Practicum in Large-Scale Data Analysis

Project 1: Real Time Proxy Indicators for SDGs

Table of Contents

I. Introduction

The United Nations' Sustainable Development Goals (SDGs) are 17 targets that the world has agreed to aim toward through policy, development, and investment. The Goals cover diverse areas such as ending poverty, ensuring access to clean water and quality education, and guaranteeing equality for all. Progress on the Goals is measured through several specific indicators into which each Goal is divided, but the data for these indicators is often difficult to find, especially for the most recent years and for the countries of greatest concern. To address this, a team of Columbia University students in the Quantitative Methods in the Social Sciences (QMSS) program was tasked with finding effective proxy measures free from time lag issues to provide updated information on selected indicators and countries on behalf of the Joint SDG Fund - United Nations.

The study will focus on three specific indicators, 2.1.1, 2.2.2, and 10.1.1, to provide an alternate proxy to measure. The first two of these indicators provide information on SDG 2, "Zero Hunger". 2.1.1 is the "Prevalence of undernourishment" in a country, and 2.2.2 is the "Prevalence of childhood malnutrition (wasting or overweight)". This is defined as the "prevalence of malnutrition among children under 5 years of age, by type (waste and overweight)". Finally, indicator 10.1.1 "Income growth inequalities" under SDG 10, "Reduced Inequalities", measures the "growth rates of the household expenditure or income per capita of the bottom 40% of the population". These indicators were chosen as the focus of the group's work after the initial stages of theoretical research and data exploration revealed them as the most promising to explore.

II. Literature Review and Theory

Scholars have conducted research into which metrics are correlated with one another and may be used to approximate one another. For both SDG 10.1.1 and SDG 2, theoretical ideas included aspects of a society and economy such as share of expenditure by different sectors, the overall inequality as measured by the Gini coefficient, and social media data. Some sectors of the economy are sometimes used as proxies for or considered related to the income distribution in a society, as they reflect both the different preferences and the different financial abilities of people in different classes. Sani (2018) clarifies the bottom 40 percent population through a predictive classification. The paper identifies the best machine learning models using Naive Bayes, Decision Tree and k-Nearest Neighbors algorithm for classifying the bottom 40 percent population. Additionally, social media data could reflect sentiments or situations people are in due to financial stress or lack thereof.

The recent research paper, *Malawi: Nutrition Profile* released by the USAID in May 2021 gives inspiration for the proxy SDG 2.2.2 specifically. In this study, educational data concerning

adults will be used (ideally between 19-40 years old) as an alternative measurement, as the research paper showed that malnutrition of babies is related to the education level of their parents. For instance, in Malawi, according to the most recent Demographic and Health Survey (DHS), among children 6–23 months born to mothers with no education, only five percent receive a minimum acceptable diet; this number increases to 13 percent among mothers with a secondary education. Research also showed that childbearing begins early in Malawi. By age 19, 59.2 percent of adolescent girls had begun childbearing in 2015–2016. To make this proxy more appropriate, a number of countries with relatively low literacy rates have been selected . Therefore, Malawi and Rwanda will be the first-stage countries of interest for the data validation of proxy.

Some researchers have found the opposite conclusion in different countries, which implies that generalizations should not be made. A validated and good indicator for one or several countries can still be a poor proxy for another due to different national conditions. For instance, Howe (2009) uses the wealth index as a proxy for consumption expenditure, which turns out to be a poor indicator. However, Abreu (2013) has the opposite evidence from African countries. Provided that asset selection is accurately made, so that assets are able to discriminate along the actual poverty-wealth spectrum, and that there is relative homogeneity of that spectrum within the population, asset indices provide good proxies of long-run wealth.

III. Indicator 2.1.1

Scope and Theoretical Support

- Background

Indicator 2.1.1 Prevalence of undernourishment (PoU) measures the percentage of individuals in the total population that are in a condition of undernourishment. Undernourishment is defined as the condition in which an individual's habitual food consumption is insufficient to provide the amount of dietary energy required to maintain a normal, active, healthy life. In current measures, it is computed using a model of probability distribution of habitual dietary energy intake levels (expressed in kcal per person per day) for an average individual in a population:

$$PoU = \int_{x<MDER} f(x|\theta)dx,$$

where MDER stands for minimum dietary energy requirements, and θ is a vector of parameters that characterizes the probability density function (assumed to be lognormal), including the mean dietary energy consumption (DEC) and its coefficient of variation (CV).

However, there are several limitations to this approach. First, it computes the estimated value of PoU based on the assumption of lognormal distribution of dietary energy intake levels among population, which could be violated in reality; second, the data for DEC and CV are sourced from nationally representative household surveys which are not conducted on an annual basis in most countries, resulting in inaccuracy of data in the gap years due to high dependence on values from previous years where there is survey data available. These limitations of inaccuracy and lag in report calls for alternative measurements of the indicator.

- Rationale

There have been ongoing empirical studies showing the effect of incomes and prices on undernutrition. The application study in Indonesia in the mid-1980s indicates staple food prices, along with average income levels, intra regional inequalities, have unambiguous effects on undernourishment (Ravallion, 1992). In the cross-country inquiry of Anríquez et al. (2012), researchers made more explicit the relationship between staple food price and undernourishment: "the spikes of staple food price not only reduced the mean energy intakes, but also further worsened the calories distribution of the diet, thus deteriorating the nutritional status of the population".

Based on the evidence from previous studies, the team proposed an approach that estimates the prevalence of undernourishment from the food demand side, which captures the variations in population with insufficient purchasing power to stay out of the state of undernourishment. This is realized mainly by computing an "undernourishment poverty line" from staple food prices and food consumption patterns of the lower income group and then combining it with poverty indicators to enhance the estimation of undernourishment from poverty rates.

- Countries of Focus & Years of Study

Indonesia was used for the initial construction of this study's measurement and validation. Indonesia was selected to optimally satisfy one major assumption behind this study's approach. This assumption entails that people's accessibility to staple foods could be properly reflected by the national average market prices of the foods, calling for a nationwide food market that functions relatively well. Compared with Uzbekistan and Barbados where undernourishment is less prevalent, and Malawi and Rwanda where the aforementioned assumption is more likely to be violated due to higher level of market imperfections, Indonesia and Cambodia are more appropriate for the purposes of this approach. The approach requires specific local data such as food price, dietary composition, and energy requirements to be compiled into the estimation model. For this reason, only one country can be approximated per model, meaning Cambodia

would be a separate project to Indonesia. A replication study based on Cambodia's data could be conducted in the future.

Due to data availability for food prices as well as official data for the original indicator for validation purposes, the years 2007 to 2019 will be studied in construction of the alternative measurement.

Data Descriptions and Preprocessing

The following datasets were used in this methodology:

1. SDG 2 Indicator 2.1.1 Dataset
   Source: UN SDG indicators database
2. Poverty rate data
   Description: percentage of poor people in total/urban/rural population, annual/half-year data;
   Source: BPS-Statistics Indonesia website
3. Datasets for calculation of the "undernourishment poverty line"
   a. Staple food prices
      Description: monthly statistics of national average actual prices of several categories of staple foods;
      Source: WFP data;
      Preprocessing: excluded those food categories in the dataset that are less contributing to energy intake;
      i. Included: rice, wheat flour, eggs, meat (beef), meat (chicken, broiler), milk (condensed), oil (vegetable);
      ii. Excluded: sugar, chili (bird's eye), chili (red)
   b. Weight structure of food items
      Description: weight structure of food items consumed by the population group with monthly expenditure of 200,000 to 299,999 rupiahs; original dataset disaggregated by groups of population falling within different monthly expenditure ranges;
      Source: BPS-Statistics Indonesia, 2019;
      Preprocessing: chose the consumption structure of the 200,000-299,999 rupiahs/month group because this amount of monthly expenditure is the closest to the national average food poverty line (according to BPS data), thus representative for the food consumption pattern of the studied group potentially faced with the risk of undernourishment
   c. Calorie content of food items
      Source: USDA National Nutrient Database for Standard Reference, Release 20;

Preprocessing: selected from the table the calorie content of items most matching the food categories of interest; calculated average value where necessary

d. Minimum Dietary Energy Requirements (MDER) by age, sex, height, weight and activity level
   Source: Human energy requirements: Report of a Joint FAO/WHO/UNU Expert Consultation;
   Preprocessing: looked up for the MDER of each sex/age group separately in the tables provided in the report above, where for adults the standard BMI corresponding to the national average height of male/female in Indonesia was adopted, and for all age/sex group the medium activity level was adopted; and then calculated weighted average MDER according to the distribution of population among each sex/age group
e. Age/gender structure of population
   Source: BPS-Statistics Indonesia, 2010 Population Census and Indonesia Population Projection 2010–2035;
   Preprocessing: No significant variances in sex ratio (approx. 1.01) during the observed years, took 1:1 in calculation; age structure data missing for most of the years, took the values from the age distribution data of 2019
f. Average height features of population
   Description: average height of male/female adults in Indonesia;
   Source: WorldData.info
g. Engel Coefficient
   Description: the proportion of food expenditure in total expenditure among the group with monthly expenditure of 200,000 to 299,999 rupiahs; Group chosen for the same reason as food consumption structure;
   Source: BPS-Statistics Indonesia, National Socioeconomic Survey March 2019
h. Consumer Price Indices (CPI)
   Description: CPI (2007=100) for inflation adjustments of computed food expenditure minimum line;
   Source: BPS-Statistics Indonesia

4. Percentage of population with sources of improved drinking water
   Source: BPS-Statistics Indonesia

Methodology

- Indicator Construction

The construction of the core proxy indicator will now be introduced, along with its two proposed applications in approximating the indicator of interest. One of these was adopted in the validation process of this report, while the other could be further developed in future studies.

More specifically, in both ways, average food price per kcal, national average MDER, and minimum required food expenditure per capita were computed step by step. In the first approach, the expenditure was adjusted for inflation to capture the variations in minimum required expenditures of the lower income group caused by real food price fluctuations, while in the second approach, the minimum required total expenditure per capita was fit into an estimated cumulative distribution function of income to directly derive the percentage of undernourished population. Expressed by mathematical formulas, the process of the derivation is as follows:

1. First, the average food price per kcal in each month can be calculated using staple food price data, calorie content data and the proportion of food item *i* in the total weight of a standard basket of staple foods according to consumption patterns of the lower income group:

$$Price/kcal_{m} = \frac{\sum_{i=1}^{k}(w_i \times P_{im})}{\sum_{i=1}^{k}(w_i \times C_i)}$$

   where *m* stands for month, *k* stands for the number of food items (which is 7 in this study, as mentioned in the data preprocessing part), and for each food item *i* in the basket, w_i stands for its share of weight, P_{im} stands for its monthly unit price per kg, and C_i stands for its energy content per kg.

2. Next, the average MDER per month can be calculated using information on the MDER function on age, sex, height, weight and activity level (where height, weight and activity level of each age/sex group and the overall sex ratio were treated as fixed values in this study: heights of male/female adults were taken as the average value in the country; weights were taken as the standard BMI weights associated with the average heights; activity level was assumed to be medium; the sex ratio was taken as 1:1), as well as the population structure by age group:

$$MDER^{E} = \frac{1}{2}\sum_{j=1}^{n}[p_j \times MDER_{female}(Age\ group_j)] + \frac{1}{2}\sum_{j=1}^{n}[p_j \times MDER_{male}(Age\ group_j)]$$

   where *n* stands for the number of age groups, and p_j stands for the proportion of the age group *j* in total population. $MDER_{female}(Age\ group_j)$ and $MDER_{male}(Age\ group_j)$ are functions that calculate the monthly MDER for each combination of age group and sex.

3. Then by multiplying monthly MDER by price per kcal, the minimum required food expenditure per capita in the month can be derived; and then the annual data can be derived by summing up the monthly data:

$$minimum\ required\ per\ capita\ food\ expenditure = \sum_{m=1}^{12} (Price/kcal_m \times MDER^E)$$

4. Following that, in the first approach, the calculated food expenditure is adjusted by inflation rates before applied in a regression model with poverty indicators:

$$adjusted\ food\ expenditure_t = \frac{minimum\ required\ per\ capita\ food\ expenditure_t}{CPI_t/100}$$

where CPI_t indicates Consumer Price Index in year *t* (2007=100).

5. In the second approach which was not adopted in the validation process of this report due to data unavailability, following step 3, the minimum required per capita income each year could be calculated using Engel Coefficient, which is the proportion of food expenditure in total expenditure:

$$minimum\ required\ per\ capita\ income_t = \frac{minimum\ required\ per\ capita\ food\ expenditure_t}{Engel\ Coefficient}$$

6. Finally, an estimation line of the cumulative distribution of income could be fitted using the annual poverty headcount ratio data at several national poverty lines, and then with this function the percentage of population living under the minimum required per capita income threshold could be calculated, i.e. the estimated prevalence of undernourishment:

$$Prevalence\ of\ undernourishment_t^E = F_t(minimum\ required\ per\ capita\ food\ expenditure_t)$$

where $F_t(x)$, an estimated function of the cumulative distribution of income in the year *t*, equals the percentage of population with an annual income less than or equal to *x*.

- Validation Process & Results

Figure 3.1 and Figure 3.2 show the computed price per kcal and adjusted minimum required food expenditure per capita in Indonesia. It is indicated that there were apparent fluctuations in monthly data of staple food prices, implying that price changes could potentially affect people's ability to gain sufficient dietary energies. It can also be seen that there was an overall increasing trend over the years from 2007 to 2012 in inflation-adjusted minimum food expenditure faced by the lower-income group, while the years after 2012 witnessed large swings in the figure.

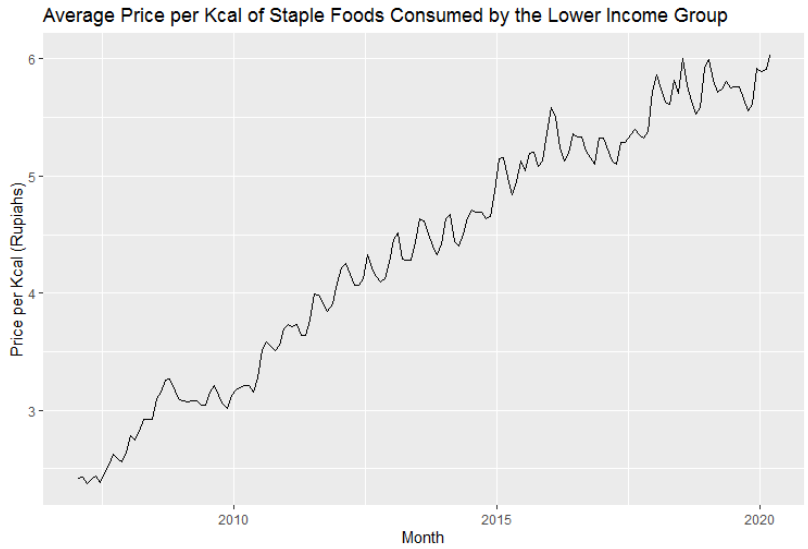


Figure 3.1

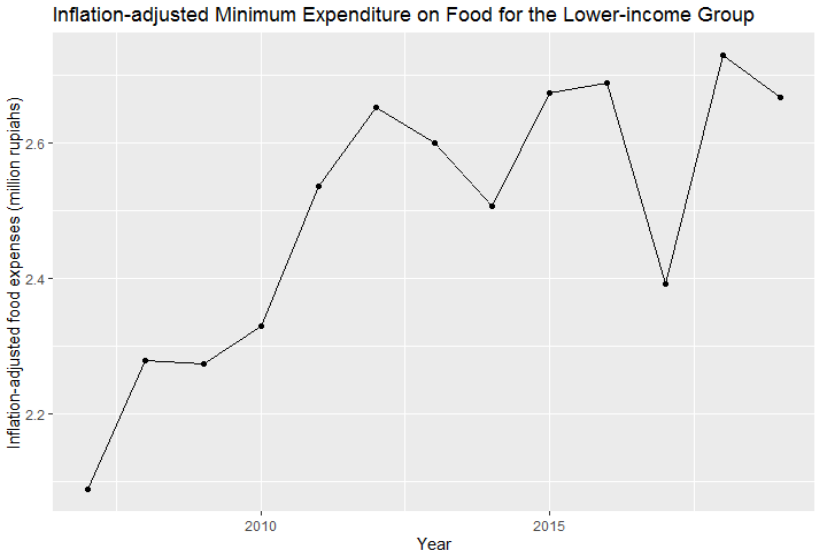


Figure 3.2

With the relative food expenditure and poverty indicators, as well as the percentage of the population with access to improved drinking water as a control variable, we conducted multilinear regression analysis on annual national data of the available years to adjust for the translation from poverty to undernourishment and to validate the results. The results in Figure 3.3 exhibited very statistically significant patterns (99% Confidence Interval) of the percent of poor population positively correlated with the dependent variable prevalence of undernourishment, with the models overall resulting in high R-squares of more than 0.95. Moreover, after including poverty rates in the rural area instead of national average rates, the predictive power of the model increased, as shown by the larger R-square in column (3) relative to column (1), whereas the urban poverty rates didn't present such effects. This could constitute

certain evidence that the majority of undernourishment was occurring in the rural parts of the country.

However, as demonstrated in the first row and sixth row of Figure 3.3, unfortunately, insignificant patterns were found for inflation-adjusted food expenses and its interaction with rural poverty rates. While poverty rates accounted for most of the satisfactory performance in approximation, the estimated effects from food price variations failed to significantly contribute to precision of the model. This could be attributed to several reasons, including possible violations of the assumptions behind the approach and small sample size due to missing data, which will be further elaborated in the next section. In addition, inaccuracy of the official data of PoU used for validation can also lead to insensitivity to price variations over the years in estimation results, since in some years the official data might be computed based on data from previous years.

	Dependent variable:			
	undernourishment			
	(1)	(2)	(3)	(4)
food_expenses	-1.933	-0.804	-0.782	
	(2.919)	(2.893)	(2.712)	
access_water	-0.001	-0.027	-0.028	-0.031
	(0.051)	(0.052)	(0.042)	(0.042)
percent_poor	1.809***			
	(0.373)			
rural_poor		1.596	1.661***	1.762***
		(1.007)	(0.297)	(0.311)
urban_poor		0.095		
		(1.392)		
rural_poor:food_expenses				-0.045
				(0.175)
Constant	-6.789	-11.409	-11.503	-13.142
	(11.645)	(11.677)	(10.937)	(7.216)
Observations	13	13	13	13
R2	0.960	0.968	0.968	0.967
Adjusted R2	0.946	0.951	0.957	0.957
Residual Std. Error	1.017 (df = 9)	0.970 (df = 8)	0.914 (df = 9)	0.915 (df = 9)
F Statistic	71.638*** (df = 3; 9)	59.627*** (df = 4; 8)	89.387*** (df = 3; 9)	89.225*** (df = 3; 9)
Note:				*p<0.1; **p<0.05; ***p<0.01

Figure 3.3

Discussions

- Strength

The main advantage of this approach is that it testifies to the strong performance of approximating prevalence of undernourishment from rural poverty rates, and proposes to adjust

for the uncaptured patterns by taking into account the effects of variations in real prices of staple foods consumed by the lower-income population. Both the poverty rates and the staple food prices data do not heavily rely on survey data that cannot be collected on an annual basis, and are relatively more prone to change than the other indicators used in the derivation process, such as food consumption patterns and population age structure. Compared with the originally adopted measurement approach, which utilizes mean dietary energy consumption and its coefficient of variation as well as assumes lognormal distribution of dietary energy intake levels among population, this approach is not directly based on household-level food consumption data, and does not make assumptions about the form of the distribution function, leading to large potential in improving timeliness and accuracy.

- Limitations and possible further study

There are also several limitations to this study. First, due to data unavailability, the sample size was too small - there were only 13 observations for 13 years, making it less feasible to conduct the standard train-test split in machine learning algorithms to get a more convincing predictive accuracy score. Second, the effectiveness of the "undernourishment poverty line" depends on multiple assumptions, the major one of which is that people under threats of undernourishment in the studied area should be faced with the same price levels, and furthermore, conditional on price, their access to the staple foods should be equal. Upon satisfaction of this assumption, most supply-side risks in cultivation, stockbreeding, etc., if resulted in fewer supplies, should be properly reflected in an increase in the calculated food price per kcal. This also means people with sufficient purchasing power should be spared from undernourishment. The model tried to mitigate the deviations caused by unequal accessibility by including the percentage of population with high-quality drinking water, which serves as a proxy negatively correlated with the proportion of population without equal access to the average-priced foods. However, it is a vague proxy for this purpose, and the resulting errors from violation of the assumption remain to be resolved. In addition, the model also assumes that the basket of selected staple foods should constitute the main energy sources for the studied population, which might not completely correspond to the real conditions.

There are several directions that the Joint SDG Fund could proceed with more available data. First, it will be feasible to conduct the same analysis using region disaggregation data within a country, or to focus on one or more particular regions of interest. At present, regional data on food price and poverty rates are readily accessible, while the data on prevalence of undernourishment among local populations are not publicly available, posing impediments to model training and validation. However, since relevant survey data were initially collected on the regional level, there should be existing records for the datasets needed that were not published. In a regional scenario, food prices and consumption data can provide more accurate information than national data on a smaller population, hence making the aforementioned assumptions more

valid and yielding higher expected accuracy. Secondly, government subsidies on food could be factored into the model to increase precision. For the Indonesian case, there is a large-scale nationally implemented rice subsidy program known as the “Raskin” program, whereas the annual data for which is not currently available. Lastly, this approach can also be potentially applied to other developing countries with relatively developed food markets.

IV. Indicator 2.2.2

Scope and Theoretical Support

- Background

SDG 2.2.2 measures the prevalence of malnutrition among children under 5 years of age, by type (wasting and overweight). Being underweight (wasting) or overweight are both defined as malnourished. According to the criteria developed by the United Nations, a child is defined as "wasted" if their weight-for-height is more than 2 standard deviations below the median of the WHO Child Growth Standards. A child is defined as "overweight" if their weight-for-height is more than 2 standard deviations above the median of the WHO Child Growth Standards. This indicator was chosen due to its similarity to the SDG 2.1.1. The team attempted several variables to approximate SDG 2.1.1, but all were hindered by the same problems of collecting a regularly updated dataset. Educational and climate data are a good source to make approximations based on, due to their availability and timeliness; however, they tend not to map strongly onto SDG 2.1.1, but can be quite predictive for SDG 2.2.2 where the scope is narrowed down to children under 5. Also, the significant amount of missing data as well as lags in reporting of SDG 2.2.2 make the work valuable to help the UN SDG Fund by providing alternative measures for this goal. Hopefully this proxy can be used for future network analysis with other indicators of interest.

- Rationale

In this study, educational-related data concerning female adults and climate data will be used as alternative measurements. Reports such as the *Malawi: Nutrition Profile* have shown that malnutrition in babies is related to the education level of their mothers. For instance, in Malawi, according to the most recent Demographic and Health Survey (DHS), among children aged 6-23 months born to mothers with no education, only five percent receive the minimum acceptable diet; this number increases to 13 percent among mothers with secondary education. The report also showed that childbearing begins early in Malawi. By age 19, 59.2 percent of adolescent girls had begun childbearing in 2015–2016. In addition to educational data, the monthly climate data as measured by the deviation from the mean temperature was incorporated as a proxy indicator since the six target countries, i.e. Barbados, Cambodia, Indonesia, Malawi, Rwanda, and

Uzbekistan rely heavily on agriculture, where production efficiency could be vulnerable to climate change.

- Countries of Focus & Years of Study

The proxies chosen for this study, educational and climate change data, may not be applicable to all countries as many developed countries are unlikely to have acute food security issues or higher illiteracy rates. Hence, countries with malnutrition levels of children under 5 that are likely to be affected by the proxies. Given this understanding, the models will be validated using only low-income countries. Moreover, there are many missing values for the SDG 2.2.2 records on the target countries of interest by the UN. Any significant pattern will narrow down the scope to target the six countries of interest to further investigate the patterns. The study will focus on a 20-year horizon from 2000 to 2019 considering the completeness of the dataset for proxies made available.

Data Descriptions and Preprocessing

- Data Source Descriptions

The following datasets will be used for this paper:

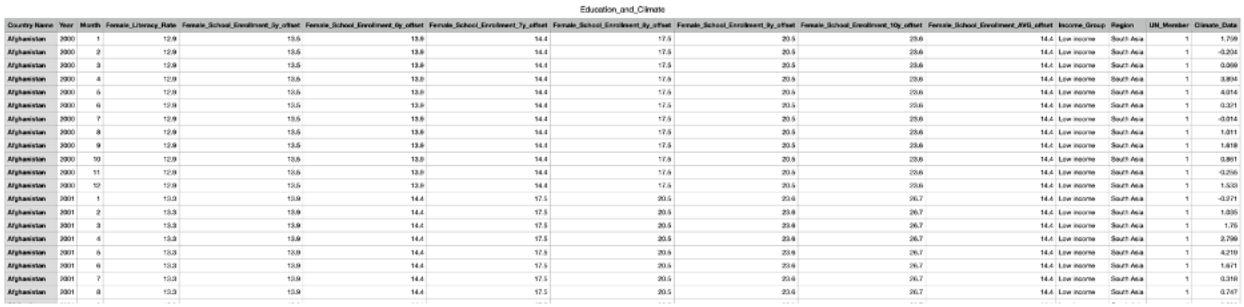
- SDG 2.2.2 Dataset from the UN
- Female Literacy Rate from the World Data Bank
- Female School Enrollment Rates from the World Data Bank
- Climate Data (temperature deviation from the average of temperature from 1980-2005) by month

- Data Preprocessing

Datasets for all countries were extracted and reorganized into a single csv file. The column names include:

- Country Name - The country's name
- Year - Year the data is from, from 2000 to 2019
- Month - Month of the year for the climate data
- Female_Literacy_Rate - Percent of the female population over the age of 15 who can read
- Female_School_Enrollment_5y_offset - Percent of girls enrolled in primary school 5 years prior to the year in question
- Female_School_Enrollment_6y_offset - see above, but based on data 6 years prior
- Female_School_Enrollment_7y_offset - see above

- Female_School_Enrollment_8y_offset - see above
- Female_School_Enrollment_9y_offset - see above
- Female_School_Enrollment_10y_offset - see above
- Female_School_Enrollment_AVG_offset - The arithmetic mean of the previous 6 columns
- Income_Group - The UN income classification of a country, including: Low Income, Middle low income, Middle high income, High income: non-OECD, and High income: OECD
- Region - The global region of the country
- UN_Member - 1 for countries that are members of the UN, 0 otherwise
- Climate_Data - The monthly standard deviation distance from the 1980-2005 mean temperature for a month (see Figure 4.1).



Education_and_Climate

Country Name	Year	Month	Female_Literacy_Rate	Female_School_Enrollment_5y_offset	Female_School_Enrollment_6y_offset	Female_School_Enrollment_7y_offset	Female_School_Enrollment_8y_offset	Female_School_Enrollment_9y_offset	Female_School_Enrollment_10y_offset	Female_School_Enrollment_AVG_offset	Income_Group	Region	UN_Member	Climate_Data
Afghanistan	2000	1	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	1.759
Afghanistan	2000	2	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	-0.204
Afghanistan	2000	3	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	0.999
Afghanistan	2000	4	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	3.834
Afghanistan	2000	5	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	4.014
Afghanistan	2000	6	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	0.321
Afghanistan	2000	7	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	-0.014
Afghanistan	2000	8	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	1.011
Afghanistan	2000	9	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	1.818
Afghanistan	2000	10	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	0.891
Afghanistan	2000	11	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	-0.256
Afghanistan	2000	12	12.9	13.5	13.9	14.4	17.5	20.5	23.6	14.4	Low income	South Asia	1	1.533
Afghanistan	2001	1	13.3	13.9	14.4	17.5	20.5	23.6	26.7	14.4	Low income	South Asia	1	-0.271
Afghanistan	2001	2	13.3	13.9	14.4	17.5	20.5	23.6	26.7	14.4	Low income	South Asia	1	1.035
Afghanistan	2001	3	13.3	13.9	14.4	17.5	20.5	23.6	26.7	14.4	Low income	South Asia	1	1.75
Afghanistan	2001	4	13.3	13.9	14.4	17.5	20.5	23.6	26.7	14.4	Low income	South Asia	1	2.799
Afghanistan	2001	5	13.3	13.9	14.4	17.5	20.5	23.6	26.7	14.4	Low income	South Asia	1	4.219
Afghanistan	2001	6	13.3	13.9	14.4	17.5	20.5	23.6	26.7	14.4	Low income	South Asia	1	1.671
Afghanistan	2001	7	13.3	13.9	14.4	17.5	20.5	23.6	26.7	14.4	Low income	South Asia	1	0.316
Afghanistan	2001	8	13.3	13.9	14.4	17.5	20.5	23.6	26.7	14.4	Low income	South Asia	1	0.747

Figure 4.1

Because a significant amount of educational data was missing, linear imputation was used between data points to estimate data that otherwise would not have been usable. While this adds some complication, it should be valid as the data is in a time series format. The column “Climate_Data” has been categorized into monthly climate data. The 12 new columns represent each month of the year, and have been added to the target and educational data. Figure 4.2 shows the head of the final dataset that will be used for modeling and the validation process.

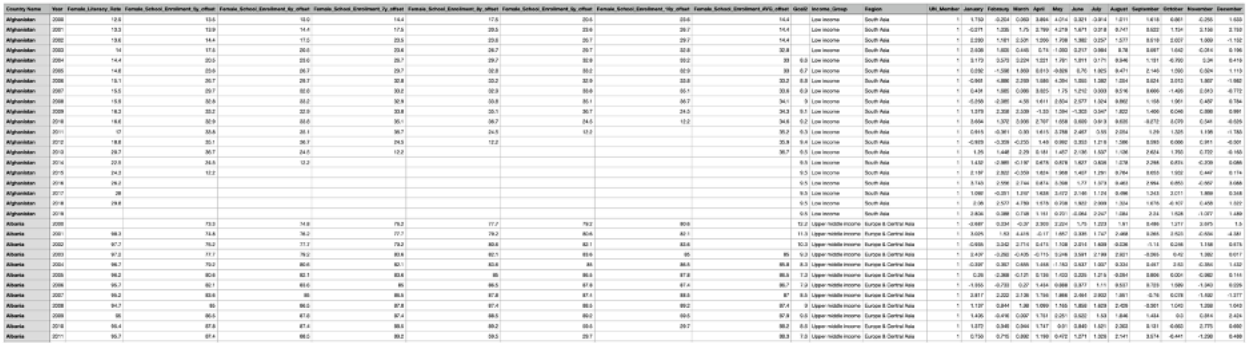


Figure 4.2

- Exploratory Data Analysis

All numeric columns in the dataset were standardized using the StandardScaler to eliminate the effect of differences in the range of data for each column. The correlation matrix heatmap is plotted as shown in Figure 4.4. Notice that the enrollment offset variables are renamed into “offset_5y”, “offset_6y”, “offset_7y”, “offset_8y”, “offset_9y”, and “offset_10y” respectively for convenience. Low-income countries were chosen to be investigated first, as they have more complete data for SDG 2.2.2 and are more relevant to the UN, as these countries have higher rates of malnutrition and all six of the countries of interest are low-income. Figure 4.3 illustrates that the average percent of children who are malnourished by year for each income group of a country with 95% confidence intervals.

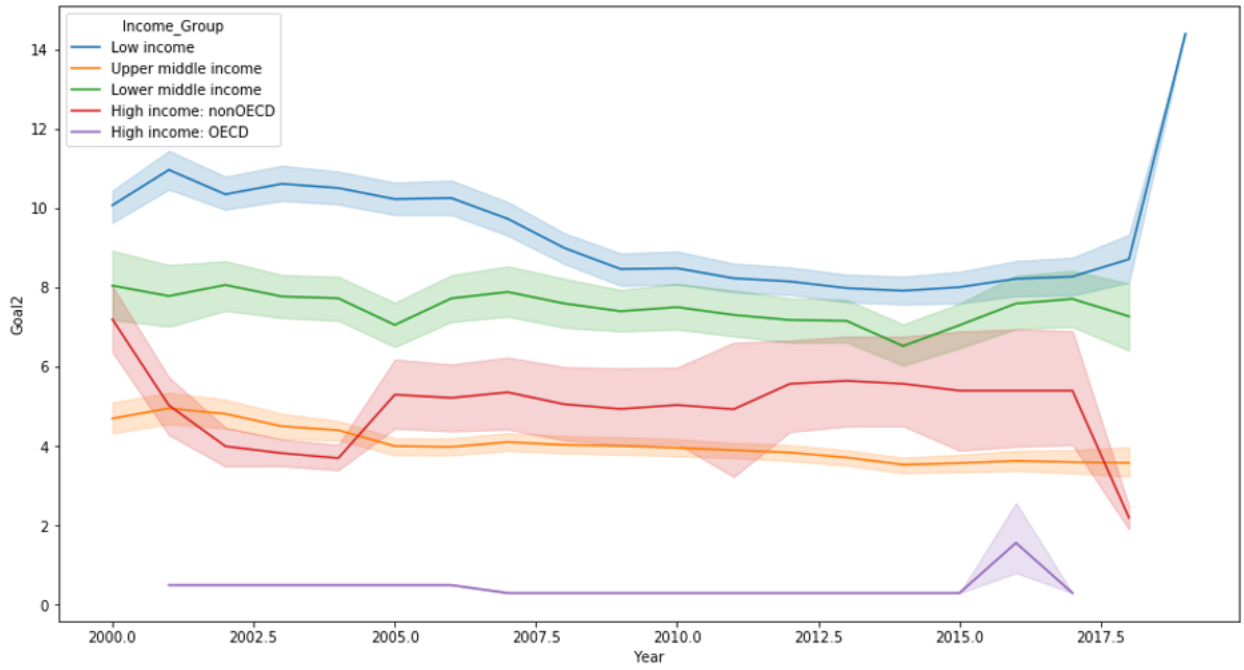


Figure 4.3

As seen from the correlation matrix in Figure 4.4, variables that have relatively high correlations with the SDG 2.2.2 proxy used in this study are “Year”, and “Female_Literacy_Rate”. These variables will be used as the base model of this study. For the other variables, it is hard to draw conclusions from the matrix heatmap. Some other interesting findings can be seen from the climate data. While the correlations are extremely low for some months, the months “January” (correlation = -0.12), “April” (correlation = 0.16), “June” (correlation = -0.15), and “July” (correlation = -0.11) have relatively more significant correlations with “Goal 2” compared to the other months, which approach to 0.

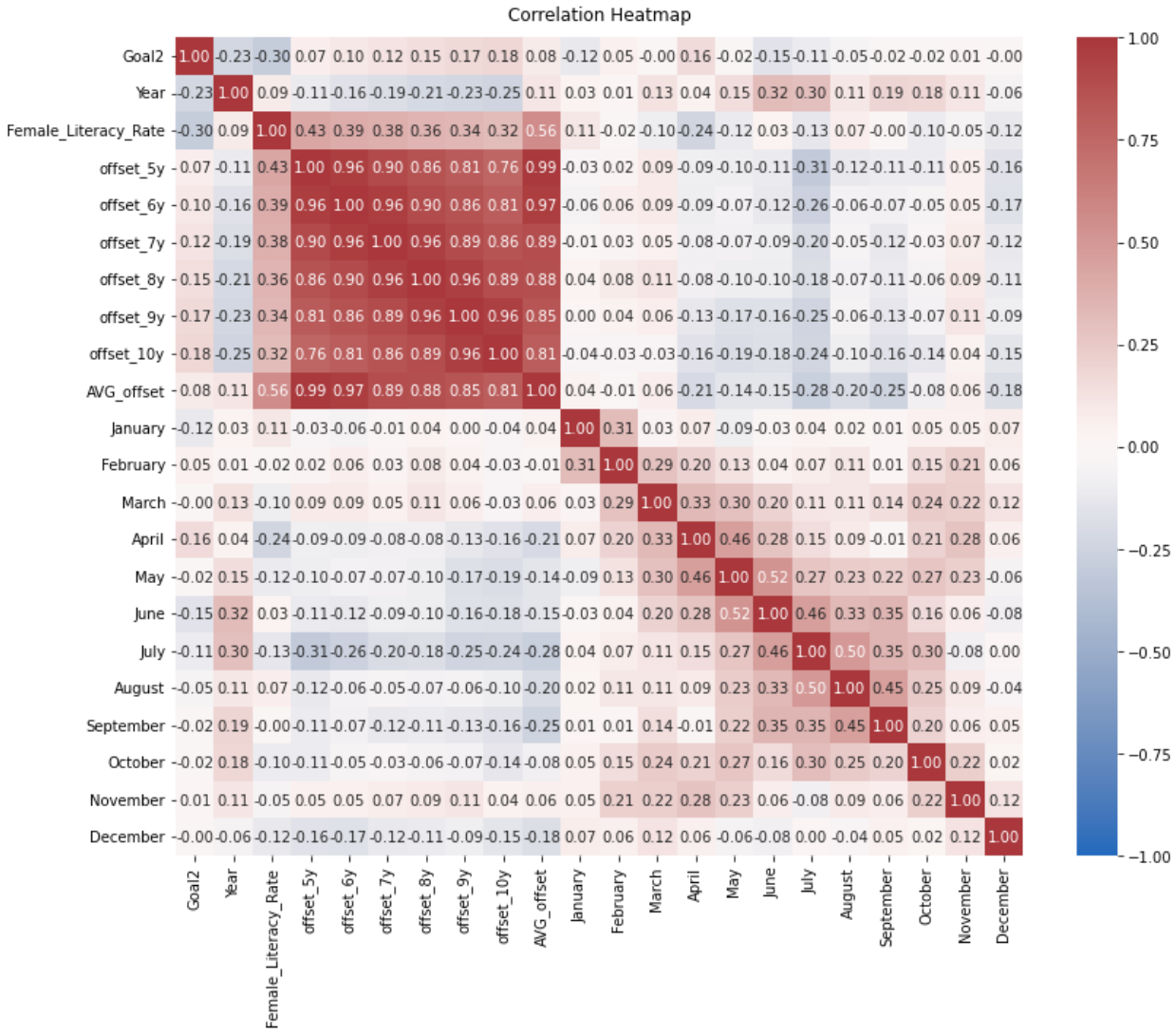


Figure 4.4

Methodology

- Modeling Algorithm

The team ran a multi-linear regression to validate the chosen proxies and see how closely they approximate the original measure of SDG 2.2.2. Ordinary Least Squares regression (OLS) was used as a model for analysis. Four different models based on the variables of interest were constructed, and an ANOVA test was used to select the optimal model out of the four. The four models were:

Model 1: Goal2 ~ Female_Literacy_Rate + Year

Model 2: Goal2 ~ Female_Literacy_Rate + AVG_offset + Year

Model 3: Goal2 ~ Female_Literacy_Rate + January + February + March + April + May + June + July + August + September + October + November + December + Year

Model 4: Goal2 ~Female_Literacy_Rate + offset_5y +offset_6y +offset_7y +offset_8y +offset_9y +offset_10y +January + February + March + April + May + June + July + August + September + October + November + December + Year

Based on the ANOVA test (p-value .001991), Model 3 was selected as the final model as it provided a statistically significant improvement over the previous models at the level of .05. The code chunk is shown in Figure 4.5 where "lin_reg_base", "lin_reg_ed", "lin_reg_cl", and "lin_reg_all" correspond to Model 1, Model 2, Model 3, and Model 4 respectively.

```
sm.stats.anova_lm(lin_reg_base, lin_reg_ed)
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	502.0	13.354723	0.0	NaN	NaN	NaN
1	241.0	6.717922	261.0	6.636801	0.912222	0.766833

```
sm.stats.anova_lm(lin_reg_base, lin_reg_cl)
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	502.0	13.354723	0.0	NaN	NaN	NaN
1	488.0	12.463742	14.0	0.890981	2.491791	0.001991

```
sm.stats.anova_lm(lin_reg_base, lin_reg_all)
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	502.0	13.354723	0.0	NaN	NaN	NaN
1	228.0	6.300210	274.0	7.054513	0.931743	0.712785

```
sm.stats.anova_lm(lin_reg_cl, lin_reg_all)
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	488.0	12.463742	0.0	NaN	NaN	NaN
1	228.0	6.300210	260.0	6.163532	0.857899	0.884442

Figure 4.5

- Validation Process & Results

OLS Regression Results

Dep. Variable:	Goal2	R-squared:	0.187
Model:	OLS	Adj. R-squared:	0.164
Method:	Least Squares	F-statistic:	8.017
Date:	Sun, 05 Dec 2021	Prob (F-statistic):	1.76e-15
Time:	01:34:29	Log-Likelihood:	216.26
No. Observations:	503	AIC:	-402.5
Df Residuals:	488	BIC:	-339.2
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5997	0.074	8.092	0.000	0.454	0.745
Female_Literacy_Rate	-0.1773	0.032	-5.579	0.000	-0.240	-0.115
January	-0.2540	0.085	-2.996	0.003	-0.421	-0.087
February	0.1417	0.073	1.950	0.052	-0.001	0.284
March	-0.0708	0.070	-1.013	0.312	-0.208	0.067
April	0.2503	0.063	3.980	0.000	0.127	0.374
May	-0.0936	0.083	-1.126	0.261	-0.257	0.070
June	-0.1615	0.080	-2.007	0.045	-0.320	-0.003
July	-0.1124	0.060	-1.870	0.062	-0.230	0.006
August	0.0352	0.071	0.494	0.621	-0.105	0.175
September	0.1366	0.070	1.960	0.051	-0.000	0.274
October	-0.0039	0.082	-0.048	0.962	-0.164	0.156
November	-0.0501	0.064	-0.785	0.433	-0.176	0.075
December	-0.0719	0.055	-1.312	0.190	-0.180	0.036
Year	-0.1011	0.030	-3.335	0.001	-0.161	-0.042

Omnibus:	50.847	Durbin-Watson:	0.391
Prob(Omnibus):	0.000	Jarque-Bera (JB):	63.944
Skew:	0.815	Prob(JB):	1.30e-14
Kurtosis:	3.630	Cond. No.	29.2

Figure 4.6

As seen from the summary statistics above in Fig 4.6, the estimated coefficients for “Female_Literacy_Rate”, “January”, “April”, “June” and “Year” are statistically significant with their p-values all less than 0.05. The adjusted R-squared for the model is .164, suggesting that 16.4% of the variation in the SDG 2.2.2 value can be explained by these variables. This seems to be reasonably strong given that not all low-income countries rely on agriculture heavily. It is also interesting to find that only climate data in January, April and June are significant, which may be attributed to the seasonal pattern of crop yields. The scatter plot in Figure 4.7 also showed a linear pattern between the SDG 2.2.2 value and the predicted values of independent variables.

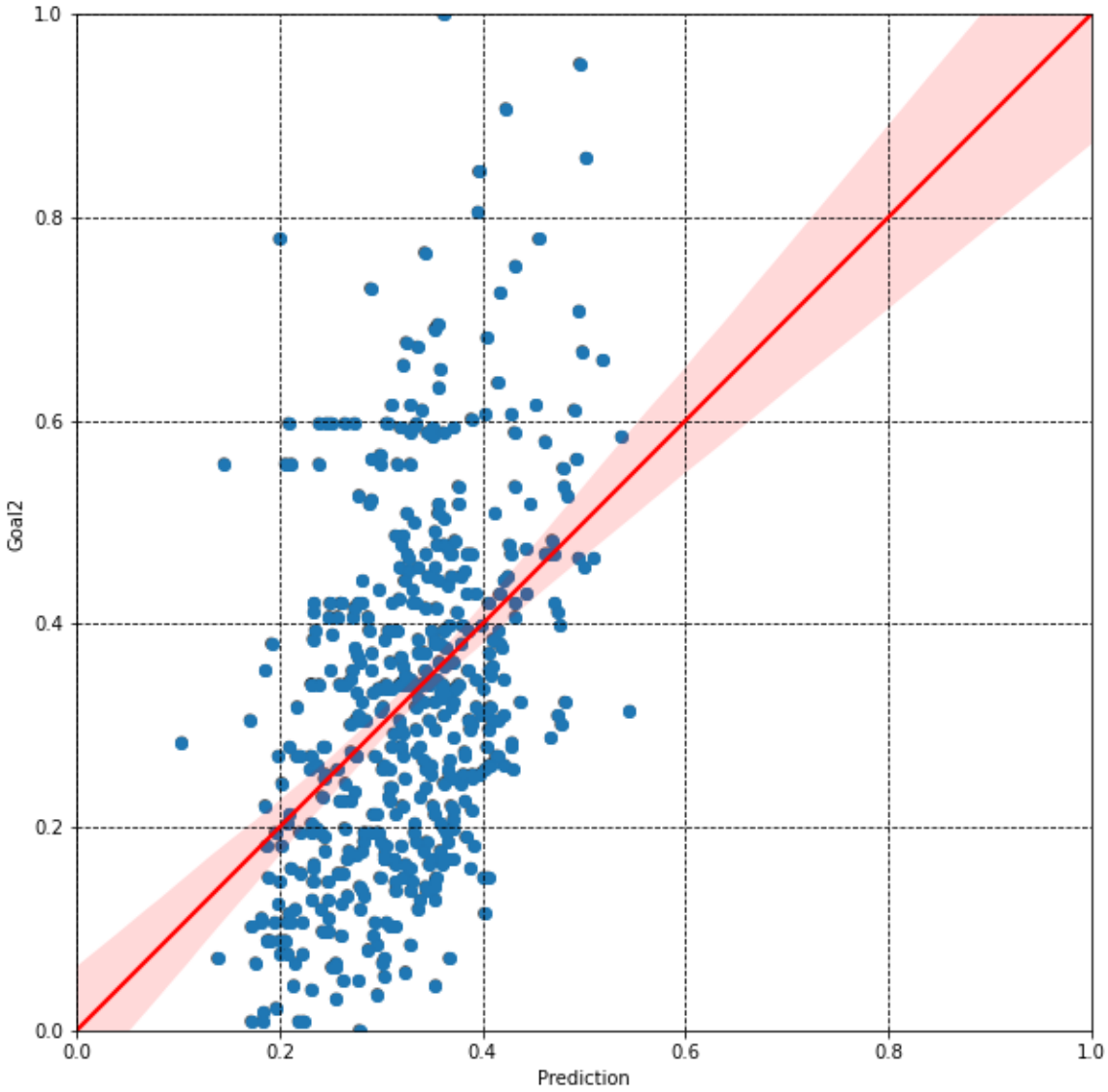


Figure 4.7

Filtering the data further, Figure 4.8 shows the results from running the analysis filtered to the six target countries of interest. The adjusted R-squared for the model increased from .164 to .379, suggesting that 37.9% of the variation in SDG 2.2.2 for these countries can be explained by the variables. As for the scatter plot, due to the limitations of available data points, the linear relationship becomes less obvious.

OLS Regression Results

Dep. Variable:	Goal2	R-squared:	0.477
Model:	OLS	Adj. R-squared:	0.379
Method:	Least Squares	F-statistic:	4.885
Date:	Sun, 05 Dec 2021	Prob (F-statistic):	2.58e-06
Time:	02:18:35	Log-Likelihood:	20.301
No. Observations:	90	AIC:	-10.60
Df Residuals:	75	BIC:	26.90
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.9330	0.277	3.362	0.001	0.380	1.486
Female_Literacy_Rate	0.3254	0.080	4.063	0.000	0.166	0.485
January	-0.2286	0.270	-0.847	0.400	-0.767	0.309
February	-0.0190	0.201	-0.094	0.925	-0.420	0.382
March	-0.3004	0.185	-1.626	0.108	-0.668	0.068
April	0.3122	0.262	1.193	0.237	-0.209	0.834
May	-0.1234	0.275	-0.449	0.655	-0.671	0.424
June	-0.5131	0.264	-1.941	0.056	-1.040	0.013
July	-0.1901	0.184	-1.036	0.304	-0.556	0.175
August	-0.2081	0.138	-1.504	0.137	-0.484	0.068
September	-0.1164	0.161	-0.721	0.473	-0.438	0.205
October	-0.3495	0.201	-1.735	0.087	-0.751	0.052
November	0.0849	0.184	0.462	0.646	-0.281	0.451
December	0.1851	0.207	0.895	0.373	-0.227	0.597
Year	-0.1153	0.098	-1.171	0.245	-0.312	0.081

Omnibus:	1.122	Durbin-Watson:	1.015
Prob(Omnibus):	0.571	Jarque-Bera (JB):	1.002
Skew:	0.002	Prob(JB):	0.606
Kurtosis:	2.483	Cond. No.	39.1

Figure 4.8

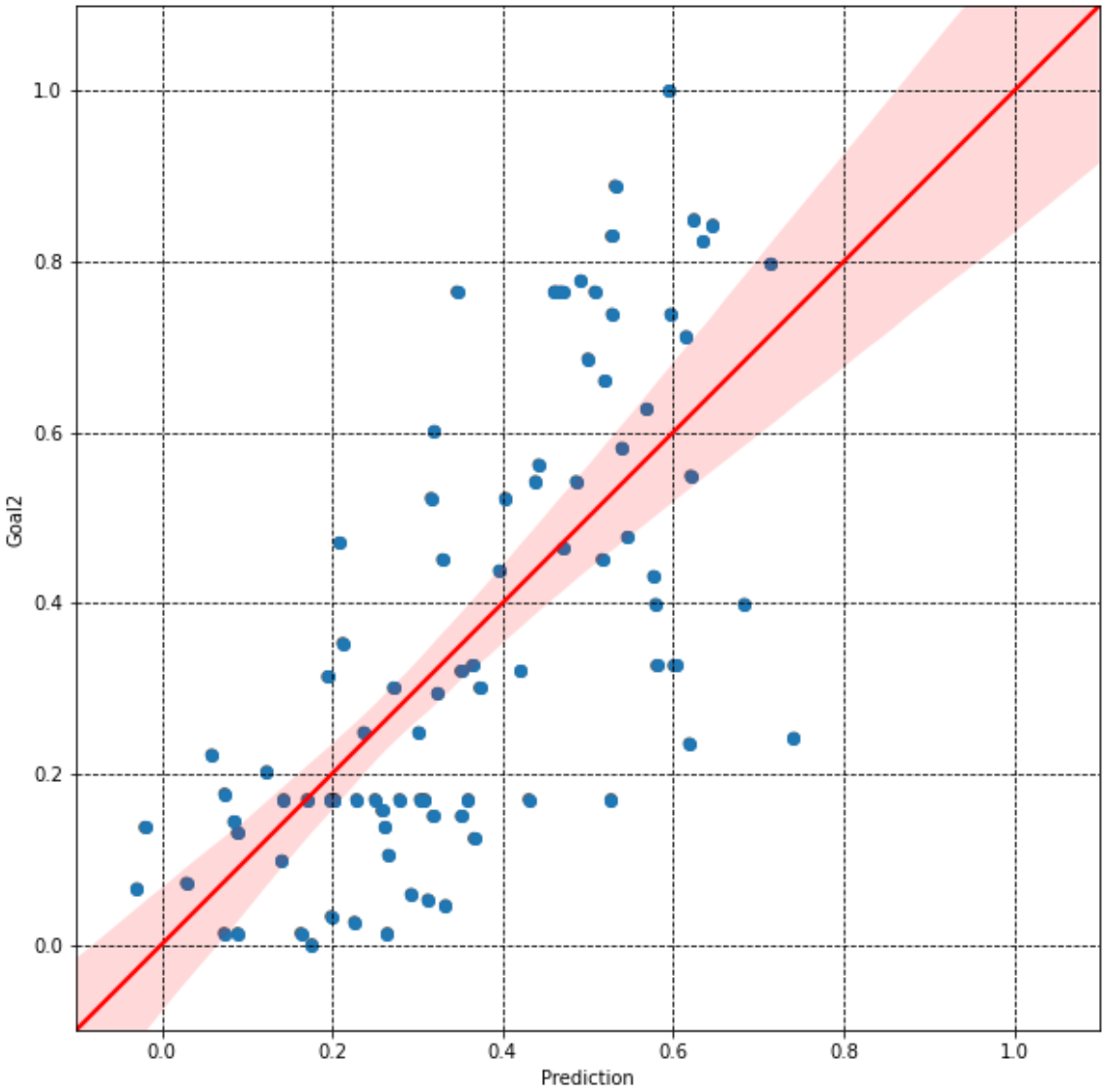


Figure 4.9

Discussions

- Strength

The selected proxies are more readily available and have more complete datasets compared to the currently available dataset for SDG 2.2.2. The new dataset fills in the missing gap by providing more recent and consistent data, whereas the alternatives contained data limited to one or two years or data that was not recent enough. In recent years, educational data for low-income countries generally and the 6 target countries in particular are becoming increasingly accessible, meaning this proxy has potential as an alternative indicator for SDG 2.2.2. Furthermore, the climate data is always up to date, meaning missingness is not an issue of

concern, which was a major problem for the original measurement adopted by the UN. These data not only served as proxies for the goal, but can also point to potential causes of delay in achieving the goal. It provides a level of understanding regarding the bottlenecks issue for each country. For example, a country underachieving its SDG 2.2.2 may be because of the persistent rise in the temperature or a series of agricultural disasters. The wide availability and relative strength of the model as measured by the adjusted R-squared value provide justification to use this model.

- Limitations

Despite the strengths, a major challenge for these proxies is that not all “low-income countries” put great emphasis on agriculture, making the overall inferential analysis less significant than expected. However, results derived from the target countries are reasonably strong. Further research can be conducted on climate data for specific months, as the analysis reveals that some months have statistically significant patterns while the others do not. Since the six target countries rely heavily on agriculture, the crop production may be more significantly impacted by certain months, which may explain this. Furthermore,climate data is very complex and other types of indicators such as the level of rainfall may be more significant in specific countries.

V. Indicator 10.1.1

Data Description and Preprocessing

Allowing the strategy to build the model to be entirely data-driven and based on any detected relationship, the team used a multitude of data sources in their investigation into indicator 10.1.1. These included both official sources such as the UN and World Bank as well as other types of organizations including Our World in Data. Further, some of the team’s research into various aspects and the relationship they may have with the income share of the bottom 40% was driven by methodology and surveys used by the Poverty Probability Index.

<u>Data source:</u>

- UN Stats - official source of data regarding SDG Indicators
- World Bank, Development Research Group. Data are based on primary household survey data obtained from government statistical agencies and World Bank country departments. Data for high-income economies are from the Luxembourg Income Study database.

  This dataset is chosen since it contains a large amount of data for many variables and covers a wide range of countries all over the world, which would be ideal for constructing a universal methodology.

- International trade center. Trade Map provides trade statistics and market access information for export development. It provides indicators on export performance, international demand, alternative markets and the role of competitors, and covers yearly trade data for 220 countries and territories and all 5,300 products of the Harmonized System.
- Our World in Data. Often combining several sources and datasets from governmental organizations and research institutions, Our World in Data makes available large amounts of information covering myriad variables and aspects of countries' economies or societies. From this site, the team recovered data regarding countries' populations' years of schooling, the Gini Coefficient, and government expenditures.

Variables:

- Dependent variable (Target): Household expenditure or income share held by the bottom 40 percent of the population

  The dependent variable can be obtained by aggregating the variables "household income share per capita held by lowest 20% of the population" and "household income share per capita held by fourth 20% of the population" in the World Bank dataset.

- Independent variables: Gini coefficient; education; government expenditures, etc.

  These variables are chosen mainly based on the correlation between them and the dependent variable, using the quasi-machine learning method. In other words, they have the highest correlation with the dependent variable.

Methodology

Attempting to find a functional proxy for indicator 10.1.1, the team took a data-based approach, allowing the availability of data to largely guide the exploration. This method took two central forms: exploration based on theory and that driven by discovered relationships with the income share of the bottom 40%. The team generally sought relationships between variables and the nominal annual income share of the bottom 40%, while the indicator, by definition, is the growth rate of this share. This is due to the significant lack of data present in the growth rate data, and if the team could build an effectively predictive model of the income share, then growth rates could be subsequently calculated by finding the growth rate of the predicted values of the income share.

The first form of exploration saw the team conduct initial searches for various aspects of a country's economy or societal features that could be related to the income share of the bottom 40%. If data could be found regarding these aspects, they were tested against the set of data about the income share of the bottom 40% that corresponded by country and time period to

identify what relationship exists between the two measures, largely through finding the correlation between them. If variables were found to have a significant relationship, which was generally decided intuitively via strong correlations, more complete data about these variables was sought and kept for possible inclusion in the final model. Though this exploration was largely directed by theory, the approach remains data-based in that which variables would be included in the final model was based on data availability and predictive functionality. The results from this stage can be seen in the table below.

Variable	Correlation with Income Share of the Bottom 40%
Gini Coefficient	-0.9798089
Proportion of Consumption on Food and Accomodation	-0.135095
Proportion of Consumption on Clothing	-0.2754005
Loan Default Rate	0.547835
Expenditure on Durable Goods	-0.9360472
Number of Commercial Bank Branches	0.2352157
Number of McDonald's	-0.05808015
Incarceration Rate	-0.182441
Trade data	Ranging from .001 to .21 for all countries; .31 to .83 for target countries
Amount of missing data in World Bank	<.011

Figure 5.1

The second form of exploration, discovered relationships, utilized the large amount of data covering hundreds of variables available through the World Bank. After gathering data regarding the income share of the bottom 40% for each of the world's countries between 1970 and 2019, several measures of the correlation between this and each of the variables included in the World Bank's data were found. For each variable, the measures of correlation found included the nominal correlation between the income share of the bottom 40% and the variable, the correlation between the annual percent change of the income share and the nominal value of the variable, and the correlation between the annual percent change of both the income share and the given variable. Using these results, variables that seemed to have a strong relationship with the income share of the bottom 40% were explored further, attempting to find additional and more complete data to use as predictors of the income share. Variables' relationship with the income share was judged generally intuitively based on several factors including the value of the correlation coefficients and the correlation's persistent strength across the different measures of the correlation.

The team tested many variables for their relationship with the income share of the bottom 40%. Through the first form, some of the variables thought to have a potential relationship with the indicator included the Gini Coefficient, the proportion of consumption on clothing, food, and

housing, the loan default rate, the expenditure share on durable goods, the number of commercial bank branches in a country, the number of McDonald's locations in a country, the incarceration rate, and the amount of imports or exports of several commodities. Correlations between these and the income share ranged in strength between .058 and .98. The strongest relationship found in this stage was the income share's correlation with the Gini Coefficient, and this measure of the inequality present in a country was included in the final model. While several other variables were found to have significant relationships with the income share of the bottom 40%, most were not included in the final model due to a lack of available data or low predictive power. This is particularly relevant, as the team conducted complete observation analysis, meaning any observation that lacked data for any variable included in the model would be omitted from analysis. Thus, not only was missingness a central challenge, but variables that contained missing data in differing patterns could largely not be included in the model, as inclusion would decrease the sample size below feasibility. Notable variables that were subject to this problem and therefore excluded from the model but could potentially be strong predictors if greater amounts of data could be utilized included the proportion of expenditure on durable goods and imports of commodities such as meat and coffee.

The second method of exploration, utilizing the World Bank data, uncovered some additional variables that have strong relationships with the income share of the bottom 40%. Some of the variables that were found here to merit further exploration included the Gini Coefficient, variables regarding education levels, governmental social protection coverage, and child labor rates. Net flows from UN agencies were also strong in their relationship with the income share of the bottom 40%, but this was excluded from analysis, as its effectiveness was dubbed dubious given the hopeful use of the predictive model. Of these, the final model would include the Gini Coefficient, average education, and governmental expenditures, of which social protection programs are a fraction by definition.

In building the final model, the team used a manual stepwise regression technique, through which they added variables as predictors in various combinations to evaluate the effectiveness of the model. For this, they used all data available for all years between 1970 and 2019 and for all countries, allowing for the maximum amount of data to be utilized in building the model. In each inception of the model, the team built a linear regression model based on 85% of the data for which all variables were recorded, and tested the model's effectiveness by evaluating the predicted values and actual values of the income share of the bottom 40% on the remaining 15% of the data. If, through the validation process, the model was found to effectively predict the values of the income share of the bottom 40%, then it can be used to fill some of the current gaps that exist in the indicator data. Of the 15% used for validation, this could be reduced to only the six countries of interest (Malawi, Rwanda, Indonesia, Uzbekistan, Barbados, and Cambodia) if desired, and analysis of both the full set and only that containing these six countries was performed. Variables' ability to improve the model were judged by the model's

adjusted R^2 value, number of observations remaining in the model, and the relationship between the model's predicted values and actual values. As noted previously, the inclusion of some variables led to the number of observations included in the analysis to diminish below effectiveness, meaning these variables could not be included based on the data with which the team was working. Through this process of including variables as data was added and evaluating the model's predictive ability, the team's methodology became fully data-driven.

The model's quality was determined based on its predictions of the values of income share of the bottom 40% in the 15% of the data not used to build the model. After randomly breaking the full dataset into the two sections, the team judged the model's predictions based on the maximum error, meaning the maximum difference between an observation's predicted value and actual value. They also used a visual of the plotted predicted and actual values to gain an intuitive understanding of the model's quality. These validation processes help determine the success of the model's ability to address the missingness, and potentially the time-lag, of data that currently exist.

Through the process of validation evaluation, the team arrived at the final model, a linear model of the income share of the bottom 40% with the predictors being the Gini Coefficient, the average number of years of education of a country's population in a given year, and the total amount of annual government expenditures by country. This model was chosen, as it allowed for a relatively large number of observations to remain in the model after omitting observations missing at least one of the included variables, and because it held a relatively strong R^2 value just over 0.74. In various simulations of the model, taking different groups of the data as the test and validation sets, the maximum difference between the predicted and actual value of the income share held by the bottom 40% ranged between 0.03 and 0.09, corresponding to a difference of between 3 and 9 percentage points. Using these variables, 1,241 observations remained in the data used to build the model and estimate the coefficients, and 218 observations were included in the set used for validation and evaluation. Here, an observation corresponds to data for an individual country in an individual year.

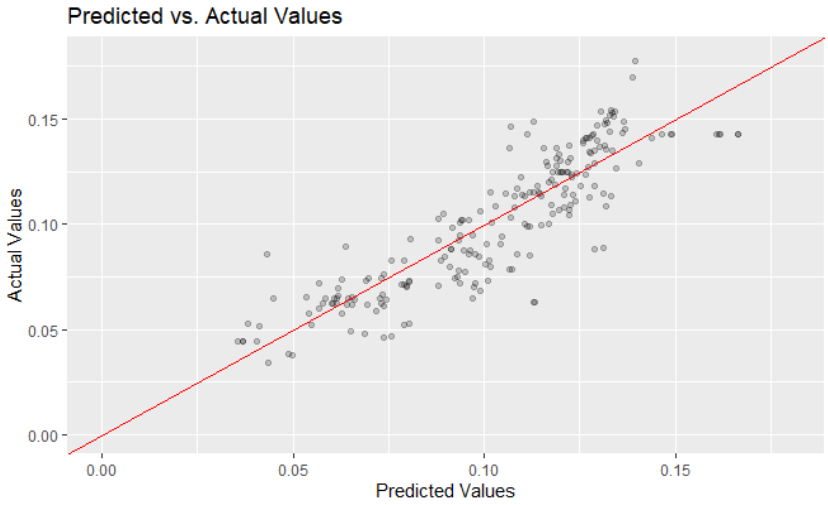


Figure 5.2

```
Call:
lm(formula = Income_Share_Bottom_40 ~ Gini.index + SchoolYears +
    govexpend, data = train)

Residuals:
      Min        1Q    Median        3Q       Max
-0.057992 -0.009505  0.000706  0.010049  0.057774

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.657e-01  3.870e-03   42.82  < 2e-16 ***
Gini.index  -2.486e-03  6.443e-05  -38.59  < 2e-16 ***
SchoolYears  1.190e-03  2.004e-04    5.94  3.7e-09 ***
govexpend    6.267e-04  4.427e-05   14.16  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01651 on 1236 degrees of freedom
Multiple R-squared:  0.7427,    Adjusted R-squared:  0.7421
F-statistic:  1189 on 3 and 1236 DF,  p-value: < 2.2e-16
```

Figure 5.3

While the initial results seem relatively promising for predicting the income share of the bottom 40% in a given country and year, the predictions of the true indicator, the growth rate of this income share, were less promising. After omission of data for which at least one of the variables used did not exist, only a small number of countries that included recorded data for two consecutive years remained, meaning that, in combination with a large amount of missing data in the official data for indicator 10.1.1, very few predicted and actual values of the growth rate could be compared. In various simulations, this number ranged between 0 and 5. Thus, the strength of the process and results relies almost entirely on the evaluation of the strength of the predictions of the income share of the bottom 40% as a nominal value. As an additional weakness of the results and method, complete observations for the six target countries were relatively rare even when only using these three variables, which contained notably more data for

these countries than most variables. When only including these three predictors, the number of observations for target countries in the set used for validation ranged between two and six.

Predictions for income share among target countries

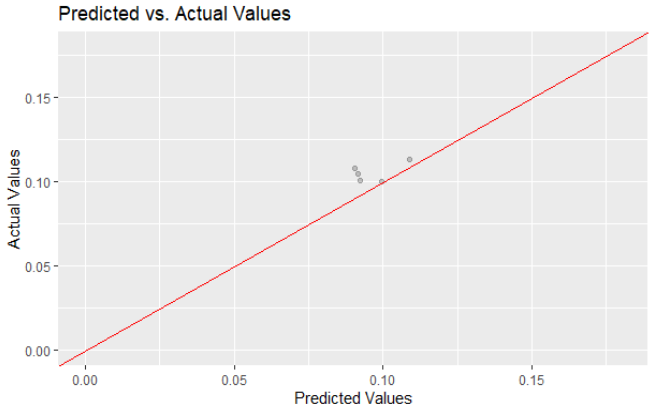


Figure 5.4

Predictions of growth rate

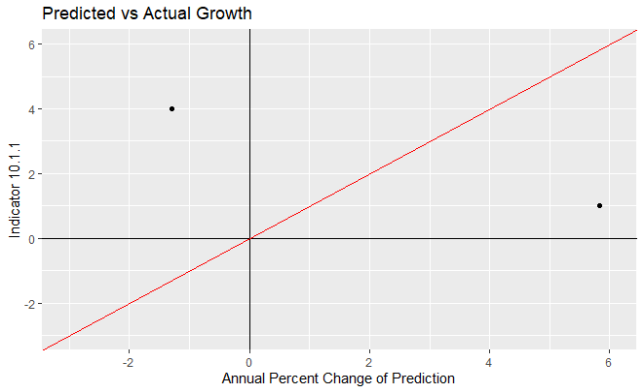


Figure 5.5

Discussions

Given the relative lack of data for indicator 10.1.1, the strength of the model built by the team lies in the predictions and predictive effectiveness of the income share of the bottom 40%. If these are deemed reliable, then predictions of the annual growth rate can be calculated by finding the annual growth rate of the predictions. However, testing the relationship between the predicted and recorded growth rates remains difficult for the same reason as the team's motivation, namely, the scarcity of this data. The model does seem relatively strong in its predictions of the annual value of the income share, but this relies on data for the Gini Coefficient (or potentially other similar measures of inequality), total government expenditures,

and the average number of years of schooling, being available. While these measures are more frequently recorded than the income share of the bottom 40%, they still suffer from patterns of missing data. Therefore, the constructed model may address some of the missingness of data in the indicator's records, but it does have notable limitations.

The primary limitation of the current model is the relatively small amount of data on which it is built and tested. This lack of data does not allow for the inclusion of many variables which also have differing patterns of missingness, if complete observation analysis is to be conducted without imputation of large amounts of data. Additionally, as in any analysis, the relatively small number of observations leads to skepticism of the accuracy of the coefficients, as they are calculated on relatively little data. Additionally, the pattern of missingness in this data is, of course, not random, with data being much more commonly available for developed countries and rare for developing ones including the six target countries. Therefore, it should be noted that the coefficients and model may be strongly influenced by countries that are different in meaningful ways from the countries of interest. Finally, the lack of data leads to relatively many of the predicted values being extrapolations and therefore potentially dubious, as the range of data in the validation set would commonly be greater than that of the data used to build the model, depending on the random selection of the 85% used as the training data.

As noted, one of the results of the small number of observations in the exclusion of several variables that may have, in larger datasets, improved the model's predictive accuracy and reliability. With the current dataset, however, including these variables led to many observations being excluded, driving down the model's overall effectiveness. Not only did the exclusion of some predictors result from the small number of observations, but this occurred in other instances due to data not being available or accessible. For instance, one variable found to have a persistently strong relationship with the income share of the bottom 40% was the rate of child labor in a country and year. However, no large amount of data regarding this could be utilized due to unavailability. Similarly, some trade patterns, as noted above, were found to be seemingly related to the income share of the bottom 40%, but these datasets often contained relatively large amounts of missing data, meaning they could not be effectively included in the model. More complete data on some of these variables and aspects of a country could greatly increase the predictive ability of the model.

Given these limitations, the model is certainly not without potential improvements. However, the model in its initial form does seem to be able to predict the income share of the bottom 40% with relative accuracy based on only three predictors, and this could be used to potentially significantly reduce the amount of missing data present in the indicator's records by calculating the annual change of the predicted values. In its current form, the model has strengths in its relative simplicity generating fairly accurate estimates of the income share held by the

bottom 40%, though the accuracy could likely be greatly improved if greater amounts of data for the predictors could be found and additional predictors included.

VI. Conclusion

This study attempted to find or build proxy measures for indicators 2.1.1, 2.2.2, and 10.1.1 to address the large gaps that exist in the data in recent years. For goal 2.1 staple food prices and climate data were used as proxy indicators. For goal 2.2, educational data regarding female adults and climate data were used. Finally, for goal 10, the team used the Gini Coefficient, the average years of schooling, and government expenditures to approximate the income share of the bottom 40%. The results were promising in all three cases, but the lack of target data, and to some extent missing data in the proxy indicators, resulted in difficulty in validating the results.

The team attempted to approximate goal 2.1.1 using a purely theoretical approach, building a modeling framework that could ideally be applied to a number of different countries in the future. The approach provided strong results but could require more tuning to be precise; the model did not take into account regional differences across the country, which would give additional insight. However, the primary reason these regional differences were not considered was the lack of regional data to validate the model, despite the fact that the proxy indicators could be found at the regional level. The inability to validate results in an indicator that may be useful, but with little chance to demonstrate this fact.

The approach taken for Goal 2.2.2 was somewhat successful, using two easily accessible data points and explaining 37.9% of the variation in the target variable for the six countries of interest. However, there were concerns as this approach required imputation of significant amounts of data. Furthermore, the R^2 value is based on a relatively small dataset, given that it is filtered to six countries that are not completely accounted for. While the results are promising, the missing data still causes any results to be somewhat imprecise at best.

For goal 10, although the data was lacking, the model using the Gini Coefficient, the average years of schooling, and government expenditures was able to predict the income share of the country's bottom 40% to gain an understanding of the levels of poverty within each country within about five percentage points of the actual value recorded already in the dataset, on average, though this was subject to the simulation. This offers a promising model to predict the bottom 40% income share for the countries of interest. The model's predictions for the growth rate of the income share was, however, difficult to validate due to large amounts of missing data. For goal 10, the next step would be to invest resources into gathering data on countries through surveys so that more recent and consistent data could be utilized in the model to make the results more generalizable.

Missingness in data can be a major hurdle in research or in performance review. As has been demonstrated, while both the theoretical and data-driven approaches had some promising features, the fundamental problem for each method was the inability to validate the findings due to insufficient data in the target. Because of the importance of the UN being able to accurately measure the SDGs, the central conclusion from this study is that approximating these indicators may prove more difficult than anticipated, and in addition to investing in progress on the goals, it may be necessary to invest more heavily in data collection going forward, because accurately measuring the progress of the indicators is of crucial importance to the SDG Joint Fund.

VII. Bibliography

1. Average height in calculating MDER

   https://www.worlddata.info/average-bodyheight.php

2. FAO, IFAD, UNICEF, WFP and WHO. *The State of Food Security and Nutrition in the World 2018: Building climate resilience for food security and nutrition*, 2018

   http://www.fao.org/3/I9553EN/i9553en.pdf

3. BPS-Statistics Indonesia

   https://www.bps.go.id/

4. Staple food prices in Indonesia

   https://data.humdata.org/dataset/wfp-food-prices-for-indonesia

5. USDA National Nutrient Database for Standard Reference, Release 20

   https://ssl.adam.com/graphics/pdf/en/19996.pdf

6. FAO, WHO and UNU. *Human energy requirements: Report of a Joint FAO/WHO/UNU Expert Consultation*, 2001

   https://www.fao.org/3/y5686e/y5686e.pdf

7. Malawi: Nutrition Profile, (updated May 2021) - USAID

   https://www.usaid.gov/sites/default/files/documents/tagged_Malawi-Nutrition-Profile.pdf

8. Climate Impacts on Agriculture and Food Supply

   https://climatechange.chicago.gov/climate-impacts/climate-impacts-agriculture-and-food-supply

9. Global Climate Data - Warming since 1960 (°C / century)

   http://berkeleyearth.lbl.gov/country-list/

10. UN Statistics

    http://data.un.org/

11. World Bank Data Center

    https://databank.worldbank.org/home.aspx

12. Our World in Data
    a. Education
       https://ourworldindata.org/global-education
    b. Inequality
       https://ourworldindata.org/income-inequality
    c. Government Expenditures
       https://ourworldindata.org/government-spending

13. Poverty Probability Index

    https://www.povertyindex.org/

14. United Nations department of economic and social affairs sustainable development

    https://sdgs.un.org/goals/goal10

15. The global goals for sustainable development

    https://www.globalgoals.org/10-reduced-inequalities

16. SDG tracker

    https://sdg-tracker.org/inequality

17. Global SDG Indicator Platform

    https://sdg.tracking-progress.org/indicator/10-1-1-growth-rates-of-household-expenditure-or-income-per-capita-among-the-bottom-40-percent-of-the-population/

18. Martin Ravallion, *Does undernutrition respond to Income and Prices, Dominance Tests in Indonesia, The World Bank Economic Review, Volume 6, Issue 1, January 1992, Pages 109–124*

19. Gustavo Anríquez, Silvio Daidone, Erdgin Mane, *Rising food prices and undernourishment: A cross-country inquiry,289022, Food and Agriculture Organization of the United Nations, Agricultural Development Economics Division (ESA)*