Machine learning can guide food security efforts when primary data are not available

Keywords: food insecurity, SDGs, non-traditional data, machine learning, forecasting.

Extended Abstract

The 2030 Agenda for Sustainable Development, adopted by all United Nations Member States in 2015, calls for urgent action to "end hunger, achieve food security and improved nutrition and promote sustainable agriculture". However, in 2019, 650 million people were still undernourished with 135 million in 55 countries and territories reported to be acutely food insecure. These numbers have significantly increased as a consequence of the COVID-19 pandemic, with at least 280 million people reported to be acutely food insecure in 2020 [1]. To address this global issue, monitoring the situation and its evolution is key. Governments and international organizations such as the World Food Programme (WFP), the Food and Agriculture Organization (FAO) and the World Bank perform food security assessments on a regular basis through face-to-face surveys or, increasingly so, through remote mobile phone surveys [2]. However, there are limitations with these approaches given their high costs in both monetary and human resources. Predictive models could therefore complement these data to nowcast food security in near real-time, allowing decision-makers to make more timely and informed decisions on policies and programs oriented towards the fight against hunger.

In this work, we propose to use machine learning to predict the current sub-national food consumption and food-based coping situation on a global scale from secondary data. Our main assumption is that, when primary data is not available, levels of insufficient food consumption and of crisis or above food-based coping can be estimated from secondary information, specifically on the key drivers of food insecurity. Experts identify three main causes for food insecurity: conflict, economic shocks and extreme weather events [1]. To build the proposed predictive models, we therefore collected historical data covering all three dimensions: data on conflict-related fatalities, economic information (prices of staple food in local markets, headline and food inflation, currency exchange rates and gross domestic product (GDP) per capita) and data on rainfall and vegetation, including anomalies with respect to historical averages. For each available historical measurement of insufficient food consumption and of crisis or above food-based coping for a given geographical area and time window (~35k data points across 78 countries and 15 years), we associated as independent variables the corresponding conflict, economic and weather situation for the same area in the previous three-month window. Moreover, we also took into account as independent variables under-nourishment and population density and the target prevalence measured during the most-recent previous food security assessment, whose time frame varies across the different areas.

For each target variable, we fitted 100 bootstrapped models using gradient boosted regression trees [3], employing the first (in temporal terms) 85% of the historical data. The proposed models are able to explain on the remaining 15% out-of-sample data (that is, corresponding to the past two months of data), 81% of the variation in the prevalence of people with insufficient food consumption and 73% of the variation in the prevalence of people using crisis or above-crisis food-based. We then trained two additional models, using the same approach but removing the prevalence from the previous assessment from the set of independent variables,

hence using secondary data only. In this case results show that the proposed models are able to explain, on the test set, 74% of the variance in the prevalence of people with insufficient food consumption and 61% of the variance in the prevalence of people using crisis or above-crisis food-based coping. We compared these results with those obtained from a naive approach that uses only the prevalence measured during the previous assessment as an independent variable. We find that this naive model can explain only 51% of the variation in the prevalence of people with insufficient food consumption and 45% of the variation in the prevalence of people using crisis or above-crisis food-based coping, demonstrating the fundamental importance played by secondary data to capture the dynamic nature of food insecurity. Having trained and validated the proposed models on historical data, we then further show that they can be used to predict the current prevalence of insufficient food consumption and of crisis or above food-based coping by comparing the data being collected in near real time by WFP during a recent two-month period with the corresponding predicted levels. Finally, we show that despite the nonlinear tree-based model structure, it is possible to provide interpretable explanations of predicted figures and of what causes changes over time, even if the models do not have an intrinsic dynamic component.

In the second part of the work, we focused on six countries in West Africa and the Middle East (Burkina Faso, Cameroon, Mali, Nigeria, Syria and Yemen) for which daily sub-national time series of the prevalence of people with insufficient food consumption are available. Here, we tackle a different problem: forecasting the future daily evolution of the prevalence of people with insufficient food consumption. Having access to reliable predictions of the evolution of insufficient food consumption levels over future weeks and months could allow governments and organizations to identify which areas should be monitored more closely and to eventually take timely decisions on resource allocation. To this aim, we combined information on the historical evolution of the target indicator with historical information on the key drivers of food insecurity and built gradient boosted decision tree models that comprise both endogenous (insufficient food consumption itself, as well as food-based coping information) and exogenous factors (conflict-related fatalities, rainfall and vegetation and their anomalies, staple food prices and Ramadan's occurrence). We showed that the proposed model makes it possible to forecast the prevalence of people with insufficient food consumption up to 30 days into the future with higher accuracy than a naive approach solely based on the last measured prevalence, at least in places where enough training data are available to inform the model, such as in Yemen (Figure 1). The spatial and temporal coverage of available historical observations proved to be a key element in forecasting success. Therefore, our study presents a simple, yet fundamental message for governments and humanitarian organizations on the power of the data they collect: collecting data on a regular basis for long enough periods of time and across enough different geographic areas does not only make it possible to monitor the evolution of a situation in near real-time but also to inform forecasting models that would make it possible to produce estimates of how the situation is likely to evolve in the near future.

References

[1] Food Security Information Network. *Global Report on Food Crises* (2021) https://www.wfp.org/publications/global-report-food-crises-2021.2021.
[2] World Food Programme. *Food Security Analysis* (2022) https://www.wfp.org/food-security-analysis
[3] Chen, T. & Guestrin, C. *XGboost: a scalable tree boosting system* (2016) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pages 785–794.

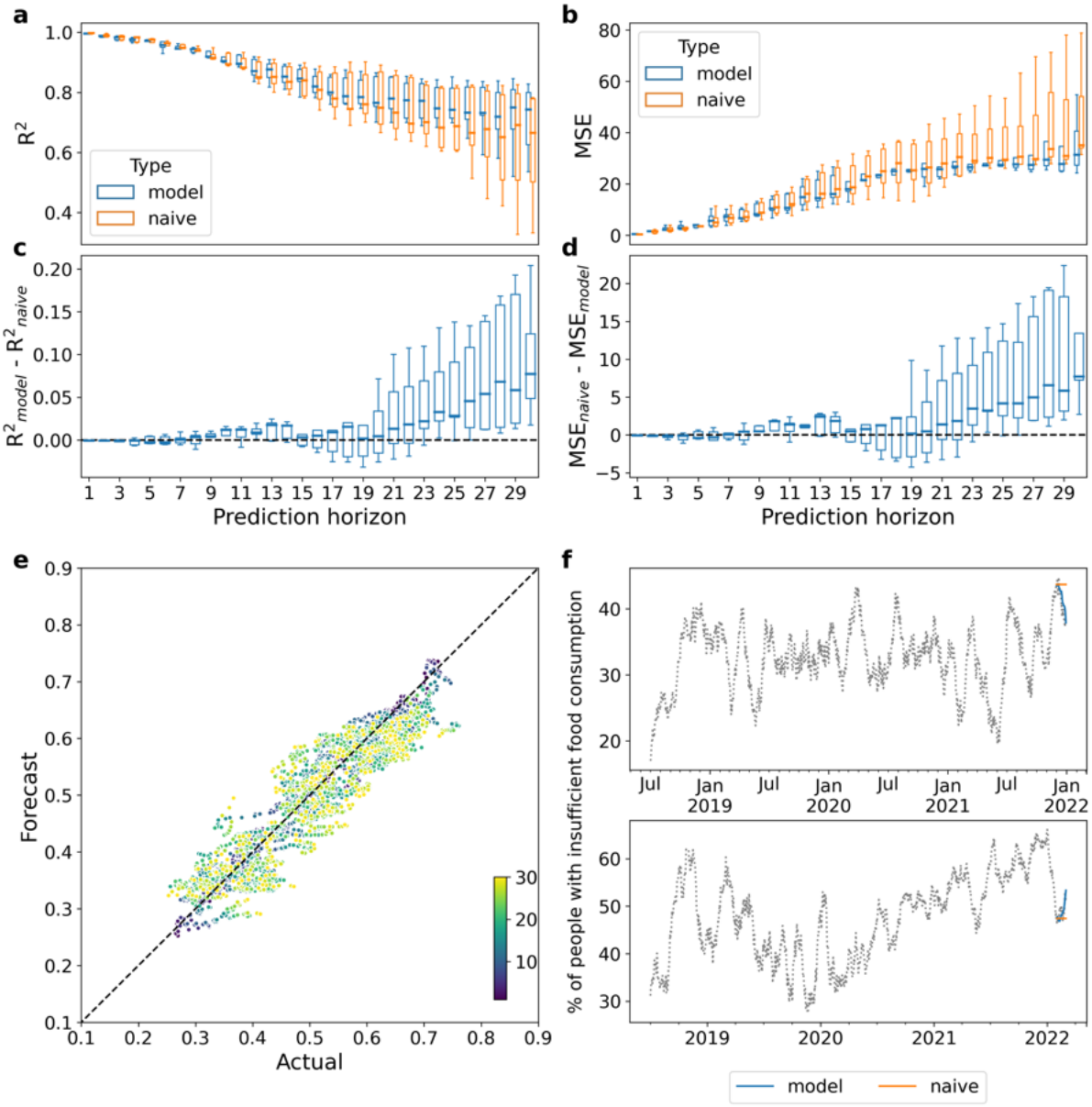

Figure 1: Forecasting the prevalence of people with insufficient food consumption in Yemen. The forecasting is performed over 5 different monthly splits of all governorates time series, from October 2021 to February 2022. (a) Box plots of the coefficient of determinations (R^2) across the 5 splits for both the proposed and the naive models (in blue and orange, respectively), for each forecasting horizon. (b) Box plots of the mean squared error (MSE) across the 5 splits for both the proposed and the naive models for each forecasting horizon. (c) Box plots of the difference between the R^2 of the proposed and of the naive model for each split. (d) Box plots of the difference between the MSE of the naive and of the proposed model for each split. (e) Predicted vs actual value for each data point in the 5 splits. Colors represent the corresponding forecasting horizon and vary from dark blue (1day) to yellow (30 days) (f) Example of forecasting results for December 2021 in Amanat Al Asimah (top) and February 2022 in Abyan (bottom).