UN SDG Indicator Proxies

Columbia University, MA Quantitative Methods in Social Sciences

Background

The United Nations' Sustainable Development Growth goals, or SDGs, have a number of indicators which they believe offer a blueprint for peace and prosperity. However, many of these indicators are difficult to measure, at least on an annual basis. Because of this difficulty, the UN would like to find proxy indicators that are both easier to collect and able to somewhat accurately predict indicators of interest. A small team at Columbia University attempted to find such proxy indicators for a small subset of the SDG indicators as a proof of concept that this plan was feasible. Unfortunately, doing so proved to be more difficult than was anticipated; the primary problem of missingness in the data also proved to be a major hindrance to finding a proxy indicator, as model validation was incredibly difficult.

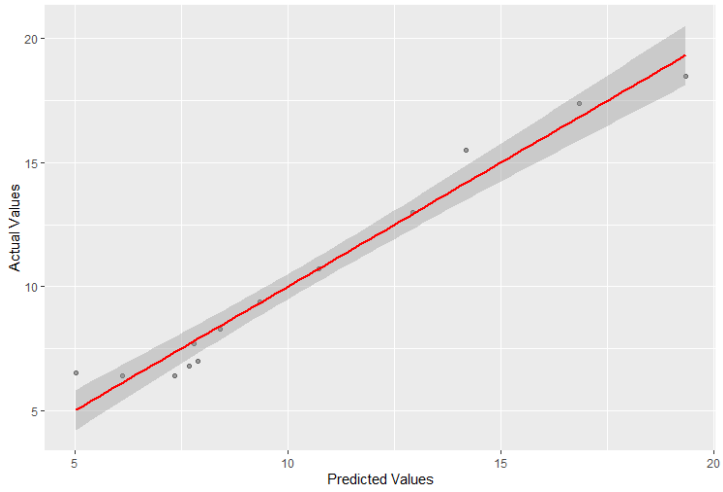
Main Objectives & Chosen Indicators

As a major goal of this research is a proof of concept, the team elected to try three different methodologies (one purely theoretical, one purely data driven, and one hybrid model) to approximate three different SDG indicators. The three selected indicators were: Goal 2.1.1, which measures the prevalence of undernourishment in a

country and was approximated in a purely theoretical way; Goal 2.2.2, which measures the prevalence of malnutrition among children under 5 years of age and was approximated in a theoretical and data driven hybrid model; and Goal 10.1.1, which measures the growth rates of household expenditure or income per capita in the bottom 40% of the population and was approximated with a purely data-driven approach. In all three cases, while the methodologies showed potentially promising results, the lack of data resulted in an inability to validate them. While proxy indicators theoretically could be used to determine the UN's progress toward reaching its goals, it would seem that the main takeaway of this study is that a concerted effort must be made to gather more data, at least in some key areas.

Methodologies

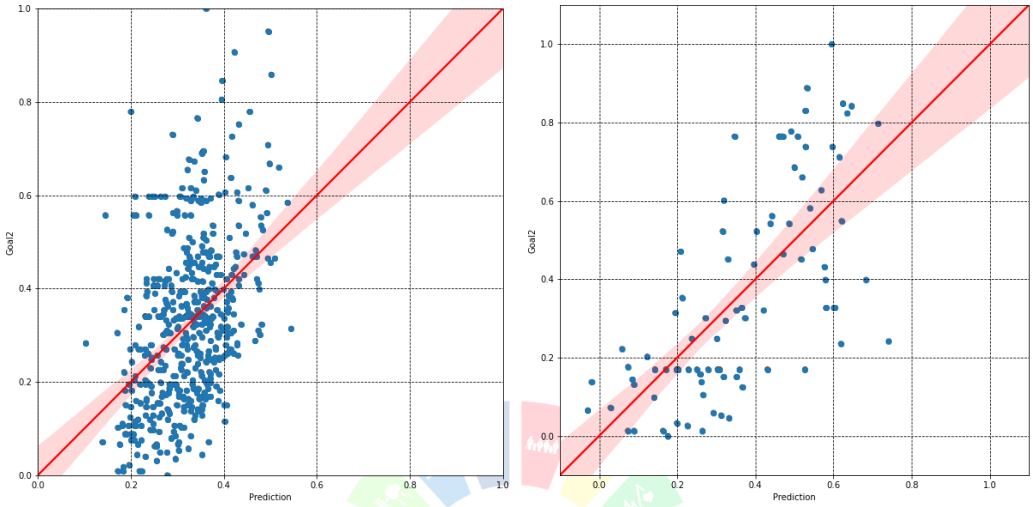
For Indicator 2.1.1, the team used a theory based approach to approximate the prevalence of undernourishment in Indonesia. The team proposed to construct a proxy from the perspective of food demand that uses staple food prices to calculate the “undernourishment poverty line” and combines it with the poverty rate to learn about the population with insufficient purchasing power to stay out of the state of undernourishment. More specifically, using the food consumption patterns of the lower income population, the team computed the weight structure of a standard basket of staple foods, the average food price per kcal, and minimum required food expenditure per capita, step by step, and then adjusted for inflation rates to capture the variations in minimum required expenditure caused by food price fluctuations.



SDG 2,1,1: <u>Model performance on Indonesian data in 2007-2019</u>

Indicator 2.2.2 is chosen as the team has previously attempted many potential proxies on SDG 2.1.1 such as mobile data and hospital data but all hindered by the same problems of collecting a promptly or even accurate dataset. We then found that educational and climate data are very great data sources to make approximations due to their availability and timeliness. However, they are not very related to SDG 2.1.1 but instead very relevant to SDG 2.2.2 when the scope narrows down to children under 5, backed. Also, the great number of missing data as well as lags in reporting of indicator 2.2.2 itself and the high prevalence of malnutrition in low-income countries make our work valuable to help the UN SDG Fund by providing alternative measures for this goal. For indicator 2.2.2, a multi-linear regression is used to validate our chosen proxies and see how closely they proximate the original measure of SDG 2.2.2, on low-income countries identified by the United Nations and the 6 target countries of interest respectively. The model

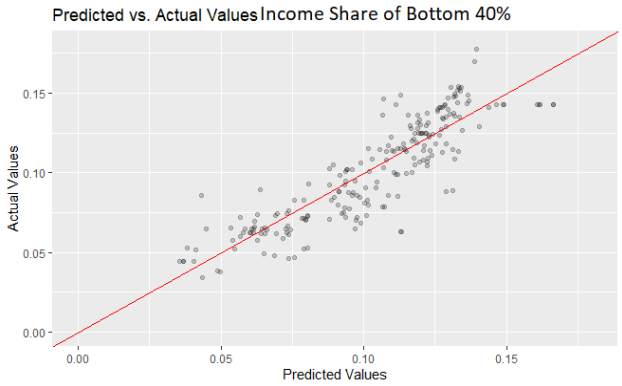
incorporated both female literacy rate and monthly climate data to approximate the prevalence of malnutrition among children under five. The model performed fairly well and resulted in an adjusted R-squared value of 37.9% for our 6 target countries.



SDG 2.2.2: Scatter plot for low-income countries (left) and 6 target countries (right)

Building a proxy measure for indicator 10.1.1, the team used entirely data-driven methods. They found the relationships that existed between the income share of the bottom 40% and hundreds of other aspects of countries' economic and social conditions, using strong relationships to guide what should be included in the then constructed model. In sum, they decided to use the Gini Coefficient, the average number of years of schooling among the population, and the total government expenditures, with each measure being annual data by country. Validating the model based on a subset of the total data, predictions were relatively accurate, with the model's R^2 value being about 0.74 and predictions being, on average, within 5 percentage points of the recorded measure. These predictions

could, then, be used to calculate the predicted annual percent change of the income share, but validating this was difficult due to scarcity of data to measure the predicted values against. Below is a chart of the predicted values of the income share of the bottom 40% against the actual recorded values



SDG 10.1.1: Predicted vs Actual Values for the Bottom 40% Income Share

Results

The three methods used by the team offer a chance to investigate the value of different approaches when investigating the relationship between indicators and potential proxies. predictions against. The main advantage of 2.1.1 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 to lower-income population. Compared with the originally adopted measurement approach, 2.1.1 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.

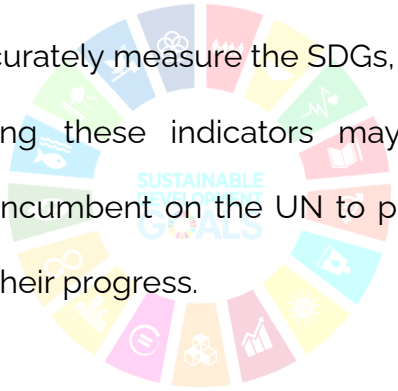
.Compared to the original measures of indicator 2.2.2, our selected proxies are more readily available and have more complete datasets. In recent years, educational data for low-income countries are becoming increasingly accessible, meaning our proxy has more potential as an alternative indicator for indicator 2.2.2. Besides, the climate data is always up to date, suggesting that 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 also the potential causes of delay in achieving the goal. It provides direction on understanding the bottlenecks issue for each country. The data-driven approach taken by the team to study indicator 10.1.1, meanwhile, makes prioritizing the validity of predictions its strength. It also avoids as much as possible the problem of missing data, as no single variable is crucial to the model, allowing the availability of data to influence what is included. Finally, this approach may allow researchers to discover relationships that were unpredicted but may be helpful in making accurate predictions.

For indicator 2.1.1, the modeling result was exasperated by potential violations in assumptions, imprecision in the official data used for validation, and the small sample size due to missing data. indicator 2.2.2 is that the climate data is complex due to its high dependency on different regions, further studies are required in order to better model the most appropriate climate data into the model. As for indicator 10.1.1., While not relying on theory to decide what variables are included in the model has its advantages, it also may be a detriment, as relationships that are discovered may, in fact, be spurious and not applicable to all time periods. Additionally, it allows the model to be built on data that refers to countries that may

behave inherently differently, as these countries are the source of much of the available data. Applying the model to countries of a different nature, then, may prove problematic.

Discussions

Missingness in data can be a major hurdle to deal with in research or in performance review. As has been demonstrated, while each methodology had some promising results, the fundamental problem for each method was an 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 main take-away from this study is that approximating these indicators may prove more difficult than anticipated, and it may be incumbent on the UN to prioritize data collection going forward to accurately track their progress.



Authors: Cunhonghu Ding, John Marion, Lingxiu Guo, Pengyun Li, Xiaofan Liu, Yanji Du, Brian Goddard, Rina Factor, Yi Hyun Kim, Xintong Tang, Yin Long, Yinghzi Zhang