Understanding Food Security, Undernourishment, and Political Stability: A Supervised Machine Learning and Data-Driven Approach

Yijie "Jerry" Yao, Shiyu Tian, Allen (Yixin) Hu

May 22, 2022

1 Introduction

Agriculture accounts for a comparatively small share of the global economy, but remains central in shaping the economic, health, and political scene of a nation. According to the Food and Agriculture Organization of the United Nations (FAO), many regions in the world with variable food supplies can be considered "food insecure", where individual persons do not consume a sufficient amount of calories (dietary energy) on a regular basis to lead a normal, active and healthy life [4]. Warfare and other types of political instability is also related to critical "staple crops" (crops such as maize, wheat, or rice), and "cash-crops" (such as coffee, cocoa, tea etc.) also plays into political instability. For example, the National Consortium for the Study of Terrorism (START) documents that the "green gold rush" of avocado production has become increasingly lucrative for cartels, and bolsters the cartel's income stream via violent extortion. [9]

Our guiding research question is: how do different agriculture-related features – prices, import and export, economic indicator of agriculture performance, environmental variables, and etc. – co-move and contribute to better understanding and prediction the undernourishment level and political stability?

Our research project ultimately uncovers critical factors driving undernourishment and political instability with our two-part econometric investigation powered by supervised machine learning techniques: 1) Interpretation of machine learning model outcomes / parameters and comparison with present literature, and 2) A diverse set of machine learning predictions of undernourishment levels and political stability levels of different nations.

1.1 Discussion on Current Literature

As mentioned previously, FAO documented strong correlation between the lack of food security and undernourishment. The mechanisms of food security effected by the agricultural sector was outlined by the FAO in their 2006 policy brief [6]. It was proposed that there are four critical pathways by which the dynamics of the agricultural industry will affect the food security of citizens:

1. Availability of food which hinges upon domestic production and/or imports.
2. Access to food which depends on individuals having sufficient resources or entitlements to obtain food, such as income and transaction costs.
3. Utilization of food, reflected by the quality and diversity of diets, and of clean water.
4. Stability which ensures that food can be accessed by individuals at all times.

However, it is unclear how these variables may affect food security (and in turn undernourishment levels) in different countries. Many research papers have used this framework as a means to understand food security, yet the results are inconclusive at best. For example, there has been a long-standing debate between cash-cropping and staple-cropping, where various studies have shown inconclusive results on whether or not one is "better" for food security than the other (see Anderman et al., 2014 and Achterbosch et al., 2014). [2][1] These papers and reports conducted their analyses based on author-aggregated values of different crops, yet their results regarding the relation between agricultural sector composition/activity and food security are sometime in conflict with one another: a common occurrence in the literature.

While these papers are well-founded in theory, we see the necessity to look at more granular data provided by the FAO database (which we use in this report), and leverage large-scale, data-driven techniques to identify what particular crops being produced, traded, supplied, and utilized, would contribute most significantly to undernourishment in a region. We want to compare and contrast our findings with supervised machine learning approaches against these finding.

Beyond this question, recall that START gave us a glimpse into how certain compositions of agricultural land may be significantly correlated with unrest within a region. Therefore, we are also interested about whether the prevalence in production and trade of particular crop types (say income-generating crops) are indeed more correlated with political instability. This is another interesting target variable to look at since we could see if the same factors that contribute to food insecurity is also prevalent for political stability.

2 Methodology

We have two main datasets, one for each target variables described below. For each dataset, we have 1000 features, and each row represents a country's features and values at an given year spanning from 2000 to 2020.

2.1 Target Variable 1: Political Stability

The **Political Stability and Absence of Violence/Terrorism Index**, constructed by Worldwide Governance Indicators, takes incidents and indices from a wide range of sources to "measure perceptions of the likelihood of political instability and/or politically- motivated violence, including terrorism."[10] Representative sources include counts of Orderly transfers, Armed conflict, Violent demonstrations, etc. gathered by EIU (Economist Intelligence Unit Riskwire Democracy Index), and Government instability, Internal and external conflicts collected from PRS (Political Risk Services International Country Risk Guide).

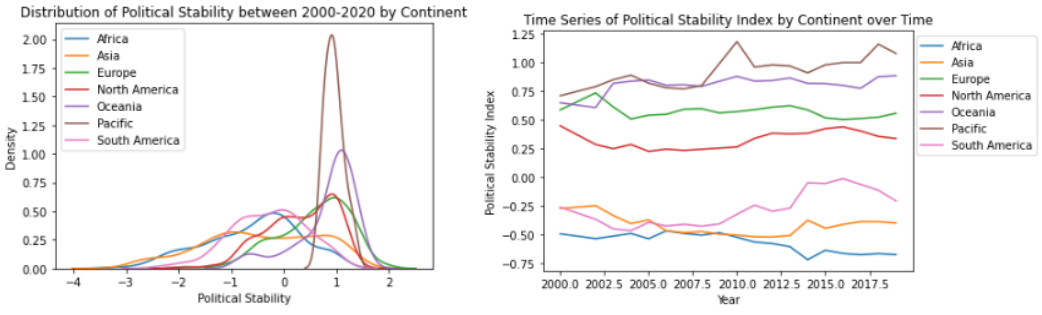

Figure 1: Distribution of undernourishment and political stability between 2000-2020.

It is worth noting that most of these distributions are skewed with a long left tail, indicating that the political stability level in many regions of the world are lower than the baseline value of this index.

We see no extreme anomalies, except that the Pacific and Oceania are more skewed towards the right, where as South America and Africa is more skewed towards the left. Turning to the time series, there are two main groups we can see visually: We see that Oceania, Europe and North America have on average higher political stability levels relative to Africa, South America, and Asia.

2.2 Target Variable 2: Undernourishment

The **Undernourished Index** measures percentage of undernourished people in a single country. FAO defines "undernourishment" as the inability "to acquire enough food to meet the daily minimum dietary energy requirements, over a period of one year".[5] In other words, the undernourishment index, or POU, is only interested in whether people have enough energy intake and disregards the dietary structure of a population.

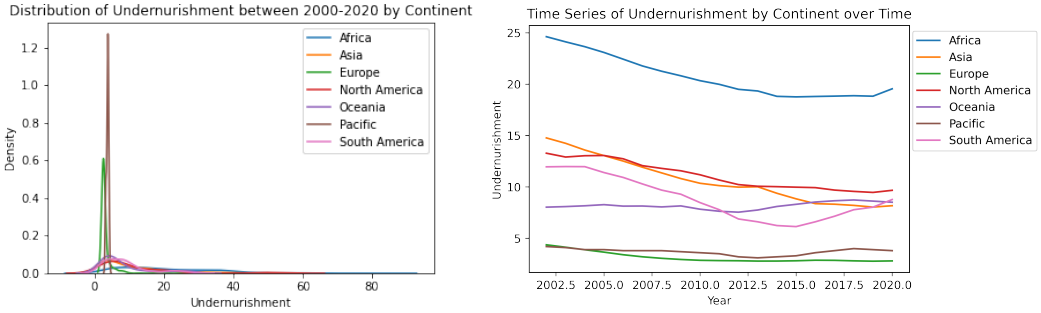

Figure 2: Time series of average undernourishment and political stability between 2000-2020

Notice that for the distribution of undernourishment (lower is better), we see a very long-tailed distribution for many nearly all continents besides Europe and the Pacific island portion of Oceania. This is because there are certain areas will undernourishment levels of nearly 60 to 70, near the far end of the right tail. This is concerning as we have large gaps in our data, where the extreme values are far greater than where the conventional distribution lies. Turning to the time series, we see that in general, undernourishment is decreasing in the world, but it can be broken up into three main "groups", Africa has consistently the highest level; Asia, the Americas, and Oceania occupy the middle, and Pacific Oceania and Europe has the lowest levels consistently.

2.3 Feature Variables

The database that we use is the most recent and complete FAOSTAT database, which is a publicly available database that the FAO provides for open use. We left join on each of the undernourishment and political stability datasets when merging the full FASTOT dataset. We ended up with more than 12,000 different features for 3000+ observations. We then picked the top 1000 features with the least amount of data missing to start with for our machine learning approaches.

2.4 Methods for Interpretation

For the first task of ML-based interpretation, we use LASSO and a Random Forest model to select important features for our target variables of undernourishment and political stability.

2.4.1 LASSO

We train LASSO regression on both political stability and undernourishment indices grouped by continents for all observations before and include year 2017. We assume the following Data Generating Process:

With the above-mentioned introduction of the Political Stability and Absence of Violence / Terrorism index in Part 2.1, we argue that such index takes a rather comprehensive look at a nation's political stability-related performances. Moreover, we recognize that the political instability of a nation is instigated by a number of finite internal/external actors taking interest in different but finite factors, such as Gross National Income (GNI), Food Availability, Employment Rate, etc. While it could be argued that some of these factors contribute to the outcome, it is inconceivable to argue that most of these factors, which include items such as amount of chicken waste left exposed, would be significant in inducing undernurishment. As most factors do not contribute significantly to the outcome, it is reasonable for us to consider the DGP of the political stability as sparse, and thus fulfilling the prerequisite of applying LASSO regression.

As discussed in Part 2.2, Undernourishment Index reflects the percentage of population in a nation whose energy intake does not meet the suggested lower limit for health. We would argue that the direct reason for such phenomenon comes from a rather small number of factors, such as the type of crops, price for local staple crops, food accessibility as well as national income level rather than the whole set of potential crops a farm can plant and thus implying an appropriate application with LASSO regression.

2.4.2 Random Forest

We train random forest on observations before and include year 2017 to provide another lens for evaluating feature importance. Feature importance of Random Forest is estimated very differently compared to LASSO. The importance of features is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.

2.5 Methods for Prediction

We will use the following models and regression techniques for our prediction tasks:

Linear Models Linear Regression, LASSO, Ridge Regression

Ensemble learning Decision Tree Regression, Random Forest Regression

Non-Linear Models Support Vector Regression, Neural Networks for Regression

All models will be trained with data before and include the year 2017. Models are tested for Out-of-Sample Prediction based on data after year 2017. Linear Models are evaluated with R squared and Mean Absolute Errors (MAEs); and the remaining models models can only be evaluated by MAEs. All MAEs are evaluated in the standard scaled data. All models have their parameters tuned to obtain the best Out-of-Sample MAE. We choose the best performing model to showcase results through and a subset of countries.

3 Results and Discussions:

3.1 LASSO for Interpretation

With the assumptions in mind, we would like to interpret the results of LASSO regression by conducting a worldwide-continental comparative study.

We can clearly observe (Figure 3) a positive correlation between the size of GDP, Gross Fixed Capital Formation and political stability, implying that stronger economies are (generally) politically more stable. However, the fact that Value Added (Agriculture, Forestry and Fishing) is the most influential negative factor worldwide indicates that nations that are more dependent on agricultural income are more likely to suffer from political instability, possibly because the nations dependent on primary

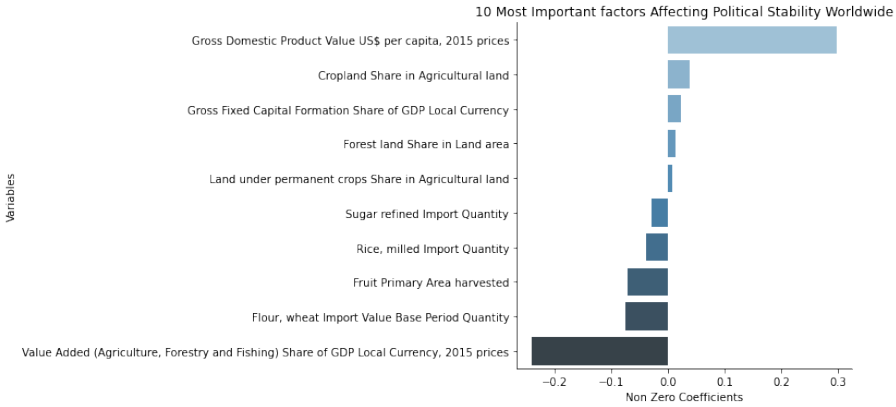

Figure 3: 10 Most Important Features Affecting Political Stability Worldwide

sector economy yield lower GDP per capita in general, or perhaps individuals could be underemployed into the agricultural sector due to the lack of sufficient educational infrastructure, which as been suggested in some parts of developmental literature[3]. This however cannot be necessarily determined with the data and models at hand, nor is it necessarily the aim of this exercise.

Furthermore, LASSO regression justifies the importance of food security with respect to political stability. Three out of five negatively significant features – quantity of imported flour, rice and sugar – marks the level of staple food independence, suggesting that the dependency for imported food may lead to food insecurity, as well as influencing political instability (note that we are not making a causal argument, rather it is interesting that these variables indeed co-move).

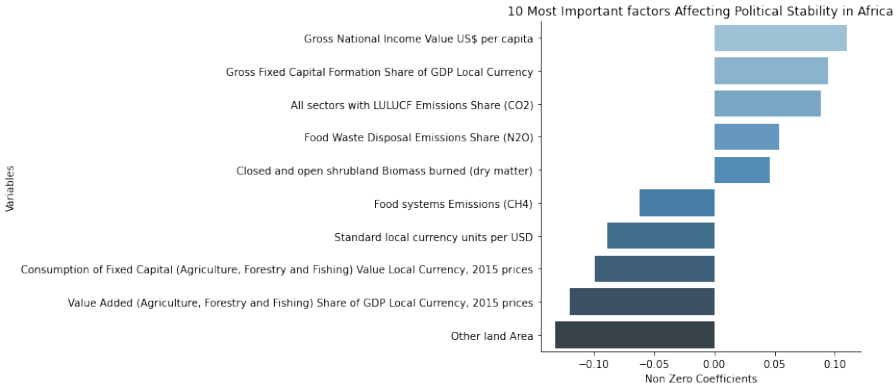

Figure 4: 10 Most Important Features Affecting Political Stability in Africa

Unlike the worldwide data, Africa's (Figure 4) most significant negative impact factor seems to be "Other Land Area": Land areas not utilized for forestry or agriculture.[8] In other words, other land area counts urban land area as well as barren land area. In the case of Africa, most of the "other land area" may be barren land such as the Sahara Desert. Hence we suspect that this is indicating that inefficient use of land can thus jeopardize political stability of a nation.

In the case of undernourishment, our LASSO regression applied on worldwide data (Figure 5) shows that food and agricultural product export unit/value index which "represent the changes in the quantity-weighted unit values of products traded between countries." [7] In short, such index measures the trading price of foods in the international market. We can subsequently validate our assumption on price fluctuation affecting undernourishment. Nevertheless, most of the covariates

negatively correlated to undernourishment are yield of agricultural products that are common on people's table but not necessarily staple-crops.

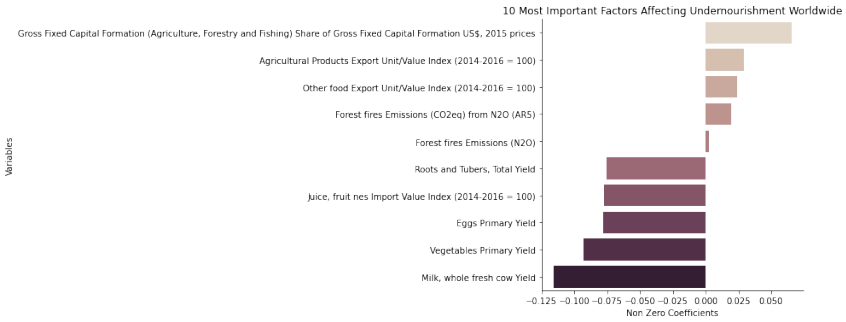

Figure 5: 10 Most Important Features Affecting Undernourishment Worldwide

Focusing on Asia (Figure 6), we can see that instead of the unit value of exported agricultural products in general (as in the worldwide case), our LASSO regression in Asia successfully singles out the most important staple crop in Asian population that has significant impact on undernourishment – wheat flour.

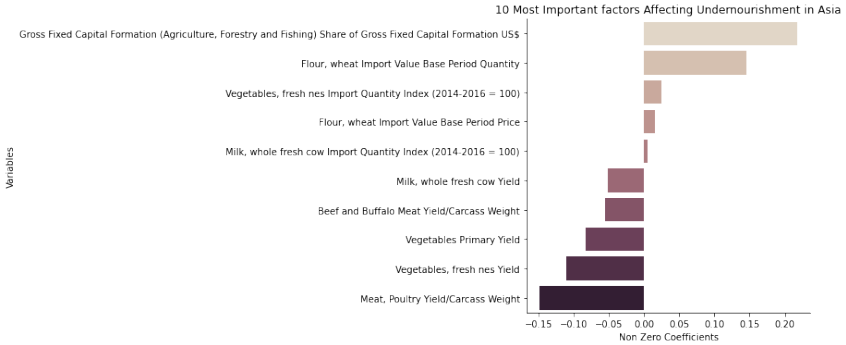

Figure 6: 10 Most Important Features Affecting Undernourishment in Asia

3.2 Random Forest for Interpretation

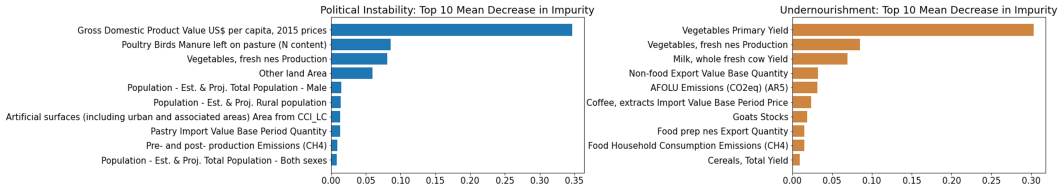

Figure 7: Random Forest Top 10 Most Important Features

Based on top 10 features selected for the entire Political Stability dataset (Left on Figure 7), Random Forest shortlisted a slightly different subset of features from top 10 most important features selected by LASSO, signified by one shared feature: Gross Domestic Product Value. So it confirms the importance of GDP in attempt to estimate political stability – stronger economies are harder to destabilize. Such argument can also be supported by the presence of population as important factors. As previously discussed in the case of Africa, we also see Other Land Area in significant features selected by Random Forest. Recall that "other land areas" is composed of artificial/urban areas and

barren areas. Since the index "artificial surfaces" is also included in the most important features, we can infer that urbanization may also have an effect on political stability.

Based on the 10 features selected for the entire undernourishment dataset (Right on Figure 7), Random Forest and LASSO shortlist the same top 3 features. Because both Random Forest considered Vegetables Primary Yield, Vegetables production, and Milk whole fresh cow Yield as the most important features contribute to estimating undernourishment, we are confident that these domestic yield of the these agricultural products are indeed important indicators. Other notable important feature that estimated by Random Forest but ignored by Lasso include: Non-Food Export Value Base Quantity, Goat Stocks, Food Household Consumption Emissions. These features shared the commonality that they serve as proxy as how much food is available and consumed on household level.

3.3 Results on Prediction with Different Regression Techniques

Political Stability: Regression In-Sample and Out-of-Sample Prediction Performance							
	Linear	LASSO	Ridge	Random Forest	Support Vector	Decision Tree	Neural Network
Parameters		Alpha = 0.12	Alpha = 5e7	N estimators = 80		Max depth = 90	5 layers
R^2 (training)	0.9266	0.4763	0.0024				
R^2 (testing)	-3.0713	0.4742	-6.7428				
MAE (training)	0.2100	0.5738	0.8170	0.0617	0.2306	0.0000	0.5188
MAE (testing)	21875503.1	0.5304	0.9187	0.1701	0.3407	0.2350	0.4848

Undernourishment: Regression In-Sample and Out-of-Sample Prediction Performance							
	Linear	LASSO	Ridge	Random Forest	Support Vector	Decision Tree	Neural Network
Parameters		Alpha = 0.06	Alpha = 1e4	N estimators = 100		Max depth = 40	5 layers
R^2 (training)	0.8462	0.4862	0.5237				
R^2 (testing)	-1.9081	0.3252	0.2883				
MAE (training)	0.2900	0.5069	0.4778	0.0403	0.2666	0.0000	0.4970
MAE (testing)	26021302.8	0.5059	0.4880	0.2099	0.5991	0.2304	0.4778

Based on the table above, we see that linear regressions performs extremely poorly, while Decision Tree Regression and Random Forest Regression consistently outperform other regression techniques.

Decision tree regressions performs well because the nature of our dataset has each row representing a country in a particular year. So the decision tree builds regression models in the form of tree structures. It breaks down a dataset into smaller and smaller subsets at the same time as an associated decision tree. The leaf node represents the decision on the numerical target. So these decision tree can essentially break down to individual country to an extent of a perfect fit as shown in the almost 0 MAE for training.

However, the single decision tree method is not the most optimal way of prediction relative to Random Forest. Recall that the single decision tree method constructs subtrees via pruning using a single, large tree as its template. This method is not exactly optimal due to many factors, but an important one discussed in class was due to the potential to overfit the data by picking up too much of the noise. Random forests constructs subtrees using a random subset of variables and makes the decision framework less sensitive to the structure of the training data. By doing so, the random forest method does not capture too much of the noise present in the training data, and thus makes for less overfitting, and hence tends to appear as a generally stronger prediction method. Thus we use Random Forest to predict the target variables for a few countries after 2017 on, presented in Figure 8 below (Notice that they tend to have a “centering” effect, whereby the model predicts countries to converge to a mean value). Of course, it would also be interesting if we could have used individual country outcomes to predict the future for *that particular* country, yet the data was simply not robust enough for us to present any meaningful observations.

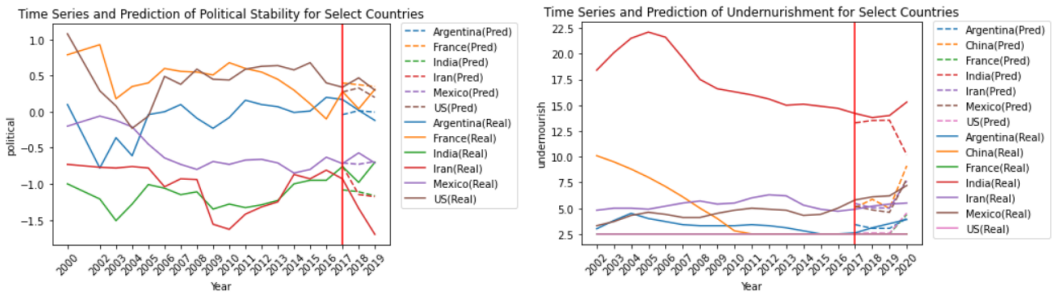

Figure 8: Predictions using Decision Tree Regressions

4 Summary and Conclusion

After the process of exploratory data analysis, interpretation of important features powered by LASSO regression and Random Forests as well as a "horse-race" of multiple linear and non-linear prediction methods, we have came to the following conclusions:

We set off to analyze whether FAO's claimed factors on food security can be justified using a data-driven approach. In Part 3.1 and 3.2, we have observed that part of FAO-identified factors are indeed obvious factors for food security and subsequently affects political stability as well as undernourishment. Indicators for the size of national economy and income, GDP and GNI in particular, are ranked as highly important features in our feature selection results. In addition, indices signifying food independence such as import and domestic production quantity of "staple crops" seems to have a heavy impact on both political stability as well as undernourishment. On the other hand, there are certain factors that the FAO doesn't seem to take into account, such as the degree of urbanization and Fixed Capital Formation for agriculture and forestry.

Through implementing a number of linear and non-linear regression methods on respective target variables, we have discovered that Decision Trees and Random Forests provides a significantly lower for in-sample and out-of-sample Mean Absolute Error compared to other methods. However, through testing our Random Forest predictions on representative countries and visualizing the result, we learned that out-of-sample prediction with train Random Forest regressor tends to exhibit a "centering" effect and converge to the mean value.

References

[1] Thom Acherbosch, Siemen van Berkum, and Gerdien Meijerink. *Cash crops and food security*. 2014. URL: https://edepot.wur.nl/305638.

[2] Tal Lee Anderman et al. "Synergies and tradeoffs between cash crop production and food security: A case study in rural Ghana". In: *Food Security* 6.4 (2014), pp. 541–554. DOI: 10.1007/s12571-014-0360-6.

[3] Atif Awad. "From school to employment; the dilemma of youth in Sub–Saharan Africa". In: *International Journal of Adolescence and Youth* 25.1 (2020), pp. 945–964. DOI: 10.1080/02673843.2020.1778492.

[4] FAO. *Hunger and food insecurity*. URL: https://www.fao.org/hunger/en/.

[5] FAO. *Prevalence of undernourishment*. 2000. URL: https://www.fao.org/sustainable-development-goals/indicators/211/en/.

[6] Agriculture FAO and Development Economics Division. *Policy brief: Food security - issue 2, June 2006*. June 2006. URL: https://reliefweb.int/report/world/policy-brief-food-security-issue-2-june-2006.

[7] FAOSTAT Database FAO. *Definitions and standards used in FAOSTAT*. 2000. URL: https://www.fao.org/faostat/en/#data.

[8] Synthesis Report FAO. *The State of the World's Land and Water Resources for Food and Agriculture*. 2021. URL: https://www.fao.org/3/cb7654en/cb7654en.pdf.

[9] *Tracking Cartels Infographic Series: The Pits: Violence in Michoacán Over Control of Avocado Trade*. 2019. URL: https://www.start.umd.edu/tracking-cartels-infographic-series.

[10] Governance Indicators World Bank. *Worldwide Governance Indicators*. 1996. URL: http://info.worldbank.org/governance/wgi/Home/Documents.

5 Appendix

We included our data description in our methodology section because we believe it's key to address and understand why we use our target and feature variables as part of our methods. The original FAOSTAT dataset including both our target and feature variables can be found here https://www.fao.org/faostat/en/.

Here is the link to our data and code repository: https://github.com/jerryyao-uofc/food_security_and_political_stability. In this repository, you will find our link to our merged datasets, notebooks that documented our data cleaning, processing, merging steps, and regression results.