Applying Small Area Estimation for Agricultural Census Data

Estimating Maize Productivity at District Level in Tanzania Mainland

Students: C. Galgo, T. Nguyen, A. Stampa, R. Wankya

Tutors: Prof. Monica Pratesi; Dr. Stefano Marchetti; Dr. Gaia Bertarelli

Course: Analysis of Survey Data and Small Area Estimation 2018

Università di Pisa

Introduction

Agriculture plays a crucial role in Tanzania's economy where it contributes around 80% of export earnings and most industries in the country are linked to the sector (Leyaro et al. 2014). The need for accurate and timely estimates even for small areas for effective policy making in agriculture is undeniable. In this case, direct estimates (DIR) usually yield little reliability, i.e. large standard errors due to small sample sizes. Small area estimation (SAE) is addresses this issue by 'borrowing strength' from related areas to increase robustness of estimators for a given area or simultaneously, for several areas (Prasad et al. 1990). In this study, the average yield of maize, which is a major staple crop in Tanzania, at the district level of mainland is investigated. To do so, we applied a Fay-Herriot (F-H) model. We calculated and compared DIR and Area Level Empirical Best Linear Unbiased Predictors (AL-EBLUP) at the district level for the harvested area and the harvested quantity of maize. Subsequently, we calculated the ratio to obtain an estimate for yield (kg/ha).

Data

Annual Agricultural Sample Survey

- POINT SAMPLE AREA FRAME METHODOLOGY 21,210 selected sample points, 15,281 complete sample points (response rate 72%)
- SEASONAL DATA from two growing seasons, stacked in a single dataframe
- STUDY VARIABLES maize production: harvested area (Hv.A.) in hectare, harvested quantity (Hv.Q.) in kg
- SMALL AREAS 159 districts
- OBSERVATIONS n = 5,422

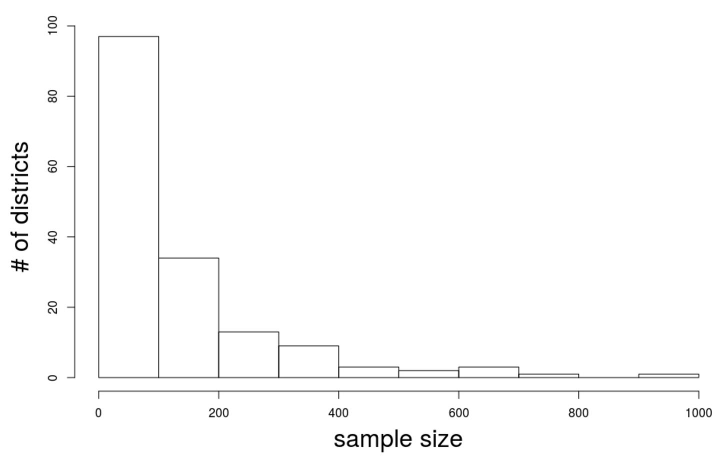


Figure 1: Histogram of district sample sizes

Auxiliary Data

- AGRICULTURAL ROUTINE DATA SYSTEM (ARDS) aggregated at district level (imputed and cleaned version)
- SATELLITE DATA ON LANDUSE contains landuse data for each of the 159 districts (Landuse classes: Forest, Grassland, Cropland, Wetland, Settlement, Otherland, Cloud, Cloudshadow, Total)

Methodology

Fay-Herriot Area Level Model

We use a F-H model to compute the Area Level Enhanced Best Linear Predictor (AL-ELUP) which is a linear combination of the DIR and a predicted component, based on a linear mixed model. Under F-H, the harvested quantity and area are related to the auxiliary data on district level. The model also accounts for within-area homogeneity.

- A linear relationship between θ_d and a set of covariates is assumed, described as:
  $\theta = X_d^T * \beta + u_d$ with
  X_d^T = vector of covariates for domain d
  β = regression coefficient vector
  u_d = domain effects assumed to be distributed with
  $\mu = 0$ and $variance = \sigma_u^2$
  The random effects account for the extra variability not explained by the auxiliary variables in the model
- With the design unbiased direct estimator
  $\hat{\theta}_d = \theta_d + e_d$ with
  $e_d s$ = sampling errors associated with direct estimators, for which
  $E(e_d|\theta_d) = 0$ [DIR is unbiased] and
  $V(e_d|\theta_d) = \varphi_d$ [variances are known],
- through combining the two equations, we obtain the linear mixed model:
  $\theta = X_d^T\beta + u_d + e_d$
  We assume normality for u and e in order to compute the mean square error (MSE) but for the estimation of the target parameter it is not necessary.
- the final AL-EBLUP formula is
  $\hat{\theta}_d^{EBLUB\_AREA} = \gamma_d\hat{\theta}_d + (1-\gamma_d)X_d^T\hat{\beta}$ with $\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\phi_d}$

- DATA PREPARATION The datasets provided for this project work are already cleaned. Therefore, data preparation only consists of merging direct estimates and estimated mean square errors with the auxiliary datasets.
- COMPUTATION OF DIRECT ESTIMATES In a first step, we computed Horvitz-Thompsonn direct estimates. Table 1 shows that the majority of direct estimates on district level for both harvested quantity and harvested area have to be considered as not sufficiently reliable (with a CV $\geq$ 16,5 used as a rule of thumb after Statistics Canada).
- SELECTION OF THE AUXILIARY VARIABLES We run a linear regression to identify the correlation of the auxiliary variables with the target variable and select the auxiliary variables for the EBLUP in three steps, see Table 2.

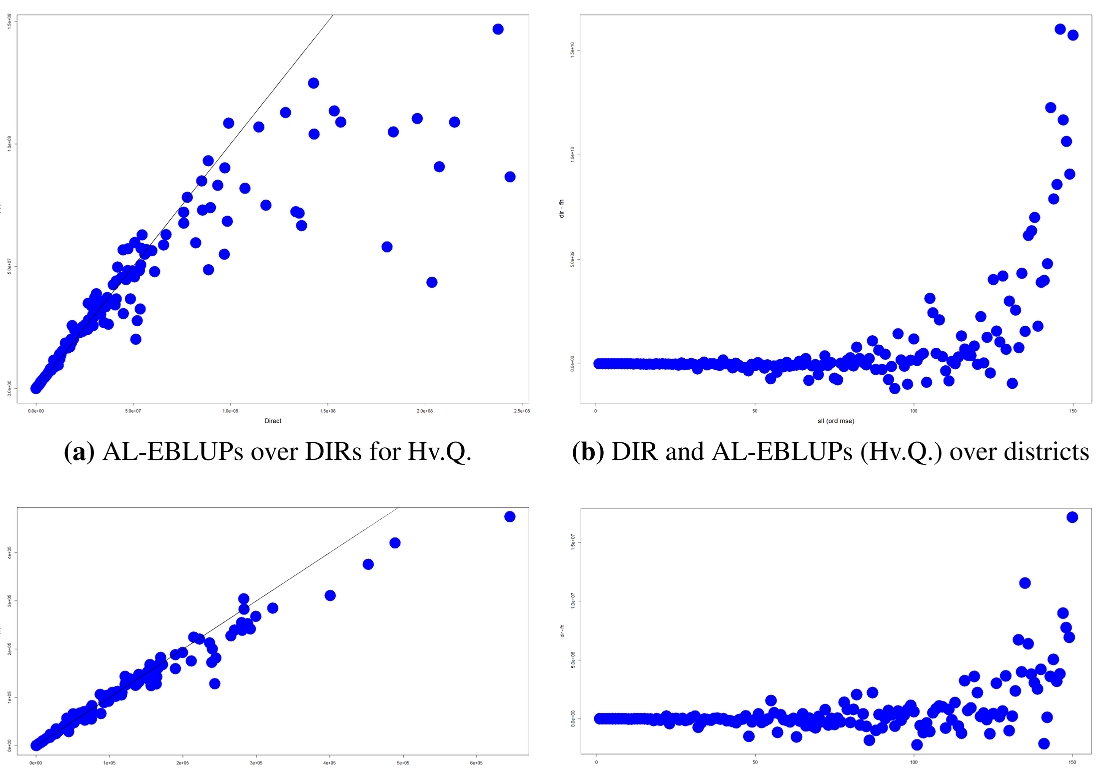
Table 1: Number of regions and districts being (a) reliable (CV $\leq$ 16.5), (b) restrictedly reliable (16.5 $\leq$ CV $\leq$ 33.3) and (c) not reliable (CV $\geq$ 33.3)

CV (%)	# of regions		# of districts	
	DIR Hv.Q.	DIR Hv.A.	DIR Hv.Q.	DIR Hv.A.
0 - 16,5	14	21	9	26
16,5 - 33,3	13	7	65	68
33,3 - 100	3	2	87	67

Table 2: Steps: Selection of auxiliary variables

1. round	choice based on **coefficient correlation** with target variable
2. round	**refinement of model** by removing auxiliary variables that do not show significant correlation (threshold $p < 0.1$)
3. round	application of **a priori** knowledge: subjective choice according to how much the variables seem to be related with maize production

Results



(a) AL-EBLUPs over DIRs for Hv.Q. (b) DIR and AL-EBLUPs (Hv.Q.) over districts

(c) AL-EBLUPs over DIRs for Hv.A. (d) DIR and AL-EBLUPs (Hv.A.) over districts

Figure 2: Comparison between DIRs and AL-EBLUPs

- Due to restricted auxiliary data availability, 150 AL-EBLUP estimates were calculated for both the harvested quantity and the harvested area.
- The diagrams of Figure 2 (a) and (c) show the AL-EBLUPs plotted against the DIR. As a general trend, the DIR seems to be larger than the AL-EBLUP. This is because the DIR systematically overestimates the true value due to the error.
- Figure 2 (b) and (d) plot the difference in value of DIR and AL-EBLUP over the districts, ordered descending by sample size. The graphs show that with decreasing sample size, the difference between the DIR and the AL-EBLUP increases.
- For the Hv.Q. AL-EBLUP, the second round proved to be most accurate (using the area-level error variance as the choice criterion, as suggested by Marchetti (2018)). Simultaneously, for the Hv.A. AL-EBLUP, the first round proves to have the best fit.

Table 3: Number of districts being (a) reliable (CV $\leq$ 16.5), (b) restrictedly reliable (16.5 $\leq$ CV $\leq$ 33.3) and (c) not reliable (CV $\geq$ 33.3) under DIR, in-sample AL-EBLUPs (SAE.in) and AL-EBLUPs including synthetic estimates (SAE.all)

CV (%)	DIR	SAE.in	SAE.all
0 - 16,5	9	29	29
16,5 - 33,3	61	65	65
33,3 - 100	80	56	61

Table 4: Estimate examples for small and big sample size districts

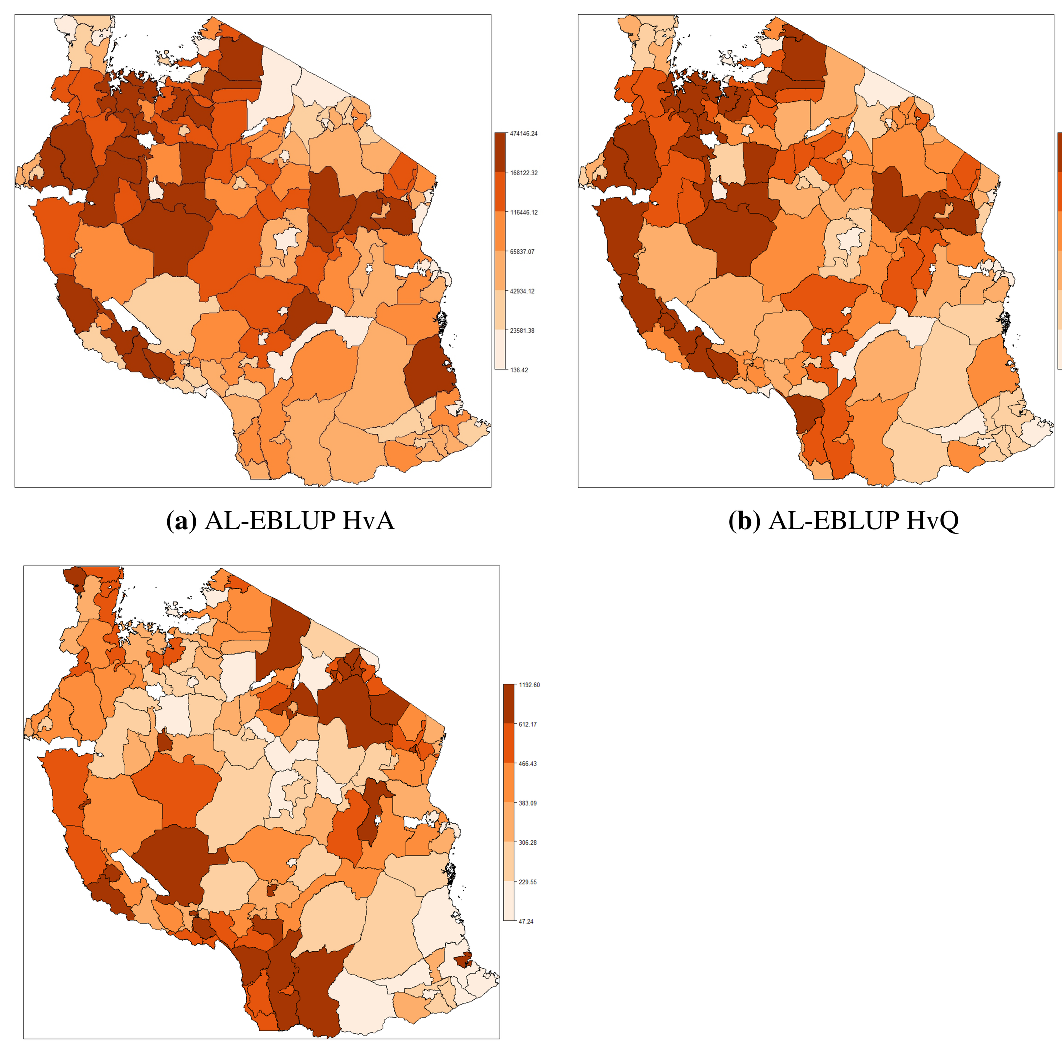
Parameter	Arusha M. (n=2)	Kwimba (n=130)
DIR Hv.Q.	2,129,661.33	85,308,096.49
AL-EBLUP Hv.Q.	2,085,950.76	84,882,270.04
DIR Hv.A.	2,700.9215	282,939.6524
AL-EBLUP Hv.A.	2,725.0957	304108.4255
DIR ratio	788.4943	301.5063
AL-EBLUP ratio	765.4596	279.1184
MSE of AL-EBLUP Hv.Q.	3.102033e+12	1.110110e+14
MSE of AL-EBLUP Hv.A.	4.906081e+06	1.310601e+09
CV DIR Hv.Q.	82.93013	13.79035
CV DIR Hv.A.	13.606908	82.064160

Normality of Area-Level Errors: Shapiro-Wilks Test

- **(Hv.A.)**: W = 0.9965 (p = 0.9791)
- **(Hv.Q.)**: W = 0.98055 (p=0.03204)

Goodness of Fit: Wald test

- **(Hv.A.)**: W = 31.95629 (p = 1)
- **(Hv.Q.)**: W = 114.1294 (p = 0.9959091)



(a) AL-EBLUP HvA **(b)** AL-EBLUP HvQ

(c) Yield ratio

Figure 4: AL-EBLUPS for Hv.Q., Hv.A. and yield ratio at the district level

Conclusions and Forthcoming Research

Using a F-H model, we computed small area estimates of harvested quantity and harvested area of maize at district level and subsequently computed the ratio to look at maize productivity in kg/ha at district level. With auxiliary data on landuse and agricultural census data, we obtained direct estimates and Area Level Enhanced Best Linear Predictors.

Concerning the quality of the estimates, we observe an increase in difference with decreasing sample size at the area level, yet the difference is not statistically significant. This is because small area estimates should not be much different from direct estimates, particularly for those obtained with a reasonable sample size, namely more than 50 or 100 observations. While carrying out our analysis, we noticed that for other important crops like paddy, cassava, sisal etc., the auxiliary data used in this analysis would not be strong enough. Future research could address these issues with more auxiliary data of better quality.

Selected References

- 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, 74, 269-277.

- Leyaro, V. et al. (2014): Food crop production in Tanzania. Evidence from the 2008/09 National Panel Survey. Internatioal Growth Centre. Reference Number F-40110-TZA-1.

- Marchetti, Stefano (2018): Instruction Manual for small area estimation in Tanzania Mainland and Zanzibar. Part of: TCP/URT/3505 - Support to the Implementation of Agriculture Statistics Strategic Plan: Improving district level data using Small Area Estimation methods-ZANZIBAR" and GCP/URT/145/IRE - Support to the Implementation of Agriculture Statistics Strategic Plan - Improvement of District Level Data using Small Area Estimation Method. Unpublished.

- Molina, Isabell & Marhuenda, Yolanda (2015): SAE: an R Package for Small Area Estimation. The R Journal Vol. 7/1, ISSN 2073-4859, https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf

- Sander Scholtus (?): EBLUP Area Level for Small Area Estimation (Fay-Herriot). Memobust Handbook on Methodology of Modern Business Statistics, https://ec.europa.eu/eurostat/cros/system/files/Weighting%20and%20Estimation-11-M-EBLUP%20Area%20Level%20for%20SAE%20v1.0.pdf)

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

- Pratesi, M. (2015): Spatial Disaggregation and Small Area Estimation Methods for Agricultural Surveys: Solutions and Perspectives, Technical Report in the Global Strategy Publications.

Special Thanks to **Dr. Stefano Marchetti** who provided us not only with his *ad hoc* formulae for the analysis, but also with a very useful overview at the intial stage of our project work.