Causal relationships among global indicators

Matej Neumann
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
matej.neumann@student.fmf.uni-lj.si

Marko Grobelnik
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
marko.grobelnik@ijs.si

ABSTRACT

It is important to know how changing one thing will affect another. This becomes even more important when the thing we are changing will affect a lot of people. Therefore, we need a way to visualize how all the things are connected. In this paper, we will demonstrate an approach that uses Granger causality to find causal relationships between global indicators. Our results show that global indicators are indeed highly interconnected however, they still need to be looked at within each country individually. We also comment how this approach can be used to help with policy making decisions.

KEYWORDS

Causality, Global indicators, Granger, Timeseries, SDGs

1 INTRODUCTION

The Sustainable Development Goals (SDGs) launched on January 1, 2016 include 17 goals, 169 targets and 232 unique indicators with the intent to help frame the policies of the United Nations' (UN) member states through 2030 [8]. Because the goals are highly interconnected, as the indicators are not independent, it is important to understand synergies, conflicts and causal relationships between them to support decisions. Without such understanding a policy to help one goal could hurt another. For example, a policy aiming to improve hunger could conflict with climate-mitigation. This paper will focus on finding such relationship with Granger causality.

Granger causality is a statistical concept of causality that is based on prediction and was traditionally only used in the financial domain however, over recent years there has been growing interest in the use of Granger causality to identify causal interactions in neural data [6].

Similar works such as [7] and [2] have already looked for causal relationships between specific SDGs. This paper confirms the previously done work and expands it by adding additional indicators and looking for causal relationship between all the indicators, not just the ones focused on SDGs.

In paper [2] the authors say that the analysis of all of the indicators country by country is without doubt impractical. Nevertheless, Table 2 shows that however impractical it may be, it is still required, as even neighboring countries have vastly different causal relationships.

2 DESCRIPTION OF DATA

2.1 United Nations Statistics Division (UNSD)

This is the official source published by the United Nations it provides information on the development and implementation of an indicator framework for the follow up and review of the 2030 Agenda for Sustainable Development [4].

2.2 The World Bank (WB)

As the data set provided by the UN itself often has missing values, which results in unhealthy timeseries and unreliable results, we decided to add the dataset "World Development Indicators" from The World Bank [5]. Although the data set might not be as official as the one provided by the UN, it does contain 1440 unique indicators for 266 different countries and groups, where each indicator contains a timeseries ranging from the year 1960 to the present time. This addition does not only make the dataset healthier, it also introduces new indicators that are not listed in the UN SDGs. Even so our new dataset still has some limitations. From Figure 1 we can see that on average a country or groups has no values for around 33% of its indicators. Therefore, from now on when talking about the indicators, we will restrict ourselves to just those ones that have at least 20 nonmissing values in their timeseries. This restriction will insure that we are always dealing with a healthy timeseries and it is justified as on average those indicators make up about 50% of all of the ones available as seen in Figure 2.

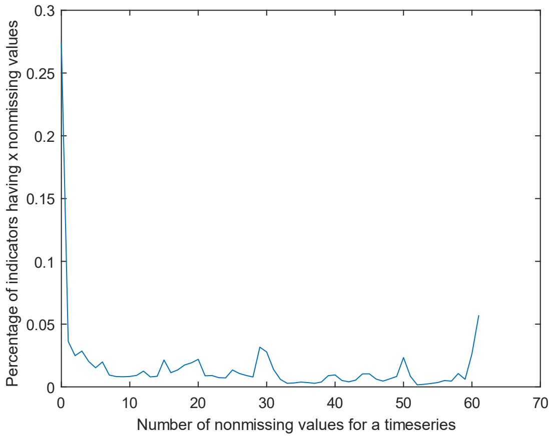

Figure 1: Percentage of indicators having x nonmissing values in its timeseries.

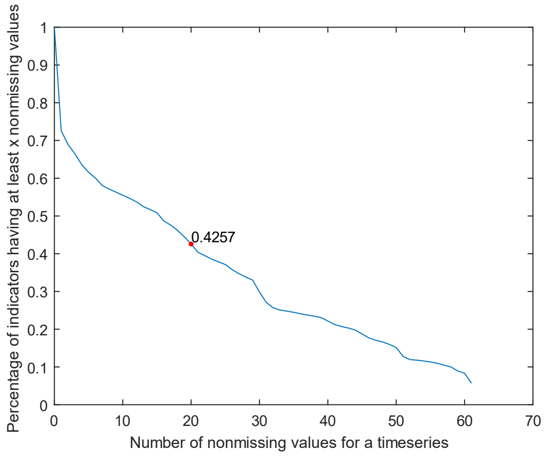

Figure 2: percentage of indicators having at least x nonmissing values in its timeseries.

To better imagine what kind of indicators we are dealing with, we can check Table 1 which shows the top 10 most common ones.

Indicator name	Frequency
Renewable electricity output (% of total electricity output)	265
Population, total	265
Population growth (annual %)	265
Nitrous oxide emissions in energy sector (thousand metric tons of CO2 equivalent)	265
Methane emissions in energy sector (thousand metric tons of CO2 equivalent)	265
Agricultural nitrous oxide emissions (thousand metric tons of CO2 equivalent)	265
Agricultural methane emissions (thousand metric tons of CO2 equivalent)	265
Urban population growth (annual %)	263
Urban population (% of total population)	263
Urban population	263

Table 1: Most common indicators and their frequency of 20 nonmissing values

3 METHODOLOGY

3.1 Granger causality

The causal relationships between indicators were determined by the Granger causality test. The Granger causality test is a statistical hypothesis test for determining whether one timeseries is useful in forecasting another. Informally we say that timeseries X Granger-causes timeseries Y if predictions of the value of Y based on its own past values and on the past values of X are better than predictions of Y based only on Y's own past values. Or in other words X Granger-causes Y if we can better explain the future values of Y with both the past values of X and Y and not just the past values of Y.

More formally, let x and y be stationary timeseries and let x(t) and y(t) be the univariate autoregression of x and y respectfully:

$$x(t) = b_0 + \sum_{i=1}^{p} b_i x(t-i) + E_2(t)$$

$$y(t) = a_0 + \sum_{i=1}^{p} a_i y(t-i) + E_1(t)$$

where p is the number of chosen lagged values included in the model, a_i and b_i are contributions of each lagged observation to the predicted values of $x(t)$ and $y(t)$ and $E_i(t)$ the difference between the predicted value and the actual value. To test the null hypothesis that x does not Granger-cause y, we augment $y(t)$ by including the lagged values of x to get:

$$y(t) = c_0 + \sum_{i=1}^{p} a_i y(t-i) + b_i x(t) + E_3(t).$$

We then say that x Granger-causes y if the coefficients b_i are jointly significantly different from zero. This can be tested by performing an F-test of the null hypothesis that $b_i = 0$ for all i.

3.2 Statistical significance and the p-value

In testing, a result has statistical significance if it is unlikely to occur assuming the null hypothesis. More precisely, a significance level α, is the probability of the test rejecting the null hypothesis, given that the null hypothesis was assumed to be true and the p-value is the probability of getting result at least as extreme, given that the null hypothesis is true. Then we say that the result is statistically significant when $p \leq \alpha$.

3.3 Limitations of the Granger causality test

As its name implies, Granger causality is not necessarily true causality. Having said this, it has been argued that given a probabilistic view of causation, Granger causality can be considered true causality in that sense, especially when Reichenbach's "screening off" notion of probabilistic causation is considered [1].

A problem may occur if both timeseries x and y are connected via a third timeseries z. In that case our test can reject the null hypothesis even if manipulation of one of the timeseries would not change the other. Other possible sources of problems can happen due to: (1) not frequent enough or too frequent sampling, (2) time series nonstationarity, (3) nonlinear causal relationship.

4 EXPERIMENTS

4.1 Setup

Due to time constraints and the limitations of my home system, we decided to limit ourselves to taking just a few countries and groups and calculating the causality relationships for them. The ones we decided on are: (1) United States, (2) China, (3) Uruguay, (4) Slovenia, (5) Austria, (6) Croatia, (7) Italy, (8) European Union and (9) OECD. Our plan was to choose

	AUS	CH	CRO	EU	ITA	OECD	SLO	UY	USA
AUS	100%	4.8%	5.1%	6.9%	6.7%	6.0%	5.9%	4.4%	7.1%
CH		100%	5.6	3.5%	4.3%	3.9%	4.2%	4.7%	4.3%
CRO			100%	4.6%	5%	3.3%	6.6%	3.8%	5.6%
EU				100%	11%	20%	5.7%	3.6%	10%
ITA					100%	6.7%	7.5%	3.8%	6.7%
OECD						100%	5%	3%	17%
SLO							100%	3.5%	5.6%
UY								100%	4.2%
USA									100%

Table 2: Percentage of same causal relationships.

a few of the major world powers and compare the differences and similarities between the causal relationships.

4.2 Modeling the dataset

Once the data was collected from the UNSD and WB website it first had to be put into a suitable form. We decided on a 3D matrix where the first component represented the country or group, the second component represented the time series and last one representing the indicator.

4.3 Parameters

As mentioned before, when searching for causal relationships in a certain country or group we limit ourselves only to those indicators who have at least 20 nonmissing values. Furthermore, we chose a significance level of 0.05 or 5% and tested for lagged values from 1 to 4.

4.4 Determining causality

Once the modeling was done and the parameters were set we first needed to make sure that the timeseries were stationary. To do that we ran the ADF-test and differenced the times series accordingly to make them stationary. Then we ran the Granger-causality test 4 times, once for each lagged value, for each of the 9 countries and groups listed in 4.1. The results for each lagged value were then saved in a 1440x1440 weighted adjacency matrix, where the (i,j) element was nonzero if and only if the i-th indicator Granger-caused the j-th indicator for all lagged values between 1 and 4 and had the weight of the average of the 4 p-values.

Once we had the weighed adjacency matrix we matched the available indicators with the 17 SDGs by comparing the most common buzzwords found in the description of the SDGs and the name of the indicators. An example of some of the buzzwords can be seen in Table 3.

5 RESULTS

With the weighted adjacency matrix in hand, it is sensible to ask ourselves whether there exist any causal relationships that hold true for each of the tested countries or groups. The answer is positive as seen in Figure 3. We can however see that the only causal relationships that survived were the ones that connected different population ages to each other. This result seems sensible as in general no two countries are exactly the same and are therefore going to have a unique set of causal relationships. That being said one can easily imagine why each population age Granger-causes

the next one. For example, if we know the percentage of people aged 4, we can pretty accurately predict what the percentage of people aged 5 is going to be in the next year.

SDG	Buzzwords
Zero Hunger	nourishment, food, stun, anemia, agriculture
Clean Water and Sanitation	water, sanitation, drinking, drink, hygiene, freshwater
Affordable and Clean Energy	energy, electricity, fuel
Climate Action	disaster, disasters, climate, natural, risk, Sendai, environment, environmental, green, developed, pollution
Good Health and Well-Being	mortality, birth, infection, tuberculosis, malaria, hepatitis, disease, cancer, diabetes, treatment, Alcohol, death, birth, health, pollution, medicine

Table 3: Some of the most common buzzwords found in SDGs

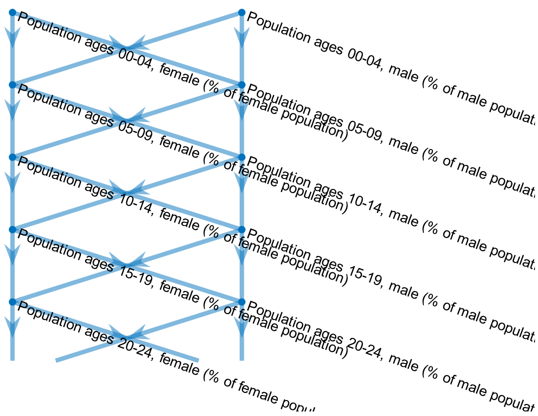

Figure 3 Only causal relationships that are true for each of the 9 countries and groups (continuous down).

On the other hand, one may assume that if we compare countries which are close to each other or are historically connected then the causal relationships should not differ by a lot.

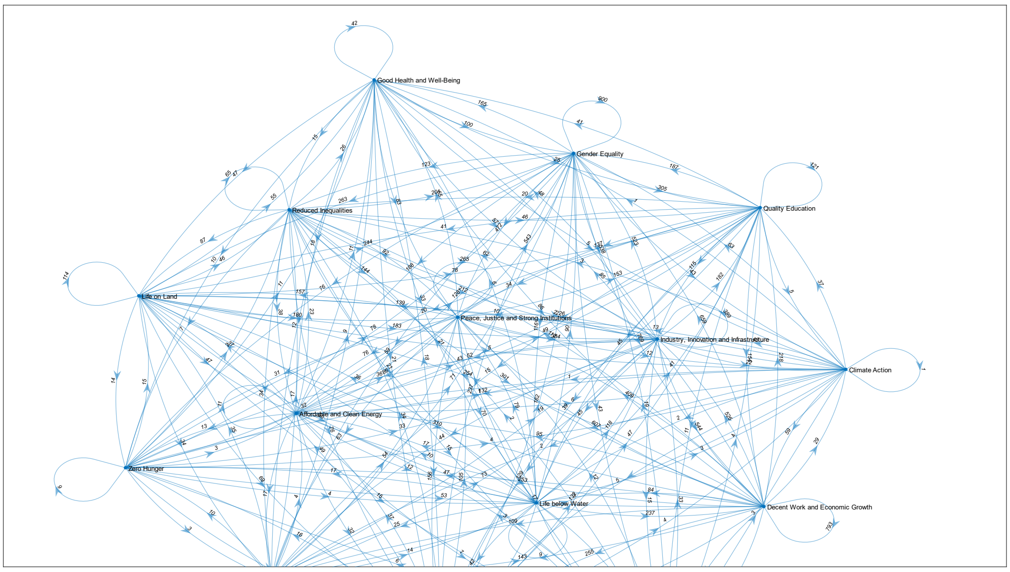

Figure 4: Interconnectedness of SDGs.

That however is not the case as can be seen in Table 2. This suggests, that when talking about causal relationships, one must look at each country or group individually.

Therefore, let's focus just on Slovenia. Due to Slovenia having 10083 positive causal relationships we will limit ourselves to just those that interact with SDGs. Figure 4 shows that indeed SDGs are not independent and in fact are highly interconnected. The presence of self-loops also suggests that there exist causal relationships between indicators of an SDG itself. This result has two consequences:

- When thinking about policies aiming to improve one goal we need to be careful to not harm another
- Instead of outright improving one goal, we can instead focus the ones that are in causal relationship with the one we wish to improve

Let's give an example. Suppose we would want to implement a policy to help to help lower the suicide mortality rate, but we are not how to do that directly. We can therefore instead check which indicators Granger-cause the one we are trying to improve. In our case the indicator "Unemployment, youth total (% of total labor force ages 15-24)" Granger-causes the suicide mortality rate. Therefore, if we improved the % of unemployed young people we would be able to also reduce the suicide mortality rate which was our initial goal.

6 CONCLUSION AND FUTURE WORK

In this paper we demonstrated an approach for calculating causality between depending global indicators and mentioned how this can help with implementing policies. We also showed that neighboring and similar countries in general don't have the same causal relationships, which makes it hard to group them together. However, finding such a grouping, if it exists, could be done in the future. The approach shown in this paper could also be implemented to find causal relationship between certain google searches and natural events. For example, we could check if there is any correlation between the increase of users searching the words "water", "rain", or "cloud" and the likelihood of a flood happening.

7 ACKNOWLEDGMENTS

This work has been supported by the Slovenian research agency.

8 REFERENCES

[1] M. Michael in S. L. Bressler, „Foundational perspectives on causality in large-scale brain networks," *Physics of Life Reviews,* pp. 107-123, 2015.

[2] G. Dörgő, V. Sebestyén in J. Abonyi, „Evaluating the Interconnectedness of the Sustainable Development Goals Based on the Causality Analysis of Sustainability Indicators," *Sustainability,* 2018.

[3] C. Stefano in S. Sangwon, „Cause-effect analysis for sustainable development policy," *NRC Research Press,* 2017.

[4] https://unstats.un.org/sdgs/indicators/database/.

[5] https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators.

[6] B. Corrado in K. Peter, „On the directionality of cortical interactions studied," *Biological Cybernetics,* 1999.

[7] K. Irfan, H. Fujun in P. L. Hoang, „The impact of natural resources, energy consumption, and population growth on environmental quality: Fresh evidence from the United States of America," *Science of The Total Environment,* 2020.

[8] H. Tomáš, J. Svatava and M. Bedřich, "Sustainable Development Goals: A need for relevant indicators," *Ecological Indicators,* pp. 565-573, 2016.