Causal Effects of Renewable and Fossil Fuel Energy Sectors

Progress report

Luca Dibattista, George Maratos

ABSTRACT

Time series play an extremely important role in data science. Granger Causality allows to study a time series from a causal point of view. In this work, we start using synthetic data for having a better understanding of the problem and then we extend our experiments to real-world datasets.

1 INTRODUCTION

In this work, we will consider time series data and investigate causal inference. Sequence data is prevalent and can be easily measured, especially with the increased access to cheap sensor components. A natural question arises whether or not this data can be useful for data mining, pattern recognition, and automated learning tasks. For example, an issue could be with a common assumption made by practitioners, that the data is sampled *i.i.d.*. There is a probable violation of the independence assumption for time-based features.

That being said, there are still many applications that have a temporal aspect, like predicting stock prices from stock tickers, anomaly detection in electro-cardiograms, or the vast array of sensor data that could be given in real-time for robotic planning. There are also many types of data that can be transformed into time series with useful results like DNA, handwriting, novels, and shapes. While we will not explore all of these exciting applications we hope to illustrate some well-known techniques for working with temporal data within the context of causality.

In causal inference, a well-known standard for time series is Granger Causality (GC) [4], and more recent work along this line focuses on explainability and predictability. We will explore and discuss the details of the GC test, and the related works below.

2 RELATED WORK

The Granger assumption [4] is that we can predict the value of Y at the current time step t using information from previous time steps. While it is interesting to explore the exact relationship between Y_t and $Y_{t-1..1}$ we can also define a Granger causal relation between X and Y if we can show that including $X_{t-1..1}$ in our computation of Y_t improves the accuracy of our prediction. Some approaches use a fixed lag model and we will discuss those first.

The simple approach is to exhaustively search for causality using this test by considering all pairs of variables. In this case, an edge is assumed if the granger test is positive. Another is by using the L_1 norm for regression, or more commonly known as Lasso [6], which has the useful property of selecting a minimal amount of coefficients for prediction. We can perform lasso regression on the target variable Y_t, using the lagged predictors $\overrightarrow{x}_{t-1}$ as predictors, and a causal relation is assumed to exist when the coefficient is non-zero. There many others, which can be found in [3], like SIN-Granger and Vector Auto-Regressive Methods.

It can be argued that the fixed lag assumption is in fact too strong of an assumption to make, particularly in real-world applications where the influence from one variable to another could be delayed at arbitrary time steps because of the stochastic nature of the process that generates the outcomes we are trying to observe. Therefore, in such scenarios where the time lag varies, we should consider methods that will optimally align the observational sequences. This is a well-known problem called the Edit-Distance problem and there exist many efficient dynamic programming algorithms to solve this, here we will consider the Edit-Distance problem under the more familiar term Dynamic Time Warping (DTW) [5] which form the basis for the Variable-Lag Granger Method [2] and its extension to the non-linear case Variable-Lag Transfer Entropy [1].

3 PROBLEM DESCRIPTION

We begin by considering the formulation from [3] for GC and the *feature causal network*. We have a set of N features $\{x_i\}_{i=1}^{N}$, where each x_i is a sequence of T observations. We wish to model their causal relationships and characterize the *causal feature network* that is assumed to exist. When there is a causal relationship, the edge between them is paired with a weight called the *lag*. Where the lag is defined as the time delay for causal influence. More concretely, if the variable x_i is a cause of x_j with a lag of k then there is a directed edge $x_i^{T-k} \rightarrow x_j^T$ where we superscript a feature to represent its position in the time sequence. The graph models the distribution over the next set of elements in the sequence conditioned on the lag variables, $P(\{x_i^T\}|\{x_i^t\}_{i=1,..,N,t=0,..,T-1})$. The distribution can be defined in many ways and in many cases they are linear Gaussian models. We will also consider a non-linear approach based on the Transfer Entropy [1].

As mentioned previously, the Granger test is the key to finding GC. If variable X improves the accuracy of predicting Y better than just previous time steps of Y then we can say X Granger Causes Y. This is formalized by considering the following two equations when have linear Gaussian models:

$$\begin{aligned} r_Y(t) &= Y^t - \textstyle\sum_{i=1}^{T} \alpha_i Y^{t-i} \\ r_{YX}(t) &= Y^t - \textstyle\sum_{i=1}^{T} \left[\alpha_i Y^{t-i} + \beta_i X^{t-i}\right] \end{aligned} \tag{1}$$

where α and β are the optimal coefficients from regressing on Y. We can compare the performance of these two models in numerous ways like the F-test or Bayesian Information Criterion (BIC) [2, 1]. If the difference in performance, when including X produces a statistically significant improvement (reduction in variance), then we say that X Granger Causes Y. Therefore, a naive approach to discovering the graph structure is to run an exhaustive Granger test on all pairs of variables.

The type of model described in the test of equation 1 is a characterization of a fixed lag model. There will be issues when regressing because α and β are fixed when the optimal might change through time. To mitigate this issue, [2] utilize Dynamic Time Warping to align the sequence from X to Y using a dynamic programming

algorithm for the minimum edit distance problem which gives an optimal time sequence alignment $\{\delta_t\}$. Then they regress with the additional sequence aligned variable X_*, and the new residual term considered in the granger test is

$$r^*_{YX}(t) = \sum_{i=1}^{T} \left[\alpha_i Y^{t-i} + \beta_i X^{t-i} + \gamma_i X^{t-i}_*\right]. \quad (2)$$

We can then say that X VL-Granger causes Y if r^*_{YX} is less than both r_Y and r_{YX}.

4 INITIAL SOLUTIONS

Starting from the data described in Section 5, we selected the dataset we were interested in. We grouped the data by country and year, and we built a table with the value of a certain indicator (e.g., GDP) for each country and year. Then we merged all the different datasets to obtain D. The main limitation of this approach is the small amount of observation that we were able to get. The number of observations (between 100 and 200) was not enough for a causal study. For this reason, we decided to look for datasets from other sources with a higher number of observations, i.e., with a higher time granularity given the same time interval.

By using the same schema, we found some of the variables that we had with yearly data, being able to run some experiments and obtain more reliable results.

5 DATA DESCRIPTION AND CLEANING

We are using multiple UN datasets[1]. Each sample of each dataset is a value for a certain indicator, and there is a row for each country and for each year.

Some datasets do not have the information for each year, but for a range (e.g. they give a value for the years 1990-1999 instead of the value for each year separately). In this case, to be able to merge these datasets with all the others, we plan to study the function that best represents the data (through interpolation) to obtain one observation for each year. Since data comes from different sources, we also intend to join the different datasets on the countries with a text similarity algorithm, since sometimes the same country has different names (e.g. “United States of America" is the same as “United States").

Human Development Index (HDI). It is an indicator that was born to emphasize that the development of a country should not be considered based on its economy only, but also on people. This is why the HDI is a function of the Life Expectancy Index, Education Index, and GNI index. This is the indicator we are interested in because it is a simple number that includes three extremely important indicators. The dataset data ranges from 1990 to 2018. Figure 2 shows the HDI of the United Stated as a function of time. As we can notice, the slope is positive, meaning that the indicator is growing with time. Since the indicator is a number between 0 and 1, we expect the slope to decrease over time, as it is possible to notice from the last 6 years. In 2019 the United Nations Development Programme (UNDP) introduced the inequality-adjusted HDI (IHDI), where inequalities are taken into account.

[1]https://data.un.org/

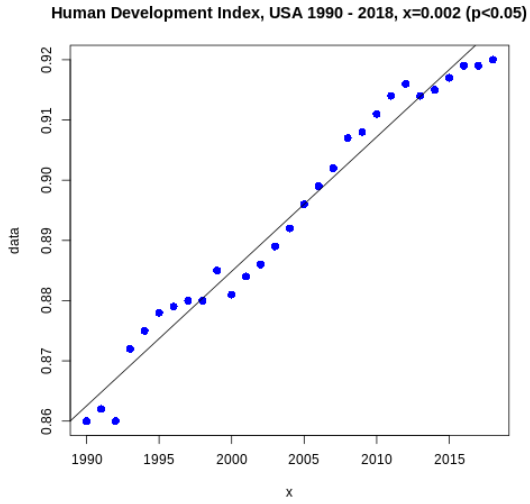

Figure 1: Human Development Index in U.S. over time.

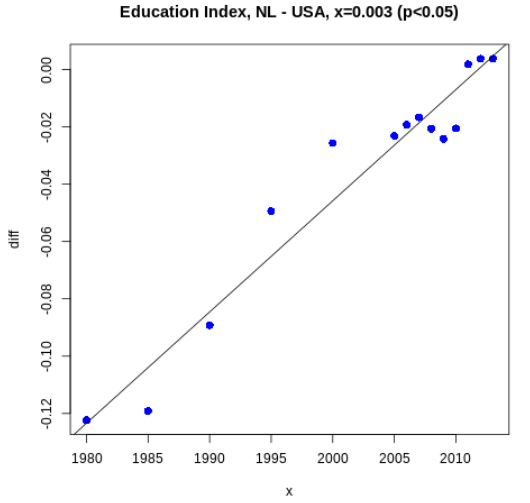

Figure 2: Difference between education index of Netherlands and U.S.A.

Education Index. It is calculated using Mean Years of Schooling (number of years of education received by people with 25 or more years) and Expected Years of Schooling (how many years a child is expected to go to attend school, university, including repetitions). Data from 1980 to 2013 is available from a list of 187 countries. To understand which countries are investing more in education, is is not enough to see how high is that value. Figure 2 shows the difference between the education indexes of Netherlands and the U.S. If the two countries both had the same improvement, we would have a zero difference (i.e., a constant line). In our case, we can notice how the Netherlands is geining higher improvements than the U.S. We have chosen Netherlands because it is known to be

a green country, and there might be a causality between the two factors (being a green country and having an high education index). We intend to investigate more in this possible causal effect.

Life expectancy at birth. It indicates what is life expectancy when a child is born. The years are bins composed of 5 years. We plan to find a function that intercepts the data we have to obtain samples of one year each.

Per capita GDP at current prices. It contains the GDP per capita measured in US dollars. Despite DGP at constant prices would have been more useful with respect to current prices[2], unfortunately, the GDP at constant price contained 6729, while the one at current price 9870.

Employment. This dataset indicates, for each country, year, and sector, the number of employees. We selected all the samples of the electric field.

Energy Statistics Database. It contains data about the energy field from 1990 to 2017. We selected some indicators such as the total demand, production, and consumption, as well as solar and wind energy production. We also included indicators about fossil fuel production (e.g. brown coal, coking coal, fuel oil, gas oil, etc.). The fact that all the energy data is coming from a single datamart is positive.

Greenhouse Gas Inventory Data. It includes data from 1990 to 2017, where each value is the amount of CO_2 equivalent produced (greenhouse gases - GHGs). The same datamart also contains information about methane, nitrogen trifluoride, and other polluting compounds.

The main problem with the UN datasets is the time granularity. Despite the high number of indicators that are available on the UNdata website in an accessible format, all of them contain values for indicators for each year. Since the year range for most of the indicators is 1980-2010, we only have 30 observations per country for each indicator. For this reason, we looked for other sources and we found some of the same indicators with a quarterly or monthly period. We decided to focus on the United States.

U.S. Energy Information Administration (EIA). It contains data about monthly energy consumption and production from January 1973 to July 2020. Energy consumption and production are divided by fossil fuels, nuclear electric power, and renewable energy. Figure 3 shows data for energy production by sector from this dataset. We may notice that despite the fossil fuel sector is currently fulfilling most of the energy demand, renewable energies have gained market share in the last decade. It is also interesting that, while fossil fuel energy production has dropped from 7.00 to 5.86 quadrillion BTU from January 2020 to May 2020 (that is when the coronavirus epidemic started), the renewable sector has slightly increased from 1.00 to 1.04 quadrillion BTU during the same time interval. The EIA database also contains data about monthly CO_2 emissions from energy consumption by source.

U.S. Bureau of Economic Analysis (BEA). This dataset provides the same number of observations for both current and constant prices. We selected the GDP at a constant price, measured in GDP in billions of chained 2009 US dollars, from 1930 to 2015. Since the data is given for each quarter of the year, when we included the GDP in our analysis, we computed, for each other variable, the total value in the same quarter of the year. For instance, if an indicator value was 1 in January, February, and March and we had the GDP for the first quarter (January, February, and March) of the year, then we summed the other indicator (sum made sense, for instance with energy production) and we obtained one row with a value of 3 for the indicator in the period Jan-Mar.

[2]Constant price takes into account the effect of inflation, current prices do not

	$\mu_{\text{fixed}} - \mu_{\text{variable}}$	$\sigma_{\mu_{\text{fixed}}-\mu_{\text{variable}}}$
VL Granger	0.1862193	0.1159677
VL Entropy	-0.0675950	0.0294300
Granger Basic	0.1172530	0.1019945

Table 1: The summary statistics from the difference between variable and fixed lag experiments. Smaller is better.

6 EXPERIMENTAL RESULTS

6.1 Synthetic Experiments

Here we perform an experiment that compares methods with fixed lag to variable lag. While these experiments were done in [1] we attempted to run another synthetic test with the goal of proving that the variable lag methods have better performance than the fixed. We followed the same format for data generation where X is sampled from a Autoregressive Moving Average stochastic process with Y's being a sum of variables that influence it. We generate two synthetic datasets, one with a fixed lag and one with a variable lag. Then we evaluate three different methods for graph discovery and measure the precision, recall, and F1 score on each dataset. We ran this test independently 30 times and took the averages of the performance metrics. Unfortunately the evidence was not very conclusive.

The results are found in Tables 2 and 1. What we would expect to see from Table 1 is the VL methods having little difference while the basic Exhaustive Granger Fixed Lag method showing a statistically significant difference. We suspect that there may be some external factor, like a bug in our code, which explains why VL-Granger is reporting a significant difference. Also the result that VL-Granger Entropy has no significant difference is not useful because, as we can see in Table 2 it is in fact the worst performing model in predicting graph structure. It could be the case that our graph was incorrectly generated, because of the fact that Basic has performed poorly on the fixed lag test. Ultimately we believe that these results are in fact inconclusive.

7 FUTURE WORK

A main issue for us is merging long sequence data so we can derive causality, with the original topic we had in mind. Since much of the un data actually turned out to be deceptively sparse, in terms of continuous sequential observations for any single country, performing inference with time series data seems to be a intractable task. A major component of our previous model had involved finding employment figures with a sufficient level of granularity but on top of the above mentioned issue there does not seem to be specific

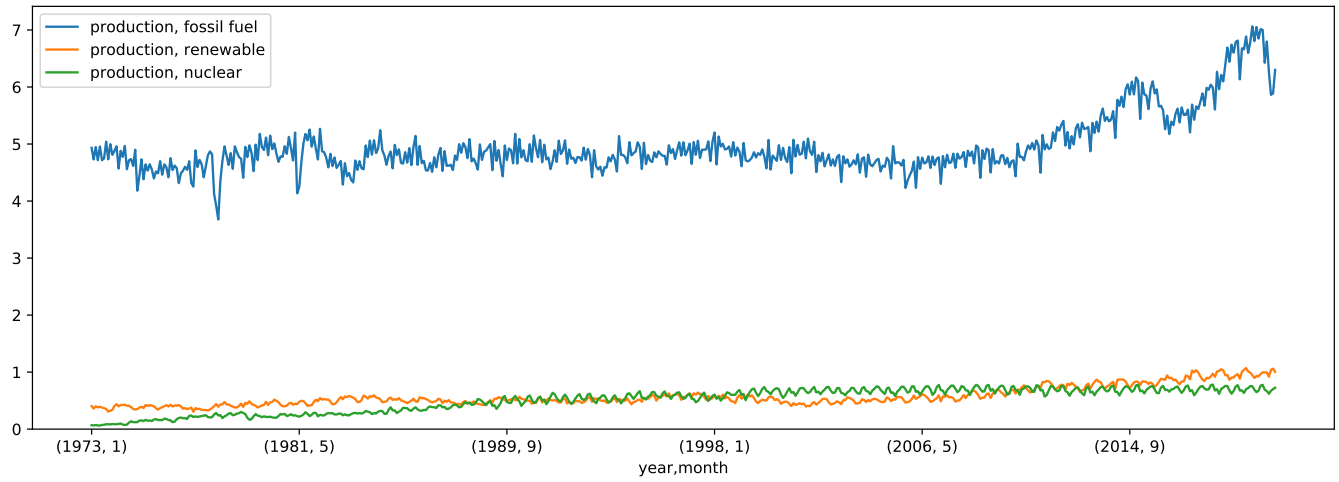

Figure 3: Energy production by sector

	fixed lag			variable lag		
	precision	**recall**	**F_1**	**precision**	**recall**	**F_1**
VL	0.79916 ± 0.06697	0.77666 ± 0.08172	0.78496 ± 0.05840	0.78027 ± 0.09506	0.59666 ± 0.13767	0.67024 ± 0.11337
Entropy	0.29214 ± 0.04607	0.81666 ± 0.12058	0.42988 ± 0.06493	0.27047 ± 0.05240	0.72333 ± 0.13308	0.39295 ± 0.07328
Basic	0.38828 ± 0.07626	0.91000 ± 0.04806	0.53972 ± 0.07163	0.25216 ± 0.02559	0.86666 ± 0.10933	0.38952 ± 0.03660

Table 2: The performance metrics for both experiments across the three tested models.

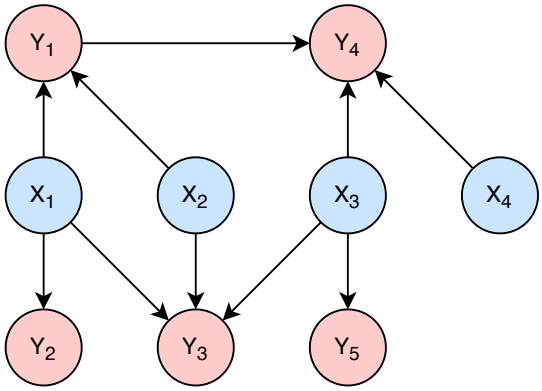

Figure 4: The graph we used to generate synthetic data.

enough information about individuals employed in these industries. This means that we might have to consider having only a single feature for total employment in the energy sector, but if we can find a sufficient amount of time series data we could run interesting experiments on how it is influenced by other *renewable vs fossil fuel* based metrics like total power production.

The first part of this report, we believe, shows how the project might have to pivot from working specifically on time series data about UN development goals to broadening our view by focusing on the methodology. This would mean that we would have more of a benchmarking style project, but hopefully touching on interesting challenges with finding causality with time series observations. We had originally hoped, for the synthetic experiment, to generate some novel evidence of why considering variable lag is important and then maybe expand on them to explore the impact of certain parameters on performance. We could investigate how well the sequences can be aligned with DTW in scenarios where graph structure is highly complex and lag is significant.

For the real world examples, we are considering just exploring causality with data from the United States because our chances of finding sufficient data are more likely. For example we might be able to get better employment figures and accurate measures on power consumption and production. We also want to consider other datasets that have time series, so we can illustrate how variable lag presents itself in real world settings. The synthetic data gives us an easy way to measure performance because we know the ground truth, but for real world data the graphs ae not so obvious. This means that it is more of an unsupervised task, related to graph discovery and reasoning about whether the results from running the Granger test make sense. The Granger tests could be seen as a way to reinforce a hypothesized graph, and combined with building architectures that accurately model real world processes, we can then truly perform causal inference.

REFERENCES

[1] Chainarong Amornbunchornvej, Elena Zheleva, and Tanya Berger-Wolf. "Variable-lag Granger Causality and Transfer Entropy for Time Series Analysis". In: *arXiv preprint arXiv:2002.00208* (2020).

[2] Chainarong Amornbunchornvej, Elena Zheleva, and Tanya Y Berger-Wolf. "Variable-lag Granger Causality for Time Series Analysis". In: *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. 2019, pp. 21–30.

[3] Andrew Arnold, Yan Liu, and Naoki Abe. "Temporal causal modeling with graphical granger methods". In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2007, pp. 66–75.

[4] Clive WJ Granger. "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica: journal of the Econometric Society* (1969), pp. 424–438.

[5] Hiroaki Sakoe and Seibi Chiba. "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE transactions on acoustics, speech, and signal processing* 26.1 (1978), pp. 43–49.

[6] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.