Food and Agriculture Organization of the United Nations

SUSTAINABLE DEVELOPMENT GOALS

Area level SAE models with R

Virtual Training on Data Disaggregation and Small Area Estimation for the SDGs
18,19,25,26 April 2023

Area-level SAE with R

Various R packages can be used to implement area-level SAE models, among others (not treated in this course):

- ***smallarea*** offers different variance estimation methods for the standard FH model and a function to estimate unknown sampling variances.
- ***JoSAE*** implements area-level EBLUP estimators and their MSE estimators, also under heteroscedasticity.
- ***saeRobust*** enables robust estimation of area-level models with spatial and/or temporal structures in the random effects.
- ***msae*** implements univariate and multivariate FH models.

Area-level SAE with R (2)

- ***hbsae*** provides functions to compute small area estimates using hierarchical Bayesian approaches.

- ***BayesSAE*** also provides functions to analyze area-level models in Bayesian context.

- ***mme*** fits Gaussian area-level multinominal mixed-effects models.

- ***saeME*** contains functions for an area-level model when the auxiliary variables are measured with error.

Most commonly used packages are **sae** and **emdi** (used in this practical session).

The R package *emdi*

The R package ***emdi*** provides tools for model building, diagnostics, presenting and exporting the results.

Beside including **the Fay-Herriot model**, the package contains various extensions:

- Transformed area-level models with back-transformations;
- Spatial and robust extensions of area-level FH;
- Adjusted variance estimation methods;
- Area-level models that account for measurement errors.

The function that implements the Fay-Herriot model and its extensions is ***fh()***, providing also the corresponding mean squared error estimators for assessing the uncertainty.

The R package *emdi* (2)

Argument	Description	Default
Fixed	Formula of fixed-effects part of linear mixed model	
Vardir	Domain-specific sampling variances of the direct estimates	
combined_data	Combined sample and census data set	
domains	Domain identifier for combined_data	NULL
method	Model fitting method	reml
interval	Lower and upper limit for the variance estimation	NULL
k	Tuning constant for robust estimation	1.345
mult_constant	Bias correction multiplyer constant for robust estimation	1
transformation	Type of transformation	No
backtransformation	Type of back-transformation	NULL
eff_smpsize	Type of back-transformation	NULL
correlation	Correlation of random effects	no
corMatrix	Proximity matrix for the spatial model	NULL
Ci	Array of the variance-covariance matrix of the explanatory variables for each area for the ME model	NULL
tol	Tolerance value for the variance estimation	0.0001
maxit	Maximum number of iteration for the variance estimation	100
MSE	MSE estimation	FALSE
mse_type	Type of MSE estimator	analytical
B	Numbers of bootstrap iteration for computation of a bootstrap MSE and information criteria	C(50,0)
seed	Seed for random number generator	123

Practical demonstration

Let's switch to R!

APPLICATION BASED ON SDG INDICATORS 2.3.1 AND 2.3.2

SDG Indicator 2.3.1

Indicator 2.3.1 - Volume of production per labour unit by classes of farming / pastoral / forestry enterprise size

This indicator refers to the value of production per labour unit operated by small scale producers in the farming, pastoral and forestry sectors. Data will be produced by classes of enterprise size. The indicator will measure progress towards SDG Target 2.3.

Target 2.3

By 2030, double the agricultural productivity and incomes of small-scale food producers, in particular women, indigenous peoples, family farmers, pastoralists and fishers, including through secure and equal access to land, other productive resources and inputs, knowledge, financial services, markets and opportunities for value addition and nonfarm employment.

SDG Indicator 2.3.2

Indicator 2.3.2 Average income of small-scale food producers, by sex and indigenous status

The indicator refers to the average income of small-scale food producers employed in the farming, pastoral and forestry sectors. Data will be disaggregated sex and indigenous status. This indicator will measure progress towards SDG Target 2.3.

Target 2.3

By 2030, double the agricultural productivity and incomes of small-scale food producers, in particular women, indigenous peoples, family farmers, pastoralists and fishers, including through secure and equal access to land, other productive resources and inputs, knowledge, financial services, markets and opportunities for value addition and nonfarm employment.

Data disaggregation

Data sources used to compute the two indicators are:

- **Agricultural surveys**
- **Household surveys**

Main issue for the production of disaggregated estimates:

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

Application of a FH area-level model to produce disaggregated estimates of SDG Indicators 2.3.1 and 2.3.2 at granular sub-national level.

Data description

The application is based on survey microdata from Mali "***Enquête Agricole de Conjoncture Intégrée aux Conditions de Vie des Ménages***", 2017.

Sampling procedure: the sample is nationally representative and covers all regions and areas (urban and rural) with the exception of Kidal and Menaka.

Region	# of circles	Sample size by region - 2.3.1	Sample size of small-scale food producers by region – 2.3.1	Average num. of sampled small-scale food producers by circle - 2.3.1	Sample size by region - 2.3.2	Sample size of small-scale food producers by region – 2.3.2	Average num. of sampled small-scale food producers by circle - 2.3.2
Kayes	7	431	384	55	572	413	59
Koulikoro	7	381	282	40	517	309	44
Sikasso	7	368	206	29	526	229	33
Segou	7	436	295	42	595	331	47
Mopti	8	323	221	28	429	265	33
Tombouctou	5	137	126	25	219	161	32
Gao	3	110	101	34	180	141	46
Kidal	4 (out of sample)	-	-	-	-	-	-
Menaka	4 (out of sample)	-	-	-	-	-	-
* Bamako	1 (Bamako)	22	22	22	760	71	71

Auxiliary variables

Multiple geospatial auxiliary data that could have a possible correlation with the indicators were assessed. These variables were providing info on the following topics:

- Soil characteristics;
- Weather and climate data;
- Land cover and elevation;
- Total population density;

Since agricultural production in Mali is largely dominated by cotton and grains (including rice, millet, sorghum, and wheat), data **on the harvested area and production of these crops were included among the auxiliary variables.**

Sources of auxiliary variables

Variable name	Spatial resolution	Temporal resolution	Source
Vol. fraction of coarse fragments	~1x1 km	Static	ISRIC: World Soil Information
Nitrogen	~1x1 km	Static	
Sand	~1x1 km	Static	
Silt	~1x1 km	Static	
Clay	~1x1 km	Static	
Soil organic carbon	~1x1 km	Static	
Minimum temperature	~4.5x4.5 km	Monthly	WorldCilm: Historical monthly weather data
Maximum temperature	~4.5x4.5 km	Monthly	
Precipitation	~4.5x4.5 km	Monthly	

Sources of auxiliary variables (2)

Variable name	Spatial resolution	Temporal resolution	Source
Direct normal irradiation (Long-term yearly average)	~0.3x0.3 km	1994-2018	Solargis
Diffuse horizontal irradiation (Long-term yearly average)	~0.3x0.3 km	1994-2018	
Air temperature (Long-term yearly average)	~1x1 km	1999-2020	
Vegetation indexes	~5.5x5.5 km	Monthly	NASA EarthData
Elevation	~1x1 km	Static	CGIAR CSI
Bare ground	~1x1 km	Annual	Zenodo
Built-up	~1x1 km	Annual	
Harvested area (major crops)	~1x1 km	Annual	MAPSPAM
Production (major crops)	~1x1 km	Annual	

Geospatial covariates processing

Spatially continuous phenomena, as remote sensing data, are typically represented with the raster data model according to which the world is divided into a grid of equally sized rectangles referred to as cells or pixels. Often the extent of input raster data is larger than the area of interest as in this case where we need the geospatial covariates to be at Mali's circle level.

To do so, the steps performed to calculate and extract the underlying values from a raster file from the irregularly shaped polygons, represented by the circles in Mali, were the following:

- Mask and crop: which it simply consists in taking from a raster only the portion that falls inside the study region, in this case Mali.
- Extract and aggregate: is the process of identifying and aggregating as means the values associated with a raster at specific selected areas.

The R software provides several packages with functions that allow reading and processing raster datasets, applying the aforementioned operations.

Practical demonstration

Let's switch to R!

Parameters to estimate

$$\theta_d = x_d^T \beta + u_d + e_d, d = 1, \dots, D$$

The unknown parameters of the area level model to be estimated are:

- the fixed-effects parameters β;
- the variance of random effects σ_u^2.

The EBLUP estimator can then be expressed as:

$$\hat{\theta}_d^{EBLUP} = \hat{\gamma}_d \hat{\theta}_d + (1 - \hat{\gamma}_d) x_d^T \hat{\beta}$$

where $\hat{\beta}$ is the weighted least squares estimators of the regression parameters and $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{e,d}^2}$ is the so-called shrinkage factor for domain d.

Variance smoothing

In the basic area-level model, the sampling error variance $\hat{\sigma}^2_{e,d}$ is assumed to be known but in most practical applications the variance are estimated and may not be stable due to the small sample sizes in the domains of interest.

Therefore, they are smoothed using generalized variance functions (GVF).

First of all, given that the parameter of interest is a ratio, a linearization approach of the variance is adopted. Indicating with $\hat{R} = \frac{\hat{X}}{\hat{Y}}$ the ratio estimator, where $\hat{X}$ and $\hat{Y}$ denote estimators of X and Y, the Taylor series estimator of the variance of $\hat{R}$ is:

$$v(\hat{R}) = \frac{1}{\hat{Y}^2}\left[v(\hat{X}) + \hat{R}^2 v(\hat{Y}) - 2\hat{R}\hat{\rho}_{XY}\sqrt{v(\hat{X})}\sqrt{v(\hat{Y})}\right]$$

where $v(\hat{X})$, $v(\hat{Y})$ and $\hat{\rho}_{XY}$ denote estimators of the variance of $\hat{X}$, the variance of $\hat{Y}$ and the linear correlation between $\hat{X}$ and $\hat{Y}$ respectively.

Variance smoothing (2)

The smoothed variance of $\hat{R}$ can be obtained by substituting to $v(\hat{X})$, $v(\hat{Y})$ their smoothed versions $\tilde{v}(\hat{X})$, $\tilde{v}(\hat{Y})$:

$$\tilde{\sigma}_{e_d}^2 = \tilde{v}(\hat{R}) = \frac{1}{\hat{Y}^2}\left[\tilde{v}(\hat{X}) + \hat{R}^2\tilde{v}(\hat{Y}) - 2\hat{R}\hat{\rho}_{XY}\sqrt{\tilde{v}(\hat{X})}\sqrt{\tilde{v}(\hat{Y})}\right]$$

The smoothing model chosen for $v(\hat{X})$ and $v(\hat{Y})$ was the following:

$$\log\left(v(\hat{X})\right) = \alpha + \beta\log(\hat{X}) \ , \log\left(v(\hat{Y})\right) = \alpha + \beta\log(\hat{Y})$$

The smoothed variances obtained $\tilde{\sigma}_{e_d}^2$ were bias-corrected by applying a ratio adjustment of a factor equal to $\Delta = \frac{\sum \hat{\sigma}_{e_d}^2}{\sum \tilde{\sigma}_{e_d}^2}$ avoiding in this way overestimation or underestimation of the direct variances $\hat{\sigma}_{e_d}^2$.

Practical demonstration

Let's switch to R!

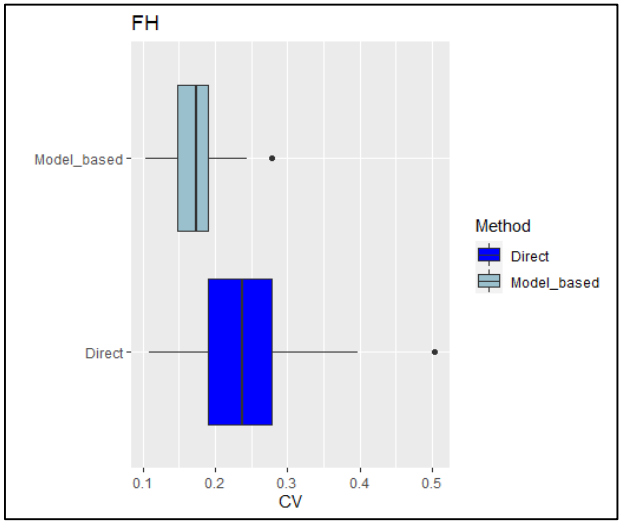
Model building – 2.3.1

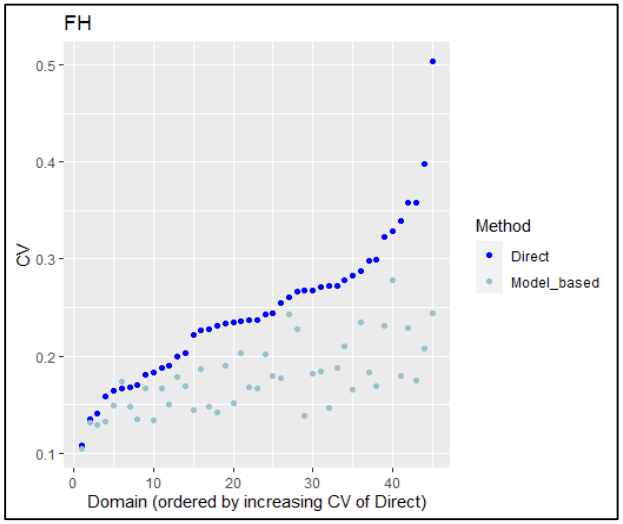
Variable of interest: **vp_lab=** value of crop production per labour day.

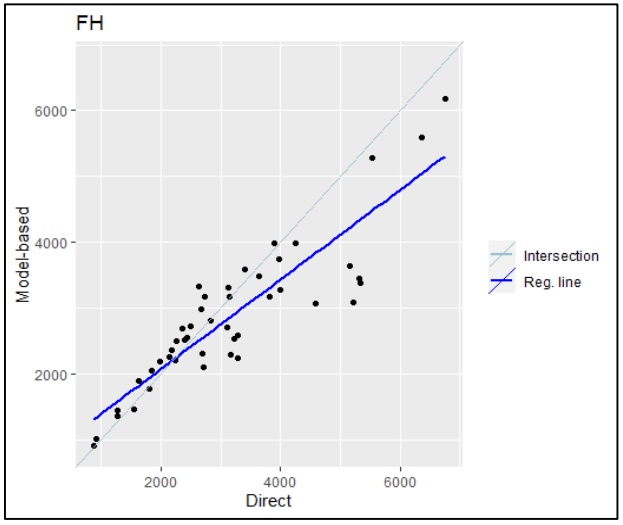
List of auxiliary variables selected by the stepwise regression:

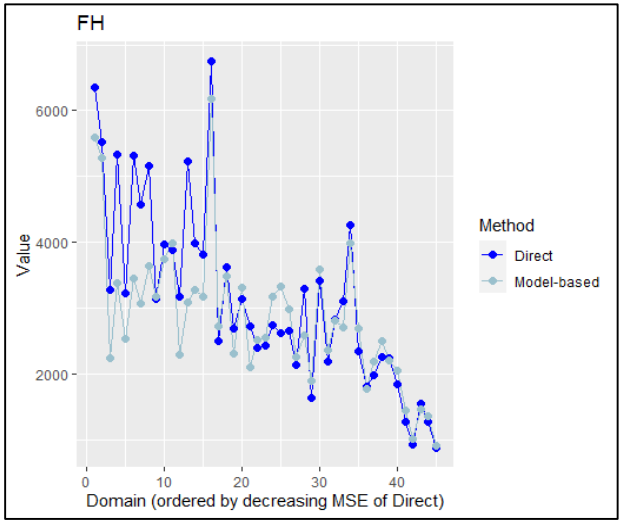
Variable name	Unit of measure	Img (%)
Production of cotton	(Metric ton)	24.0
Direct normal irradiation	(kWh/m2)	16.1
Production of wheat	(Metric ton)	14.6
Production of rice	(Metric ton)	11.7
Production of sorghum	(Metric ton)	11.4
Vol. fraction of coarse fragments	(%)	9.8
Soil organic carbon	(g kg-1)	8.7
Harvested area of rice	(hectare)	3.7

Model validation – 2.3.1

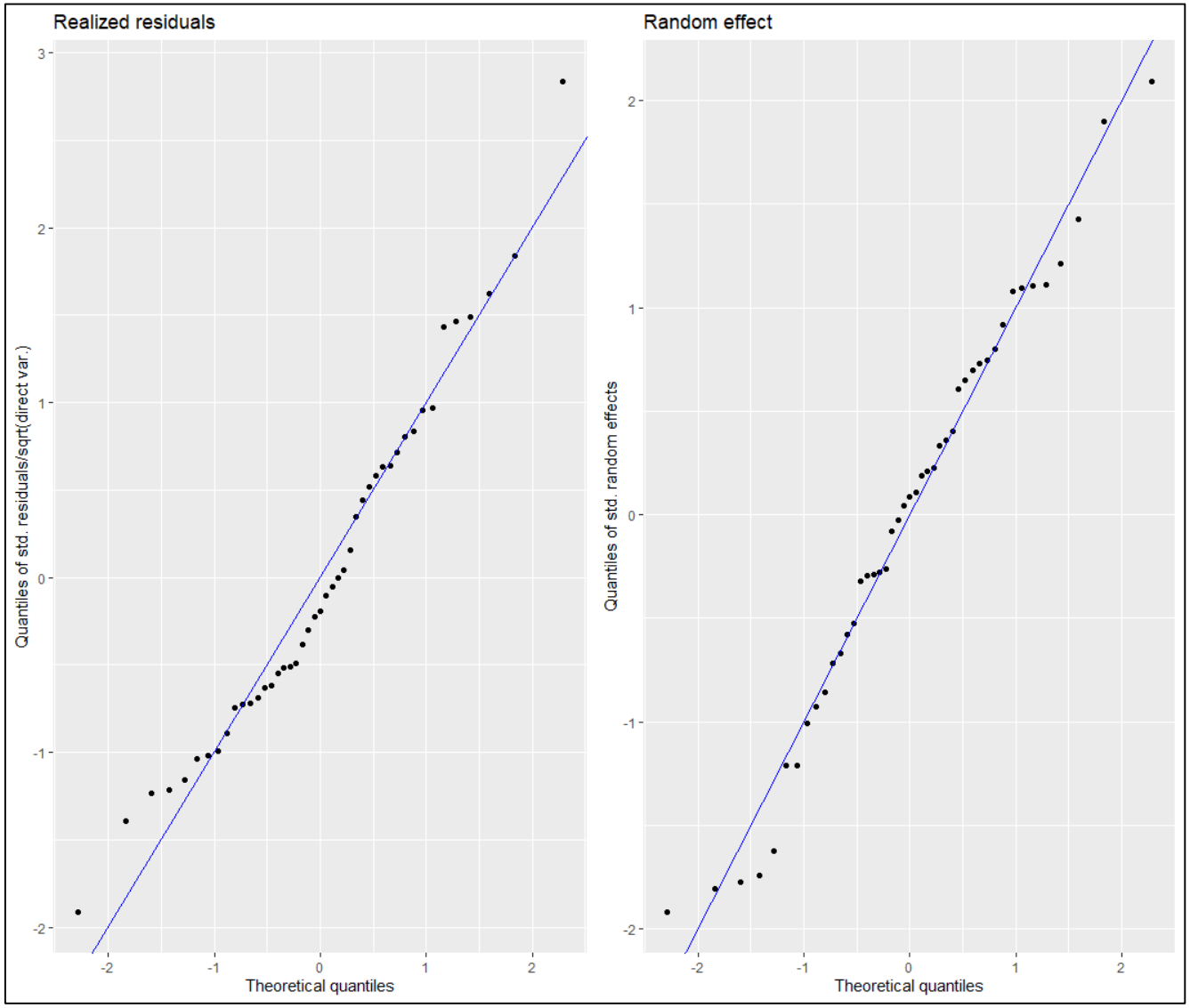








Model validation – 2.3.1 (2)



Practical demonstration

Let's switch to R!

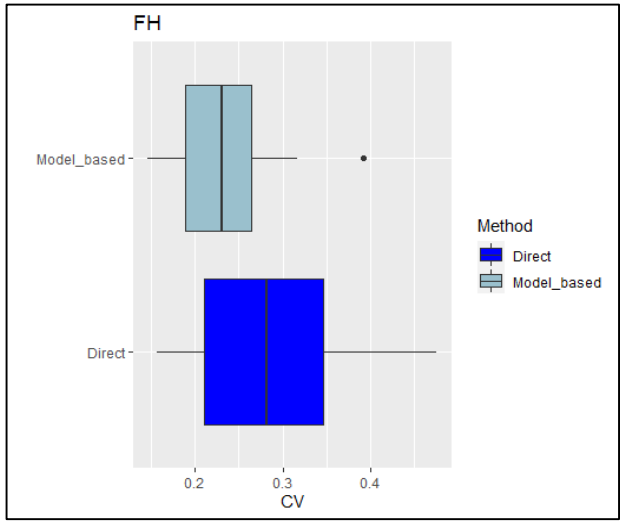
Model building – 2.3.2

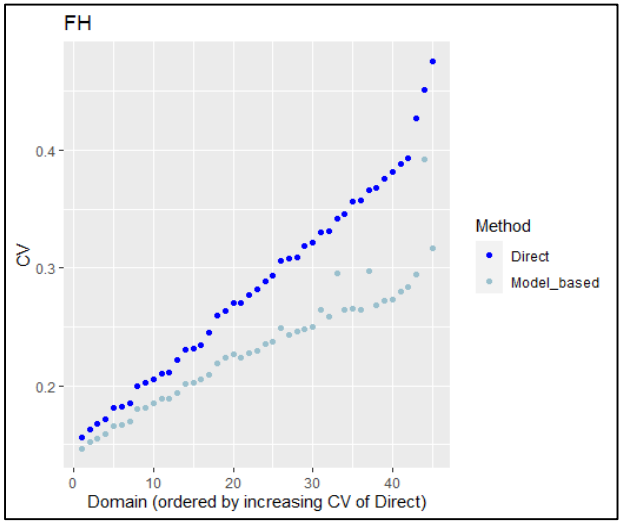
Variable of interest: **agrincome=** represents total household income derived from agricultural activities i.e crop, livestock, fishery and forestry production.

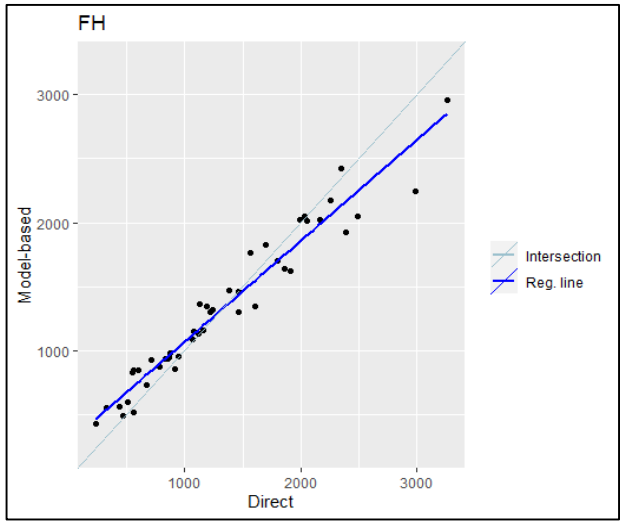
List of auxiliary variables selected by the stepwise regression:

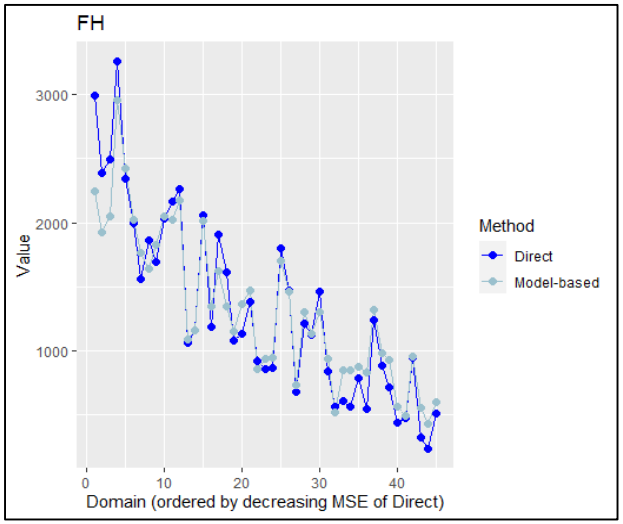
Variable name	Unit of measure	Img (%)
Minimum temperature	(Celsius °C)	69.2
Production of cotton	(Metric ton)	16.5
Total population	(number)	14.3

Model validation – 2.3.2

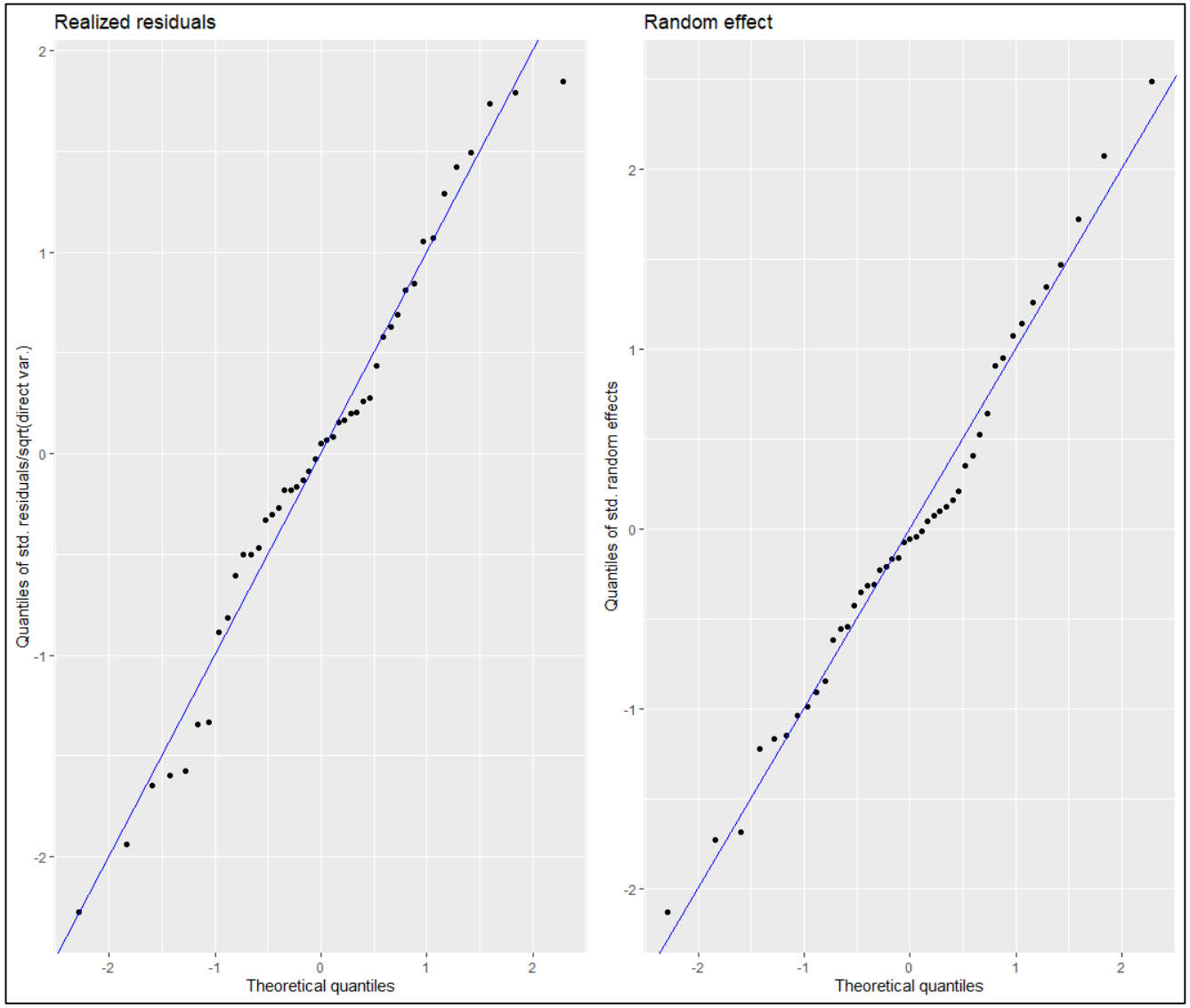








Model validation – 2.3.2 (2)



Practical demonstration

Let's switch to R!

Thank you!