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

Daniel A. Seussler Becerra

Master's thesis
supervised by Dr. Cornelius Fritz

Abstract

In low- and middle-income countries, nationally representative household surveys such as the Demographic and Health Surveys provide a wealth of primary data on health, nutrition, and socio-economic outcomes. For epidemiological studies, the survey data is often drawn upon to identify health risk factors, both at the individual and geographical levels. In practice, the functional form of the risk factors is not known beforehand. For instance, an effect could be linear or non-linear, if included at all. Furthermore, the increased availability of remotely sensed data provides a new data source that can be integrated into the analyses of health conditions but is not necessarily informative. The increased dimensionality of such analyses demands methods of variable selection and model choice, both to remain interpretable and generalise well to future observations. In this thesis, I employ component-wise boosting to identify risk factors of two prevalent health conditions in sub-Saharan Africa. The approach is applied in two case studies, where risk factors of individual-level outcomes of chronic childhood malnutrition and environmental correlates of the geographic prevalence of malaria are modelled. The flexible estimation of linear, non-linear and spatial effects is found to be central in the understanding of both outcomes, even improving on other non-parametric models in terms of predictive capacity. When estimating malaria risk, component-wise boosting allows for response distributions that account for excess variability at the cluster level while being superior in interpretability compared to competing approaches proposed in the literature on predictive disease mapping.

Keywords and phrases — component-wise boosting, DHS surveys, chronic childhood malnutrition, small area estimation, malaria risk

Department of Statistics, University of Munich
January 2023

The idea is to go from numbers to information to understanding.
—Hans Rosling

Contents

1 Introduction

The identification of health risk factors is central to the epidemiological understanding of disease burdens. Especially in developing countries, epidemiologists and public health researchers rely on data collected through household surveys to study risk factors of common diseases. Risk factors, or determinants, are variables that are associated with an increased risk, and protective factors with a decreased risk of the disease. Ideally, the findings of such research designs can inform public health policies and interventions aimed at reducing health risks in vulnerable populations. When studying multiple risk factors that are potentially related to the health condition studied, it is often not clear whether a variable shows a linear or non-linear effect, or if it should be included at all in the statistical model. Intrinsic model selection approaches provide an integrated approach to his modelling issue, allowing for variable selection and model choice of a set of possible factors.

Based on data from the DHS surveys, the objective is to identify background characteristics that are potentially predictive of the risk of two of the most prevalent health impairments of sub-Saharan Africa (SSA). In this manuscript, I showcase an approach from the intersection of statistical and machine learning thinking. To identify risk factors, I use component-wise boosting, which allows for intrinsic variable choice and model selection in complex additive models. Specifically, for a set of possible covariate effects, the model can identify relevant linear, non-linear or spatial effects. Also, the framework is versatile in the modelling of many statistical tasks, such as survival, quantile or cost-sensitive regression and can easily be extended to situations where joint outcomes are of interest.

Recent contributions to the literature have explored whether the nonparametric methods from the machine learning field can be adapted to such tasks. These approaches have shown great generalisation performance in a variety of applied settings. For instance, when mapping indicators such as disease prevalence, predictive performance is often desired. Moreover, such approaches tend to scale well in higher dimensional data settings. Yet, such approaches present other drawbacks, most notably that the inner structure can be considered a 'black box' and inference on model parameters is difficult. In two applications, I show that the component-wise boosting framework promises competitive performance compared to boosted trees, which are often used as the default and that the inclusion of non-linear and spatial effects underscores the necessity to consider a broader range of effect types. For malaria risk prediction, I contrast the approach to previous literature in the outcome prediction and variable selection, showing that the component-wise boosting approach proposed herein compares favourably to the modelling decisions and statistical methods provided therein.

For both analyses, I draw from household survey data collected by the Demographic and Health Survey Program (DHS). Nationally representative household surveys, such as the DHS from ICF International and UNICEF's Multiple Cluster Indicator Surveys (MICS) collect and disseminate data on important population health and socio-demographic characteristics such as nutrition, malaria, childhood mortality and family planning. The target population are generally women of reproductive age (15-49) with additional information collected for children under five years. In many low- and middle-income countries (LMICs) such surveys often represent the only source of accurate and reliable data in otherwise data-scarce settings, providing a wealth of primary data for research topics ranging from public health and epidemiology to demography and economics.

Recent years have seen a renewed interest in the statistical modelling of socio-economic and health indicators, in particular in developing countries, to monitor progress for the Sustainable Development Goals (SDGs) and provide guidance on evidence-based public interven-

tions.[1] This has motivated a large body of literature, chiefly in the area of national and subnational estimation of health and development indicators. Since much information in LMICs stems from household survey data, many of the analyses are based on the modelling of the individual or aggregated survey responses.

For the remainder of the introduction, I provide a selective review of this literature and highlight different strains of research. After a brief review of component-wise boosting I introduce the two case studies that motivated the assessment of component-wise boosting in the context of identifying health risk factors from household survey data.

Monitoring sustainable development

Data from household surveys underlay much of the current knowledge on maternal and child mortality, fertility and nutrition in low- and middle-income countries and is therefore central to the development of appropriate statistical tools. To understand and model changes and trends in population health statistics, Bayesian hierarchical models are often the approach of choice. In settings where data sparsity is common, and data points exhibit known measurement error – such as in aggregate statistics from survey data – these models allow smooth and stable estimates across space and time. On the national level, child and maternal mortality statistics are arguably one of the most widely followed health indicators for developing countries. Alkema and New (2014) propose a model for child mortality, Alexander and Alkema (2018) and Wang et al. (2022) for the neonatal mortality and stillbirth rate, respectively. A model for educational attainment by school completion rates is proposed in Dharamshi et al. (2022).

Household surveys include questionnaires on a wide array of individual topics, allowing one to study life or health patterns jointly. For example, Wade et al. (2022) discuss a multivariate regression approach to study life patterns jointly, for continuous, categorical and censored variables. Hohberg et al. (2021) provide a study of multidimensional poverty in Indonesia by modelling income and education with copulas. Furthermore, household surveys provide one very common type of survey data, to collect data on specific population characteristics, different types exist. For marked presence-only data of vulnerable populations in Malawi, Laga, Niu, and Bao (2022) provide a model-based approach to estimate the total population size.

Subnational indicators and small area estimation (SAE)

For policy research and formulation, it is often desirable to obtain estimates at subnational levels, as policies generally are implemented at administrative levels one or two below the national level. For child mortality, Mercer et al. (2015) propose a model for the admin 1 level. Subsequently extended to admin 2 (Wakefield et al. 2019) and to include census collected data (Godwin and Wakefield 2021). Dong and Wakefield (2021b) introduce a model to disaggregate immunisation coverage from routine services and supplementary vaccination campaigns to inform the latter. The subnational coverage of Measles-containing-vaccine first-dose (MCV1) immunisation off of household surveys is an extension of the space-time smoothing approach proposed in Mercer et al. (2015).

1. In the 2030 Agenda for Sustainable Development, the General Assembly of the United Nations laid out the 17 Sustainable Development Goals (SDGs), a framework to globally mobilise efforts to eradicate poverty and foster economic, social and environmental development. Specifically, the resolution calls for a systematic review of the implementation, stating "[t]hey will be rigorous and based on evidence, informed by country-led evaluations and data which is high-quality, accessible, timely, reliable and disaggregated by income, sex, age, race, ethnicity, migration status, disability and geographic location and other characteristics relevant in national contexts" (United Nations 2015, p. 32).

Admin 1 is feasible as household survey data most often allow for design-based estimates one level below the national level. For one administrative level below, admin 2, this is often not attainable. Here, model-based geo-statistical approaches are widely used to obtain geographically fine-scaled predictions at the grid-cell level. See Giorgi and Diggle (2021) and Diggle and Giorgi (2016) for a comprehensive introduction to the topic. This is in line with the push for 'precision public health', the (geographic) targeting of small populations with specific health interventions (Dowell, Blazes, and Desmond-Hellmann 2016; Desmond-Hellmann 2016). For country mappings of metrics such as vaccine coverage, this has led to a sheer explosion of research designs. But many statistical approaches did not appropriately account for the complex survey design of the survey data. Dong and Wakefield (2021a) provide recommendations and I further discuss those in light of the second case study in section 5. Briefly, geographic mapping of health indicators can be categorised in either design-based or model-based approaches. The former provides estimates for sub-regions based on the survey design, while the model-based approach informs estimates with additional covariates and often spatial effects to borrow information from nearby observations. In general, design-based estimates of population statistics should be preferred since those are compliant with the complex survey design. But, DHS surveys are most often only designed to provide estimates at admin 1, the design-based approach may not be applicable for admin 2 because of data sparsity. See Fuglstad, Li, and Wakefield (2022) for an in-depth discussion of this topic and Paige et al. (2022), where different approaches are evaluated on simulated surveys.

The aforementioned literature on (model-based) approaches to estimate local health conditions requires a selection of explanatory variables that are included. Especially when a large number of possible covariates are available, derived for example from survey answers or remotely sensed covariates for ecological correlates, one often aims for a sparse model which includes only the most relevant variables. Furthermore, beyond variable selection, it might be desirable to identify whether a continuous covariate has a linear or non-linear effect and whether a varying coefficient term or spatially varying terms is appropriate. The framework of component-wise boosting allows for intrinsic variable choice and model selection. Note, however, that the framework is not intended to uncover causal relationships.

Component-wise boosting and competing approaches

Boosting originated in the machine learning literature as a method for classification tasks (Freund and Schapire 1996), has since been extended to other contexts and widely adopted due to its superior performance in many real-world applications (Chen and Guestrin 2016). Friedman, Hastie, and Tibshirani (2000) and Friedman (2001) described boosting in terms of functional gradient descent, connecting the method to the more conventional statistical framework of maximum likelihood estimation. Briefly put, a weak learner – or base learner – is fitted iteratively to the negative gradient of a pre-specified loss function and the estimated learner is added to the additive predictor. In practice, the weak learner is often chosen to be a shallow tree. But the approach can also be model-based, with a (penalised) least squares regression used as weak learner (Bühlmann and Yu 2003). If those are fit component-wise, only the best-fitting learner is selected in each iteration. This, if some learners are never selected, yields an intrinsic selection of included learners.

Thus, by carefully selecting the set of possible learners, complex models can be fitted for a variety of response distributions with data-driven variable selection. Since a semi-parametric model is obtained as a result, this approach can also be viewed through an interpretable machine learning lens. I discuss component-wise boosting in section 3.

Alternatively, intrinsic effect selection for generalised additive models can be accomplished

Country	Type	Year	Fieldwork
Madagascar	Standard DHS	2021	March 2021 - July 2021
Mali	Malaria Indicator Survey	2021	September 2021 - November 2021

Table 1: *Selected surveys for the two case studies.*

in a Bayesian framework with spike-and-slab priors, see Scheipl, Fahrmeir, and Kneib (2012) and Klein et al. (2021) for more information. While providing the benefit of straightforward uncertainty quantification, such approaches tend to not scale well in moderate to high-dimensional settings.

Case studies

I present two case studies that resemble typical research designs. In both case studies, variable selection and model choice is of particular interest. First, individual-level risk factors of chronic malnutrition as indicated by low height-for-age for children under five years. I use data from the Madagascar 2021 Standard DHS. Second, the identification of environmental and climatic predictors of cluster-level malaria prevalence, as tested with Rapid Diagnostic Tests (RDTs) in Mali. I use data from the Mali 2021 Malaria Indicator Survey (MIS). Table 1 shows the surveys selected for the two case studies.

Both topics have been treated extensively in the literature, mapping risk most often with a Bayesian model-based geo-statistical approach.[2] In this manuscript, I employ the component-wise boosting approach, which has been proven useful in similar studies (Fenske, Kneib, and Hothorn 2011; Torres Munguía and Martínez-Zarzoso 2021). For mapping geographic malaria risk, in particular, one is especially interested in the predictive accuracy of the model, as such risk maps can be used to inform local public health interventions and elimination campaigns. Malaria transmission risk is highly dependent on environmental and climatic factors, hence it is convenient to use remotely sensed covariates to inform local estimates. To improve prediction accuracy, Bhatt et al. (2017) propose a stacked ensemble approach with multiple common statistical and machine learning approaches embedded in a geo-statistical regression. However, the authors rely on ad-hoc transformations to adopt the cluster-level count data to common statistical software packages. The component-wise boosting discussed herein, where the model choice of smooth and spatial effects is intrinsic, I argue, achieves similar performance while accommodating the binomial nature of the survey data.

The remainder of the manuscript is structured as follows. In section 2, I provide an overview of survey data from the DHS and discuss the survey design commonly used in such surveys. In section 3, I discuss component-wise boosting and different resampling strategies for hyper-parameter selection. The two case studies introduced above are discussed in section 4 and section 5. Finally, in section 6, I discuss the findings and offer thoughts on future research.

2 Demographic and Health Surveys

The Demographic and Health Surveys (DHS) are nationally representative household surveys in developing countries. To date, more than 400 surveys in over 90 countries were conducted

2. See, for example, Aheto et al. (2017), Kinyoki et al. (2020), Egbon, Belachew, and Bogoni (2022), and Uwiringiyimana et al. (2022) for childhood malnutrition. Diggle et al. (2002) provides an early application of model-based geo-statistics to malaria prevalence, see Weiss et al. (2019), Ejigu (2020), and Nzabakiriraho and Gayawan (2021) for more recent discussions.

with the assistance of the DHS program. For low- and middle-income countries surveys are organised in 3 to 5-year intervals and data on a broad range of health and socio-economic outcomes are surveyed. In particular, the surveys collect information on demographic characteristics such as fertility and mortality, health outcomes such as reproductive health and nutrition, and economic indicators. While some components of the survey vary depending on the need of the particular country, the components themselves are standardised and allow for comparisons across time and countries. This consistency is one of the key reasons data that DHS data is ubiquitous in many research designs.[3] See, for example, Corsi et al. (2012) for a profile of DHS data in epidemiology. For research purposes, the micro-data from the surveys is available upon request from the corresponding website.[4]

Though the following discussion is centred around the DHS, much of it applies to the Multiple Cluster Indicator Surveys (MICS) done by UNICEF, since those follow similar survey designs (Khan and Hancioglu 2019). To achieve valid estimates of population-level indicators, the DHS surveys usually employ a two-stage cluster sampling approach, which will be described next.

2.1 Survey sample design

The target population of the survey are women aged 15-49 and children below the age of five in residential households. Often men (or a random subsample of men) aged 15-59 are interviewed too. To obtain a probability sample from the population, a DHS survey generally employs a two-stage cluster sampling design. In the first stage, primary sampling units (PSUs), also called clusters, are drawn from a pre-specified sampling frame. This is most often a recent population census, where the census enumeration areas (EAs) are the units drawn with probability proportional to size (PPS). If no recent population census is available, sampling frames can be adapted from alternative sources such as remotely sensed night-time light intensity imagery. To pre-select households in each chosen cluster, a complete enumeration of households is done. In the second stage, a fixed number of households are drawn randomly from the enumerated households and selected for the survey (ICF International 2012).

Additionally, DHS surveys are commonly stratified by design domains to improve the precision of the estimates for sub-populations. DHS surveys are generally stratified by administrative regions crossed with urbanicity, that is, urban or rural populations. Some strata may account for a very low share of the population and, as a consequence, survey estimates are not sufficiently precise. Therefore, in some DHS surveys, some strata are oversampled. In every survey, each sampled individual is assigned a survey weight, which loosely quantifies the relative number of individuals the observed individual represents. In the DHS surveys, this is generally computed as the product of the inverse sampling probabilities at each stage, adjusted for non-response patterns. Due to privacy concerns, only the weight is disclosed and not the individual sampling probabilities.

2.2 Design-based estimates

To estimate population-level totals or means from survey data with a complex sampling design, it is imperative to use the provided survey weights to account for the complex sur-

3. Even if administrative data is available, household survey data may provide an unbiased assessment of population indicators. See, for example, Sandefur and Glassman (2015). The authors compare vaccination coverage and school enrolment rates from administrative and DHS data. The authors find significant discrepancies in the reported statistics and argue that this misrepresentation is not merely due to the lack of analytical capacity, but to weak state capacity and incentive structures of donors in highly aid-dependent countries.

4. See https://dhsprogram.com.

vey design. The estimate of the country-level prevalence of an indicator can be obtained with the Horvitz–Thompson (HT) estimator (Horvitz and Thompson 1952). For instance, let $i = 1, \ldots, N$ index the sampled children and w_i designate the survey weight of children i. Then,

$$\hat{p} = \frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i}$$

where y_i is a 0/1 indicator for child i, indicating the absence or presence of the condition under study. Furthermore, one can obtain design-based estimates of the variance for $\hat{p}$. The design-based estimates can be computed with the `survey` package (Lumley 2004).

3 Methodology

In this section, I discuss component-wise gradient boosting for (distributional) regression. Boosting can be seen as a general framework with specialised extensions for different tasks. Originally proposed in Friedman, Hastie, and Tibshirani (2000) and Bühlmann and Yu (2003), functional gradient boosting can be understood as an optimisation method to fit Generalised Additive Models (GAMs) (Hastie and Tibshirani 1986; Wood 2017). A comprehensive treatment can be found in Bühlmann and Hothorn (2007).[5]

3.1 Component-wise boosting

Let $(y_i, \mathbf{x}_i),\ i = 1, \ldots, N$ be observations where y is the response variable and $\mathbf{x}$ a vector of explanatory covariates. In a structured additive regression framework, the mean outcome is modelled through an additive predictor $\eta(\mathbf{x})$ with an inverse link function h. Then,

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

In particular, the additive predictor can include components f_j that account for linear and smooth effects, varying coefficient terms or spatial effects. The components are derived from a pre-selected set of base learners. These are commonly simple models such as linear regression or penalized splines with a pre-specified low degree of freedom. I provide an overview below. For a suitable loss function ρ, boosting seeks to estimate the function η through functional gradient descent. The corresponding optimisation problem is defined as

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

The loss function is frequently chosen to be $(y - \eta)^2$ for (mean) regression or the negative log-likelihood in more general cases. For example, we will use the negative binomial log-likelihood for logistic regression in one of the case studies below. To implement the approach in practice, the expectation is replaced by the empirical risk,

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

5. See also Coors et al. (2021) for component-wise boosting in an interpretable automated machine learning framework.

Algorithm 1: Component-wise boosting (Bühlmann and Hothorn 2007).

Data: $(y_i, \mathbf{x}_i),\ i = 1, \ldots, N$.

Setup: Let $\mathcal{B} = \{b_1 \ldots, b_L\}$ be the pre-specified set of base learners.

Step 1: Set $m = 1$. Initialise $\hat{\eta}^{[0]}$, a common option is to set

$$\hat{\eta}^{[0]} = \arg\min_c \sum_{i=1}^{N} \rho(y_i, c).$$

Step 2: Compute the negative functional gradient

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

and fit each of the base learner b_l for $l = 1, \ldots, L$ to the current value.

Step 3: Select the best fitting base learner l^*, i.e., the base learner which minimises the residual sum of squares:

$$l^* = \arg\min_l \sum_{i=1}^{n} \Big(u_i - \hat{b}_l(\mathbf{x}_i)\Big)^2.$$

Step 4: Update the estimate of the additive predictor by the selected base learner b_{l^*},

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

The step-length ν induces shrinkage and is commonly fixed at a small value, typically $\nu = 0.1$ or $\nu = 0.01$. Increase m by 1.

Step 5: Repeat steps 2 - 4 until $m > m_{\texttt{stop}}$.

After initialisation of the additive predictor $\hat{\eta}$ and the pre-specification of the set of base learners, boosting iteratively evaluates the negative gradient of the loss function at the predicted values of the previous iteration. Each base learner is then separately fitted to the negative gradient, hence the name component-wise. In each iteration, only the best-fitting learner is multiplied with a shrinkage parameter and added to the additive predictor. This procedure is reiterated until the number of maximum boosting iterations is reached. The details of the algorithm are provided in algorithm 1.

If left to run until convergence, the algorithm will recover the maximum likelihood estimates of each base learner. However, this may result in suboptimal generalisation performance. The model is said to over-fit the data and the capacity of the model to predict unseen data is reduced. Therefore, selecting an optimal value of boosting iterations $m_{\texttt{stop}}$ for the expected generalisation performance is crucial. In practice, the algorithm is left to run a high number of iterations T, then pruned to an earlier iteration $m_{\texttt{stop}} < T$, where the resampled risk is minimised. I discuss common resampling methods in subsection 3.3.

Note, by selecting exactly one base learner to be added to the model in each iteration (Step 3), the selection of components into the model is implicit. If one base learner is never selected, the partial contribution is estimated to be zero and the component is effectively excluded from the model. By specifying univariate components and decomposing non-linear effects into their

linear and non-linear deviation, variable selection and model choice is intrinsic. I expand on this further in subsection 3.2.

Given the regularised estimates of coefficients and the variable selection properties of the algorithm, it is difficult to obtain inferences for model parameters similar to a conventional regression framework. Hofner, Kneib, and Hothorn (2016) propose the construction of confidence intervals by refitting the model to bootstrap samples of the original data set. Subsampling replications achieve a similar objective. Alternatively, stability selection provides finite sample control on the expected number of falsely selected variables. (Meinshausen and Bühlmann 2010; Shah and Samworth 2013). See also Hofner, Boccuto, and Göker (2015) for a discussion in the context of boosting. In practice, the approach tends to be very conservative, as noted in the simulations in Hofner, Boccuto, and Göker (2015) and Thomas et al. (2017).

The boosting approach as formulated optimisation problem allows for a very flexible implementation of many regression settings. Besides the general case described above, the setup allows for quantile regression (Fenske, Kneib, and Hothorn 2011), survival analysis (Bühlmann and Hothorn 2007), and cost-sensitive boosting (Kriegler and Berk 2010). Furthermore, complex joint likelihoods can be modelled, see Strömer, Klein, et al. (2022) for multivariate distributions, Hans et al. (2022) for copulas, and Griesbach, Groll, and Bergherr (2021) for boosting longitudinal and survival data jointly. I describe one particular extension next, boosting distributional regression models.

Boosting distributional regression

In some settings, it may be more informative to model the response distribution of interest beyond the mean. This is commonly referred to as distributional regression.[6] The algorithm described above can be extended to fit distributional models, as described next. In distributional regression, each parameter of the response distribution can be modelled separately, thereby allowing inference not only on the mean but also on other properties of the response distribution. Let

$$f_{\texttt{dens}}(y|\mu, \sigma, \nu, \tau)$$

be the density of interest. The parameters μ, σ, ν, τ are commonly referred to as location, scale, shape and kurtosis, respectively. For brevity, I denote the possible parameters of the distribution in a vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$. For each, a structured additive predictor η_k and an appropriate, fixed inverse link function h_k is defined:

$$\begin{aligned} \theta_k &= h_k(\eta_k(\mathbf{x})) \\ \eta_k(\mathbf{x}) &= \beta_{0k} + \sum_{j=1}^{J_k} f_{jk}(\mathbf{x}), \quad k = 1, \ldots, 4. \end{aligned}$$

Analogously to the non-distributional case before, the loss function ρ is set as the negative log-likelihood and the optimisation problem can be defined as

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

where $\eta = (\eta_{\theta_1}, \eta_{\theta_2}, \eta_{\theta_3}, \eta_{\theta_4})$ is the vector of the additive predictors. Mayr et al. (2012) first

6. Rigby and Stasinopoulos (2005) and Klein et al. (2015) provide a frequentist and Bayesian treatment, respectively. Distributional regression has also received ample interest from the machine learning field, as it can be used to quantify prediction uncertainty. See Duan et al. (2020) for boosting score functions and Schlosser et al. (2019), who propose distributional forests as an extension to random forests in a distributional setting.

Algorithm 2: Non-cyclical component-wise boosting (Thomas et al. 2018).

Data: $(y_i, \mathbf{x}_i),\ i = 1, \ldots, N$.

Setup: Let $\mathcal{B}_k = \{b_{k1} \ldots, b_{kL}\}$ for $k = 1, \ldots, 4$ be the pre-specified set of base learners.

Step 1: Set $m = 1$. Initialise $\hat{\eta}_{\theta_k}^{[0]}$, a common option is to set

$$\hat{\eta}_{\theta_k}^{[0]} = \arg\min_c \sum_{i=1}^{N} \rho(y_i, c).$$

Step 2: For each $k = 1, \ldots, 4$, do:

1. Compute the negative functional gradient

$$u_{ki} = \frac{1}{\partial \eta_{\theta_k}} \rho(y, \eta)\Big|_{\eta = \hat{\eta}^{[m-1]}(\mathbf{x}_i)}, \quad i = 1, \ldots, N.$$

   and fit each of the base learner $b_{kl}, l = 1, \ldots, L$ to the current value.

2. Select the best fitting base learner l^* by the inner loss:

$$l^* = \arg\min_l \sum_{i=1}^{N} \left(u_{ki} - \hat{b}_{kl}(\mathbf{x}_i)\right)^2$$

3. Calculate the change in the outer loss

$$\Delta\rho_k = \sum_{i=1}^{N} \rho\left(y_i, \hat{\eta}_{\theta_k}^{[m-1]}(\mathbf{x}_i) + \nu \hat{h}_{kl^*}(\mathbf{x}_i)\right)$$

Step 3: Select the parameter that results in the lowest risk:

$$k^* = \arg\min_k (\Delta\rho_k).$$

Step 4: Update the additive predictor for the parameter where the largest loss reduction was achieved with the best fitting learner for this parameter:

$$\hat{\eta}_{\theta_{k^*}}^{[m]} = \hat{\eta}_{\theta_{k^*}}^{[m-1]} + \nu\, \hat{h}_{k^* l^*}.$$

The hyper-parameter ν induces shrinkage and is commonly fixed at a small value, typically $\nu = 0.1$ or $\nu = 0.01$. Increase m by 1.

Step 5: Repeat steps 2 - 4 until $m > m_{\texttt{stop}}$.

proposed a cyclical approach to estimation, where the parameters are updated consecutively, each conditioning on the previous iterations up to the current state for each parameter. The following non-cyclical approach was proposed in Thomas et al. (2018) and in each iteration, both the base learner and the parameter of the distribution are selected jointly for an update based on the largest risk reduction. The details of the algorithm are provided in algorithm 2.

3.2 Base learner, variable selection and model choice

The set of base learner $\mathcal{B}$ determines the components that can be selected into the additive predictor in each boosting iteration, and thus, the specification of the additive model to be fitted. In practice, a variety of linear, non-linear, and spatial effects can be included. For the case studies, I focus on the following effect types.

- *Linear effects* can be included by simple univariate linear regressions. Categorical effects can be encoded in either treatment or 'dummy' coding.
- *Smooth effects* to model non-linear associations of continuous covariates can be included with P-Splines (Eilers and Marx 1996; Schmid and Hothorn 2008).
- *Spatial effects* can be included for continuous spatial data or areal data with a first-order neighbourhood structure. In the former, bivariate P-splines can be used to model smooth interaction surfaces, in the latter Markov random field effects (see Kneib, Hothorn, and Tutz (2009) and Sobotka and Kneib (2012), respectively).

In each boosting iteration, exactly one base learner is which yields a selection of included components, as some may never be selected.[7] Note, that the selection into the model is determined greedily by the relative contribution of the variable to the risk reduction, and not by some concept of statistical significance. To achieve an unbiased selection of linear and non-linear effects for continuous variables, Kneib, Hothorn, and Tutz (2009) introduce a decomposition of the non-linear effect into a parametric and a non-linear deviation from the parametric component:

$$f_{\texttt{smooth}}(\mathbf{x}) = b_{\texttt{param}}(\mathbf{x}) + b_{\texttt{centred}}(\mathbf{x}) \tag{1}$$

This decomposition allows the algorithm to select none, only the parametric or only the non-linear component. See Kneib, Hothorn, and Tutz (2009) for details. In addition, the decomposition allows for a pre-specification of a comparable complexity of base learners, which is required for the unbiased selection (Hofner et al. 2011).

3.3 Early stopping and resampling methods

Two hyper-parameters control the amount of regularization applied in the estimation. First, the step-length ν controls the contribution of the selected base learner in each iteration, therefore inducing shrinkage of the estimates. In practice, the value is of secondary importance, so long it is chosen to be small enough (Schmid and Hothorn 2008). Of higher importance is the number of boosting iterations $m_{\texttt{stop}}$. Model selection, by choosing the amount of regularization, is commonly achieved by resampling methods such as cross-validation, bootstrap

7. Especially in low-dimensional regression cases component-wise boosting is known to select too many variables. These additional variables are characterised by their small contribution to the prediction accuracy and small coefficients. Strömer, Staerk, et al. (2022) propose a method to deselect base learners based on their contribution to loss (risk) reduction. Since the share of reduced risk from the overall risk reduction can be interpreted as 'importance', the base learners with the least importance are excluded and the model refit. The authors show that this heuristic successfully removes noise variables erroneously selected by the boosting procedure while maintaining its predictive capacity. However, the reasonableness of the approach depends on the distribution of attributable risk to the base learners and the assumption of sparsity. If the model is indeed composed of many covariates, of which each is of low importance then the approach might deselect meaningful – albeit of low importance in terms of risk attribution – covariates from the model.

validation or subsampling. Specifically, such methods provide an estimate of the average prediction error (risk) across unseen data sets (Bates, Hastie, and Tibshirani 2022; Hastie, Tibshirani, and Friedman 2009). The optimal boosting iteration $m_{\texttt{stop}}$ is selected as the iteration where the minimum average risk over the holdout folds or data is attained. I provide a selective overview, for a comprehensive treatment see Raschka (2020) and Bischl et al. (2012).

In cross-validation, the data is split into several folds K, typically 5 or 10. The model is then refit repeatedly to the data each time leaving out a different fold. Formally, let κ be the function assigning a fold to each data point,

$$\kappa : \{1, \dots, N\} \to \{1, \dots, K\},$$

and denote with $\hat{m}^{-k}$ the fitted model trained with the kth fold of the data excluded. The cross-validated estimate of the risk is

$$\text{CV}(\hat{m}) = \frac{1}{N} \sum_{i=1}^{N} \rho(y_i, \hat{m}^{-\kappa(i)}(\mathbf{x}_i)).$$

For bootstrap resampling, the subset generation is done by random sampling with replacement from the original dataset and of the same size as the original dataset. For the estimate of the predictive risk, one then only considers observations not selected into the bootstrap sample.[8] Denote the set of bootstrap samples where observation i is *not* contained in C^{-i}, then:

$$\text{BTS}(\hat{m}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|C^{-1}|} \sum_{b \in C^{-1}} \rho(y_i, \hat{m}^b(\mathbf{x}_i)).$$

Here $\hat{m}^b$ is the model fitted to the bootstrap sample b. A common number of bootstrap samples employed are 25, 50 and 100. The bootstrap sample exhibits oversampling of some observations, which can induce a pessimistic bias on the estimate. One possible remedy was proposed in Efron (1983) with the 0.632 bootstrap, where the estimate is a weighted average of the training error and hold-out error. In cases where a model strongly over-fits, for example achieves perfect training error, this estimate is clearly problematic. For this case, Efron and Tibshirani (1997) propose a variant where the weights are determined adaptively.

In subsampling, a random subset of size $\lfloor N/2 \rfloor$ is drawn from the data without replacement. The model is fitted on the subset and evaluated on the data points not included. One then averages the risk over the holdout sets of the repeated sub-samples. A common number of subsampling samples employed are 25, 50 and 100.

Lastly, the holdout validation splits the data once, fits the model on the training data and estimates the generalisation performance on the validation set. Common proportions are 70/30 or 80/20. For performance estimation and model selection, the data can be split into training/validation/test sets (three-way holdout method). This approach, however, is often not recommended for smaller data sets, as fewer samples in the holdout set increase the variance of the estimator (Raschka 2020).

If variable selection is of primary interest, 'probing' provides a computationally cheap option to stop the algorithm (Thomas et al. 2017). In this method additional noise variables – unrelated to the outcome – are included as explanatory variables, once one of the noise variables is selected into the model, the algorithm is stopped.

8. On average, about a third of the observations will not be selected into the bootstrap: $P(\text{observation in bootstrap sample}) = 1 - (1 - 1/N)^N \approx 1 - \exp^{-1} = 0.632$.

In general, there are few clear recommendations in the literature to model evaluation and selection and the use of resampling methods should be assessed on a case-by-case basis. Kohavi (1995) provides a simulation study, recommending (stratified) 10-fold cross-validation as default for many applications. For complex survey data, Wieczorek, Guerin, and McMahon (2022) argue to account for the design in the fold structure of cross-validation. That is, in the case of two-stage cluster sampling stratify folds by the survey strata and select in only complete clusters of observations. In practice, many strata consist of only very few clusters, so the second recommendation sets a restrictive limitation on the number of folds if not completely prohibiting it.

4 Childhood malnutrition in Madagascar

In the first case study, the objective is to identify individual risk factors of malnutrition. Chronic childhood malnutrition remains endemic in large parts of sub-Saharan Africa (Roser and Ritchie 2019) and the identification of relevant risk factors may guide interventions and policy. Similar analyses are ubiquitous in the literature and this case study was motivated by the research done in Fenske, Kneib, and Hothorn (2011) and Fenske et al. (2013) using data from the 2005/2006 India National Family Health Survey. Here, I use recent data from the Madagascar 2021 Standard Demographic and Health Survey.

4.1 Introduction

Madagascar is an island country off the coast of East Africa. A former French colony, Madagascar today is counted towards the least-developed countries in the United Nations Classification. From an estimated 28.5 million inhabitants, as of 2021, 61% are characterised as rural and 40% below the age of 14 years (World Bank 2022). The country has one of the highest shares of undernourished children, as per 2021 DHS a share of 39.8% classifies as stunted. More recently, Madagascar became infamous for experiencing famine-like conditions of what experts have called the first famine caused by Global Warming.[9] With the primary economic activity being subsistence farming, food security is highly dependent on rainfall seasons, in particular in rural regions.

Childhood malnutrition is commonly assessed by three measures: stunted, wasted and underweight. A child classified as stunted has a *height* that is two standard deviations below the median height-for-age as determined by the World Health Organization's Child Growth Standards. Wasting refers to the condition that a child is too thin for the respective height and underweight to the condition of too low weight for age. Unlike the latter two, stunting is largely considered to be irreversible after the first 1000 days of life and is therefore often taken to be the primary indicator for chronic childhood malnutrition (Dewey and Begum 2011; Victora et al. 2021). As a health condition, impaired growth has been linked to reduced cognitive development, educational and economic outcomes, long-term effects that have put the reduction of childhood malnutrition in the focus of development targets (Dewey and Begum 2011). McGovern et al. (2017) provide a review of literature linking chronic malnutrition to economic outcomes.

Correlates of childhood malnutrition are certainly not an understudied subject, publications that employ a variety of statistical approaches have been mentioned above. In the literature, two related trends can be identified. First, the shift to include remote sensed or satellite data to inform local environmental conditions and proxy economic shocks (see, for example,

9. See https://time.com/6081919/famine-climate-change-madagascar/.

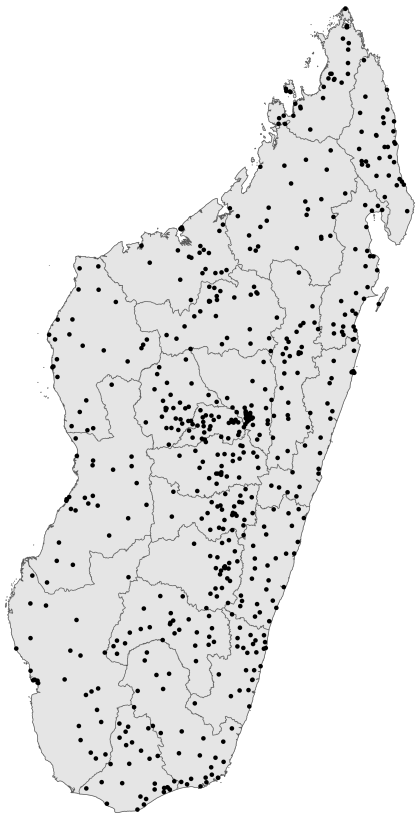

Figure 1: *Childhood malnutrition in Madagascar: designated survey regions and cluster locations of the Madagascar 2021 Standard DHS.*

Grace et al. (2022), van der Merwe, Clance, and Yitbarek (2022), and Seiler et al. (2021)). Specifically, as research attempts to assess the effects of anthropogenic climate change on socioeconomic and health outcomes, climatological variables present an important source of information. See Phalkey et al. (2015) for a review of the research concerning malnutrition. Second, the shift to employing predictive approaches that stem from the machine learning community. This coincides with the need to include a larger number of explanatory variables facilitated by the availability of remotely sensed covariates. For example, Browne et al. (2021) suggest a multivariate random forest approach to predict wealth and malnutrition scores originating from DHS data jointly. Kim et al. (2021) map malnutrition indicators at high spatial resolution by first estimating cluster-level probabilities in a hierarchical regression and in a second step employing a semi-supervised regression approach to assign clusters to villages.

The component-wise boosting approach discussed herein is favourably suited to include a large number of (potentially uninformative) covariates to obtain effect selection for covariates of interest.

Madagascar 2021 Standard DHS

The Madagascar 2021 Standard DHS was designed to provide estimates of population health indicators at the national level, the 22 administrative regions plus the capital, and for urban and rural populations. With the capital being urban only, this resulted in 45 strata. Based on

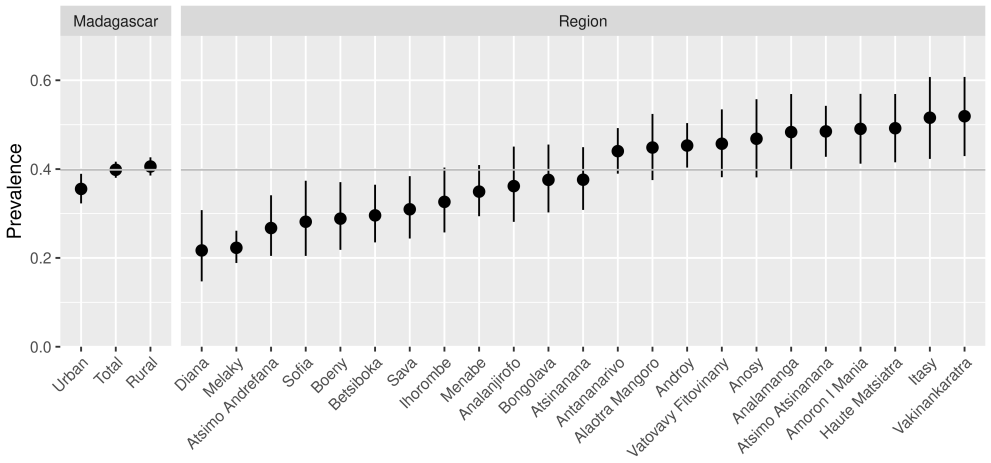

Figure 2: *Childhood malnutrition in Madagascar: design-based estimates of prevalence of moderately or severely stunted children under 5 years. Error bars indicate 95% confidence intervals. Data from the Madagascar 2021 Standard DHS.*

the population census of 2018, a total of 657 clusters were drawn with probability proportional to population size, in each stratum independently. In each cluster, 34 households were drawn at random. This resulted in a sample size of 20'510 households, whereof 5'146 were urban and 15'364 rural (Institut National de la Statistique (INSTAT) and ICF 2022). Figure 1 plots the boundaries of the survey regions and the locations of the survey clusters.

The DHS surveys take anthropometric measurements from children under five years and provide the z-score of the height-for-age of each child. The analysis includes all children under five years and uses the binary indicator *stunted* for all classified as moderately or severely stunted, that is, all children whose z-score is at least two standard deviations from the reference value. Further details on the compilation of the statistics can be found in Croft, Marshall, and Allen (2020). Figure 2 plots the design-based estimates of the regional and country-level prevalence with the corresponding confidence intervals.

Explanatory variables

As potential risk factors (or, conversely, protective factors), I include variables that previous research has considered in similar analyses. Specifically, the selection of explanatory variables is guided by the selection as discussed in Fenske, Kneib, and Hothorn (2011) and Fenske et al. (2013) and the references therein. Also, I further extend the selection of community-level factors with modelled covariates. Those are access to cities (i.e., access to markets and economic opportunities) and healthcare facilities. The covariates are provided at the grid-cell level and matched to the cluster location of an individual as the nearest grid-cell average. The selection is completed with a food security indicator provided by FEWS NET (2022) in the month preceding the survey. Table 2 provides the description, type and corresponding source of all included covariates.

Covariate (Type)	Source
Individual	
Age of the child in months (continuous)	Survey
Duration of breastfeeding in months (continuous)	Survey
Gender of the child (categorical: male, female)	Survey
Indicator for twin children (categorical: no, yes)	Survey
Position of the child in the birth order (categorical: 1, ..., 8+)	Survey
Body mass index of the mother (continuous)	Survey
Age of the mother in years (continuous)	Survey
Years of education of the mother (continuous)	Survey
Employment status of the mother (categorical: no, yes)	Survey
Religion of the mother (categorical)	Survey
Number of dead children (categorical: 0, 1, 2, 3+)	Survey
Household	
Number of household members (continuous)	Survey
Source of drinking water (categorical: unimproved, improved, piped)	Survey
Type of toilet facility (categorical: unimproved, improved)	Survey
Wealth index (categorical: poorest, poorer, middle, richer, richest)	Survey
Household has electricity supply (categorical: no, yes)	Survey
Household has a radio (categorical: no, yes)	Survey
Household has a television (categorical: no, yes)	Survey
Household has a refrigerator (categorical: no, yes)	Survey
Household has a bicycle (categorical: no, yes)	Survey
Household has a motorcycle (categorical: no, yes)	Survey
Household has a car (categorical: no, yes)	Survey
Community	
Administrative region (categorical: 23 with neighbourhood structure)	Survey
Place of residence (categorical: rural, urban)	Survey
Walking time to healthcare facilities (continuous)	Weiss et al. (2020)
Travel-time to cities (continuous)	Weiss et al. (2018)
Food security classification (categorical: 1-minimal, 2-stressed, 3-crisis)	FEWS NET (2022)

Table 2: *Childhood malnutrition in Madagascar: individual, household and community-level predictors for chronic childhood malnutrition.*

4.2 Modelling

Model specification

To estimate risk factors of stunting, I fit a logistic regression with the component-wise boosting described in section 3. Let y_i denote the classification status (0/1) of the child i and model the outcome with a Bernoulli model

$$y_i \sim \text{Bernoulli}(\mu_i), \quad i = 1, \dots, N, \tag{2}$$

where $i = 1, \dots, N$ are the children observed in the survey that are included in the analysis.[10] The relative probability of being stunted is estimated with a logistic additive predictor

$$\begin{aligned} \text{logit}(\mu_i) &= \eta_i \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \sum_{k=1}^{q} f_{\texttt{smooth}}(x_{ki}) + f_{\texttt{spatial}}(x_{\text{region[i]}}), \quad i = 1, \dots, N. \end{aligned} \tag{3}$$

The loss function is specified as the negative binomial log-likelihood. Categorical covariates are included as linear effects, and treatment coded. Smooth effects for continuous covariates are cubic P-splines with second-order differences and 20 inner knots. To allow for effect selection between linear and non-linear effects for continuous covariates, the smooth effects are included in the parametric and centred decomposition described before. Lastly, to account for the cumulative effect of unidentified covariates on a regional level, I include a spatial effect for the survey regions by adding a Markov random field effect induced by the first-order neighbourhood structure.[11]

In the following, I also briefly consider several alternative specifications of the additive predictor. First, as a sensible baseline, I consider a simple linear model. Second, to test a successive increase in implemented model complexity, an augmented linear model where all first-order interactions are added is considered. Third, I consider Equation 3 where each covariate additionally interacts with either (1) the gender of the child or (2) the urbanicity status. As a non-parametric approach, I also test boosted regression trees with a maximum depth of four, and with all other parameters left at their default values. The latter model should provide insights of whether higher-order interactions are present in the data.

Model evaluation and selection

As discussed in section 3, early stopping is applied to the algorithm to prevent over-fitting and obtain a model that generalises well on unseen data. For the following results, given the larger number of samples, I use the 2 or 3-way holdout method to fit the model on a training set, select the number of boosting iterations by where the minimum predictive risk is attained on the validation set, and use test-set performance to select between models. For the model evaluation part, I use approximately 70/20/10 splits and for the final results in the following

10. After the removal of data with missing characteristics a total of $N = 5722$ observations are included in the following analyses.

11. As discussed in section 2, the individual survey responses are assigned a sampling weight. In the literature, the use of survey weights in regression is debated. See for instance Winship and Radbill (1994) and Gelman (2007). The parameter estimates estimated by boosting are regularised and therefore likely not consistent concerning the true population value. Furthermore, it is not clear how to correct the weights when subsampling the survey data, since the sampling weights are post-stratified to correct for non-response. Though it is possible to include observation weights in this approach – all learners can be estimated by penalised weighted least squares – I omit those for the following analyses.

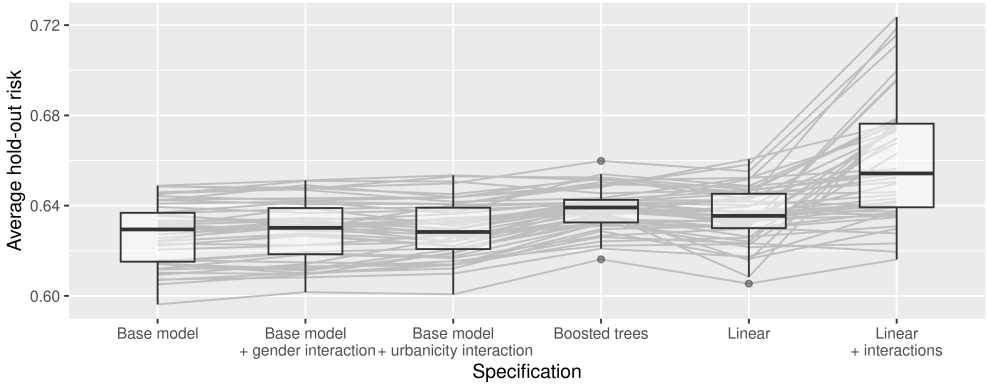

Figure 3: *Childhood malnutrition in Madagascar: comparison of different model specifications. Lower is better. Grey lines indicate the average test-set error (predictive risk). Empirical distributions over 50 replications.*

part 80/20. All folds are stratified by the survey strata to acknowledge the survey design. Throughout this section, I fix the hyper-parameter $\nu = 0.1$.

Figure 3 plots the average test-set risk of 50 replications. Seemingly, compared to the base model, the additional flexibility of the augmented models does not translate into improved generalisation performance. The linear model does not yield the same performance as the base specification, underscoring the necessity to capture non-linearities in associated risk factors. That coincides with previous literature (Fenske, Kneib, and Hothorn 2011; Kandala et al. 2009). But the differences are also not markedly different. Clearly, the linear model with all interactions seems unstable, and the hold-out risk varies strongly. Interestingly, boosting trees do not yield the same performance, but are more stable (lower variance) across replications. Altogether this provides evidence against higher-order interactions in the data. For the remainder of this analysis, I turn to the base specification as provided in Equation 3.

4.3 Results

For the following results, I show for each covariate the estimated partial effect from 50 replications of the train-validation split. This subsampling approach allows for an idea of the stability of the estimated parameters (Meinshausen and Bühlmann 2010). The inclusion frequency over the replications is provided in the supplementary material.

Beginning with the continuous covariates we find a negative partial effect of access to cities, indicating that higher distance to cities is associated with an increased risk of stunting (Figure 4). Access to healthcare is inversely U-shaped but often estimated to be zero. The linear effect was included only in 36% of the replications and the non-linear deviation in 68%. The effect estimates of the children's age are very stable, with very young ages having a protective factor, the partial effect increasing until the age of 20 months. The mother's age is inversely U-shaped with very young ages being a risk factor. For higher ages (>40 years) the estimation of the effect is much more variable. The partial effect of stunting is decreasing with the mother's BMI and increases with the number of household members, both estimated (mostly) to be linear, although the latter with much more variability in the slope. The mother's education shows an inversely U-shaped partial effect, with lower values accumulated around zero and a protective risk factor for a higher number of years in education.

Figure 5 shows the mean and standard deviation of the estimated discrete spatial effect. Eastern regions of the island show an excess risk towards stunting. Generally, the south (-

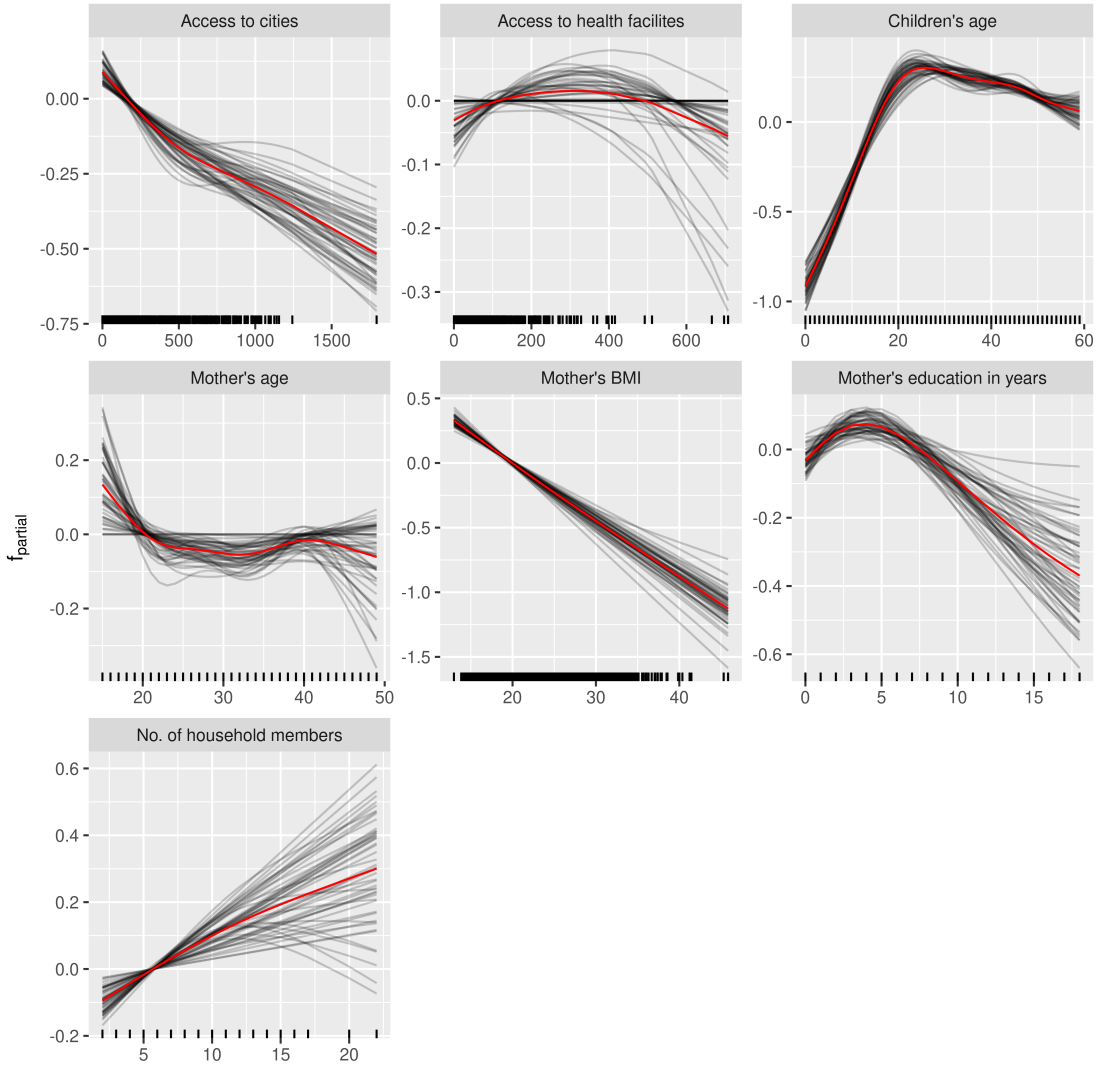

Figure 4: *Childhood malnutrition in Madagascar: partial effects of the continuous covariates. The red line indicates the pointwise average and the grey lines indicate the estimated effects from 50 replications.*

east) is considered the most afflicted area in terms of food security (FEWS NET 2022).

Lastly, a review of selected categorical effects. For most indicators associated with wealth, those are a protective factor (car, radio, television, etc.). Similarly, being in the richest quantile of the population (DHS wealth category). The children's gender (female vs. male) has a negative partial effect, i.e., decreased risk of being stunted, which is in line with the observation that the prevalence of stunting is higher among boys than girls (UNICEF 2013, p. 10). Higher birth order is associated with a higher log-odds ratio of being stunted. Note, the effect of the birth order can not be interpreted as a within-household effect because the anthropometric measures in the DHS sample are age-censored.[12]

In the supplementary material, I present the empirical distributions of the 'importance' of each component of the additive predictor. The highest contribution to the reduced risk is the

12. See Spears, Coffey, and Behrman 2022 for further discussion on this topic.

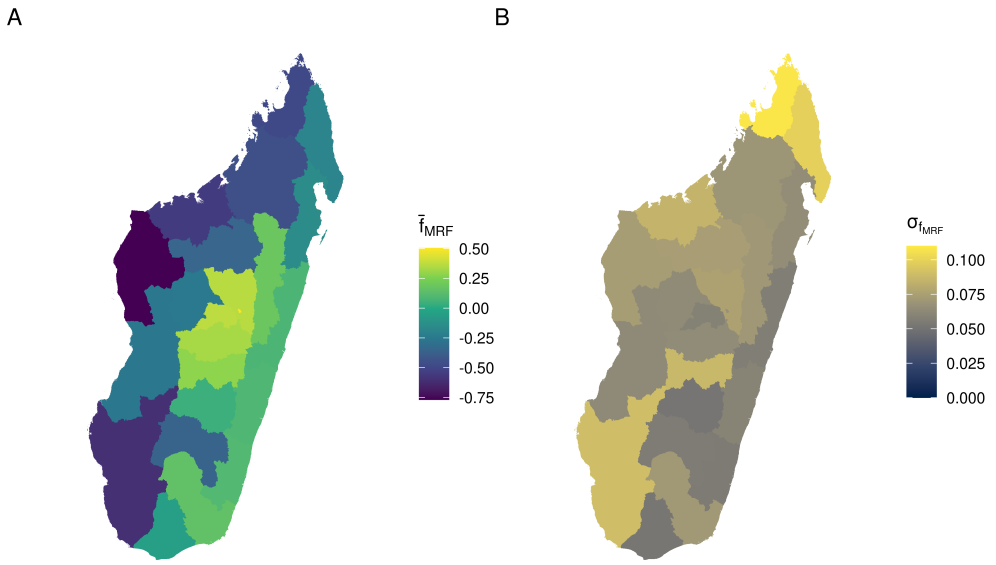

Figure 5: *Childhood malnutrition in Madagascar: estimated discrete spatial effect. Panel A: mean effect over 50 replications. Panel B: standard deviation of the estimated effects.*

spatial effect indicating the need for an improved understanding of regional-level risk factors of chronic childhood malnutrition. Other variables that contributed the most to the reduction were the children's age, the mother's BMI, access to cities, and the children's gender.

The component-wise boosting approach allows for important insights into the model, and, in this setting, allows for inference about the functional form of an effect. For instance, the child's age as a risk factor is estimated non-linear across the range. Furthermore, stability selection can be employed to obtain type-1 error controls for variable selection (Meinshausen and Bühlmann 2010; Shah and Samworth 2013). Thus, the preliminary results for the interested reader are presented in the enclosed code repository.

5 Geographic malaria risk in Mali

In the second case study, we turn to cluster-level estimates of malaria prevalence.[13] Specifically, the objective of this case study is to identify environmental correlates of local malaria prevalence for children below the age of five and predict the estimated risk at a high spatial resolution. A similar study was presented in Giardina, Sogoba, and Vounatsou (2016), where the authors propose a Bayesian variable selection based on Dirichlet priors and a non-stationary Gaussian spatial process to model residual spatial variation at the transition of ecological zones. In this case study, I employ the component-wise boosting approach to obtain effect selection.

5.1 Introduction

Malaria is an infectious disease for humans, caused by the *Plasmodium spp.* when transmitted by a bite of an infected female *Anopheles* mosquito. Most common are the *Plasmodium falciparum* and *Plasmodium vivax* species, the two variants account for the majority of cases (Phillips et al. 2017). Untreated, severe malaria can be fatal, with the highest burden among

13. Note, the term *prevalence* refers to the empirical proportion of the population experiencing the condition while the term *risk* refers to the analogue of the hypothetical infinite population (Fuglstad, Li, and Wakefield 2022).

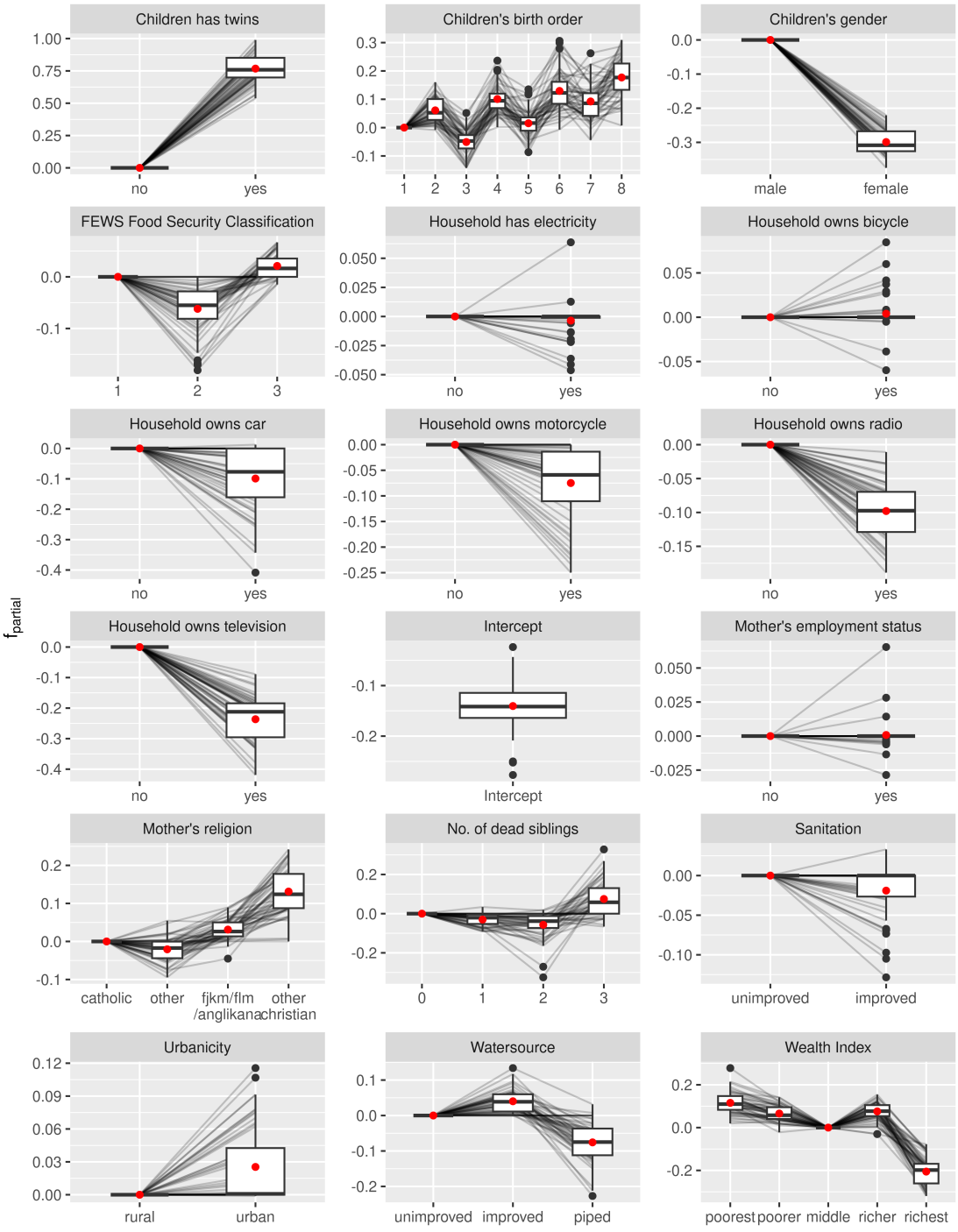

Figure 6: *Childhood malnutrition in Madagascar: estimated categorical coefficients. Grey lines indicate the 50 replications, and the red point indicates the pointwise average of the estimated coefficients.*

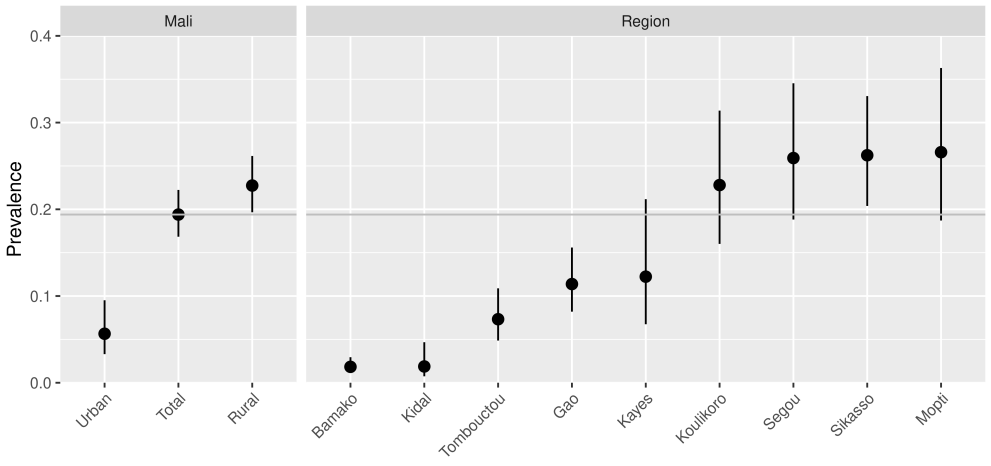

Figure 7: *Geographic malaria risk in Mali: design-based estimates of malaria prevalence in children 6-59 months. Error bars indicate 95% confidence intervals. Data from the Mali 2021 MIS.*

young children below the age of five. Cases and deaths have declined since 2000, albeit slowly, with a recent up-tick likely due to the COVID-19 pandemic. For 2020, the latest data at the time of writing, the WHO estimates 241 million cases and 627'000 deaths (World Health Organization 2021). Globally, high-transmission countries in sub-Saharan Africa accounted for a large share of malaria cases in the 85 countries where malaria is endemic: 29 countries accounted for 96% of all the cases and deaths. Of all child deaths below the age of five, malaria is estimated to account for 7.8%. The country in this analysis, Mali, whilst only with a population of about 22 million, is estimated to represent 3% of the *global* caseload. Undoubtedly, the health and economic burden on households and countries is substantive and long-lasting. In country-level studies, micro-evidence from early eradication campaigns of malaria suggests that exposure early in life does shape (long-term) economic outcomes (Bleakley 2010; Cutler et al. 2010; Lucas 2010; Hong 2011).

Household surveys routinely collect information on malaria-related indicators, such as the ownership and use of insecticide-treated mosquito nets. The Malaria Indicator Surveys (MIS) from the DHS program are a format analogous to the standard DHS surveys but tailored to collect malaria-related information. In selected countries, additional samples are taken from populations at risk to establish the infection of the individual. In particular, those surveys are conducted during the high malaria transmission season of the respective country. The following study is based on the Mali 2021 Malaria Indicator Survey (MIS) where children between 6-59 months were tested for malaria with Rapid Diagnostic Tests (RDTs).[14]

Mali 2021 Malaria Indicator Survey

The Mali 2021 Malaria Indicator Survey (MIS) was designed to provide estimates of key malaria-related indicators at the national level, for urban and rural populations, and each of the eight

14. Some household surveys use Polymerase Chain Reactions (PCR) tests identify malaria infection rather than Rapid Diagnostic Tests (RDTs). PCR tests generally have a higher sensitivity than RDT tests at the expense of higher costs. Therefore, prevalence estimates from both testing methods should not be compared directly. Florey (2014) assess differences in estimates in DHS surveys where both test methods were used, concluding that RDTs may be sufficient to identify populations with higher risk but should not be used to assess the effectiveness of health interventions.

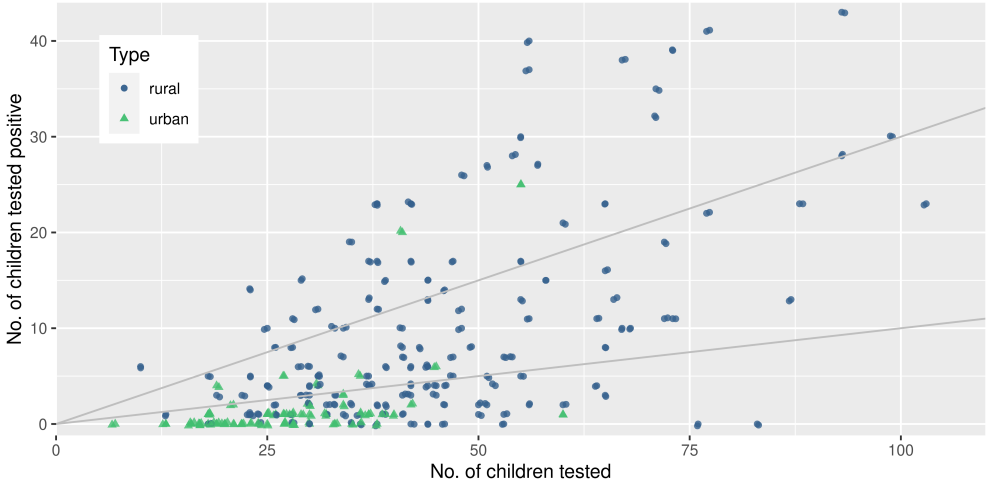

Figure 8: *Geographic malaria risk in Mali: number of children tested positive against cluster size. Grey lines indicate a raw prevalence of 30% and 10%, upper and lower lines, respectively. A small amount of noise was added to the values to enhance visualisation. Data from the Mali 2021 MIS.*

administrative regions plus the capital Bamako. The latter is exclusively urban, yielding 17 strata (regions crossed with urbanicity). A total of 216 clusters were drawn from a sampling frame based on the 2009 population census. If the number of enumerated households per cluster surpassed 300, the area was partitioned and only one partition was selected for complete enumeration. In each, 26 households were selected randomly and all women (15-49) usually living in the selected households and present the night before the interview were eligible for the questionnaire. Additionally, all children aged 6-59 months were eligible for a rapid diagnostic test on malaria infection and anaemia (Institut National de la Statistique (INSTAT), Programme National de Lutte contre le Paludisme (PNLP), and The DHS Program 2022).

Figure 9 shows the cluster locations of the Mali 2021 MIS. Northern Mali is characterised by a sparse and highly rural population. The realised survey cluster locations are therefore located predominantly in the south of the country. The design-based estimates and the corresponding confidence intervals of malaria prevalence in children 6-59 months are shown in Figure 7. Strong subnational differences are discernible, both between regions as well as urban and rural designated areas. The urban-rural divide is particularly evident in the raw cluster-level prevalences. Figure 8 plots the count of children who tested positive against the number of children in each cluster. Almost all urban clusters are below the 10% line, and the rural clusters show much more variability in the raw estimates. Furthermore, even though the number of households sampled in each cluster is the same (26), the effective number of children tested per cluster varies greatly. In particular, any modelling strategy should account for different realised sample sizes at each cluster location.

Environmental variables

To inform local risk of malaria, the analysis includes environmental and climatic predictors that have been associated with malaria risk in previous work (see Weiss et al. 2015; Millar et al. 2018; Weiss et al. 2019; Mohammed et al. 2022, and the references therein). Transmission risk is dependent on the species distribution of the *Anopheles* mosquitoes, which are sensitive to climatic factors. The environmental variables for this case study are precipitation (annual aggregate) (Funk et al. 2015), elevation (static) (Jarvis et al. 2008), land surface temperature (annual

mean, day and night) (Didan, MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V061). Additionally, I include two vegetation indices, the Enhanced Vegetation Index (EVI) and Normalised Difference Vegetation Index (NDVI) (Wan, Hook, and Hulley, MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V061).

As discussed in Dong and Wakefield (2021a) and Paige et al. (2022), model-based approaches to small area estimation should include the stratification of the survey design, as omission may produce biased estimates. Therefore, to distinguish between urban and rural at unseen locations, I include an urban-rural indicator based on the degree of urbanisation from the Global Human Settlement Layers (GHSL) project (Schiavina, Melchiorri, and Pesaresi, GHS-SMOD R2022A - GHS Settlement Layers, Application of the Degree of Urbanisation Methodology (Stage I) to GHS-POP R2022A and GHS-BUILT-S R2022A, Multitemporal (1975-2030)). This variable is constructed based on estimated population counts and remotely sensed built-up grids.[15] Lastly, I include population counts (Schiavina, Freire, and MacManus, GHS-POP R2022A - GHS Population Grid Multitemporal (1975-2030)) and the climate classification by Köppen-Geiger (Beck et al. 2018). Mali is covered by three distinct climate regions, see Figure 9.

Matching georeferenced data

For each cluster, the locations of the interviewed households are recorded and the centroid is taken as cluster-level information. To ensure the privacy of the respondents, a randomisation procedure is applied to the coordinates, where urban locations are randomly displaced up to 2km and rural up to 5km with an additional 1% of the observations up to 10km (Burgert et al. 2013). This naturally introduces measurement error and mismatching when integrated with other data sources based on the geographic location.[16] Additional information on the data and the matching procedure is given in the supplementary material.

5.2 Modelling

Model specification

From the survey micro-data, one obtains for each cluster c a count of children n_c that were tested, and a count y_c that tested positive. A natural way to model such data is with a Binomial likelihood in a generalised additive model framework. To estimate the model, I use the component-wise boosting approach described herein. Thus, let

$$y_c \sim \text{Binomial}(n_c, \mu_c), \quad c = 1, \ldots, N.$$

15. An alternative approach was proposed in Dong and Wakefield (2021a) and Paige et al. (2022), where the authors construct an urbanicity variable with information about the sampling frame published in the survey reports. The surveys publish the percentage of urban population for each region and the country from the used primary sampling frame, and with population density layers, a threshold to obtain an urban-rural indicator per grid cell can be inferred. In the case of Mali, this resulted in very few urban locations. Ultimately, both approaches are likely sensitive to the chosen population density layer, which has been shown to induce substantial differences in applications (see, for example, Hierink et al. 2022). For further discussion on the modelling of urban-rural fractions, and a comparison of alternative approaches, see Wu and Wakefield (2022).

16. In and of itself, there is not much to be done to counter this induced error. Wu and Wakefield (2022) discuss a Bayesian approach to account for the displacement of urban regions. Michler et al. (2022) assesses the impact of different spatial anonymisation techniques for a similar series of household surveys from the World Bank. Comparing estimates of measures of agricultural productivity for rural and agricultural households, the authors conclude the anonymisation method introduces limited error, but care should be taken in the selection of remote sensing products. In this case study, I take the cluster locations as provided.

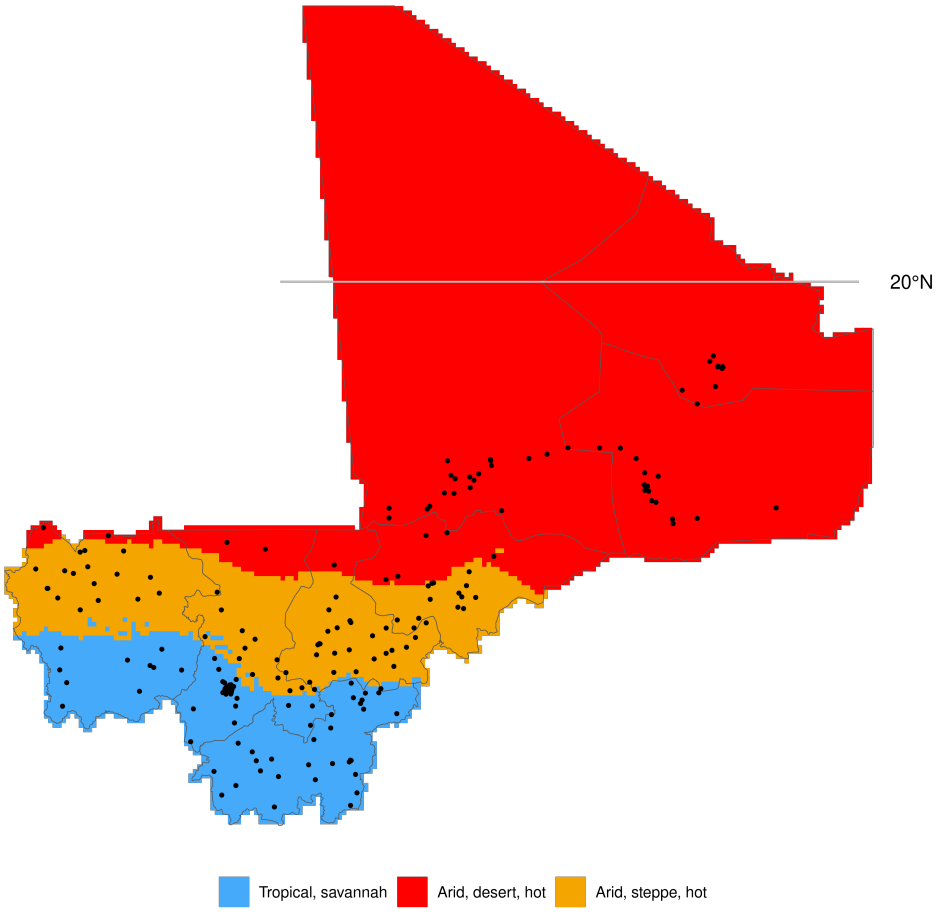

Figure 9: *Geographic malaria risk in Mali: climate classification by Köppen Geiger and survey cluster locations from the Mali 2021 MIS. Regions north of the* 20° *degree are omitted due to data sparsity.*

The relative probability of occurrence μ_c is estimated with a linear predictor

$$\begin{aligned} \text{logit}(\mu_c) &= \eta_c \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{jc} + \sum_{k=1}^{q} f_{\texttt{smooth}}(x_{kc}) + f_{\texttt{spatial}}(x_{\text{lon},c}, x_{\text{lat},c}), \quad c = 1, \ldots, n. \end{aligned} \tag{4}$$

The loss function is taken to be the negative log-likelihood of the binomial distribution. In this base model, I include linear effects for the categorical variable (urbanicity) and smooth effects using cubic P-splines with second-order differences and 20 inner knots. To ensure unbiased effect selection these are included in the parametric and centred decomposition discussed before. To account for spatial effects, I include a bivariate smooth with a similar decomposition. For details see Kneib, Hothorn, and Tutz (2009).

As discussed in Dong and Wakefield (2021a), count data aggregated to cluster-level often exhibit more variability than be accounted for in a binomial response distribution. The data is said to be *overdispersed*. To accommodate possible within-cluster variation, the authors suggest

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

Table 3: *Geographic malaria risk in Mali: model validation based on 10-fold cross-validation stratified by survey region. Values rounded rounded to the nearest hundredth.*

a beta-binomial likelihood

$$y_c \sim BB(n_c, \mu_c, \sigma_c), \quad c = 1, \ldots, N.$$

The beta-binomial distribution can be thought of first drawing a probability from the beta distribution $p_c \sim \text{Beta}(\mu_c, \sigma_c)$ and then the cluster-level count $y_c \sim \text{Binomial}(n_c, p_c)$. Hence, it accommodates greater variability in the data than the binomial distribution. For $\sigma \to 0$ the limiting distribution is the binomial distribution, hence greater values of σ_c correspond to a higher degree of overdispersion. Additional information on the response distributions is included in the supplementary material.

The beta-binomial distribution can be modelled in a boosting framework by moving to the distributional regression approach, which allows additive predictors for both the mean and the degree of overdispersion (i.e., the location and shape of the distribution). Thus, let $\eta_{\mu,c}$ be the predictor defined in Equation 4 and

$$\begin{aligned} \log(\sigma_c) &= \eta_{\sigma,c} \\ &= \beta_0 + \beta_1 x_{\texttt{urban}}, \quad c = 1, \ldots, N. \end{aligned} \tag{5}$$

Although it would be simple to add additional terms to model the degree of overdispersion of the conditional distribution, an intercept and urbanicity keeps model complexity low and are motivated by the urban-rural divide.

Besides the comparison of the two different choices of the data model, I consider alternative specifications of the additive predictor for the beta-binomial model. First, a simple linear model as a baseline. Two versions of the linear model, one augmented with a bivariate spatial smooth, and a second specification of the linear model augmented with the spatial effect and all linear covariates interacted with the climate classification. I also consider a model with spatially-varying coefficients, where each continuous covariate enters the additive predictor linearly, as a modifier of a bivariate smooth. Finally, as an alternative to the pre-specified semi-parametric effects, I also evaluate the performance of regression trees as base learners with a maximum interaction depth of 4 and with all other parameters left at their default values.[17]

I employ a bootstrap approach to quantify variability in the estimated effects. To make ideal use of the data I draw bootstrap samples from the individual-level data stratified by cluster. The model is refitted and the estimates are compared to the 'main' model, the model fitted on the original data set.

Model evaluation and selection

Given the lower sample size ($N = 216$), the holdout method is not reasonable and I select the number of boosting iterations based on where the minimum of the cross-validated risk is

17. It would be interesting to see if the fully nonparametric specification could be further improved by the use of oblique coordinates (Møller et al. 2020).

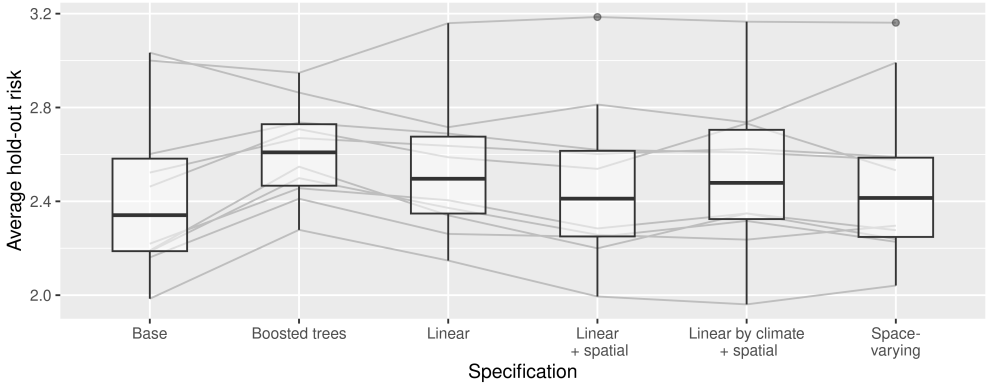

Figure 10: *Geographic malaria risk in Mali: comparison of different model specifications. Lower is better. Grey lines indicate the average hold-out risk of 10-fold cross-validation stratified by survey regions.*

attained. Cross-validation estimates are based on $K = 10$ and folds are stratified by region, as the number of clusters per strata does not allow for stratification by survey design strata. The step-length ν is fixed at 0.1.

To compare the different model specifications proposed above, I employ nested cross-validation. The outer loop consists of 10-fold cross-validation, and the inner loop for early stopping is likewise based on 10-fold cross-validation both stratified by survey regions. First, I validate the data model for the prevalence data. Table 3 displays three common regression evaluation metrics for the point predictions, that is Bias, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). To assess the quality of the prediction intervals, I also include the average coverage of the $100(1-\alpha)\%$ prediction interval (PI) based on the conditional distribution. The $100(1-\alpha)\%$ PI is defined as

$$PI_{1-\alpha}(x) = \left[Q_{\alpha/2}(x), Q_{1-\alpha/2}(x)\right]$$

where Q_α is the α-quantile of the conditional distribution. Bias, MAE and RMSE are calculated with respect to the target of inference $\hat{\mu_c}$ and the raw prevalence at cluster-level y_c/n_c, and the coverage of the PI is calculated at the observed counts. The presented results are the average values over the outer hold-out folds.

Concerning the first three metrics, the three models show only minor differences. All tend to slightly overestimate the observed rate (bias). Given that the comparison is only based on 10-folds with a comparatively small sample size of 216 clusters, the differences should be considered marginal. However, that is not the case for the coverage of the prediction intervals. Here the binomial model does markedly worse than the model with the beta-binomial likelihood. Where the 90% PIs only cover on average around 70 % of the observations, the beta-binomial achieves a coverage above 90%. Clearly, the prevalence data show more variability than can be accommodated in the binomial distribution, indicating the beta-binomial model has a superior fit. The following results, therefore, are based on the data models with the beta-binomial model.

Next, Figure 10 provides the comparison of the hold-out risk of the alternative model specifications. Clearly, likely due to the data limitations, the variability in the generalisation performance is high. Interestingly, the plain linear model does provide a tenable baseline, and the addition of a spatial effect tends to improve holdout risk. Augmenting the model with interactions by the climatological zone introduces instability. Interestingly, boosted trees fare the

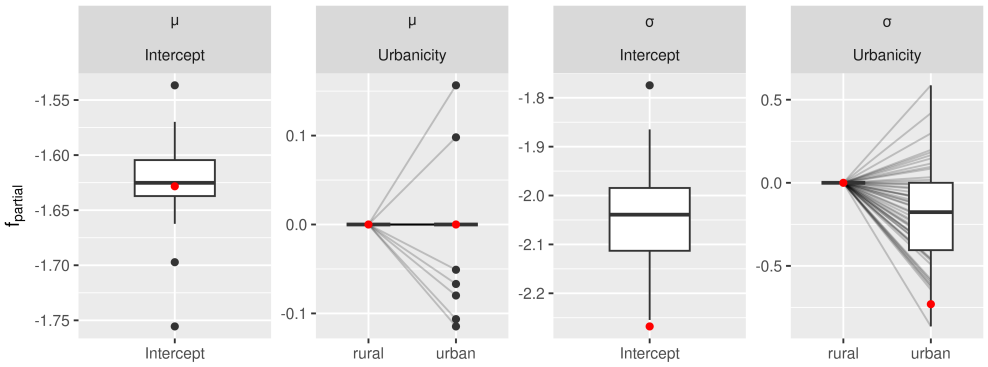

Figure 11: *Geographic malaria risk in Mali: estimated coefficients of the intercept and urbanicity covariate. The red points indicate the estimated effect of the main model and boxplots the empirical distribution of the estimated coefficients over the 50 bootstrap samples.*

worst in this comparison, but with less variability. Overall, the base specification as in Equation 4 provides the best generalisation performance. Based on this model, I will discuss the estimated factors and grid-cell predictions below.

5.3 Results

Figure 11 shows the coefficients of the categorical covariates included in the model. Note that there – as for all the following partial effect estimates – the estimates are provided on the scale of the link function. Thus, for the conditional mean the partial effects are on the log-odds ratio of μ, and for the conditional scale of the beta-binomial distribution on the log scale of σ.[18]

For μ, urbanicity is estimated to be zero in the main model, with a few deviations in the bootstrap samples. For σ, the urbanicity coefficient is estimated to be negative, with the third quartile below zero.

Figure 12 plots the partial effects of the continuous covariates. The main model is indicated by the red line, and the models replicated on 50 bootstrap samples are shown in the background. For elevation, the estimated effect is linear, higher altitudes correspond to a decreased malaria risk. The Enhanced Vegetation Index (EVI) is estimated clearly in a U-shaped form, in particular, an increased risk towards the upper end of the scale. This is in line with the interpretation of the index, where higher values indicate dense vegetation. Likewise, the Normalised Vegetation Index (NDVI) shows an increased risk for higher values and can be interpreted similarly. The effect of land surface temperature during the day is mostly estimated to be zero, as in the main model, with only some bootstrap replicates showing different results. For the temperature at night, the effect is estimated to be negatively associated with malaria risk in the lower range, peaking between 20–22 degrees and declining for higher values. For values toward 26°, it shows an up-tick, but not indistinguishable from zero. The (log) population counts are relatively flat for the first half of the support then drastically decrease, i.e., higher population counts are associated with a lower risk of malaria. Since a higher population equate with urban areas, this is plausible. The annual aggregated precipitation shows an increasing effect up until 600, then decrease. For values beyond 1200, the estimation is supported by few data points (as indicated by the rug plot), correspondingly, the estimates show

18. Note, the effects of ecological covariates should not be interpreted on an individual level, as one risks to commit an ecological fallacy (Piantadosi, Byar, and Green 1988).

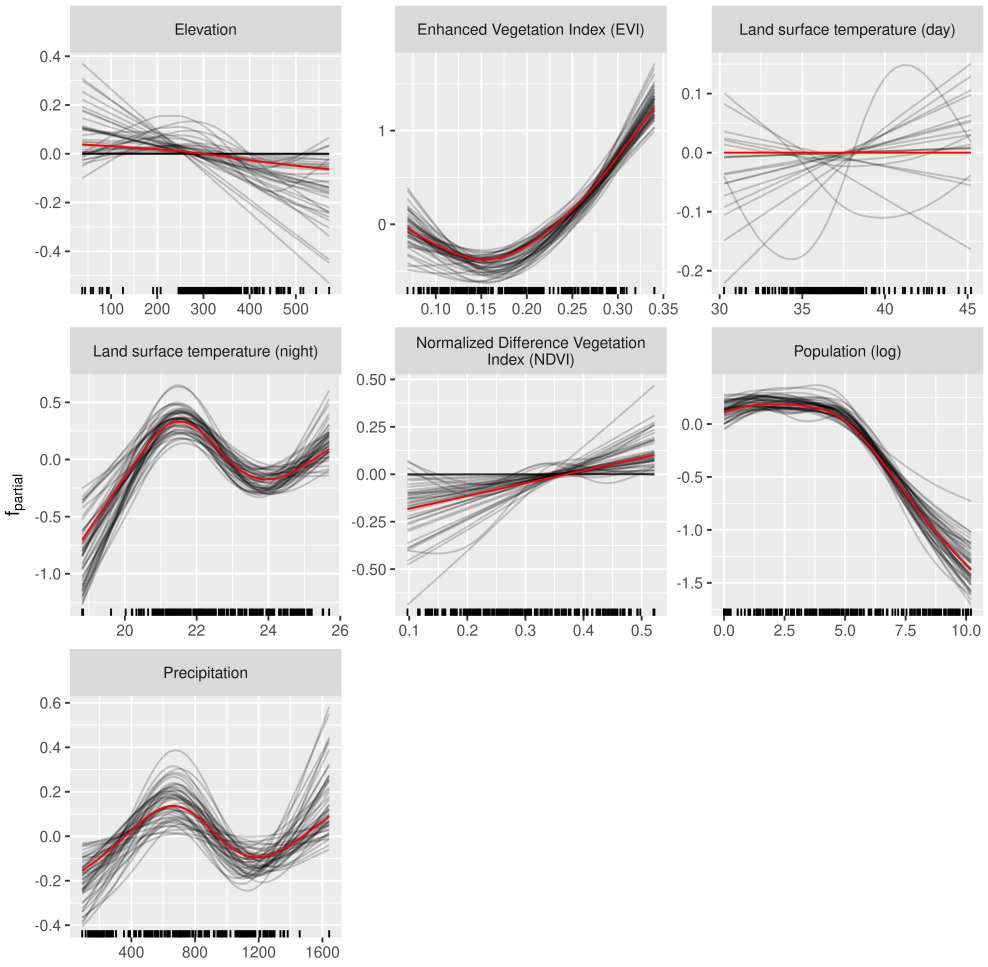

Figure 12: *Geographic malaria risk in Mali: partial effects of the continuous covariates. The red line indicates the main model and the grey lines indicate the estimated effects from 50 bootstrap samples.*

much more variability. Again, the effect selection properties of the component-wise boosting algorithm yield informative selections about the functional form of a given covariate.

The partial effect of the spatial smooth is plotted in Figure 13. For the country's west and north-eastern part a negative partial effect can be identified, in the mid-eastern part, towards the border with Burkina Faso, a positive partial effect. The standard error of the spatial effects is shown in Figure 14. The estimated effect is estimated stable in the southern part (where the majority of cluster locations are located) and shows a higher standard error in the northern parts. Given the lack of data north of the 18°, only the grid-cells south of the 20° are plotted (c.f. the observed cluster locations in Figure 9).

Based on the model, it is feasible to construct maps of the predicted risk of malaria throughout the country on a fine-scaled grid. Figure 15 plots the predicted values $\hat{\mu}$. I provide additional details on the construction of the grid in the supplementary material. The north, dominated by desert or semi-desert lands is estimated at very low risk. The lower half shows higher mean prevalences, particularly on the border towards Guinea (southwest). Furthermore, distinct geographical features such as the capital Bamako in the southwest and parts of the Niger river in the east can be distinguished from the estimated near zero risk. The predicted risk for each

Figure 13: *Geographic malaria risk in Mali: estimated partial effect of the bivariate P-splines.*

Figure 14: *Geographic malaria risk in Mali: standard error of the estimated partial effect based on 50 bootstrap samples.*

Figure 15: *Geographic malaria risk in Mali: predicted risk of malaria for children below the age of five.*

location can be accessed in the enclosed code repository.

Maps of predicted risk at such high resolution may mask high uncertainty in the predictions (Dong and Wakefield 2021a). Therefore, I include maps of the 10% and 90% quantile as well as the standard error of the predictions over the bootstrap samples. The figures can be found in the supplementary material.

Construction of subnational prevalence estimates

The predicted grid-level risk estimates can be aggregated to obtain an estimate for the subnational prevalence at the admin level 1 or 2. Assuming a fixed proportion of under five-year-olds to the population throughout the country, the estimate is obtained by the average risk, weighted by the population over the area of interest. To assess how well these estimates compare to the design-based estimates, Figure 16 shows the design-based and model-based estimates side-by-side. At the mean, the model-based estimates track closely the design-based estimates for five of the nine regions and recover the design-based estimates. The confidence intervals are the quantiles from the bootstrapped models. Certainly, the confidence intervals of the model-based approach are over-confident and likely provide poor coverage of the true prevalence.

The differences between the estimates for the regions Koulikoro, Mopti, Segoú and Sikasso, nevertheless, demand further investigation. Two possible sources of error are the urban-rural indicator (specifically, if the baseline prevalence in each is substantially different) and the gridded population maps.[19] For example, for Koulikoro, the share of the urban population provided

19. Rather than the GHSL gridded population map employed in this analysis one could have used WorldPop (Bondarenko et al. 2020). However, a complete comparison is beyond the scope of this manuscript.

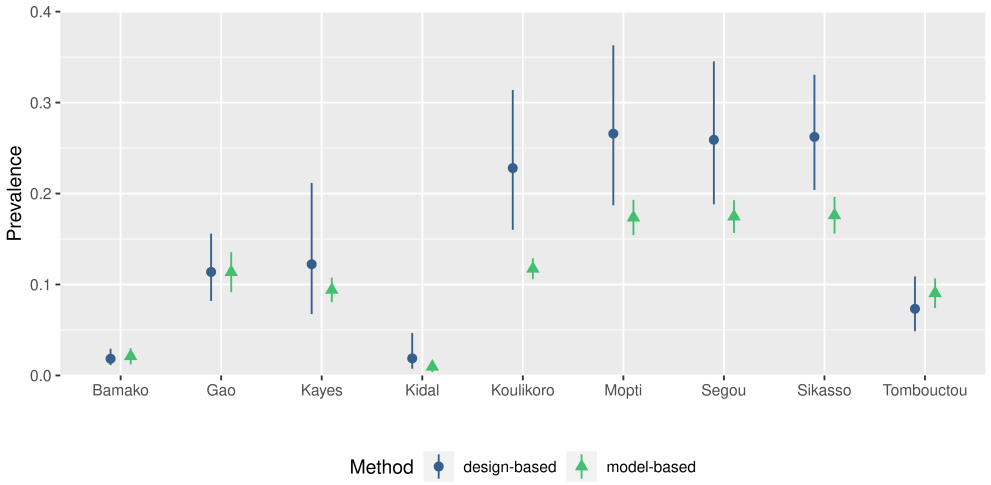

Figure 16: *Geographic malaria risk in Mali: comparison of design-based estimates and model-based estimates for admin 1. Error bands are 95% confidence intervals, for the model-based approach based on bootstrap quantiles.*

in the survey report is 5.5% while the estimate based on the gridded population maps is 41.7%. Though the population census is from 2009, and it is not inconceivable that a fast urbanization process leads to strong differences in urban-rural fractions, the hypothesis of underestimating the rural population is consistent with the strong underestimation of the malaria prevalence in those regions. The dependence of such approaches on gridded population maps warrants further research. Therefore, the estimates for admin 2 should be interpreted with caution and can be accessed in the enclosed code repository.

Comparison to alternative approaches

In a similar malaria mapping study, Bhatt et al. (2017) propose a stacked ensemble approach to improve prediction accuracy. Briefly, the fitted values of different state-of-the-art algorithms (gradient-boosted trees, random forests, etc.) are weighted and included in a geo-statistical model which accommodates spatial variation unaccounted for in the first level. The authors find the approach to be highly competitive in predictive tasks such as malaria prevalence estimation. The specifics can be found in Bhatt et al. (2017).

The approach proposed herein can account for non-linear and spatial effects in a single framework without the need for additional calibration of intermediate levels. Furthermore, the data is binomial, provided per cluster c as the number of children y_c that tested positive out of a total number of tested children n_c. The authors compute the cluster-level prevalences $\hat{p}_c = y_c/n_c$ and logit transform the estimates to obtain values with continuous support. This allows the use of common software packages for gradient-boosting trees and random forests, however, it does not respect the binomial nature of the data. In the proposed approach, the binomial distribution can be modelled directly as response distribution and allow inference about underlying risk. In addition, we obtain a semi-parametric model that can uncover important ecological correlations.

Similarly, Weiss et al. (2015) study environmental correlates of malaria. The authors compile a set of more than 50 million (!) possible predictors by an array of transformations and combinatorial interactions. The set of possible terms was successively reduced using information criteria, to identify relevant risk predictors and the corresponding functional form.

Though the authors use binomial GLMs, variable selection based on information criteria is generally considered as unstable (see, for example, the discussion in Mayr et al. 2012, Section 2.2). Employing a component-wise boosting approach such as described herein avoids such pitfalls and allows for competitive predictive performance and effect selection.

6 Discussion

In this manuscript, I demonstrated the framework of component-wise boosting to identify risk factors of two common prevalent health conditions in sub-Saharan Africa. Boosting, a method stemming originally from the machine learning literature for classification has been continuously developed in the last decade to address a variety of common statistical tasks and provide inference in moderate to high-dimensional regression settings. I have reviewed those in section 3. In particular, boosting compares favourably with other approaches if the variable selection is desired or necessary, either because of the dimensionality induced by covariates or because the understanding of the functional form of effects is crucial.

In the case study on chronic childhood malnutrition, the analysis underscored the necessity to consider the non-linear effects of potential risk factors, such as the child's age. Studying the environmental correlates of malaria reiterated this point, as the specification outperformed alternative definitions and even boosted trees, a method often employed as the default. In contrast to approaches presented in the literature on disease prediction, the component-wise boosting approach yields an interpretable model and allows for response distributions appropriate for the prevalence data at hand. It would be an interesting task to extend the malaria risk analysis to evaluate the predictive performance on cross-country data sets with multiple survey rounds and an extended set of explanatory variables with comparison to the distributional random forests approach (Schlosser et al. 2019).

Recent research has studied the merits of including 'alternative' data sources in the estimation of poverty or other local development indicators.[20] For instance, Steele et al. (2017) explored embedding mobile operator call detail records (CDRs) for the small area estimation of poverty. Those data are usually provided in a collection of statistics that can be derived from call data and aggregated at the unit of interest. But it is not clear which statistics are informative for the task. In such situations, one obtains tasks that boosting tends to handle very well. While the default approach in predictive modelling usually encompasses nonparametric decision trees, simpler additive models may perform competitively. This has also been shown elsewhere (Kapoor and Narayanan 2022). Besides, interpretable models are often desired if policy decisions are based on them (Rahman and Keseru 2021).

Even so, by employing boosting, one trades off conventional statistical concepts such as the quantification of uncertainty in the estimates by improved generalisation performance and intrinsic effect selection. If this trade-off is useful, will depend on a case-by-case basis. In these case studies, I used subsampling and bootstrap replications to assess variability in the estimates. While these approaches are straightforward to implement, they are computationally very intensive and do not necessarily guarantee coverage of the intervals derived, making statistical inference difficult. If a proper uncertainty quantification of model parameters is required, Bayesian hierarchical models such as those commonly employed in the geo-statistical framework are likely advantageous. Here, a clear conceptualisation of the objective of each research design is imperative, also to maintain trust in decisions derived from complex statistical approaches Broderick et al. (2021).

20. See, for example, Jean et al. (2016), Pape and Wollburg (2019), E. Aiken et al. (2022), Ziulu et al. (2022), and E. L. Aiken et al. (2023).

The literature on monitoring population health statistics has developed incredibly fast in recent years, providing models for the most followed statistics. But there still seems to be a wider gap in the treatment of data-sparse and conflict settings. The absence of official data has motivated much work in the inclusion of remotely sensed data. Obviously, there is a clear limit on what can be inferred from such data to understand population health statistics. Since the association between predictors and outcomes is likely noisy, much more promising is the merge of multiple data sources, such as household-, high-frequency phone surveys, and local assessments. There, careful modelling of the bias is important to draw correct inferences,[21] and provides many important avenues for research.

There is a broader case to be made for predictive modelling for development and epidemiological applications (e.g., Greenough and Nelson 2019). The component-wise boosting approach discussed herein can make some important contributions in this direction by, for example, employing appropriate response distributions, improving interpretability through model-building, and estimating non-linear and spatial effects of interest, particularly where generalisation performance is crucial. The appropriateness of statistical learning methods, however, will depend on the inferential objective of the research design.

Acknowledgements

I would like to express my gratitude to my supervisor, Cornelius Fritz, who guided me throughout this project. I would also like to thank Yanchun Zhang for enabling my internship at the Human Development Report Office, United Nations Development Programme, in the spring of 2022. It was there where I first learned about the literature on the statistics behind monitoring socio-economic development that has inspired much of the work done in this manuscript. Camila and Paul, I would like to thank you for many thoughtful discussions on political science, statistics, and the philosophy of science, I appreciated them very much at this university. Finally, Yaiza, for keeping me sane throughout this period, thank you very much.

21. For a combination of survey data and 'novel data' see, for instance, Alexander, Polimis, and Zagheni (2022) and Hsiao et al. (2023).

References

Aheto, Justice Moses K., Benjamin M. Taylor, Thomas J. Keegan, and Peter J. Diggle. 2017. "Modelling and Forecasting Spatio-Temporal Variation in the Risk of Chronic Malnutrition among under-Five Children in Ghana." *Spatial and Spatio-temporal Epidemiology* 21:37–46. https://doi.org/10.1016/j.sste.2017.02.003.

Aiken, Emily, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E. Blumenstock. 2022. "Machine Learning and Phone Data Can Improve Targeting of Humanitarian Aid." *Nature* 603 (7903): 864–870. https://doi.org/10.1038/s41586-022-04484-9.

Aiken, Emily L., Guadalupe Bedoya, Joshua E. Blumenstock, and Aidan Coville. 2023. "Program Targeting with Machine Learning and Mobile Phone Data: Evidence from an Anti-Poverty Intervention in Afghanistan." *Journal of Development Economics* 161:103016. https://doi.org/10.1016/j.jdeveco.2022.103016.

Alexander, Monica, and Leontine Alkema. 2018. "Global Estimation of Neonatal Mortality Using a Bayesian Hierarchical Splines Regression Model." *Demographic Research* 38:335–372. https://doi.org/10.4054/DemRes.2018.38.15.

Alexander, Monica, Kivan Polimis, and Emilio Zagheni. 2022. "Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States." *Population Research and Policy Review* 41 (1): 1–28. https://doi.org/10.1007/s11113-020-09599-3.

Alkema, Leontine, and Jin Rou New. 2014. "Global Estimation of Child Mortality Using a Bayesian B-spline Bias-reduction Model." *The Annals of Applied Statistics* 8 (4): 2122–2149. https://doi.org/10.1214/14-AOAS768.

Bates, Stephen, Trevor Hastie, and Robert Tibshirani. 2022. *Cross-Validation: What Does It Estimate and How Well Does It Do It?,* arXiv:2104.00673. https://doi.org/10.48550/arXiv.2104.00673.

Beck, Hylke E., Niklaus E. Zimmermann, Tim R. McVicar, Noemi Vergopolan, Alexis Berg, and Eric F. Wood. 2018. "Present and Future Köppen-Geiger Climate Classification Maps at 1-Km Resolution." *Scientific Data* 5 (1): 180214. https://doi.org/10.1038/sdata.2018.214.

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.

Bischl, Bernd, Olaf Mersmann, Heike Trautmann, and Claus Weihs. 2012. "Resampling Methods for Meta-Model Validation with Recommendations for Evolutionary Computation." *Evolutionary computation* 20 (2): 249–275. https://doi.org/10.1162/EVCO_a_00069.

Bleakley, Hoyt. 2010. "Malaria Eradication in the Americas: A Retrospective Analysis of Childhood Exposure." *American Economic Journal: Applied Economics* 2 (2): 1–45. https://doi.org/10.1257/app.2.2.1.

Bondarenko, Maksym, David Kerr, Alessandro Sorichetta, Andrew Tatem, and WorldPop. 2020. *Census/Projection-Disaggregated Gridded Population Datasets, Adjusted to Match the Corresponding UNPD 2020 Estimates, for 51 Countries across Sub-Saharan Africa Using Building Footprints.*

Broderick, Tamara, Andrew Gelman, Rachael Meager, Anna L. Smith, and Tian Zheng. 2021. *Toward a Taxonomy of Trust for Probabilistic Machine Learning,* arXiv:2112.03270. https://doi.org/10.48550/arXiv.2112.03270.

Browne, Chris, David S. Matteson, Linden McBride, Leiqiu Hu, Yanyan Liu, Ying Sun, Jiaming Wen, and Christopher B. Barrett. 2021. "Multivariate Random Forest Prediction of Poverty and Malnutrition Prevalence." *PLOS ONE* 16 (9): e0255519. https://doi.org/10.1371/journal.pone.0255519.

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): 324–339. https://doi.org/10.1198/016214503000125.

Burgert, Clara R., Josh Colston, Thea Roy, and Blake Zachary. 2013. *Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys,* DHS Spatial Analysis Reports No. 7. Calverton, Maryland, USA: ICF International.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785.

Coors, Stefan, Daniel Schalk, Bernd Bischl, and David Rügamer. 2021. *Automatic Componentwise Boosting: An Interpretable AutoML System,* arXiv:2109.05583. https://doi.org/10.48550/arXiv.1811.12808.

Corsi, Daniel J, Melissa Neuman, Jocelyn E Finlay, and SV Subramanian. 2012. "Demographic and Health Surveys: A Profile." *International Journal of Epidemiology* 41 (6): 1602–1613. https://doi.org/10.1093/ije/dys184.

Croft, Trevor N., Aileen M. J. Marshall, and Courtney K. Allen. 2020. *Guide to DHS Statistics DHS-7.* Rockvile, Maryland, USA: ICF.

Cutler, David, Winnie Fung, Michael Kremer, Monica Singhal, and Tom Vogl. 2010. "Early-Life Malaria Exposure and Adult Outcomes: Evidence from Malaria Eradication in India." *American Economic Journal: Applied Economics* 2 (2): 72–94. https://doi.org/10.1257/app.2.2.72.

Desmond-Hellmann, Sue. 2016. "Progress Lies in Precision." *Science* 353 (6301): 731–731. https://doi.org/10.1126/science.aai7598.

Dewey, Kathryn G., and Khadija Begum. 2011. "Long-Term Consequences of Stunting in Early Life." *Maternal & Child Nutrition* 7:5–18. https://doi.org/10.1111/j.1740-8709.2011.00349.x.

Dharamshi, Ameer, Bilal Barakat, Leontine Alkema, and Manos Antoninis. 2022. "A Bayesian Model for Estimating Sustainable Development Goal Indicator 4.1.2: School Completion Rates." *Journal of the Royal Statistical Society: Series C (Applied Statistics),* rssc.12595. https://doi.org/10.1111/rssc.12595.

Didan, Kamel. 2021. (MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V061; accessed November 4, 2022). https://doi.org/10.5067/MODIS/MOD13A2.061.

Diggle, Peter, Rana Moyeed, Barry Rowlingson, and Madeleine Thomson. 2002. "Childhood Malaria in the Gambia: A Case-Study in Model-Based Geostatistics." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 51 (4): 493–506. https://doi.org/10.1111/1467-9876.00283.

Diggle, Peter J., and Emanuele Giorgi. 2016. "Model-Based Geostatistics for Prevalence Mapping in Low-Resource Settings." *Journal of the American Statistical Association* 111 (515): 1096–1120. https://doi.org/10.1080/01621459.2015.1123158.

Dong, Tracy Qi, and Jon Wakefield. 2021a. "Modeling and Presentation of Vaccination Coverage Estimates Using Data from Household Surveys." *Vaccine* 39 (18): 2584–2594. https://doi.org/10.1016/j.vaccine.2021.03.007.

Dong, Tracy Qi, and Jon Wakefield. 2021b. "Space-Time Smoothing Models for Subnational Measles Routine Immunization Coverage Estimation with Complex Survey Data." *The Annals of Applied Statistics* 15 (4): 1959–1979. https://doi.org/10.1214/21-AOAS1474.

Dowell, Scott F., David Blazes, and Susan Desmond-Hellmann. 2016. "Four Steps to Precision Public Health." *Nature* 540 (7632): 189–191. https://doi.org/10.1038/540189a.

Duan, Tony, Anand Avati, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. 2020. "NGBoost: Natural Gradient Boosting for Probabilistic Prediction." In *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org.

Efron, Bradley. 1983. "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation." *Journal of the American Statistical Association* 78 (382): 316–331. https://doi.org/10.1080/01621459.1983.10477973.

Efron, Bradley, and Robert Tibshirani. 1997. "Improvements on Cross-Validation: The 632+ Bootstrap Method." *Journal of the American Statistical Association* 92 (438): 548–560. https://doi.org/10.1080/01621459.1997.10474007.

Egbon, Osafu Augustine, Asrat Mekonnen Belachew, and Mariella Ananias Bogoni. 2022. "Modeling Spatial Pattern of Anemia and Malnutrition Co-Occurrence among under-Five Children in Ethiopia: A Bayesian Geostatistical Approach." *Spatial and Spatio-temporal Epidemiology* 43:100533. https://doi.org/10.1016/j.sste.2022.100533.

Eilers, Paul H. C., and Brian D. Marx. 1996. "Flexible Smoothing with B-splines and Penalties." *Statistical Science* 11 (2): 89–121. https://doi.org/10.1214/ss/1038425655.

Ejigu, Bedilu Alamirie. 2020. "Geostatistical Analysis and Mapping of Malaria Risk in Children of Mozambique." *PLOS ONE* 15 (11): e0241680. https://doi.org/10.1371/journal.pone.0241680.

Fenske, Nora, Jacob Burns, Torsten Hothorn, and Eva A. Rehfuess. 2013. "Understanding Child Stunting in India: A Comprehensive Analysis of Socio-Economic, Nutritional and Environmental Determinants Using Additive Quantile Regression." *PLoS ONE* 8 (11): e78692. https://doi.org/10.1371/journal.pone.0078692.

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.

FEWS NET. 2022. "FEWS NET Data Center: Food Security Classification Data." Accessed November 26, 2022. https://fews.net/data.

Florey, Lia. 2014. *Measures of Malaria Parasitemia Prevalence in National Surveys: Agreement between Rapid Diagnostic Tests and Microscopy,* DHS Analytical Studies No. 43. Rockville, Maryland, USA: ICF International.

Freund, Yoav, and Robert E. Schapire. 1996. "Experiments with a New Boosting Algorithm." In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning,* 148–156. ICML'96, Bari, Italy.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2000. "Additive Logistic Regression: A Statistical View of Boosting (With Discussion and a Rejoinder by the Authors)." *The Annals of Statistics* 28 (2): 337–407. https://doi.org/10.1214/aos/1016218223.

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.

Fuglstad, Geir-Arne, Zehang Richard Li, and Jon Wakefield. 2022. *The Two Cultures for Prevalence Mapping: Small Area Estimation and Spatial Statistics,* arXiv:2110.09576v2. https://doi.org/10.48550/arXiv.2110.09576.

Funk, Chris, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, et al. 2015. "The Climate Hazards Infrared Precipitation with Stations—a New Environmental Record for Monitoring Extremes." *Scientific Data* 2 (1): 150066. https://doi.org/10.1038/sdata.2015.66.

Gelman, Andrew. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22 (2): 153–164. https://doi.org/10.1214/088342306000000691.

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.

Giorgi, Emanuele, and Peter J. Diggle. 2021. *Model-Based Geostatistics for Global Public Health: Methods and Applications.* Routledge.

Godwin, Jessica, and Jon Wakefield. 2021. "Space-time Modeling of Child Mortality at the Admin-2 Level in a Low and Middle Income Countries Context." *Statistics in Medicine* 40 (7): 1593–1638. https://doi.org/10.1002/sim.8854.

Grace, Kathryn, Andrew Verdin, Molly Brown, Maryia Bakhtsiyarava, David Backer, and Trey Billing. 2022. "Conflict and Climate Factors and the Risk of Child Acute Malnutrition Among Children Aged 24–59 Months: A Comparative Analysis of Kenya, Nigeria, and Uganda." *Spatial Demography* 10 (2): 329–358. https://doi.org/10.1007/s40980-021-00102-w.

Greenough, P. Gregg, and Erica L. Nelson. 2019. "Beyond Mapping: A Case for Geospatial Analytics in Humanitarian Health." *Conflict and Health* 13 (1): 50. https://doi.org/10.1186/s13031-019-0234-9.

Griesbach, Colin, Andreas Groll, and Elisabeth Bergherr. 2021. "Joint Modelling Approaches to Survival Analysis via Likelihood-Based Boosting Techniques." *Computational and Mathematical Methods in Medicine* 2021:1–11. https://doi.org/10.1155/2021/4384035.

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.

Hastie, Trevor, and Robert Tibshirani. 1986. "Generalized Additive Models." *Statistical Science* 1 (3): 297–310. https://doi.org/10.1214/ss/1177013604.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition.* 2nd ed. Springer Series in Statistics. New York, NY: Springer.

Hierink, Fleur, Gianluca Boo, Peter M. Macharia, Paul O. Ouma, Pablo Timoner, Marc Levy, Kevin Tschirhart, et al. 2022. "Differences between Gridded Population Data Impact Measures of Geographic Access to Healthcare in Sub-Saharan Africa." *Communications Medicine* 2 (1): 117. https://doi.org/10.1038/s43856-022-00179-4.

Hofner, Benjamin, Luigi Boccuto, and Markus Göker. 2015. "Controlling False Discoveries in High-Dimensional Situations: Boosting with Stability Selection." *BMC Bioinformatics* 16 (1): 144. https://doi.org/10.1186/s12859-015-0575-3.

Hofner, Benjamin, Torsten Hothorn, Thomas Kneib, and Matthias Schmid. 2011. "A Framework for Unbiased Model Selection Based on Boosting." *Journal of Computational and Graphical Statistics* 20 (4): 956–971. https://doi.org/10.1198/jcgs.2011.09220.

Hofner, Benjamin, Thomas Kneib, and Torsten Hothorn. 2016. "A Unified Framework of Constrained Regression." *Statistics and Computing* 26 (1): 1–14. https://doi.org/10.1007/s11222-014-9520-y.

Hohberg, Maike, Francesco Donat, Giampiero Marra, and Thomas Kneib. 2021. "Beyond Unidimensional Poverty Analysis Using Distributional Copula Models for Mixed Ordered-continuous Outcomes." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 70 (5): 1365–1390. https://doi.org/10.1111/rssc.12517.

Hong, Sok Chul. 2011. "Malaria and Economic Productivity: A Longitudinal Analysis of the American Case." *The Journal of Economic History* 71 (3): 654–671. https://doi.org/10.1017/S0022050711001872.

Horvitz, D. G., and D. J. Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47 (260): 663–685. https://doi.org/10.1080/01621459.1952.10483446.

Hsiao, Yuan, Lee Fiorio, Jonathan Wakefield, and Emilio Zagheni. 2023. "Modeling the Bias of Digital Data: An Approach to Combining Digital With Official Statistics to Estimate and Predict Migration Trends." *Sociological Methods & Research,* 004912412211401. https://doi.org/10.1177/00491241221140144.

ICF International. 2012. *Demographic and Health Survey Sampling and Household Listing Manual.* MEASURE DHS, Calverton, Maryland, U.S.A: ICF International.

Institut National de la Statistique (INSTAT) and ICF. 2022. *Enquête Démographique et de Santé à Madagascar (EDSMD-V) 2021.* Antananarivo, Madagascar et Rockville, Maryland, USA.

Institut National de la Statistique (INSTAT), Programme National de Lutte contre le Paludisme (PNLP), and The DHS Program. 2022. *Enquête sur les Indicateurs du Paludisme au Mali 2021.* Bamako, Mali et Rockville, Maryland, USA.

Jarvis, A., H.I. Reuter, A. Nelson, and E. Guevara. 2008. *Hole-Filled SRTM for the Globe Version 4, Available from the CGIAR-CSI SRTM 90m Database.*

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353 (6301): 790–794. https://doi.org/10.1126/science.aaf7894.

Kandala, Ngianga-Bakwin, Ludwig Fahrmeir, Stephan Klasen, and Jan Priebe. 2009. "Geo-Additive Models of Childhood Undernutrition in Three Sub-Saharan African Countries: Childhood Undernutrition in Africa." *Population, Space and Place* 15 (5): 461–473. https://doi.org/10.1002/psp.524.

Kapoor, Sayash, and Arvind Narayanan. 2022. *Leakage and the Reproducibility Crisis in ML-based Science,* arXiv: 2207.07048. https://doi.org/10.48550/arXiv.2207.07048.

Khan, Shane, and Attila Hancioglu. 2019. "Multiple Indicator Cluster Surveys: Delivering Robust Data on Children and Women across the Globe." *Studies in Family Planning* 50 (3): 279–286. https://doi.org/10.1111/sifp.12103.

Kim, Rockli, Avleen S. Bijral, Yun Xu, Xiuyuan Zhang, Jeffrey C. Blossom, Akshay Swaminathan, Gary King, et al. 2021. "Precision Mapping Child Undernutrition for Nearly 600,000 Inhabited Census Villages in India." *Proceedings of the National Academy of Sciences* 118 (18): e2025865118. https://doi.org/10.1073/pnas.2025865118.

Kinyoki, Damaris K., Aaron E. Osgood-Zimmerman, Brandon V. Pickering, Lauren E. Schaeffer, Laurie B. Marczak, Alice Lazzar-Atwood, Michael L. Collison, et al. 2020. "Mapping Child Growth Failure across Low- and Middle-Income Countries." *Nature* 577 (7789): 231–234. https://doi.org/10.1038/s41586-019-1878-8.

Klein, Nadja, Manuel Carlan, Thomas Kneib, Stefan Lang, and Helga Wagner. 2021. "Bayesian Effect Selection in Structured Additive Distributional Regression Models." *Bayesian Analysis* 16 (2): 545–573. https://doi.org/10.1214/20-BA1214.

Klein, Nadja, Thomas Kneib, Stefan Lang, and Alexander Sohn. 2015. "Bayesian Structured Additive Distributional Regression with an Application to Regional Income Inequality in Germany." *The Annals of Applied Statistics* 9 (2): 1024–1052. https://doi.org/10.1214/15-AOAS823.

Kneib, Thomas, Torsten Hothorn, and Gerhard Tutz. 2009. "Variable Selection and Model Choice in Geoadditive Regression Models." *Biometrics* 65 (2): 626–634. https://doi.org/10.1111/j.1541-0420.2008.01112.x.

Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2,* 1137–1143. IJCAI'95, Montreal, Quebec, Canada. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Kriegler, Brian, and Richard Berk. 2010. "Small Area Estimation of the Homeless in Los Angeles: An Application of Cost-Sensitive Stochastic Gradient Boosting." *The Annals of Applied Statistics* 4 (3): 1234–1255. https://doi.org/10.1214/10-AOAS328.

Laga, Ian, Xiaoyue Niu, and Le Bao. 2022. "Modeling the Marked Presence-Only Data: A Case Study of Estimating the Female Sex Worker Size in Malawi." *Journal of the American Statistical Association* 117 (537): 27–37. https://doi.org/10.1080/01621459.2021.1944873.

Lucas, Adrienne M. 2010. "Malaria Eradication and Educational Attainment: Evidence from Paraguay and Sri Lanka." *American Economic Journal: Applied Economics* 2 (2): 46–71. https://doi.org/10.1257/app.2.2.46.

Lumley, Thomas. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9 (8): 1–19. https://doi.org/10.18637/jss.v009.i08.

Mayr, Andreas, Nora Fenske, Benjamin Hofner, Thomas Kneib, and Matthias Schmid. 2012. "Generalized Additive Models for Location, Scale and Shape for High Dimensional Data—a Flexible Approach Based on Boosting." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61 (3): 403–427. https://doi.org/10.1111/j.1467-9876.2011.01033.x.

McGovern, Mark E, Aditi Krishna, Victor M Aguayo, and Sv Subramanian. 2017. "A Review of the Evidence Linking Child Stunting to Economic Outcomes." *International Journal of Epidemiology* 46 (4): 1171–1191. https://doi.org/10.1093/ije/dyx017.

Meinshausen, Nicolai, and Peter Bühlmann. 2010. "Stability Selection: Stability Selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (4): 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x.

Mercer, Laina D., Jon Wakefield, Athena Pantazis, Angelina M. Lutambi, Honorati Masanja, and Samuel Clark. 2015. "Space–Time Smoothing of Complex Survey Data: Small Area Estimation for Child Mortality." *The Annals of Applied Statistics* 9 (4). https://doi.org/10.1214/15-AOAS872.

Michler, Jeffrey D., Anna Josephson, Talip Kilic, and Siobhan Murray. 2022. "Privacy Protection, Measurement Error, and the Integration of Remote Sensing and Socioeconomic Survey Data." *Journal of Development Economics* 158:102927. https://doi.org/10.1016/j.jdeveco.2022.102927.

Millar, Justin, Paul Psychas, Benjamin Abuaku, Collins Ahorlu, Punam Amratia, Kwadwo Koram, Samuel Oppong, and Denis Valle. 2018. "Detecting Local Risk Factors for Residual Malaria in Northern Ghana Using Bayesian Model Averaging." *Malaria Journal* 17 (1): 343. https://doi.org/10.1186/s12936-018-2491-2.

Mohammed, Kamaldeen, Mohammed Gazali Salifu, Evans Batung, Daniel Amoak, Vasco Ayere Avoka, Moses Kansanga, and Isaac Luginaah. 2022. "Spatial Analysis of Climatic Factors and Plasmodium Falciparum Malaria Prevalence among Children in Ghana." *Spatial and Spatio-temporal Epidemiology* 43:100537. https://doi.org/10.1016/j.sste.2022.100537.

Møller, Anders Bjørn, Amélie Marie Beucher, Nastaran Pouladi, and Mogens Humlekrog Greve. 2020. "Oblique Geographic Coordinates as Covariates for Digital Soil Mapping." *SOIL* 6 (2): 269–289. https://doi.org/10.5194/soil-6-269-2020.

Nzabakiriraho, Jean Damascene, and Ezra Gayawan. 2021. "Geostatistical Modeling of Malaria Prevalence among Under-Five Children in Rwanda." *BMC Public Health* 21 (1): 369. https://doi.org/10.1186/s12889-021-10305-x.

Paige, John, Geir-Arne Fuglstad, Andrea Riebler, and Jon Wakefield. 2022. "Design- and Model-Based Approaches to Small-Area Estimation in a Low- and Middle-Income Country Context: Comparisons and Recommendations." *Journal of Survey Statistics and Methodology* 10 (1): 50–80. https://doi.org/10.1093/jssam/smaa011.

Pape, Utz, and Philip Wollburg. 2019. *Estimation of Poverty in Somalia Using Innovative Methodologies,* Policy Research Working Paper. Washington, DC: World Bank. https://doi.org/10.1596/1813-9450-8735.

Phalkey, Revati K., Clara Aranda-Jan, Sabrina Marx, Bernhard Höfle, and Rainer Sauerborn. 2015. "Systematic Review of Current Efforts to Quantify the Impacts of Climate Change on Undernutrition." *Proceedings of the National Academy of Sciences* 112 (33): E4522–E4529. https://doi.org/10.1073/pnas.1409769112.

Phillips, Margaret A., Jeremy N. Burrows, Christine Manyando, Rob Hooft van Huijsduijnen, Wesley C. Van Voorhis, and Timothy N. C. Wells. 2017. "Malaria." *Nature Reviews Disease Primers* 3 (1): 17050. https://doi.org/10.1038/nrdp.2017.50.

Piantadosi, Steven, David P. Byar, and Sylvan B. Green. 1988. "The Ecological Fallacy." *American Journal of Epidemiology* 127 (5): 893–904. https://doi.org/10.1093/oxfordjournals.aje.a114892.

Rahman, Zara, and Julia Keseru. 2021. *Predictive Analytics for Children: An Assessment of Ethical Considerations, Risks, and Benefits,* Innocenti Working Papers 2021-08. Florence, Italy: UNICEF Office of Research - Innocenti.

Raschka, Sebastian. 2020. *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning,* arXiv: 1811.12808. https://doi.org/10.48550/arXiv.1811.12808.

Rigby, R. A., and D. M. Stasinopoulos. 2005. "Generalized Additive Models for Location, Scale and Shape (with Discussion)." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54 (3): 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x.

Roser, Max, and Hannah Ritchie. 2019. "Hunger and Undernourishment." *Our World in Data.*

Sandefur, Justin, and Amanda Glassman. 2015. "The Political Economy of Bad Data: Evidence from African Survey and Administrative Statistics." *The Journal of Development Studies* 51 (2): 116–132. https://doi.org/10.1080/00220388.2014.968138.

Scheipl, Fabian, Ludwig Fahrmeir, and Thomas Kneib. 2012. "Spike-and-Slab Priors for Function Selection in Structured Additive Regression Models." *Journal of the American Statistical Association* 107 (500): 1518–1532. https://doi.org/10.1080/01621459.2012.737742.

Schiavina, Marcello, Sergio Freire, and Kytt MacManus. 2022. (GHS-POP R2022A - GHS Population Grid Multitemporal (1975-2030); accessed November 25, 2022). https://doi.org/10.2905/D6D86A90-4351-4508-99C1-CB074B022C4A.

Schiavina, Marcello, Michele Melchiorri, and Martino Pesaresi. 2022. (GHS-SMOD R2022A - GHS Settlement Layers, Application of the Degree of Urbanisation Methodology (Stage I) to GHS-POP R2022A and GHS-BUILT-S R2022A, Multitemporal (1975-2030); accessed November 25, 2022). https://doi.org/10.2905/4606D58A-DC08-463C-86A9-D49EF461C47F.

Schlosser, Lisa, Torsten Hothorn, Reto Stauffer, and Achim Zeileis. 2019. "Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain." *The Annals of Applied Statistics* 13 (3): 1564–1589. https://doi.org/10.1214/19-AOAS1247.

Schmid, Matthias, and Torsten Hothorn. 2008. "Boosting Additive Models Using Component-Wise P-Splines." *Computational Statistics & Data Analysis* 53 (2): 298–311. https://doi.org/10.1016/j.csda.2008.09.009.

Seiler, Johannes, Kenneth Harttgen, Thomas Kneib, and Stefan Lang. 2021. "Modelling Children's Anthropometric Status Using Bayesian Distributional Regression Merging Socio-Economic and Remote Sensed Data from South Asia and Sub-Saharan Africa." *Economics & Human Biology* 40:100950. https://doi.org/10.1016/j.ehb.2020.100950.

Shah, Rajen D., and Richard J. Samworth. 2013. "Variable Selection with Error Control: Another Look at Stability Selection: *Another Look at Stability Selection*." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (1): 55–80. https://doi.org/10.1111/j.1467-9868.2011.01034.x.

Sobotka, Fabian, and Thomas Kneib. 2012. "Geoadditive Expectile Regression." *Computational Statistics & Data Analysis* 56 (4): 755–767. https://doi.org/10.1016/j.csda.2010.11.015.

Spears, Dean, Diane Coffey, and Jere R. Behrman. 2022. "Endogenous Inclusion in the Demographic and Health Survey Anthropometric Sample: Implications for Studying Height within Households." *Journal of Development Economics* 155:102783. https://doi.org/10.1016/j.jdeveco.2021.102783.

Steele, Jessica E., Pål Roe Sundsøy, Carla Pezzulo, Victor A. Alegana, Tomas J. Bird, Joshua Blumenstock, Johannes Bjelland, et al. 2017. "Mapping Poverty Using Mobile Phone and Satellite Data." *Journal of The Royal Society Interface* 14 (127): 20160690. https://doi.org/10.1098/rsif.2016.0690.

Strömer, Annika, Nadja Klein, Christian Staerk, Hannah Klinkhammer, and Andreas Mayr. 2022. *Boosting Multivariate Structured Additive Distributional Regression Models,* arXiv:2207.08470. https://doi.org/10.48550/arXiv.2207.08470.

Strömer, Annika, Christian Staerk, Nadja Klein, Leonie Weinhold, Stephanie Titze, and Andreas Mayr. 2022. "Deselection of Base-Learners for Statistical Boosting—with an Application to Distributional Regression." *Statistical Methods in Medical Research* 31 (2): 207–224. https://doi.org/10.1177/09622802211051088.

Thomas, Janek, Tobias Hepp, Andreas Mayr, and Bernd Bischl. 2017. "Probing for Sparse and Fast Variable Selection with Model-Based Boosting." *Computational and Mathematical Methods in Medicine* 2017:e1421409. https://doi.org/10.1155/2017/1421409.

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): 673–687. https://doi.org/10.1007/s11222-017-9754-6.

Torres Munguía, Juan Armando, and Inmaculada Martínez-Zarzoso. 2021. "Examining Gender Inequalities in Factors Associated with Income Poverty in Mexican Rural Households." *PLOS ONE* 16 (11): e0259187. https://doi.org/10.1371/journal.pone.0259187.

UNICEF. 2013. *Improving Child Nutrition: The Achievable Imperative for Global Progress.* New York: United Nations Children's Fund.

United Nations. 2015. *General Assembly Resolution 70/1: Transforming Our World: The 2030 Agenda for Sustainable Development.*

Uwiringiyimana, Vestine, Frank Osei, Sherif Amer, and Antonie Veldkamp. 2022. "Bayesian Geostatistical Modelling of Stunting in Rwanda: Risk Factors and Spatially Explicit Residual Stunting Burden." *BMC Public Health* 22 (1): 159. https://doi.org/10.1186/s12889-022-12552-y.

Van der Merwe, Eduard, Matthew Clance, and Eleni Yitbarek. 2022. "Climate Change and Child Malnutrition: A Nigerian Perspective." *Food Policy,* 102281. https://doi.org/10.1016/j.foodpol.2022.102281.

Victora, Cesar G, Parul Christian, Luis Paulo Vidaletti, Giovanna Gatica-Domínguez, Purnima Menon, and Robert E Black. 2021. "Revisiting Maternal and Child Undernutrition in Low-Income and Middle-Income Countries: Variable Progress towards an Unfinished Agenda." *The Lancet* 397 (10282): 1388–1399. https://doi.org/10.1016/S0140-6736(21)00394-9.

Wade, Sara, Raffaella Piccarreta, Andrea Cremaschi, and Isadora Antoniano-Villalobos. 2022. "Colombian Women's Life Patterns: A Multivariate Density Regression Approach." *Bayesian Analysis* 17 (2): 405–433. https://doi.org/10.1214/20-BA1256.

Wakefield, Jon, Geir-Arne Fuglstad, Andrea Riebler, Jessica Godwin, Katie Wilson, and Samuel J Clark. 2019. "Estimating Under-Five Mortality in Space and Time in a Developing World Context." *Statistical Methods in Medical Research* 28 (9): 2614–2634. https://doi.org/10.1177/0962280218767988.

Wan, Zhengming, Simon Hook, and Glynn Hulley. 2021. (MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V061; accessed August 8, 2022). https://doi.org/10.5067/MODIS/MOD11A2.061.

Wang, Zhengfan, Miranda J. Fix, Lucia Hug, Anu Mishra, Danzhen You, Hannah Blencowe, Jon Wakefield, and Leontine Alkema. 2022. "Estimating the Stillbirth Rate for 195 Countries Using a Bayesian Sparse Regression Model with Temporal Smoothing." *The Annals of Applied Statistics* 16 (4). https://doi.org/10.1214/21-AOAS1571.

Weiss, Daniel J, Tim C D Lucas, Michele Nguyen, Anita K Nandi, Donal Bisanzio, Katherine E Battle, Ewan Cameron, et al. 2019. "Mapping the Global Prevalence, Incidence, and Mortality of Plasmodium Falciparum, 2000–17: A Spatial and Temporal Modelling Study." *The Lancet* 394 (10195): 322–331. https://doi.org/10.1016/S0140-6736(19)31097-9.

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.

Weiss, Daniel J, A. Nelson, H. S. Gibson, W. Temperley, S. Peedell, A. Lieber, M. Hancher, et al. 2018. "A Global Map of Travel Time to Cities to Assess Inequalities in Accessibility in 2015." *Nature* 553 (7688): 333–336. https://doi.org/10.1038/nature25181.

Weiss, Daniel J, A. Nelson, C. A. Vargas-Ruiz, K. Gligorić, S. Bavadekar, E. Gabrilovich, A. Bertozzi-Villa, et al. 2020. "Global Maps of Travel Time to Healthcare Facilities." *Nature Medicine* 26 (12): 1835–1838. https://doi.org/10.1038/s41591-020-1059-1.

Wieczorek, Jerzy, Cole Guerin, and Thomas McMahon. 2022. "K-Fold Cross-Validation for Complex Sample Surveys." *Stat* 11 (1): e454. https://doi.org/10.1002/sta4.454.

Winship, Christopher, and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods & Research* 23 (2): 230–257. https://doi.org/10.1177/0049124194023002004.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. Second edition. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton: CRC Press/Taylor & Francis Group.

World Bank. 2022. "Data Bank." Accessed December 4, 2022. https://data.worldbank.org/.

World Health Organization. 2021. *World Malaria Report 2021*. Geneva: World Health Organization.

Wu, Yunhan, and Jon Wakefield. 2022. *Modeling Urban / Rural Fractions in Low- and Middle-Income Countries*, arXiv: 2209.10619. https://doi.org/10.48550/arXiv.2209.10619.

Ziulu, Maria Virginia, Jessica Marie Meckler, Gonzalo Hernandez Licona, and Jozef Leonardus Vaessen. 2022. *Poverty Mapping : Innovative Approaches to Creating Poverty Maps with New Data Sources*. Working Paper, IEG Methods and Evaluation Capacity Development Working Paper Series 174763. Washington, DC: Independent Evaluation Group, World Bank.

A Supplementary Material

A.1 Data sources, availability and additional details

Survey regions and country borders were retrieved from the Spatial Data Repository (ICF International 2022) and the Database of Global Administrative Boundaries (GADM) (Global Administrative Areas 2022). The remotely sensed covariates in section 5 are mean values over the year preceding the start date of the survey fieldwork. All raster files are open-access and were retrieved from the Google Earth Engine API (Gorelick et al. 2017) or the respective provider as described in the respective publications at a spatial resolution of 1km x 1km.

To create a fine-scaled country grid I use the spatial unit indexing system *H3: A Hexagonal Hierarchical Geospatial Indexing System* (Uber Technologies 2022). The grid level estimates are produced for a grid of resolution 7, hexagons with an area of approximately 5km^2. I extract for each cluster location or hexagon centroid the value interpolated from the values of the four nearest raster cells. One exception is the population data, where exact areal extraction is used to obtain a consistent disaggregation of population totals.

A.2 Computational implementation

All analyses were conducted in `R` 4.2.2 (R Core Team 2022) and Python 3.9.13. The code files and data requirements to fully replicate this work along with additional results are included in the corresponding GitHub repository.[22]

The described models were fitted using the `mboost` and `gamboostLSS` packages (Hothorn et al. 2022; Hofner, Mayr, and Schmid 2016), see also Hothorn et al. (2010) for an introduction. The following `R` packages provided helpful functions for evaluation metrics, raster extraction and survey data analysis: Hamner and Frasco (2018), Pfeffer et al. (2018), Watson, FitzJohn, and Eaton (2019), Lumley (2020), Hijmans, Ghosh, and Mandel (2022), and Baston (2022).

A.3 Distributions

Binomial distribution

$$Y \sim \text{Binomial}(n, \mu)$$

For $y = 0, 1, \ldots, n$ and $0 < \mu < 1$, the probability density function of the binomial distribution is

$$f(y|n, \mu) = \frac{n!}{y!(n-y)!}\mu^y(1-\mu)^{n-\mu} \tag{6}$$

where the first and second moments are

$$E(Y) = n\mu,$$
$$Var(Y) = n\mu(1-\mu).$$

Beta-binomial distribution

$$Y \sim BB(n, \mu, \sigma)$$

22. See https://github.com/danielseussler/ssahealthriskfactors.

For $y = 0, 1, \ldots, n$, $0 < \mu < 1$, and $\sigma > 0$, the probability density function of the beta-binomial distribution is

$$f(y|n, \mu, \sigma) = \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \frac{\Gamma(\frac{1}{\sigma})\Gamma(y+\frac{\mu}{\sigma})\Gamma(n+\frac{(1-\mu)}{\sigma}-y)}{\Gamma(n+\frac{1}{\sigma})\Gamma(\frac{\mu}{\sigma})\Gamma(\frac{1-\mu}{\sigma})}. \tag{7}$$

The corresponding first and second moments are

$$\begin{aligned} E(Y) &= n\mu, \\ Var(Y) &= n\mu(1-\mu)[1+\sigma(n-1)/(1+\sigma)]. \end{aligned}$$

See also Rigby et al. (2019) for further information.

A.4 Additional Results

In this section, I present additional figures and tables for the two case studies.

Base-learner	Frequency
cage	1.00
csex	1.00
ctwin	1.00
cbord	1.00
mbmi	1.00
mage	0.74
medu	1.00
memployed	0.22
mreligion	0.98
nodead	0.86
hmembers	1.00
watersource	0.90
sanitation	0.44
wealth	1.00
electricity	0.30
radio	1.00
television	1.00
bicycle	0.28
motorcycle	0.82
car	0.70
urban	0.52
healthaccess	0.36
cityaccess	1.00
fews	0.86
f(cage)	1.00
f(mage)	0.90
f(mbmi)	0.40
f(medu)	1.00
f(hmembers)	0.48
f(healthaccess)	0.68
f(cityaccess)	0.98
f(dhsregion)	1.00

Table 4: *Childhood malnutrition in Madagascar: selection frequencies of base learners over the 50 replications. The name indicates the linear effect only, f(·) is the non-linear deviation from the linear effect.*

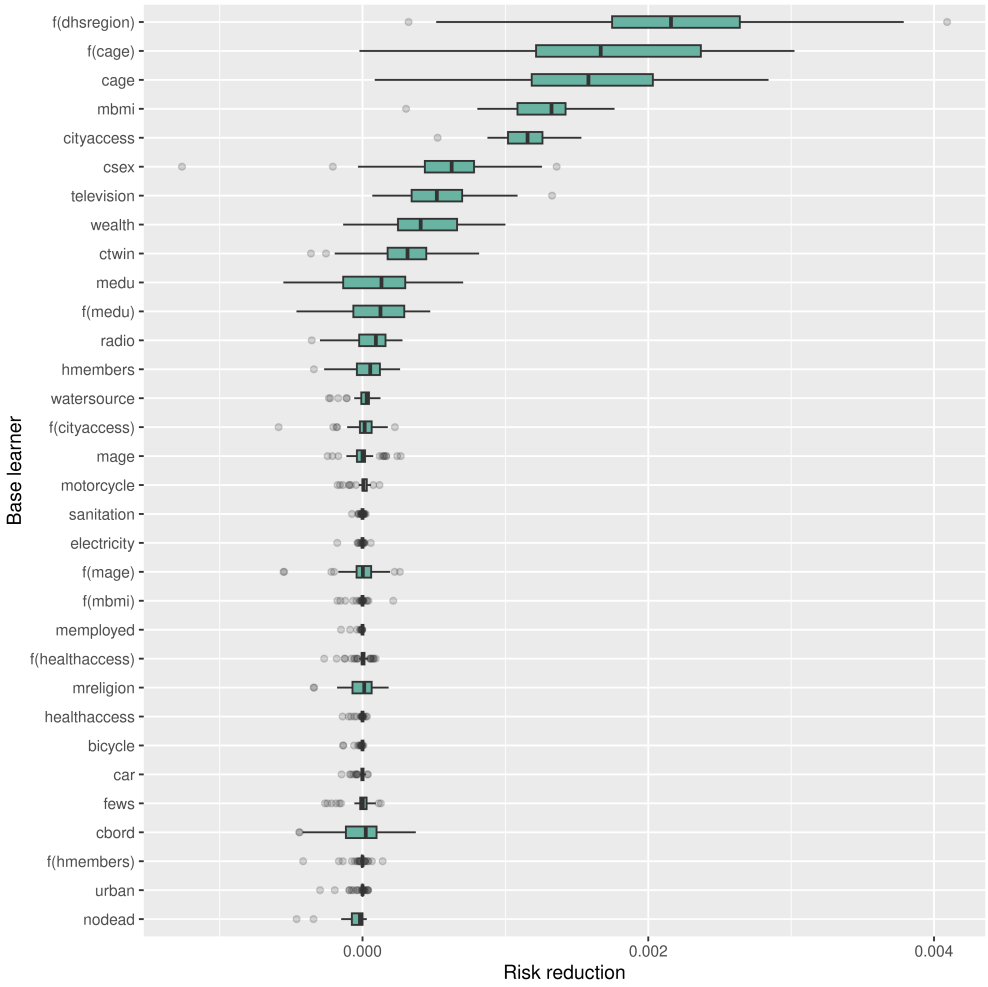

Figure 17: *Childhood malnutrition in Madagascar: empirical distributions of the variable importance of each base-learner by attributed risk reduction over 50 replications. Note, (small) negative risk reduction can be obtained as a fitting artefact if boosting iterations are extended beyond the maximum likelihood estimates.*

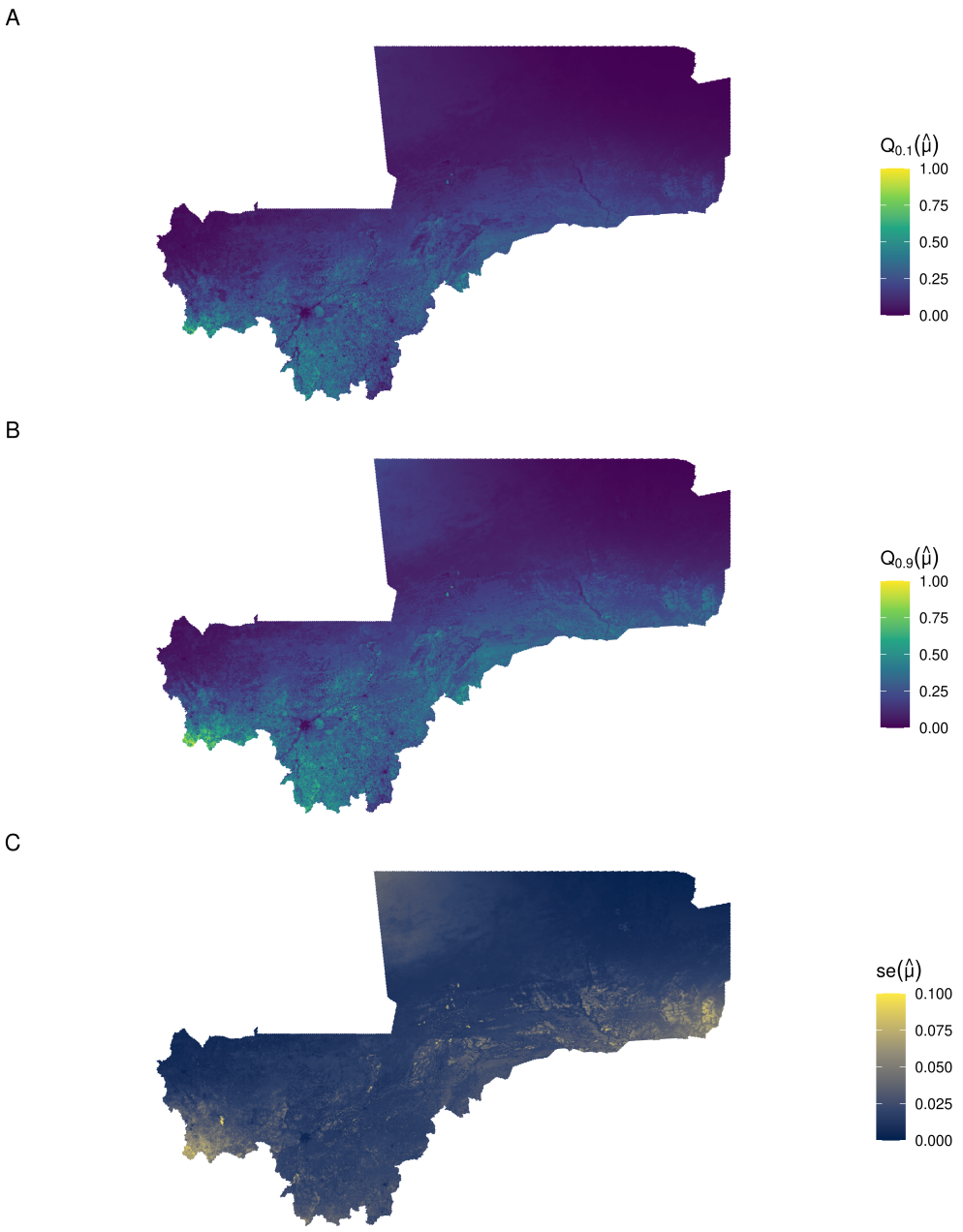

Figure 18: *Geographic malaria risk in Mali: lower, upper quantiles and standard error of the predicted risk $\hat{\mu}$ based on 50 bootstrap samples.*

References

Baston, Daniel. 2022. *Exactextractr: Fast Extraction from Raster Datasets Using Polygons*. R package version 0.9.1.

Global Administrative Areas. 2022. "GADM Database of Global Administrative Areas, Version 4.1." Accessed November 26, 2022. www.gadm.org.

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. "Google Earth Engine: Planetary-scale Geospatial Analysis for Everyone." *Remote Sensing of Environment* 202:18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Hamner, Ben, and Michael Frasco. 2018. *Metrics: Evaluation Metrics for Machine Learning*. R package version 0.1.4.

Hijmans, Robert J., Aniruddha Ghosh, and Alex Mandel. 2022. *Geodata: Download Geographic Data*. R package version 0.4-9.

Hofner, Benjamin, Andreas Mayr, and Matthias Schmid. 2016. "gamboostLSS: An R Package for Model Building and Variable Selection in the GAMLSS Framework." *Journal of Statistical Software* 74 (1): 1–31. https://doi.org/10.18637/jss.v074.i01.

Hothorn, Torsten, Peter Buehlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. 2022. *Mboost: Model-Based Boosting*. R package version 2.9-7.

Hothorn, Torsten, Peter Bühlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. 2010. "Model-Based Boosting 2.0." *Journal of Machine Learning Research* 11 (71): 2109–2113.

ICF International. 2022. "Spatial Data Repository, The Demographic and Health Surveys Program." Accessed November 26, 2022. https://spatialdata.dhsprogram.com/.

Lumley, Thomas. 2020. *Survey: Analysis of Complex Survey Samples*. R package version 4.0.

Pfeffer, Daniel A., Timothy C. D. Lucas, Daniel May, Joseph Harris, Jennifer Rozier, Katherine A. Twohig, Ursula Dalrymple, et al. 2018. "malariaAtlas: An R Interface to Global Malariometric Data Hosted by the Malaria Atlas Project." *Malaria Journal* 17 (1): 352. https://doi.org/10.1186/s12936-018-2500-5.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rigby, Robert A., Mikis D. Stasinopoulos, Gillian Z. Heller, and Fernanda De Bastiani. 2019. *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. 1st ed. Chapman and Hall/CRC. https://doi.org/10.1201/9780429298547.

Uber Technologies. 2022. "H3: A Hexagonal Hierarchical Geospatial Indexing System." https://h3geo.org/.

Watson, Oliver J., Rich FitzJohn, and Jeffrey W. Eaton. 2019. "Rdhs: An R Package to Interact with The Demographic and Health Surveys (DHS) Program Datasets." *Wellcome Open Research* 4:103. https://doi.org/10.12688/wellcomeopenres.15311.1.

Declaration of Authorship

I hereby certify that I have written the present thesis entitled *Identification of Health Risk Factors in Developing Countries using Intrinsic Model Selection Approaches* independently and that the work contained herein is my own. All formulations and concepts taken verbatim or in substance from printed or unprinted material or the Internet have been cited according to the rules of good scientific practice and indicated by exact references to the original source. The same applies to all illustrations. The present thesis has not been submitted to another university for the award of an academic degree in this form. I understand that the provision of incorrect information may have legal consequences.

Date, Signature