Food and Agriculture Organization of the United Nations

SUSTAINABLE DEVELOPMENT GOALS

Unit level SAE models with R

2022 Virtual Training on Data Disaggregation and Small Area Estimation for the SDGs –
22-25 November 2022, *INStaD Benin, INSTAT Mali, Stats South Africa, Statistics Botswana*

Small Area Estimation using R

Multiple **R packages** are available to implement **unit-level SAE**.

Two packages used for applications presented in this session: **emdi** and **mind**

Other examples (not treated in this course):

- ***sae*** provides EBLUP estimators based on unit-level nested error model, as well as EB method for estimation of general non-linear parameters.
- ***rsae*** enables robust prediction of the area-level means under the basic unit-level model.
- ***Hbsae*** provides functions to compute small area estimates using hierarchical Bayesian approaches.
- ***JoSAE*** implements unit-level EBLUP estimators and their MSE estimators, also under heteroscedasticity.

Small Area Estimation using R (2)

Packages used in this practical session:

- **emdi:** the function ***ebp()*** of *this* package produces estimates using the **Empirical Best Prediction Approach** by Molina and Rao (2010). Additionally, mean squared error (MSE) estimation can be conducted by using a parametric bootstrap approach.

- **mind:** its main function is ***mind.unit()*** which fits unit level multivariate linear mixed models. Beyond the possibility of considering multivariate qualitative dependent variables, this package allows to specify a model with more than one random effects.

Package *emdi*

Argument	Description	Default
Fixed	Formula of fixed-effects part of linear mixed model	
pop_data	Population data set, that needs to comprise the explanatory variables and the variable that indicates the domains	
pop_domains	Domain identifier for pop_data	
smp_data	Survey data, that needs to comprise the explanatory variables and the variable that indicates the domains	
smp_domains	Domain identifier for smp_data	
L	Number of Monte-Carlo simulations.	50
threshold	Threshold for poverty indicators	NULL
transformation	Type of transformation	“box.cox”
interval	vector containing a lower and upper limit determining an interval for the estimation of the optimal parameter	“default”
MSE	MSE estimation	FALSE
B	Numbers of bootstrap populations in the parametric bootstrap approach for the MSE estimation	50
seed	Seed for random number generator	123
boot_type	Type of bootstrap: “parametric” or “wild”	“parametric”
parallel_mode	Mode of parallelization	Automatic
cpus	Number of kernels for parallelization	1
custom_indicator	Customized indicators	NULL
na.rm	Deletion of observations with missing No deletion values	FALSE
weights	Sampling weights	NULL

Package *mind*

mind.unit(formula, dom, data, universe, weights = NA, broadarea = NA, max_iter = 200, max_diff = 1e-05, phi_u0 = 0.05, REML = TRUE)

- the fixed part of the model using a *formula* (~) where the variable of interest is placed on the left-hand side and the desired unit level covariates separated by "+" on the right-hand side.
- the domain of interest.
- the survey *data* frame containing the variables in the model.
- the population data frame containing the list of the units belonging to the target population, along with the corresponding values of the auxiliary variables, *universe*.

Survey weights can be included in the fitting process by specifying the parameter *weights*; similarly the parameter *broadarea* can be used to specify if a broadarea is required in the model.

Practical demonstration

Let's switch to R!

APPLICATION BASED ON SDG INDICATOR 5.A.1

SDG Indicator 5.a.1

Indicator 5.a.1 - (a) Percentage of people with ownership or secure rights over agricultural land (out of total agricultural population), by sex; and (b) share of women among owners or rights-bearers of agricultural land, by type of tenure

This indicator is divided in two sub-indicators. Part (a) is an incidence measure. It measures how prevalent ownership or secure rights over agricultural land are in the reference population. Part (b) measures the share of women among owners or rights-bearers of agricultural land. Therefore it can be used to monitor the under-representation of women among the owners or holders of agricultural land. This is a de facto indicator which will measure progress towards SDG Target 5.a.

Target 5.a

Undertake reforms to give women equal rights to economic resources, as well as access to ownership and control over land and other forms of property, financial services, inheritance and natural resources, in accordance with national laws.

SDG Indicator 5.a.1 (2)

Data sources used to compute the two sub-components of the Indicator are:

- **Agricultural surveys** (AGRISurvey)
- **Multi-topic household surveys** (LSMS-ISA)
- **Agriculture and population censuses** with the necessary questions included in their questionnaires. However, as censuses are conducted only every 5/10 years, they are not the preferred vehicle to collect data needed to report the indicator on an annual basis.

Administrative data to proxy the here considered indicators is not recommended as they often do not allow isolating the target population, i.e. adult individuals living in agricultural households.

Data disaggregation of SDG Indicator 5.a.1

Mandatory dimensions:

- 5.a.1.a: disaggregated by gender;
- 5.a.1.b: disaggregated by type of tenure right.

Additional levels for future disaggregation can be considered to better inform the policy-making process, e.g.:

- Disaggregation at sub-national level;
- Disaggregation by urban/rural location of the household;
- Disaggregation by the type of legally recognized document;
- Disaggregation by income level and age class of individuals.

Data disaggregation of SDG Indicator 5.a.1 (2)

As the number of countries reporting on these two sub-indicators is still minimal, little attention has been paid so far at the additional challenges posed by the need of producing **disaggregated** estimates.

The main difficulties:

- traditional sampling designs impose limitations on the production of reliable disaggregated estimates, especially for what concerns small sub-populations or geographical areas.

- in cases where multi-topic household surveys are used instead of agricultural surveys, it is first necessary to perform a screening of agricultural households.

Possible solution is the use of Small Area Estimation techniques.

Data sources

1. Survey Microdata: The Uganda National Panel Survey 2013-2014

The Ugandan National Panel Survey (UNPS) of 2013-2014, is a survey carried out annually by the Ugandan Bureau of Statistics (UBOS) under the assistance of the World Bank (WB) Living Standard Measurement Study (LSMS).

The Survey is based on a multi-stage stratified sampling design, meaning that the selection of the sample is implemented in the two following steps:

- Enumeration areas (EAs) or primary sampling units (PSUs) are randomly selected without replacement from each stratum.
- Individuals are randomly selected without replacement from each PSU.

2. Census microdata: The Uganda Population and Housing Census of 2014

As second data source, used to retrieve auxiliary information to implement SAE techniques, the case study considered the Uganda Population and Housing Census of 2014.

The 10% sample was selected with simple random sampling without replacement from the IPUMS web-catalogue.

Data description

- The territory of Uganda is divided in 4 administrative regions, which are further divided in sub-regions, districts, counties, sub-counties and parishes.
- In 2014, Uganda had 115 districts and one city (Kampala), which have been selected as the level of geographic disaggregation for this study. Hence, combining the geographic disaggregation with the gender of individuals results in a total of 232 disaggregation domains for which estimates need to be produced.
- The sample of the UNPS provided information on 101 of the 115 districts and Kampala, leaving 15 districts out of the sample.

Table: Number of observations by district and sex

	District-Total	District-Male	District-Female
Number of domains with 0 observations (out of sample)	15	15	15
Number of domains with up to 5 observations	10	11	11
Number of domains with 6-15 observations	3	21	17
Number of domains with 16-25 observations	11	17	20
Number of domains with more than 25 obs.	77	52	53
Range of number of observations	0-187	0-90	0-98

Identification of auxiliary variables

An important prerequisite for the implementation of SAE is the presence of **common auxiliary variables** in the two data sources to be integrated.

Process-flow for **feature selection**:

1) The questionnaires of the survey and the census need to be reviewed in order to identify common questions and the answer options envisaged by the two data collection instruments.
2) The variables resulting from these common questions have to be identified in the two datasets, and tabulated to assess the available answer options.
3) Then, this subset of variables needs to be recoded to produce a set of auxiliary variables sharing the same structure in the two datasets.
4) Finally, the relationship between the variable of interest and the identified auxiliary information need to be assessed using the survey data, in order to identify a sub-set of variables with the power of predicting the variable of interest in a larger dataset.

Identification of auxiliary variables (2)

As feature selection method, a **stepwise regression** has been implemented using as response variable the dummy ***land_owner_SDG5a1***.

- Only 11 variables out of 14 were considered as significant by the selection method. The relative importance of variables has been assessed considering the **LMG method**.
- After this first layer of selection, the generalized variance-inflation factor (**GVIF**) of all the tested auxiliary variable was assessed, to ensure the absence of linear dependencies among the independent variables included in the regression. To make GVIFs comparable across dimensions, suggest to use the following measure:

$$GVFI^* = GVIF^{\frac{1}{2}*df}$$

A value of $GVFI^*$ below 2 ensures the absence of multicollinearity.

Variable	lmg (%)	GVIF	Df	GVFI*
Relationship	39.6	9.7	7	1.2
Marital	23.0	7.9	4	1.3
Age class	21.7	3.2	3	1.2
HH_labor	3.8	3.4	2	1.4
Tot_employment	3.5	1.5	1	1.2
Dem_dep_ratio	3.3	1.8	1	1.3
HH_size	1.7	2.6	2	1.3
Region	1.4	1.1	3	1.0
Ownhome	1.3	1.1	1	1.0
Sex	0.3	1.7	1	1.3
Ec_dep_ratio	0.3	1.3	1	1.1

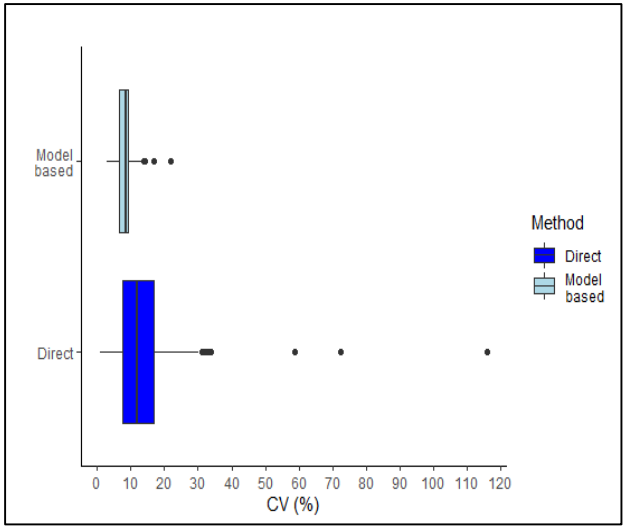
Model validation

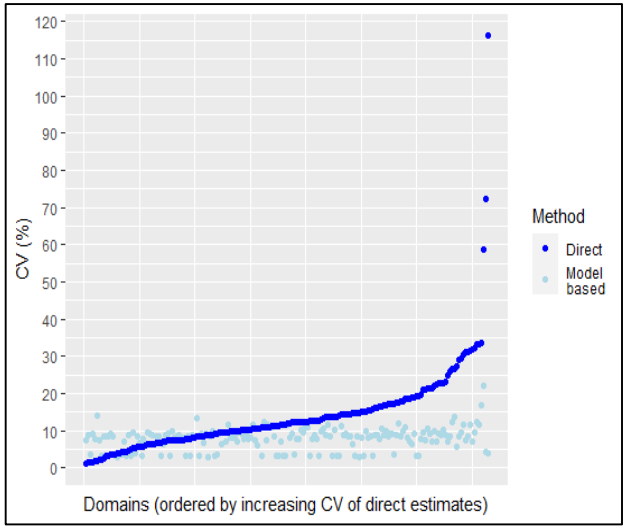
- After producing direct and indirect small-area estimates, their properties need to be assessed. In particular, for estimation domains with sampled observations, direct and indirect estimates can be pared along with their accuracy measures.

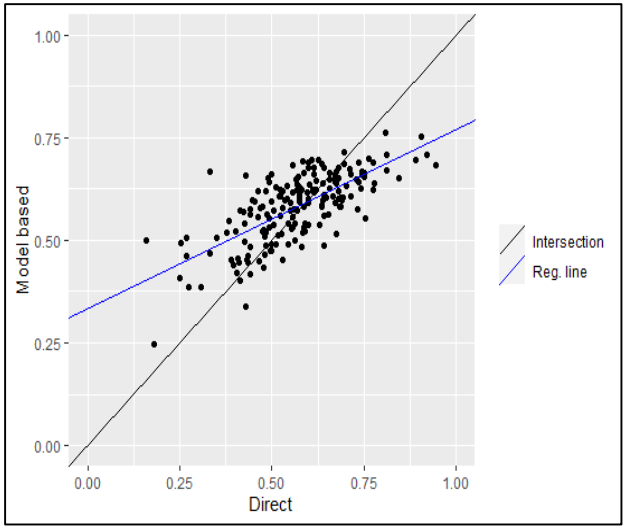
- Generally speaking, especially for domains with many observations, direct and indirect estimates are expected to be correlated, i.e. the two approaches should produce similar results. In terms of quality, the MSE of model-based small area estimates is expected to be considerably lower than the MSE of direct estimates.

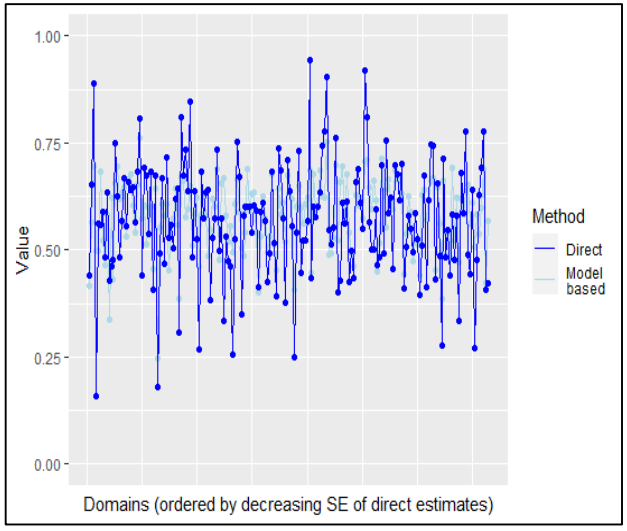
Model validation (2)

5.a.1.a



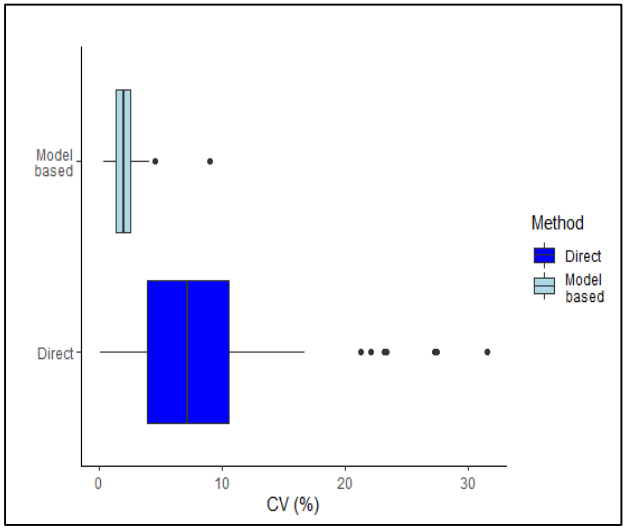


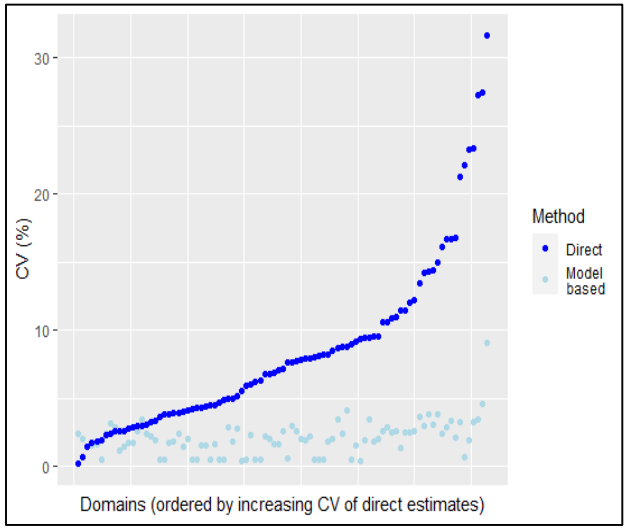


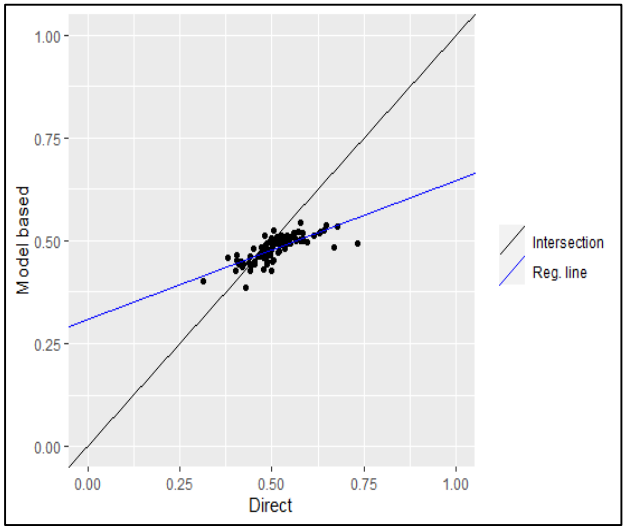


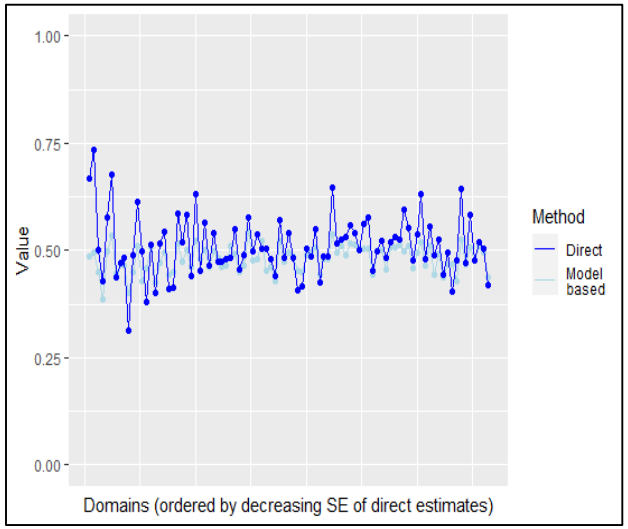
Model validation (3)

5.a.1.b



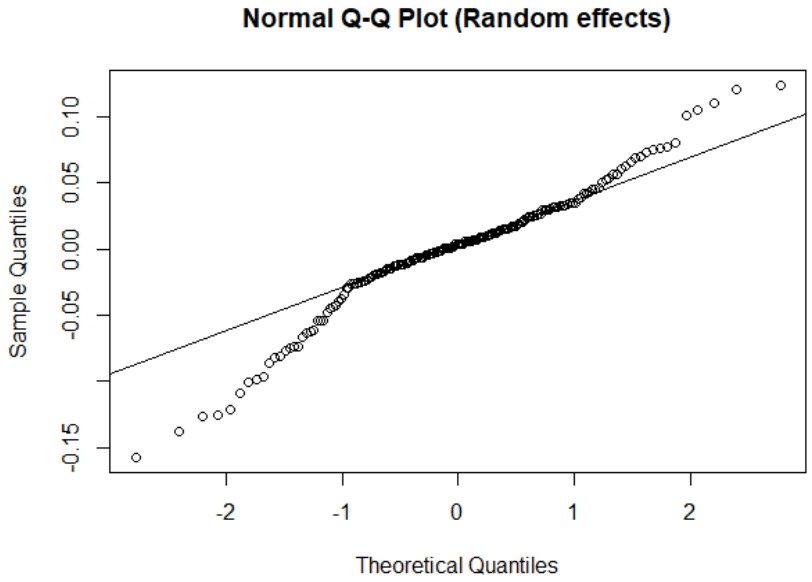


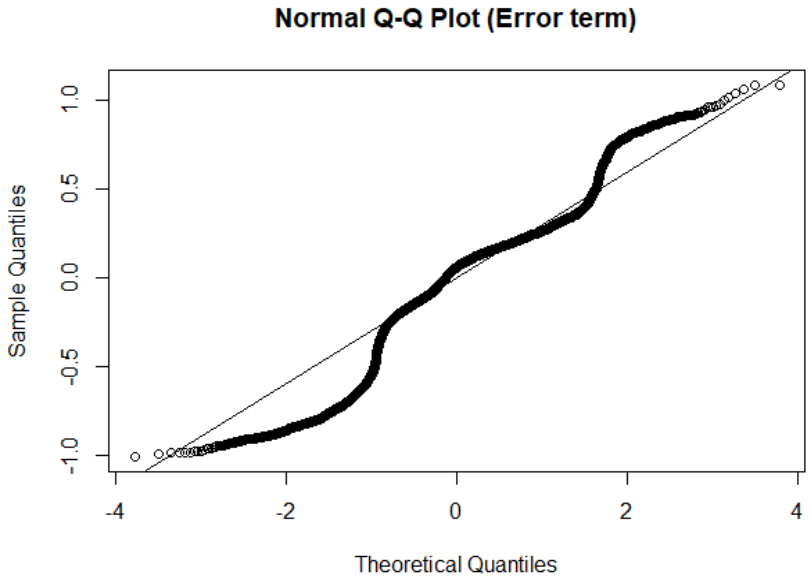




Model validation (4)

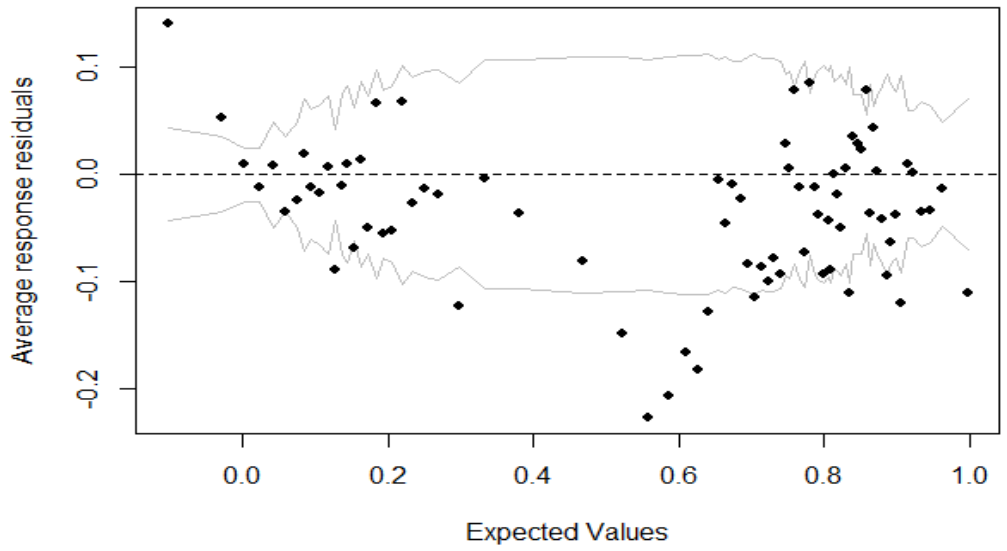
Mixed-effects SAE models are based on complex fitting procedures and rely on several assumptions, especially about the distribution of residuals and random effects.



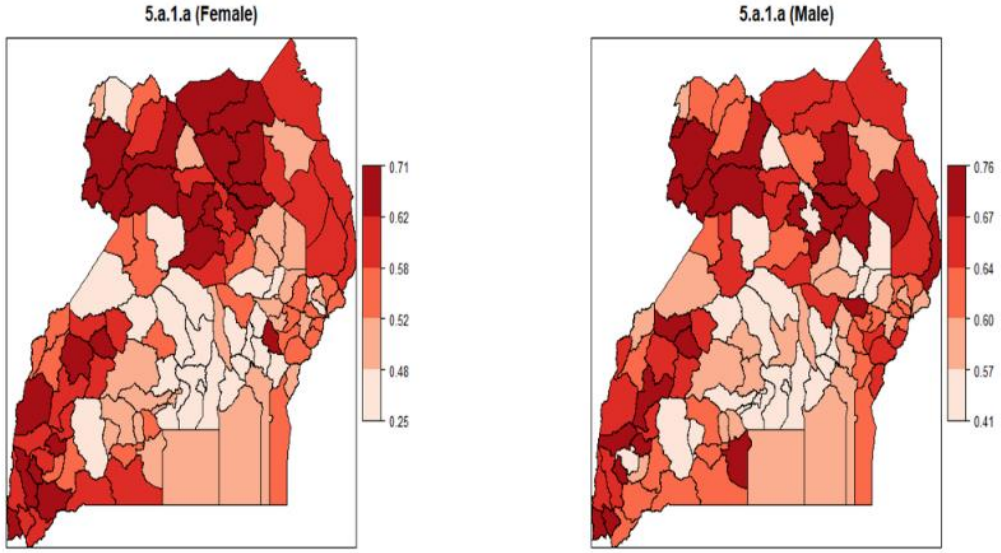


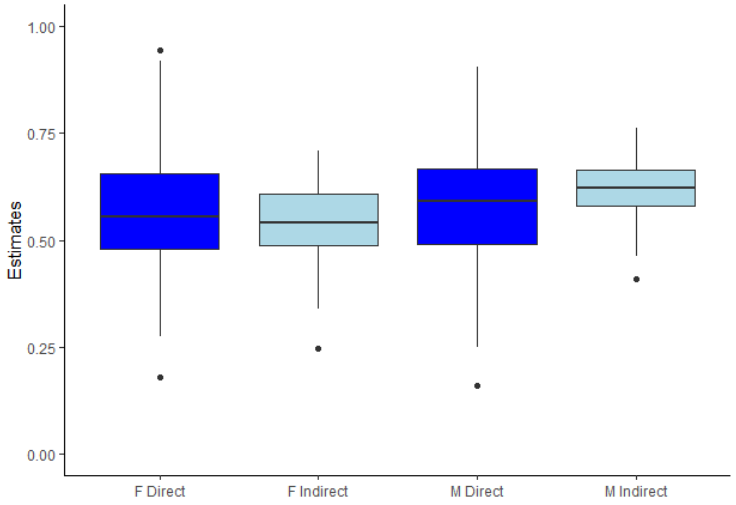
Model validation (5)

Finally, the binned residual plot can be used to assess the overall fit of regression models for binary outcomes.

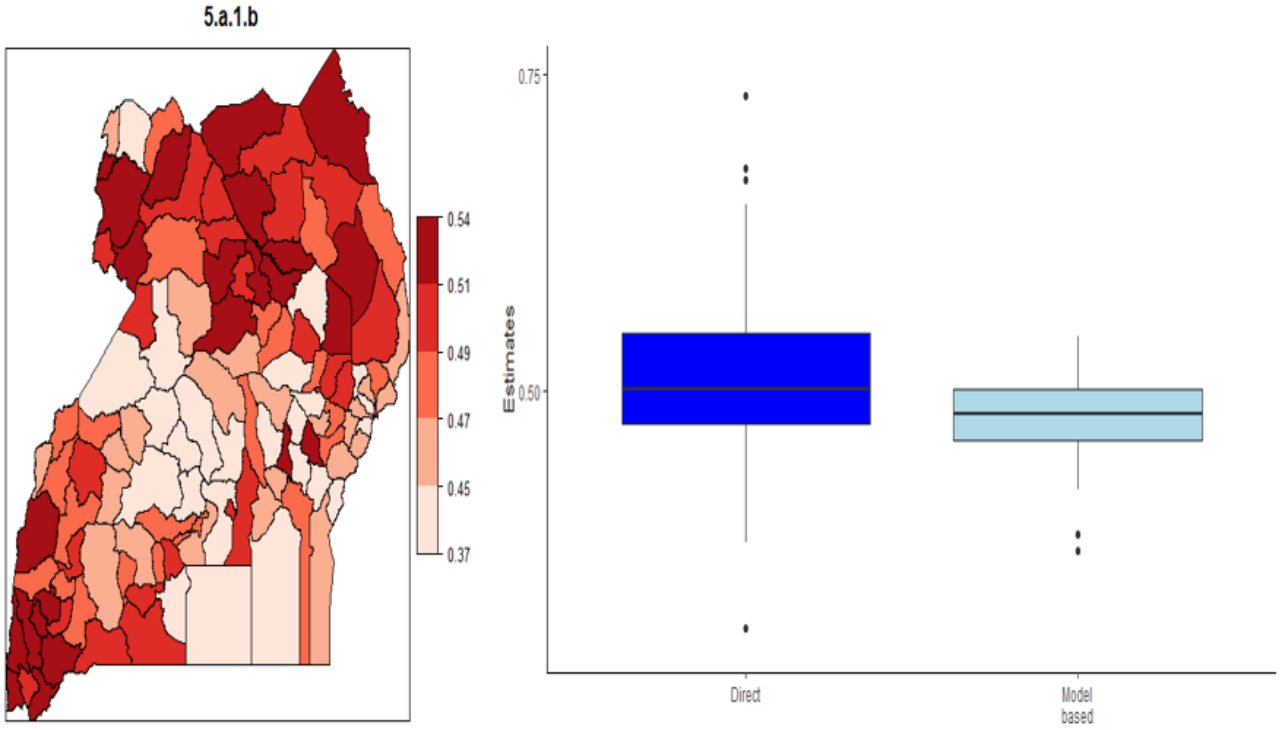


Results – 5.a.1.a





Results – 5.a.1.b



Practical demonstration

Let's switch to R!

Thank you!