear Model with Spatial Auxiliary Information

BY DAMTEW BERHANU

A Thesis Submitted to the Department of Statistics in Partial Fulfillment of the Requirements for the Degree of Master of Science in Statistics (Applied Statistics).

JUN, 2016

ADDIS ABABA

Addis Ababa University

College of Natural and Computational Sciences

Graduate Regular Program

Department of Statistics

Name:	Damtew Berhanu
Degree:	Master of Science in Applied Statistics
Title of Thesis:	**Small Area Estimation of Maize Yield of Wereda - level Using Mixed Effect Linear Model with Spatial Auxiliary Information**

Approved by Board of Examiners

Prof. M.K. Sharma Advisor	Signature	Date
Birhanu Teshome, (PhD) Internal Examiner	Signature	Date
Emmanuel Gebreyohannes, (PhD) External Examiner	Signature	Date

Abstract

Small area estimation (SAE) plays an important role in survey sampling due to the growing demands of such statistics by both government and private sectors for various decision making and planning purposes. The main purpose of this study was to produce small area estimates of maize yield at wereda level based on a unit level mixed effect model approach and compare the results with direct estimator and other model based indirect estimators. In order to achieve this goal we used annual agricultural sample survey data from Central Statistical Agency (CSA) and other set of spatial auxiliary information from CSA, Ministry of Agriculture and Natural Resources (MoANR) and web sites. The AGSS data for 238 weredas of Oromia Region included in the sample survey was used to compute direct and model based estimators. The model based estimators compared were: EBLUP_B based on unit level mixed model, SEBLUP_A based on spatial Fay Herriot's model, SYN_SLM based on simultaneous autoregressive lag dependent linear model and SYN_SACLM based on spatial simultaneous autoregressive SAC linear model. Using four diagnostic metrics the study revealed in general that EBLUP_B estimator show better performance than other estimators.

Key words and phrases: *Small area estimator, direct estimator, small area model, wereda, empirical best linear unbiased predictor, spatial dependence, maize yield, weight matrix*

Acknowledgements

First of all, I would like to thank my advisor Prof. M.K Sharma for his continuous follow-up excellent advice and offering comments, suggestions and important reference materials that led to significant improvements during planning and execution of my research work.

I would like to thank my organization Central Statistical Agency, for all their cooperation they offered me just to mention a few; providing me an opportunity to study in Addis Ababa University, their financial support particularly for this research, and providing me data and materials to do my research work.

My special and very big gratitude extended to Demisse Bemirew for his moral and technical support he offered me while I have been doing this research and also for his valuable materials support to do this research.

I wish to thank Department of Statistics, Addis Ababa University for allowing me to do this research and for the financial support they offered.

I wish to address my heartfelt thanks to all my families for their unreserved assistance and encouragement throughout my educational life.

'ABOVE ALL 'THANKS GOD'

Table of Contents

List of Tables

List of Figures

Acronyms and Abbreviations

AGSS	Agricultural Sample Survey
ANOVA	Analysis of Variance
ARB	Absolute Relative Bias
ARE	Absolute Relative Error
BHF	Battese, Harter and Fuller
BLUP	Best Linear Unbiased Predictor
CkD	Cook's distance
CSA	Central Statistical Agency
CV	Coefficient of Variation
DE	Direct Estimator
DEM	Digital Elevation Model
EA	Enumeration Area
EBLUP	Empirical Best Linear Unbiased Predictor
EBLUP_B	Empirical Best Linear Unbiased Predictor based on Unit Level Model
EFF	Relative Efficiency
ESRI	Environmental Science Research Institute
EVI	Enhanced Vegetation Index
FH	Fay and Herriot
FAO	Food and Agricultural Organization of the United Nations
GLS	General Least Square
MoANR	Ministry of Agriculture and Natural Resources
MoARD	Ministry of Agriculture and Rural Development
ML	Maximum Likelihood
MODIS	Moderate Resolution Imaging Spectroradiometer
MSE	Mean Squared Error
NDVI	Normalized Differences Vegetation Index
PSU	Primary Sampling Units
REML	Residual Error Maximum Likelihood
RMSE	Root Mean Squared Error
SAC	Spatial Autocorrelation
SAE	Small Area Estimation
SAR	Simultaneous Autoregressive Regressive
SEBLUP_A	Area Level Spatial Empirical Best Linear Unbiased Predictor
SFH	Spatial Fay and Herriot
SSU	Secondary Sampling Units
SYN	Synthetic Estimator
SYN_SACLM	Synthetic Estimator based on Simultaneous Autoregressive Regressive SAC Model
SYN_SLM	Synthetic Estimator based on Spatial Lag Model
USGS	United States Geological Survey

Chapter One

INTRODUCTION

1.1. Background

In recent years one of the most difficult problems faced by the official statistical institutes is the information delivery about small geographical areas. Various government agencies of countries (e.g., the United States Census Bureau, USDA's National Agricultural Statistics Service (NASS), Statistics Canada and the Central Statistical Office of the United Kingdom) have a requirement to produce reliable small-area statistics (Bellow & Lahiri, 2011). Small domain or area refers to a population for which reliable statistics of interest cannot be produced due to certain limitations of the available data. In the context of agricultural surveys these domains denote a geographical region (e.g. a wereda, rural kebele, census divisions, etc.). One of the primary objectives of producing small area estimates is to provide summary statistics to central or local governments so that they can plan for immediate or future resource allocation. Small area estimates of crop parameters such as harvested area, production and yield are used by farmers, agribusinesses and government agencies for local agricultural decision making.

Small area estimation (SAE) was first studied at Statistics Canada in the seventies. Small area estimates have been produced using administrative files or surveys enhanced with administrative auxiliary data since the early eighties (Hidiroglou, 2007). Demographers have long been using a variety of indirect methods for small area estimation of population and other characteristics of interest in post-census years. Purcell and Kish (1980) categorize these methods under the general heading of Symptomatic Accounting Techniques (SAT). Such techniques utilize current data from administrative registers in conjunction with related data from the latest census (as cited in Gohosh & Rao, 1994).

Small area estimation models vary widely with respect to available auxiliary information and with respect to the relationship of this information to the variables of interest. There is no useful general model which will accommodate all small area estimation problems. Nevertheless, many of the basic relationships can be approximated reasonably well by simple linear regression models. SAE methods can be divided broadly into "design-based" and "model-based" methods.

The latter methods use either the frequentist approach or the full Bayesian methodology, and in some cases combine the two, known in the SAE literature as "empirical Bayes." Domain estimates that are computed using only the sample data from the domain are known as direct estimates. Although direct estimates have several desired design-based properties, direct estimates often lack precision when domain sample sizes are small. Domains for which direct estimates of adequate precision cannot be produced are known as small areas (Rao, 2003). Survey designs usually focus on achieving a particular degree of precision for estimates at a much higher level of aggregation than that of small areas; therefore, the sample sizes for small areas are typically small. Producing estimates for small areas with an adequate level of precision often requires indirect estimators that use auxiliary data or values of the variable of interest from related areas, or both.

In making estimates for small areas with adequate level of precision, it is often necessary to use "indirect" estimators that "borrow strength" by using values of the variable of interest, y, from related areas and/or time periods and thus increase the "effective" sample size. These values are brought into the estimation process through a model (either implicit or explicit) that provides a link to related areas and/or time periods through the use of supplementary information related to y, such as recent census counts and current administrative records.

The Central Statistical Agency of Ethiopia conducts agricultural surveys annually both in Meher and Belg seasons that cover all regions and publishes summary of agricultural statistics at national, regional, and zonal level (CSA, 2013). However, these surveys do not provide accurate wereda level direct estimates of crop yield due to sample size limitations and the associated large standard errors (Bocci et al., 2012). The demand of such data in small areas like wereda has greatly increased during the past few years. This increase is due to the usefulness of these data in government policy and program development, allocation of various funds and regional planning (Rao, 2004).

Often many developing countries like Ethiopia conduct agricultural censuses on a sample basis and where a country has statistical data from which to compute indicators of land use, crop production, livestock, farm structure, incomes and living conditions, their quality at the local level is not homogenous. Therefore, when the survey data provides few observations and cost constraints prevent additional surveys or additional sampling of the study area, existing

information must be integrated and harmonized to produce credible statistics on the dynamics of change at the local level. In this case, an estimate of the target variable for local domains can be obtained from reliable data relating to a larger domain that includes the domains in question. In short, the available aggregate data for broad areas must be disaggregated at the local level for small areas (Monica, 2015).

Most researchers in the areas of small area estimation propose an integration of survey data with known larger and smaller area level auxiliary information to produce more reliable estimate of the target variable at the local level (Monica, 2015). Spatial auxiliary information is crucial in many applications of the estimation method for local areas, because it can increase the efficiency and effectiveness of the estimations (Michael & James, 2015). Spatial auxiliary information can be derived from administrative archives and maps of the territory under study, and geographical information systems can provide spatial data relating to coverage, perimeters, extensions and distances.

Mixed-effect unit level small area models that explicitly consider spatial correlation as a function of distance between small areas of study have been recently proposed by some researchers and have received significant attention (e.g. Pedro & Luis (2011) and Monica (2015)).

In this study we aimed at evaluating unit level mixed effect small area model that utilizes annual agricultural sample survey (main season) data and spatial auxiliary information to provide reliable model based indirect estimates of maize yield at wereda level. In order to evaluate the precision of the estimates obtained from the proposed model, we used various diagnostic approaches for small area estimators including CV, RRMSE and other bias diagnostic measures. The estimates from the mixed-effect unit level model were assessed in comparison with direct estimators, spatial area level linear mixed-effect model and spatial synthetic estimators based on linear fixed effect models.

1.2. Statement of Problem

Annual agricultural sample surveys could not provide reliable direct wereda level crop yield estimates due to lack of adequate representative samples for each wereda. Cost and time constraints also dispossess to increase sample sizes or conduct additional surveys to get wereda level estimates. In recent years there has been a growing demand for reliable small area

agricultural statistics, owing to their usefulness in local agricultural decision making, planning and policy implementation by government and other organizations. Therefore, an appropriate small area estimation approach that employs implicit and explicit models approach should be applied in order to address the availability of reliable small area statistics.

A lot of researches have been conducted all over the world regarding small area estimation techniques and applications in the agriculture sector. But the methods used by different national statistical offices vary mainly due to the availability of auxiliary information and the context of agricultural practice of each country. In the case of Ethiopia, there is no available current literature that justifies an optimal small area estimation model suitable for estimation of yield and other crop parameters.

According to Sirvastava (2010) presentation to FAO, Fay and Herriot's area level model have been utilized by CSA from 2008 – 2009. Based on this model CSA produced area and yield estimate for some selected crops namely teff, barley, maize, sorghum, wheat and finger millet at wereda level. The 2001 Ethiopian Agricultural Census Enumeration results and the former Ministry of Agriculture and Rural Development's (MoARD) wereda level estimates obtained by an aggregative approach for the year 2007 and 2008 used as auxiliary information in the model. But CSA did not use the Fay and Herriot's area level model in the years after 2009 up to now.

In the past several years the Fay-Herriot's area level model has been applied in many empirical studies conducted in various countries. The results are generally satisfactory given the characteristics of each case-study. However, in the case of Ethiopia, the quality of small area estimates obtained depends on the timeliness and accuracy of wereda level auxiliary information used as covariates in the model. None of the auxiliary information used in the model possesses such qualities to predict crop yield. The main reason is that several factors determine crop yield in small rain-fed farming practices for instance current environmental factors, use of modern technology and type of seed. In addition to this quality of MoARD's wereda level estimates used as auxiliary information is unknown. The reason is that the ministry did not use any objective statistical methodology to collect such data.

In this study, we explored an application of unit level mixed effect small area estimation models to derive model-based estimates of maize yield at wereda levels in the Oromia Region by linking

data generated under agricultural sample survey (main season) scheme by CSA and spatial and remote sensing auxiliary information.

1.3. Objectives of Study

1.3.1. General Objective

Obtaining accurate estimates or predictions from available data are one of the important goals in applied statistical research. Therefore, the main purpose of this research was to produce estimates of maize yield using an appropriate unit level mixed effect small area model and evaluate the results with direct estimators and other model based estimators on their CV, RRMSE and other diagnostics approaches.

1.3.2. Specific objectives

More specifically the study aimed at addressing the following specific objectives.

The study attempted to:

i. Assess various spatial auxiliary information gathered from remote sensing data, GIS maps and administrative sources that best correlates with maize yield output per hectare.
ii. Fit and evaluate various spatial and non-spatial small area models based on linear fixed effects only and linear mixed effects models to estimate crop yield at wereda level using highly correlated auxiliary information and annual agricultural sample survey data.

1.4. Significance of the Study

The conclusions and recommendations drawn from the findings of this study can be used by concerning organizations, practitioners or researchers who want to involve in the areas of agricultural researches specifically in small area estimation of maize yield.

1.5. Limitations of the Study

This study has encountered various constraints from its design to final stage. The major constraints were: i) Shortage of reference materials in relation to small area estimation in Ethiopian context ii) Reliability of model based small area estimators are heavily dependent on accuracy of spatial auxiliary information at wereda and enumeration area level. However, among a set of auxiliary variables that could be considered for small area estimation of maize yield, we only used four auxiliary variables that satisfied this requirement iii) Limitation of the available data to fit unit level spatial mixed effect model restricted the researcher only to try area level

SFH model. This is because unit level SFH model requires unit (enumeration area) level variance inputs.

1.6. Scope of the Study

The study area included 252 weredas of the Oromia Region. The annual agricultural sample survey (AGSS) 2013 data provides direct estimators of maize yield only for 238 weredas which have greater than zero sample sizes. This small area estimation study was conducted to evaluate the proposed small area model using only main agricultural season productions by small agricultural holders.

1.7. Thesis Organization

This thesis is divided in to five chapters, the first chapter is introduction, which covers about the basic aspects of the research including, background, statement of the problem, objectives, significance, limitations, and scope of the study. The second chapter incorporates review of related literature. The third chapter deals with the methodologies followed and materials used to conduct this research. In the fourth chapter, analysis of the results and discussion of outputs are presented. The final chapter covers summary, conclusions and recommendations.

Chapter Two

LITERATURE REVIEW

Chapter two provides review of related literature organized under the following subtitles (1) Small area estimation approaches (2) Small area models, (3) Spatial extensions small area models (4) Determinants of maize productivity (5) Applications of small area estimation models in agriculture sector

2.1. Small Area Estimation Approaches

2.1.1. Design-based Approach

The design-based approach to small area estimation or, more generally, to domain estimation is based on the traditional probability sampling theory (e.g. Cochran 1977, Cassel et al. 1977), which rests on the assumption of finite and fixed population and drawing random samples from it with selection probabilities defined by the sampling design. In the case of sampling without replacement consider that the sample s_d is drawn without replacement within domain U_d, d = 1, …,D. Let π_{dj} be the inclusion probability of j^{th} unit from d^{th} domain in the corresponding domain sample s_d and let $W_{dj} = \pi_{dj}^{-1}$ be the corresponding sampling weight. The unbiased estimator of $\bar{Y}_d$ is the Horvitz-Thompson estimator (Horvitz & Thompson, 1952), given by

(2.1) $$\hat{\bar{Y}}_{d,DIR} = N_d^{-1} \sum_{j \in s_d} \frac{Y_{dj}}{\pi_{dj}} = N_d^{-1} \sum_{j \in s_d} w_{dj} Y_{dj}$$

Now let π_{djk} be the inclusion probability of unit j and k pairs from d^{th} domain in the sample s_d. The sampling variance is given by

(2.2) $$V_\pi\left(\hat{\bar{Y}}_{d,DIR}\right) = \frac{1}{N_d^2}\left\{\sum_{j=1}^{N_d} \frac{Y_{dj}^2}{\pi_{dj}}(1 - \pi_{dj}) + 2\sum_{j=1}^{N_d} \sum_{\substack{k=1 \\ k>j}}^{N_d} \frac{Y_{dj}}{\pi_{dj}} \frac{Y_{dk}}{\pi_{dk}} (\pi_{djk} - \pi_{dj}\pi_{dk})\right\}$$

If $\pi_{djk} > 0$, $\forall$ (j, k), an unbiased estimator of this variance is given by

(2.3) $$\hat{V}_\pi\left(\hat{\bar{Y}}_{d,DIR}\right) = \frac{1}{N_d^2}\left\{\sum_{j \in s_d} \frac{Y_{dj}^2}{\pi_{dj}}(1 - \pi_{dj}) + 2\sum_{j \in s_d} \sum_{\substack{k \in s_d \\ k>j}} \sum_{j=1}^{N_d} \frac{Y_{dj}}{\pi_{dj}} \frac{Y_{dk}}{\pi_{dk}} \frac{(\pi_{djk} - \pi_{dj}\pi_{dk})}{\pi_{djk}}\right\}$$

Sample survey data are extensively used to provide reliable direct estimates of totals and means for the whole population and large areas or domains. A direct estimator is one that

uses values of the variable of interest, y, only from the sample units in the domain of interest. However, a major disadvantage of such estimators is that unacceptably large standard errors may result: this is especially true if the sample sizes within the domain is small or zero (Lohr & Prasad, 2010).

Model-based methods have also been used to develop direct estimators and associated inferences. Such methods provide valid conditional inferences referring to the particular sample that has been drawn, regardless of the sampling design (Rao, 2003).

The variance of design-based estimator differs from that for the model based estimator. The discrepancy is due to the different definitions of variance. In design-based sampling, the variance is the average squared deviation of the estimate from its expected value, averaged over all samples that could be obtained using a given design. If we are using a model, the variance is again the average squared deviation of the estimate from its expected value, but here the average is over all possible samples that could be generated from the population model (Cochran , 1977).

The model-based estimator uses a prediction approach in which the values of y not in the sample are predicted using the model (Lohr, 2010). The model-based estimates are only model-unbiased that is, they are unbiased only within the structure of that particular model.

2.2. Small Area Models

Small area models are classified into two broad types: (i) Aggregate (or area) level models that relate small area direct estimators to area-specific covariates. Such models are necessary if unit (or element) level data are not available. (ii) Unit level models that relate the unit values of a study variable to unit-specific covariates.

A critical assumption for the unit level models is that the sample values obey the assumed population model, that is, sample selection bias is absent . For area level models, we assume the absence of informative sampling of the areas in situations where only some of the areas are selected to the sample, that is, the sample area values (the direct estimates) obey the assumed population model (Rao, 2003).

2.2.1. Model-Based Indirect Estimation

2.2.1.1. Implicit Models Approach for Small area Estimation

This approach includes three statistical techniques of indirect estimation – which are synthetic, composite and demographic estimations (Azizur, 2008).

Robinson (1991) described some models such as EBLUP (empirical best linear unbiased prediction), that involve design-based and model-based random variables, and analyze random area effects through the use of area level and unit level mixed linear statistical models.

Gohosh and Rao (1994) presented most recurrent method of small area estimation based on models that consider composite estimates. These models combine synthetic and direct estimates, balancing the potential bias of the synthetic estimates with the instability of the direct estimates by a weighted average of the two functions.

Rao (2003) provide extensive overviews on some of the most widely used indirect estimators based on implicit models as the synthetic estimator, the regression-adjusted synthetic, the composite estimator, and the sample-dependent estimator.

Michael (2010) described the synthetic estimator that uses reliable information of a direct estimator for a large area that spans several small areas, and this information is used to obtain an indirect estimator for a small area. It is assumed that the small areas have the same characteristics as the large area.

2.2.1.2. Explicit Models Approach

This class of small area estimation approaches is mainly using different explicit models and in the literature it is termed 'small area models'. Rao (1999) and (2003) classified available small area models as the basic area level models and the basic unit level models. In the first type of models, information on the response variable is available only at the small area level, and in the second type of models, data are available at the unit or respondent level. In addition, the general linear mixed model consists all of these area and unit levels models. A brief summary of these small area models is given below.

Basic area level model

Fay and Herriot (1979) introduced area level models to obtain small area estimators of median income in small places in the U.S. These models are well known in the literature of small area

estimation (SAE) and are the basic tool when only aggregated auxiliary data at the area level are available. They used a two-level Bayesian model which is currently well-known as the Fay-Herriot model or basic area level mixed model. The basic FH model is defined in two stages.

First, since true values population mean are not observable, our data will be the direct estimates $\widehat{\bar{Y}}_{d,DIR}$. These estimates have an error and this error might be different for each area because samples sizes in the areas are generally different.

Rao (1999) pointed out that in the basic area level models, direct survey estimators of the small area population mean that is, $\widehat{\bar{Y}}_{d,DIR}$ are available whenever the sample sizes of the d^{th} area is greater and/or equal to one Thus, in the first stage, we assume the following model representing the error of direct estimates,

(2.4) $$\hat{\theta}_d = \theta_d + \varepsilon_d, d= 1, 2,\ldots, D,$$

and is referred to as sampling model, where $\hat{\theta}_d = \hat{g}(.)$ is the function of direct estimators of the population of interest at d^{th} small area, and $\varepsilon_d's$ are the sampling error terms - assumed to be independently and normally distributed with mean zero and known sampling variance ψ_d.

Rao (2003 distinguished these two assumptions of the matching model and considered them as the limitations of the basic area level model. He argues that the assumption of known sampling variances for the matching sampling model is restrictive and the assumption of a zero mean may not be tenable if the small area sample size is very small and the relevant functional relationship is a nonlinear function of the small area mean. To estimate the sampling variance ψ_d, Fay and Herriot (1979) utilized the generalized variance function (GVF) method that uses some external information in addition to the survey data.

In the second stage, true values of population mean are assumed to be linearly related with a vector of auxiliary variables. Let $\mathbf{x}_d = (x_{1d}, x_{2d}, \ldots, x_{pd})^T$ represents an area-specific auxiliary data for the d^{th}area (d= 1, 2,…,D) and $\bar{Y}_d$ represents the population mean of small area d. The basic area level model can be expressed by the following mathematical way:

(2.5) $$\theta_d = \mathbf{x}_d^T\boldsymbol{\beta} + \varepsilon_d, d = 1,2, \ldots, D,$$

where $\theta_d = g(.)$ is a linear function of Y_d assumed to be related to $\mathbf{x}_d$ through the above linear model, $\boldsymbol{\beta}$ is the vector of regression parameters, and ε_d's uncorrelated errors of the random small area effects which are assumed to be normally distributed with mean zero and variance σ_ε^2. The parameters of this model, $\boldsymbol{\beta}$ and σ_ε^2, are generally unknown and are estimated from the available data. In some applications, not all areas are selected in the sample. Suppose that we have M areas in the population and only D areas are selected in the sample. We assume a model of the form (2.5) for the population, that is, $\hat{\theta}_d = \mathbf{x}_d^T\boldsymbol{\beta} + \epsilon_d, \text{for } d = 1,2,\dots,M$. We further assume that the sample areas obey the population model, that is, the bias in the sample selection of areas is absent so that (2.5) holds for the sampled areas.

Now by combining the matching sampling model (2.4) with linking model (2.5), we get the following form of the standard mixed linear model

(2.6) $$\hat{\theta}_d = \mathbf{x}_d^T\boldsymbol{\beta} + u_d + \epsilon_d, d = 1,2,\dots,D$$

Note that model (2.6) involves both model based random errors u_d and design based random errors ϵ_d. The parameter σ_u^2 is a measure of homogeneity of the areas after accounting for the covariates $\mathbf{x}_d$. Since σ_u^2 is unknown, in practice it is replaced by a consistent estimator $\hat{\sigma}_u^2$. Several estimation methods for σ_u^2 are considered including a moment estimator called Fay-Herriot (FH) method, maximum likelihood (ML) and restricted (or residual) ML (REML).

Basic unit level model

On the other hand, Battese et al. (1988) introduced the basic unit level model based on unit level auxiliary variables. These are related to the unit level values of response through a nested error linear regression model, under the assumption that the nested error and the model error are independent to each other and normally distributed with common mean zero and common or different variances. Rao (2003) represented this type of model by the following mathematical equation

(2.7) $$y_{dj} = \mathbf{x}_{dj}^T\boldsymbol{\beta} + \mathbf{u}_d\mathbf{1}_{\mathbf{n_d}} + \varepsilon_{dj}\ j = 1,2,\dots,N_d, d = 1,2,\dots,D,$$

where $\mathbf{x}_{dj} = (x_{dj1}, x_{dj2}, \dots, x_{djp})^T$ represents unit-specific auxiliary data, which are available for areas $j = 1,\dots,N_d$ and small areas $d = 1,\dots,D$ as N_d is the number of population units in the d^{th} area, and $\boldsymbol{\beta}$ represents the vector of regression parameters. The unit responses y_{dj} are

assumed to be related to the auxiliary values x_{dj} through the above (2.7) nested error regression equation. The $\mathbf{u}_d$'s are normal, independent, and identically distributed with mean zero and variance σ_u^2. The ε_{dj}'s are independent of $\mathbf{u}_d$'s and as well as independently normally distributed with mean zero and variance σ_ε^2.

2.3. Spatial Extension Small Area Models

All the above model based estimators did not account for spatial information. However there are situations prticularly in agricultural and environmental data, where the relationship between y and x is not constant over the study area, that is, the regression coefficients vary spatially across the geography of interest. Tzavidis et al. (2012) referred to this phenomenon as spatial nonstationarity.

Cressie (1993) used spatial models for random area effects in order to take into account the correlation between neighbouring areas.

Petrucci and Pratesi (2014) and Tzavidis et al. (2012) similarly proposed using spatial nonstationary extension to the linear mixed effect and linear synthetic models in the case of spatial nonstationarity

Petrucci and Pratesi (2014) also suggested that the use of spatial information is likely to be most productive when the available model covariates are weak. In this case, spatial information can substantially strengthen prediction for non-sampled areas - provided there is significant spatial correlation.

Most often we consider a linear regression model with spatial dependence in the error structure. In particular, a Simultaneous Autoregressive (SAR) error process (Anselin, 1992), where the vector of random area effects $\mathrm{v} = (v_i)$ satisfies

(2.8) $$\mathrm{v} = \rho\mathbf{W}\mathrm{v} + \mathrm{u},$$

were ρ is a spatial autoregressive coefficient, $\mathbf{W}$ is a proximity matrix of order m and $\mathrm{u} \sim \mathrm{N}(0, \sigma_u^2\mathbf{I_m})$. Since $\mathrm{v} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathrm{u}$ with $\mathrm{E}(\mathrm{u}) = 0$ and $\mathrm{Var}(\mathrm{u}) = \sigma_u^2\mathbf{I_m}$, we have $\mathrm{E}(\mathrm{v}) = 0$ and $\mathrm{Var}(\mathrm{v}) = \sigma_u^2[(\mathbf{I_m} - \rho\mathbf{W})(\mathbf{I_m} - \rho\mathbf{W}^{\mathrm{T}})]^{-1} = \mathbf{G}$. The $\mathbf{W}$ matrix describes how random effects from neighbouring areas are related, whereas ρ defines the strength of this spatial relationship.

The simplest way to define **W** is as a contiguity matrix. That is, the elements of **W** take non-zero values only for those pairs of areas that are adjacent. Generally, for ease of interpretation, this matrix is defined in row-standardized form in which case ρ is called the spatial autocorrelation parameter (Banerjee et al., 2004).

Formally, the element w_{jk} of a contiguity matrix takes the value 1 if area j shares an edge with area k and 0 otherwise. In row-standardised form this becomes

(2.9)
$$w_{jk} = \begin{cases} d_j^{-1} \text{ if j and k are contigious} \\ 0 \text{ otherwise} \end{cases}$$

where d_i is the total number of areas that share an edge with area j (including area j itself). Contiguity is the simplest but not necessarily the best specification of a spatial interaction matrix. It may be more informative to express this interaction in a more detailed way, e.g. as some function of the length of shared border between neighbouring areas or as a function of the distance between certain locations in each area. Furthermore, the concept of neighbours of a particular area can be defined not just in terms of contiguous areas, but also in terms of all areas within a certain radius of the area of interest.

2.4. Determinants of Maize Productivity

Braimoh and Vlek (2006) investigated the most important variables affecting maize yield as: soil quality index, fertlizer use, household size, distance from main market and the interaction between fallow length and soil quality index.

Cai et al. (2014) found that temperature tends to have negative effects on U.S. corn yields in warmer regions and positive effects in cooler regions, with spatial heterogeneity at a fine scale. The spatial pattern of precipitation effects is more complicated. In further analysis they revealed that precipitation effects are sensitive to the existence of irrigation systems.

Tsedeke et al. (2015) pointed out that the expansion and productivity change in maize production in Ethiopia is attributable to multiple factors. These include a) increased availability of modern varieties, b) increased commitment to enhance farmer access to and use of modern inputs through better research-extension linkages, c) wider adaptability of the crop and modern varieties, d) better production conditions and low production risks and e) growing consumption

demand and market access for producers to support market-based production to absorb surplus supply.

Josephson et al. (2014) also conducted an emperical study to estimate the impacts of rural population density on agricultural intensification, productivity and farm income in Ethiopia. Their study revealed that rural population density has a significant impact on maize and teff yields.

2.5. Applications of Small Area Estimation Models in Agriculture

Battese et al. (1988) used the nested error unit level regression model (2.7) for the first time to model county crop areas in USA. The authors have used the normally distributed common errors variance assumption and revealed that based on the fitting-of-constants method the estimates of errors variances are slightly different from each other. They also demonstrated some techniques for validating their model on the basis of unit level auxiliary variables. First, they introduced two quadratic terms for the two covariates they used in their model and tested the null hypothesis that the regression coefficients associated with the quadratic terms are zero. The null hypothesis was not rejected at the 5% level. Secondly, they tested the null hypothesis that the error terms $\mathbf{u}_d$ and ε_{dj} are normally distributed by using the transformed residuals $(y_{dj} - \hat{\alpha}_d \bar{y}_d) - (x_{dj} - \hat{\alpha}_d \bar{\mathbf{x}}_d)^T \hat{\beta}$ with $\hat{\alpha}_d = 1 - (1 - \hat{\gamma}_d)^{1/2}$. Under the null hypothesis, the transformed residuals are independent normal with mean 0 and variance σ_ε^2;. The well-known Shapiro-Wilk W statistic, applied to the transformed residuals, gave p-values equal to 0.921 and 0.299 for corn and soybeans respectively, suggesting the tenability of the normality assumption.

Chandra et al. (2007) used real data and design-based simulation techniques to evaluate the performance of EBLUP, MBDE, SEBLUP and SMBDE in the context of a real population and realistic sampling methods using the ISTAT farm structure survey in Tuscany. They used sample farms to generate a population of N = 22,977 farms by sampling with replacement from the original sample of 529 farms, with probabilities proportional to their sample weights. The small areas of interest are defined by the 23 local economic systems of northern Tuscany. Sample sizes in these areas were fixed to be the same as in the original sample. Their aim was to estimate average olive production in quintal (100 kg units) in each local economic system using the surface utilized for olives in hectares as the auxiliary variable. The results of their study showed

that EBLUP and SEBLUP are unstable in a few small areas, mainly because there is little or no variability in the variable of interest in these areas. In contrast, the MBDE and SMBDE methods appear unaffected by such behaviour. The median relative bias of MBDE is smaller than that of EBLUP. In contrast, the median relative root mean squared error (RRMSE) of EBLUP is smaller than that of MBDE. The median relative bias and median RRMSE of SEBLUP is marginally smaller than that of EBLUP.

Salvati et al. (2009) applied EBLUP, SEBLUP and a spatial version of the MQ predictor to estimate the average production of olives per farm in quintals for each of the small areas making up the local economic systems in Tuscany. In this application the authors employed data from the 2003 ISTAT farm structure survey, which is carried out every two years to collect information on farmland by type of cultivation, amount of animal production and structure and amount of farm employment on 55,030 farms. The GIS Atlas of Coverage of the Tuscany Region provided information on coordinates, surface area and positions of the small areas of interest. The centroid of each area is the spatial reference for all the units residing in the same small area. The auxiliary variable they employed in the models was the surface area used for olive production

Bellow and Lahiri (2011) applied an empirical best linear unbiased prediction (EBLUP) approach to estimation of crop parameters such as harvested area (and potentially other crop parameters) of county level for seven Midwestern United States for the year 2008. Their method assumes a linear mixed model that relates survey reported harvested area to both unit (farm) and area (county) level covariates, with variance components estimated using a technique which ensures strictly positive consistent estimation of the model variance. The EBLUP estimator was tested for corn and soybeans in 2008 using two auxiliary variables: 1) farm level size variable, and 2) county level FSA planted acreage figures. Synthetic estimation was used to adjust for undercoverage due to missing values of the first covariate. For each state, the EBLUP estimator was compared with four survey estimators commonly used when population level auxiliary information is available - 1) the simple ratio estimator (SR), 2) combined ratio estimator (CRE), 3) simple regression estimator (SRGE), and 4) combined regression estimator (CRGE) (Cochran, 1977). They were used official NASS county level harvested acreage estimates for 2008 as the 'gold standard' for assessing estimation accuracy in the study. Five efficiency metrics they

computed for each of the five estimate types were - average absolute deviation (AAD), average squared deviation (ASD), average absolute relative deviation (AARD), average squared relative deviation (ASRD) and percentage below official (PBO). They defined AAD as the mean of absolute deviations between county estimates and corresponding 2007 official estimates, ASD the corresponding mean of squared deviations, AARD the mean of ratios between absolute deviations and official values and ASRD the corresponding mean of squared ratios. And PBO is defined as the proportion of counties with estimate less than the corresponding 2008 official estimate. Values of PBO below (above) 0.5 suggest overestimation (underestimation) tendencies for an estimator (direction of bias). Based on five metrics they used, the result of their study found that EBLUP estimator outperforms the other five estimators in general for both crops (especially corn).

Coelho and Pereira (2011) described the design of the Monte Carlo simulation study, and present empirical results on the performance of the direct and indirect estimators using a real dataset from an agricultural survey conducted by the Portuguese Statistical Office. In particular, the authors analyzed the performance of the EBLUP with random small-area effects to present a spatial covariance structure following an isotropic exponential model. To explore the behavior of the small-area predictors the authors built a pseudo-population obtained from a real dataset containing the responses to the 1993 Agricultural Structure Survey, which is carried out by the Portuguese Statistical Office between agricultural censuses. The responses for the variable total production of cereals were extracted and circumscribed to the small areas of the Alentejo Region. The total sample size was 7,060 and the population size 47,049. Production in 1989 was used as an auxiliary variable in the models applied in the simulation. Geographical coordinates associated with the centroids of administrative divisions are recorded. The results of their simulation experiment revealed that when the data display spatial variability, the estimators that reflect the spatial correlation between observations tend to present reductions in bias and bias ratio when compared with estimators that ignore this variability. These reductions are usually accompanied by a modest loss of precision, resulting in bias ratios that are generally substantially lower than those obtained for the other estimators.

Tzavidis et al. (2012) also applied both well known small area estimators, such as the empirical best linear unbiased predictor (EBLUP), and more recently proposed small area estimators, for

example, tile M-quanlile, the robust EBLUP, and the Ansed Direct estimators using a real agricultural business survey dataset. They generated a population of N = 81982 farms by bootstraping the origional AAGIS sample.That is, the origional 1652 farms in the origional sample were themselves sampled with replacement 81982 times using selection probabilities proportional to a farm's AAGIS sample weight. An independent sample of farms was taken from this population using stratified random sampling, with regions defining the strata and with strata sample allocations equal to those in the origional AAGIS sample. They replicated the selected samples and then used for exploratory model specification.They estimated the following small area estimators; (1) the direct estimator (regional sample mean), (2) the EBLUP based on random intercepts model (RI/EBLUP), (3) the EBLUP based on random slopes model (RS/EBLUP), (4) the model-based direct estimator based on the random intercepts model (RI/MBDE), (5) the model-based direct estimator based on the random slopes model(RS/MBDE), (6) the M-quintile naïve estimator (MQ/Naïve), (7) the M-quantile CD estimator (MQ/CD), (8) the M-quintile bias corrected estimators (MQ/BC), (9) the robust EBLUP estimator based on the random intercepts model (REBLUP) and (10) the robust bias corrected EBLUP based on random intercepts model (REBLUP/BC). They also estimated MSE of estimators using analytic and bootstrap (parametric and nonparametric) estimators. For assessing different estimators they used a set of diagnostics methods. Using the diagnostic plots they noted that all model based-estimators have similar consistency with the direct estimates except for two regions. The GoF diagnostic results indicate that all model-based estimates are not statistically different from the direct estimates apart from the REBLUP and the REBLUP/BC. Taking into account the results from the coefficient of variation, the bias diagnostic plots and the GoF diagnostic, they suggested that the EBLUP estimator that is based on the random slopes model appears to provide some efficiency gains over the direct estimator.

Chapter Three

MATERIALS AND METHODS

This chapter describes the study area, types and sources of data used to analyze the results, summary of the response and auxiliary information used in model estimation, direct and model-based estimators considered in the analysis, model selection and verification method, assessment of small area estimators, and software's used for the data analysis.

3.1. Study Area

This study was conducted in the Oromia Region of Ethiopia. Oromia Region is one of the most populous administrative regions of Ethiopia and is found approximately between 3.7^{o} - 10.5^{o} latitude and 36^{o} - 45^{o} longitude. The region has 253 rural weredas and considered as the main agrarian and grain crop producing administrative regions of the country (CSA, 2013). Most agricultural activities are under small holder rain fed agriculture system which constitutes 95% of the total production (CSA, 2013). This research focused on evaluating small area estimation models by utilizing main season agricultural survey data and spatial auxiliary information in the case of small agricultural holders. The domain of estimation was administrative weredas of the specified region.

3.2. Data

3.2.1. Survey Data

Annual agricultural sample survey 2013 main season (Meher) was the main source of data used to find the target variable of the study which is yield for maize crop. The AGSS used a total of 709 enumeration areas and 14,180 households selected following a two stage stratified probability sampling technique with zones as strata. At the second stage of sampling 20 households were selected from each enumeration area, and among these, 10 households were used for crop cutting experiments. Among the 709 enumeration areas 655 enumeration areas contain holders who produced maize crop. A total of 22,222 maize fields were registered during the AGSS data collection. Therefore, in order to find the average maize yield per hectare at wereda, kebele and enumeration area level for our study, we aggregated the field data based on sample weights of the AGSS design.

3.2.2. Population Data

We used population data to obtain rural population density of each wereda, kebele and EA's of the Oromia Region. The most recent population census 2007 data obtained from central statistical agency was projected to 2013 by annual population growth rate of 2.9. The projected data was then used to calculate population density per square kilometer for each geographical area.

3.2.3. Spatial Data

Spatial polygon shapefile data at EA and kebele level for the whole Oromia Region was obtained from CSA. The shapefile data was prepared for population and housing census of 2007 by central statistical agency. The data was used to map administrative boundary, analyze the maize yield data spatially and fit the appropriate spatial model. In addition to this the administrative boundary shapefile data was used to mask satellite images in order to extract spatial related information and calculate areas used to obtain population density. Elevation, slope and aspects which provide topographic features of the entire region at the EA and kebele level were extracted from USGS DEM raster image at 20m resolution. The USGS DEM raster image was also obtained from GIS and cartography directorate, CSA.

3.2.4. Remote Sensing Data

Remotely sensed data were mainly used to extract normalized difference vegetation indexes. NDVI provides a crude estimate of vegetation health and a means of monitoring changes in vegetation over time. The possible range of values is between -1 and 1, but the typical range is between about -0.1 (NIR less than VIS for a not very green area) to 0.6 (for a very green area). Terra Moderate resolution Imaging Spectroradiometer (MODIS) Enhanced Vegetation Index (EVI) is sensitive to the amount of chlorophyll in any given pixel. The Enhanced Vegetation Index (EVI) and similar indexes are commonly used to estimate plant productivity and health in agricultural applications. The Enhanced Vegetation Index (EVI) is often employed as an alternative to NDVI because it is less sensitive to these limitations, but requires information on reflectance in the blue wavelengths, which is not available on some satellites and is difficult to extract from broadband radiation measurements. For the purpose of this study we used normalized difference vegetation indexes data that was obtained from MODIS satellite 16 day MOD13Q1 product (Didan & Huete, 2006). Data was accessed through NASA's Land Processes Distributed Active Archive Center (LP DAAC) after signing up as member. Here we used the

250x250 meter resolution MODIS EVI product, which is averaged every 16 days for the whole area of interest. NDVI were calculated from reflectance in the Near InfraRed (ρNIR : 841–876 nm) and red (ρRED : 620–670 nm) wavelengths, (Huete et al., 1994; Jiang et al., 2008).

$$\mathrm{NDVI} = \frac{\rho\mathrm{NIR} - \rho\mathrm{RED}}{\rho\mathrm{NIR} + \rho\mathrm{RED}}$$

3.2.5. Climate/Weather Variables

Climate / weather related information such as rainfall, minimum temperature; maximum temperature and precipitation for the study area was extracted from a continuous climate grid layers at the desired esri raster format through https://www.worldclim.org. The resolution of the image is 30 arc second (often referred to as 1-km spatial resolution) which is adequate to get the desired data at kebele and wereda level. For many applications, data at a fine (≤ 1 km 2) spatial resolution are necessary to capture environmental variability that can be partly lost at lower resolutions, particularly in mountainous and other areas with steep climate gradients. Climate models were created by thin-plate smoothing spline algorithm implemented in the ANUSPLIN package for interpolation, using latitude, longitude, and elevation as independent variables (Hijmans et al., 2005).

3.2.6. Agro-ecological Variables

The agro-ecological data for the Oromia Region was obtained from the ministry of agriculture and natural resources. The polygon shapefile data was intersected with enumeration area shapefiles obtained from CSA to categorize enumeration areas and rural kebeles in to six major ecological zones (Alemayehu, 2006). These ecological zones are described in Table 1.

Table 1. Major agro-ecological zones for maize in Oromia Region (Alemayehu 2006; MOA 2005)

Zone	Altitude (m)	Mean Rainfall (mm)	Temperature (^{0}C)	code
Bereha (dry-hot)	500–1500	<900	>22	1
Weinadega (dry- warm)	1500–2500	900–1000	18–20	2
Erteb Kola (sub moist warm)	500–1500	900–1000	18–24	3
Dega (cold)	2500–3500	900–1 000	14–18	4
Erteb dega (moist cold)	2500–3500	>1 000	10–14	5
Wurch (very cold or alpine)	>3500	>1 000	<10	6

3.3. Description of Variables Used for Model Estimation

After testing linearity of relation between dependent and auxiliary variables, and correlation of auxiliary variables with each other we selected only four auxiliary variables to drive model based estimators. Table 2 summarizes the dependent and auxiliary variables used to estimate model parameter.

Table 2. Summary of Dependent and Auxiliary Variables (Aggregated to Enumeration Area Level)

Variable/Description	Units	Mean	SD	Min	Max
Maize Yield (dependent variable)	Quintal per Hectare	29.30	14.80	2.15	133.41
Slope (Measures average rise in degree of the land surface)	Degree	11.22	3.96	3.34	26.08
Pop_Density (Population Density)	Number of inhabitants in one kilometer square area	162.80	114.24	10.63	667.66
NDVI (Normalized Difference Vegetation Index)	NDVI is an indices numbers measured remotely with satellites, calculated from incident and reflected solar and Photosynthetically Active Radiation (PAR)	0.23	0.078	-0.004	0.419

Factor Auxiliary Variable

Variable/Description	Code
eco_factor (Agro Ecological Zone location of enumeration area)	Factor covariate having six categories coded as 1= Bereha -(dry-hot) 2 =Weinadega- (dry- warm) 3= Erteb Kola -(sub moist warm) 4 = Dega -(cold) 5 = Erteb dega - (moist cold) 6 = Wurch - (very cold or alpine)

3.4. Small Area Estimators of Maize Yield

3.4.1. Direct Estimator

Direct estimator of maize yield at wereda level was obtained based on sample weights information of the AGSS. The annual agricultural sample survey main season was designed to collect basic quantitative information on the country's agriculture. Since the lowest reporting domain is administrative zones it doesn't provide wereda level reliable crop yield estimate due to the sample size limitation.

3.4.1.1. Direct Domain Estimator of Maize Yield

We utilized sample weights, first stage and second stage inclusion probabilities in order to find wereda direct estimates of maize yield for Oromia Region. The sample weight for each selected sample households is obtained by multiplying the inverse of first order inclusion probability with the inverse of second order inclusion probability.

The first order selection probability is the inclusion probability of primary sampling units i.e. enumeration areas within each zone. And the second order probabilities are inclusion probability of each selected agricultural household within the selected enumeration area.

Therefore, the final weight for the j^{th} household within the selected k^{th} enumeration area in d^{th} wereda was computed as:

(3.1) $$w_{dkj} = \pi_{dk}^{-1} * \pi_{kj}^{-1}$$

Where π_{dk} and π_{kj} are 1^{st} and 2^{nd} stage selection probabilities defined as $\pi_{dk} = \frac{n_h . m_{dk}}{M_h}$ and $\pi_{dkj} = \frac{h_{hk}}{H_{hk}}$, respectively.

Then,

(3.2) $$w_{dkj} = \frac{M_h H_{dk}}{n_h m_{dk} h_{dk}}$$

Where h represents zone, n_h is the total number of sample EAs successfully covered in the h^{th} zone, M_h is the total households in the h^{th} zone as obtained from the sampling frame, m_{dk} is the measure of size of the k^{th} sample EA in the d^{th} wereda obtained from the sampling frame, H_{dk} is the total number of agricultural households of the k^{th} sample EA in the d^{th} wereda obtained from a fresh listing of households at the beginning of the survey, and h_{dk} is the number of sample agricultural households successfully covered in the k^{th} sample EA in the d^{th} wereda.

For estimating yields for maize crop in Wereda d:

(3.3) $$\hat{\bar{Y}}_d = \frac{\hat{P}_d}{\hat{A}_d},$$

where $\widehat{P}_d = \sum_{k=1}^{n_d}\sum_{j=1}^{h_k} w_{dkj}\, P_{dkj}$ and $\widehat{A}_d = \sum_{k=1}^{n_d}\sum_{j=1}^{h_k} w_{dkj}\, A_{dkj}$ are total production and area of maize for wereda d respectively, k represents EAs, n_d represents total EAs sampled in d^{th} wereda, h_k total number of households covered within k^{th} sample EA.

3.4.1.2. Sampling Variance of Direct Estimator

By adopting CSA's approach, sampling variance for the estimate of maize yield for wereda d was estimated by the following formulas:

(3.4) $$\mathrm{Var}(\widehat{Y}_d) = \frac{1}{\widehat{A}_d^2}\left[\mathrm{Var}\left(\widehat{P}_d\right) + \widehat{Y}_d^2\,\mathrm{Var}\left(\widehat{A}_d\right) - 2\widehat{Y}_d\,\mathrm{Cov}\left(\widehat{P}_d,\ \widehat{A}_d\right)\right]$$

Where,

(3.5) $$\mathrm{Var}(\widehat{A}_d) = (1-f_h)\frac{n_h}{n_h-1}\sum_{d=1}^{n_h}\left(\widehat{A}_{dk} - \frac{\widehat{A}_d}{n_h}\right)^2 + f_h\sum_{k=1}^{n_d}(1-f_{dk})\left(\frac{h_{hd}}{h_h-1}\right)\sum_{j=1}^{h_{hd}}\left(\widehat{A}_{dj} - \frac{\hat{\widehat{A}}_{hd}}{h_{hd}}\right)^2$$

(3.6) $$\mathrm{Var}(\widehat{P}_d) = (1-f_h)\frac{n_h}{n_h-1}\sum_{k=1}^{n_d}\left(\widehat{P}_k - \frac{\widehat{P}_d}{n_d}\right)^2 + f_h\sum_{k=1}^{n_d}(1-f_{dk})\left(\frac{h_{dk}}{h_{dk}-1}\right)\sum_{j=1}^{h_k}\left(\widehat{P}_{kj} - \frac{\widehat{P}_{dk}}{h_{dk}}\right)^2$$

(3.7) $$\mathrm{Cov}\left(\widehat{P}_d,\ \widehat{A}_d\right) = (1-f_h)\frac{n_h}{n_h-1}\sum_{k=1}^{n_d}\left(\widehat{A}_{dk} - \frac{\widehat{A}_d}{n_d}\right)\left(\widehat{P}_{dk} - \frac{\widehat{P}_{dk}}{n_d}\right) +$$

$$f_h\sum_{k=1}^{n_d}(1-f_{dk})\left(\frac{h_{dk}}{h_{dk}-1}\right)\sum_{j=1}^{h_k}\left(\widehat{A}_{dkj} - \frac{\hat{\widehat{A}}_{dk}}{h_{dk}}\right)\left(\widehat{P}_{dkj} - \frac{\widehat{P}_{dk}}{h_{dk}}\right),$$

where f_h is average first stage probability of selection of EAs within zone obtained from the sample design, $f_{dk} = \frac{h_{dk}}{H_{dk}}$ is average second stage probability of selection of households within the k^{th} sample EA in wereda d, and ($\widehat{A}_{dkj}$, $\widehat{P}_{dkj}$) are weighted values of area and production of maize respectively from j^{th} agricultural household, k^{th} EA and d^{th} wereda.

In estimating the sampling variance by the above formula, selection of EAs within a stratum is assumed to be with replacement. By so doing the variance estimate may be slightly over estimated but it greatly simplifies the estimation procedure.

3.4.1.3. Coefficient of Variation (CV) of Direct Estimates

Coefficient of Variation (CV) in percentage of estimate of maize yield in a specific wereda d was obtained by:

(3.8) $$\mathrm{CV}\left(\hat{\bar{\mathrm{Y}}}_{\mathrm{d}}\right) = \frac{\sqrt{\mathrm{Var}(\hat{\bar{\mathrm{Y}}}_{\mathrm{d}})}}{\hat{\bar{\mathrm{Y}}}_{\mathrm{d}}} * 100$$

3.4.2. Model Based Indirect Estimators

3.4.2.1. Synthetic Estimators by Linear Fixed Effect Models

The synthetic estimator is based on assuming a (linear) model for the data so that the values of the areas that have not been sampled are estimated from the model using only information for available covariates. For the maize yield, the synthetic estimator is estimated based on the following general model:

(3.9) $$\mathbf{y}_{\mathrm{dj}} = \mathbf{x}_{\mathrm{dj}}{}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_{\mathrm{dj}},\ \mathrm{j} = 1, \dots, N_{\mathrm{d}},\ \mathrm{d} = 1, \dots, \mathrm{D},$$

where $\varepsilon_{\mathrm{dj}}$ is an area-based random error, which is normally distributed with zero mean and variance σ_{u}^2, $\mathbf{y}_{\mathrm{dj}}$ denotes maize yield for enumeration area j for wereda d, of $\mathbf{x}_{\mathrm{dj}}$ denote a vector of p auxiliary variables for enumeration areas $\mathrm{j} = 1, \dots, \mathrm{N}_{\mathrm{d}}$, and wereda $\mathrm{d} = 1, \dots, \mathrm{D}$,

In matrix notation model **(3.9)** can be rewritten as:

(3.10) $$\mathbf{y} = \mathbf{X}^{\boldsymbol{T}}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ and $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \sigma^2\mathbf{I}$.

The synthetic estimator of maize yield at wereda level was obtained by using the estimate of $\boldsymbol{\beta}$ from linear regression of the individual level sample data and computing

(3.11) $$\hat{\bar{Y}}_{\mathrm{d,SYNTH}} = \mathbf{X}_{\boldsymbol{d}}{}^{\mathrm{T}}\hat{\boldsymbol{\beta}},$$

where $\hat{\bar{\mathrm{Y}}}_{d,\mathrm{SYNTH}}$, is the synthetic estimate in wereda d and $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{T}\mathrm{y}$, is the regression coefficients estimated by maximum likelihood estimation technique.$\mathbf{x}_{\boldsymbol{d}}$ is a D x p matrix of auxiliary variables.

3.4.2.2. Synthetic Estimator Based on Spatial lag Model

An extension of the above model **(3.10)** for spatially lagged dependent variable (Dependent variable influenced by neighbors) is a simultaneous autoregressive spatial lag model. Given a spatial weight matrix **W** and auxiliary variables **X**, this model can be written:

(3.12) $$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}^{\mathbf{T}}\boldsymbol{\beta} + \varepsilon,$$

(3.13) $$(\mathbf{I_n} - \rho\mathbf{W})\mathbf{y} = \mathbf{X}^{\mathbf{T}}\boldsymbol{\beta} + \varepsilon,$$

Once again to predict y, we could pre-multiply by $(\mathbf{I_n} - \rho\mathbf{W}y)^{-1}$

(3.14) $$\mathbf{y} = (\mathbf{I_n} - \rho\mathbf{W}\mathbf{y})^{-1}\mathbf{X}^{\mathbf{T}}\boldsymbol{\beta} + (\mathbf{I_n} - \rho\mathbf{W}\mathbf{y})^{-1}\varepsilon,$$

with $[\varepsilon\varepsilon^{T}] = \sigma^2\mathbf{I}_n$, $\varepsilon \sim N(0, \sigma^2\mathbf{I}_n)$, where ρ is an autoregressive lag coefficient. Here ρ is estimated by maximum likelihood estimation, $\boldsymbol{\beta}$ and other parameters by generalized least squares subsequently. Note: ρ is estimated first by numerical optimization.

Estimating the Spatially Lagged Model with Maximum Likelihood

(3.15) $$\ln(\beta, \sigma^2, \rho) = \ln|I - \rho\mathbf{W}| - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{(y-\rho\mathbf{W}-\mathbf{X}\beta)'(y-\rho\mathbf{W}-\mathbf{X}\beta)}{2\sigma^2}$$

(3.16) $$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(I - \rho\mathbf{W})y$$

W is a (n by n) non-stochastic spatial weights matrix, with zeros on the main diagonal and non-negative values for W_{ij}, $i \neq j$. Conventionally, **W** is normalized so that rows sum to 1. It is also assumed that each observation has at least one neighbor. In addition **W** is symmetric and real, so that its eigenvalues are all real.

In sample and out of sample prediction formulae

There are two types of prediction situations: the in-sample and out-of-sample cases. In the in-sample prediction problem, we have n spatial units for which we observe the dependent variable Y as well as the independent variables X and we want to predict the value of Y at the observed sites after fitting the model which is the same as computing the fitted value of Y. In the out-of-sample case, we have two types of spatial units: the in-sample units for which we observe the dependent variable Y_S as well as the independent variable X_S and the out-of-sample units for which we only observe the independent variable X_O and we want to predict the variable Y_O from the knowledge of Y_S, X_S and X_O. In the out-of-sample case, we will further distinguish according to the number of spatial units to be predicted simultaneously: if there is only one such unit, we will talk about a single out-of-sample prediction case, otherwise about a multiple out-of-sample prediction case.

In-sample prediction formulae

We used an optimal predictor introduced by Haining (1990) and detailed by Bivand (2002). This predictor is called as the "trend-signal-noise" predictor. We used an R package spdep to obtain this predictor. It is given by

(3.17) $$\widehat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TS}} = \mathbf{X_S}\widehat{\boldsymbol{\beta}} + \hat{\rho}\mathbf{W}\mathbf{Y_S}$$

Out-of-sample prediction formulae

The trend-signal-noise predictor $\widehat{Y}^{TS}$ cannot be defined in the case of out-of-sample prediction since it requires some values of Y_O which are unobserved. However in the case of a single prediction on unit o, it is possible to compute it because of the zeros on the diagonal of **W** which yields

(3.18) $$\widehat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TS1}} = \mathbf{X_O}\widehat{\boldsymbol{\beta}} + \hat{\rho}\mathbf{W_{OS}}\mathbf{Y_S}$$

The trend-corrected strategy can be applied here because it only involves the values of X (and not Y) for the out-of-sample units

(3.19) $$\widehat{\mathbf{Y}}^{\mathbf{TC}} = (\mathrm{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{X}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}} \\ \widehat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} \end{pmatrix}$$

and

(3.20) $$\widehat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} = -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}\mathbf{X_S}\widehat{\boldsymbol{\beta}} + (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{X_O}\widehat{\boldsymbol{\beta}}$$

(3.21) $$\widehat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{X_S}\widehat{\boldsymbol{\beta}} - (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}\mathbf{X_O}\widehat{\boldsymbol{\beta}}$$

$$\text{for } (\mathbf{I} - \hat{\boldsymbol{\rho}}\mathbf{W}) = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{I_S} - \hat{\rho}\mathbf{W_S} & -\hat{\rho}\mathbf{W_{SO}} \\ -\hat{\rho}\mathbf{W_{OS}} & \mathbf{I_O} - \hat{\rho}\mathbf{W_O} \end{pmatrix},$$

where the subscripts s denotes sampled observations, o denotes out of sample observations and os denotes for both in sample and out of sample observations.

3.4.2.3. Synthetic Estimator by SAC Model

Another extension of the above model (3.9) for spatially correlated random effects and a spatial lag in the endogenous variable y is a spatial simultaneous autoregressive "SAC/SARAR" model of the form:

(3.22) $\mathbf{y}_{dj} = \rho\mathbf{W}_1\mathbf{y}_{dj} + \mathbf{X}_d\boldsymbol{\beta} + u_d$, $u_d = \lambda\mathbf{W}_2 u_d + \varepsilon$, d= 1,2,…,D, j = 1,2,…., N_d,

with $\varepsilon \sim N(0, \sigma^2\mathbf{I}_n)$.

(3.23) $\mathbf{y}_{dj} = (\mathbf{I}_n - \rho\mathbf{W}_1)^{-1}\mathbf{X}_d\boldsymbol{\beta} + (\mathbf{I}_n - \rho\mathbf{W}_1)^{-1} u_d$

(3.24) $u_d = (\mathbf{I}_n - \lambda\mathbf{W}_2)^{-1}\varepsilon$

(3.25) $\mathbf{y}_{dj} = (\mathbf{I}_n - \rho\mathbf{W}_1)^{-1}\mathbf{X}_d\boldsymbol{\beta} + (\mathbf{I}_n - \rho\mathbf{W}_1)^{-1}(\mathbf{I}_n - \lambda\mathbf{W}_2)^{-1}\varepsilon$,

where $\mathbf{y}_{dj}$ is maize yield for wereda d and observational units (EA's), λ is simultaneous autoregressive error coefficient, ρ is simultaneous autoregressive lag coefficient and $\mathbf{I}_n$ is n by n identity matrix, $\mathbf{W}_1\mathbf{y}_{dj}$ is the spatially lagged dependent variable for weights matrix $\mathbf{W_1}$. $\mathbf{W_1}$ and $\mathbf{W_2}$ are a n by n spatial weights matrices based on k =5 and k=24 nearest neighborhood distances respectively.

Here ρ and λ are found by maximum likelihood estimation, and $\boldsymbol{\beta}$ and other parameters by generalized least squares subsequently

Maximum Likelihood Estimation of SAC model parameters (3.22)
First we define,

(3.26) $\epsilon = \frac{1}{\sigma}(\mathbf{I}_n - \lambda\mathbf{W}_2)[(\mathbf{I_n} - \rho\mathbf{W}_1)\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]$,

with $\epsilon \sim N(0, \sigma^2\mathbf{I}_n)$. The corresponding determinant of the Jacobean $J \equiv \det\frac{\partial\epsilon}{\partial y}$ can be written as

(3.27) $J \equiv \det\frac{\partial\epsilon}{\partial y} = |\frac{1}{\sigma}[\mathbf{I}_n - \lambda\mathbf{W}_2]||[\mathbf{I_n} - \rho\mathbf{W}_1]|$

Employing the fact that $\epsilon \sim N(0, \sigma^2\mathbf{I}_n)$ we can write the log-likelihood for the joint distribution as

(3.28) $\ln(L) = C - \frac{n}{2}\ln(\pi\sigma^2) + \ln|\mathbf{I}_n - \rho\mathbf{W}_1| + \ln|\mathbf{I}_n - \lambda\mathbf{W}_2| - \frac{\epsilon^T\epsilon}{2\sigma^2}$

Using the expression in (3.28), we can evaluate the log likelihood for values of ρ and λ. The values of the other parameters $\boldsymbol{\beta}$ and σ^2 are calculated as a function of the maximum likelihood values of, λ and the sample data y, X. therefore the generalized least squares estimates of $\boldsymbol{\beta}$ and σ^2 are obtained as:

(3.29) $$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}(\mathbf{I_n} - \rho\mathbf{W}_1)^{\mathbf{T}}(\mathbf{I_n} - \rho\mathbf{W}_1)\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}(\mathbf{I_n} - \rho\mathbf{W}_1)^{\mathbf{T}}(\mathbf{I_n} - \rho\mathbf{W}_1)(\mathbf{I}_n - \lambda\mathbf{W}_2)y$$

(3.30) $$\varepsilon = (\mathbf{I}_n - \lambda\mathbf{W}_2)y$$

(3.31) $$\sigma^2 = (\varepsilon'\varepsilon)/n$$

Spatial weights

Creating spatial weights is a necessary step in using areal data, perhaps just to check that there is no remaining spatial pattern in residuals. The first step is to define which relationships between observations are to be given a non-zero weight that is to choose the neighbor criterion to be used; the second is to assign weights to the identified neighbor links. Neighborhoods can be defined in a number of ways: Contiguity (common boundary), Distance (distance band, K-nearest neighbors), and General weights (social distance, distance decay). Due to its suitability here we applied distance based neighbors (k nearest neighbors) for our AGSS data set. In order to determine k we fitted various spatial models using weights based on different k. After fitting these models we conducted spatial dependence diagnostics using Global Moran I and Lagrange Multiplier tests in order to see the presence of spatial dependence at those values of k. In addition to this we also used correlogrms, Moran and Geary statistics plots to see the spatial lag points for maize yield (see Appendix I-III).

Spatial Weights Styles

Once our list of neighbors has been created, we need to assign spatial weights to each relationship. The weights to be assigned can be binary or may take on a number of values between 0 and 1. In order to overcome the issue of no-neighbor observations, we followed a row standardized weighting style (W). Row standardization is used to create proportional weights in cases where features have an unequal number of neighbors which is the case in AGSS data. It is useful if we want to compare spatial parameters across different data sets with different connectivity structures. Row-standardized weights increase the influence of links from observations with few neighbors. Those with many neighbors are up-weighted compared to those with few. We used spdep R package in order to prepare k-nearest neighbor weights. The k-Nearest Neighbor Weights are described as follows:

Let centroid distances from each spatial unit i to all units $j \neq i$ be ranked as follows: $d_{ij(1)} \leq d_{ij(2)} \leq \cdots \leq d_{ij(n-1)}$. Then for each k = 1,…, n-1, the set $N_k(i) = \{j(1), j(2), \ldots, j(k)\}$ contains

the k closest units to i. For each given k, the k-nearest neighbor weight matrix, W, then has spatial weights of the form:

(3.32) $$w_{ij} = \begin{cases} 1, j \in N_k(i) \\ 0, \text{otherwise} \end{cases} \text{ (Standard form)}$$

Alternatively, one can consider a symmetric version in which positive weights are assigned to all i and j pairs for which at least one is among the k-nearest neighbors of the other:

(3.33) $$w_{ij} = \begin{cases} 1, & j \in N_k(i) \text{ or } i \in N_k(j) \\ 0, & \text{otherwise} \end{cases} \text{ (Symmetric Form)}$$

Therefore, following the above procedures we used two weights for SAC model namely $\mathbf{W_1}$ and $\mathbf{W_2}$, where $\mathbf{W_1}$ is based on nearest neighborhood distance of k=5 to weight the spatial dependence in the dependent variable, and $\mathbf{W_2}$ is based on nearest neighborhood distance of k=24 to account for the spatial dependence in the disturbance terms left over after eliminating the spatial dependence in the dependent variable. In the case of spatial lag model we used only $\mathbf{W_1}$.

3.4.3. Model Based Small Area Estimators Using EBLUP approach

Mixed-effects models are widely used to improve small area estimation. One important difference between synthetic estimation and model-based small area estimation is in the inclusion of an area specific random effect term that accounts for between area variations not explained by the auxiliary variables included in $\mathbf{X}_d$ (Datta, 2009). Spatial patterns can be accounted for by means of random effects. The EBLUP based on unit-level data is the standard tool for producing small-area estimates. As with the area-level specification, it can be extended to correlate random-area effects to obtain the SEBLUP. The chart below describes how unit level small area models are used to produce small area estimates with an optimal MSE.

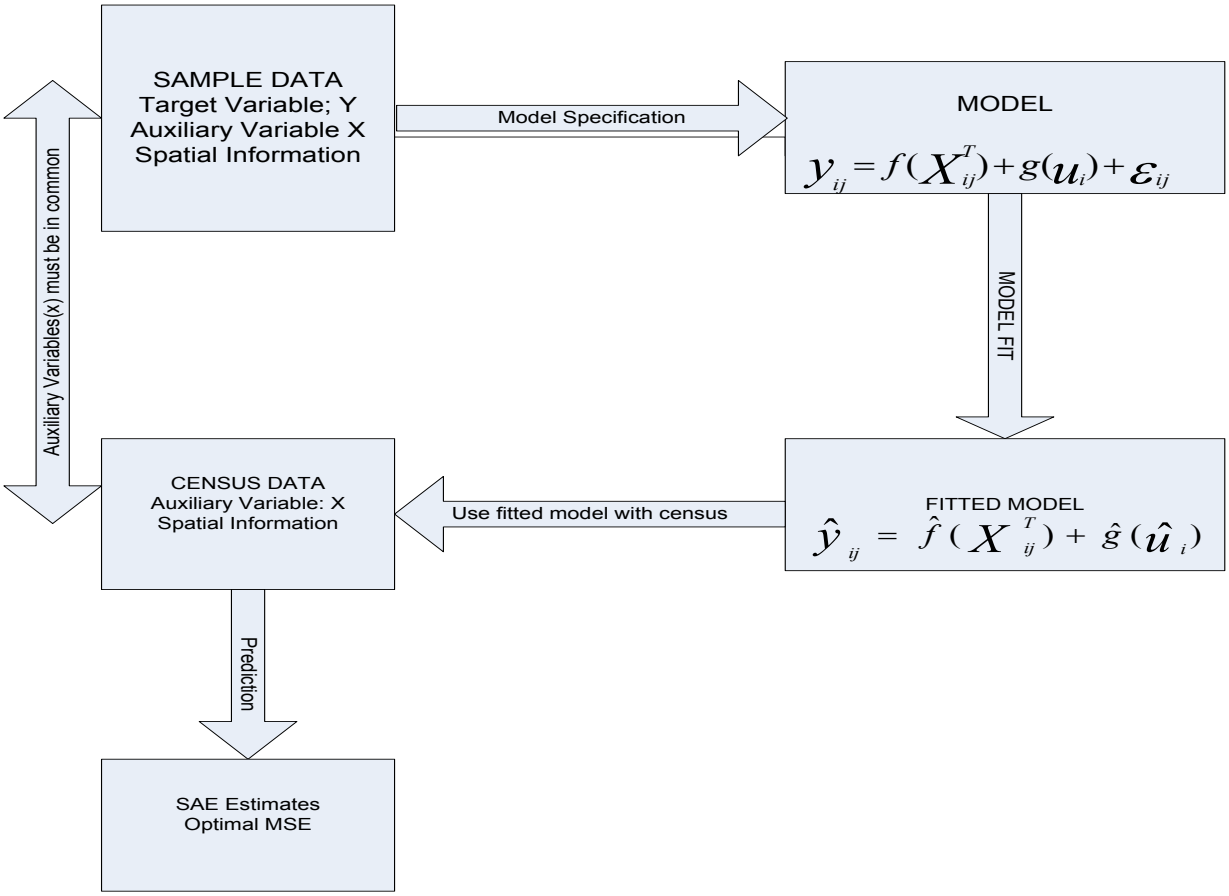


Figure 1. How an SAE unit level model works

3.4.3.1. EBLUPs Based on Mixed effects Regression Model

When auxiliary data are available at the unit level, unit-level models are likely to provide more efficient small area estimators than area level models, because they make use of the much richer information offered by micro data.

Let x_{dj} denote a vector of p auxiliary variables for each population unit j in small area d and assume that information for the variable of interest y is available only from the sample. The aim is to use the data to estimate various area-specific quantities. A popular approach is to use unit level mixed-effects models with random-area effects. The unit model originates with Battese et al., (1988). They used the nested error regression model to estimate county crop areas using sample survey data in conjunction with satellite information. Adopting this model in the case of wereda level maize yield data:

(3.34) $$y_{dj} = \mathbf{x}_{dj}^{\mathbf{T}} \boldsymbol{\beta} + \mathbf{u}_d + \boldsymbol{\varepsilon}_{dj}, j = 1, \dots, N_d, d = 1, \dots, D,$$

where y_{dj} is the observed maize yield for the j^{th} enumeration area in d^{th} wereda, $\mathbf{x}_{dj} = (1, x_{dj1}, x_{dj2}, \dots, x_{djk})^T$ is a p -vector of corresponding covariates for each EAs of wereda d , $\boldsymbol{\beta} = (\alpha, \beta_1, \dots, \beta_k)'$ is a p-vector of unknown regression coefficients, u_d denotes a random-area effect that characterizes differences in the conditional distribution of **y** given **x** between the D weredas, and ϵ_{dj} is the error term associated with the j^{th} enumeration area in d^{th} wereda (Rao, 2014).

Conventionally, $\mathbf{u_d} \sim_{\mathbf{iid}} N(0, \sigma_u^2)$, $\boldsymbol{\epsilon_{dj}} \sim_{\mathbf{iid}} N(0, \sigma_\epsilon^2)$, with $\{\mathbf{u_d}\}$ and $\{\boldsymbol{\epsilon_{dj}}\}$ mutually independent.

We assume that a sample of EAs (s_d)of size n_d is taken from the N_d units in the d^{th} wereda (d= 1,…,D) and that the sample values obey the assumed model (3.34). Then, the domain vectors y_d are independent and follow the model:

(3.35) $$y_d = \mathbf{X_d}\boldsymbol{\beta} + \mathbf{u}_d \mathbf{1}_{n_d} + \boldsymbol{\epsilon_d}, d = 1, \dots, D,$$

with $\mathbf{u_d} \sim_{\mathbf{iid}} N(0, \sigma_u^2)$, and $\boldsymbol{\epsilon_d} \sim_{\mathbf{iid}} N(0, \sigma_\epsilon^2 \mathbf{I}_{n_d})$.

Under the assumption of model (3.35) the conditional population mean is given by:

(3.36) $$\bar{\mathbf{Y}}_d = \bar{\mathbf{X}}_d^T \boldsymbol{\beta} + u_d,$$

where $\bar{\mathbf{Y}}_d$ and $\bar{\mathbf{X}}_d$ are the population means of the associated n_d observations (y_{dj}, x_{dj}) in the d^{th} sampled wereda s_d .

Then the EBLUP estimate of $\bar{Y}_d$ (Battese et al., 1988; Rao, 2003) is a composite estimate of the form:

(3.37) $$\hat{\bar{y}}_d = \hat{\gamma}_d(\bar{y}_{ds} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{ds})^T \hat{\boldsymbol{\beta}} + (1 - \hat{\gamma}_d)\bar{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}}, \ d = 1,2, \dots, D$$

where $\hat{\gamma}_d = \hat{\sigma}_u^2 / (\hat{\sigma}_u^2 + \hat{\sigma}_\epsilon^2 / n_d)$ with estimated variance components $\hat{\sigma}_u^2$ and $\hat{\sigma}_\epsilon^2$ and $\hat{\boldsymbol{\beta}}$ the weighted least square estimate of $\boldsymbol{\beta}$ which depends on $\hat{\sigma}_u^2$ and $\hat{\sigma}_\epsilon^2$. $\bar{\mathbf{X}}_\mathbf{d}$ is the vector of means of the p auxiliary variables for all EAs in d^{th} wereda, $\bar{\mathbf{x}}_{ds}$ the mean of the p auxiliary variables for the sampled EAs in d^{th} wereda and $\bar{y}_{ds}$ mean maize yield for the sampled EAs in d^{th} wereda.

The weighted least square estimate of $\boldsymbol{\beta}$ is obtained as:

(3.38) $$\hat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \mathbf{X}_d^T \hat{\mathbf{V}}_{ds}^{-1} \mathbf{X}_d\right)^{-1} \left(\sum_{d=1}^{D} \mathbf{X}_d \hat{\mathbf{V}}_{ds}^{-1} \mathbf{y}_d\right),$$

with $\widehat{\mathrm{V}}_{\mathrm{ds}} = \widehat{\sigma}_{\mathrm{u}}^2 \mathbf{1}_{\mathrm{n_d}} \mathbf{1}'_{\mathrm{n_d}} + \widehat{\sigma}_{\epsilon}^2 \mathbf{I}_{\mathrm{n}}, \mathrm{d} = 1, \ldots, \mathrm{D}$ is the estimated covariance matrix of y_d

Estimation of Variance Components

The BLUP assumes that the variance components are known. In practice of course, this is hardly ever the case. We therefore need to estimate these variance components from the sample data. The empirical best linear unbiased prediction (EBLUP) method replaces the unknown variance components in BLUP by these estimates (Prasad & Rao, 1999).

Henderson (1975) showed that substituting estimated values of variance components in the BLUP led to biased predictions. However, Kackar and Harville (1981) showed that a two-stage approach (first estimate variance components, then use these to estimate and predict fixed parameters and random components) leads to unbiased predictors provided the distribution of the data vector is symmetric about its expected value and provided the variance component estimators are translation invariant and are even functions of the data vector. They showed that the ML and REML variance component estimators have these properties.

Let us rewrite model (3.35) in matrix form as

(3.39) $$\mathbf{y} = \mathbf{X_d}\boldsymbol{\beta} + \boldsymbol{Z_d}\mathbf{u}_{\mathrm{d}} + \boldsymbol{\epsilon}_{\mathbf{d}}$$

In mixed model in particular, the aim is to predict/estimate a linear function $\mathbf{a}'\mathbf{y}$ of the population y-values given this model. The vector y in model (3.39) can be partitioned as $\mathbf{y} = (\mathbf{y}_{\mathbf{s}}', \mathbf{y}_{\mathbf{r}}')'$ where subscripts of s and r corresponding to sample and non-sample enumeration areas from all weredas.

Thus, the vector a is partitioned conformably as $\mathrm{a} = [\mathrm{a}_{\mathrm{s}}', \mathrm{a}_{\mathrm{r}}']'$. The linear function $\mathbf{a}'\mathbf{y}$ can then be written as:

(3.40) $$\mathbf{a}'\mathbf{y} = \mathbf{a}_{\mathbf{s}}'\mathbf{y}_{\mathbf{s}} + \mathbf{a}_{\mathbf{r}}'\mathbf{y}_{\mathbf{r}}$$

The first term in (3.40) depends only on the sample values and is known after the sample is observed. The second term, which depends on the non-sample values, and is unknown. The problem of using the sample values $\mathbf{y}_{\mathbf{s}}$ to predict the linear function $\mathbf{a}'\mathbf{y}$ therefore becomes the problem of predicting the linear function $\tau = \mathbf{a}_{\mathbf{r}}'(\mathbf{X}_{\mathbf{r}}\boldsymbol{\beta} + \mathbf{z}_{\mathbf{r}}\mathbf{u})$ given the linear mixed model

(3.41) $$\mathbf{y}_s = \mathbf{X}_s\boldsymbol{\beta} + \mathbf{Z}_s\mathbf{u} + \boldsymbol{\epsilon}_s,$$

where the vector of random effects u are partitioned into D subsectors $\mathbf{u}=[u_1', u_2', \dots, u_D']$, with $\mathbf{Z_s}$ partitioned conformably as $\mathbf{Z_s} = [\mathbf{z'_{1s}}, \mathbf{z'_{2s}}, \dots, \mathbf{z'_{Ds}}]$, and $\text{var}(\mathbf{u}) = \boldsymbol{\sigma^2\Omega} = \sigma^2 \text{blk} - \text{diag}(\varphi_i \Omega_i)$ where $\varphi_i = \frac{\sigma_i^2}{\sigma^2}$ is the variance component corresponding to i^{th} sampled EA from a total of n EAs from all weredas, σ_i^2 is area specific random variance for i^{th} sampled EA, σ^2 is unknown constant of proportionality. Note that: blk denotes block diagonal.

Here $\mathbf{Z_s} = \text{diag}(\mathbf{1_s})$ is a matrix that "picks out" the component of u corresponding to any particular wereda d and u is a random vector with zero mean and unknown covariance matrix **Ω.** The random errors ϵ_s are assumed to be independent of u with zero mean and variance $\sigma^2\mathbf{W_s}$.

Under this assumptions the variance covariance matrix of y_s is therefore,

(3.42) $$\sigma^2(\mathbf{W_s} + \mathbf{z_s \Omega z_s'}) = \sigma^2 \mathbf{\Sigma_s},$$

where $\mathbf{W_s} = \sigma^2 \mathbf{I_n}$, $\mathbf{I_n}$ denotes the identity matrix of order n (total number of EAs in the sample) with constant of proportionality σ^2, ω_d is the weight associated for the d^{th} wereda, $\omega' = (\omega_1, \omega_2, \dots, \omega_D)$, $\mathbf{\Omega} = \text{diag}(\omega_1, \omega_2, \dots, \omega_D)$, $\mathbf{W_s} = \sum_{d=1}^{n_d} \omega_d$.

Maximum Likelihood Estimation of Variance Components (ML)

The ML approach requires parametric specification of the distribution of the random component in a mixed model. A standard assumption is that this component is normally distributed. With this extra assumption, we can write down the log-likelihood function generated by the observation vector $\mathbf{y_s}$ under the linear mixed model (3.41) as:

(3.43) $$l = -\left(\frac{1}{2}\right)[n\ln(2\pi\sigma^2) + \ln|\textstyle\sum_{\cdot s}| + \sigma^{-2}(\mathbf{y}_s - X_s\boldsymbol{\beta})' \textstyle\sum_{\cdot \mathbf{s}}^{-\mathbf{1}} (\mathbf{y_s} - \mathbf{X_s\beta})]$$

Differentiation of this log-likelihood function with respect to the parameters **β,** σ^2 and φ_i leads to the ML score functions

(3.44) $$\partial/\partial\beta = \sigma^{-2}\mathbf{X}_s' \textstyle\sum_{\cdot s}^{-1} (\mathbf{y}_s - \mathbf{X_s\beta})$$

(3.45) $$\partial/\partial\sigma^2 = -(\tfrac{1}{2})[n\sigma^{-2} - \sigma^{-4}(\mathbf{y}_s - \mathbf{X_s\beta})' \textstyle\sum_{\cdot s}^{-1} (\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})$$

(3.46) $$\partial/\partial\varphi_i = -\left(\frac{1}{2}\right)\text{tr}(\textstyle\sum_{\cdot s}^{-1} \mathbf{Z_{si}}\Omega_i\mathbf{Z'_{si}})\text{-}\sigma^{-2}(\mathbf{y_s} - \mathbf{X_s\beta})' \textstyle\sum_{\cdot s}^{-1} \mathbf{Z_{si}}\Omega_i\mathbf{Z'_{si}} \textstyle\sum_{\cdot s}^{-1} (\mathbf{y_{ds}} - \mathbf{X_s\beta})]$$

Equating these score functions to zero yields the ML estimating equations for $\boldsymbol{\beta}$, σ^2 and φ_i. Given the MLE$_s$ for σ^2 and φ_i, and hence the MLE $\widehat{\sum}_{\boldsymbol{s}}$ for $\sum_{\boldsymbol{s}}$, it is clear that the MLE for $\boldsymbol{\beta}$ is just the GLS estimator of this parameter defined by $\widehat{\sum}_{\boldsymbol{s}}$. That is,

(3.47) $$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \left(\mathbf{X}_s'\widehat{\sum}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\widehat{\sum}_s^{-1}\mathbf{y}_s.$$

However, the estimating equations for the variance components have no analytic solution and so have to be solved numerically.

Residual Maximum Likelihood of Variance Components (REML)

For the linear mixed model (3.42), the expectations of the component score functions for $\boldsymbol{\sigma^2}$ and $\boldsymbol{\varphi_i}$ defined by (3.47) above are zero. However, if $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is substituted for $\boldsymbol{\beta}$ in these functions, then these expectations are no longer zero. For example, the expected value of the score function for $\boldsymbol{\sigma^2}$ when $\boldsymbol{\beta}$ is replaced by $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$, is $-(p/2)\sigma^2$ and so the ML estimator for this parameter is biased. On the other hand, an unbiased estimator for $\boldsymbol{\sigma^2}$ is obtained by solving the equation defined by

(3.48) $$(n\text{-}p)\sigma^{-2} - \sigma^{-4}\left(y_s - \mathbf{X_s}\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}\right)'\sum_{\cdot s}^{-1}\left(y_s - \mathbf{X_s}\boldsymbol{\beta}_{\mathrm{ML}}\right) = 0$$

where n is sample size and p is the rank of the matrix $\mathbf{X_s}$. Similarly, we can show that when $\boldsymbol{\beta}$ is replaced by $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$, the ML estimator of $\boldsymbol{\sigma^2}$ is biased. On the other hand, an unbiased estimator for $\boldsymbol{\sigma^2}$ is obtained by solving the equation defined by

(3.49) $$\frac{\partial \ln|\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s|}{\partial\varphi_i} + \frac{\partial \ln|\boldsymbol{\Sigma}|}{\partial\varphi_i} + \frac{1}{\sigma^2}\left(\mathbf{y_s} - \mathbf{X_s}\widehat{\boldsymbol{\beta}}_{\mathbf{ML}}\right)'\frac{\partial\Sigma_s^{-1}}{\partial\varphi_i}\left(\mathbf{y_s} - \mathbf{X_s}\widehat{\boldsymbol{\beta}}_{\mathbf{ML}}\right) = 0$$

The Residual Maximum Likelihood (REML) approach uses the above idea to reduce bias when estimating variance components. In particular, this approach starts by first transforming $\mathbf{y_s}$ into two independent vectors $\mathbf{y_{s1}} = \mathbf{K_1}\mathbf{y_s}$ and $\mathbf{y_{s2}} = \mathbf{K_2}\mathbf{y_s}$. The $\mathbf{y_{s1}}$ vector has a distribution that does not depend on the fixed effect $\boldsymbol{\beta}$ and hence satisfies $\mathbf{E}\ (\mathbf{K_1}\mathbf{y_s}) = 0$, i.e. $\mathbf{K_1}\mathbf{X_s} = 0$, while the $\mathbf{y_{s2}}$ vector is independent of $\mathbf{y_{s1}}$ and satisfies $\mathbf{K_1}\boldsymbol{\Sigma_s}\mathbf{K_2'} = 0$. The matrix $\mathbf{K_1}$ is chosen to have maximum rank, i.e. n-p, and so the rank of $\mathbf{K_2}$ is p. The likelihood function of y_s is then the product of the likelihoods of $\mathbf{y_{s1}}$ and $\mathbf{y_{s2}}$. REML estimators of variance components are maximum likelihood estimators based on $\mathbf{y_{s1}}$. That is, the REML method estimates variance components by maximizing the log-likelihood function defined by $\mathbf{y_{s1}} = \mathbf{K_1}\mathbf{y_s}$,

(3.50) $$l_{REML} = -\left(\frac{1}{2}\right)[n(n-p)\ln(2\pi\sigma^2) + |\mathbf{K_1\Sigma_s K_1}| + \sigma^{-2}\mathbf{y_s' K_1(K_1\Sigma_s K_1)^{-1} K1 y_s}],$$

where the matrix $\mathbf{K_1} = \mathbf{W_s^{-1}} - \mathbf{W_s^{-1}X_s(X_s'W_s^{-1}X_s)^{-1}X_s^{-1}X_s'W_s^{-1}}$. Note that if $\mathbf{K_1\Sigma_s K_1}$ is not of full rank, then $|\mathbf{K_1\Sigma_s K_1}|$ must be interpreted as the determinant of its linearly independent rows and columns.

Given this definition of $\mathbf{K_1}$, the matrix $\mathbf{K_2}$ is defined as $\mathbf{K_2} = \mathbf{X_s'\Sigma_s^{-1}}$. The log-likelihood function defined by $\mathbf{y_{s2}} = \mathbf{K_2 y_s}$ is

(3.51) $$l_L = -(\tfrac{1}{2})[p\ln 2\pi\sigma^2 + \ln|\mathbf{K_2\Sigma_s K_2'}| + \sigma^{-2}\big(\mathbf{y_{s2}} - \mathbf{E(y_{s2})}\big)'(\mathbf{K_2\Sigma_s K_2'})^{-1}\big(y_{s2} - E(y_{s2})\big)]$$

$$=-(1/2)[p\ln 2\pi\sigma^2 + \ln\left|\mathbf{X_s'\Sigma_s^{-1}X_s}\right| + \sigma^{-2}(\mathbf{y_s - X_s\beta})'\mathbf{\Sigma_s^{-1}X_s(X_s'\Sigma_s^{-1}X_s)^{-1}X_s\Sigma_s^{-1}(y_s - X_s\beta)}]$$

For given values of the variance components, $\boldsymbol{\beta}$ is estimated by maximizing l_L, leading to

(3.52) $$\widehat{\boldsymbol{\beta}} = ((\mathbf{X_s'\Sigma_s^{-1}X_s})^{-1}\mathbf{X_s'\Sigma_s^{-1}y_s}.$$

Estimation of Mean Square Error (MSE)

Analytic MSE estimator sometimes leads to under/over coverage (depending on the method used to estimate the MSE). To overcome this problem, bootstraps techniques have been proposed in literature (Hindmarsh, 2013). Bootstrap estimates of the MSE are more stable. Often, in practical application the parametric bootstrap is used. The parametric bootstrap is a resampling-based method of estimating the MSE. That is, an estimate of the MSE of the original data is obtained by accessing the variability between the replicates created by re-sampling and re-fitting the model to each replicate sample. R function pbmseBHF gives a parametric bootstrap MSE estimate for the EBLUP under the BHF model (3.34). The function applies the parametric bootstrap procedure for finite populations introduced by Gonzalez-Manteiga et al. (2008) particularized to the estimation of means.

Model Assumptions: The main assumptions of the unit level mixed effect model are the linearity of the relation between the dependent and the auxiliary variables and the independence of random area effects (Rao, 2003).

Out-of-sample predictions

In the mixed model above the synthetic mean predictor for out-of-sample wereda i is based on the estimated parameters of the linear mixed effects model and on the X auxiliary information:

(3.53) $$\hat{\vartheta}_i^{MX/SYNTH} = N_i^{-1} \sum_{j \epsilon U_i} \mathbf{X}_{ij}^T \hat{\boldsymbol{\beta}},$$

where u_i denotes list of all enumeration areas from the sampling frame for i^{th} wereda having a sample size of zero in the survey and j denotes enumeration area.

Note: that all variation in the area-specific predictions comes from the area-specific auxiliary information. The conventional synthetic estimation for out-of-sample areas can potentially be improved by using a model that borrows strength over space.

3.4.3.2. EBLUPs based on a spatial Fay-Herriot Model

One popular approach to small area estimation when data are spatially correlated is to employ Simultaneous Autoregressive Regressive (SAR) random effects models to define an extension to the Empirical Best Linear Unbiased Predictor namely, the Spatial Empirical Best Linear Unbiased Predictor (SEBLUP) (Salvati & Petrucci, 2008). Spatial models are special cases of the general linear mixed model defined in section (3.3.3.1). The objective is to estimate domain parameters $\delta_d = h_d(\mathbf{y}_d), d = 1, \dots, D$, based on a FH model with spatially correlated area effects.

Let the deviations v from the fixed part of the model $\mathbf{X\beta}$ be the result of an autoregressive process with parameter ρ and proximity matrix **W**, then $\mathbf{v} = \rho\mathbf{W}v + \mathbf{u} \Rightarrow \mathbf{v} = (\mathbf{I}_D - \rho\mathbf{W})^{-1}\mathbf{u}$, where $\mathbf{u} = (u_1, \dots, u_D)'$ satisfies $u \sim N(0_D, A\mathbf{I}_D)$ for A unknown. We assume that the matrix $(\mathbf{I}_D - \rho\mathbf{W})$ is non-singular. Hence $(\mathbf{I}_D - \rho\mathbf{W})^{-1}$ exists, where $\mathbf{I}_D$ denotes the D × D identity matrix, ρ is an autoregression parameter $\in [-1, 1]$ and **W** is proximity matrix. We consider that the proximity matrix **W** is defined in row standardized form; that is, **W** is row stochastic. The model with spatially correlated errors can be expressed as:

(3.54) $$\mathbf{y} = \mathbf{X\beta} + \mathbf{D}(\mathbf{I_D} - \boldsymbol{\rho}\mathbf{W})^{-1}\mathbf{u} + \varepsilon$$

(3.55) $$\mathbf{y} = \mathbf{X\beta} + \mathbf{Dv} + \varepsilon,$$

where ε is independent of **v**, **y** is the D x 1 vector of the direct estimates of maize yield for D weredas, $\mathbf{X} = (x_1, \dots, x_D)'$ is a Dxp matrix containing in its columns the values of p auxilary variables for the D weredas, $\mathbf{v} = (v_1, \dots, v_D)'$ is the vector of area effects and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_D)'$ is

the vector of independent sampling errors, independent of **v**, with $\varepsilon \sim N(o_D, \psi)$, where the covariance matrix $\psi = diag(\psi_1, \ldots, \psi_D)$ is estimated from AGSS data. Under the above model, the spatial best linear unbiased predictor of the small-area mean and its empirical version – SEBLUP – are obtained following Henderson (1975). In particular, the SEBLUP of the wereda-level mean maize yield is:

(3.56) $$\hat{\vartheta}_d^{MX|SAR} = N_d^{-1}\left[\sum_{j \in s_d} y_{dj} + \sum_{j \in r_d} \hat{y}_{dj}\right],$$

where $\hat{y}_{dj} = \boldsymbol{X_{dj}^T}\,\widehat{\boldsymbol{\beta}} + \mathbf{d_{dj}}\hat{\mathbf{v}}_d$, $\hat{\mathbf{v}}_d = \mathbf{b}_d^T\widehat{\mathbf{G}}\mathbf{D}\widehat{\mathbf{V}}(y - \mathbf{X}\widehat{\boldsymbol{\beta}})$,

(3.57) $$\widehat{\mathbf{G}} = \hat{\sigma}_u^2\,[(\mathbf{I}_D - \hat{\rho}\mathbf{W}^T)(\mathbf{I}_D - \hat{\rho}\mathbf{W})]^{-1},$$

(3.58) $$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}y,$$

(3.59) $$\widehat{\mathbf{V}} = \sigma_\varepsilon^2 I_n + \mathbf{DGD}^T,$$

The error term **v** have the D x D SAR covariance matrix: $\widehat{\mathbf{G}}(\delta) = \hat{\sigma}_u^2\,[(\mathbf{I}_D - \hat{\rho}\mathbf{W}^T)(\mathbf{I}_D - \hat{\rho}\mathbf{W})]^{-1}$ and the covariance matrix of $\hat{\vartheta}$ in equation (3.56) is given by $\widehat{\mathbf{V}}(\delta) = \sigma_\varepsilon^2 I_n + \mathbf{DGD}^T$, where $\delta = (\sigma_u^2, \rho)$, $(\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2)$ are asymptotically consistent estimators of the parameters obtained by ML or REML estimation.

The SEBLUP of the small-area mean $\hat{\vartheta}_d^{MX|SAR}$ in equation (3.56) is then:

(3.60) $$\hat{\vartheta}_d^{MX|SAR}(\hat{\delta}) = \mathbf{x}_i\hat{\boldsymbol{\alpha}} + \mathbf{b}_i^T\widehat{\mathbf{G}}\,\mathbf{D}^T\{\sigma_\varepsilon^2 I_n + \mathbf{DGD}^T\}^{-1}(\widehat{\boldsymbol{\vartheta}} - \mathbf{X}\hat{\boldsymbol{\alpha}}),$$

where $\hat{\boldsymbol{\alpha}} = (\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\widehat{\boldsymbol{\vartheta}}$ and and $\mathbf{b}_d^T$ is 1 x m vector (0,0,…,1,…,0) with value 1 in the d^{th} position. The predictor is obtained from Henderson's (1975) results for general linear mixed models (LMMs) involving fixed and random effects.

In the SEBLUP estimator the value of $\hat{\delta} = (\hat{\sigma}_u^2, \hat{\rho})$ is obtained by maximum likelihood(ML) or restricted maximum likelihood (REML) methods based on the normality assumption of the random effects(Singh et al., 2005; Pratesi and Salvati, 2008).

The spatial random effects can follow different structures, but in general they are relied on the adjacency between the areas. For example, neighbouring regions should have a higher

correlation than regions that are further apart. The spatial random effects are modeled at the area level.

The proximity matrix for the above model is obtained using Arc Gis version 10.2 by inverse weighting method. The inverse weighting method assigns larger weights for nearby points and smaller weights for points which are far apart from the sampled areas. In order to find the XY coordinate points of the sampled points center of the polygons were used as a representative for that sampled wereda.

Table 3. Weighting matrix for the first ten weredas

WE_ID	40101	40102	40103	40104	40105	40106	40107	40108	40109	40110	….
40101	0.0000	0.0743	0.0000	0.0427	0.1436	0.0513	0.0588	0.0846	0.0804	0.0537	….
40102	0.0427	0.0000	0.0352	0.0509	0.0860	0.0906	0.0577	0.1005	0.0283	0.0728	….
40103	0.0000	0.0231	0.0000	0.0574	0.0165	0.0378	0.0242	0.0209	0.0000	0.0327	….
40104	0.0427	0.0509	0.0574	0.0000	0.0402	0.0842	0.0398	0.0395	0.0000	0.0747	….
40105	0.1008	0.1049	0.0306	0.0402	0.0000	0.0541	0.0537	0.0969	0.0427	0.0533	….
40106	0.0237	0.0727	0.0462	0.0842	0.0356	0.0000	0.0492	0.0511	0.0000	0.0928	….
40107	0.0211	0.0361	0.0231	0.0398	0.0275	0.0383	0.0000	0.0616	0.0214	0.0777	….
40108	0.0395	0.0815	0.0258	0.0395	0.0644	0.0516	0.0799	0.0000	0.0305	0.0681	….
40109	0.0606	0.0372	0.0000	0.0000	0.0459	0.0000	0.0450	0.0494	0.0000	0.0363	….
40110	0.0203	0.0478	0.0326	0.0747	0.0287	0.0759	0.0816	0.0551	0.0182	0.0000	….

……

For out of sampled areas, spatial EBLUP of θ_i is

(3.61) $$\hat{\theta}_i^{\text{spatial EBLUP}} = \mathbf{X}_{ij}^T \hat{\boldsymbol{\beta}}^s$$

An approximately unbiased estimator of the MSE of the SEBLUP (3.61) is under normality of random effects and errors, the MSE of the Spatial EBLUP can be decomposed as:

(3.62) $$\widehat{\text{MSE}}\left(\hat{\theta}_d^{\text{spatial EBLUP}}\right) = g_{1d}^{(s)}(\hat{\boldsymbol{\psi}}) + g_{2d}^{(s)}(\hat{\boldsymbol{\psi}}) + 2g_{3d}^{(s)}(\hat{\boldsymbol{\psi}}) - \mathbf{B}_d^{(s)T}(\hat{\boldsymbol{\psi}})\nabla g_{1d}^{(s)}(\hat{\boldsymbol{\psi}})$$

Where the first term$g_{1d}^{(s)}(\hat{\boldsymbol{\psi}})$ is due to the estimation of random area effects and is of order O(1) while the second term is due to the estimation of $\boldsymbol{\beta}$ and is of order $O(D^{-1})$ for large D. The third term $g_{3d}^{(s)}(\hat{\boldsymbol{\psi}})$ is due to the estimation of the variance component. Finally, the last term

$\mathbf{B}_d^{(s)T}(\hat{\boldsymbol{\psi}})\nabla g_{1d}^{(s)}(\hat{\boldsymbol{\psi}})$ is the bias when ML method of estimation is used for variance component. This term is negligible and thus ignored when REML or method of moment is used for parameter estimation. Various terms of (3.62) are:

(3.63) $$g_{1d}^{(s)}(\hat{\boldsymbol{\psi}}) = a_d^T(\hat{\Omega} - \Omega_z^T\hat{V}_z^{-1}\hat{\Omega})a_d,$$

(3.64) $$g_{2d}^{(s)}(\hat{\boldsymbol{\psi}}) = (\mathbf{X}_d^T - C_d^T\mathbf{X})(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}(\mathbf{X}_d^T - C_d^T\mathbf{X})^T,$$

(3.65) $$g_{3d}^{(s)}(\hat{\boldsymbol{\psi}}) = \mathrm{tr}\{(\nabla_{C_d}^T)\hat{\mathbf{V}}(\nabla_{C_d})\hat{\mathbf{V}}(\hat{\boldsymbol{\psi}})\},$$

where $C_d^T = a_d^T\hat{\boldsymbol{\Omega}}_Z^T\hat{\mathbf{V}}^{-1}$, $\nabla_{C_d}^T = \frac{\partial C_i^T}{\partial \boldsymbol{\psi}}$, $\hat{V}(\hat{\boldsymbol{\psi}})$ is the estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\psi}}$ defined by the inverse of the relevant observed information matrix and $\mathbf{B}_d^{(s)T}(\hat{\boldsymbol{\psi}})\nabla g_{1d}^{(s)}(\hat{\boldsymbol{\psi}})$ is the bias correction due to ML estimator of $\boldsymbol{\psi}$.

The percentage standard error (% RMSE) of the estimator $\hat{\theta}_d$ in wereda d is calculated by:

(3.66) $$\%\mathrm{RMSE}_d = 100 \text{ x } \frac{\mathrm{RMSE}(\hat{\theta}_d)}{\hat{\theta}_d} \text{ ; } d = 1, \dots, D.$$

Model assumptions: This model assumes the stationarity of spatial correlation given the contiguity matrix.

3.5. Model Selection and Diagnostics

Model-based small area estimation heavily depends on the validity of the assumed model for the sample data. It is therefore important to use appropriate methods for model selection and then do checking of the selected model through residual analysis, influential diagnostics, etc. Under this section we describe the procedures followed in model selection and validation of fitted models.

3.5.1. Spatial Model Selection and Goodness of fit Diagnostics

We started by estimating an initial model $\mathbf{y}_{dj} = x_{dj}^T\boldsymbol{\beta} + u_{dj}$, which goes in line with the stated assumption of model (3.9). Second, we followed the standard approach towards detecting the presence of spatial dependence in the fitted regression model by apply diagnostic tests. The best known test statistics against spatial autocorrelation are, Moran's I and Geary's C statistic for spatial autocorrelation applied to the regression residuals. Thirdly, on the basis of the estimated model we used more specific formal tests, Lagrange Multiplier (LM) tests, on the basis of model (3.9) are conducted to test in favor of a selection of the appropriate model, SAR, SEM, SDM or

SAC models. If the null hypothesis under each LM tests of no spatial correlation is rejected, then spatial dependence matters and an appropriate spatial model should be estimated.

The Moran I and Geary C tests for spatial error autocorrelation are general tests; but the LM tests are more specific. They provide a basis for choosing an appropriate spatial regression model. We conducted five LM tests in order to fit an appropriate spatial regression model namely, LM error, LM lag, RLM error, RLM lag and SARMA tests. Significance of LM error, LM lag, SARMA tests points to a spatial error model, spatial lag model, and a spatial simultaneous autoregressive SAC model, respectively, while significance of RLM error test and RLM lag test points to a mixed spatial error and a spatial lag models, respectively.

Based on the model selection procedures described above we have fitted the spatial lag and the spatial Autoregressive Regressive (SAC) models. The next step is to conduct significance tests on fitted spatial regression models and hypothesis tests on the assumptions of the model. This is achieved through Likelihood ratio tests, Wald tests and Lagrange Multiplier tests.

3.5.2. Model Selection and Diagnostics for Mixed effect Linear Models

Model selection and diagnosis for the model based estimators (linear mixed-effect model) have been conducted based on the minimum cAIC criteria. The cAIC is applicable to mixed models where the focus is on prediction at the level of clusters or areas (Vaida and Blanchard, 2005). It is defined as cAIC = −2 L + 2p, where L is the conditional log-likelihood and p a penalty based on a measure for the model complexity. In the case of a fixed effects model, p is the number of model parameters. The random part of a mixed model also contributes to the number of model degrees of freedom p with a value between 0 in the case of no domain effects (i.e. $\hat{\sigma}_u^2 = 0$) and the total number of domains D in the case of fixed domain effects (i.e. $\hat{\sigma}_u^2 \to \infty$). In the expression of the cAIC, p is the effective degree of freedom of the mixed model and is defined as the trace of the hat matrix H, which maps the observed data to the fitted values, i.e. $\hat{y}$= Hy. When comparing models, the one with the lowest cAIC value is preferred. In addition to cAIC we employed ANOVA method to compare the null and full model by refitting the models by maximum likelihood method. The intra-class correlation coefficient was also another indicator of the appropriateness of the fitted mixed effect model. Therefore we computed this coefficient to see the appropriateness of the random component in our model.

Moreover, the map based cluster and outlier analysis was also conducted to see the spatial dependence of maize yield for the sample data. As a result of these tests a Spatial EBLUP estimator based in a SAR specification, Petrucci and Salvati (2006) was used to model the observed spatial dependence.

3.6. Assessment of Estimators

Assessment of small area estimators was conducted to validate the reliability of estimators generated under design and model based approaches. They are used to investigate if the model based estimates are less extreme when compared to the direct survey estimates. It demonstrated the typical SAE outcome of shrinkage more extreme values towards the mean. In this study we employed three approaches namely area specific measures, global measures, and diagnostic methods. The estimators used in our analysis are: Direct Estimator (DE), Spatial Lag Synthetic Estimator (SYN_SLM), Spatial Autoregressive Regressive Synthetic Estimator (SYN_SACLM), Unit Level Empirical Best Linear Unbiased Predictor (EBLUP_B), and Area Level Spatial Empirical Best Linear Unbiased Predictor (SEBLUP_A). In the following sections, we describe an overview of these evaluation approaches aimed to select the best wereda-level maize yield estimator.

3.6.1. Area Specific Measures

The aim of this approach consists in identifying the best estimator of wereda-level maize yield estimator comparing CVs and RRMSE values relating to indirect estimators considered. The MSE of estimators were computed by following the methodology described in previous sections.

(3.67) % Relative Root Mean Squared Error: RRMSE = $\left[\sqrt{\frac{1}{D}\sum_{d=1}^{D}\frac{MSE_d}{\hat{Y}_d}}\right]$ *100, d= 1, 2,…, 238,

where MSE and $\hat{Y}_d$ refers to the estimated mean squared error and the corresponding estimate of maize yield for wereda d by one of the estimators respectively.

3.6.2. Global Measures

The second approach involves methods that supply global measures averaged over small areas to evaluate empirical properties of estimators considered; for this aim, we employed the Monte Carlo Simulation techniques. A Monte Carlo analysis can also be very useful for determining small-sample bias in an estimator and, by setting N large, for determining that an estimator is actually consistent (Inglese et al., 2008). Therefore, under customary repeated sampling, R = 500

samples, each of size n = 709 equal in size as the original AGSS, from the overall population of N = 25,249 enumeration areas were selected by simple random sampling from each stratum. The stratum sizes allocations were fixed with same size as the original AGSS.

Since the true population value of maize yield is unknown given there is no recent agricultural census conducted, we created a synthetic population by replicating the household level survey data by their corresponding sampling weight. The Global bias diagnostic measures considered are, $\overline{\mathrm{ARB}}$, $\overline{\mathrm{ARE}}$ and $\overline{\mathrm{EFF}}$.

From each simulated sample, the following evaluation criteria have been computed (Michele et al., 2011).

(3.68) % Average Absolute Relative Bias: $\overline{\mathrm{ARB}} = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{R}\sum_{r=1}^{R}\frac{\hat{Y}_{dr}-Y_d}{Y_d}\right|$

(3.69) % Average Relative Error: $\overline{\mathrm{ARE}} = \frac{1}{\mathrm{D}}\sum_{d=1}^{D}\frac{1}{R}\sum_{r=1}^{R}\left|\frac{\hat{Y}_{dr}-Y_d}{Y_d}\right|$

(3.70) % Average Relative Efficiency: $\overline{\mathrm{EFF}} = \sqrt{\overline{\mathrm{MSE}}(\hat{\mathrm{Y}}_{\mathrm{DE}})/\overline{\mathrm{MSE}}(\hat{\mathrm{Y}}_{\mathrm{dr}})}$

where $\hat{\mathrm{Y}}_{\mathrm{dr}}$ refers to the estimated yield for wereda d and r^{th} simulated sample (r= 1,2,…R) by one of the estimators and $\mathrm{Y_d}$ the corresponding synthetic population estimate of maize yield for wereda d, $\overline{\mathrm{MSE}}(\hat{\mathrm{Y}}) = \frac{1}{\mathrm{D}}\sum_{\mathrm{d=1}}^{\mathrm{D}}\frac{1}{\mathrm{R}}\sum_{\mathrm{r=1}}^{\mathrm{R}}(\hat{\mathrm{Y}}_{\mathrm{dr}} - \mathrm{Y_d})^2$ is the average relative mean squared error of the model based estimator and $\overline{\mathrm{MSE}}(\hat{\mathrm{Y}}_{\mathrm{DE}})$ is the average relative mean squared error of the direct estimator. Here for model comparison purpose we used D (238) weredas where the direct survey estimates were available.

Note that ARB measures the bias of an estimator, whereas $\overline{\mathrm{EFF}}$ and ARE both measure the accuracy of an estimator.

3.6.3. Diagnostic Methods

The last evaluation approach considered use observations relating to all small areas gathered from one sample and provide an appraisal of bias. These methods may be considered as preliminary internal evaluation and are useful when there are some doubts about the assumptions underpinning small area model (Inglese et al., 2008). The crucial assumption here is that the

direct estimates of the small area values of interest are unbiased and the confidence intervals associated with these estimates achieve their nominal coverage levels. Below we describe only GoF diagnostic we used among three diagnostic methods available under this approach.

3.6.3.1. Goodness of fit Diagnostic

The aim of goodness of fit diagnostic is to check for unconditional bias in the model-based estimates; for that we use a Wald goodness of fit statistic, W, to test whether there is a significant difference between the expected values of the direct estimates and the model-based estimates. Since small area model-based estimates and direct estimates will be usually approximately uncorrelated, we can express W (Michele et al, 2011) as:

(3.71) $$W = \sum_{d=1}^{D} \frac{[\hat{Y}_d^{DE} - \hat{Y}_d^{MB}]^2}{\widehat{VAR}(\hat{Y}_d^{DE})p + \widehat{VAR}(\hat{Y}_d^{MB})m},$$

where $\hat{Y}_d^{DE}$, $\hat{Y}_d^{MB}$ are direct and model based estimates of maize yield for wereda d respectively, $\widehat{VAR}(\hat{Y}_d^{DE})p$, $\widehat{VAR}(\hat{Y}_d^{MB})m$ are variances of direct and model based maize yield estimators for wereda d respectively.

Under the hypothesis that the model-based estimates are equal to the expected values of the direct estimates, and provided the sample sizes in the small areas are sufficient to justify central limit assumptions, it is possible to conduct a parametric significance test of bias of model-based estimates relative to their precision, since W will then have a χ^2 distribution with degrees of freedom equal to the number of small areas in the population.

3.7. Software Used

The R statistical package version 3.3.1 for 64 bit windows has been used to estimate model parameters, mean squared error of estimators, model selection, diagnostics, model significance tests, graphical plots and other statistical analysis (R Core Team, 2016). Other statistical package software's, such as SPSS version 20, Excel 2010, Access 2010 has also been used for data preparation, data cleaning, imputation, summarizing, and joining auxiliary information with AGSS data. GIS software's namely: Arc GIS 10.2, ERDAS IMAGINE 9.1, ENVI 4.5, and IDRISI SILVA 17.0 were also used to process spatial and remote sensing data.

Chapter Four

RESULTS AND DISCUSSIONS

In this study, design based direct estimator and model based indirect estimators was considered for comparison to see how reliable and optimal in estimating maize yield at wereda level. The direct estimator was obtained by utilizing sample weights of the AGSS design. The synthetic estimators was based on fixed effect linear model assumptions between the response and auxiliary variables and area level spatial autocorrelation. The EBLUP_B estimator based on mixed effects model accounted for the area level random effects in to the random component of the model. The spatial SEBLUP extension of the mixed effect model was included to overcome the disadvantages of the spatial autocorrelations detected in the random components of the model.

Among various auxiliary variables considered to have significant effect on maize yield only, eco_factor, slope, Pop_density and NDVI were used to drive model based estimators. The main reasons for excluding other auxiliary variables such as rain fall, precipitation, temperature, soil type and agricultural area were collinearity between variables and non-linear relationships with the dependent variable.

Table 4 summarizes the estimated parameters of the four models considered for this study and significance of their coefficients at $\alpha = 0.05$ level.

4.1. Summary of Estimated Model Parameters

Table 4. Estimated model parameters and significance tests

Fixed Effects	SYN_ SLM	SYN_ SACLM	EBLUP_A	SEBLUP_B
Intercept				
Coefficient	18.384	15.571	28.485	22.356
Std. Error	3.579	4.784	3.798	4.433
z-value	5.137	3.254	-------	--------
t-value	--------	--------	7.499	5.0426
P-value	0.000*	0.001*	0.000*	0.000*
eco-factor				
Coefficient	-2.084	-2.077	-2.154	-1.497
Std. Error	0.973	1.083	1.062	1.400
z-value	-2.142	-1.917	-------	--------
t-value	---------	--------	-2.029	-1.069
P-value	0.032*	0.055	0.043*	0.285
Slope				
Coefficient	-0.276	-0.197	-0.304	-0.441
Std. Error	0.137	0.152	0.149	0.295
z-value	-2.001	-1.303	-------	-------
t-value	--------	-------	-2.040	-1.494
P-value	0.044*	0.192	0.042*	0.135
Pop-Density				
Coefficient	0.013	0.018	0.013	0.026
Std. Error	0.005	0.006	0.005	0.009
z-value	2.774	3.137	--------	-------
t-value	--------	--------	2.335	2.739
P-value	.005*	0.002*	0.020*	0.006*
NDVI				
Coefficient	22.671	25.224	34.152	38.848
Std. Error	6.991	8.828	8.243	14.019
z-value	3.243	2.857	--------	--------
t-value			4.143	2.771
P-value	0.001*	0.004*	0.000*	0.005*
		Random Effects		
$\hat{\sigma}_{\varepsilon}^2$	181.25	179.63	153.36	122.855
$\hat{\sigma}_{u}^2$			54.83	32.453
$\hat{\rho}$	0.424	0.438		0.827
$\hat{\lambda}$		0.313		

Significance Codes: 0.05 '*'

Model Fit Statistics

Statistics	Estimators			
	SYN_ SLM	**SYN_ SACLM**	**EBLUP_A**	**SEBLUP_B**
AIC	5292.5	5283.3	5317.8	1870.10
Loglike	-2639.23	-2633.66	-2651.89	-928.37
BIC			5345.12	1894.40
AIC for lm	5355.6	5355.6		
Lambda (for lm)	0	*0*		
Estimation Method	ML	ML	REML	REML

4.2. Assessment of fitted Models

Under this section we describe different types of diagnostics to validate the reliability of the model based estimators. Those are the diagnostics used to verify if the model assumptions were satisfied, the diagnostics for appropriateness of the fitted models and the diagnostics for the quality of the small area estimates.

4.2.1. Model Diagnostics Results for Synthetic Estimators

Synthetic estimators we used for wereda level maize yield estimation were spatial linear models. We assumed that our observational units (EAs) have the same characteristics as the weredas. We started by fitting the standard linear regression model, and after performing various spatial autocorrelation tests, we selected an appropriate spatial model. Under this section we presented the pre-estimation and post estimation diagnostics conducted to check for the appropriateness of fitted spatial models.

4.2.1.1. Pre-estimation Tests

In order to choose between standard linear regression, simultaneous autoregressive error, spatial lag, spatial simultaneous autoregressive SAC models, we tested for spatial dependence in the residuals of the fitted standard regression model by use of three tests namely, Moran's Global I, Geary's C and Lagrange Multiplier diagnostics tests. If the disturbances are spatially correlated, the assumption of a spherical error covariance matrix $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}.\boldsymbol{\varepsilon}') = \boldsymbol{\sigma^2}.\mathbf{I}$ is violated.

Global test for spatial autocorrelation

The standardized Moran coefficient follows a standard normal distribution under the null hypothesis of no spatial dependence. As shown in Table 5 we applied Moran I statistic under normality for testing spatial autocorrelation in the regression residuals. The result shows a strong and significant spatial dependence in the residuals. Similarly testing the same hypothesis using

Geary's C statistics also confirmed there is a significant spatial autocorrelation in the residuals of the standard linear model. The cause of spatial dependence is unspecified, i.e. the underlying spatial process is not specified. Thus the Moran I test and Geary C test are general tests for detecting spatial autocorrelation.

Table 5. Global Spatial Autocorrelation test for Standard Linear Model

Test	Statistic	P-value
Moran I	0.209	< 2.2e-16*
Geary C	0. 778	<2.2e-16*

Significance Codes: 0.05 '*'

Lagrange Multiplier test for Spatial Dependence

Moran I test and Geary C test for spatial autocorrelation are general tests, but the LM tests are more specific. The LM tests provide a basis for choosing an appropriate spatial regression model. In order to specify the nature of spatial dependence and fit an appropriate spatial regression model, we conducted simple LM tests, namely, LM error and LM lag. Furthermore we also tested significance of their robust versions, that is, RLM error and RLM lag tests.

Unlike the Moran test Lagrange Multiplier tests rely on well-structured hypotheses. The Lagrange Multiplier test for spatial dependence (LM error test) is based on the estimation of the linear regression (3.9) model with spatially auto-correlated errors $u_{dj} = \lambda \mathbf{W}_1 u_{dj} + \varepsilon$. The LM hypothesis we tested here is about the significance of a spatial autoregressive coefficient λ (H_0: $\lambda = 0$ versus H_1: $\lambda \neq 0$). The test statistic is distributed as χ^2 (chi-square) with one degree of freedom.

Spatial dependence in regression models may not only be reflected in the error. Instead it may be accounted by entering a spatial lag $\mathbf{W}_1 y$ in the endogenous variable y. In this case the regression model reads $y = \rho \mathbf{W_1} y + \boldsymbol{\beta} \mathbf{X}_d + \varepsilon$.Under the null hypothesis H_0: $\rho = 0$ the standard regression model (3.9) holds, while under the alternative hypothesis H_1: $\rho \neq 0$ the extended regression model $y = \rho \mathbf{W_1} y + \boldsymbol{\beta} \mathbf{X}_d + \varepsilon$ would be valid. For conducting the Lagrange Multiplier test for spatial lag dependence (LM lag test) again only the standard regression model (3.9) is to be estimated. The test statistic is distributed as χ^2 (chi-square) with one degree of freedom.

Table 6. Lagrange Multiplier diagnostics for spatial dependence

	Statistic	Parameter	p-value
LM error test	81.295	1	0.000 *
LM lag test	82.947	1	0.000 *
RLM error test	0.208	1	0.648
RLM lag test	1.861	1	0.172
SARMA test	83.156	2	0.000 *

Significance Codes: 0.05 '*'

As we can see from Table 6, the LM error, LM lag, and SARMA tests are significant, indicating the presence of spatial dependence both in the error terms and dependent variable. Significance of LM error, LM lag tests points to a spatial error model, and spatial lag model respectively. In order to choose between spatial lag and spatial error model we followed a right hand rule (that is, LM lag test statistic > LM error test statistic), therefore we fitted the spatial lag model (Torben, 2008). The SARMA test, which tests for both lag and error showed significant value but it is not practically important since it is highly significant if either of the two simple LM tests is significant. The SARMA test was performed only for the sake of completeness (Torben, 2008).

4.2.1.2. Post-estimation Assessment of Fitted Spatial Simultaneous Autoregressive Lag Model

The overall goodness of fit test analysis result (see Table 7) for the spatial autoregressive lag model showed that the fitted model is appropriate. Rho reflects the spatial dependence inherent in our sample data, measuring the average influence on observations by their neighboring observations. It has a positive effect and is significant.

Table 7. Overall Goodness of fit test: Spatial Simultaneous Autoregressive Lag Model

Parameter		Asymptotic standard error	Test	Statistic	P-value
Rho:	0.42402	0.047454	Z	8.9354	< 2.22e-16*
			LR	65.106	6.66e-16*
			Wald	79.842	< 2.22e-16*

Significance Code: 0.05 '*'

Zero order correlation matrixes between model parameters is given in Table 8. As can be seen from the results there is no high intra-class correlation between fixed effects. But the variable eco-factor is negatively correlated with the intercept term.

Table 8. Correlation of Model Coefficients

	sigma	rho	(Intercept)	eco_factor	Pop_Density	slope
rho	-0.15					
(Intercept)	0.05	-0.36				
eco_factor	-0.01	0.07	-0.75			
Pop_Density	0.01	-0.10	0.07	-0.26		
slope	-0.01	0.04	-0.27	-0.08	-0.12	
NDVI	0.03	-0.18	-0.37	0.11	0.00	-0.19

Homogeneity of Variance Test Results

Hypothesis: H_0: $\sigma_e^2 = 0$ vs H_1: $\sigma_e^2 \neq 0$

We tested if individual specific variance components were zero with Studentized Breusch-Pagan test. The result of the test (BP: 8.986, df = 4, p-value: 0.061) indicate that we do not reject the null hypothesis that the variances are homogenous, that is, there is no heteroskedasticity in the residuals. We also confirmed the result by LM test for residual autocorrelation (LM: 0.446, p-value: 0.504).

Assumptions of Normality test Results

We used histogram and QQ normal quintile plots to assess the normality of residuals as shown in Figure 2. Both plots suggest the deviation from normality. Similarly, Shapiro-Wilk and Pearson chi- square tests (see Table 9.) also confirmed the violation of normality. In order to deal with this issue we tried Boxcox and log transformation but unfortunately it would not solve the problem.

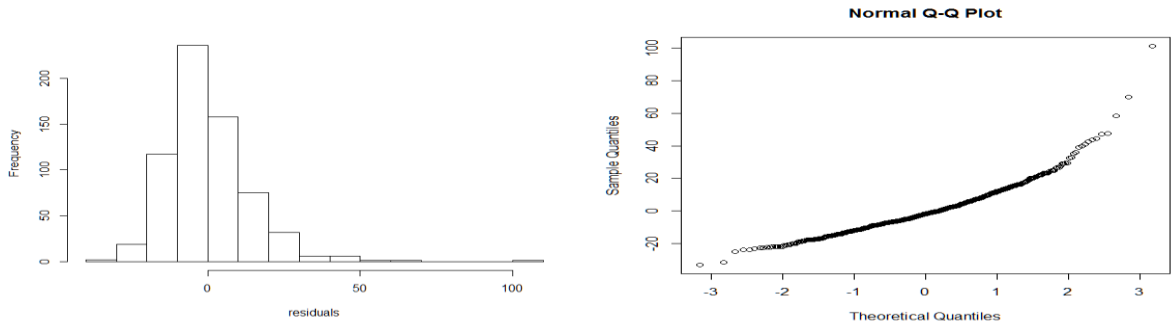


Figure 2. Histogram and Q-Q plots of residuals for spatial simultaneous autoregressive lag model

Table 9. Normality Test for Spatial Simultaneous Autoregressive Lag Model

Type of normality test	Statistic	p-value
Shapiro-Wilk	W = 0.922	< 2.2e-16*
Pearson Chi-square	P = 73.183	7.13e-07*

Significance Code: 0.05 '*'

4.2.1.3. Post-estimation Assessment of Fitted Spatial Simultaneous Autoregressive SAC Model

The overall goodness of fit test analysis result for spatial simultaneous autoregressive model shown in Table 10 indicates that the fitted model is appropriate. Both parameters of the fitted SAC model are positive and significant indicating the presence of spatial dependence both in the dependent variable and residuals.

Table 10. Goodness of fit test: Spatial Simultaneous Autoregressive Model

Parameters		Asymptotic standard error	z-value	p-value
Rho:	0.438	0.109	3.987	6.69e-05*
Lambda:	0.310	0.065	0.065	1.85e-06*
LR test value:	76.254			< 2.22e-16*

Significance Code: 0.05 '*'

Table 11 shows correlations between model parameters of fitted SAC model. There is no strong intra-class correlation between fixed effects except that the eco-factor variable is negatively correlated with the intercept term.

Table 11. Correlation of coefficients for SAC model

	sigma	rho	lambda	(Intercept)	eco_factor	Pop_Density	slope
rho	-0.01						
lambda	-0.09	-0.56					
(Inter cept)	0.01	-0.59	0.33				
eco_factor	0.00	0.06	-0.03	-0.63			
Pop_Density	0.00	-0.08	0.04	0.04	-0.23		
slope	0.00	-0.04	0.02	-0.2	-0.09	-0.05	
NDVI	0.00	-0.21	0.12	-0.29	0.09	-0.01	-0.14

Homogeneity of Variance Test Results

Hypothesis: H_0: $\sigma_e^2 = 0$ vs H_1: $\sigma_e^2 \neq 0$

In order to test if individual specific variance components were zero we used Studentized Breusch-Pagan test. The result of the test (BP: 3.567, df = 4, p-value: 0.468) indicate that we do not reject the null hypothesis that the variances are homogenous.

Assumptions of Normality test Results

In the same way as the spatial lag model, we assessed weather or not the normality assumption of the SAC model is fulfilled using histogram and Q-Q normal quintile plots as shown in Figure 3. Both plots suggest deviation from normality. The well-known Shapiro-Wilk and Pearson chi-square tests (see Table 12) also confirmed the violation of normality assumption.

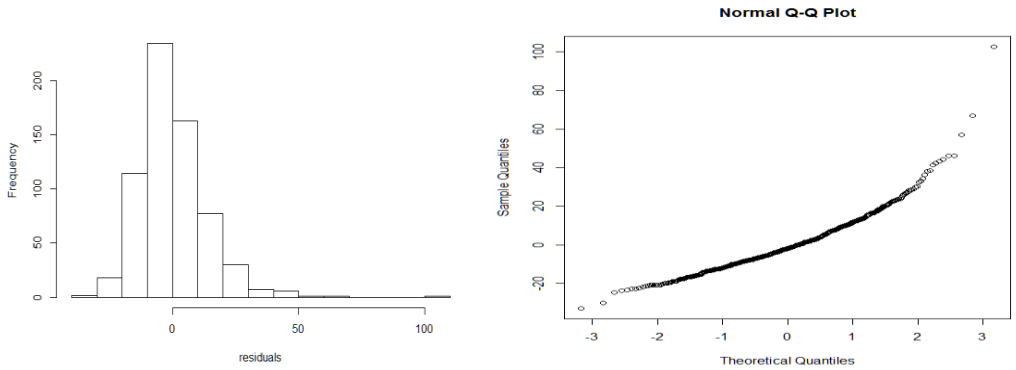


Figure 3. Histogram and Q-Q normal plots of residuals for spatial simultaneous autoregressive SAC Model

Table 12. Normality test of Spatial Autoregressive Regressive SAC Model

Type of normality test	Statistic	p-value
Shapiro-Wilk	W = 0.921	< 2.2e-16*
Pearson Chi-square	P = 73.183	7.13e-07*

Significance Code: 0.05 '*

4.2.2. Model Diagnostics for Mixed effect Linear Models

4.2.2.1. Unit level Mixed effects Linear Model (EBLUP_B)

In the mixed model approach one common way to test the model's fit is to rerun the analysis but include only the intercept terms which is often called the null model and compare the conditional AIC (cAIC) of that model to the hypothesized (full) model. Therefore, the cAIC of the full model 5267.7 that is smaller than the null model that is 5286.4 suggesting the full model is appropriate. In order to confirm this, we refitted the model by ML method and conducted ANOVA test. The results shown in Table 13 clearly indicate that at $\alpha = 0.05$ level the two models are significantly different.

Table 13. Goodness of fit statistics for mixed effect model, refitting the model by ML

Model	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
NULL	3	5336.9	5350.4	-2665.5	5330.9			
FULL	7	5314.5	5345.9	-2650.2	5300.5	30.414	4	4.03e-06 *

Significant code, "*" 0.05

We also conducted analysis of variance test which returns F statistics corresponding to the sequential decomposition of the contributions of fixed-effects terms. The relative magnitudes of the four sums of squares as shown in Table 14 indicate that NDVI, Pop_Density, eco-factor, and slope terms explain highest to lowest variation successively. The magnitudes of the F statistics for NDVI, Pop_Density, and eco-factor strongly suggest significance at $\alpha = 0.05$ level.

Table 14. ANOVA Test of Fixed Effects: Mixed effect linear model (EBLUP_B)

Fixed Effects	numDF	denDF	Sum Sq	Mean Sq	F value	p-value
eco_factor	5	412	2086.46	417.29	2.759	0.020*
Pop_ Density	1	412	1236.29	1236.29	8.175	0.005*
Slope	1	412	334.30	334.30	2.204	0.140
NDVI	1	412	2498.45	2498.45	16.531	0.000*

Significance code 0.05(*)

Another measure of goodness of fit test of the model was performed by looking at the proportion of total variability attributed to wereda level random effect. The variance estimates are of interest here because we can add them together to find the total variance (of the random effects) and then divide that total by the wereda level random effect to see what proportion of the random effect variance is attributable to wereda-level random effect (similar to R^2 in traditional regression). The estimated variance components for this model are $\hat{\sigma}_u^2 = 54.83$ and $\hat{\sigma}_\varepsilon^2 = 153.36$ for random effect and residuals respectively. Therefore dividing the total variance by the random effect variance we can see that 26.33% of the total variance of the random effects is attributed to the wereda level random effect. The observed percentage indicates presence of the area level random effect in the model and appropriateness of the fitted mixed linear model.

4.2.2.2. Diagnostic for Assessing Assumptions of mixed effect linear model

Various types of diagnostic plots were performed to see whether the assumptions of the model was satisfied

The histogram and Q-Q plots shown in Figure 4 indicated that the normality assumption of the model is violated. This is a common phenomenon in agro-environmental variables (Monica, 2015). When the assumption is satisfied, the EBLUPs show a substantial gain in efficiency for spatially stationary, SAR stationary and spatially non-stationary models (Rao, 2003). Therefore the effect of non-normality to small area estimators must be checked using various diagnostic approaches.

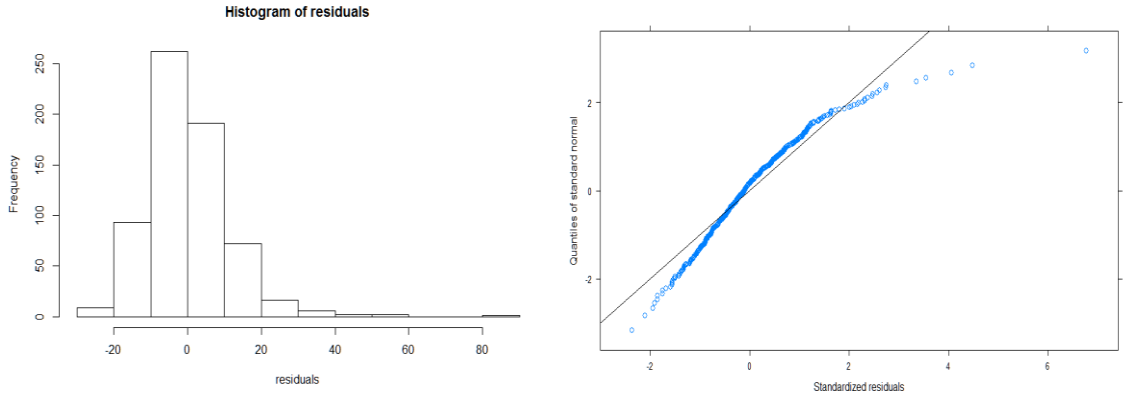


Figure 4. Residual plots for mixed effect linear model (EBLUP)

Influential Diagnostics Results

In order to see the presence of influential observations we used case delete and cooks distance. We run first case_delete () using R package to obtain diagnostics from a full refit of the model for each deletion to extract all the necessary information from the model, after which the same influence functions can be called on the result. Case delete is important to identify influential observations on variance components, and we deleted two observations that are influential. In the case of Cook's distance we used the proposed cut point value of CkD, CkD >1 identifies cases that might be influential. We run this for both grouping variables wereda and EA level. According to the Cook's distance diagnostic approaches none of the observations are considered as influencial (see these diagnostic results in Appendix IV and Appendix V)

4.2.2.3. Spatial Autoregressive Mixed effect Linear Model (EBLUP_A)

We extended unit-level mixed effect linear model to the spatial Fay-Herriot's area-level model to account for the spatial dependence in the response variable. As we can see from the unit-level spatial lag and spatial autoregressive regressive (SAC) linear models described above, there was a significant spatial dependence (autocorrelation) in the dependent and residual terms. Therefore these spatial dependences have to be accounted by fitting an appropriate spatial mixed model. In

this study we only considered area-level spatial Fay-Herriot's. Below we provide the pre-estimation diagnostics conducted to see the appropriateness of the fitted spatial Fay-Herriot's area-level model.

Spatial Autocorrelation Test Results

Spatial autocorrelation is the formal property that measures the degree to which near and distant things are related. In order to see whether there is a spatial autocorrelation or not in the residuals of the mixed effect model, two types of formal tests, Global Moran's I test under randomization and Geary's c test under randomization was used. The hypothesis tested here is that $H_o: \rho = 0$ V_s $H_A: \rho > 0$. Both test results confirmed that there is a significant spatial autocorrelation. The results are presented in Table 15 below.

Table 15. Moran's I and Geary's C tests for Spatial Autocorrelation

Test	Statistic	P-value
Moran's I test under randomization	0.123	< 2.2e-16*
Geary's c test under randomization	0.865	1.30e-12*

In addition to the above tests, map based clustered outlier and spatial dependence of the geo-referenced variable maize yield for the sample data has been examined.

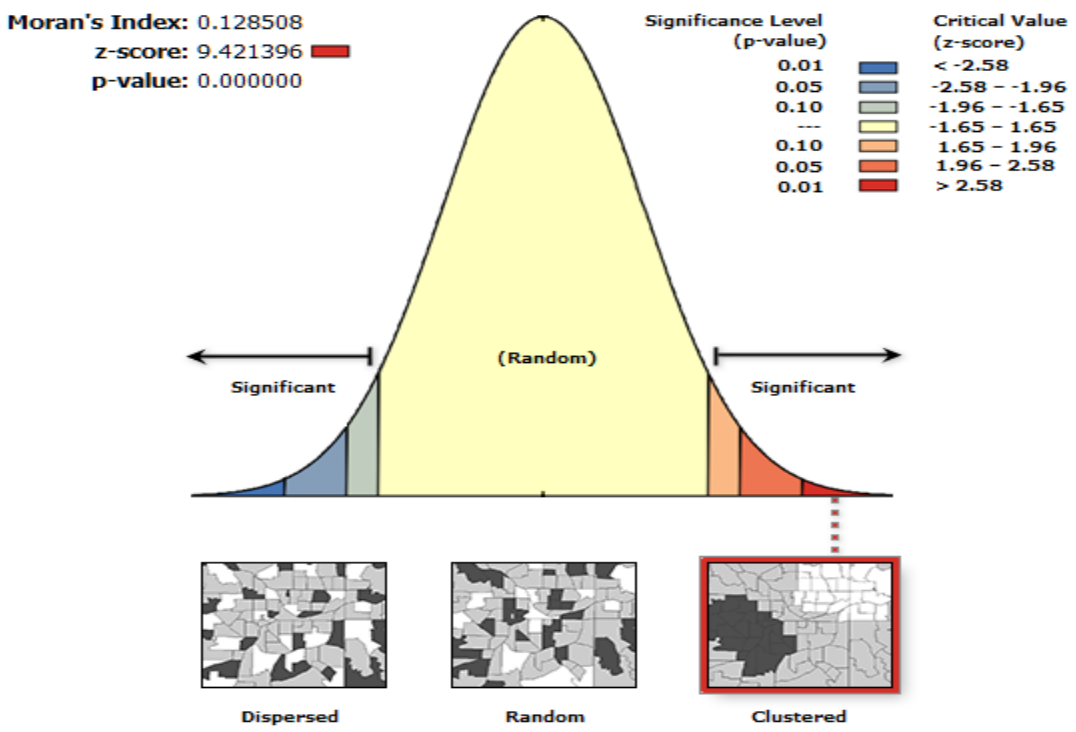


Figure 5. Spatial Autocorrelation Report for Maize Yield: Global Moran's I Summary (Using Inverse Weighting & Row Standardized method)

Given the z-score of 9.42 provided in Figure 5, there is less than 1% likelihood that this clustered pattern could be the result of random chance.

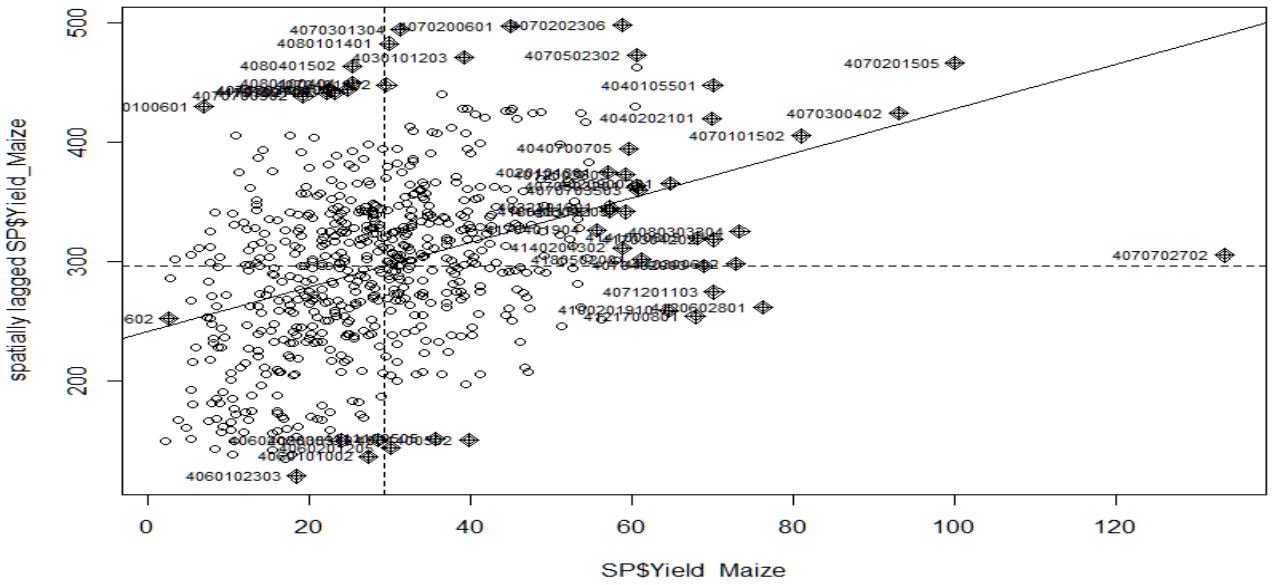


Figure 6. Moran's Spatial Dependence Scatter Plot K=5

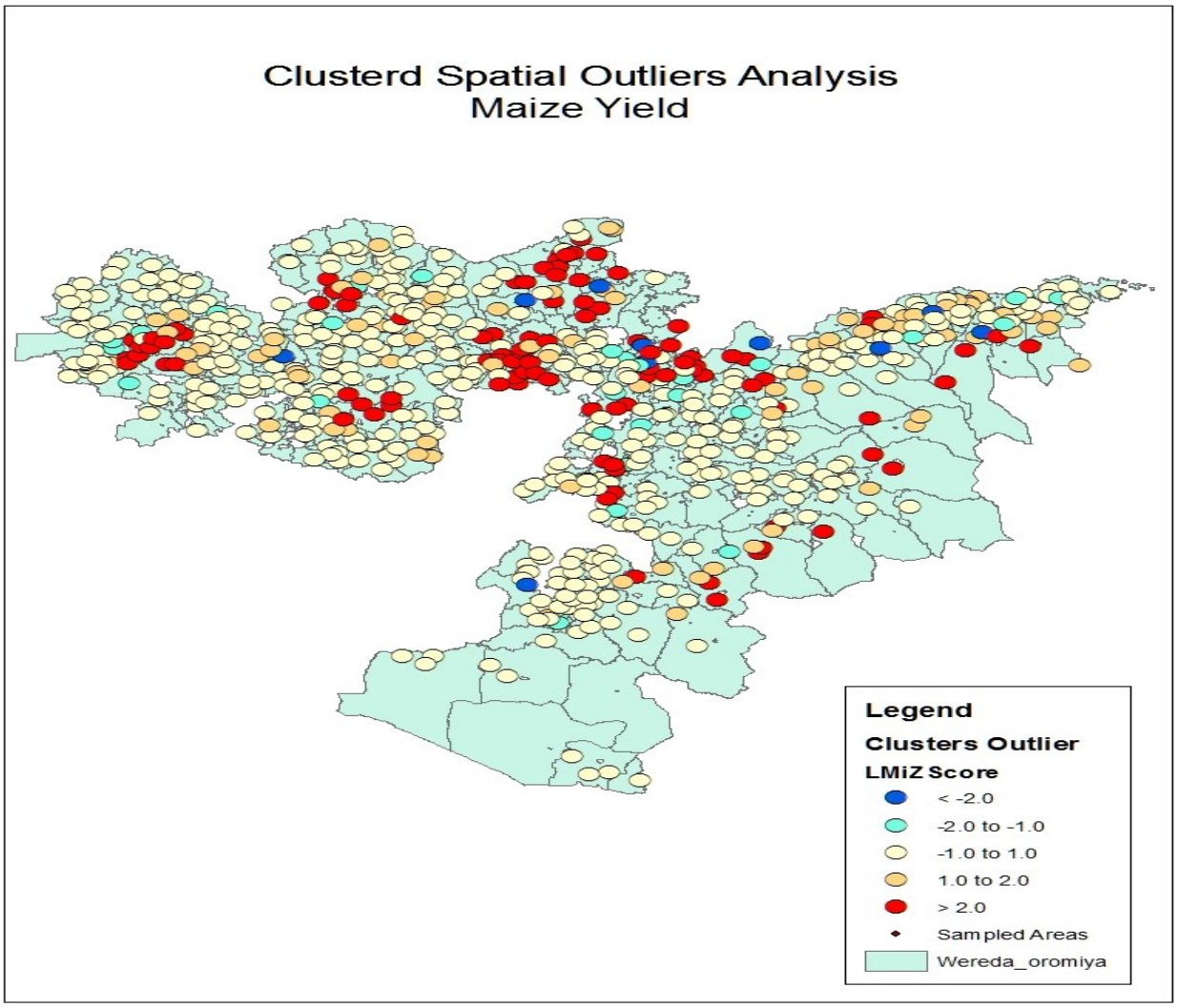


Figure 7. Map Based Cluster Outliers Analysis: Maize Yield

The map based cluster analysis of maize yield shown in the map (Figure 7) confirmed the spatial dependence of maize yield. Therefore an appropriate spatial extension of mixed effect model has to be fitted in order to account for these relations. Further analysis of this spatial dependences was also conducted by Riples K function and Morans I spatial dependence plot for k=5 nearest neighborhood distances.

Multi-Distance Spatial cluster analysis (Riples K function) shown in Figure 8 determines whether maize yield exhibit statistically significant clustering or dispersion over a range of distances. Both methods confirmed there is clustering's of points at shorter distances.

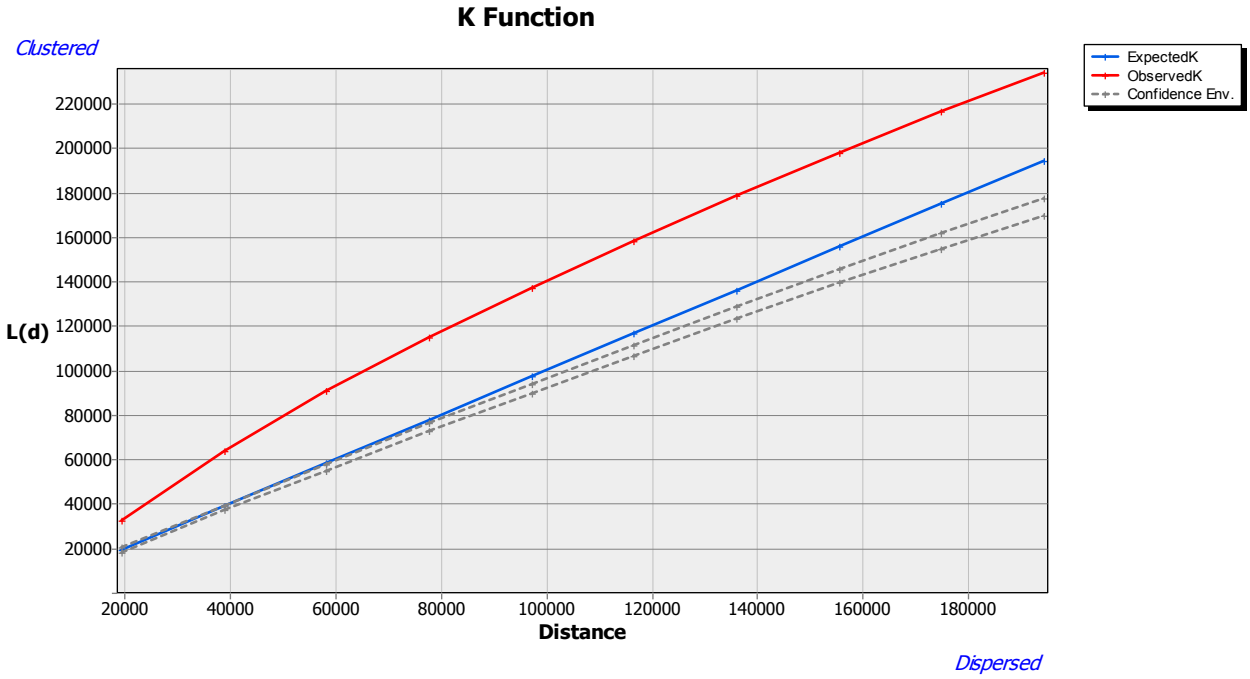


Figure 8. Multi-Distance Spatial cluster analysis (Riples K-function: 999 permutations)

All the above pre-estimation tests regarding the spatial dependence of maize yield suggested that the spatial extension of mixed effect model has to be considered for wereda-level maize yield estimation. One such model is the spatial EBLUPs under a spatial Fay-Herriot model. Therefore we used the spatial extension of area level Fay-Herriot model in order to estimate maize yield at wereda level. The main reason why area level spatial Fay-Herriot's model was preferred rather than unit level spatial mixed effect model was due to lack of direct variance estimates at unit level. Spatial mixed effect models require direct variance estimates in addition to proximity matrix as in put to model spatial dependence.

4.3. Assessment of Small Area Estimators

Small area estimation techniques are useful for subpopulation (domain) analysis when direct domain estimators do not have adequate precision due to small sample sizes. Indirect estimation for small areas uses statistical models and auxiliary variables to borrow strength from similar areas. In this study we used wereda level indirect estimators assuming spatial linear, unit level mixed effect and area level spatial mixed effect models. All model based estimators do not provide reliable estimates; some of the reasons include failed distributional assumptions, model specification problems, presence of outliers and influential observations. With regard to this we

devoted to assess various aspects of the fitted models by use of graphical plots and formal statistical tests. The assessment results were generally good except that the normality assumption has failed. In the next section we provide validation results for model based and direct estimators.

4.3.1. Area-specific measures

4.3.1.1. Coefficient of Variation of Estimators

The validation of the model-based estimators were conducted by computing the coefficient of variation (CV) to assess the improved precision of the model based estimators when compared to the direct survey estimators. Coefficient of variation of estimator shows the sampling variability as a percentage of the estimate. The estimates with large CVs are considered unreliable. Although there are no internationally accepted gold standard to judge what is "too large" the estimated CVs (Figure 9, Figure 10, Figure 11 and Figure 12) indicated that all model-based estimates have a higher degree of reliability than the direct survey estimates.

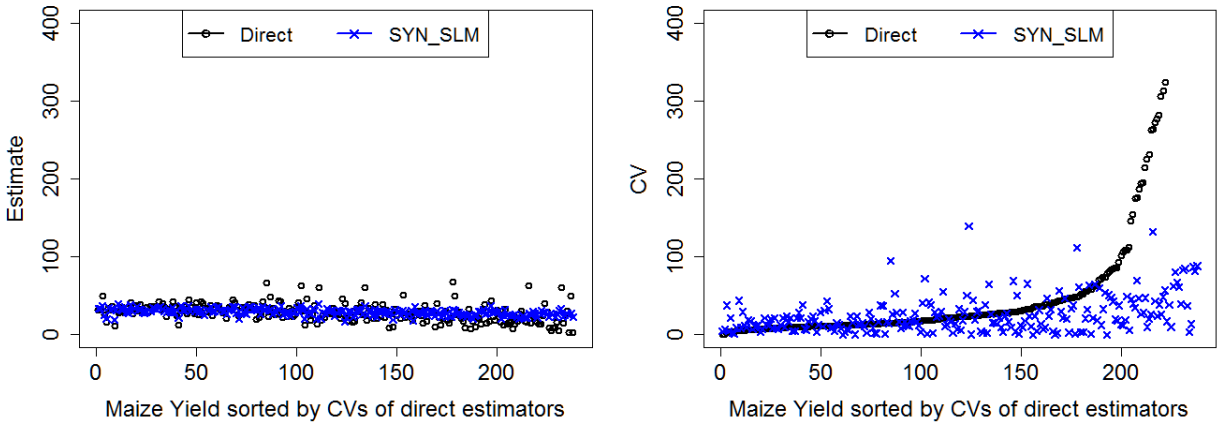


Figure 9. SYN_SLM and direct estimates of maize yield for each weredas (left), CVs of SYN_SLM and Direct estimators for each weredas (right).

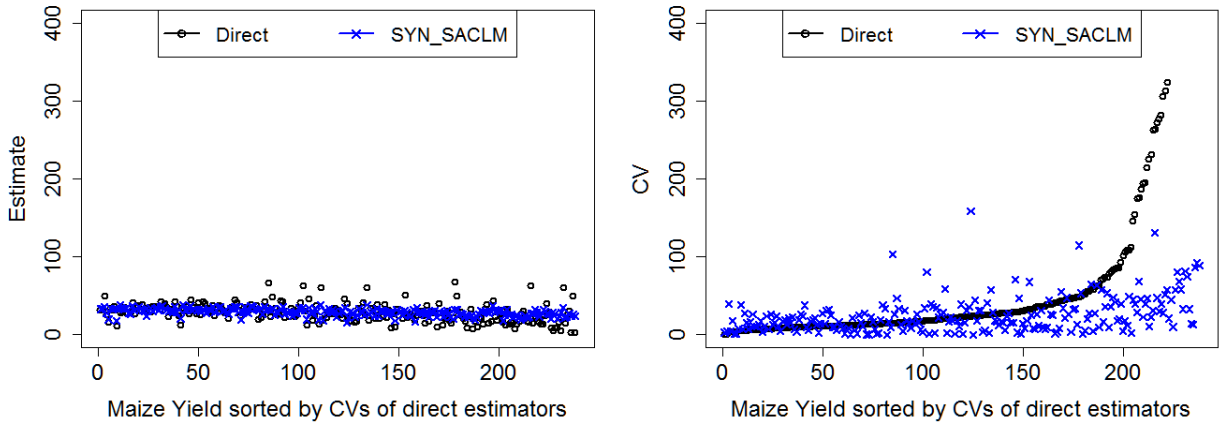


Figure 10. SYN_SACLM and direct estimates of maize yield for each weredas (left), CVs of SYN_SACLM and direct estimators for each weredas (right).

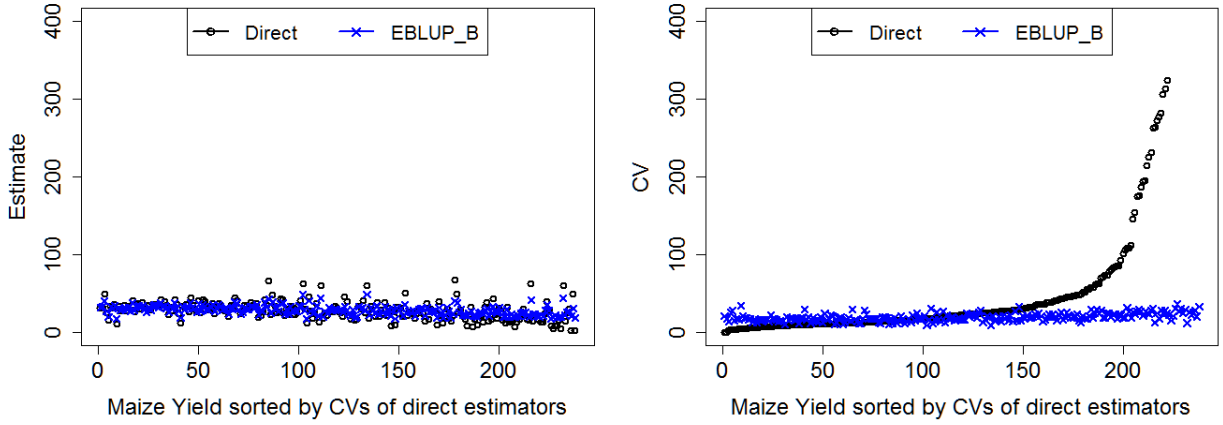


Figure 11. EBLUP_B and direct estimates of maize yield for each weredas (left), CVs of EBLUP_B and Direct estimators for each weredas (right)

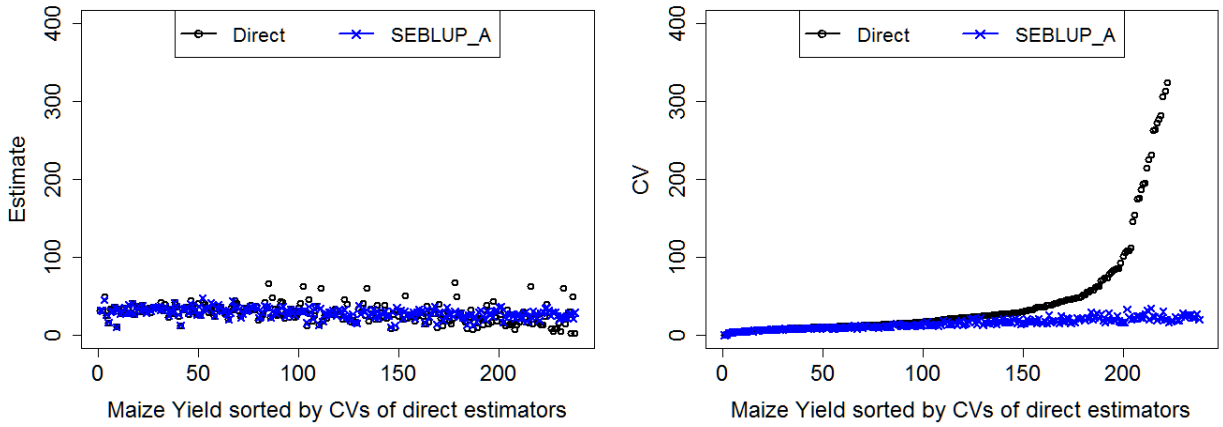


Figure 12. SEBLUP_A and direct estimates of maize yield for each weredas (left), CVs of SEBLUP_A and Direct estimators for each weredas (right).

4.3.1.2. RRMSE of Estimators

Table 16 reports %RRMSE for each estimator calculated based on the methodology described in section (3.6.1).The results show that SEBLUP_A and EBLUP_B (based on Unit level mixed effect model and Spatial Fay Herriot model) have relatively small values of RRMSE (15.16% and 19.90% respectively). But the SYN_SLM and SYN_SACLM estimators have relatively higher values of %RRMSE (38.22% and 38.26%, respectively).

Table 16. Rate of best RRMSE

ESTIMATOR	%Rate of best RRMSE
SYN_SLM	38.22
SYN_SACLM	38.26
EBLUP_B	19.90
SEBLUP_A	15.16

4.3.2. Global measures

Regarding the second diagnostic approach, we present the results of the Monte Carlo analysis. Table 17 reports the percentage values of $\overline{\text{ARB}}$, $\overline{\text{ARE}}$ and $\overline{\text{EFF}}$ for direct and model-based estimators.

EBLUP_B based on unit level mixed effect linear model performs better than other estimators in terms of $\overline{\text{ARB}}$, $\overline{\text{ARE}}$ and $\overline{\text{EFF}}$. But the SEBLUP_A (spatial Fay-Herriot model) performs poor relative to synthetic estimators SYN_SLM, and SYN_SACLM (simultaneous autoregressive lag

and spatial autoregressive regressive linear models) in terms of $\overline{\text{EFF}}$. Regarding $\overline{\text{ARB}}$ and $\overline{\text{ARE}}$, SEBLUP_A performed slightly better than SYN_SLM and SYN_SACLM.

Table 17. Quality measures by direct and model-based estimators

Estimators	Quality Indicator		
	$\%\overline{\text{ARB}}$	$\%\overline{\text{ARE}}$	$\%\overline{\text{EFF}}$
DIRECT	3.37	4.50	100.000
SYN_SLM	30.90	44.36	165.325
SYN_SACLM	30.60	44.27	160.632
EBLUP_B	23.85	32.27	227.145
SEBLUP_A	29.80	40.22	125.055

Note: *The Variance of the Direct Estimator used here is computed following a multistage probability sampling Design and weighting approaches (AGSS). The estimated variances are obtained under the approximation that the selection probabilities of the ultimate sampling units are the product of second order inclusion probabilities and first order inclusion probabilities. Therefore it is biased compared to the variance estimates under the assumption of simple random sampling.*

4.3.3. Diagnostic methods

The results of the last evaluation approach considered use observations relating to all small areas gathered from one sample and provide an appraisal of bias. These methods are considered as preliminary internal evaluation, and are useful when there is some doubt about the assumptions underpinning small area model, that is not necessarily nested.

4.3.3.2. Goodness of fit Diagnostic Results

Table 18 shows goodness of fit diagnostic for each model-based estimator. The results are however similar, that is, all model based estimators did not exhibit a significant bias from the direct estimator at $\alpha = 0.05$ level of significance.

Table 18. Wald test for Goodness of fit diagnostics

	Estimators			
	SYN_SLM	SYN_SACLM	EBLUB_B	SEBLUP_A
W	95.57	94.95	71.55	110.44
df	238	238	238	238
P.value	0.999	0.999	0.999	0.999

Chapter Five

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

This chapter summarizes the study by highlighting the research conducted on small area estimation of maize yield at wereda level under mixed effect linear model. The conclusions given were drawn from the outcomes of the empirical study and analysis of different small area estimators produced by design based and model based approaches. Moreover, recommendations were made from the findings and conclusion of the study.

5.1. Summary

The researcher conducted an empirical research to analyze the performance of small area estimators of maize yield at wereda level under simultaneous autoregressive lag dependent linear model, spatial simultaneous autoregressive SAC model, linear mixed effect model and spatial autoregressive linear mixed effect model (spatial Fay Herriot's model). The general purpose of this study was to examine the relative performance of linear mixed effect model in comparison to direct design based estimator and other models that do not account for area level random effects and spatial dependence between random effects. The small area estimation was conducted for all weredas found in Oromia Region of Ethiopia. In order to achieve this goal it was understandable to collect micro level auxiliary data from satellite imagery, administrative maps, and census data and link this data with agricultural survey data. The main season (meher) survey data for Oromia Region was obtained from CSA for the year 2013. Population size and administrative boundary shapefiles including enumeration area shapefiles was also obtained from CSA. The set of auxiliary data considered initially for model specification include, rainfall, temperature (min, max), precipitation, percentages of agricultural land under EA, agro ecology factors, normalized difference vegetative index (NDVI), population density, elevation, slope and aspects. However, examining the linear relationships between auxiliary and target variable we finally selected four covariates namely; agro ecology factors, slope, population density and NDVI to fit the models. Before fitting the models, kebeles with missing value or outliers (mainly due to cloudy weather and steep slopes) in the auxiliary variables have been imputed by smoothing method. After a serious of model selection and diagnostics processes we fitted the best models that are competent to be in use as a small area estimator of maize yield at wereda level.

The estimators computed were, SYN_SLM (simultaneous autoregressive lag dependent linear model), SYN_SACLM (spatial simultaneous autoregressive SAC linear model), EBLUP_B (unit level mixed effect linear model), and SEBLUP_A (spatial Fay Herriot's model). The diagnostics analysis of wereda-level maize yield estimators was also performed to assess the assumptions of each model, appropriateness of fitted models and compare robustness of each estimator relative to other estimators. The validation of fitted models was conducted involving various diagnostics methods including spatial autocorrelation tests, cAIC, ANOVA, formal residual analysis and diagnostic plots. On the other hand assessment of wereda level yield estimates was conducted based on four diagnostic metrics namely, graphical plot of CVs of each model based estimators against CVs of the direct estimator, %RRMSE, global measures (%$\overline{\text{ARB}}$, %$\overline{\text{ARE}}$ and %$\overline{\text{EFF}}$), Goodness of fit test diagnostics. The relative quality of each model was analyzed taking in to account design based direct estimator as a reference.

5.2. Conclusions

The study revealed augmenting the survey data with accurate remote sensing data and recent census results using implicit and explicit linking models would likely produce reliable estimates of maize yield at wereda-level. In this study we investigated the relative performance of four model based estimators fitted with fixed effects only and mixed - effects approach. The result of the study confirmed that the entire model based estimators showed less extreme values compared to direct survey estimator with regard to their CV's. The study also tried to further examine the relative performance of these model based estimators obtained under mixed effect linear model and fixed effect only spatial linear model approaches. The first estimator (SYN_SLM) was produced using a simultaneous autoregressive lag dependent linear model, the second estimator (SYN_SACLM) was produced using a spatial simultaneous autoregressive SAC linear model, the third model (EBLUP_B) was produced using unit level mixed-effect linear model, and the last estimator (SEBLUP_A) was produced using an area level spatial autoregressive mixed effect model. The result of the study show EBLUP_B and SEBLUP_A estimators have relatively less relative bias measures in terms of best %RRMSE and global bias measures compared to SYN-SLM and SYN-SACLM. In further comparison of the relative efficiency of EBLUP_B with SEBLUP_A the study revealed that EBLUP_B has relatively less bias and reliable estimator of wereda level maize yield.

Finally, we concluded that based on CV, RRMSE, Global bias and GoF diagnostics methods the study used, EBLUP estimator (unit level linear mixed effect model) appears to provide some efficiency gains over the direct estimator.

5.3. Recommendations

The following recommendations are offered for related research in the areas of small area estimation of maize yield using mixed effect model approach.

1. Given the necessities of reliable and optimal small area estimators of crop yield at wereda level in real life situations, a serious of continuous researches that accounted for other important auxiliary information such as fertilizer type, use of improved seed and other environmental factors is required before using for predictions.

2. Replication of the same research in other regions would be important to compare the results and justify the appropriateness of the proposed model

3. Use of some recently proposed alternative approaches based on linear M-quantile regression model under frequentist methodology and computation of current estimators under a full Bayesian methodology, which is not done in this research due to space limitations; have to be explored before using for practical purposes.

References

Alemayehu, M. (2006). *Country Pasture/Forage Resource Profiles: Ethiopia.* Rome: FAO.

Anselin, L. (1988a). *Spatial Econometrics: Methods and Models.* Netherlands: KluwerAccademic Publishers Group.

Anselin, L. (1992). *Spatial Econometrics: Method and Models.* Boston: Kluwer Academic Publishers.

Azizur, R. (2008). *A Review of Small Area Estimation Problems and Methodological Developments.* Australia : NATSEM, University of Canberra.

Banerjee, S., Carlin, B., & Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data.* New York: Chapman & Hall.

Battese, G., Harter, R., & Fuller, W. (1988). An Error Component Model for Prediction of County Crop Areas Using Survey and Satelite Data. *Journal of the American Statistical Association*(83), 28-36.

Bellow, M. E., & Lahiri, P. S. (2011). An Empirical Best Linear Unbiased Prediction Approach to Small-Area Estimation of Crop Parameters. *JSM* (pp. 3976-3986). Maryland: National Agricultural Statistics Service.

Bivand, R. (2002). Spatial econometrics functions in R: Classes and methods. *Journal of Geographical Systems, 4*, 405-421.

Bocci, C., Petrucci, A., & Rocco, E. (2012). Small Area Methods for Agricultural Data: A Two Part Geoadditive Model to Estimate the Agrarian Region Level Means of the Grapevines Production in Tuscany. *Journal of the Indian Society of Agricultural Statistics*, 135-144.

Braimoh, A. K., & Vlek, P. (2006). Soil quality and other factors influencing maize yield in northern Ghana. *Soil Use and Managment*(22), 165-171.

Cai, R., Yu, D., & Oppenheimer, M. (2014). Estimating the Spatially Varying Responses of Corn Yields to Weather Variations using Geographically Weighted Panel Regression. *Journal of Agricultural and Resource Economics, 39*(2), 230-252.

Cochran, W. G. (1977). *Sampling Techniques.* New York: John Wiley & Sons, Inc.

Cressie, N. (1993). *Statistics for Spatial Data* (revised ed.). New York: Wiley.

CSA. (2013). *Report on area and production of major crops.* Addis Ababa: Central Statistical Agency.

D' Alo, M., Consiglio, L. D., & Stefano Falorsi, M. G. (2011). Use of Spatial Information in Small Area Models for Unemployment Rate Estimation at Sub-Provincial Area in Italy. *The Indian Society of Agricultural Statistics*, 43-53.

Datta, G. S. (2009). Model-Based Approach to Small Area Estimation. In D. Pfeffermann, & C. R. Rao, *Sample Surveys: Inference and Analysis* (p. 255). Elsevier B.V.

Didan, K., & Huete, A. (2006). *MODIS Vegetation Index Product Series Collection 5 Change Summary.Tucson, Arizona.* Retrieved 2016, from https://lpdaac.usgs.gov/products/modis_products_table/mod13q1.

Fay, R., & Herriot, R. (1979). Estimates of income for small places: an application of James–Stein procedures to census data. *Journal of the American Statistical Association*(74), 269–277.

Fischer, M. M., & Wang, J. (2011). *Spatial Data Analysis Models, Methods and Techniques.* London New York: Springer Heidelberg Dordrecht.

Gohosh, M., & Rao, J. (1994). Small Area Estimation: An appraisal. *Statistical Science*, 55-93.

Gonzalez-Manteiga, W., Lombardia, M., Molina, I., Morales, D., & L., S. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Computational Statistics and Data Analysis, 52*, pp. 5242-5252.

Haining, R. (1990). *Spatial data analysis in the social and environmental sciences.* Cambridge: Cambridge University Press.

Henderson, C. (1975). Best linear unbiased estimation and prediction under selection model. *Biometrics*, 423-447.

Hidiroglou, M. (2007). Small-Area Estimation: Theory and Practice. *Section on Survey Research Methods* (pp. 3445-3456). Philadelphia, PA: American Statistical Association.

Hidiroglou, M. (2010). *Small-Area Estimation: Theory and Practice.* Ottawa: Statistics Canada.

Hijmans, R., Cameron, S., Parra, J., Jones, P., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 1965-1978.

Hindmarsh, D. M. (2013, March 18). Small area estimation for health surveys. *Doctor of Philosophy thesis, School of Mathematics and Applied Statistics*. Wollongong, New South Wales, Australia: University of Wollongong, http://ro.uow.edu.au/theses/3746.

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

Inglese, F., Monica, R., & Russo, A. (2008). Different approaches for evaluation precision Small Area Model-Based Estimators. *Proceedings of Q2008 European Conference on Quality in Official Statistics* (pp. 1-13). Roma,Italy: Italian National Statistical Institute.

Josephson, A. L., Ricker-Gilbert, J., & Florax, R. J. (2014). How does population density influence agricultural intensification and productivity? Evidence from Ethiopia. *Food Policy*(48), 142-152.

Kacker, R., & Harville, D. (1981). Unbiased of two-stage estimation and prediction procedure for mixed linear models. *Communications in Statistics- Theory and Methods*, 1249-1261.

Lohr, S. L. (2010). *Sampling: Design and Analysis.* Boston, MA 02210, USA: Brooks /Col ,Cengage Learning.

Lohr, S. L., & Prasad, N. (2010). *Small Area Estimation with Auxiliary Survey Data.* Tempe, AZ 85287-1804: Arizona State University.

Michael, M., & James, W. (2015). *Ethiopian Wheat Yield and Yield Gap Estimation: A Small Area Integrated Data Approach* . Addis Ababa, Ethiopia: International Food Policy Research Institute (IFPRI) .

Monica, P. (2015). *Spatial Disaggregation and Small-Area Estimation Methods for Agricultural Surveys:Solutions and Perspectives.* United Nations Statistical Commission.

Petrucci, A., & Pratesi, M. (2014). spatial models in small area estimation in the context of official statistics. *Italian Journal of Applied Statistics, 24*(1), 9-27.

Prasad, N., & Rao, J. N. K. (1999). The estimation of mean squared error of small-area estimators. *Journal of the American Statistical Association, 85*, 163–171.

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria: https://www.R-project.org/.

Rao, J. N. K. (2003). *Small Area Estimation.* (R. M., Groves, K. Graham, R. J. N. K., N. Schwarz, & C. Skinner, Eds.) Hoboken, New Jersey, USA: John Wiley & Sons, Inc.

Rao, J. N. K. (2003a). Some New Developments in Small Area Estimation. *JIRSS, 2*, 145-169.

Rao, J. N. K. (2014). Inferential issues in model-based small area estimation: some new developments. *STATISTICS IN TRANSITION new series and SURVEY METHODOLOGY, 16*(Small Area Estimation), 491-510.

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

Salvati, & Petrucci, A. M. (2008). Small area estimation: the eblup estimator based on spatially correlated random area effects. *Statistical Methods & Applications, 17*, 113–141.

Sirvastava, A. (2010). *Small area Estimation: Case of Ethiopia.* Retrieved December 10, 2015, from Food and Agriculture Organization of the United Nations (FAO): http://www.fao.org>ess>Technical-Document

Torben, K. (2008). Spatial model selection and spatial knowledge spillovers:A regional view of Germany. *Centre for European Economic Research (ZEW)* (pp. 1-66). Augsburg: University of Augsburg.

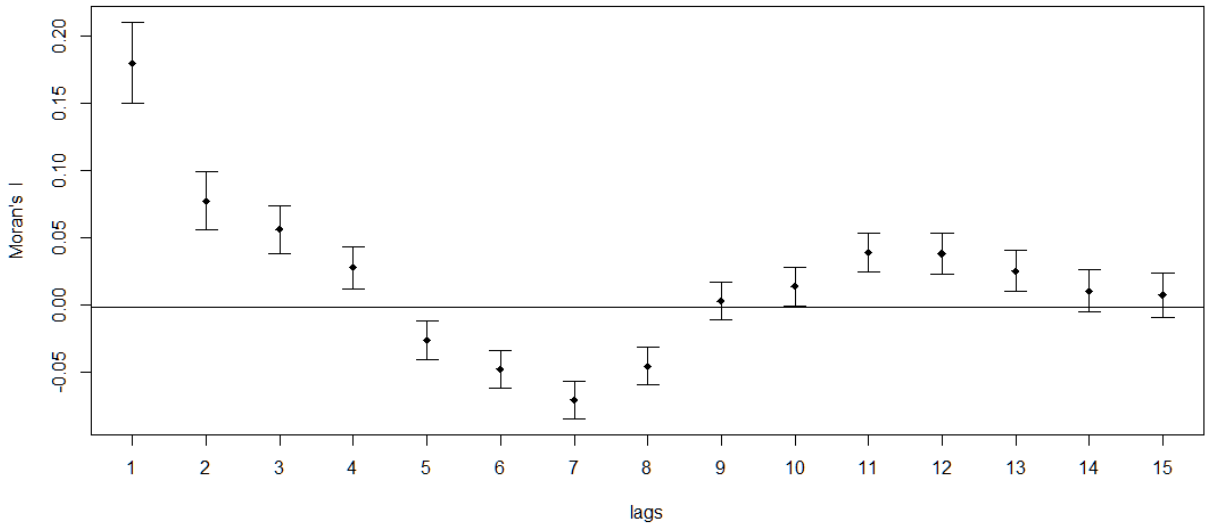
Tsedeke, B., Shiferaw, A., Abebe, M., Dagne, W., Yilma, K., Kindie, T., et al. (2015, July 13). *Factors that transformed maize productivity in Ethiopia.* Retrieved Jun 20, 2016, from SpringerLink: https://www.researchgate.net

Tzavidis, N., Chambers, R. L., Salvati, N., & Chandra, H. (2012). Small area estimation in practice an applocation to agricultural business survey data. *Journal of the Indian Society of Agricultural Statistics, 1*(66), 213-228.

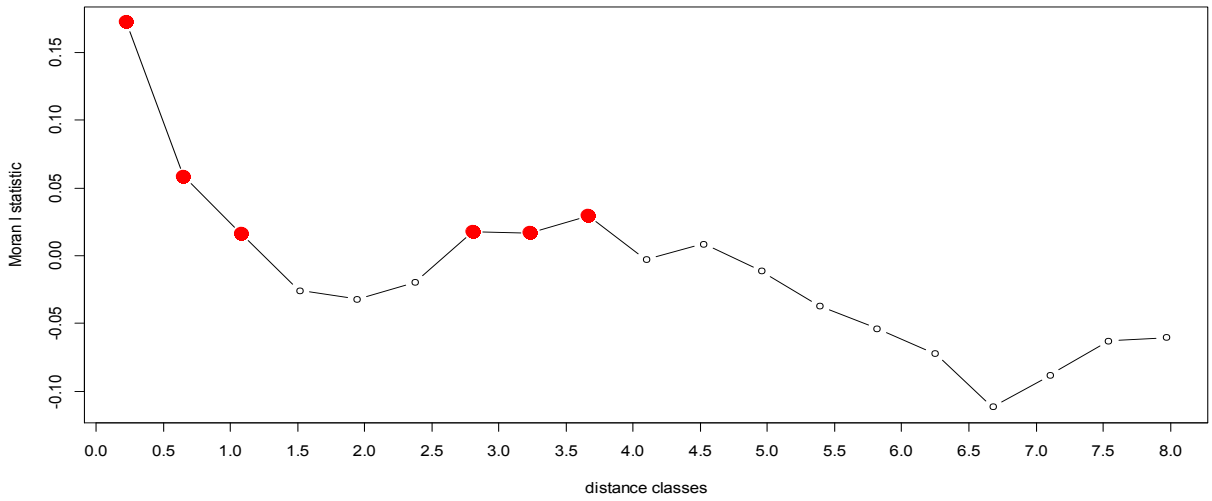
Vaida, F., & Blanchard. (2005). Conditional Akaike information for mixed effect models. *Biometrika*(92), 351−370.

APPENDIXES

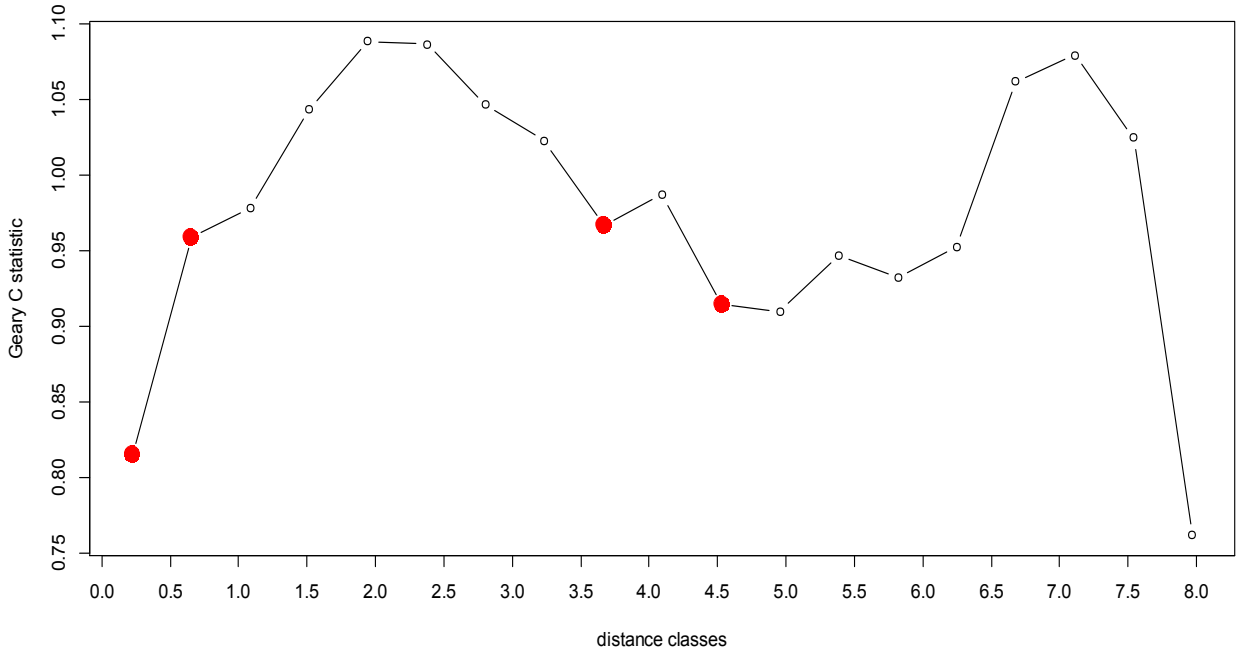
Appendix I. Correlograms for Maize Yield: Values of Moran's I for fifteen successive lag orders of contiguous neighbors



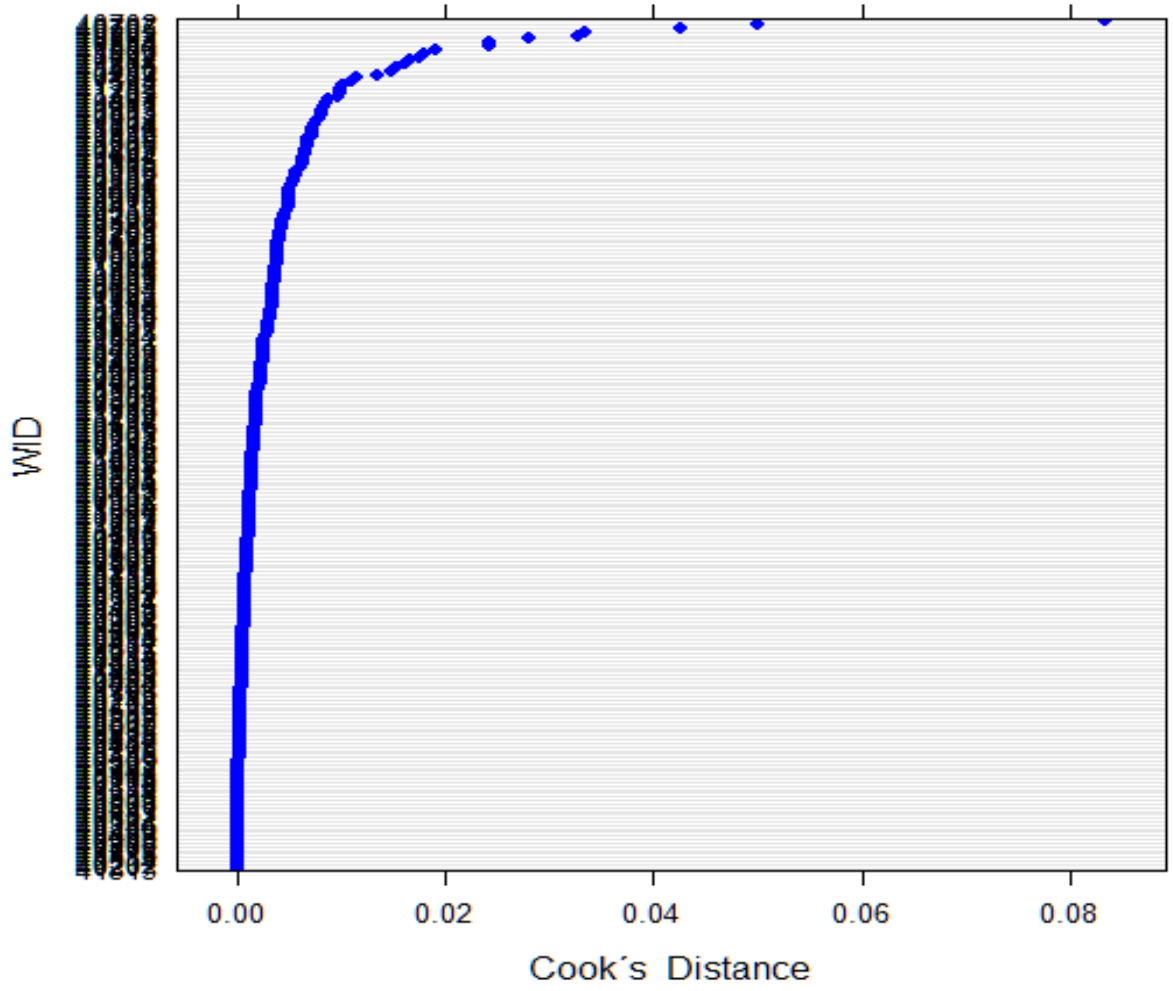
Appendix II. Moran I statistic = f(distance classes)



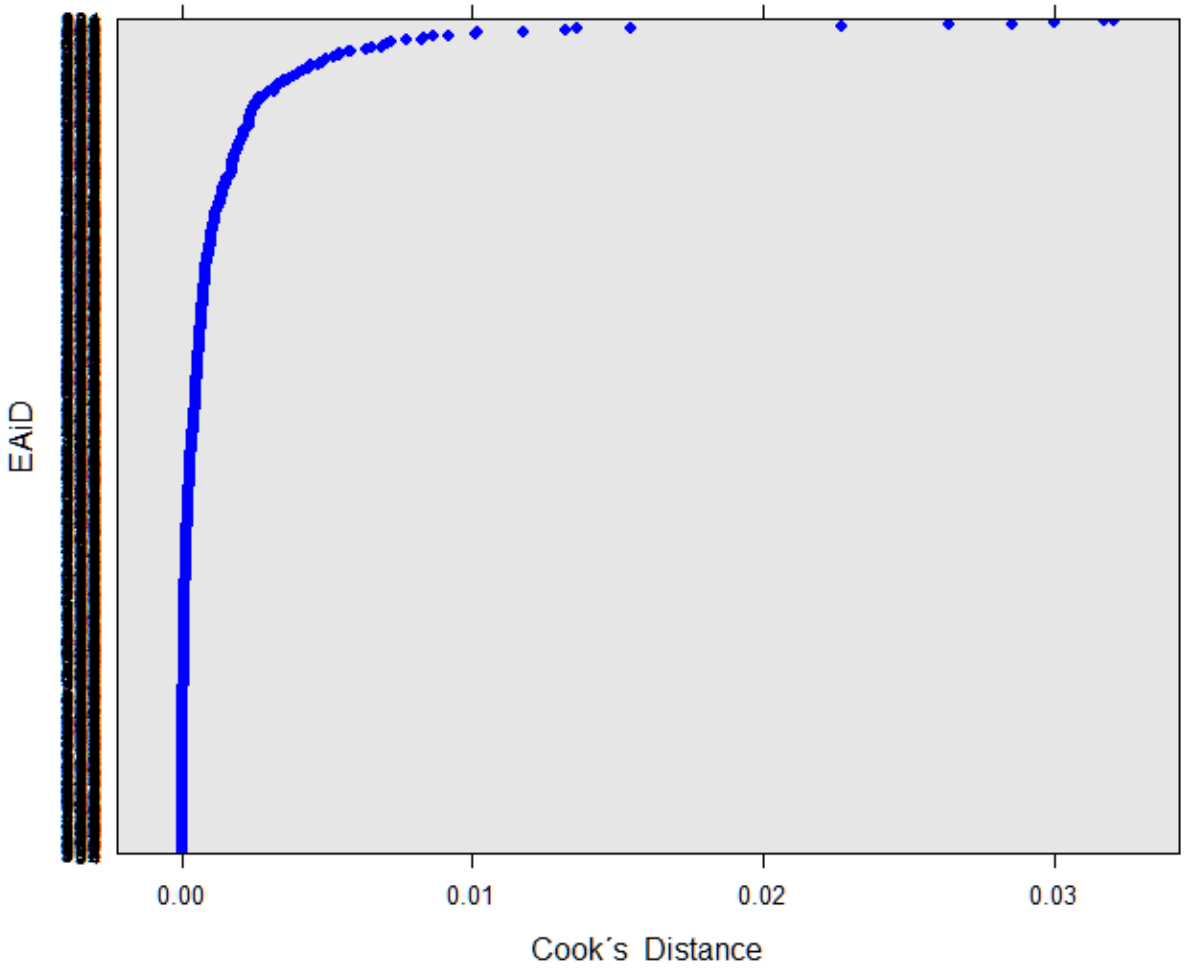
Appendix III. Geary C statistic = f(distance classes)



Appendix IV. Cook's Distances for Grouping Variable WID (wereda id)



Appendix V. Cook's Distance at Observations (enumeration area level)



Declarations

I, the undersigned, declare that this is my work and that all the sources of material used for this thesis have been duly acknowledged.

Damtew Berhanu Gebremariam _____

Name of student Signature

M.K. Sharma _____

Name of Advisor Signature