Predicting Wasting Burden in the Sahel Region

Team 17

April 20, 2023

Abstract

Malnutrition is one of the leading causes of infant mortality in the Sahel region in Africa. Predicting hot-spots of child wasting is then crucial for policymakers to intervene in a decisive and timely fashion. This paper proposes a multiclass classification model to predict wasting, using sub-national information on nutrition and health. We find that eXtreme Gradient Boosting provides the best performance, when evaluated against competing methods. Particularly, XGBoosting predicts wasting with an accuracy of 84% on the test set. We also show that enriching the data with information on consumer and producer local food prices significantly increases the models' predictive power. Finally, we highlight that local food prices and climatic conditions emerges as the main fundamental determinants of infant severe malnutrition in the Sahel region.

1 Introduction

Prevention of malnutrition is one of the main tasks tackled by humanitarian organizations. According to the latest Unicef estimates[1], around 3.6 million children in Eastern and Southern Africa are in urgent need of life-saving treatment for severe wasting. Malnutrition has been shown to be the primary cause of immunodeficiency in infants (Katona and Katona-Apte, 2008), hindering to psycho-motor development (McDonald et al., 2013) and detrimental to long-run health (Victora et al., 2008). It appears evident that accurate prediction of possible hotspots for malnutrition is of pivotal importance not only to alleviate this dramatic issue, but to also prevent it from causing irreparable, long-lasting damage to individuals and societies alike.

This paper responds to the need for a reliable classification algorithm to predict the burden of severe infant (6-59 months) malnutrition, i.e. wasting, in the Sahel region of Africa. We also tackle the issue of identifying the most important drivers of this phenomenon, providing policy makers with better information to design food provision and malnutrition relief policies.

To do so, we first build a data set at the sub-national level (ADM-2) for 8 Sub-saharian countries observed from 2020 to 2022. We collect data on burden of wasting and its potential health, natural and socio-economic determinants from various sources. Moreover, we enrich the starting data set with gridded-level information on climatic conditions, and consumer and producer local food prices. Specifically, for climatic conditions we use a well-know index of measuring soil dryness and wetness, the Standardized Precipitation Evapotranspiration Index (SPEI). We compute both a 12-month and 48-month moving average to capture both short- and medium-run effects of changing in the amount of water in the land. As for prices, we use consumer and producer food price indexes, proposed and compiled by McGuirk and Burke (2020), which allows us to also include local food price shocks as covariates. Before the empirical analysis, we also tackle the large number of missing values of the data set using imputation, whenever possible. Specifically, we impute data at the administrative level (ADM-2) by averaging, using population weights at higher administrative areas (ADM-1 and ADM-2), the observed information at ADM-2.

Second, we compare three different machine learning tools to identify the most accurate at predicting infant wasting. Initially, we perform a preliminary analysis through a multivariate logit model along with a group-Lasso, weighted for the number of children from 6 to 59 months at the local level, to identify the key features for prediction. As expected, the model can be sensibly improved in terms of accuracy, but reinforces the importance, among well-established variables related to nutrition and exposure to risk, also of external variables in predicting child malnutrition. We extend our analysis by applying a Random Forest classification algorithm aimed at minimising the classification error rate of the model, which also allows us to assess the relative contribution of each variables for prediction. Lastly, we implement eXtreme Gradient Boosting (XGBoosting) to further improve the predictive power of our model. Each model is evaluated on a test set of 40% of the original data, with hyperparameters optimally chosen by means of appropriate cross-validation techniques. Overall, we find that XGBoosting outperforms the competing models with an overall accuracy of 84%. We also show that both the SPEI and price indexes are significant predictor of child malnutrition, boosting model prediction, hence suggesting important channels spanning from economic forces.

Our paper contributes to the literature on how to predict the burden of severe malnutrition (Bulti et al., 2017; Dale et al., 2017; Isanaka et al., 2021). From one stand point, we propose machine learning multiclass classification algorithms which outperform standard statistical techniques in

[1]Unicef press release, January 2022. Available at: https://www.unicef.org/esa/press-releases/children-lack-life-saving-treatment-for-severe-wasting

predicting infants with severe malnutrition. Moreover, differently from previous works, we include local producers and consumers' prices, which prove to be key determinants in identifying areas with severe child malnutrition. We show that they not only sensibly increase the accuracy of the model, but they also emerge as fundamental drivers of wasting. This is in line with existing literature showing that local malnutrition is deeply intertwined with global price dynamics (Cudjoe et al., 2010), with food price shocks associated to more severe malnutrition (Cornia et al., 2016), ultimately affecting individuals which are most exposed to risk (Bloem et al., 2009).

The rest of the paper is structured as follows: Section 2 introduces the data used in the analysis and explains the process of imputing and data-augmenting undertaken preliminary to the analysis. Section 3 presents the empirical framework, with insights on the importance of both the algorithm choice and the selection of covariates, with a focus on external measures. Section 4 presents the results obtained in our empirical specification, which are further discussed in Section 5, along with potential limitations of our approach and further extensions.

2 Data

This section presents the data used in our analysis. To address our research questions, we require data with specific features: first, we need sub-national information on infant nutrition to measure the level of child wasting and more in general malnutrition in a specific area. Second, we also require geographically disaggregated information on health conditions, institutional settings, climatic variables and food prices to enrich the data set.

2.1 Nutrition and Health Data

Our main source for the analysis is a data set on Administrative Level-2 (ADM-2) sub-national regions for 8 Subsaharian countries covering the years from 2020 to 2022. We collect this data from three main reference sources: (i) the Hotspot Analysis Data (HA), which provides malnutrition burden figures for each administrative sub-national areas; (ii) the Nutrition Survey Data (NS), with information on both nutrition and physical deterioration (diseases) at the individual level, that we then aggregate at the subnational level (ADM-2); (iii) the INFORM Sahel data[2], which collects more than 40 indicators on several measures of inequality, exposure to risk and political instability at the ADM-1 level.

Through this data collection we build our main variable of interest: a categorical variable indicating the level[3] of severe malnutrition, particularly wasting, based on burden[4] of infants between 6 and 59 months in each sub-national region.

At this stage, the data set consists of 1203 observations for the countries of interests.

2.2 Food Prices

Relevant missing determinants of infant malnutrition from the previous data source are food prices for both production and consumption. Food production prices are particularly crucial in areas where subsistence farming is still one of the main source for food supplies, like in the Sub-Saharan region (Hilson, 2016).

[2]It is initiated by the Emergency Response and Preparedness Group of regional Inter-Agency Standing Committee (IASC)

[3]The categorical variable is classified as follows: "Low" (0), "Medium" (1), "High" (2) and "Very High"(3)

[4]We distinguish burden from prevalence in our analysis. Particularly, we use the former as a more adequate measure of the level of malnutrition in a geographical area. Burden is obtained as follows: $Burden = PrevalentCases + IncidentCases = Population_{6-59months} \times Prevalence(1 + ICF)$, where $ICF = 1.6$

To include prices in our analysis we would require food price data with three general properties: spatial variation across sub-national levels, variation over time and variation that captures shocks on local nutrition level of infants. However, collecting this data is not trivial, as they usually come either at global level or they have to be collected using specific survey data[5] in the country of interest. We therefore decide to use food price data from McGuirk and Burke (2020). They construct gridded ($0.5^\circ \times 0.5^\circ$) standardised food prices indexes for both consumer and producers for the whole African continent. They do so first combining temporal variation in global crop prices[6] with local-level spatial variation in crop production and consumption patterns (relative importance of a crop for consumption and production in each grid).
As their data only cover the years 1989–2013, for our analysis we restrict the information to the period 2010-2012. We then extract an average of each price index for each sub-national level included in our data set. To do so, we match each grid cell with the sub-national region where most of its area falls.[7]

2.3 Global Drought Monitor

One drawback from INFORM Sahel data is that they provide ADM-1 rather than ADM-2 sub-national information. As we cannot geographically improve all the indexes that are included in the INFORM Sahel data, we decide to enhance only few well-known drivers of infant malnutrition for which the data are easily accessible: land degradation, floods and droughts. To do so, we collect a climatological measure, the Standardized Precipitation Evapotranspiration Index (SPEI) from the Global SPEI Dataset[8], which provides long-term, robust information about dryness conditions of the soil at the global scale, with a 0.1°spatial resolution and a monthly time resolution. This index compares the amount of precipitation and potential evapotranspiration to obtain measures of drought, land deterioration and floods based on water balance. The SPEI takes on values ranging from -3 (extreme drought) to +3 (flood). We use this information to recover two variables: *spei12*, a 12-month moving average of the index, which we average over the calendar year, and *avg_spei48*, which is a 48-month average instead.[9] These two variables allow to capture both long-run (land deterioration through absence of groundwater) and short-run (extreme events) consequences of absence and abundance of precipitation. Similar to food prices, we match each grid cell with the sub-national region where most of its area falls. We then extract population[10] weighted averages of the two variables for each ADM-2 regions.

2.4 Imputation

Even after enriching the data set with climatic and food prices information, the data set still presents a large number of missing values. Thus, we need to make proper adjustments before starting the analysis. First, we apply a re-balancing of the data set imputing missing values at the most disaggregated administrative level (ADM-2) by averaging, for each country j and year t, the non-missing ADM-2 observations at the higher administrative level (ADM-1), weighting by the share of population of the ADM-2 ($POP_{j,t}^{\text{ADM-2}}$) over the total population at the ADM-1 level ($POP_{j,t}^{\text{ADM-1}}$). Mathematically, the imputed values $\hat{\mathbf{X}}$ for variables at the ADM-2 level for

[5] Such as the Living Standard Measurement Studies from World Bank.

[6] The prices are taken from the IMF (International Monetary Fund) International Finance Statistics series and the World Bank Global Economic Monitor.

[7] We recover two shapefiles: (i) for the Sahel area from Humdata and (ii) for Sierra Leone from GADM.

[8] We employ version 2.8. Available at: https://spei.csic.es/database.html

[9] As crops planting and agricultural planning typically happens several months in advance to the actual harvest and tends to have lasting implications, we focus on the index for the calendar year preceding the year of interest for a given observation in the main data set.

[10] It is standard in the climate economics literature to weight by population climatic variables (Dell et al., 2014). We use gridded population data from Gridded Population of the World (GPW), v4 (https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count-rev11)

year t and country j are then:

$$\hat{\mathbf{X}}_{j,t}^{\text{ADM-2}} = \sum_{\text{ADM-1}} \mathbf{X}_{j,t}^{\text{ADM-2}} \cdot \frac{POP_{j,t}^{\text{ADM-2}}}{POP_{j,t}^{\text{ADM-1}}} \tag{1}$$

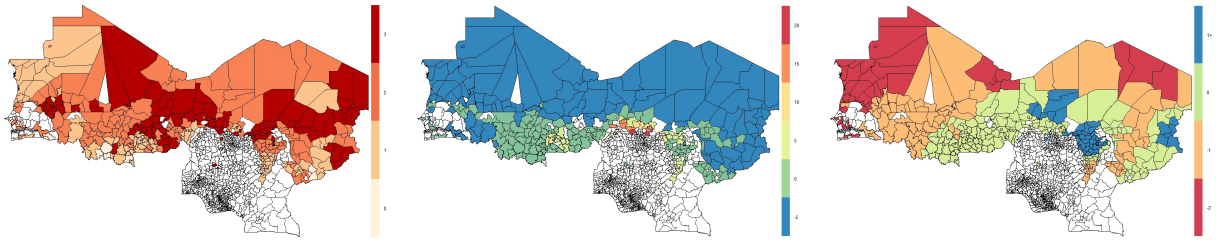
As we are still left with several missing values, we repeat the same procedure averaging at country-level. That is, that the imputed values are obtaines as follows:

$$\hat{\mathbf{X}}_{j,t}^{\text{ADM-2}} = \sum_{\text{ADM-0}} \mathbf{X}_{j,t}^{\text{ADM-2}} \cdot \frac{POP_{j,t}^{\text{ADM-2}}}{POP_{j,t}^{\text{ADM-0}}} \tag{2}$$

After this imputing procedure, we are left with 711 observations which can be effectively used for the empirical analysis. The target variable for the remaining observations are balanced across the four classes of interest, with a prevalence of people at high (3) and very high (4) level of risk. This means that if the predictive model were not performing well, the results obtained would be skewed toward a more pessimistic classification in terms of risk of malnutrition.

Figure 1 provides a visualization of some of the data features discussed in this section. Panel (1a) shows the distribution of wasting burden across the Sahel region. The missing values are concentrated in South and Eastern Nigeria, as well as in Southern Chad. Panel (1b) shows the distribution of producer prices in 2012. Unsurprisingly, the closer a region is to the Sahara desert, the lower the index is. Producer indexes are instead higher towards the Atlantic coast and spike the highest in Northern Nigeria. Moreover, we also notice that areas with high production prices tends to have high severe infant malnutrition. Panel (1c) shows the distribution of long-term soil dryness across Sub-Saharan Africa: specifically, areas with wetter soils (above average) are plotted in blue, while areas that suffered droughts are plotted in red.

Figure 1: Geographical distribution of wasting burden (2022), producer price index (2012) and SPEI 48 (2021) index in the Sahel region



(a) Wasting burden **(b)** Producer Price Index **(c)** SPEI48

3 Methodology

The purpose of this section is to assess the most effective machine learning methods to predict severe malnutrition, i.e. wasting, using all the possible meaningful information in the data we collect. Moreover, we aim at defining the covariates that explain most of the variance in severe malnutrition, and therefore they have higher influence in identifying the hot-spots of wasting in the Sahel region. Finally, we want to test how much the inclusion of food prices as covariate may improve the estimation.

3.1 Multinomial Logistic Lasso

We start our analysis by proposing a benchmark model, in particular, we focus on a standard multivariate Logit model, that is:

$$P\left(Y=r \mid \boldsymbol{x}_i\right)=\frac{\exp\left(\beta_{r0}+\boldsymbol{x}_i'\boldsymbol{\beta}_r\right)}{\sum_{s=1}^{k}\exp\left(\beta_{s0}+\boldsymbol{x}_i'\boldsymbol{\beta}_s\right)} \tag{3}$$

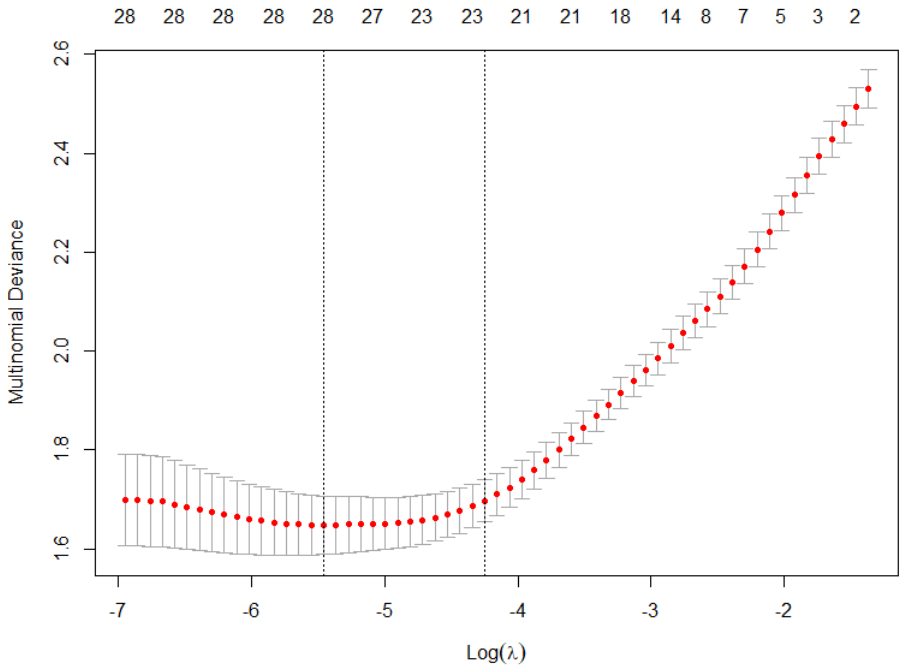
where $\mathbf{x}_i$ is the full matrix of covariates for the i-th observation for the categorical variable Y. We perform features' selection by applying a weighted group Lasso (Meier et al., 2008) to shrink the coefficients of the least-predictive covariates. The weights for the Lasso are based on the population of children aged 6 to 59 months in each administrative area (ADM-2), in order to account for potential size differences across countries. This leads to perform a penalized likelihood problem which takes the form (Tutz et al., 2015):

$$\widehat{\boldsymbol{\theta}}=\underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmax}} l_p(\boldsymbol{\theta})=\underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}}(-l(\boldsymbol{\theta})+\lambda J(\boldsymbol{\theta})) \tag{4}$$

where l_p is the log-likelihood for the multivariate logistic distribution, while $J(\boldsymbol{\theta})$ is a l_1 penalty of the form $J(\boldsymbol{\beta})=\sum_{g=1}^{G} s\left(\mathbf{df}_g\right)\left\|\boldsymbol{\beta}_g\right\|_2$.

For the feature selection we include most of the covariates available in the data set. Indeed, we first exclude any variables that directly measures malnutrition (such as "Malnutrition", "gam_prevalence" and "sam_prevalence") as they create spurious predictability with the dependent variable. Second, we keep only the disagreggated indicator from the INFORM Sahel data, dropping the aggregated indexes.[11]

Figure 2: Choice of Tuning Parameter Lambda (10-fold Cross Validation)



[11] For instance, we include as covariate "Political Violence" and "Conflict Probability", but we exclude "Human" which is simply the weighted average of the two.

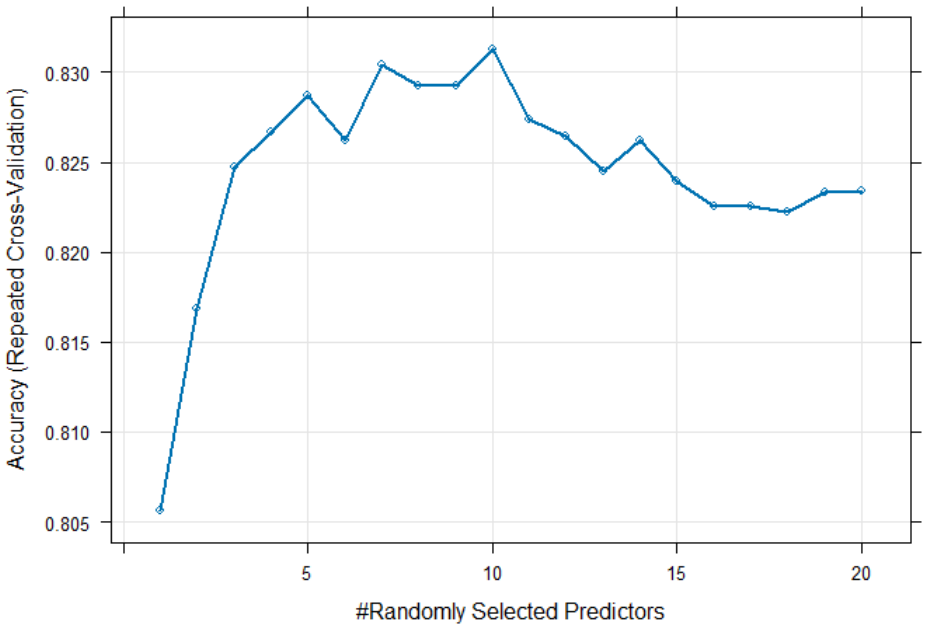
We train the model on 60% of the sample and test it on the remaining 40%, a conservative choice to mitigate the problem of over-fitting.
The tuning parameter λ is chosen by means of 10-fold cross-validation on the training set maximising the Area (AUC) under the Receiver Operating Characteristics (ROC), and based on the "one-standard error rule" (Figure 2). The reason is the one-standard error rule allows to choose the simplest model whose accuracy is comparable with the best model (Krstajic et al., 2014).
At the end of estimation, we are left with 23 predictors over the original 31. Variable selection confirms the importance of external measures of climate and price shocks, which motivates their inclusion in subsequent models, along with measures of health, nutrition and exposure to conflict and natural risk well-documented in the literature (Katona and Katona-Apte, 2008). The selected coefficients are presented in Table 2 in the Appendix.

3.2 Random Forests

As a next step, we extend our analysis with a random forest predictive algorithm. Random forests are a flexible estimation technique which improves on the previous analysis by ensembling different weak models in order to produce a more precise prediction in terms of variance.

As for Lasso, we train the model on 60% of the sample and test it on the remaining 40%. To fully exploit the variability in the data, we generate $B = 1000$ independent bootstrap samples and fit a classification tree to each of these data sets, for which only $m \ll p$ predictor are included and allow to leverage the information stemming from independent fits and minimize the variance of the model. The main tune parameter for random forest is the number of variables randomly sampled as candidates at each split and the number of trees. We determine the optimal number of predictors based on accuracy after a 10-fold cross-validation repeated 5 times.[12]

Figure 3: Optimal Number of Predictors at Each Split (10-fold Cross Validation)



[12]Note that we can also tune the number of trees. However, as it is computationally intensive, we prefer to focus only the number of features at each split.

Figure 3 shows the number of predictors in the model as a function of accuracy. Ultimately, we select the number of predictors which maximize accuracy, that amounts to $m = 10$ predictors. The performance of the model is ultimately assessed on a test set consisting of 40% of the original data.

3.3 Extreme Gradient Boosting

Random forests should help to a sensible improve over the first benchmark specification, increasing the accuracy of the model. This is particularly relevant in our setting, where we have a large number of highly correlating covariates which could bias our conclusion. However, we can further boost accuracy while minimizing the variance of the model through a boosting classification algorithm to further exploit the complexity of the data. Similarly to random forests, boosting algorithm are ensemble averages of weak learners, i.e. trees, which are aggregated to reduce the variance of the model. The main difference with respect to random forests is that the boosting algorithm works consequentially, with each tree learning from the residuals of the previous ones. This feature is particularly attractive in our setting, where we work with spatial data where complexity and non-linearities challenge features' extraction. In our framework, we employ the eXtreme Gradient Boosting algorithm, where the sequence of trees is derived by minimizing an objective function as a function of the input classes. Given the multiclass classification problem of interest, the four classes of interest are assigned a normalized probability distribution through a softmax function of the form:

$$\sigma(\mathbf{z})_i \;=\; \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{5}$$

where $\mathbf{z}$ is the input vector of interest for the $i = 1, \ldots, 4$ classes. Using these softmax probabilities, the algorithm minimizes the cross-entropy loss:

$$L_{\text{CE}} \;=\; -\sum_{i=1}^{n} y_i \log\left(\sigma_i\right) \tag{6}$$

The algorithm minimizes the objective function by means of gradient-descent optimization improving the classification error as the number of rounds of minimization increases.

To tune the great amount of hyperparameters[13] used in XGBoosting, we run 5-fold cross validation 100 times on the training set (60% of the original data set), each time with random parameters.

The results obtained with the optimal hyperparameters are then evaluated on a test set of 40% of the original data set.

[13] This includes the number of rounds, the step size shrinkage used in update to prevents overfitting, the maximum depth of a tree, the inimum sum of instance weight (hessian) needed in a child, the maximum delta step we allow each leaf output to be, the best seed, and so on.

4 Results

4.1 Model Perfomances

We now turn to the comparison of the models proposed in the previous section.

Table 1: Performance of the Prediction of Severe Malnutrition on the Test Set - Comparisons across Machine Learning Methods

Metrics	Wasting Class	Lasso (1)	Random Forest (2)	XGBoost (3)
Accuracy		0.593	0.800	0.835
Cohen's K		0.406	0.711	0.748
Sensitivity				
	Low	0.368	0.632	0.818
	Medium	0.690	0.798	0.915
	High	0.561	0.785	0.735
	Very High	0.587	0.867	0.844
Specificity				
	Low	1.000	0.996	0.996
	Medium	0.766	0.901	0.896
	High	0.702	0.871	0.907
	Very High	0.924	0.938	0.943
AUC				
	Low	-	-	-
	Medium	0.959	0.989	.
	High	0.836	0.949	.
	Very High	0.701	0.908	.

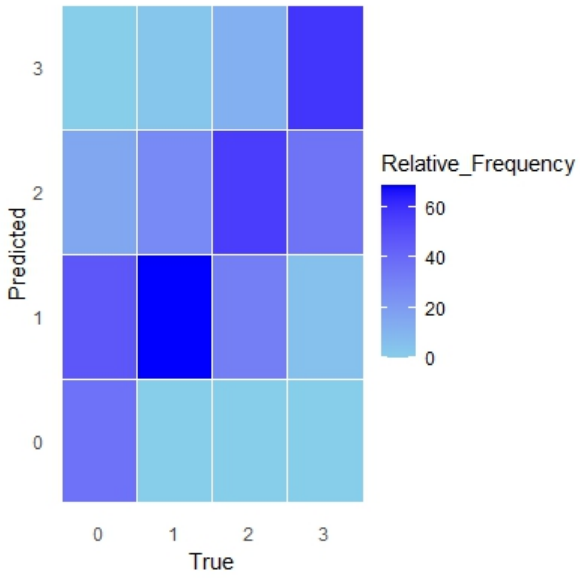
Notes: AUC stays for Area under the Receiving Operating Characteristic (ROC) curve. For AUC the reference category is "Low" Wasting.

Table 1 shows several performance measures compared across the three models of interest, while Figure 4 displays the confusion matrices for the three models.

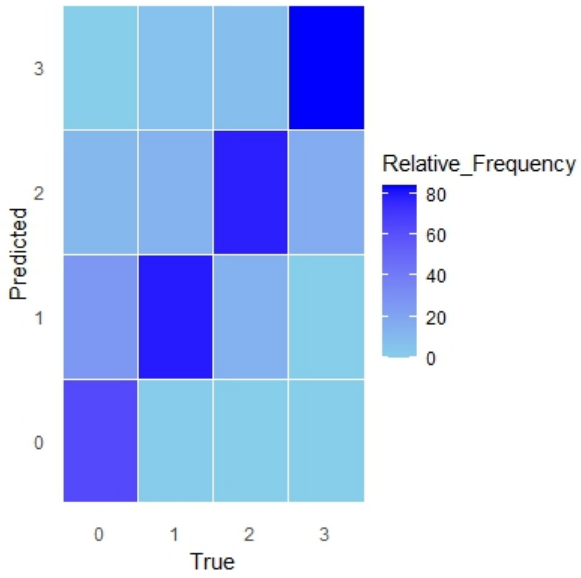
As expected, the benchmark model is not suitable for prediction, with an overall accuracy of 59%. The limitations of this approach, which motivate further analysis, can be seen by looking at the sensitivity of the prediction: much of the areas that are affected by high or very high malnutrition status are incorrectly captured by the algorithm. Looking at the confusion matrix, we can see that there is a lot of error between similar classes, so that administrative areas with severe malnutrition are incorrectly classified as ones with a lower malnutrition level, and the same holds for lower classes. However, there is also a lot of error between with high and medium malnutrition areas, with misclassification in both directions. This is not satisfactory for two reasons: first, we do not want to underestimate the burden of wasting, since it should be the first target

for policymakers; second, we do not want to overestimate the malnutrition status of children either, since this could lead to a misallocation of resources, which can be particularly scarce.

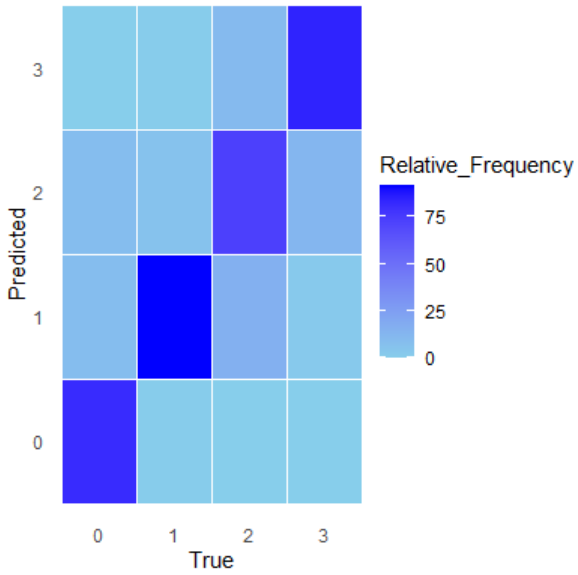
Figure 4: Confusion matrices: Lasso, Random Forests and eXtreme Gradient Boosting



(a) Lasso



(b) Random Forests



(c) XGBoost

These problem are sensibly mitigated when looking at the results obtained with random forests, as can be seen in panel 4c. Aside from an important gain in overall accuracy, we can see how the algorithm is able to consistently disentangle the areas with severe malnutrition from others, and to mitigate the variability of the prediction for the highest categories. The improvement for said groups can be seen looking at the AUC of the models in Table 1, which is a the broadest indicator of accuracy of the model: Random Forests improve on the naive benchmark with a $\approx$ 10 p.p. increase in the AUC for areas with a high level malnutrition and a $\approx$ 20 p.p. increase in the AUC for areas with severe malnutrition. Finally, we turn to our main estimation strategy, based on

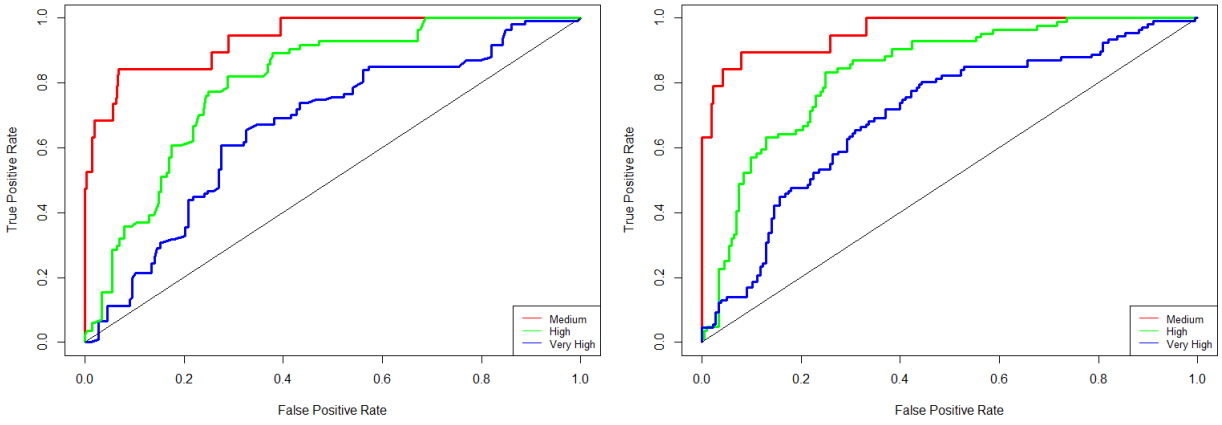
the eXtreme Gradient Boosting algorithm. Overall, the algorithm provides the most satisfying results in terms of accuracy, with an improvement of 3.5 p.p. over Random Forests. The rationale for this further implementation can be seen looking at the sensitivity and specificity of the model. If the sensitivity is almost unchanged, or slightly lower for areas with a high or very high level of wasting, there is a sensible improvement in the identification of medium-low levels of malnutrition, with a considerable increase in the sensitivity for these two categories. We argue that this is a non-negligible feature for a model which aims at predicting severe levels of wasting, correctly identifying people at risk and finally targeting mitigation policies. Furthermore, the XGBoost algorithm is fully calibrated by means of cross-validation techniques: indeed, each parameter of the model is selected through extensive cross-validation exercises, which improves the stability and out-of-sample performance of the algorithm.

Overall, the main prediction strategy, based on a fully cross-validated XGBoost algorithm, has proven to effectively predict the burden level for wasting in children across two dimensions: (i) identifying areas with high and very high malnutrition levels, which should be the main target for policymakers and (ii) identifying areas with an overall better health environment, which can allow organizations to consistently better allocate resources, by prioritizing areas with a higher level of malnutrition.

4.2 Including Food Prices and Climate Information

We test whether the inclusion of local producer and consumer food prices, as well of the SPEI climatic variables, helps enhancing the prediction. Figure 5 shows the two Receiving Operating Characteristic (ROC) obtained from the LASSO classificaiton for the three categories excluding (Panel 5a) and including (Panel 5b) these variables. We find a substantial improvement in the AUC. For the very high wasting category, the AUC moves from 0.66 to 0.70, whereas for the high category the AUC increases by 0.03 including the variables.

Figure 5: Receiving Operating Characteristic (ROC) with and without local food prices and SPEI variables - LASSO

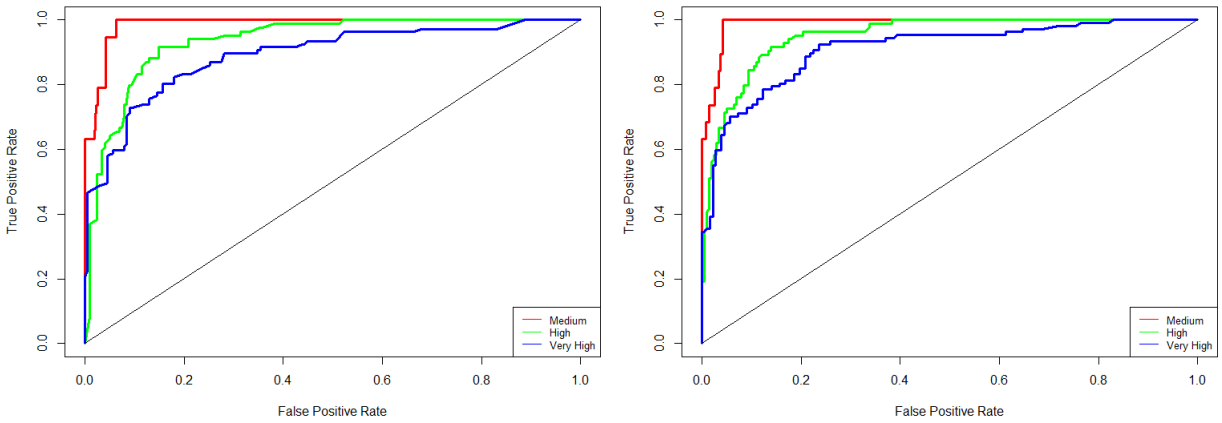


(a) Without Food Prices and SPEI

(b) With Food Prices and SPEI

We find similar results with the Random Forest (Figure 6). For each category the AUC is always greater when we include food prices and SPEI.

Figure 6: Receiving Operating Characteristic (ROC) with and without local food prices and SPEI variables - Random Forest



(a) Without Food Prices and SPEI

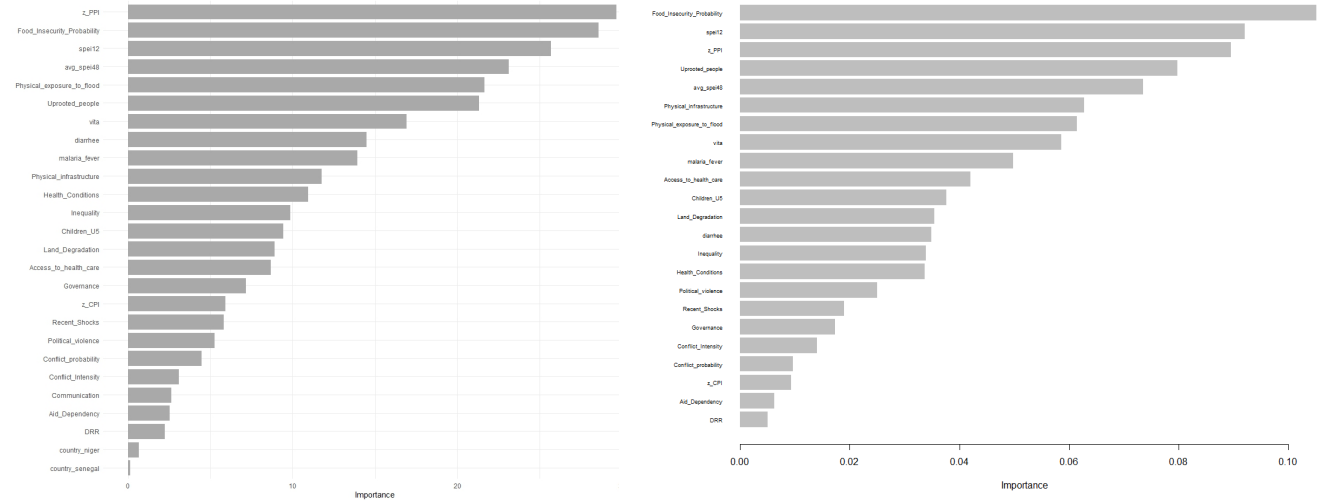
(b) With Food Prices and SPEI

Overall, we advocate that the inclusion of these additional variables is key for the increasing predictive power of the model. However, to have further assurance on this side, we need to determine the relevance importance of each variable in our data set.

4.3 Relative Importance

The proposed estimation strategy allows us to disentangle the relative importance of the predictors in the Random Forest and eXtreme Gradient Boosting algorithm, which greatly informs about the relevance of each variable in predicting malnutrition.

Figure 7: Relative importance of coefficients, Random Forest and XGBoost



(a) Random Forest

(b) XGBoost

Figure 7 shows the relative importance of each predictor in the Random Forest and XGBoost algorithms. The variables are presented in order of decreasing importance. First, we can notice how both the producer price index z_PPI and the SPEI are indeed important predictors of

wasting, while a less important role is played by consumer prices, z_CPI. The findings highlight the importance of these variable in predicting child malnutrition, and therefore motivate their inclusion in the dataset. These findings are coherent with the literature on the effects of climate shocks on local populations and with the literature on the role of global price shocks on local populations, especially those with a high share of people exposure to risk (Bloem et al., 2009; Vellakkal et al., 2015). Furthermore, we confirm the importance of well-established predictors in the health literature, such as diahrrea and health conditions (Katona and Katona-Apte, 2008), along with other socio-political indicators which capture the instability of local areas (Kalu and Etim, 2018).

5 Discussion and conclusion

We propose an effective multiclass classification models based on machine-learning tools to predict the presence of severe malnutrition in Sub-Saharan Africa. Using appropriate features' selection procedure, we are able to identify the key predictors of severe child malnutrition. The best model, both in terms of accuracy and true-positive rates, is based on an eXtreme Gradient Boosting procedure applied on a relevant subset of predictors, providing an accuracy of 84% along with information on the contribution of each of the variables of interest. Crucially, we show that the inclusion of local food prices and climate indexes significantly increases the accuracy of our model.

We argue that the inclusion of these variables provides a significant improvement in terms of predicting and hence preventing malnutrition. Specifically, we identify some main advantages of our methodology. First, given that the price index proposed by McGuirk and Burke (2020) uses global price trends of agricultural goods, forecasts of these trends are easily accessible and can be therefore used for policy. Under the fair assumption of a slowly-changing agricultural structure, global price trends can be effectively used to infer areas particularly exposed to children malnutrition. Secondly, we have shown that medium-run trends of climate risk also can inform in advance about areas which could be exposed to malnutrition problem. Lastly, given the often uncertain nature of the political and institutional environment in Africa, we argue that the ability to rely on gridded, geographically coded and easily measurable variables for prediction constitutes an important improvement.

Our work presents important limitations. The main drawback of the model is related to data quality and availability. The lack of local variability in several predictors proved to be an important obstacle in the prediction of children malnutrition, and it required us to work with imputed data, which rely on more aggregate information. The imputation process may introduce measurement error and bias our estimates since lead to the loss of fundamental local dynamics. Furthermore, we employ price indexes which are not updated with the latest price trends. They only serve as a proxy of current price trends. An immediate extension would be to build a novel index based on more recent price series, also controlling for potential changes in the local agricultural shares. Finally, we need to acknowledge the complete lack of causality in the estimation framework. While the algorithms provided are useful to identify key predictors of children malnutrition and to have an initial indication of where policies needs to be targeted, there is no clear answer on which policies should be preferred or which channels should be prioritized for prevention. In this sense, the large recent literature on tackling causality in machine-learning models (Pearl, 2019; Prosperi et al., 2020) can further improve the analysis by offering guidance on the causal drivers of malnutrition, which should be prioritized by policymakers.

References

Bloem, M. W., Semba, R. D., and Kraemer, K. (2009). Castel Gandolfo Workshop: An Introduction to the Impact of Climate Change, the Economic Crisis, and the Increase in the Food Prices on Malnutrition. *The Journal of Nutrition*, 140(1):132S–135S.

Bulti, A., Briend, A., Dale, N. M., De Wagt, A., Chiwile, F., Chitekwe, S., Isokpunwu, C., and Myatt, M. (2017). Improving estimates of the burden of severe acute malnutrition and predictions of caseload for programs treating severe acute malnutrition: experiences from nigeria. *Archives of Public Health*, 75:1–8.

Cornia, G. A., Deotti, L., and Sassi, M. (2016). Sources of food price volatility and child malnutrition in niger and malawi. *Food Policy*, 60:20–30. Towards a food secure future: Ensuring food security for sustainable human development in Sub-Saharan Africa.

Cudjoe, G., Breisinger, C., and Diao, X. (2010). Local impacts of a global crisis: Food price transmission, consumer welfare and poverty in ghana. *Food Policy*, 35(4):294–302.

Dale, N. M., Myatt, M., Prudhon, C., and Briend, A. (2017). Using cross-sectional surveys to estimate the number of severely malnourished children needing to be enrolled in specific treatment programmes. *Public health nutrition*, 20(8):1362–1366.

Dell, M., Jones, B. F., and Olken, B. A. (2014). What do we learn from the weather? the new climate-economy literature. *Journal of Economic literature*, 52(3):740–798.

Hilson, G. (2016). Farming, small-scale mining and rural livelihoods in sub-saharan africa: A critical overview. *The Extractive Industries and Society*, 3(2):547–563.

Isanaka, S., Andersen, C. T., Cousens, S., Myatt, M., Briend, A., Krasevec, J., Hayashi, C., Mayberry, A., Mwirigi, L., and Guerrero, S. (2021). Improving estimates of the burden of severe wasting: analysis of secondary prevalence and incidence data from 352 sites. *BMJ Global Health*, 6(3):e004342.

Kalu, R. and Etim, K. (2018). Factors associated with malnutrition among underfive children in developing countries: A review. *Global Journal of Pure and Applied Sciences*, 24(1):69–74.

Katona, P. and Katona-Apte, J. (2008). The interaction between nutrition and infection. *Clinical Infectious Diseases*, 46(10):1582–1588.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6:1–15.

McDonald, C. M., Manji, K. P., Kupka, R., Bellinger, D. C., Spiegelman, D., Kisenge, R., Msamanga, G., Fawzi, W. W., and Duggan, C. P. (2013). Stunting and wasting are associated with poorer psychomotor and mental development in hiv-exposed tanzanian infants. *The Journal of nutrition*, 143(2):204–214.

McGuirk, E. and Burke, M. (2020). The economic origins of conflict in africa. *Journal of Political Economy*, 128(10):3940–3997.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1).

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60.

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., and Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375.

Tutz, G., Pößnecker, W., and Uhlmann, L. (2015). Variable selection in general multinomial logit models. *Computational Statistics Data Analysis*, 82:207–222.

Vellakkal, S., Fledderjohann, J., Basu, S., Agrawal, S., Ebrahim, S., Campbell, O., Doyle, P., and Stuckler, D. (2015). Food price spikes are associated with increased malnutrition among children in andhra pradesh, india. *The Journal of nutrition*, 145(8):1942–1949.

Victora, C. G., Adair, L., Fall, C., Hallal, P. C., Martorell, R., Richter, L., and Sachdev, H. S. (2008). Maternal and child undernutrition: consequences for adult health and human capital. *The lancet*, 371(9609):340–357.

Appendix

Table 2: Variables from Lasso selection

	variables	cat0	cat1	cat2	cat3
1	(Intercept)	2.352	-0.786	-0.127	-1.438
2	Access_to_health_care	-0.081	-0.075	0.203	-0.047
3	Aid_Dependency	-0.123	0.021	0.195	-0.093
4	avg_spei48	0.173	-0.670	0.189	0.308
5	Children_U5	-0.037	-0.109	-0.046	0.193
6	Conflict_Intensity	-0.082	0.093	0.014	-0.024
7	country_niger	-0.069	-0.001	0.034	0.036
8	country_senegal	1.631	-0.297	-0.262	-1.072
9	diarrhee	0.003	-0.001	0.006	-0.008
10	Food_Insecurity_Probability	-0.582	-0.498	0.293	0.788
11	Health_Conditions	-0.083	0.271	-0.022	-0.167
12	Inequality	0.423	0.487	-0.369	-0.541
13	Land_Degradation	-0.171	0.184	0.042	-0.056
14	malaria_fever	0.000	0.003	-0.002	-0.001
15	Physical_exposure_to_flood	-0.232	-0.075	0.149	0.158
16	Physical_infrastructure	-0.133	-0.002	0.020	0.115
17	Political_violence	0.009	-0.052	-0.090	0.133
18	Recent_Shocks	-0.056	-0.058	0.053	0.061
19	spei12	0.829	0.322	-0.564	-0.587
20	Uprooted_people	-0.053	-0.144	0.061	0.136
21	vita	-0.026	0.015	-0.006	0.017
22	z_CPI	0.444	-0.351	-0.019	-0.074
23	z_PPI	-0.000	0.013	0.012	-0.025