ROBUSTLY MODELING THE NONLINEAR IMPACT OF CLIMATE CHANGE ON AGRICULTURE BY COMBINING ECONOMETRICS AND MACHINE LEARNING

Benedetta Francesconi
Vrije Universiteit Amsterdam, NL
benedetta.francesconi.1993@gmail.com

Ying-Jung C. Deweese
Descartes Labs & Georgia Institute of Technology, US
yingjungcd@gmail.com

ABSTRACT

Climate change is expected to have a dramatic impact on agricultural production; however, due to natural complexity, the exact avenues and relative strengths by which this will happen are still unknown. The development of accurate forecasting models is thus of great importance to enable policy makers to design effective interventions. To date, most machine learning methods aimed at tackling this problem lack a consideration of causal structure, thereby making them unreliable for the types of counterfactual analysis necessary when making policy decisions. Econometrics has developed robust techniques for estimating cause-effect relations in time-series, specifically through the use of cointegration analysis and Granger causality. However, these methods are frequently limited in flexibility, especially in the estimation of nonlinear relationships. In this work, we propose to integrate the nonlinear function approximators with the robust causal estimation methods to ultimately develop an accurate agricultural forecasting model capable of robust counterfactual analysis. This method would be a valuable new asset for government and industrial stakeholders to understand how climate change impacts agricultural production.

1 INTRODUCTION

The climate system is becoming more extreme, with an increase of heavy precipitation events and long dry spells [14]. Such drastic changes in weather events will inevitably have a significant impact on crop yield. The precise effect of different climate variables on yield is complicated due to known variation among plant types and intricate unknown causal structures [13, 8, 10]. For instance, a naïve correlational analysis of Figure 1 shows increasing temperature to increase vegetable production, as they both have steadily increased in the last 50 years. However, this first impression can be misleading.

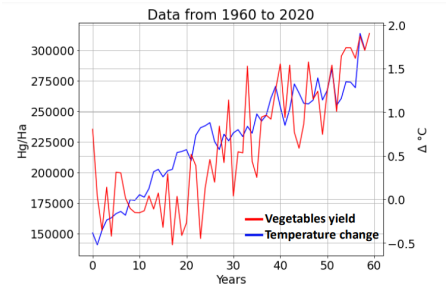

FIGURE 1: Annual temperature change and annual vegetables yield in Italy

Figure 2 is an illustration of the potential problematic causal graph, depicting an unobserved causal effect. If we consider temperature (T), yield (Y) and a third variable (C) (potentially population or pollution), it can be the case that C acts causally on both temperature and yield, thereby confounding initial estimates of how temperature may impact yield. A mistaken causal estimation would have significant negative impact for policy decisions. Indeed, confounders in the climate science literature have lead to misleading correlations [28]. Thus, accurately assessing the causal effects of climate change on agricultural production is critical for correctly forming climate change adaptation and mitigation policies. In prior work, researchers in econometrics and machine learning have examined the effects of climate change on agriculture

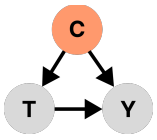

FIGURE 2: C may bias the estimated relations between T & Y.

[11, 7, 16, 18, 26, 15, 29]; however these approaches have yet to capture the causal influence of environmental and socioeconomic variables on agriculture.

In this work, we propose to couple the econometric techniques of cointegration analysis and Granger causality with machine learning, thereby combining the robust econometric methods of causal discovery with the nonlinear modeling capacity of machine learning. We propose to apply this method to assess the effects of climate change on agricultural production, which, to the best of our knowledge, has not been done before in the literature. The contributionos of this proposal are: (i) the application of powerful nonlinear causal analysis to a new agriculture dataset, and (ii) the improved robustness of these causal methods through the use of cointegration analysis.

2 Prior work

In econometrics, importance is given to the techniques of cointegration and the estimation of Vector Error Correction Models (VECM) [17, 5] or Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models [22]; however, such models come with several known limitations [21]. For example, cointegration tests are not flexible enough to handle nonlinear relations, and in general make use of basic statistical models, thereby lacking a consideration causality by default [24]. Despite this, cointegration methods are useful to robustly unveil hidden relations between seemingly unrelated variables. Moreover, such statistical models are reliable only if the initial conditions and assumptions of the models do not change. As we are working with factors in climate system that have been changing over last decades, the initial conditions have been changing throughout time. On the other hand, machine learning models are also usually not able to detect causality between variables. Traditionally machine learning models focus on prediction, but not on causality detection.

3 Approach

Methodology We propose an approach where the econometric tests of cointegration and Granger causation are coupled with deep autoregressive model (DeepAR) and expanded through a counterfactual analysis. In causal inference, counterfactual analysis allows one to uncover cause and effects mechanisms between variables. This is possible because the same model is estimated first with the original variables and then it is estimated again without including those same variables. The differences in the two estimations allow us to establish the presence of a causal relation. DeepAR, Granger causality and counterfactural analysis have already been used in combination together, but without performing a preliminary cointegration analysis [6]. We propose to expand the work in [6] by performing cointegration tests before working with Granger causality and DeepAR. The presence of cointegration is essential for two or more variables to be involved in a Granger causality relation [19]. In absence of cointegration, the relation between the variables might not exist. As a further difference with the study of [6], we propose to model the impact of climate change on agricultural data. The climate variables used in this study are: precipitation, temperature, wind and radiation flux. Soil moisture is also included, while the target variable will be agricultural production. The ultimate goal is to assess the presence of causal effects of the mentioned climatic variables or soil moisture over crops yield.

In our approach, the first step involves the application of several cointegration techniques: Johansen cointegration, threshold cointegration and time-varying cointegration. Cointegration signals the presence of a linear relation between variables seemingly unrelated. When present, it can have a time-varying or time-invariant nature or appear only at specific periods. If a cointegration relation is found, then a Granger causality test can be performed. Granger causality allows to understand if certain variables can be useful to predict the value of another variable. Granger casuality only confirms that the inclusion of a variable improves the forecasting of another variable. Also, Granger causality test can deal only with stationary and linear models. Assuming that climatic variables and crops yield are linearly related is an unrealistic limitation. For this reason, in our approach the Granger causality test will not be performed using a Vector Autoregressive (VAR) model, as it is standard practice, but rather a deep autoregressive model (DeepAR) [27]. The DeepAR model is chosen as it is able to deal with non-linear and non-stationary time series, better fitting the assumptions of the variables at hand. The model is also able to extract hidden features in the data, such as seasonality or other hidden patterns. Given

the seasonality of climatic phenomena and potentially climate change related trends observed in the last decades, such a powerful model will be better suited to identify hidden variables which cannot be observed from the data.

In order to establish if climatic variables or soil have causal effects over crops yield we will use knockoffs [12]. Knockoffs are variables that will have the same distribution as the original climatic and soil variables, but that will be independent of the model output. This means that the knockoffs are generated such that they have no causal relation on the crops yield. In order to assess causality, the DeepAR model will be estimated both using the original climatic and soil variables and then by using the knockoff variables. The two DeepAR models will then be compared following the procedures of the Granger causality test: if significant differences are found between the two model estimations, then we can conclude that the original independent variables have a causal effect over crops yield. If no differences are measured, then we cannot conclude that our climatic or soil variables cause effects over crops yield. To validate the above procedure, we first propose the use of a synthetic dataset where the causal relations between the variables are already known. We can then measure the agreement of the causal relations and strengths estimated from the above procedure with the true causal structure to evaluate the strength of the approach. It will be then possible to quantitatively measure the increased robustness induced by our proposed introduction of cointegration analysis. Following this model validation, we propose to apply these innovative techniques to the agricultural dataset outlined below, thereby providing a realistic case-study for the proposed method.

Data We propose to use the average monthly temperature and the average monthly precipitation data from the NOAA dataset [1]. This dataset provides data from 1895 until 2022, at the 5x5 lat/lon scale. For wind data, we will use the NOAA dataset [2] spanning from 1979 to 2022 at a 1.9x1.9 resolution level. From NOAA we will also obtain the data relative to soil moisture [3], available from 1948 until 2022 at a 0.5x0.5 spatial resolution level and data relative to radiation flux (shortwave and longwave) [4], available from 1979 to 2022 at a 0.3 degrees resolution level. The crops data will require additional manipulation. Monthly crops data are extremely rare and the dataset that is usually used, the MIRCA2000, has data only until the year 2000. For this reason, we will create the necessary data using the approach specified in [20]. The authors propose a methodology based on using publicly available data from the FAOSTAT database and the GAEZ Version 4 global gridded dataset. Through them, they generate circa 2015 annual crop harvested area, production, and yields by crop production system (irrigated and rainfed) for 26 crops and crop categories globally at 5-minute resolution. We will replicate the approach of the authors to generate data for also all the years before 2015 up to 1979 and for those after 2015 until 2022. In this way we will be able to use a dataset of variables spanning from 1979 to 2022, which is the timespan common among all different data sources. Since the selected data in this work are at different resolution, we will upsample these data and aggregate data at the county level. The counties under consideration will be those of California, US.

4 Limitations

Our proposal has some limitations, mainly related to the dataset to be used. The main limitation is that agricultural production is not solely influenced by climate, even when this plays a big role. Other phenomena such as technological advancements, production techniques and shocks provoked by wars just to name a few have a non negligible influence. These variables are all extremely hard to capture. For this reason, the risk of an incorrect estimation of the relation between the variables persists. In future work, we may consider the introduction of instrumental-variables as a method to handle such unobserved confounders [25, 23].

5 Conclusion

Models of climate change's effect on agricultural production are currently lacking a fundamental component – the estimation of causal relationships. Despite the research done in econometrics and machine learning, there is still little importance given to the analysis of causal effects between the variables involved. Through this work, we suggest a new approach which integrates econometric and machine learning methods to robustly estimate causal relationships, and propose to apply this model to agricultural data for the first time. As a result, we believe this model and proposed experimental validation would offers a reliable new technique, providing guidance for policy-making decisions among government agencies and industries.

REFERENCES

[1] NOAA National Oceanic and Atmospheric Administration. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00332. Accessed: 2023-01-27.

[2] NOAA National Oceanic and Atmospheric Administration. https://www.ncei.noaa.gov/access/monitoring/wind/. Accessed: 2023-01-27.

[3] NOAA National Oceanic and Atmospheric Administration. https://psl.noaa.gov/data/gridded/data.cpcsoil.html. Accessed: 2023-01-28.

[4] NOAA National Oceanic and Atmospheric Administration. https://psl.noaa.gov/data/gridded/data.narr.html. Accessed: 2023-01-29.

[5] Melaku Adinew and Gebrekirstos Gebresilasie. Effect of climate change on agricultural output growth in ethiopia: Co-integration and vector error correction model analysis. *Budapest International Research in Exact Sciences (BirEx) Journal*, 14(4):132–143, 2019. URL https://doi.org/10.33258/birex.v1i4.461.

[6] Wasim Ahmad, Maha Shadaydeh, and Joachim Denzler. Causal inference in non-linear time-series using deep networks and knockoff counterfactuals. 09 2021. doi: 10.1109/ICMLA52953.2021.00076.

[7] Faiza Ahsan, Abbas Chandio, and Wang Fang. Climate change impacts on cereal crops production in pakistan: Evidence from cointegration analysis. *International Journal of Climate Change Strategies and Management*, ahead-of-print, 02 2020. doi: 10.1108/IJCCSM-04-2019-0020.

[8] Samuel Asumadu-Sarkodie and Phebe Asantewaa Owusu. The causal nexus between carbon dioxide emissions and agricultural ecosystem—an econometric approach. *Environmental Science and Pollution Research*, 24(2):1608–1618, October 2016. doi: 10.1007/s11356-016-7908-2.

[9] Mohammad Taha Bahadori and Yan Liu. An examination of practical granger causality inference. pp. 467–475. doi: 10.1137/1.9781611972832.52.

[10] Imran Baig, Farhan Ahmed, Md. Abdus Salam, and Shah Khan. An assessment of climate change and crop productivity in india: A multivariate cointegration framework. *Test Engineering and Management*, 83:3438–52, 08 2020.

[11] Imran Baig, Farhan Ahmed, Md. Abdus Salam, and Shah Khan. An assessment of climate change and crop productivity in india: A multivariate cointegration framework. *Test Engineering and Management*, 83:3438–52, 08 2020.

[12] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015. ISSN 00905364. URL http://www.jstor.org/stable/43818570.

[13] A. Bonfante, A. Impagliazzo, N. Fiorentino, G. Langella, M. Mori, and M. Fagnano. Supporting local farming communities and crop production resilience to climate change through giant reed (arundo donax l.) cultivation: An italian case study. *Science of The Total Environment*, 601-602:603–613, 2017. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2017.05.214.

[14] Michele Brunetti, Letizia Buffoni, Franca Mangianti, Maurizio Maugeri, and Teresa Nanni. Temperature, precipitation and extreme events during the last century in italy. *Global and Planetary Change*, 40(1):141–149, 2004. ISSN 0921-8181. doi: https://doi.org/10.1016/S0921-8181(03)00104-8. Global Climate Changes during the Late Quaternary.

[15] Evan J. Coopersmith, Barbara S. Minsker, Craig E. Wenzel, and Brian J. Gilmore. Machine learning assessments of soil drying for agricultural planning. *Computers and Electronics in Agriculture*, 104:93–104, 2014. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2014.04.004.

[16] Andrew Crane-Droesch. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, 13, 11 2018. doi: 10.1088/1748-9326/aae159.

[17] Hasan Gökhan Doğan1 and Arzu Kan1. The effect of precipitation and temperature on wheat yield in turkey: a panel fmols and panel vecm approach. *Environment, Development and Sustainability (2019) 21:447–460*, 21(21):447–460, 2019. URL https://doi.org/10.1007/s10668-018-0298-5.

[18] Georgios Giannarakis, Vasileios Sitokonstantinou, Roxanne Suzette Lorilla, and Charalampos Kontoes. Personalizing sustainable agriculture with causal machine learning. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*, 2022. URL https://www.climatechange.ai/papers/neurips2022/112.

[19] C.W.J. Granger. Causality, cointegration, and control. *Journal of Economic Dynamics and Control*, 12(2):551–559, 1988. ISSN 0165-1889. doi: https://doi.org/10.1016/0165-1889(88)90055-3.

[20] Steve Wisser Dominik Prusevich Alex Glidden Stanley Grogan Danielle, Frolking. Global gridded crop harvested area, production, yield, and monthly physical area data circa 2015. 9(15):1–5, 2022. doi: https://doi.org/10.1038/s41597-021-01115-2.

[21] Maria-Carmen Guisan. Causality and cointegration between consumption and gdp in 25 oecd countries: limitations of cointegration approach. *Applied Econometrics and International Development*, 1(1):39–61, 2001. URL https://EconPapers.repec.org/RePEc:eaa:aeinde:v:1:y:2001:i:1_2.

[22] Boris E. Okello David Deom Carl Li, Aizhen; Bravo-Ureta and Naveene Puppala. *Working Papers 017*. URL http://dx.doi.org/10.22004/ag.econ.148353.

[23] Matthew L. Maciejewski and M. Alan Brookhart. Using Instrumental Variables to Address Bias From Unobserved Confounders. *JAMA*, 321(21):2124–2125, 06 2019. ISSN 0098-7484. doi: 10.1001/jama.2019.5646.

[24] Brady Neal. Introduction to causal inference pages 1:18. 2020.

[25] Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003. doi: https://doi.org/10.1111/1468-0262.00459.

[26] X.E. Pantazi, D. Moshou, T. Alexandridis, R.L. Whetton, and A.M. Mouazen. Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, 121:57–65, 2016. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2015.11.018.

[27] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2019.07.001.

[28] Jürgen Scheer and Esteban R. Reisin. Simpson's paradox in trend analysis: An example from el leoncito airglow data. *Journal of Geophysical Research: Space Physics*, 118(8):5223–5229, 2013. doi: https://doi.org/10.1002/jgra.50461.

[29] S. Veenadhari, Bharat Misra, and CD Singh. Machine learning approach for forecasting crop yield based on climatic parameters. pp. 1–5, 2014. doi: 10.1109/ICCCI.2014.6921718.

6 Appendix

6.1 Cointegration

6.1.1 Johansen cointegration

A general Vector Autoregressive model (VAR) with Gaussian errors can be written in the Error Correction Model (ECM) form as:

$$\Delta Y_t = \sum_{i=1}^{p-1} \Gamma_i \Delta Y_{t-i} + \Pi Y_{t-p} + \Phi D_t + \mu + \epsilon_t \tag{1}$$

where Y_t is the vector of the time-series in consideration. In the problem exposed in this work, it is constituted by the variables in consideration, such as: wind, soil moisture, temperature,

precipitation, and the target yield. $\Delta Y_t = Y_t - Y_{t-1}$ is a vector in $\mathbb{R}^k$, D_t is a vector of seasonal dummy variables orthogonal to the constant term μ, Φ is the coefficients of the D_t variables, $\epsilon_t \sim N_k(0, \Lambda)$ represents the Gaussian errors, Γ_i and $\Pi = \alpha\beta^T \in \mathbb{R}^{k\times k}$, where $\alpha \in \mathbb{R}^{k\times r}$ and $\beta \in \mathbb{R}^{k\times r}$. The considered total lags back in time is indicated by p.

The Johansen appoach is based on the analysis of the rank of the Π matrix, which is called the impact matrix and where r represents the maximum number of independent vectors within this matrix. If the rank $r = 0$, the Π matrix would collapse and so the error correction term would disappear, meaning that there is no long term relationship between the variables involved. On the other hand, if the rank is $0 < r < k$, then the variables are said to be cointegrated; if $r = k$, then Π is a full rank matrix and the variables are all linearly independent. In the first and last case no cointegration can be devised.

Given a stationary ΔY_t and a nonstationary Y_t, if there exists a β matrix such that the linear combination between them in the form of $\beta^T Y_t$ is stationary, then the elements in Y_t and ΔY_t are said to be cointegrated. The space spanned by β is the space spanned by the rows of Π and it represents the cointegrating space, while α is the adjustement coefficient, as in the work by Agunloye et al. (2014). In order to infer the number of r cointegrating vectors and their significance, the Johansen approach proposes two tests, the trace test and the maximum eigenvalues test.

6.1.2 Threshold cointegration

The threshold cointegration methodology proposed in this work spans from the ones presented by Balke and Fomby (1997) to those of Hansen and Seo (2002). Given two variables that are suspected of being cointegrated and characterised by an Error Correction Model (ECM), it is assumed that the cointegrating relationship (and so the tendency to move towards the long-run equilibrium) is not present at each time t but instead takes place only when the equilibrium between the variables involved trespasses one or more threshold levels. To illustrate this concept, Balke and Fomby (1997) make use of a bivariate model of the type:

$$\begin{cases} y_t + \alpha x_t = z_t & \text{where } z_t = \rho^{(i)} z_{t-1} + \epsilon_t. \\ y_t + \beta x_t = B_t & \text{where } B_t = B_{t-1} + \eta_t. \end{cases} \tag{2}$$

where both ϵ_t and η_t are iid random variables with mean 0. The first equation in the system represents the equilibrium relationship between y_t and x_t, with z_t called the equilibrium error and being the deviation from the equilibrium level and $(1, \alpha)$ the cointegrating vector; the B_t equation represents instead the common stochastic trend of y_t and x_t.

Any departure from the equilibrium z_t is supposed to follow a threshold autoregression as specified above ($z_t = \rho^{(i)} z_{t-1} + \epsilon_t$), where:

$$\begin{cases} \rho^{(i)} = 1 & \text{if } |z_{t-1}| \le \theta. \\ \rho^{(i)} = \rho, & \text{if } |z_{t-1}| > \theta. \end{cases} \tag{3}$$

where θ represents a critical threshold. As long as the equilibrium value is within the threshold level, the system does not mean revert towards the equilibrium level but as soon as the threshold level is surpassed, the cointegration relation takes place and the system drifts back to the equilibrium level. In the words of Engle and Granger (1987), "while locally z_t may have a unit root, globally this series is stationary". In case there are multiple threshold levels, their distance will also influence the long-term dynamics of the system: the more they are far apart, the longer it will take for the system to reach them and so the longer they will be characterised by a non-stationary behaviour.

6.1.3 Time-varying cointegration

For the purpose of this study, we propose to use the method proposed by Bierens and Martins (2010) to test the hypothesis of standard, time invariant cointegration against the hypothesis of time-varying cointegration. The reason for this is that there are several assumptions which presume that the relationship spanning between pollution, crop yields and climatic variables has not always been and will most likely not be invariant in the future.

To derive the test used for testing the presence of a time-varying cointegration, let's start by taking a Vector Error Correction Model (VECM) of the type:

$$\Delta Y_t = \Pi_t^T Y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta Y_{t-j} + \epsilon_t \tag{4}$$

where $t = 1,2,...,T$ the total number of observations, $\Pi_t^T = \alpha\beta_t^T$ where α and β_t are both (kxr) matrices, $\Delta Y_t = Y_t - Y_{t-1}$ a (kx1) matrix, $Y_t \in \mathbb{R}^k$, $\epsilon_t \sim i.i.d.N_k(0,\Omega)$ and Ω and Γ_j both fixed (kxk) matrices; given such a model, the objective of the test is to test the null hypothesis of time invariant cointegration, such as $\Pi_t^T = \Pi^T = \alpha\beta^T$ and both α and β are fixed (kxr) matrices with $rank(\Pi_t^T) = r < k$, against the alternative hypothesis of time-varying cointegration, such as $\Pi_t^T = \alpha\beta_t^T$ and α is still a fixed $(k \times r)$ matrix but β_t time variant, even though keeping the same dimension $(k \times r)$; in this second case we still have that $rank(\Pi_t^T) = r < k$.

In order to estimate β_t, which is an unknown function of time, Chebyshev Polynomials are used. Chebyshev polynomials are polynomials defined as:

$$P_{0,T}(t) = 1 \tag{5}$$

$$P_{1,T}(t) = \sqrt{2}cos(i\pi\left(\tfrac{t-0.5}{T}\right) \tag{6}$$

with $t = 1,2,...T$ being the time periods and $i = 1,2,3...,m$ the order of the polynomial. These polynomials are smooth functions of i and for all integers i, j we have that $\frac{1}{T}\sum_{t=1}^{T} P_{i,T}(t)P_{j,T}(t) = \mathbf{1}(i = j)$, which gives this polynomial its orthonormality characteristic. The orthonormality property allows to model any function $g(t)$ of discrete time as:

$$g(t) = \sum_{i=0}^{T-1} \xi_{i,T} P_{i,T}(t) \tag{7}$$

where

$$\xi_{i,T} = \frac{1}{T}\sum_{t=1}^{T} g(t)P_{i,T}(t) \tag{8}$$

are Fourier coefficients. In case β_t is smoothly evolving thorough time, as assumed in this framework, then β_t can be expressed as $\beta_t(m) = g_{m,T}(t) = \sum_{i=0}^{m} \xi_i P_{i,T}(t)$, which in turns allows us to get the following time-varying VECM specification:

$$\Delta Y_t = \alpha(\sum_{i=0}^{m} \xi_i P_{i,T}(t))_t^T Y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta Y_{t-j} + \epsilon_t \tag{9}$$

The time invariant cointegration corresponds to having $\xi^T Y_{t-1}^m = \beta^T Y_{t-1}^0$ where $\xi^T = (\beta^T, O_{r,k.m})$ and $Y_{t-1}^0 = Y_{t-1}$. Now, given that we have disclosed a way to estimate the time-varying VECM, testing the null hypothesis of time invariant cointegration against the alternative of time-varying cointegration boils down to using the following Likelihood Ratio (LR) test to determine which model performs better:

$$LR^{tvc} = -2[\hat{l}_T(r,0) - \hat{l}_T(r,m)] \tag{10}$$

where $\hat{l}_T(r,0)$ is the log likelihood of the time invariant VECM (and that is why $m = 0$), while $\hat{l}_T(r,m)$ is the log likelihood for the time varying VECM. In both cases, r represents the cointegrating rank as mentioned above.

6.2 Granger causality

Granger causality is based on two main principles: [i] the cause usually preceeds its effects and [ii] the cause makes unique changes in the effects. We say that a variables X Granger causes Y, if the past values of X support in predicting the future values of Y beyond what could have been done with the past values of Y only. The stationarity of the time series under consideration is a fundamental assumption of the Granger causality analysis. If the series involved are not stationary, they have to be made so through differencing or other techniques, such as taking the log of the series. Specifically, given two stationary time series X and Y we have two different information sets: [i] $I^*(t)$ the set of all available information up to time t and [ii] $I^*_{-X}(t)$ the set of all available information up to time t excluding the information provided by X. If X really aids

the prediction of future values of Y, the conditional distribution of the future values of Y should differ under the information set $I^*(t)$ and under $I^*_{-X}(t)$ [9]. Then, X is defined to Granger cause Y if

$$\mathbb{P}\big[Y(t+1) \in A | I^*(t)\big] \neq \mathbb{P}\big[Y(t+1) \in A | I^*_{-X}(t)\big] \tag{11}$$

for some measurable set $A \subseteq \mathbb{R}$ and $t \in \mathbb{Z}$, with A being the set of future realisations of $Y(t)$.

However, modeling the distribution of multivariate time series can be highly complicated, especially when using functions with non-convex loss landscapes such as deep neural networks. Moreover, the Granger causality definition does not give exact assumptions on the data generating process of the variables involved. For this reason, a usual approach to test for the presence of Granger causality is through the estimation of linear models, which tend to be easy to estimate and yet robust in their estimation. The VAR model is one of such models and one of the most used. The idea is the following. Given several time series $X_1, ..., X_v$, we estimate the following VAR for each of the X_j time series:

$$X_j(t) = \sum_{i=1}^{V} \beta_{j,i}^T \mathbf{X}_i^{t,Lagged} + \epsilon_j(t) \tag{12}$$

where $X_j^{t,Lagged} = [X_i(t-L), ..., X_i(t-1)]$ is the history of X_i up to time t, L is the maximal time lag and $\beta_{j,i} = [\beta_{j,i}(1), ..., \beta_{j,i}(L)]$ is the vector of coefficients modeling the effects of X_i on the target time series. The Granger causality is tested estimating the model with and without all the possible X_i values with $i = 1, ..., V$. If the conditional probability of the target variable X_j does not change under the different models, then there is no Granger causality as expressed in 12.

6.3 DeepAR

Given a target time series $\mathbf{z}_{i,1:t_0-1} = [z_{i,1}, ..., z_{i,t_0-2}, z_{i,t_0-1}]$ and wanting to estimate its future values $\mathbf{z}_{i,t_0:T} = [z_{i,t_0}, z_{i,t_0+1}, ..., z_{i,T}]$, we need to model the conditional distribution $P(\mathbf{z}_{i,t_0:T} | \mathbf{z}_{i,1:t_0-1}, \mathbf{x}_{i,1:T})$, where $\mathbf{x}_{i,1:T}$ is a time series of covariates assumed to be known at all time points. Assuming that the model distribution consists of a product of likelihood factors like

$$Q_\theta(\mathbf{z}_{i,t_0:T} | \mathbf{z}_{i,1:t_0-1}, \mathbf{x}_{i,1:T}) = \Pi_{t=t_0}^T Q_\theta(z_{i,t} | \mathbf{z}_{i,1:t-1}, \mathbf{x}_{i,1:T}) = \Pi_{t=t_0}^T p(z_{i,t} | \theta(\mathbf{h}_{i,t}, \Theta)) \tag{13}$$

where $\mathbf{h}_{i,t}$ is the output of an autoregressive recurrent network $\mathbf{h}_{i,t} = h(\mathbf{h}_{i,t-1}, z_{i,t-1}, \mathbf{x}_{i,t}, \Theta)$. The h is a function implemented by a multi layer Recurrent Neural Network (RNN) estimated with a Long Short Term Memory (LSTM) model and parametrized by Θ [27]. This model can be used in place of the VAR and then proceeding in the testing of Granger causality. Specifically, it's possible to assess differences in the DeepAR model estimation before and after using specific variables, assumed to Granger cause the target variables, by using the following causal significance score (CSS):

$$CSS_{i \to j} \ln \frac{MAPE_j^i}{MAPE_j} \tag{14}$$

where $MAPE_j^i$ is the mean absolute percentage error between the $\hat{z_{j,t}}$ and the real $z_{j,t}$ using the variable $z_{i,t}$ and $MAPE_j$ without using $z_{i,t}$.

6.4 Knockoff counterfactual

The knockoff counterfactual technique was first proposed in 2015 [12]. The idea of the technique is to swap the original variables with some fake ones and checking if the model estimations change. Given the set of the original variables Z such that $Z = Z_1, Z_2, ..., Z_n$, with distribution P_z, the knockoffs are created such that they are in-distribution null variables. The knockoffs have the same distribution as the original variables but they do not contain any information about the target variable, and for this reason they can be swapped with the original variables to check how the model estimation change. Moreover, the knockoffs have the same covariance structure and the correlation between the knockoffs is the same as the correlation between the original variables.