Meteorological Data Science: exploiting causality discovery in time-series for knowledge discovery and improved forecasting

Gkikas A.[1], Maragoudakis M.[1,*]

1 Department of Information and Communication Systems Engineering, University of the Aegean, Karlovasi, Samos, 83200, Greece

*corresponding author:Maragoudakis M.: e-mail: mmarag@aegean.gr

Abstract

Climate change and its impact on everyday life still remains one of the greatest challenge of our era. The complex nature of climate data addresses the use of data science techniques to provide predictive analytics to the task at hand. While most existing approaches exploit correlation between observations and features to improve forecasting, the present work deals with causality, a principle that enhances robustness and provides better insight to domain experts. More specifically, a novel framework for causality discovery is proposed, based on statistical (i.e. Granger causality tests) as well as on non-linear state space reconstruction algorithms (i.e. Convergent Cross Mapping, a very effective algorithm in dynamic systems, such as the task at hand) in order to find the causal relations between meteorological time series. Furthermore, the framework also supports methods for graph analysis, thus providing informative visualizations on the influential levels of causality. Experiment results on a dataset of real observations from different cities of Greece, obtained through crawling of Internet sites of Davis weather stations demonstrate the ability to model and visualize the relations of the meteorological parameters amongst the cities. Moreover, by utilizing such causal inference knowledge, the forecasting performance for each city is significantly improved, since only relevant and informative features were taken into consideration.

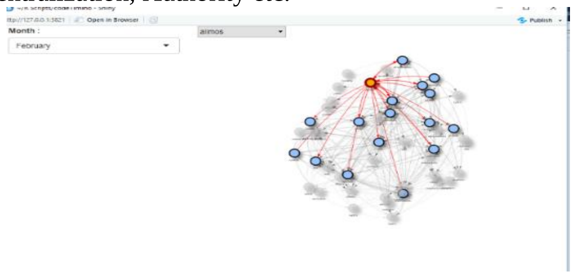
Keywords: Data Science, Causal Inference, Time-Series Analysis, Graph Analysis, Feature Selection.

1. Introduction

In many scientific fields, there is a need to explain the interactions of the indices we are studying and finding out about their causality inference, so we could better understand how correlated they are. the Granger Causality Test is a well-known method for the causation discovery between time series. It has been successfully applied in the field of finance (Akinboade, O. A. and Braimoh, L. A.,2010), but also in neuroscience, since it helps to find parts of the brain that affect others (Bressler, S. L. and Seth, A. K., 2011). A novel approach (Convertino, V. A. et al.,2015), also done in the climate science, for detecting causality inference between PM2.5(atmospheric particulate matter) and meteorological factors in region of Jing-Jin-Ji using CCM algorithm. Over the years, there have been many variations of the classic Granger Test to meet the needs for better performance of the causal discovery process. The test can be carried out both in the time domain and in the spectral domain, giving scientists a wide range of applications. We need to remind that correlation does not imply causation (Aldrich, J., 1995). In Granger Test, a relation between two time series in the form of: X→ Y denotes that past observations of X can help towards predicting Y. The relation between two variables could be unidirectional(X→Y) or bidirectional (X←→Y). A significant parameter for the test is lag order, which is the number of past observations to be taken account. Other methodologies that discover causality include Timino (Peters, J., et al.,2013), which outputs a DAG and also avoids wrong conclusions, such as existent of cofounders or Instantaneous effects, and CCM (Convergent Cross Mapping) (Sugihara, G., et al.,2012) for modeling non-linear dynamic systems, as is the ecosystem, using convergent cross mapping.. Our main goal is to create a framework with the adding ability of a visualized graph, in order to provide users with visualization capabilities to better understand the extracted results.

2.Framework

The build up framework is based on R, a powerful programming language, ideal for mathematical purposes. It supports two algorithms, Granger causality tests and EDM analysis (with the support of rEDM package). Users can input data frame of time series and choose the algorithm that is best suited for the data. Upon the analysis of causality, the output can be a visualized as a network (with the use of visNetwork and Shiny libraries) with connections showing the causes between nodes. Each node represents a variable/timeserie in the model we study and each arc, represents a directed connection between two nodes showing which variable driven by the other. One of the features provided by the framework is the choice for splitting the timeseries into chunks (time windows) and separately run analysis for each (Figure 1). For example, when a phenomenon of extreme weather events occurs, it would be better to separate it from the whole time series and to study it individually. This gives the opportunity to a clearer picture of how the network evolves over time, which is very important when handle dynamic changing variables, reducing the likelihood of false positives relations between the variables. Furthermore, output data can be saved in format ready for importing in Gephi, a java application with SNA (social network analysis) capabilities. Some popular SNA metrics that are supported include PageRank, Centralization, Authority etc.



Month :
February

alimos

Figure 1.Showing Temperature relations in Atiiki region.

3.Experiments

For our experiments we used a dataset, which was contributed by Meteo*, validated network of meteorological stations in Greece. It is containing one-year (2017) observations from stations in region of Attiki, with sample rate of 1 hour. Furthermore, we gathered meteorological data of two months, like temperature, wind, humidity, etc., from large cities (with population of over 15.000) of Greece including islands, keeping the same sample rate. Since the ecological system is dynamic, the preprocessing of data has verified that time series of environmental factors are non-linear, resulting in the decision of proceeding with non-linear model to perform causal inference. Moreover, since weather stations often fail to give correct information, due to some network failure, there are discontinuities in time series, which Granger causality cannot manage. So, during preprocess, the missing values where filled with the previous known value.

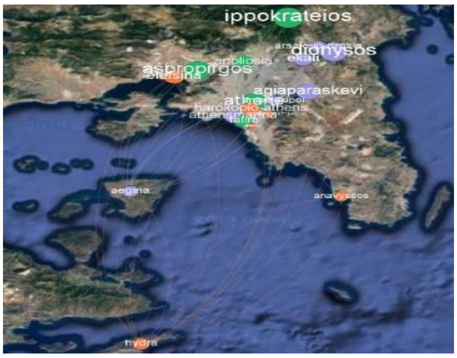


***Figure 2.** Gephi graph overlayed in google maps.Shows ,which regions are more casual infrerence by others.*

As shown in Figure 2, each node corresponds to the geographic location of a Davis station,overlayed in Google Maps, for better understanding of how each micro-climate affects the other. We use the metric of authority, which belongs to a link analysis algorithm (called HITS or else Hubs and authorities), where it classifies the nodes. Statistically, authority shows how important a node is in the network, taking as a measure the number of incoming arcs. A good Hub is the one which has many outcoming edges. As we can see, It appears, that the areas which are affected most by other areas, are Aspropyrgos, Dionysos,Ippoktaeio and Athens Center. A small causality, show remote areas like Aegina, Hydra and Anavyssos, but even more strange it seems that nearby Alimos and Hymmetos areas have the smallest authority, as they have 2 and 3 inbound connections respectively.

4.Future Work

The framework may display causality between nodes and the strength of each node in the given network. In the future, we going to concentrate our work to build a forecast model in our framework using as input our results, and evaluate it with already existing frameworks, for any improvements.

References

Akinboade, O. A., & Braimoh, L. A. (2010). International tourism and economic development in South Africa: A Granger causality test. *International Journal of Tourism Research*, *12*(2), 149-163.

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. Statistical science, 10(4), 364-376.

Bressler, S. L., & Seth, A. K. (2011). Wiener–Granger causality: a well established methodology. Neuroimage, 58(2), 323-329.

Convertino, V. A., Howard, J. T., Hinojosa-Laborde, C., Cardin, S., Batchelder, P., Mulligan, J., ... & MacLeod, D. B. (2015). Individual-specific, beat-to-beat trending of significant human blood loss: the compensatory reserve. Shock, 44, 27-32.

Peters, J., Janzing, D., & Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. In Advances in Neural Information Processing Systems (pp. 154-162).

Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. science, 1227079.

*www.meteo.gr