LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Identification of Health Risk Factors in Developing Countries using Intrinsic Model Selection Approaches

Daniel Seussler
Master‘s Thesis

supervised by Dr. Cornelius Fritz
Department of Statistics

Introduction

Introduction

Objective:

- Identification of risk factors of two common disease burdens in sub-Saharan Africa.
- Statistically, how to approach moderate dimensional regression where effect selection is necessary, and how to approach predictive modelling with binomial data where inference on the underlying risk is desired.

Methodology:

- Component-wise (model-based) boosting with intrinsic variable selection and model choice.

Data Source: Demographic and Health Surveys (DHS)

- Large N household surveys in low- and middle income countries collect data on important population health and socio-economic characteristics.
- Often used in the epidemiological literature due to it’s temporal compatibility and representative samples, e.g., in the research of determinants of disease burdens.

- I use the Madagascar 2021 Standard DHS and the Mali 2021 Malaria Indicator Survey (MIS).

Component-wise boosting for variable selection and model choice

Component-wise boosting

Functional gradient boosting: estimate function by numerical optimisation in function space (Friedmann 2001).

Very popular in the literature given it’s superiority in many tasks (XGBoost). Generally boosted trees but a model based approach is also possible (Bühlmann and Yu 2003, Bühlmann and Hothorn 2007).

Boosting additive models (component-wise). In a generalised additive regression:

$$E(y|\mathbf{x}) = h(\eta(\mathbf{x}))$$

$$\eta(\mathbf{x}) = \beta_0 + \sum_{j=1}^{J} f_j(\mathbf{x}).$$

We define the following optimisation problem

$$\arg\min_{\eta} = E(\rho(y, \eta))$$

where ρ is the negative log-likelihood (i.e., loss) and replace the expectation with the empirical risk

$$R = n^{-1} \sum_{i=1}^{n} \rho(y_i, \eta).$$

Component-wise boosting

Init $\hat{\eta}^{[0]}$ and define a set of base-learners $b_l, l = 1, \dots, L$. In each iteration m refit each to the negative gradient

$$u_i = -\frac{\partial}{\partial \eta} \rho(y_i, \eta) \Big|_{\eta = \eta^{[m-1]}(\mathbf{x}_i)}$$

and select the best-fitting learner b_{l*} by residual-sum-of-squares.

Update $\hat{\eta}^{[m]}$

$$\hat{\eta}^{[m]} = \hat{\eta}^{[m-1]} + \nu \hat{b}_{l*}$$

where the parameter ν is $0 < \nu \ll 1$.

Early stopping (to avoid overfitting and improve generalisation error) can be achieved through holdout sets and resampling methods (k-fold cv, bootstrap etc.).

- Intrinsic selection of learners as in each iteration exactly *one* learner is added to the additive predictor.
- Effect selection by decomposition of smooth effects (Kneib et al. 2009). Allows to choose none, linear, non-linear effects.
- Many statistical tasks covered: survival analysis, quantile regression (Fenske et al. 2011), multivariate distributions (Strömer et al. 2021) and copulas (Hans et al. 2022).

Component-wise boosting

Distributional regression (or GAMLSS: Generalised Additive Models for Location, Scale and Shape)

➔ Allows modelling of each distribution parameter with a seperate additive predictor.

➔ Algorithm: non-cyclical boosting. In each iteration both the learner *and* the parameter are selected jointly for an update. (see Thomas et al. 2018)

Uncertainty quantification

➔ Subsampling replications or bootstrap replications of the model, but: no theoretical guarantees of coverage.

➔ Use it to assess stability of estimated coefficients.

➔ Stability selection (Meinshausen and Bühlmann 2010) allows for finite sample control of Type 1 error for variable selection, but: very conservative.

Case Studies

Case Study 1: Childhood malnutrition in Madagascar

Objective: Identification of risk factors (or determinants).

Outcome: child is stunted yes / no
(height-for-age score is more than 2σ below the reference)

Explanatory Covariates:

- Individual-level risk factors (age, gender, mother’s bmi, if twin, etc.) ,
- household- (no. of household members, DHS wealth index, etc.),
- and community-level (urbanicity, distance to healthcare facilities etc.).

Additional Specifications:

- Decomposition of smooth effects for continuous covariates to allow for effect selection.
- Markov random field for administrative regions (first-order neighbourhood structure).

Case Study 2: Geographic malaria risk in Mali

Objective: Identification of environmental correlates of malaria prevalence and inference of risk.

Outcome: *k* children out of *n* tested positive in cluster *c*, i.e., binomial.

- Following Dong and Wakefield 2021, I model the outcome with a beta-binomial distribution that accounts for cluster overdispersion due to within-cluster variability.

Explanatory Covariates:

- Elevation, land surface temperature (day/night), rainfall - aggregated annually (mean / sum).
- Vegetation indices, NDVI and EVI.
- Population (Global Human Settlement Layer Data).
- Urbanicity variable to account for survey design.

Additional Specifications:

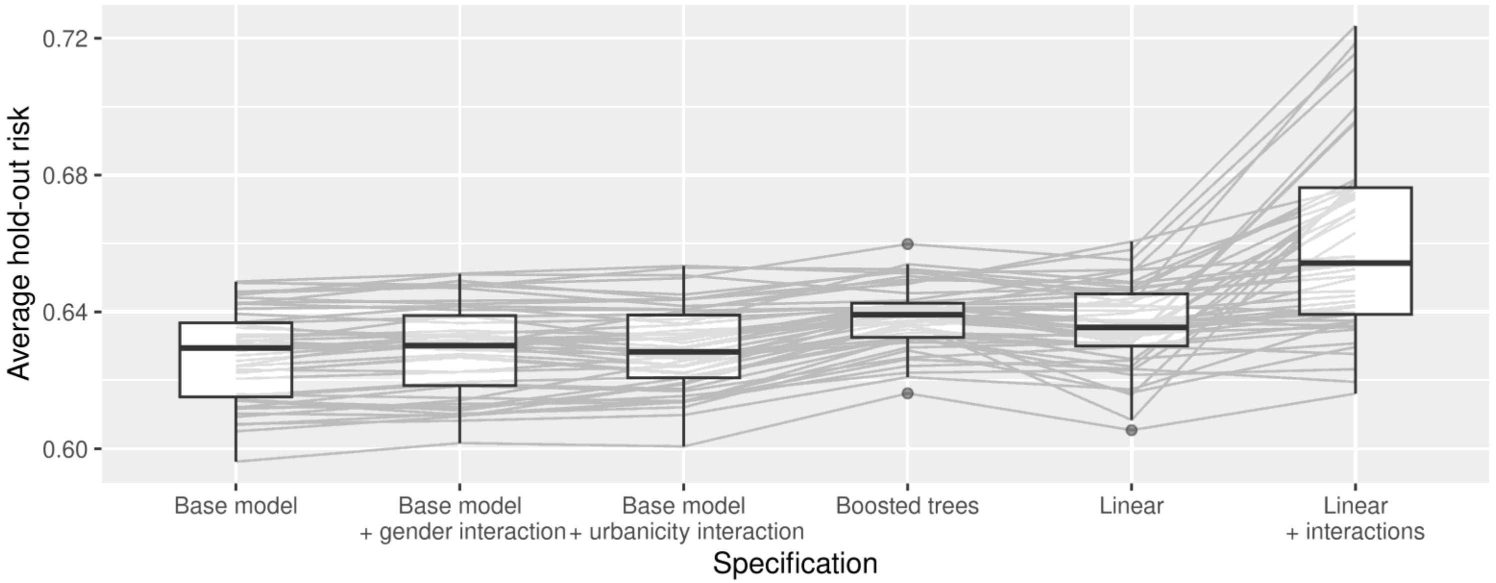
- Decomposition of smooth effects for continuous covariates to allow for effect selection.
- Bivariate tensor P-spline to model spatial effects.

Results

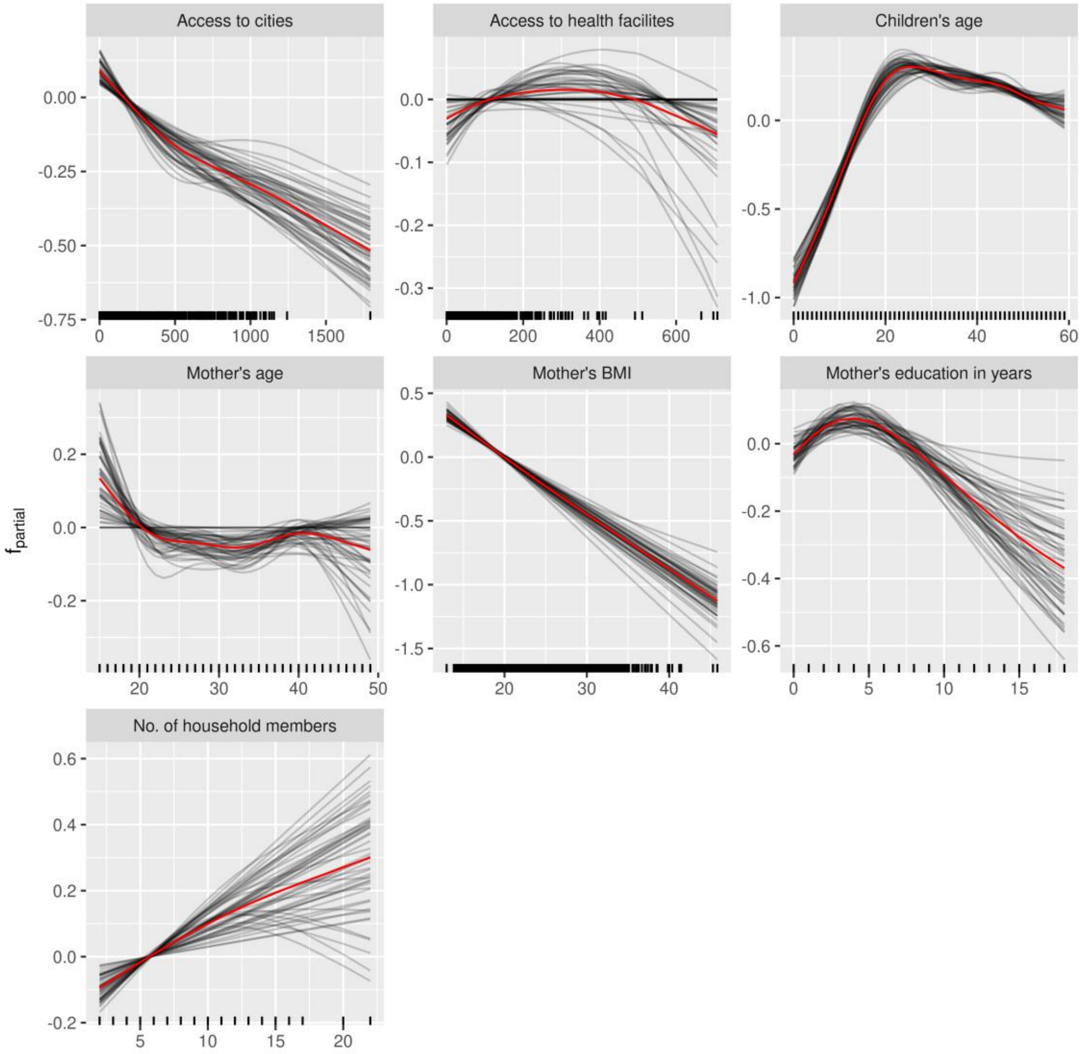
Childhood malnutrition in Madagascar

Model selection

- by three-way holdout method (70/20/10), no. of boosting iterations selected optimal performance on validation set.
- Following results based on 50 subsampling replications. Red line indicates pointwise average.

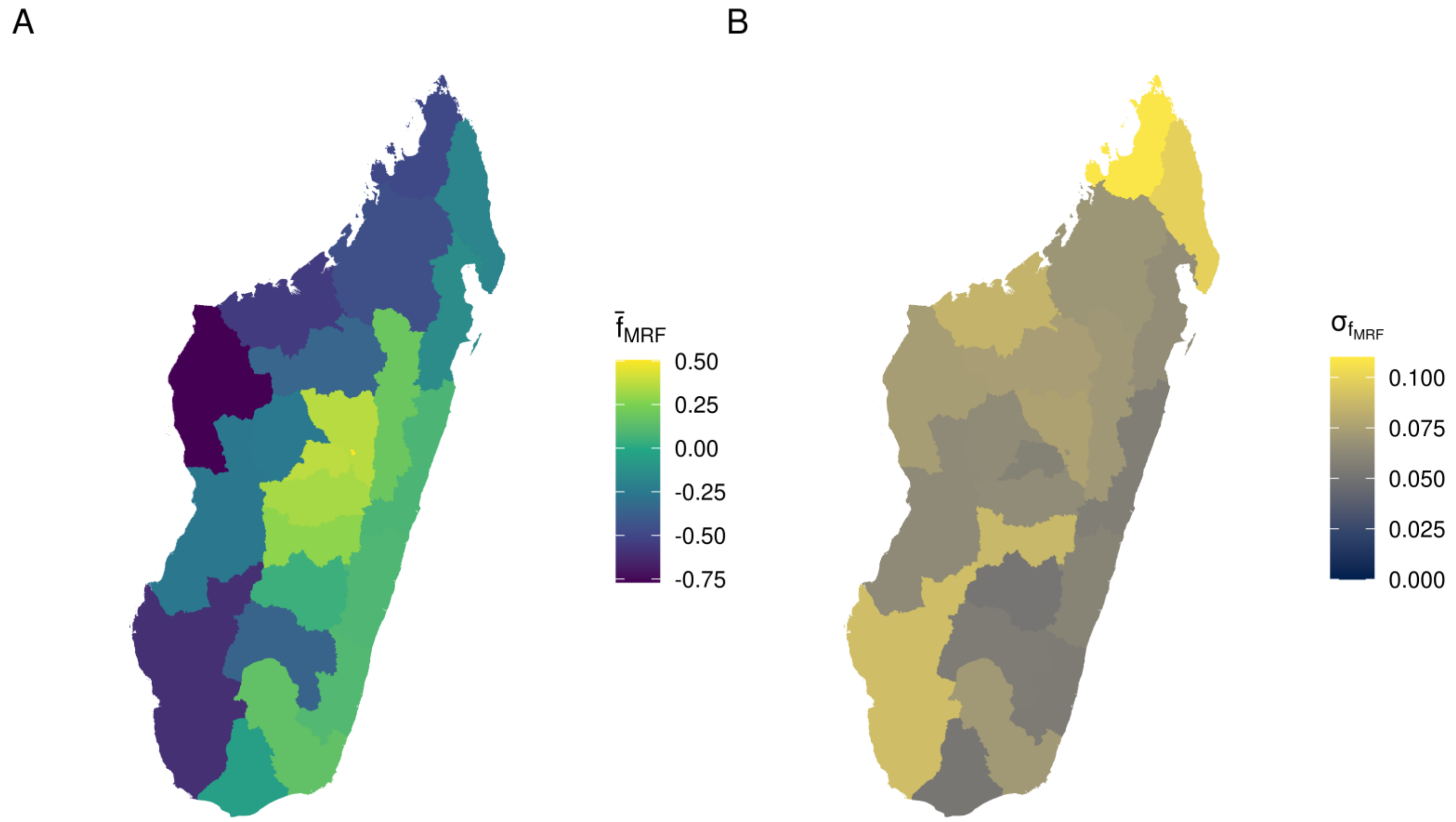


Childhood malnutrition in Madagascar



- Estimated effects of the continuous covariates.

Childhood malnutrition in Madagascar



- Markov random field effect for admin 1 regions, mean (A) and standard deviation (B) of the subsampling replications.

Geographic malaria risk in Mali

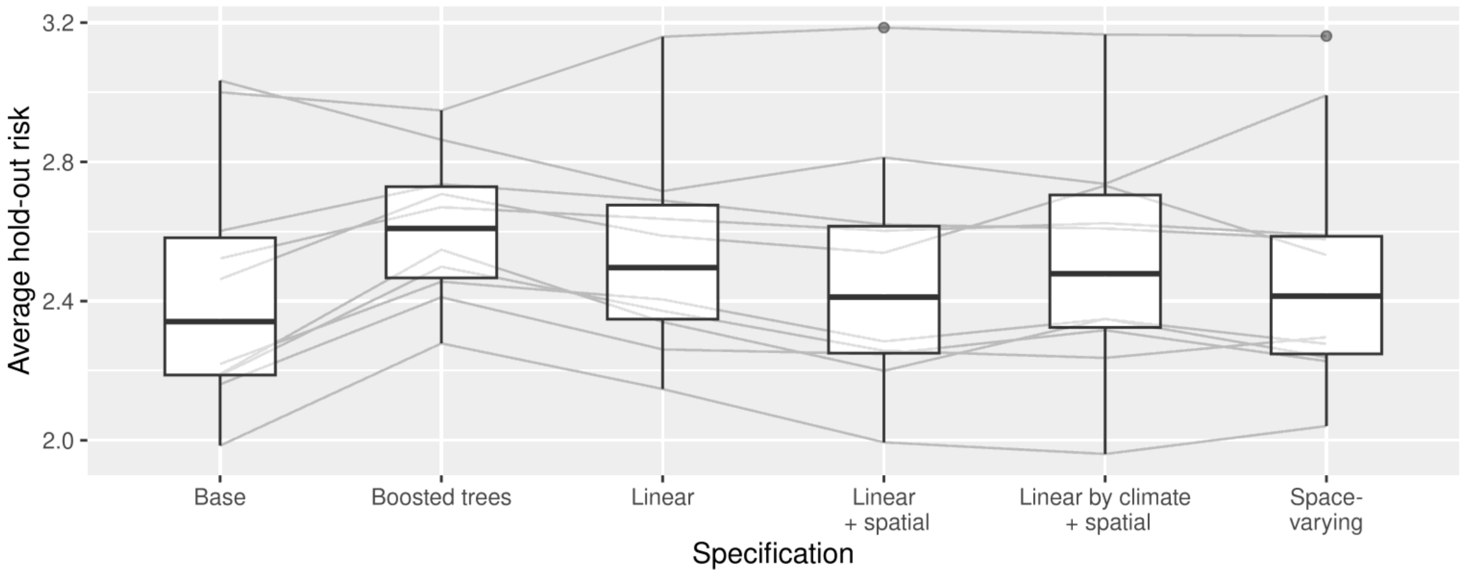
Model validation and selection

- by nested cross-validation (10-fold, stratified by region)
- comparison of base specification to boosted trees and
  - Linear
  - Linear + bivariate tensor P-spline
  - Linear + climate interactions + bivariate tensor P-spline
  - Linear + spatially-varying coefficient

Early stopping by 10-fold cross-validation. Uncertainty quantification with bootstrap samples.

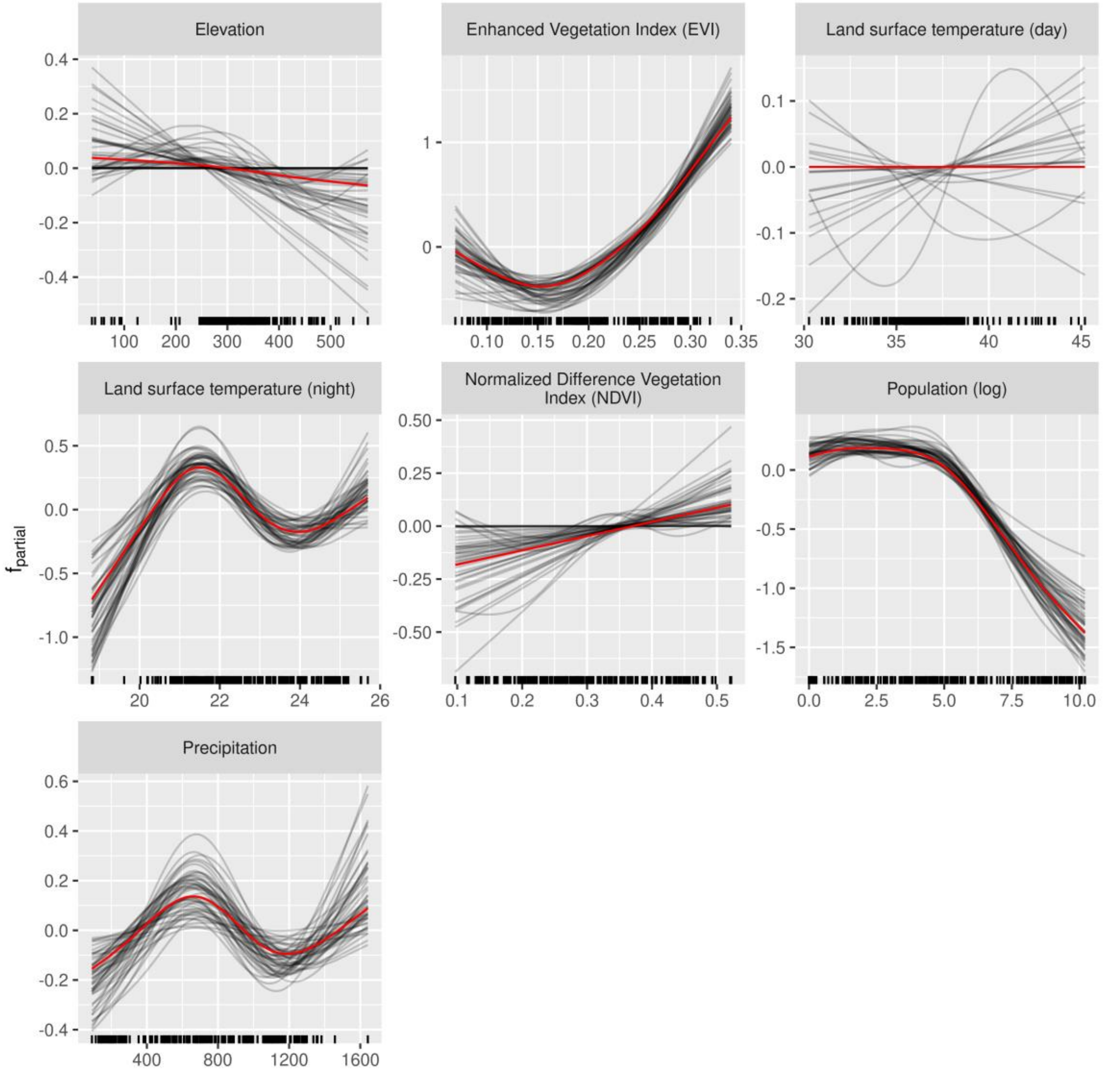
Geographic malaria risk in Mali

Model	Bias	MAE	RMSE	80% PI	90% PI	95% PI
Beta binomial	-0.005	0.093	0.131	0.861	0.944	0.963
Binomial	-0.007	0.091	0.131	0.597	0.708	0.759

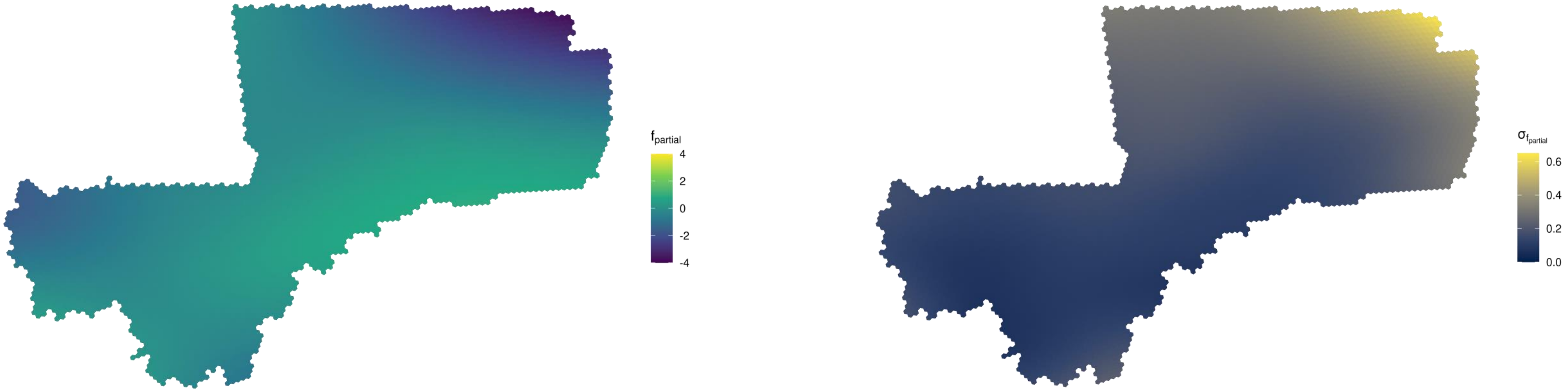


Geographic malaria risk in Mali

- Red line / dot indicates estimate of the principal model.
- Uncertainty quantification based on 50 bootstrap samples.

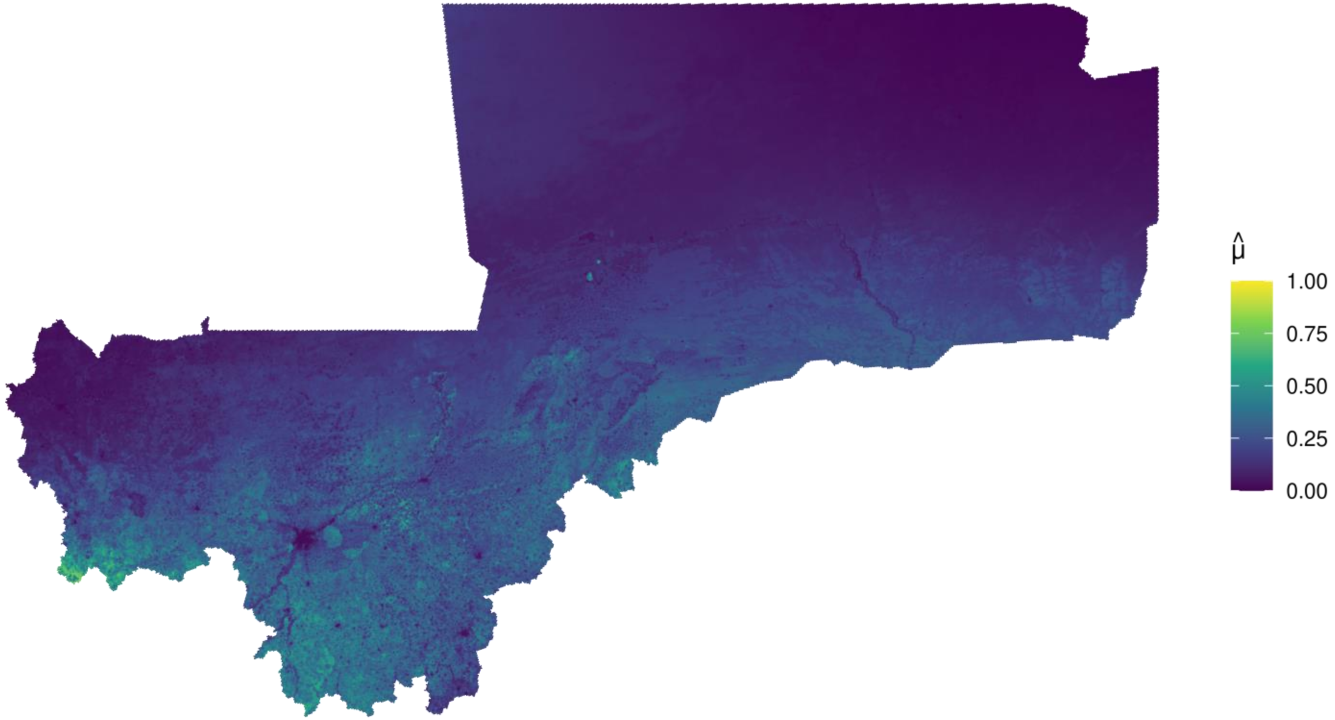


Geographic malaria risk in Mali



- Spatial effect (left) and standard error of bootstrap samples (right).

Geographic malaria risk in Mali



- Predicted malaria risk.

Discussion

Discussion

Component-wise boosting for risk identification?

➔ Yes, can allow for interesting insights in moderate dimensional datasets where a priori the functional form of an effect is not known, but difficult if rigorous notion of statistical inference is required.

Case Study 1 (Malnutrition)

➔ If effect *type* selection of interest – great. But difficult to achieve controls for variable selection: stability selection is very conservative in low dimensional settings.

➔ Probably more useful: hierarchical modelling of main factors to distinguish regional differences and deviations from country-level relationship.

Case Study 2 (Malaria)

➔ Likely competitive and interpretable approach compared to Bhatt et al. 2017 and Weiss et al. 2015 with respect to variable selection and predictive accuracy (and: appropriate response distribution).

➔ High accuracy masks large uncertainties in predicted risk, if admin 2 estimates are the objective probably not the right approach (since uncertainty is central).

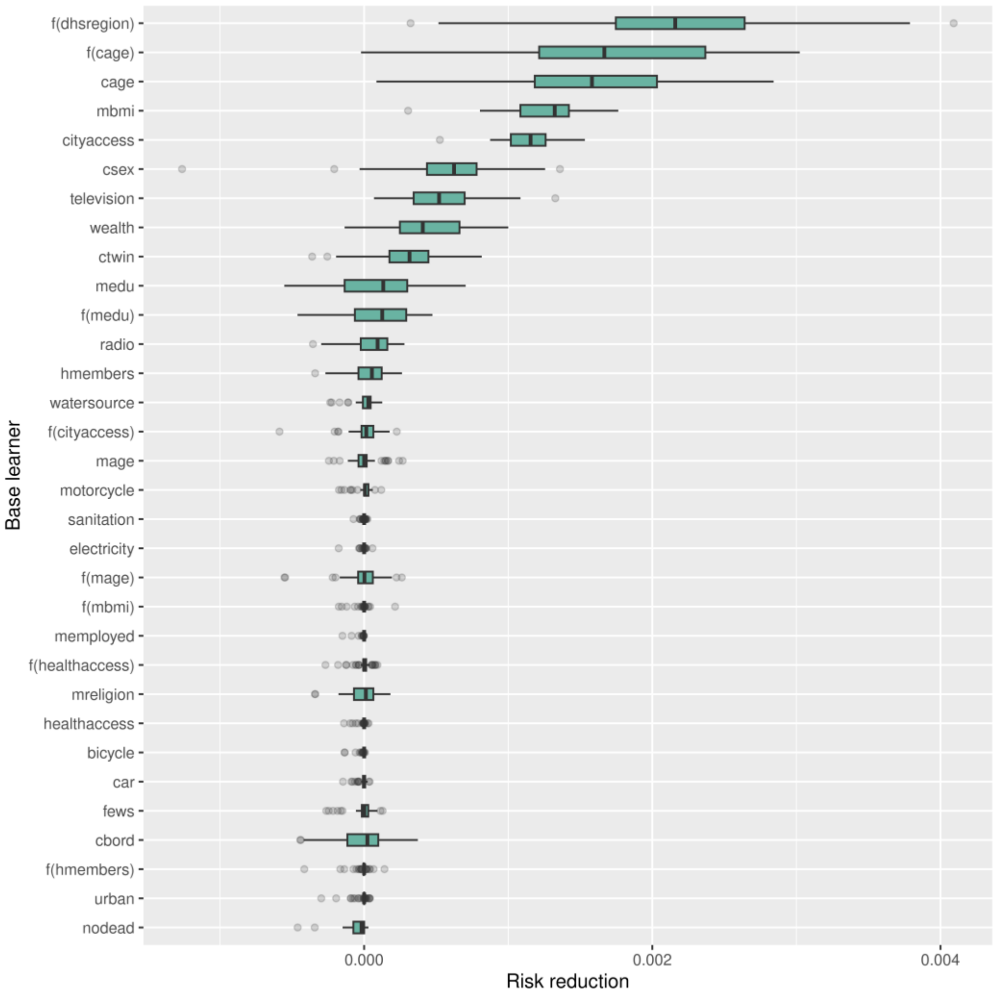
Q&A

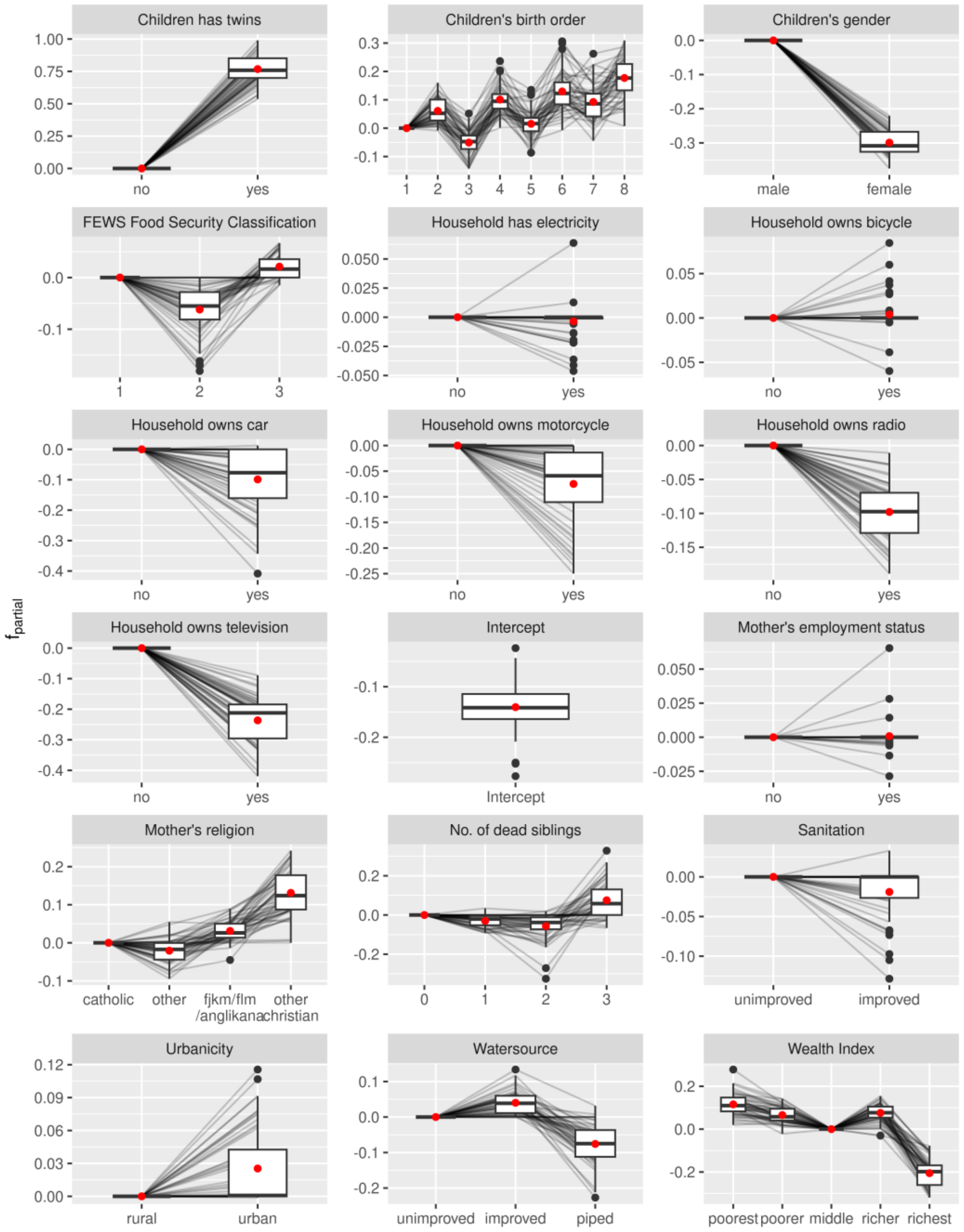
References

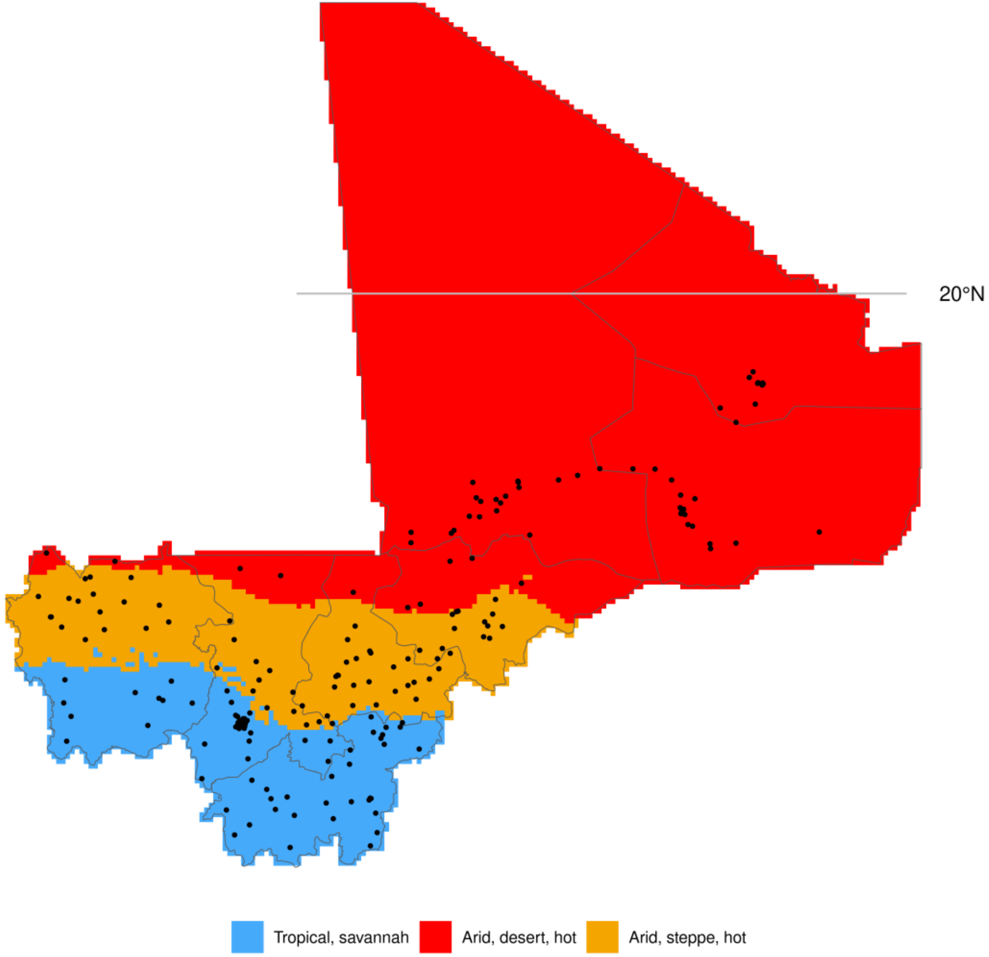
References

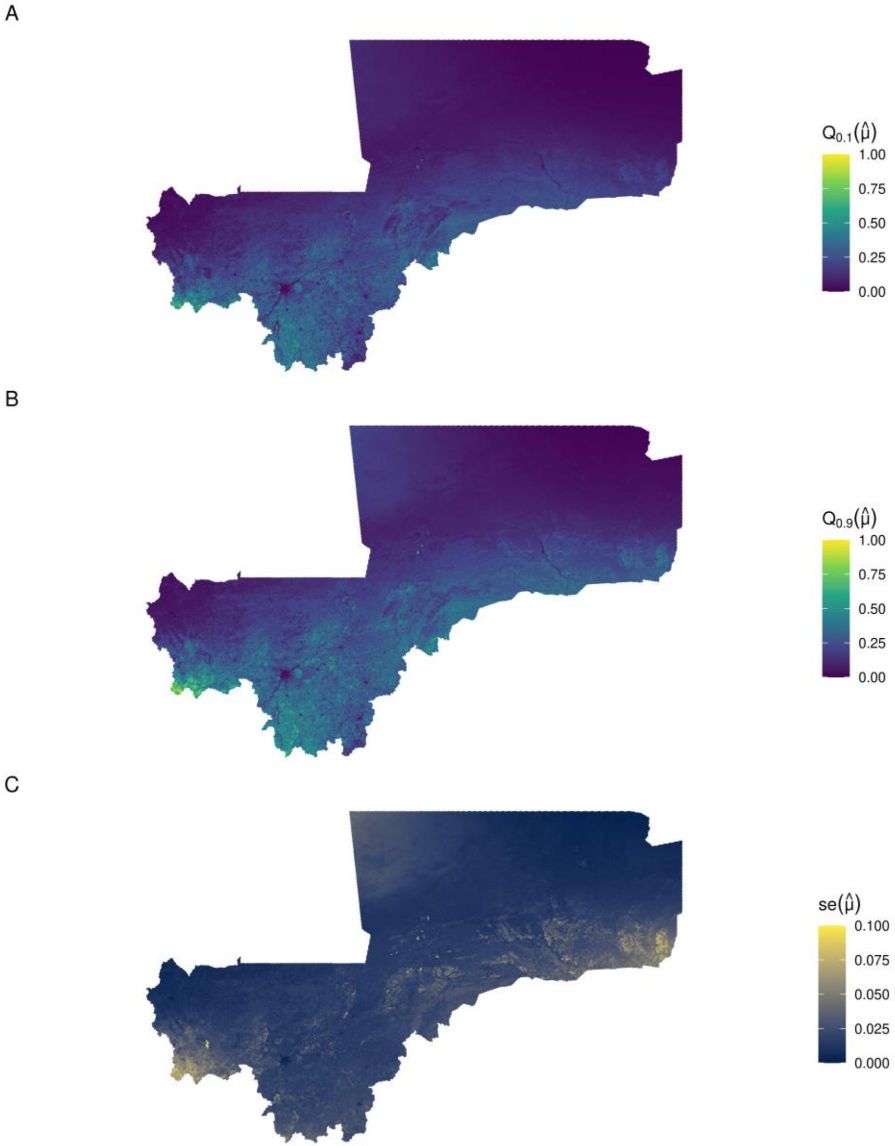
- Bhatt, Samir, Ewan Cameron, Seth R. Flaxman, Daniel J. Weiss, David L. Smith, and Peter W. Gething. 2017. ‘Improved Prediction Accuracy for Disease Risk Mapping Using Gaussian Process Stacked Generalization’. *Journal of The Royal Society Interface* 14 (134): 20170520. https://doi.org/10.1098/rsif.2017.0520.
- Bühlmann, Peter, and Torsten Hothorn. 2007. ‘Boosting Algorithms: Regularization, Prediction and Model Fitting’. *Statistical Science* 22 (4): 477–505. https://doi.org/10.1214/07-STS242.
- Bühlmann, Peter, and Bin Yu. 2003. ‘Boosting With the L_2 Loss: Regression and Classification’. *Journal of the American Statistical Association* 98 (462): 324–39. https://doi.org/10.1198/016214503000125.
- Dong, Tracy Qi, and Jon Wakefield. 2021. ‘Modeling and Presentation of Vaccination Coverage Estimates Using Data from Household Surveys’. *Vaccine* 39 (18): 2584–94. https://doi.org/10.1016/j.vaccine.2021.03.007.
- Fenske, Nora, Thomas Kneib, and Torsten Hothorn. 2011. ‘Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression’. *Journal of the American Statistical Association* 106 (494): 494–510. https://doi.org/10.1198/jasa.2011.ap09272.
- Friedman, Jerome H. 2001. ‘Greedy Function Approximation: A Gradient Boosting Machine.’ *The Annals of Statistics* 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.
- Giardina, Federica, Nafomon Sogoba, and Penelope Vounatsou. 2016. ‘Bayesian Variable Selection in Semiparametric and Nonstationary Geostatistical Models: An Application to Mapping Malaria Risk in Mali’. In *Handbook of Spatial Epidemiology*. Chapman and Hall/CRC.
- Hans, Nicolai, Nadja Klein, Florian Faschingbauer, Michael Schneider, and Andreas Mayr. 2022. ‘Boosting Distributional Copula Regression’. *Biometrics*, biom.13765. https://doi.org/10.1111/biom.13765.
- Kneib, Thomas, Torsten Hothorn, and Gerhard Tutz. 2009. ‘Variable Selection and Model Choice in Geoadditive Regression Models’. *Biometrics* 65 (2): 626–34. https://doi.org/10.1111/j.1541-0420.2008.01112.x.
- Meinshausen, Nicolai, and Peter Bühlmann. 2010. ‘Stability Selection: Stability Selection’. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (4): 417–73. https://doi.org/10.1111/j.1467-9868.2010.00740.x.
- Strömer, Annika, Nadja Klein, Christian Staerk, Hannah Klinkhammer, and Andreas Mayr. 2022. ‘Boosting Multivariate Structured Additive Distributional Regression Models’. https://doi.org/10.48550/arXiv.2207.08470.
- Thomas, Janek, Andreas Mayr, Bernd Bischl, Matthias Schmid, Adam Smith, and Benjamin Hofner. 2018. ‘Gradient Boosting for Distributional Regression: Faster Tuning and Improved Variable Selection via Noncyclical Updates’. *Statistics and Computing* 28 (3): 673–87. https://doi.org/10.1007/s11222-017-9754-6.
- Weiss, Daniel J, Bonnie Mappin, Ursula Dalrymple, Samir Bhatt, Ewan Cameron, Simon I Hay, and Peter W Gething. 2015. ‘Re-Examining Environmental Correlates of Plasmodium Falciparum Malaria Endemicity: A Data-Intensive Variable Selection Approach’. *Malaria Journal* 14 (1): 68. https://doi.org/10.1186/s12936-015-0574-x.

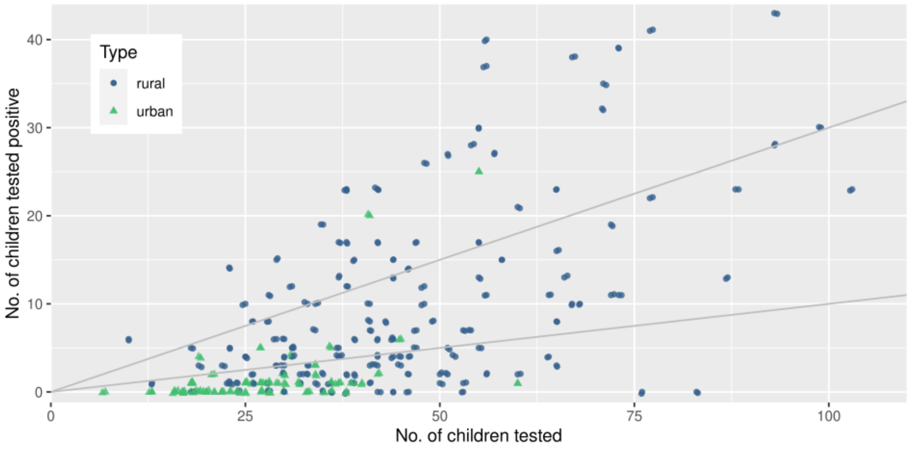
Additional Figures

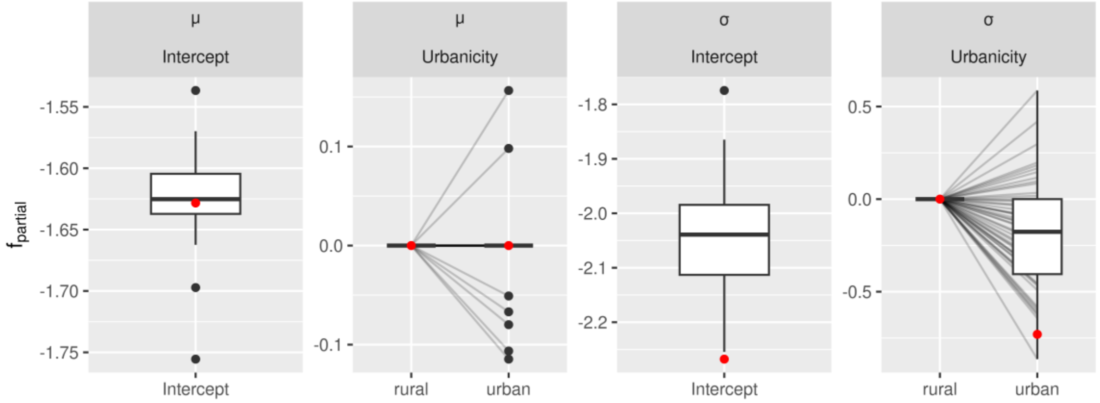

f(dhsregion)
f(cage)
cage
mbmi
cityaccess
csex
television
wealth
ctwin
medu
f(medu)
radio
hmembers
watersource
f(cityaccess)
mage
motorcycle
sanitation
electricity
f(mage)
f(mbmi)
memployed
f(healthaccess)
mreligion
healthaccess
bicycle
car
fews
cbord
f(hmembers)
urban
nodead
Base learner
0.000
0.002
0.004
Risk reduction

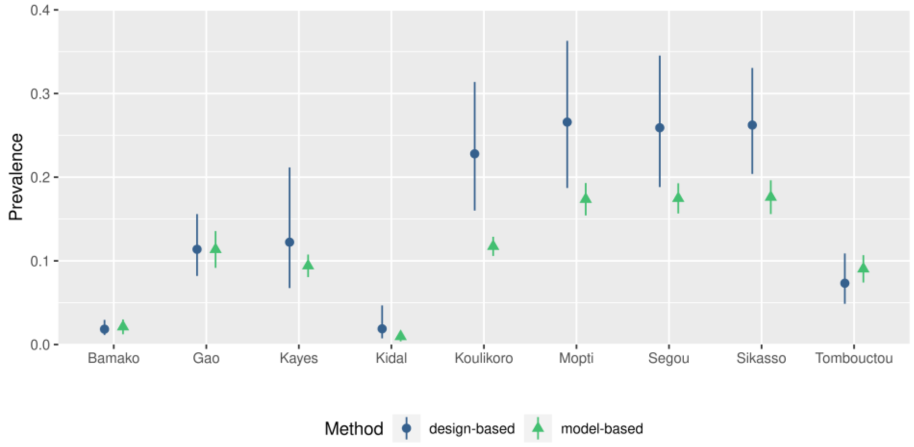

f_partial
Children has twins
1.00
0.75
0.50
0.25
0.00
no
yes
Children's birth order
0.3
0.2
0.1
0.0
-0.1
1
2
3
4
5
6
7
8
Children's gender
0.0
-0.1
-0.2
-0.3
male
female
FEWS Food Security Classification
0.0
-0.1
1
2
3
Household has electricity
0.050
0.025
0.000
-0.025
-0.050
no
yes
Household owns bicycle
0.05
0.00
-0.05
no
yes
Household owns car
0.0
-0.1
-0.2
-0.3
-0.4
no
yes
Household owns motorcycle
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
no
yes
Household owns radio
0.00
-0.05
-0.10
-0.15
no
yes
Household owns television
0.0
-0.1
-0.2
-0.3
-0.4
no
yes
Intercept
-0.1
-0.2
Intercept
Mother's employment status
0.050
0.025
0.000
-0.025
no
yes
Mother's religion
0.2
0.1
0.0
-0.1
catholic
other
fjkm/flm /anglikana
other christian
No. of dead siblings
0.2
0.0
-0.2
0
1
2
3
Sanitation
0.00
-0.05
-0.10
unimproved
improved
Urbanicity
0.12
0.09
0.06
0.03
0.00
rural
urban
Watersource
0.1
0.0
-0.1
-0.2
unimproved
improved
piped
Wealth Index
0.2
0.0
-0.2
poorest
poorer
middle
richer
richest

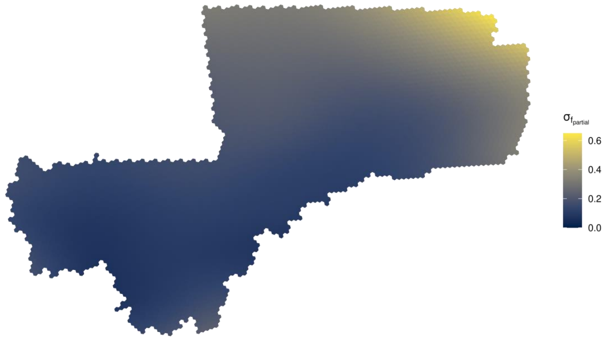

20°N
Tropical, savannah
Arid, desert, hot
Arid, steppe, hot


A
Q0.1(μ̂)
1.00
0.75
0.50
0.25
0.00
B
Q0.9(μ̂)
1.00
0.75
0.50
0.25
0.00
C
se(μ̂)
0.100
0.075
0.050
0.025
0.000


Type
rural
urban
No. of children tested positive
0
10
20
30
40
0
25
50
75
100
No. of children tested


μ
Intercept
μ
Urbanicity
σ
Intercept
σ
Urbanicity
$f_{partial}$
-1.55
-1.60
-1.65
-1.70
-1.75
Intercept
0.1
0.0
-0.1
rural
urban
-1.8
-1.9
-2.0
-2.1
-2.2
Intercept
0.5
0.0
-0.5
rural
urban


Prevalence
0.0
0.1
0.2
0.3
0.4
Bamako
Gao
Kayes
Kidal
Koulikoro
Mopti
Segou
Sikasso
Tombouctou
Method
design-based
model-based


$\sigma_{f_{partial}}$
0.6
0.4
0.2
0.0