AJAS / SPAS

African Journal of Applied Statistics
Vol. 3 (1), 2016, pages 69–90.
DOI: http://dx.doi.org/10.16929/ajas/2016.69.203

ISSN 2316-0861

Crop yield estimation at district level for agricultural seasons 2014 in Rwanda

Innocent Ngaruye^{1,3,*}, **Dietrich von Rosen**^{2,3} **and Martin Singull**^{3}

^{1}Department of Mathematics, College of Science and Technology, University of Rwanda
^{2}Department of Energy and Technology, Swedish University of Agricultural Sciences, Sweden
^{3}Department of Mathematics, Linköping University, Sweden

Received 2016/11/30; Accepted 2016/12/15; Published Online 2016/12/29

Abstract. In this paper, we discuss an application of Small Area Estimation (SAE) techniques under a multivariate linear regression model for repeated measures data to produce district level estimates of crop yield for beans which comprise two varieties, bush beans and climbing beans in Rwanda during agricultural seasons 2014. By using the micro data of National Institute of Statistics of Rwanda (NISR) obtained from the Seasonal Agricultural Survey (SAS) 2014 we derive efficient estimates which show considerable gain. The considered model and its estimates may be useful for policy-makers or for further analyses.

Résumé. Dans cet article, nous abordons une application des techniques d'estimation sur petits domaines en utilisant le model de régression linéaire multivariée pour des données de mesures répétées en vue to produire des estimations du rendement de récolte de haricots bruts et haricots grimpants pour les saisons agriculturales 2014 au Rwanda. En utilisant les microdonnées de l'Institut National des Statistiques au Rwanda (NISR) obtenues de l'ênquete agricultural saisonier de 2014, nous déduisons des estimations efficaces avec un gain considérable. Le modèle considéré et ses estimations peuvent être utiles aux décideurs politiques ou à une analyse plus approfondie.

Key words: Crop yield, Model-based estimates, Multivariate linear model, Small Area Estimation.
AMS 2010 Mathematics Subject Classification : 62F10, 62H12, 62D05

Presented by Dr Serigne Lo, University of Sydney, Australia,
Member of the Corresponding Editors Board

*Corresponding author Innocent Ngaruye : innocent.ngaruye@liu.se
Dietrich von Rosen : dietrich.von.rosen@liu.se
Martin Singull : martin.singull@liu.se

1. Introduction

Agriculture in Rwanda is currently one of the pillars of Rwanda's economy and the agricultural production of the majority of Rwandese households is dominated by crop production. In recent years, the Government of Rwanda through its Ministry of Agriculture and Animal Resources (MINAGRI) has implemented several programs and policies to increase productivity of the agricultural sector. With this regard, the vision of MINAGRI is "to modernize Agriculture and Livestock to achieve food security. The vision is a transformation of the agriculture sector from subsistence to a productive high value, market oriented farming that is environmentally friendly and has an impact on other sectors of the economy".

In order to achieve these goals, the Government of Rwanda regularly gather up-to-date information for monitoring the progress on agriculture programs and policies, about crop production and livestock and other agricultural statistics which can be used for food and agriculture policy formulation and planning. In this framework, Seasonal Agricultural Surveys (SAS) are annually carried out to provide new agricultural statistics. Most of the cases, these surveys are designed to provide efficient estimates of parameters of interest at national level and do not provide reliable direct estimates at district level because of small sample size connected to the district. The need of crop production estimates is not only at national level but also at district level to show the distribution of crop production all over the country in order to facilitate the development of agriculture sector with the focus on specific regions.

The problem of how to produce reliable estimates of characteristics of interest for sub populations or domains for which the direct estimates are not of high precision because of small sample sizes taken from these domains and the assessment of estimation or prediction error is known as the Small Area Estimation (SAE) problem (e.g., see Pfeffermann (2013)). Because of a growing demand for small area statistics worldwide, SAE has received a lot of attention in recent years. Several authors have discussed the SAE methodology. Among others, one can refer to Pfeffermann (2002, 2013); Jiang and Lahiri (2006); Rao (2003); Rao and Molina (2015) for comprehensive reviews and accounts of methods connected to SAE. The most commonly and widely used approach to handle SAE problems is the linking of statistical models of direct estimates with auxiliary information which is also known as model-based methods. In particular, we mention the area level model which was originally proposed by Fay and Herriot (1979) for the prediction of mean per-capita income in small geographical areas within counties in United States and the nested linear regression model presented by Battese *et al.* (1988) which was used to estimate the area under crop for 12 Iowa counties in United States.

In this paper, we apply a multivariate linear model for repeated measures data described in Ngaruye *et al.* (2016) to produce crop yield estimates at district level for SAS 2014 in Rwanda. The crop yield used here refers to the measure of yield of a crop per unit area of land cultivation. The annually SAS in Rwanda covers three agricultural seasons A, B and C. However, the agricultural seasons considered in this study that fit the proposed model are Season A and Season B which cover the period of September 2013 to February 2014 and March 2014 to June 2014, respectively. Moreover, the study is strictly limited to the crop of beans having two varieties, bush beans and climbing beans.

The considered model in this paper presents several advantages such as enabling to produce small area means at each time point, for each group units and for all time points. By group units, we mean a set of units belonging to the same category with specific characteristics, for example gender categories, crop varieties, age groups of individuals, etc. Indeed, it can even handle more complicated situations than the one which has been considered in the present paper. Although this model is complicated, it is fairly realistic and maybe empirically verified. Depending on the type of variable of interest, it is very important to follow the evolution of the characteristic of interest by modeling the trends over time and very often one might be interested in a given group of the population. This model accounts for such cases.

We also note that the current study is limited to the estimation of beans yield at district level, the results could not be extended at smaller administrative unit such as at sector level since the response values and auxiliary information about covariates were only available at district level. Furthermore, the trend of beans yield estimates at district level was not investigated since the sampled segments included in the SAS were all applicable for only two seasons A and B, and not for season C.

The paper is organized as follows, after the first section of general introduction, follows the second section about the description of the data used. Then we discuss the estimation methodology in the third section and the SAE for multivariate linear model for repeated measures data in section four. The fifth and sixth sections are devoted to the main results and discussions while the last section gives a general conclusion. The summary of detailed results are presented in the appendix before the references.

2. Description of the data

In this study we use micro data pertaining to the National Institute of Statistics of Rwanda (NISR). The variable of interest is crop yield. The intention is to estimate average yield for beans (bush beans and climbing beans varieties) at district level during two agricultural seasons A and B, 2014 in Rwanda. We note that the country of Rwanda is divided into five provinces: Kigali city, Northern, Southern, Western and Eastern provinces. Constituent of provinces are districts and the country is subdivided into 30 districts. Districts are divided into sectors, sectors into cells and cells into villages which are the smallest administrative units.

Rich in protein, iron and other micronutrients, beans represent an important component solution to malnutrition and hunger in developing countries. Both bush and climbing beans are staple crops in Rwanda. According to the final report of SAS 2014 (NISR, 2014), beans were the third main crop grown in 2014 seasons A and B in Rwanda after cassava and banana, covering the area of 461,339 hectares with crop production of 412,681 megatons.

2.1. Seasonal Agricultural Survey (SAS)

The Seasonal Agricultural Survey (SAS) is carried out in Rwanda every year with the main objective of providing accurate and reliable agricultural statistics in terms of land use, crop production and livestock for monitoring the agriculture sector and food supply conditions. The SAS 2014 covered the entire country and using satellite imagery, area frames were constructed by professionals in Geographic Information System (GIS). The SAS 2014 comprises season A that started in November 2013 and ended in March 2014, season B covering the period of April 2014 up to July 2014 and season C which started from September 2014 and continued up to October 2014. The total land was firstly divided into 12 non homogeneous land-use strata spread across the country according to land-use characteristics. Secondly, 5 strata defined by

(i) intensive cropland for season A and B,
(ii) intensive cropland for season A and B with potential to be used for season C,
(iii) marshland cultivated during season A, B and C,
(iv) marshland potentially cultivated with paddy rice and
(v) rangeland

were used for sampling survey and were split into sampling units of an area frame called segments (plots). The sampling design used was a two-stage sampling scheme. At the first stage, 540 primary sampling units (PSU) were selected from the 5 strata using probability proportional to size (PPS) sampling where the size of measure was area under crop. At the second stage, one secondary sampling unit (SSU) was selected randomly. A PSU was a domain of area under crop chosen on the basis of crop intensity having a size between 100 and 200 hectares and a SSU was a subdivision of PSU having around 10 hectares for the four first strata and around 50 hectares in the 5-th stratum. For each selected sampling segment, intervening agriculture operators and large scale farmers were interviewed using farm questionnaires. The survey provides direct estimates for several agricultural statistics at stratum and national levels as presented in Table 2 in the Appendix. However, we notice that the sample sizes taken from districts were too small to produce direct stable estimates of good precision at district level.

2.2. Statistics of covariates

According to the final report of SAS 2014 (NISR, 2014), the crop yield in Rwanda depends on several factors. The first influential factors were found to be the agricultural inputs such as type of seeds (traditional versus improved seeds) and the use of organic and inorganic fertilizers. The second factors are the agriculture practices which vary with seasons and type of plots. There were two type of plots: pure and mixed stand if the land was planted with one crop or with various crop, respectively. Those practices were irrigation and anti-erosion activities due the mountainous landscape of Rwanda. Moreover, the Ministry of Agriculture and Animal Resources (MINAGRI), in collaboration with the National Institute of Statistics of Rwanda conducted a Crop Assessment Survey (CAS) for agricultural season 2013 B and produced the crop yield district level estimates among others. The CAS was a two-stage stratified sampling design where strata were administrative districts.

Therefore, five auxiliary significant variables were chosen to be relevant covariates to include in the SAE for the multivariate linear model. Those are

(i) crop yield estimates for the agricultural season 2013 as reported in Crop Assessment Survey 2013 season B,
(ii) proportion of usage of organic fertilizers by area frame unit,
(iii) proportion of usage of inorganic fertilizers by area frame unit,
(iv) proportion for usage of irrigation and
(v) proportion for usage of anti erosion practices.

We note that except the crop yield estimates for the agricultural season 2013, season B which were in kilograms per hectare, the other statistics of covariates provided in the microdata were average usage of farmers given as proportions (percentages) per stratum and per district.

3. Estimation of population mean and sampling variance

We refer to the general methodology of estimation of weighted mean for a two-stage sampling where for the first stage primary sampling units are selected with probability proportional to size and for the second stage secondary sampling units are selected with simple random sampling. About the estimation procedure, one can refer to Cochran (2007); Bethlehem (2009); Korn and Graubard (2011) or Lehtonen and Pahkinen (2004) among several authors.

Given 5 strata, let A_{hi} be the size (area under crop) of the i-th PSU in stratum h and A_h be the total size (cumulative total area under crop) of stratum $h = 1, \ldots, 5$. Sampled segments and area under crop per stratum can be found in Table 1 in the Appendix.

Then the selection probability for PSU i in stratum h is given by the formula

$$\pi_{hi} = n_h \frac{A_{hi}}{A_h}, \quad h = 1, \ldots, 5, \ i = 1, \ldots, n_h,$$

where n_h is the total number of PSUs selected in stratum h. At the second stage, the selection probability of one sampling segment from the i-th selected PSU is given by

$$\pi_{h(1|i)} = \frac{1}{m_{hi}}, \quad h = 1, \ldots, 5, \ i = 1, \ldots, n_h,$$

where m_{hi} is the total number of segments in PSU i in stratum h as reported in the sampling frame. Therefore, the probability of selection of a given sampling segment coming from stratum h is given by

$$\pi_{hi1} = \frac{n_h A_{hi}}{m_{hi} A_h}, \quad h = 1, \ldots, 5, \ i = 1, \ldots, n_h,$$

and the corresponding sampling weight equals $w_{hi1} = 1/\pi_{hi1}$. For simplification of notation, let the sampling weight w_{hi1} be denoted by w_{hi} and the crop yield value for a sampling segment from i-th PSU in stratum h be denote by y_{hi} . The estimator of the population mean, within a given district, is the weighted mean given by

$$\bar{y} = \frac{\sum_{h=1}^{5}\sum_{i=1}^{n_h} w_{hi}y_{hi}}{\sum_{h=1}^{5}\sum_{i=1}^{n_h} w_{hi}}. \tag{1}$$

The corresponding sampling variance estimator of the weighted mean can be expressed as

$$\widehat{\mathrm{var}}(\bar{y}) = \frac{\sum_{h=1}^{5}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left[w_{hi}(y_{hi}-\bar{y}) - \frac{1}{n_h}\sum_{j=1}^{n_h} w_{hj}(y_{hj}-\bar{y})\right]^2}{\left(\sum_{h=1}^{5}\sum_{i=1}^{n_h} w_{hi}\right)^2}. \tag{2}$$

4. SAE with a Multivariate linear model for repeated measures data

We consider crop yield y as the variable of interest whose values are recorded for p seasons ($t = t_1, ..., t_p$), from the land-use under crop covering the entire country divided into N segments. For the empirical data in this article, $p = 2$ and $N = 45120$. The country is divided into $m = 30$ districts which we will call *small areas* having segments $N_i, i = 1, \ldots, m$. The crop under study has $k = 2$ varieties, bush beans and climbing beans, with corresponding area frame units. For each sampling segment in the present study, only one variety of beans is observed, i.e. either bush beans or climbing beans. Therefore, in the forthcoming model settings, we have two types of segments. A sample $s = s_1, \ldots, s_m$ of segments was selected from the population (land-use), where s_i is the sample of size n_i observed from district i. The sampled segments remain the same during the two seasons.

Denote by $\boldsymbol{y}_{ij}$ to be the p-vector of crop yield values on the j-th segment, in the i-th area (district), $j = 1, \ldots, N_i,\ \ i = 1, \ldots, m$. We make an assumption that the mean growth of the jth segment in district i for each crop type is a polynomial in time of degree $q - 1$ (here we consider $q = 2$). We also consider auxiliary data $\boldsymbol{x}_{ij}$ of $r = 5$ covariables, $\boldsymbol{x}_{1ij}, \ldots, \boldsymbol{x}_{rij}$ whose values are known for all segments in all m small areas (districts). The auxiliary variables considered are crop yield estimates for the agricultural season B of 2013, proportion of usage of organic and inorganic fertilizers by segment and proportion for usage of irrigation and anti erosion practices in all segments. In general settings, for arbitrary N_i, m, p and r, for each one of the k groups, the unit level regression model for j-th unit coming from the small area i at time t is expressed as

$$y_{ijt} = \beta_0 + \beta_1 t + \boldsymbol{\gamma}'\boldsymbol{x}_{ij} + v_{ijt}, \quad j = 1, \ldots, N_i; \quad i = 1, \ldots, m; \quad t = t_1, \ldots, t_p,$$

where β_0, β_1 are unknown parameters for time dependency, $\boldsymbol{\gamma}$ is a vector of fixed regression coefficients of covariables.

The random error v_{ijt} associated with the crop yield y_{ijt} is given by

$$v_{ijt} = u_{it} + e_{ijt},$$

where u_{it} is the random effect of the i-th district due to time characteristics such as soil, climate, etc not accounted for by auxiliary variables and e_{ijt} is the random effect associated with the j-th unit of the i-th district at time t as result of using a sample from the population rather than conducting a complete enumeration of the population. The random area by time effect u_{it} is assumed to be independent identically distributed with mean zero and variance σ^2_{ut} and the random error e_{ijt} is assumed to be independent identically normally distributed with mean zero and known variance σ^2_e independent of u_{it}.

However, from the survey data described in Section 2 the sample units are selected with unequal probabilities within small areas. This results in informative sampling with possibility of having large variations of survey weights w_{ij}. Following Verret *et al.* (2014) and Burgard *et al.* (2014), an augmented model which includes design weights is fitted to the survey data

$$y_{ijt} = \beta_0 + \beta_1 t + \boldsymbol{\gamma}'\boldsymbol{x}_{ij} + \theta w_{ij} + u_{it} + e_{ijt}, \; j = 1, \dots, N_i; \; i = 1, \dots, m; \; t = t_1, \dots, t_p, \quad (3)$$

where θ is an additional unknown regression coefficient accounting for weight effect on the variable of interest. Model (3) is equivalent to

$$y_{ijt} = \beta_0 + \beta_1 t + \boldsymbol{\gamma}'_\omega \boldsymbol{x}_{\omega ij} + u_{it} + e_{ijt}, \quad j = 1, \dots, N_i; \quad i = 1, \dots, m; \quad t = t_1, \dots, t_p,$$

where

$$\boldsymbol{\gamma}_\omega = \begin{bmatrix} \gamma \\ \theta \end{bmatrix} \quad \text{and} \quad \boldsymbol{x}_{\omega ij} = \begin{bmatrix} \boldsymbol{x}_{ij} \\ w_{ij} \end{bmatrix}.$$

The model for all time points is written in matrix form as follows

$$\boldsymbol{y}_{ij} = \boldsymbol{A\beta} + \boldsymbol{1}_p \boldsymbol{\gamma}'_\omega \boldsymbol{x}_{\omega ij} + \boldsymbol{u}_i + \boldsymbol{e}_{ij}, \quad j = 1, \dots, N_i; \quad i = 1, \dots, m,$$

where

$$\boldsymbol{A} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_p \end{pmatrix}, \quad \boldsymbol{1}_p \;\; \text{is a } p\text{-vector of ones, and} \;\; \boldsymbol{u}_i \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_u),$$

where $\boldsymbol{\Sigma}_u$ is supposed to be positive definite. The small area level model for k groups (varieties of beans in this study) is given by

$$\boldsymbol{Y}_i = \boldsymbol{ABC}_i + \boldsymbol{1}_p \boldsymbol{\gamma}'_\omega \boldsymbol{X}_i + \boldsymbol{u}_i \boldsymbol{z}'_i + \boldsymbol{E}_i, \quad (4)$$

where $\boldsymbol{Y}_i = (\boldsymbol{y}_{i1}, \cdots, \boldsymbol{y}_{iN_i})$, $\boldsymbol{B} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_k) : 2 \times k$, $\boldsymbol{X}_i = (\boldsymbol{x}_{\omega i1}, \cdots, \boldsymbol{x}_{\omega iN_i})$, $\boldsymbol{z}_i = \frac{1}{\sqrt{N_i}} \boldsymbol{1}_{N_i}$ and $\boldsymbol{E}_i = (\boldsymbol{e}_{i1}, \cdots, \boldsymbol{e}_{iN_i})$,

$$\boldsymbol{C}_i = \begin{pmatrix} \boldsymbol{1}'_{N_{i1}} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{1}'_{N_{ik}} \end{pmatrix}.$$

According to model (4), $\boldsymbol{Y}_i$ is a $p \times N_i$ data matrix; $\boldsymbol{B}$ is $2 \times k$ unknown parameter matrix; $\boldsymbol{A}$ and $\boldsymbol{C}_i$ are $p \times 2$, $2 \leq p$ and $k \times N_i$ known *within individuals* and *between individuals design matrices for fixed effects* with $\text{rank}(\boldsymbol{C}_i) + p \leq N_i$; $\boldsymbol{X}_i$ is a $(r+1) \times N_i$ known matrix taking the values of the covariates and design weights and $\boldsymbol{E}_i \sim N_{p,N_i}(\boldsymbol{0}, \boldsymbol{\Sigma}_e, \boldsymbol{I}_{N_i})$ stands for the matrix normal distribution with mean zero and with the essential assumption of a known positive definite covariance matrix between rows $\boldsymbol{\Sigma}_e = \sigma^2_e \boldsymbol{I}_p$ and independent columns.

Combining the all m small areas from model (4) such that

$$\boldsymbol{Y} = [\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_m], \quad \boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{C}_m \end{pmatrix}, \quad \boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_m], \quad \boldsymbol{H} = \left(\boldsymbol{I}_k : \cdots : \boldsymbol{I}_k\right),$$
$$\boldsymbol{E} = [\boldsymbol{E}_1, \cdots, \boldsymbol{E}_m], \quad \boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m],$$

we obtain the working model given in the the following definition.

Definition 1. The multivariate linear model for repeated measures data on the response variable of interest can be written as

$$\boldsymbol{Y} = \boldsymbol{ABHC} + \mathbf{1}_p\gamma'_\omega\boldsymbol{X} + \boldsymbol{UZ} + \boldsymbol{E}, \tag{5}$$

where $\boldsymbol{Y} : p \times N$ is the data matrix, $\boldsymbol{A} : p \times 2,\ 2 \leq p$ is the within individual design matrix indicating the time dependency within individuals, $\boldsymbol{B} : 2 \times k$ is unknown parameter matrix, $\boldsymbol{C} : mk \times N$ with $\text{rank}(\boldsymbol{C}) + p \leq N$ and $p \leq m$ is the between individual individual design matrix accounting for group effects, the matrix $\boldsymbol{U} : p \times m$ is a matrix of random effect whose columns are assumed to be independently distributed as a multivariate normal distribution with mean zero and a positive dispersion matrix $\boldsymbol{\Sigma}_u$, i.e. $\boldsymbol{U} \sim N_{p,m}(\mathbf{0}, \boldsymbol{\Sigma}_u, \boldsymbol{I}_m)$, $\boldsymbol{Z} : m \times N$ is a design matrix for random effect and the columns of the error matrix $\boldsymbol{E}$ are assumed to be independently distributed as p-variate normal distribution with mean zero and and known covariance matrix $\boldsymbol{\Sigma}_e$, i.e. $\boldsymbol{E} \sim N_{p,N}(\mathbf{0}, \boldsymbol{\Sigma}_e, \boldsymbol{I}_N)$. The matrix $\boldsymbol{H}$ is included in the model for technical issues of estimation by stacking as column blocks the m data matrices of model (4) together in a new matrix.

Moreover, $\text{vec}(\boldsymbol{Y}) \sim N_{pN}\Big(\text{vec}(\boldsymbol{ABHC}+\mathbf{1}_p\gamma'_\omega\boldsymbol{X}), \boldsymbol{\Sigma}\Big)$ for $\boldsymbol{\Sigma} = \boldsymbol{Z}'\boldsymbol{Z}\otimes\boldsymbol{\Sigma}_u + \boldsymbol{I}_N\otimes\boldsymbol{\Sigma}_e$, where the symbol $\otimes$ denotes the Kronecker product and vec() is the column-wise vectorization operator.

4.1. Estimation of the mean and covariance

Model (5) can be considered as a random effects growth curve model with covariates. For a comprehensive review of different considerations of random effects growth curve model, see for e.g., Yokoyama and Fujikoshi (1992); Yokoyama (1995); Nummi (1997); Pan and Fang (2012). The estimation of the mean and covariance is performed with a likelihood based approach though model decomposition.

In what follows, $\boldsymbol{A}^o$ stands for any matrix of full rank spanning $\mathcal{C}(\boldsymbol{A})^\perp$, i.e., $\mathcal{C}(\boldsymbol{A}^o) = \mathcal{C}(\boldsymbol{A})^\perp$, where $\mathcal{C}(\boldsymbol{A})$ denotes the column vector space generated by the columns of the matrix $\boldsymbol{A}$ and $\mathcal{C}(\boldsymbol{A})^\perp$ is its orthogonal complement. Moreover, $\boldsymbol{A}^-$ denotes an arbitrary generalized inverse of the matrix $\boldsymbol{A}$ such that $\boldsymbol{AA}^-\boldsymbol{A} = \boldsymbol{A}$. We also denote by $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^-\boldsymbol{A}'$ and $\boldsymbol{Q}_{\boldsymbol{A}} = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}}$ the orthogonal projection matrices onto the column space $\mathcal{C}(\boldsymbol{A})$ and onto its orthogonal complement $\mathcal{C}(\boldsymbol{A})^\perp$, respectively. For positive definite matrix $\boldsymbol{S}$, we have projections which are denoted by $\boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{SA})^-\boldsymbol{A}'\boldsymbol{S}$ and $\boldsymbol{Q}_{\boldsymbol{A},\boldsymbol{S}} = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}}$. More details about the growth curve model can be found for example in Kollo and von Rosen (2005) and derivation of estimators and predictors of model (5) are developed in Ngaruye *et al.* (2016).

We consider the partition $\boldsymbol{\Gamma} = [\boldsymbol{\Gamma}_1 : \boldsymbol{\Gamma}_2]$ of the matrix whose columns are eigenvectors associated to the corresponding eigenvalues 0 and 1 of the idempotent matrix $(\boldsymbol{C}\boldsymbol{C}')^{-1/2}\boldsymbol{C}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1/2}$ such that $\boldsymbol{\Gamma}_1$ corresponds to the block $\boldsymbol{I}_m$ and $\boldsymbol{\Gamma}_2$ corresponds to the block $\mathbf{0}$ and let

$$\begin{aligned}
\boldsymbol{K}_1 &= \boldsymbol{H}(\boldsymbol{C}\boldsymbol{C}')^{1/2}\boldsymbol{\Gamma}_1, \quad \boldsymbol{K}_2 = \boldsymbol{H}(\boldsymbol{C}\boldsymbol{C}')^{1/2}\boldsymbol{\Gamma}_2,\\
\boldsymbol{R}_1 &= \boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1/2}\boldsymbol{\Gamma}_1, \quad \boldsymbol{R}_2 = \boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1/2}\boldsymbol{\Gamma}_2.
\end{aligned}$$

The details of the estimation are developed in Ngaruye *et al.* (2016). The following corollary summarizes the results for the particular choice $p = q$ which has been considered in the empirical data analysis.

Corollary 1 (Ngaruye *et al.* (2016)). *Consider the model (5) and suppose that $p = q$, then the within design matrix $\boldsymbol{A}_{p\times p}$ is non singular and the corresponding estimators for γ_ω, $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_u$ are expressed as*

$$\begin{aligned}
\widehat{\gamma}_\omega =& \frac{1}{p}(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{P}\boldsymbol{Y}'\mathbf{1}_p,\\
\widehat{\boldsymbol{B}} =& \boldsymbol{A}^{-1}\left(\boldsymbol{Y} - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\boldsymbol{Y}\boldsymbol{P}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}')^{-1}\boldsymbol{X}\right)\boldsymbol{R}_2\boldsymbol{K}_2'(\boldsymbol{K}_2\boldsymbol{K}_2')^{-}\\
&+ \boldsymbol{A}^{-1}\boldsymbol{V}_3\boldsymbol{K}_1'\boldsymbol{K}_2^o(\boldsymbol{K}_2^{o\prime}\boldsymbol{K}_1\boldsymbol{K}_1'\boldsymbol{K}_2^o)^{-1}\boldsymbol{K}_2^{o\prime},\\
\widehat{\boldsymbol{\Sigma}}_u =& \frac{1}{m}\boldsymbol{V}_3\boldsymbol{Q}_{\boldsymbol{K}_1'\boldsymbol{K}_2^o}\boldsymbol{V}_3' - \boldsymbol{\Sigma}_e,
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{P} =& \boldsymbol{C}'^{o}(\boldsymbol{C}'^{o})' + \boldsymbol{R}_2\boldsymbol{Q}_{\boldsymbol{K}_2'}\boldsymbol{R}_2',\\
\boldsymbol{V}_3 =& \boldsymbol{Y}\boldsymbol{R}_1 - \boldsymbol{Y}\boldsymbol{R}_2\boldsymbol{K}_2'(\boldsymbol{K}_2\boldsymbol{K}_2')^{-}\boldsymbol{K}_1 - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\boldsymbol{Y}\boldsymbol{P}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{R}_1\\
&+ \frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\boldsymbol{Y}\boldsymbol{P}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{R}_2\boldsymbol{K}_2'(\boldsymbol{K}_2\boldsymbol{K}_2')^{-}\boldsymbol{K}_1 \quad \textit{and} \quad \widehat{\boldsymbol{\Sigma}}_u \quad \textit{assumed to be positive definite.}
\end{aligned}$$

4.2. Prediction of random effects and small area means

Since the model in Definition 1 will be applied, the random matrix $\boldsymbol{U}$ has to be predicted. For the prediction of random effects, as pointed out by Nummi (1997) following Henderson's approach (Henderson, 1973), the prediction of random effects $\boldsymbol{U}$ from model (5) is derived by maximizing the joint density $f(\boldsymbol{Y}, \boldsymbol{U}) = h(\boldsymbol{U})g(\boldsymbol{Y}|\boldsymbol{U})$ assuming $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ to be known, which yields

$$\widetilde{\boldsymbol{U}} = (\boldsymbol{\Sigma}_e\boldsymbol{\Sigma}_u^{-1} + \boldsymbol{I}_p)^{-1}(\boldsymbol{Y} - \boldsymbol{A}\widehat{\boldsymbol{B}}\boldsymbol{H}\boldsymbol{C} - \mathbf{1}_p\widehat{\gamma}_\omega'\boldsymbol{X})\boldsymbol{Z}',$$

where $\widehat{\boldsymbol{B}}$ and $\widehat{\gamma}_\omega$ are presented in Corollary 1. The covariance matrix for random effects, $\boldsymbol{\Sigma}_u$, is however unknown and is therefore replaced by its estimator presented in Corollary 1 yields

$$\widehat{\boldsymbol{U}} = (\boldsymbol{\Sigma}_e\widehat{\boldsymbol{\Sigma}}_u^{-1} + \boldsymbol{I}_p)^{-1}(\boldsymbol{Y} - \boldsymbol{A}\widehat{\boldsymbol{B}}\boldsymbol{H}\boldsymbol{C} - \mathbf{1}_p\widehat{\gamma}'_\omega\boldsymbol{X})\boldsymbol{Z}'. \tag{6}$$

Now, for k group units in all small areas, we consider the partition of N_i units into $N_{ig}, g = 1, \cdots, k$, and n_i units into n_{ig} such that $N_i = \sum_{g=1}^k N_{ig}$ and $n_i = \sum_{g=1}^k n_{ig}$. Similarly the matrix $\boldsymbol{Y}_i$ from model (4) is divided firstly into corresponding blocks for for k group units and secondly into blocks corresponding to n_i sampled and $(N_i - n_i)$ non sampled observations $\boldsymbol{Y}_i^{(s)} = (\boldsymbol{y}_{i1}, \cdots, \boldsymbol{y}_{in_i}) : p \times n_i$ and $\boldsymbol{Y}_i^{(r)} = (\boldsymbol{y}_{in_{i+1}}, \cdots, \boldsymbol{y}_{iN_i}) : p \times (N_i - n_i)$, respectively. In the next Proposition 1, the target small area means at each time point for each k group units are presented.

Proposition 1. *Let the repeated measures data on the variable of interest from a finite population divided into small areas be described by the model (5) given in Definition 1. The target small area means at each time point for each group units can be expressed by*

$$\widehat{\boldsymbol{\mu}}_{ig} = \frac{1}{N_{ig}}\left(\boldsymbol{Y}_{ig}^{(s)}\mathbf{1}_{n_{ig}} + \widehat{\boldsymbol{Y}}_{ig}^{(r)}\mathbf{1}_{N_{ig}-n_{ig}}\right), \quad i = 1, \cdots, m \quad g = 1, \cdots, k. \tag{7}$$

The first term of the expression (7) on the right side corresponding to sampled observations is known and the second term which corresponding to non-sampled observations is unknown and is predicted using the considered model, i.e.,

$$\widehat{\boldsymbol{Y}}_{ig}^{(r)}\mathbf{1}_{N_{ig}-n_{ig}} = \left(\boldsymbol{A}\widehat{\boldsymbol{\beta}}_g\mathbf{1}'_{N_{ig}-n_{ig}} + \mathbf{1}_p\widehat{\gamma}'\boldsymbol{X}_{ig}^{(r)} + \widehat{\boldsymbol{u}}_i\boldsymbol{z}_{ig}^{(r)'}\right)\mathbf{1}_{N_{ig}-n_{ig}}\Big), \quad i = 1, \cdots, m \quad g = 1, \cdots, k,$$

where $\widehat{\gamma}, \widehat{\boldsymbol{\beta}}_g$ and $\widehat{\boldsymbol{u}}_i$ are estimators computed from the Corollary 1 and the expression (6), respectively using observed data and taking $\widehat{\boldsymbol{\beta}}_g$, $g = 1, 2$ to be the column of the estimated parameter matrix $\widehat{\boldsymbol{B}}$ corresponding bush beans and climbing beans. Moreover, $\boldsymbol{X}_i^{(r)}$ and $\boldsymbol{z}_i^{(r)}$ are the corresponding matrix of auxiliary information and design vector for non sampled segments, respectively while the predicted vector $\widehat{\boldsymbol{u}}_i$ is the i-th column of the predicted matrix $\widehat{\boldsymbol{U}}$.

We recall that regarding the considered empirical data, the k group units correspond to two varieties of beans, the bush beans and climbing beans, while the m small areas are the 30 districts of Rwanda.

5. Main results

This section presents the results obtained from the analysis of the data described in Section 2 where the theory presented in Section 3 and Section 4 is applied. For the present empirical study, the land-use was divided into 45120 segments scattered into 30 districts. Direct district level estimates and model-based district level estimates are presented in the Appendix in Table 3 and Table 4, respectively. The map in Figure 1 shows the distribution of average beans yield model-based estimates during agricultural seasons A and B, year

2014, per district.

The model-based estimates obtained by improving the unreliable direct estimates through incorporating relevant auxiliary information and accounting for district random effect and time variations agree with the direct estimates from SAS 2014 in terms of ranking district per beans yield. The calculation of an analytic expression for the mean squared errors of the crop yield estimates at district level is not an easy task. However, we believe that our estimates have smaller mean squared errors compared to direct estimates following the results of simulations of the estimated mean squared errors as shown in Ngaruye *et al.* (2016).

The results from the current study reveal the highest beans yields in the districts of the Northern and southern West provinces (with exception of two districts Rutsiro and Karongi), while the lowest beans yields were observed in the districts of South and southern Est provinces. In the whole country, Rubavu district is the district with highest beans yield with 1,689 kilograms per hectare while the district with lowest beans yield is Huye with 604 kilograms per hectare. With exception of Bugesera district, the first districts having the highest proportions of erosion control, namely Nyabihu, Musanze and Rubavu districts are among top six districts to have highest crop yield. In all districts, the crop yield has declined during season B compared to season A for both varieties of beans. The results show also that climbing beans offered higher yields than bush beans in 23 districts during season A and in all 30 districts during season B.

6. Discussion and assessment of results

The assessment of reliability of the obtained model-based estimates is conducted via model checking of underling assumptions and bias diagnostics.

Statistical diagnostics are carried out to compare direct survey estimates and model-based estimates. The closeness of the model-based estimates to the true small area values is determined by their unbiasedness with respect to direct estimates. As pointed out by Brown *et al.* (2001), since the direct estimates of small area means are unbiased of the trues small area means, if the trues small area means were known and plotted on the X axis against the direct estimates of small area means on the Y axis, then their regression fit line would coincide with the regression line $Y = X$. To ensure that, the regression of the direct estimates on model-based estimates should be similar. The bias scatter plots are presented in the Appendix in Figure 1, Figure 2, Figure 3 and Figure 4 where direct estimates are plotted on the X-axis and model-based estimates on the Y-axis. The plots show that there is no large bias influencing on the model-based small area estimates and model-based small area means for season B tend to be better estimated than those for season A.

To check model mis-specification, we plot residuals against model-based estimates. Theoretically, if the model assumptions hold, we expect the residuals to be randomly scattered around their expected mean equal to zero. That is, the residuals do not exhibit any systematic structure. The scatter plot of residuals versus model-based estimates presented in the Appendix in Figure 5 indicates that residuals are random. The results show big variations among segments compared to the sampling variations.

7. Concluding remarks

This paper illustrates how a multivariate linear model for repeated measures data can be applied to Small Area Estimation techniques to produce reliable district level crop yield estimates per group crop variety and per time point. The techniques are applied to micro data from National Institute of statistics of Rwanda of Seasonal Agricultural Survey 2014 to produce beans crop yield estimates at district level in 30 districts of Rwanda which has two categories of beans, the bush beans and climbing beans. All the diagnostic measures conducted show that the model-based estimates produced at district level are reliable and representative of the corresponding districts. With these district level estimates available, it is straightforward to deduce the corresponding crop production since the area under crop is known for all districts.

However, we note that the Seasonal Agricultural Survey 2014 in Rwanda was composed of three agricultural seasons A,B, and C but season C was excluded in the study since not all the selected 540 segments during season A and B were followed during season C and this does not fit into our developed multivariate linear regression model. This case of drop out of units may be considered for our future works. Because of these limited data, the trends of crop yield by varieties have not been investigated. Moreover, with background information available at lower lever, it is of great importance to extend these results at smaller administrative units such as at sector and cell levels.

Note that the Table 4 of model-based point estimates and all figures presented in the Appendix are obtained using the MATLAB software, Version 9.0.0.341360 (The MathWorks, In. USA).

Appendix

A Sample sizes and detailed results of the empirical study

According to the final report of the SAS 2014 (NISR, 2014), the crop land-use was stratified into 12 strata from which 5 strata were used in the SAS 2014. The report also provides the shares (in percentage) of area occupied by strata within all 30 districts. Therefore, we have summarized the data per district and the direct point estimates presented below in Table 3 are computed using formula (1) while the Model-based point estimates presented in Table 4 are computed using formula (7).

Acknowledgements

We thank Theodomir Mugiraneza, researcher at University of Rwanda-Centre for GIS and Remote Sensing and visiting scholar at KTH, Royal Institute of Technology, Geoinformatics Division, Stockholm, Sweden for designing and producing the statistical cluster map for seasonal crop yield per district in Rwanda using obtained estimates.

The research of Innocent Ngaruye is supported by the Swedish International Development and Cooperation Agency (SIDA) in collaboration with the University of Rwanda. The research of Dietrich von Rosen is supported by the Swedish Foundation for Humanities and Social Sciences. All involved institutions are acknowledged.

Fig. 1. Beans yield estimation during agricultural seasons A and B, 2104

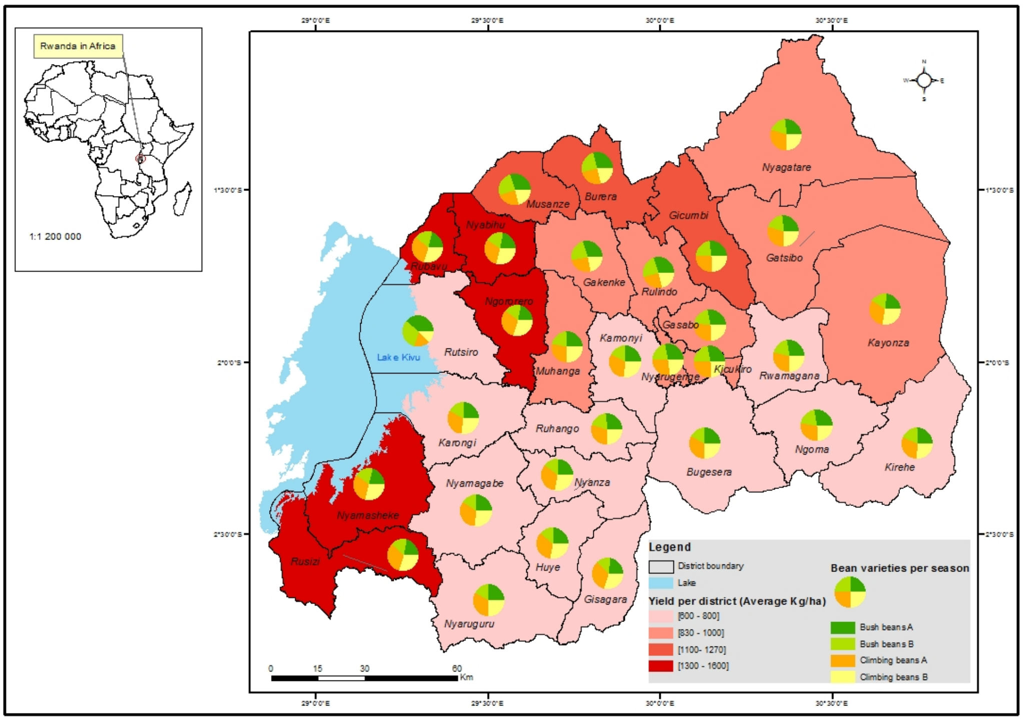


Table 1. Selected sampling segments per Stratum

Stratum (h)	**Total area in hectare** (N_h)	**Number of sampled units** (m_h)
1.1	1,479,081	340
1.2	148,388	48
2.1	95,820	64
2.2	20,200	40
3.0	28,763	48
Total	**1,806,102**	**540**

Table 2. Beans yield estimates per stratum reported by NISR for SAS 2014 (kg/ha)

Stratum	Bush beans		Climbing beans	
	Season A	Season B	Season A	Season B
1.1	938	1,080	709	922
1.2	—*	817	651	1,256
2.1	1,067	1,541	742	930
2.2	1,184	742	939	648
3.0	825	930	704	2,184
National level	942.38	1,066.09	712.90	932.60

* Bush beans variety was not cultivated in land-use from stratum 1.2 during agricultural season A

Table 3. Direct point estimates (kg/ha) according to equation (1)

	District	Bush beans		Climbing beans	
		Season A	Season B	Season A	Season B
1	Nyarugenge	1,012	741	1,012	836
2	Gasabo	1,009	739	1,021	844
3	Kicukiro	1,022	748	1,001	828
4	Nyanza	674	493	985	814
5	Gisagara	650	469	920	760
6	Nyaruguru	555	406	856	708
7	Huye	575	420	881	727
8	Nyamagabe	565	414	896	740
9	Ruhango	734	538	882	729
10	Muhanga	979	717	1,093	903
11	Kamonyi	611	448	907	750
12	Karongi	654	479	1,003	829
13	Rutsiro	1,075	789	442	375
14	Rubavu	1,476	1,082	2,027	2,023
15	Nyabihu	1,201	883	1,590	1,672
16	Ngororero	1,279	939	1,835	1,540
17	Rusizi	1,268	927	1,795	1,482
18	Nyamasheke	1,327	972	1,839	1,520
19	Rulindo	1,247	913	980	810
20	Gakenke	1,225	897	986	815
21	Musanze	1,437	1,052	1,035	1,072
22	Burera	1,382	1,012	1,113	920
23	Gicumbi	1,319	966	1,247	1,031
24	Rwamagana	743	544	865	715
25	Nyagatare	968	715	833	694
26	Gatsibo	1,000	739	868	725
27	Kayonza	800	595	950	797
28	Kirehe	673	493	938	776
29	Ngoma	761	557	859	710
30	Bugesera	804	588	839	694

Table 4. Model-based point estimates (kg/ha) given by equation (7)

	District	Bush beans		Climbing beans		Average
		Season A	Season B	Season A	Season B	
1	Nyarugenge	955	722	964	826	867
2	Gasabo	950	717	968	829	866
3	Kicukiro	956	724	965	827	868
4	Nyanza	658	432	901	769	690
5	Gisagara	587	365	834	705	623
6	Nyaruguru	644	421	828	698	648
7	Huye	585	363	799	671	604
8	Nyamagabe	656	432	870	740	674
9	Ruhango	759	532	850	717	715
10	Muhanga	935	702	1,050	912	900
11	Kamonyi	686	462	877	746	693
12	Karongi	725	499	976	844	761
13	Rutsiro	1,059	830	486	352	682
14	Rubavu	1,406	1,187	2,141	2,022	1,689
15	Nyabihu	1,295	1,093	1,798	1,698	1,471
16	Ngororero	1,194	951	1,786	1,637	1,392
17	Rusizi	1,080	836	1,663	1,514	1,273
18	Nyamasheke	1,175	931	1,743	1,593	1,360
19	Rulindo	1,138	901	957	815	953
20	Gakenke	1,136	899	967	825	957
21	Musanze	1,425	1,215	1,201	1,092	1,233
22	Burera	1,239	999	1,079	934	1,063
23	Gicumbi	1,200	960	1,221	1,075	1,114
24	Rwamagana	755	528	828	695	701
25	Nyagatare	978	747	871	734	832
26	Gatsibo	972	739	874	736	830
27	Kayonza	870	644	970	838	831
28	Kirehe	741	515	920	787	741
29	Ngoma	779	551	831	697	714
30	Bugesera	741	513	778	645	669

Fig. 1. Bias diagnostic plot for Direct estimates versus Model-based estimates Season A (Linear regression fit line for Bush beans)

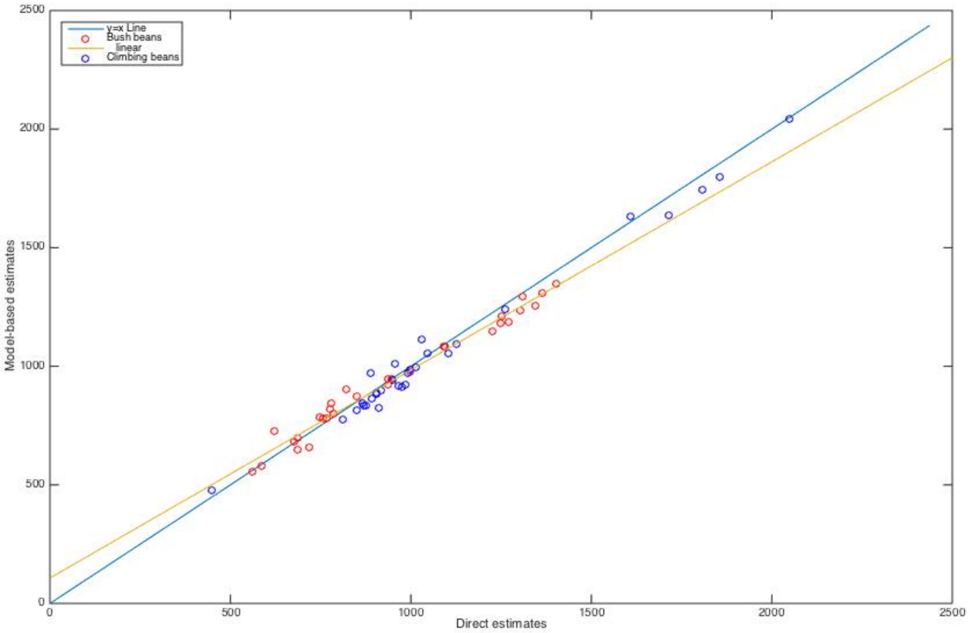


Fig. 2. Bias diagnostic plot for Direct estimates versus Model-based estimates Season A (Linear regression fit line for Climbing beans o)

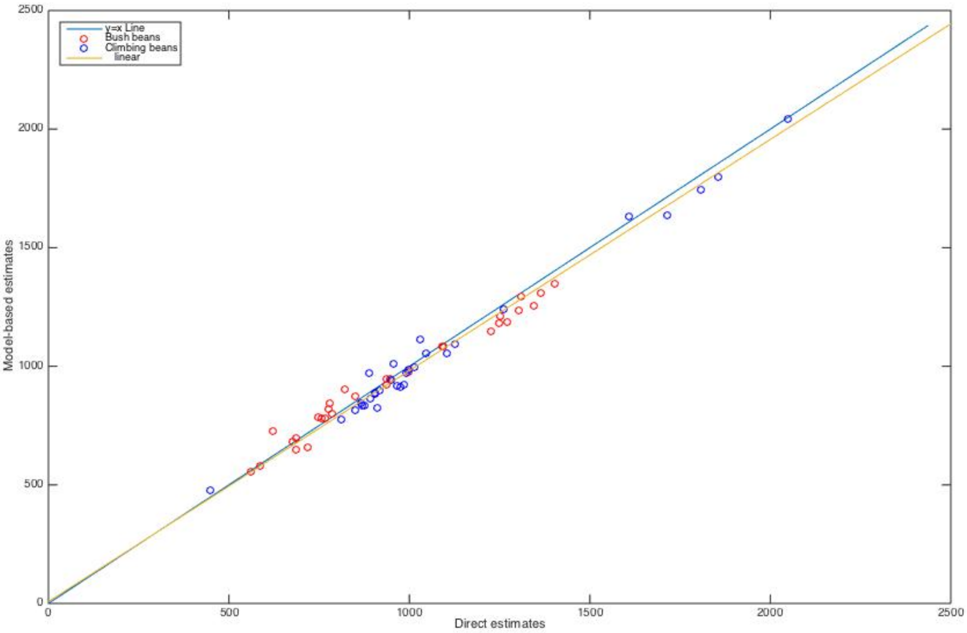


Fig. 3. Bias diagnostic plot for Direct estimates versus Model-based estimates Season B (Linear regression fit line for Bush beans)

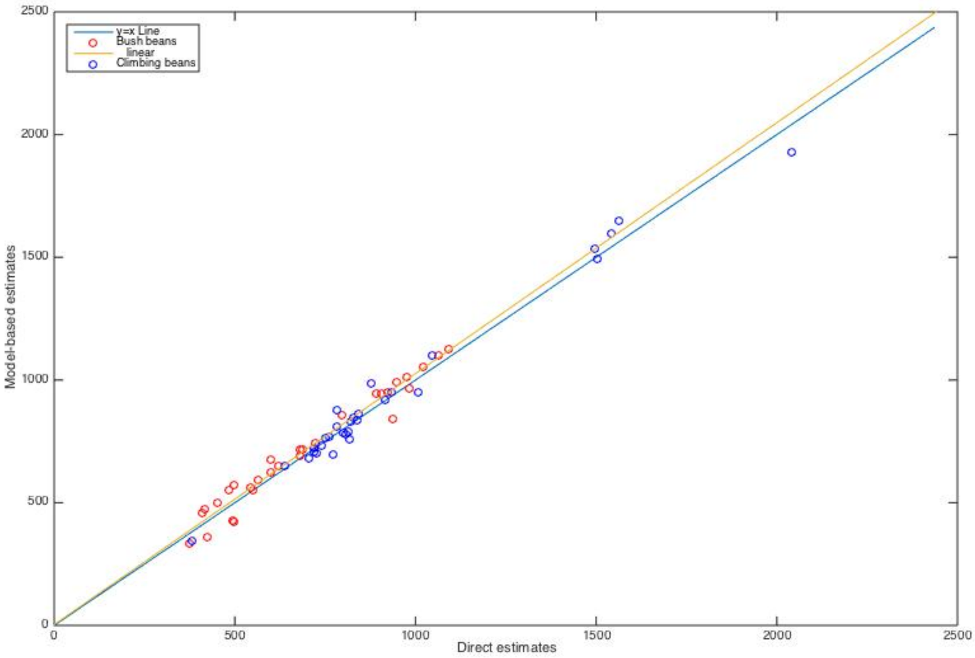


Fig. 4. Bias diagnostic plot for Direct estimates versus Model-based estimates Season B (Linear regression fit line for Climbing beans)

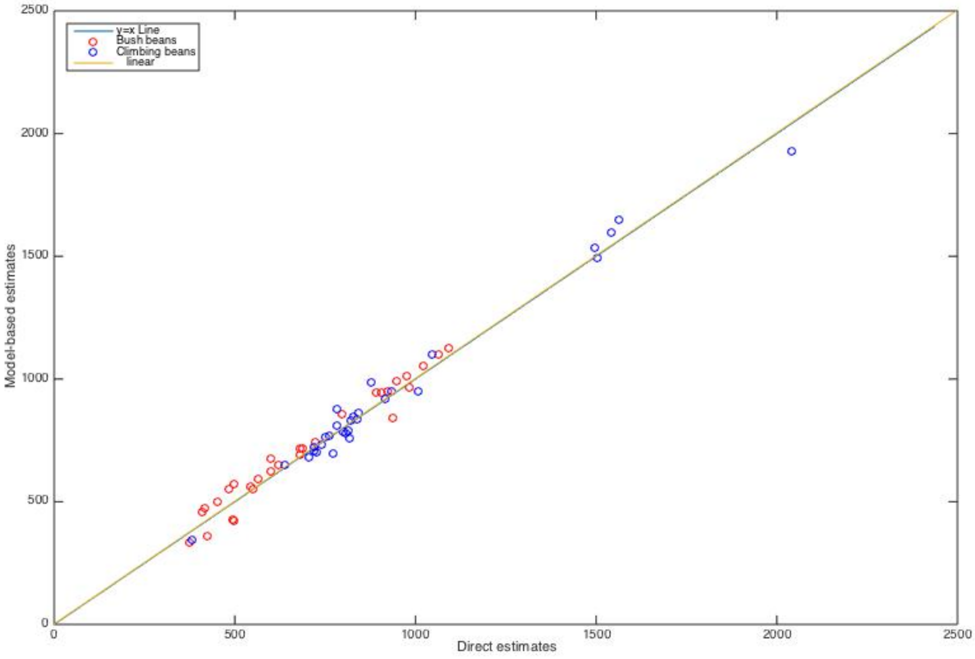
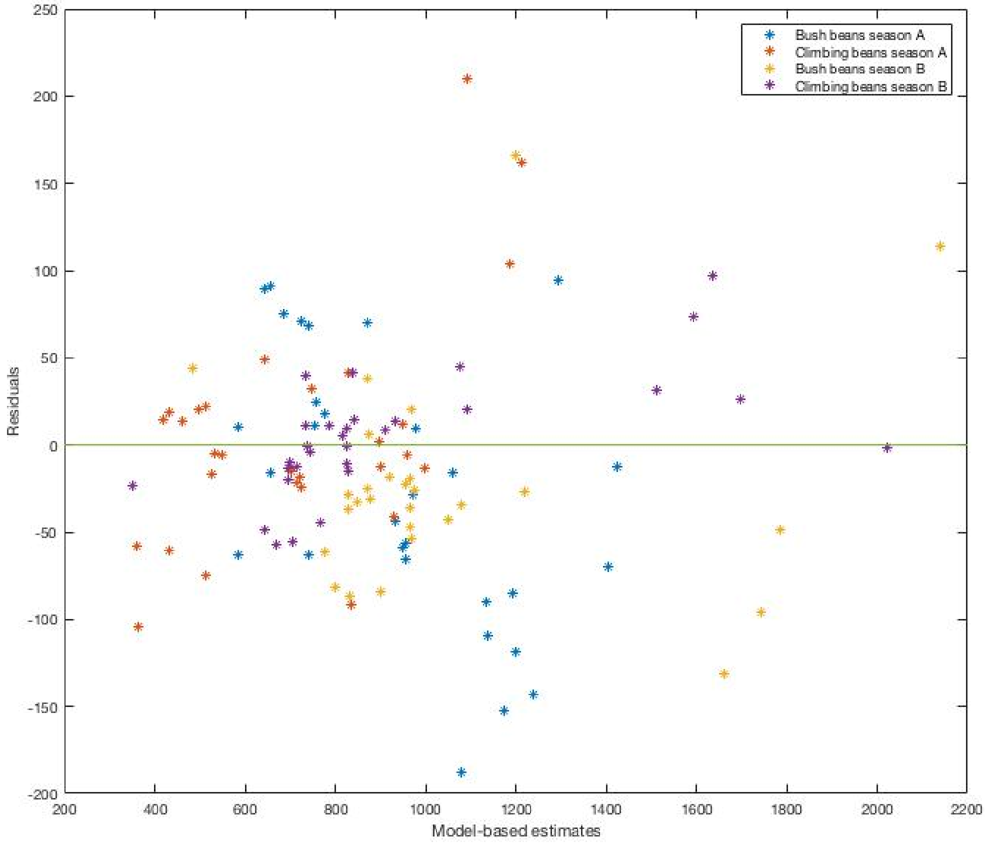


Fig. 5. Residuals versus Model-based estimates



References

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

Pfeffermann, D., (2002). Small Area Estimation - New Developments and Directions. *International Statistical Review* Vol. 70(1) , pp. 125–143

Pfeffermann, D., (2013). New important developments in small area estimation. *Statistical Science* Vol. 28(1) , pp. 40–68

Rao, J. N. K., (2003). *Small Area Estimation.* John Wiley and Sons. New York

Rao, J. N. K. and Molina,I., (2015). *Small area estimation.* John Wiley & Sons. New York

Kollo, T. and von Rosen D., (2005). *Advanced Multivariate Statistics with Matrices.* Springer. Dordrecht

Henderson, C. R., (1973). Sire evaluation and genetic trends. *Journal of Animal Science* Vol. 1973 (Symposium) , pp. 10–41

Fay, R. E. and Herriot, R. A., (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* Vol. 74(366a) , pp. 269–277

Lehtonen, R. and Pahkinen, E., (2004). *Practical methods for design and analysis of complex surveys.* John Wiley & Sons. New York

Brown, G., Chambers, R., Heady, P. and Heasman, D., (2001). *Evaluation of small area estimation methods-An application to unemployment estimates from the UK LFS.* In : Proceedings of Statistics Canada Symposium (2001) Achieving Data Quality in a Statistical Agency: A Methodological perspective . Statistics Canada

Burgard, J. P., Münnich, R., and Zimmermann, T., (2014) The impact of sampling designs on small area estimates for business data. *Journal of Official Statistics* Vol. 30(4), pp. 749–771

NISR, (2014). *Seasonal Agricultural Survey.* National Statistical Office of Rwanda. Kigali, Rwanda

Bethlehem, J., (2009). *Applied survey methods: A statistical perspective.* John Wiley & Sons. New York

Korn, Edward L. and Graubard, B. I., (2011). *Analysis of health surveys.* John Wiley & Sons. New York

Verret, F. and Rao, J. N. K. and Hidiroglou, M. A., (2014). Model-based small area estimation under informative sampling. *Survey Methodology* Vol. 2, pp. 12–001

Jiang, J. and Lahiri, P., (2006). Mixed model prediction and small area estimation. *Test* Vol. 15(1), pp. 1–96

Nummi, T., (1997). Estimation in a random effects growth curve model. *Journal of Applied Statistics* Vol. 24(2), pp. 157–168

Pan, J.-X. and Fang, K.-T., (2012). *Growth curve models and statistical diagnostics.* Springer Science & Business Media. New York

Yokoyama, T., (1995). Statistical inference on some mixed MANOVA-GMANOVA models with random effects. *Hiroshima Mathematical journal* Vol. 25(2) , pp. 441–474

Cochran, W. G., (2007). *Sampling techniques.* John Wiley & Sons. New York

Yokoyama, T. and Fujikoshi, Y., (1992). Tests for random-effects covariance structures in the growth curve model with covariates. *Hiroshima Mathematical journal* Vol. 22(2) , pp.

195–202

Ngaruye, I., Nzabanita, J., von Rosen, D. and Singull, M. (2016). Small Area Estimation under a Multivariate Linear Model for Repeated measures Data. *Communications in Statistics - Theory and Methods*. Accepted for publication. http://dx.doi.org/10.1080/03610926.2016.1248784.