Predicting food insecurity in Chad with machine learning: a granular data approach

De Alberti, Lissona, Pagliero, Pavanello [*]

April 21, 2023

Abstract

Food insecurity is one of the major threats to development, mainly in vulnerable regions in sub-Saharan Africa. Predicting the areas which are mostly affected by food insecurity is crucial to target appropriate policy interventions, to ultimately achieve the goal of ending world hunger. This paper proposes a multiclass classification model to predict food insecurity at the sub-national level in Chad, using gridded-level information on socio-economic, climate and health conditions from 2014 to 2021. We find that eXtreme Gradient Boosting provides the best performance, when evaluated against competing methods. We evaluate the model across three time different periods. First, we extract the training and test samples from the full dataset. Second, we train the model with data up to 2018 to predict food insecurity in 2019, net of the influence of COVID-19. Lastly, we train the model with data up to 2019 to predict food insecurity in 2020 and 2021. We find that the XGBoost algorithm mitigates the impact of the COVID shock in prediction. Finally, we show that the COVID shock did not affect the structural relationship between key predictors and food instability, which is strongly driven by climatic events.

[*]Department of Economics, University of Bologna

1 Introduction

According to the latest UN estimates[1], food insecurity is a widespread problem around the world, which particularly affects poor and vulnerable countries. The sub-Saharan region of Africa has the highest levels of food insecurity in the world, which has been further exacerbated by the COVID19 pandemic (Pereira and Oliveira, 2020) and the subsequent global supply chain disruption (Nekmahmud, 2022). Accurately predicting areas highly affected by food insecurity is then crucial to ensure targeted policy interventions, to efficiently allocate resources and to progress toward the goal of ending world hunger by 2030[2].

This paper proposes a reliable classification algorithm to predict areas which are most affected by food insecurity in Chad. We tackle the issue of identifying the most important drivers of this phenomenon by employing granular-level data, providing policy makers with better information to design food provision and malnutrition relief policies. We show that our proposed procedure mitigates the impact of the COVID shock on the performance of the model, without altering the structural relationship between key predictors and food insecurity.

To do so, we first build a data set at the sub-national level (ADM-2) for Chad, covering the years from 2014 to 2021. We first collect data on food insecurity at the ADM-2 level from the Cadre Harmonise data set. Then, we augment the accuracy and quality of the data set with high quality geo-referenced gridded-level information on climatic conditions, health conditions, socio-political instability and economic distress. Specifically, for climatic conditions we use a well-know index of measuring soil dryness and wetness, the Standardized Precipitation Evapotranspiration Index (SPEI). We compute a 3-month, 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. We so capture key determinants of food insecurity through disruption in local agricultural production (De Haen and Hemrich, 2007), such as droughts, floods and land degradation. Furthermore, we employ data on precipitation and temperature to better describe local climatic conditions. We also include information on the number of hospitals per areas to capture health access and quality, as well as the intensity of conflicts to measure violence. As economic predictors, we use gross domestic product (GDP) to control for the economic conditions in each sub-national region, and we get consumer and producer food price indexes from McGuirk and Burke (2020) to identify the role of local food price shocks. Ultimately, we include a monthly index of global supply chain distress build by the Federal Reserve in order to capture global shocks which may have spillovers at the local level.

Second, we compare two competing machine learning model to assess which is more suitable for prediction of food insecurity, based on the level of accuracy. Initially, we perform a preliminary analysis through a multivariate logit model along with a group-Lasso, weighted for the population at the ADM-2 level, to identify the key features for prediction, using the full dataset available. We extend our analysis by applying an eXtreme Gradient Boosting (XGBoosting) algorithm to improve the predictive power of our model. The 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 76%. We then repeat the exercise to perform two additional predictions: (i) a pre-COVID prediction, where we train the model with data up to 2018 to predict food insecurity in 2019, (ii) a post-COVID prediction, where the model is trained with data up to 2019 to predict food insecurity in 2020 and 2021. We find that, compared with the group-Lasso approach, the proposed XGBoost algorithm is better suited to account for the COVID shock in

[1] The sustainable development goals report 2021. Available at https://unstats.un.org/sdgs/report/2021/extended-report/ Goal%20(2)_final.pdf.

[2] Target 2.1 of UN goals. Report available at https://sdgs.un.org/goals/goal2

prediction. Ultimately, we provide evidence of a strong role of climatic factors in predicting food insecurity.

Our paper contributes to the literature on how to predict food insecurity (Dale et al., 2017; Nica-Avram et al., 2020) and on the identification of the key drivers for early detection of areas with a high incidence of food insecurity (Hansen et al., 2022; Mason-D'Croz et al., 2019; Pereira and Oliveira, 2020). From one stand point, we propose a machine learning multiclass classification algorithm which outperforms standard statistical techniques in predicting food insecurity. Differently from other works, we also employ granular data to build covariates, which have proven to be important drivers of food instability, related to socio-economic factors, health status, and climatic conditions in the local areas. We argue that the choice to employ granular data to build the predictors is a key driver of our results, for at least three reasons. First, granular data are often built by international organization employing cutting-edge techniques, so to have an accuracy which is often much better than competing reported or survey-based measures. Second, working with aggregate data we would need to downscale predictors at the local level by reweighting, which could introduce measurement errors, biasing our results. Last, aggregate data include several missing values, whose imputation would further increase the uncertainty of our estimates. We provide evidence for the effectiveness of the predictive power of the algorithm when accounting for the COVID period, compared to competing models. Ultimately, in line with the recent literature (Hansen et al., 2022), we confirm the strong role of climate factors for the prediction of food insecurity,

The rest of the paper is structured as follows: Section 2 introduces the data used in the analysis and explains the aggregation process performed to build the covariates. Section 3 presents the employed methodologies, with insights on the importance of both the algorithm choice and the selection of covariates. 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 enrich data quality and availability, we employ granular-level spatial data to build the covariates which will be used by the algorithms. Given the quality of the data, the variable are closely representative of local dynamics, avoiding any measurement error which would derive by down-scaling available indicators at more aggregate levels.

2.1 Food Insecurity Data

Our main source for the analysis is a data set on Administrative Level-2 (ADM-2) sub-national regions for Chad covering the years from 2014 to 2021. We collect this data from the Cadre Harmonise Data (CH), which provides food insecurity figures for each administrative sub-national areas (ADM-2) for three reference periods (January to May, June to August, September to December). We implement it with a country-specific public health facility database taken from Maina et al. (2019), that was developed through a systematic and iterative process of data assembly and the

Our main variable of interest is a categorical that describes the overall level of food insecurity in each ADMIN-2 administrative region. The values for the levels of food insecurity are in increasing order of distress and are coded as follows: 1 is ”minimal”, 2 is ”stressed” and 3 is ”crisis”. At this stage, the data set consists of 3,094 observations, covering the period 2014-2021.

2.2 Consumer and Producer Food Prices Indexes

Relevant determinants of food insecurity, and so calories' intake, 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).

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[3] in the country of interest. In this sense, our initial data set contains information on prices of specific crops. However, these measures present two issues. Firstly, there are a lot of missing values. Secondly, these prices do not allow to discriminate between consumer and producer side price pressures, which can quite plausibly affect food security through different channels. For these reasons we recover gridded-level food price data from McGuirk and Burke (2020). They construct granular ($0.5^\circ \times 0.5^\circ$) 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[4] 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 2004-2011. 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. Our price indexes then vary across sub-national regions (ADM-2) and yearly.

2.3 Climate Data

Next, we recover spatial data on environmental and climatic variables. First, we collect a well-known climatic measure, the Standardized Precipitation Evapotranspiration Index (SPEI) from the Global SPEI Dataset[5], 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 three variables: *spei03*, *spei12*, and *spei48*, 3, 12 and 48 moving averages respectively. We then average these three variables over the months of interest, based on the *reference_code* variable. These variables allow to capture both long-run (land deterioration through absence of groundwater), short-run and seasonal (extreme events) consequences of absence and abundance of precipitation.

Second, we obtain average monthly temperatures and cumulative precipitations data at 0.5° resolution from the Climate Research Unit (CRU TS v4.07) of the University of East Anglia (Harris et al., 2014).

Similar to food prices, we match each grid cell with the sub-national region where most of its area falls. We then extract population[6] weighted averages of all the climatic variables for each

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

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

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

[6] 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)

ADM-2 regions. Overall, our climatic information will vary both spatially (ADM-2) and across reference periods.

2.4 Conflict

Conflicts are commonly identified as one of the major contributor in poor food security condition since it affects farming, livelihood and markets, as well as being crucial in limiting international support (Africa Center for Strategic Studies, 2022). The Uppsala Conflict Data Program[7] (UCDP) provides geocoded information on organized violence and armed conflict. We gather this data to recover two variables: a dummy, which takes value 1 if any conflict took place within the area of interest, and a variable for conflict intensity (the number of casualties), which takes into account the impact of a given conflict. Thanks to the geocoded and time information provided for each conflict, we are able to construct these measures at the subnational and reference period level.

2.5 Gross Domestic Product

We obtain granular information on Gross Domestic Product (GDP) from Murakami and Yamagata (2019). They provide downscaled estimates gross domestic product (GDP) into 0.5-degree grids for each decade, for the periods 2010-2100. Their downscaling approach has the nice property of well capturing the difference between urban and non-urban areas. We use the gridded information from 2010 to 2020, and we employ an exponential imputation to determine the level of GDP at each cell in each year for the period 2014-2021. We then apply the usual methodology of extracting the spatial information at the ADM-2 level. Our measure of GDP then vary across sub-national regions (ADM-2) and yearly.

2.6 Health facilities

An important missing feature from the initial data set are information on health access in Chad. We gather data on the geographical distribution of hospitals within Chad from Maina et al. (2019). The authors build the data set through a variety of sources, such as the Ministry of Health (MoH) website or the United Nations Office for the Coordination of Humanitarian Affairs' (UNOCHA) Humanitarian Data Exchange (HDX) portal. Similarly to the UCDP dataset, we are provided with the geocode information of each hospital. We can so construct two variables as proxies of the access to healthcare at the sub-national level (ADM-2): i) a measure of the intensity of hospital and ii) a dummy for the presence of a hospital. Overall, our health variables vary only spatially across sub-national regions (ADM-2).

2.7 Supply Chain Distress

Following Akinci et al. (2023), supply factors are some of the main drivers of the global increasing inflation after the COVID-19 period. To tackle this indicators, the Federal Reserve Bank of New York built the Global Supply Chain Price Index with measures of chain-related costs from manufacturing firms through data from Baltic Dry Index (BDI), the Harpex index, as well as the U.S. Bureau of Labor Statistics and the Purchasing Managers' Index (PMI) surveys.[8] This may be useful for our purpose since it captures food provision issue related to global shocks that affect food security in a particular sub-national level.

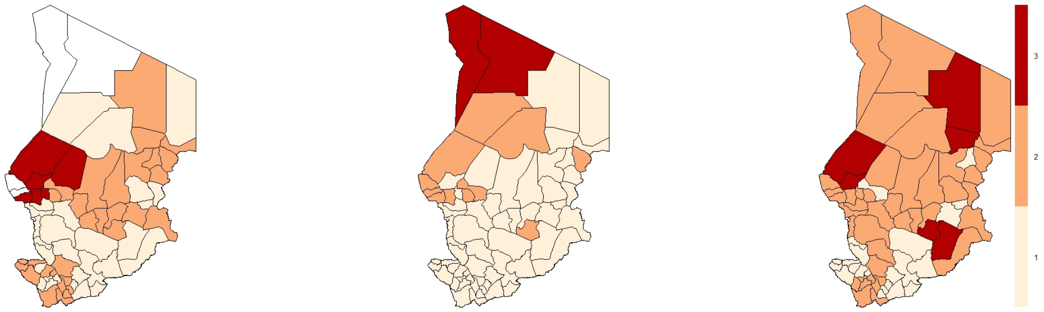
[7] Available at: https://ucdp.uu.se/

[8] Available at: https://www.newyorkfed.org/research/gscpi.html

2.8 Descriptive Analysis

Figure 1 provides the evolution of our main variable over time, describing the trend of food insecurity in Chad for the years 2015, 2018 and 2021 in the last period of the year, namely from September to December. The distribution is quite diverse across the three years. In 2015 (1a) half of the region can be considered under food insecurity distress, since it belongs to class two or higher. Despite showing a partial improvement in 2018, the map tells a worrying story for 2021, as conditions have worsened significantly for the whole of Chad, with new food insecurity hot-spots emerging in the South of the country. As documented by the literature (Pereira and Oliveira, 2020), food insecurity experienced a dramatic increase as a consequence of the COVID19 pandemics. Panel (1c) summarises this result showing that the majority of the administrative areas are now classified as a level 2 or above, while only few regions are in the level 1 class.

Figure 1: Geographical evolution of food security at ADM-2 level



(a) 2015 **(b)** 2018 **(c)** 2021

3 Methodology

The purpose of this section is to assess the most effective machine learning methods to predict areas with high food insecurity, 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 food insecurity, and therefore they have higher influence in identifying the hot-spots with the highest threat to food security. Each method is applied on three different samples. First, we employ the full dataset to extract the train and the test sample to perform evaluation. Second, we train the model on the sample up to 2018 to predict food insecurity for 2019, to net out the effects of COVID19. Last, we use data up to 2019 as the train set to test the model on data for 2020 and 2021, to asses the accuracy of the model in predicting food insecurity in the post-COVID period.

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:

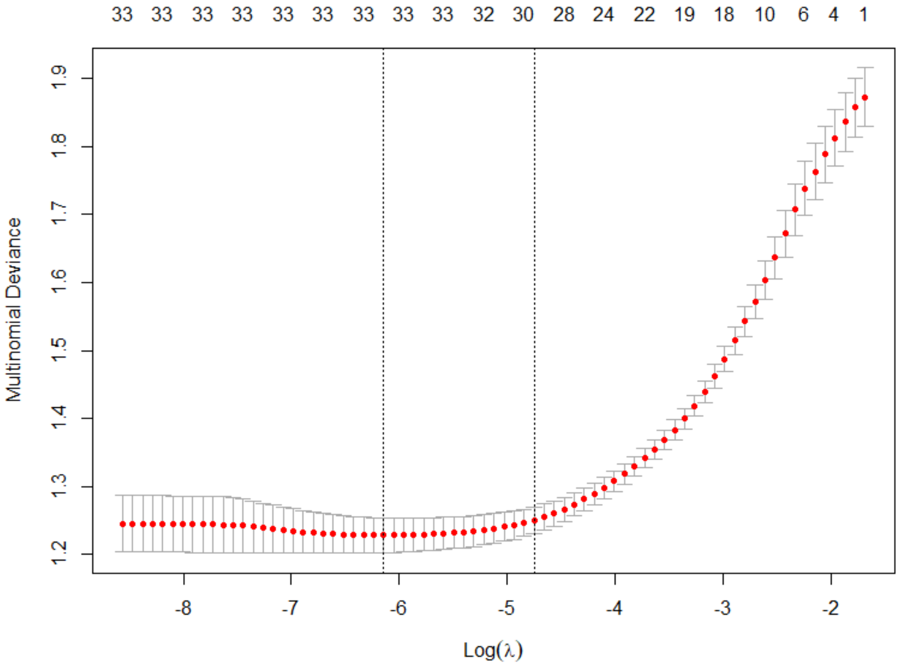
$$P\left(Y = r \mid \boldsymbol{x}_i\right) = \frac{\exp\left(\beta_{r0} + \boldsymbol{x}_i'\boldsymbol{\beta}_r\right)}{\sum_{s=1}^{k} \exp\left(\beta_{s0} + \boldsymbol{x}_i'\boldsymbol{\beta}_s\right)} \tag{1}$$

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 in each administrative area (ADM-2), in order to account for potential size differences across counties. This leads to perform a penalized likelihood problem which takes the form (Tutz et al., 2015):

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmax}} l_p(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}}(-l(\boldsymbol{\theta}) + \lambda J(\boldsymbol{\theta})) \tag{2}$$

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

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



In the full sample case, 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 **28** predictors. 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 and exposure to conflict and natural risk well-documented in the literature (Katona and Katona-Apte, 2008). The selected coefficients for the total sample are presented in Table 3 in the Appendix.

3.2 Extreme Gradient Boosting

As a next step, we extend our analysis with an eXtreme Gradient Boosting (XGBoost) algorithm, a flexible technique which improves on the previous analysis by ensembling different weak models in order to produce a more precise prediction in terms of variance, thereby improving over the first benchmark specification, increasing the accuracy of the model. Boosting algorithms are ensemble averages of weak learners, i.e. trees, which are aggregated to reduce the variance of the model. An alternative to boosting would be to apply random forests. However, in a companion paper[9], we show that XGBoost proves to be more accurate in a similar setting, which may be imputed to the advantage that the boosting algorithm works sequentially, 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 XGBoost 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:

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{3}$$

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:

$$L_{\mathrm{CE}} = -\sum_{i=1}^{n} y_i \log(\sigma_i) \tag{4}$$

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[10] 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.

[9] Case A

[10] This includes the number of rounds, the step size shrinkage used in update to prevents overfitting, the maximum depth of a tree, the minimum 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 across different choices for training and test samples. Table 1 shows several performance measures compared across the three models of interest.

Table 1: Performance of the Prediction of Food Insecurity - Comparisons across Lasso training and test sets

Metrics	Food insecurity	Lasso: total sample (1)	Lasso: pre-COVID (2)	Lasso: post-COVID (3)
Accuracy		0.67	0.56	0.37
Cohen's K		0.43	0.23	0.03
Sensitivity				
	Minimal	0.64	0.66	0.98
	Stressed	0.76	0.49	0.00
	Crisis	0.42	0.91	0.15
Specificity				
	Minimal	0.83	0.71	0.04
	Stressed	0.61	0.67	0.99
	Crisis	0.98	0.89	0.99
AUC				
	Minimal	-	-	-
	Stressed	0.83	0.72	0.50
	Crisis	0.74	0.58	0.50

Notes: AUC stays for Area under the Receiving Operating Characteristic (ROC) curve. For AUC the reference category is "Minimal" Food Insecurity.
Pre-COVID refers to the period 2014-2018, with 2019 as a test set.

As can be seen in column (1), the benchmark model is not very well suited for prediction, with an overall accuracy of 67%. The limitations of this approach emerge from sensitivity results: much of the areas that are affected by "crisis" level of food insecurity are incorrectly captured by the algorithm. This is not an encouraging result for designing policy intervention.

A crucial take-away from this exercise are the differences in performance when changing the training and test sample for our prediction algorithm appear glaring (Columns (2) and (3)). When we focus on pre-COVID data the predictive power of the model drops, as we around 20% of the total observations. As per column (3) of Table 1, almost all measures are worse, with the accuracy dropping as low as 37%. This suggests that Lasso does not perform when a huge shock like COVID19 highly changes the context (test set) where the prediction has to be conducted.

Table 2: Performance of the Prediction of Food Insecurity - Comparisons across XGBoost training and test sets

Metrics	Food insecurity	XGB: total sample (1)	XGB: pre-COVID (2)	XGB: post-COVID (3)
Accuracy		0.76	0.60	0.57
Cohen's K		0.60	0.24	0.28
Sensitivity				
	Minimal	0.79	0.71	0.89
	Stressed	0.75	0.47	0.44
	Crisis	0.70	0.27	0.23
Specificity				
	Minimal	0.82	0.58	0.54
	Stressed	0.78	0.71	0.74
	Crisis	0.98	0.96	0.99

Notes: Pre-COVID refers to the period 2014-2018, with 2019 as a test set.

Table 2 displays the performance of the XGBoost algorithm across our different specifications. Overall, we notice that the eXtreme Gradient Boosting performs much better than Lasso over all the three choices for training and test sets, with an accuracy of 76% in the full sample estimate. This supports our choice to employ such an algorithm for classification, as our preferred specification performs quite well in predicting food insecurity.

Aside from an important gain in overall accuracy, we can see how the algorithm is able to consistently disentangle the areas with ”crisis” level of food insecurity, which are the most crucial to identify. An additional rationale for this further implementation can be seen looking at the sensitivity and specificity of the model for the same category. If the specificity is almost unchanged with respect to the Lasso algorithm, there is a sizeable improvement in sensitivity. We argue that this is a non-negligible feature for a model which aims at predicting critical levels of food insecurity, correctly identifying areas at risk and finally targeting mitigation policies.

Additionally, the concerns about using pre-COVID data to predict post-COVID outcomes are mitigated. As can be seen by comparing columns (2) and (3), the difference in performance is now much smaller, with accuracy dropping of only 3%. This suggests the predictive power of XGBoosting, despite the likely presence of structural breaks due to the pandemic shock.

4.2 Relative Importance

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

Figure 3: Top-5 Predictors by Relative Importance - Total sample, XGBoost

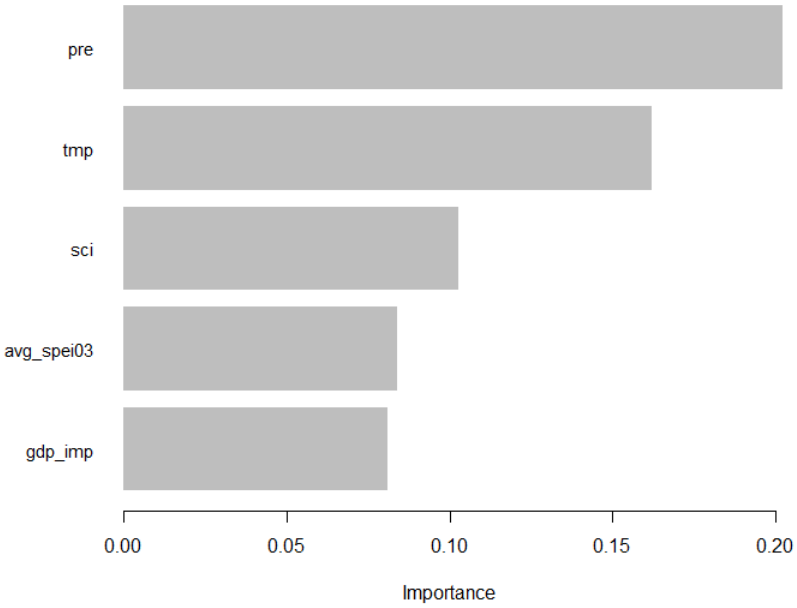


Figure 3 shows the five most predictors in the XGBoost algorithm. The variables are presented in order of decreasing importance. First, we can notice how weather variables, precipitation (*pre*) and temperature (tmp) are the most relevant variables for food insecurity. A less important role is played by short-term SPEI (avg_spei03), GDP (gdp) and the supply chain index (*sci*). 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.

Critically, if we consider the other two samples, the relevant importance of these variables remain quite the same. This means there are no changes in the main determinants of food security even when a structural break like COVID19 may occur.

5 Discussion and conclusion

We propose an effective multiclass classification model based on machine-learning tools to predict food insecurity in Chad. Using appropriate features' selection procedure, we are able to identify the key predictors of food insecurity. Our model of choice, which performs well 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 76%, along with information on the contribution of each of the variables of interest.

We find that XGBoosting proves to be effective at partially mitigating the prediction issues that necessarily emerge in the presence of structural breaks. This improvement is especially notable when comparing the algorithm with a Lasso model. We argue that this is a good tool for policy makers that need to target policy, especially in response to a crisis. The ability of the model to overcome a significant and multi-dimensional shock like COVID-19, and to still retain its predictive power is a remarkable feature that should be kept in mind by humanitarian organizations and institutions alike.

We show that the inclusion of granular, gridded data is an effective and often viable choice to achieve high prediction accuracy. We argue that the inclusion of these variables provides a significant improvement in terms of predicting and hence preventing food insecurity. 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. Third, 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 over other existing methods. Lastly, a potential extension of our work would be to run these machine learning methods directly on the spatial grids rather at the sub-national level. This would likely improve the estimates, as there would be more observations, and it would include more spatial correlated features.

Our work presents important limitations. The main drawback of the model is related to the lack of access to data on local impact of the COVID-19 pandemic, which could be informative for the diffusion of food uncertainty in local areas in the post-COVID period. A possible extension in this direction would be to enrich the data with plausible proxies of COVID impact, such as excess deaths or the number of infected individuals. An additional drawback of the model is related to data quality and availability. 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. Additionally, the data we use for GDP estimates is not available for the whole time window of our sample. Hence, some data need to be imputed, which can be a source of bias and uncertainty. It is also difficult to argue that GDP, albeit granular and population weighted, necessarily represents the best choice for estimating income in rural and isolated areas, such as some regions of Chad. Similarly, the data we gather from Akinci et al. (2023) on supply chain stress are only available globally. Indeed, one can reasonably expect that such a shock might have drastically different impacts across regions. Ideally, this measure would include considerations about, for instance, quality of local infrastructure.
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

Africa Center for Strategic Studies (2022). Conflict Remains the Dominant Driver of Africa's Spiraling Food Crisis.

Akinci, O., Benigno, G., Clark, H. L., Cross-Bermingham, W., Nourbash, E., et al. (2023). Global supply chain pressure index: The china factor. Technical report, Federal Reserve Bank of New York.

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.

De Haen, H. and Hemrich, G. (2007). The economics of natural disasters: Implications and challenges for food security. *Agricultural economics*, 37:31–45.

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.

Hansen, J., List, G., Downs, S., Carr, E., Diro, R., Baethgen, W., Kruczkiewicz, A., Braun, M., Furlow, J., Walsh, K., et al. (2022). Impact pathways from climate services to sdg2 ("zero hunger"): a synthesis of evidence. *Climate Risk Management*, page 100399.

Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations–the cru ts3. 10 dataset. *International journal of climatology*, 34(3):623–642.

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.

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.

Maina, J., Ouma, P. O., Macharia, P. M., Alegana, V. A., Mitto, B., Fall, I. S., Noor, A. M., Snow, R. W., and Okiro, E. A. (2019). A spatial database of health facilities managed by the public health sector in sub Saharan Africa. *Scientific Data*, 6(1):134.

Mason-D'Croz, D., Sulser, T. B., Wiebe, K., Rosegrant, M. W., Lowder, S. K., Nin-Pratt, A., Willenbockel, D., Robinson, S., Zhu, T., Cenacchi, N., et al. (2019). Agricultural investments and hunger in africa modeling potential contributions to sdg2–zero hunger. *World development*, 116:38–53.

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).

Murakami, D. and Yamagata, Y. (2019). Estimation of gridded population and gdp scenarios with spatially explicit statistical downscaling. *Sustainability*, 11(7):2106.

Nekmahmud, M. (2022). Food consumption behavior, food supply chain disruption, and food security crisis during the covid-19: The mediating effect of food price and food stress. *Journal of Foodservice Business Research*, pages 1–27.

Nica-Avram, G., Harvey, J., Goulding, J., Lucas, B., Smith, A., Smith, G., and Perrat, B. (2020). Fims: Identifying, predicting and visualising food insecurity. In *Companion Proceedings of the Web Conference 2020*, pages 190–193.

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

Pereira, M. and Oliveira, A. M. (2020). Poverty and food insecurity may increase as the threat of covid-19 spreads. *Public Health Nutrition*, 23(17):3236–3240.

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.

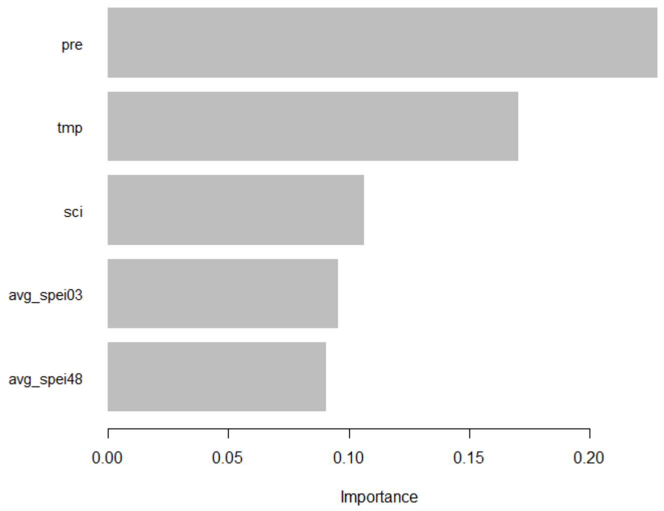
Tutz, G., Pößnecker, W., and Uhlmann, L. (2015). Variable selection in general multinomial logit models. *Computational Statistics Data Analysis*, 82:207–222.

Appendix

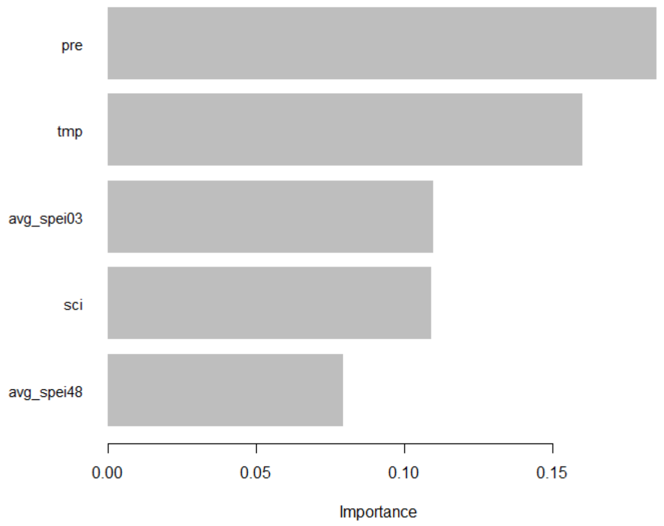
Table 3: Lasso variable selection - Total sample

	variables	cat0	cat1	cat2
1	(Intercept)	12.260	2.356	-14.616
2	adm1_pcod2_TD02	-0.123	-0.178	0.301
3	adm1_pcod2_TD03	1.916	-1.444	-0.472
4	adm1_pcod2_TD04	-1.024	-0.480	1.504
5	adm1_pcod2_TD06	-2.313	-0.394	2.707
6	adm1_pcod2_TD07	-1.532	-0.351	1.883
7	adm1_pcod2_TD08	0.506	-0.315	-0.191
8	adm1_pcod2_TD09	-0.047	0.206	-0.159
9	adm1_pcod2_TD10	0.272	-0.118	-0.154
10	adm1_pcod2_TD11	1.428	-0.793	-0.634
11	adm1_pcod2_TD12	0.661	-0.387	-0.274
12	adm1_pcod2_TD14	-0.723	0.727	-0.004
13	adm1_pcod2_TD15	0.463	-0.355	-0.108
14	adm1_pcod2_TD16	0.653	-0.318	-0.335
15	adm1_pcod2_TD17	-2.539	-0.628	3.167
16	adm1_pcod2_TD19	-1.423	-0.705	2.128
17	adm1_pcod2_TD20	-0.219	0.214	0.005
18	adm1_pcod2_TD21	0.321	-0.170	-0.151
19	adm1_pcod2_TD22	-1.992	0.188	1.804
20	adm1_pcod2_TD23	-0.912	-0.385	1.297
21	avg_spei03	0.030	-0.020	-0.011
22	avg_spei12	-0.045	0.217	-0.172
23	avg_spei48	0.119	-0.187	0.068
24	conflict_d	0.177	0.734	-0.911
25	gdp_imp	0.000	-0.000	-0.000
26	pre	-0.003	0.000	0.003
27	sci	0.015	-0.026	0.011
28	tmp	-0.389	-0.039	0.428
29	z_PPI	0.113	-0.100	-0.013

Figure 4: Relative importance of coefficients



(a) pre-Covid



(b) post-Covid