https://doi.org/10.5194/bg-2021-231
Preprint. Discussion started: 6 September 2021

Biogeosciences Discussions

Comment on "Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach"

Jarmo Mäkelä[1], Laila Melkas[1], Ivan Mammarella[2], Tuomo Nieminen[2,3], Suyog Chandramouli[1], Rafael Savvides[1], and Kai Puolamäki[1,2]

[1]Department of Computer Science, P.O. Box 68, FI-00014 University of Helsinki, Helsinki, Finland
[2]Institute for Atmospheric and Earth System Research / Physics, P.O. Box 64, FI-00014 University of Helsinki, Helsinki, Finland
[3]Institute for Atmospheric and Earth System Research / Forest Sciences, P.O. Box 27, FI-00014 University of Helsinki, Helsinki, Finland

Correspondence: Jarmo Mäkelä (jarmo.makela@helsinki.fi)

Abstract. This is a comment on "Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach" by Krich et al., Biogeosciences, 17, 1033–1061, 2020, which gives a good introduction to causal discovery, but confines the scope by investigating the outcome of a single algorithm. In this comment, we argue that the outputs of causal discovery algorithms should not usually be considered as end results but starting points and hypothesis for further study. We illustrate how not only different algorithms, but also different initial states and prior information of possible causal model structures, affect the outcome. We demonstrate how to incorporate expert domain knowledge with causal structure discovery and how to detect and take into account overfitting and concept drift.

1 Main text

In a recent paper Krich et al. (2020) tested and applied a newly developed PCMCI algorithm (Runge, 2020; Runge et al., 2019) in order to detect causal links in geophysical data. The algorithm is used on flux tower eddy covariance data and related meteorological measurements of six variables in order to detect which variables can be seen to steer the behaviour of others. The paper can be viewed as a proof-of-concept and is a good introduction to causality and underlying problems, given the novelty of applying these types of methods to better understand biosphere-atmosphere interactions. However, we feel that contribution of Krich et al. covers only one part of practical application of causal discovery algorithms. There were items that in our opinion are significant for practical application of such causal discovery methods and which were only briefly mentioned or not at all addressed in the paper by Krich et al. (2020). These are:

- The outcomes (models) of causal structure discovery (CSD) algorithms are, in many cases, interchangeable: it is very difficult to identify the "correct" model purely based on data.
- The choice of initial state affects the final model. Due to their setup, Krich et al. (2020) employed an empty graph, but other choices are also possible.

https://doi.org/10.5194/bg-2021-231
Preprint. Discussion started: 6 September 2021


Biogeosciences Discussions Open Access EGU

- Utilising the knowledge of the domain experts and user interaction can be used to improve the models.
- Overfitting and concept drift. Overfitting means that the analysis relies too much on the training data. Usually this happens when the amount of data is too small, resulting the causal model fitting to noise. Concept drift means that the underlying data distribution changes, rendering the causal model irrelevant. An example of a concept drift is that a model trained on a certain location may not describe relations in another location; it is important to be able to take this phenomena into account.

These comments are based on our recent workshop paper in the KDD 2021 conference (Melkas et al., 2021). Since many experts in Earth system sciences are not likely to follow said conference, we wanted to convey the main findings via this reply to Krich et al. (2020) as it also originally inspired us to explore the topic. In short, our workshop paper presents a procedure on how to utilise prior knowledge of the domain experts in finding causal structure discovery (CSD) models and how a user might incorporate this knowledge with CSD algorithms. This knowledge can be characterised by a prior distribution over all possible causal structures. We use both synthetic data as well as flux tower eddy covariance variables – same variables as in Krich et al. (2020)) – measured at the SMEAR II station at Hyytiälä, Finland (Mammarella, 2020). We simulate the user's choices with a greedy search from the neighboring states of the current model. By "neighbourhood" we mean the models that can be reached from the current model by simple edits and "greedy" we mean that the user always chooses the best model from the neighbourhood of the current model, and this process is iterated, until the current model is at least as good as any of the neighbours – see Melkas et al. (2021) for details. The outcomes are also compared to a model produced by actual domain experts. The takeaway message is that instead of using expert knowledge to merely quality check the final model produced by a CSD algorithm, the prior knowledge should be incorporated into the process.

2 Differences in CSD algorithms

While Krich et al. (2020) have focused on PCMCI, it is worthwhile to note that different CSD algorithms have varied outputs (models) for the same input data (Druzdzel, 2009) since each algorithm makes different assumptions about the underlying data. Additionally, even if the modelling assumptions in the causal discovery process are correct, insufficient or biased data may result in skewed results. Therefore, the model gained from any one of these algorithms should not be viewed as the end result, but rather a starting point for further analysis. Often it is not clear, which among the discovered models is the "best", although we can argue that some of them are more plausible (Runge et al., 2019), given the expert's knowledge. In some algorithms, inputting this prior knowledge (e.g., probabilities of certain structures) is possible, but the ability to iteratively refine this background knowledge during the data analysis process nor the possibility to express uncertainty in the prior information have not been built in. These caveats hinder the usability of many CSD algorithms.

https://doi.org/10.5194/bg-2021-231
Preprint. Discussion started: 6 September 2021

Biogeosciences Discussions Open Access EGU

3 The choice of initial state

As different algorithms produce different models, so does the choice of initial state affects the outcome. These states can be, for example, empty graphs (as in Krich et al. (2020)), states produced by sampling methods, or states that reflect certain expert knowledge. Depending on the choice of initial state and on how uncertain the prior information is, different locally optimal models that fit the data may be found. Intuitively, it would be interesting to have a set of initial states that would cover all local optima, which could give rise to a global maximum-a-posteriori (MAP) solution. The underlying problem here would be to find a representative set of starting points for the exploration.

We demonstrate the combined effect of utilising multiple initial states and different levels of prior knowledge (k) with synthetic data (Fig. 1). The initial states are generated by four different CSD algorithms and are complemented by an empty graph and the correct model, which we know as the data is synthetic. The user knowledge is reflected by parameter k, where $k = 1$ indicates that user has full knowledge of the causal structure and $k = 1/3$ means that the user has no prior information (see Melkas et al. (2021) for details; values of $k > 1/2$ lead to near constant results). The structural Hamming distance (SHD) indicates how many modifications to a model have to be made in order to end up with another model. Even with a small amount of prior information, the end result after user interactions (greedy search) becomes much more stable – the spread of SHD diminishes as k increases (Fig. 1).

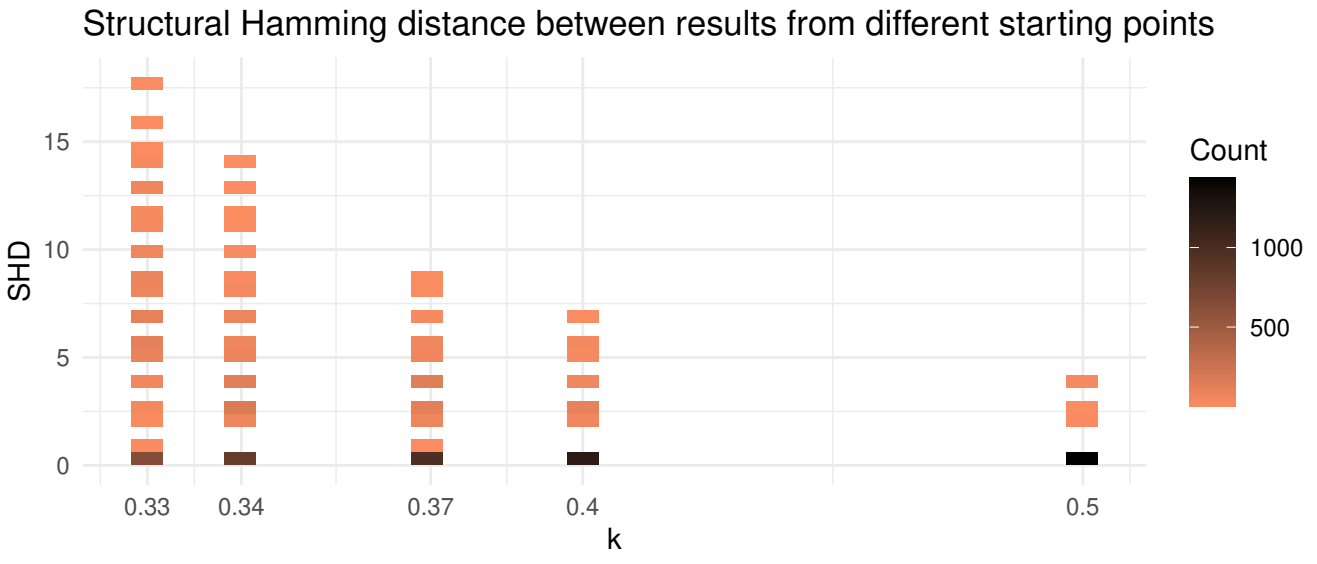


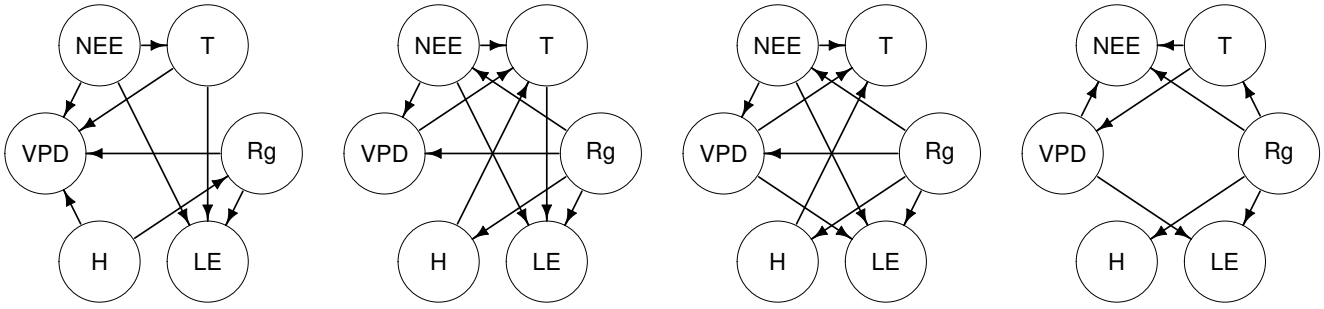
Figure 1. Pairwise structural Hamming distances when running analysis on the same data starting from different initial models. Variance in the distances show that the final model is affected by choice of initial model. Additionally, the spread of distances decreases rapidly with increasing prior knowledge.

https://doi.org/10.5194/bg-2021-231
Preprint. Discussion started: 6 September 2021

Biogeosciences Discussions Open Access EGU

4 Utilising expert knowledge and user interactions

The knowledge of the domain experts is classically used to provide suitable initial states for the CSD algorithms or to quality check the outcomes, but this knowledge should also be used to steer the CSD processes via user interactions and to allow reassessment of both user's own prior knowledge and related uncertainty as well as the algorithm process. When this knowledge is disregarded and the data is blindly trusted, any CSD algorithm or user (e.g., our greedy search) can uncover erroneous connections and miss relevant ones (Fig. 2). For example, the expert model (d) identifies four direct and well-established causal links from downwelling shortwave radiation (Rg) to latent and sensible heat fluxes (LE,H), temperature (T) and net ecosystem exchange (NEE). Two of these links (T and NEE) are missing from the best scoring model among the CSD algorithms (a), which also erroneously asserts that H is a driving force behind Rg. Both user models (b,c) find a new unrealistic link from Rg to vapour pressure deficit (VPD) and indicate that Rg is affecting T only indirectly through NEE.



(a) Initial model from algorithms. (b) Final model starting from (a). (c) Final model starting from an empty graph. (d) Expert model.

Figure 2. The user (greedy search) finds slightly different models (b,c) whether we start the search from the best scoring model among our CSD algorithms (a) or an empty graph. The underlying causal structures were given a uniform prior. Also shown is the expert model, produced before these experiments. The SHD from the expert model to (a),(b) and (c) are ten, seven and five.

5 Overfitting and concept drift

Overfitting the model to the data is a common problem in statistical modelling, but to the best of our knowledge this problem has not been addressed in the context of CSD. In Melkas et al. (2021) we demonstrate how to detect overfitting using k-fold blocked cross-validation (Bergmeir and Benítez, 2012). The same method is also applicable in detecting concept drift, which we induced by including a set of measurements taken in August 2015 to calibration data containing measurements taken in April 2013–2015 – this violates causal stationarity stemming from seasonality.

https://doi.org/10.5194/bg-2021-231
Preprint. Discussion started: 6 September 2021


Biogeosciences Discussions Open Access EGU

6 Concluding remarks

Novel CSD algorithms, and more generally many machine learning methods, offer new insights in Earth system sciences. We argue that combining these methods with already abundant knowledge of the domain experts will yield more robust results. We also argue that while there are plethora of CSD algorithms that has been applied in earth sciences the question of how to use them in practice is still open. We have briefly presented here one fairly simple approach as how to achieve this, demonstrated its effectiveness and highlighted some pitfalls. Hopefully, this will encourage developers to implement and study further interactive workflows. We direct anyone interested in a more detailed presentation to see Melkas et al. (2021).

Author contributions. JM prepared the comment, while LM ran the simulations and prepared the KDD manuscript under supervision of KP. IM and TN provided the domain expert knowledge and together with SC and RS commented the original paper and this comment.

Competing interests. The authors declare that they have no competing of interests.

Acknowledgements. We thank Helsinki Institute for Information Technology, Future Makers Funding Program, and Finnish Center for Artificial Intelligence for support.

https://doi.org/10.5194/bg-2021-231
Preprint. Discussion started: 6 September 2021

References

Bergmeir, C. and Benítez, J. M.: On the use of cross-validation for time series predictor evaluation, Information Sciences, 191, 192–213, https://doi.org/https://doi.org/10.1016/j.ins.2011.12.028, 2012.

Druzdzel, M. J.: The role of assumptions in causal discovery, in: Workshop on Uncertainty Processing (WUPES-09), pp. 57–68, University of Pittsburgh, http://d-scholarship.pitt.edu/6017/, 2009.

Krich, C., Runge, J., Miralles, D. G., Migliavacca, M., Perez-Priego, O., El-Madany, T., Carrara, A., and Mahecha, M. D.: Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach, Biogeosciences, 17, 1033–1061, https://doi.org/10.5194/bg-17-1033-2020, 2020.

Mammarella, I.: Drought 2018 Fluxdata Preview Selection, Hyytiälä, 1995-12-31–2018-12-31, https://hdl.handle.net/11676/EBmVEuoJaOmOw8QmUyyh6G-n, 2020.

Melkas, L., Savvides, R., Chandramouli, S., Mäkelä, J., Nieminen, T., Mammarella, I., and Puolamäki, K.: Interactive Causal Structure Discovery in Earth System Sciences, arXiv:2107.01126 [physics.data-an], 2021.

Runge, J.: Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets, in: Conference on Uncertainty in Artificial Intelligence, edited by Peters, J. and Sontag, D., vol. 124 of *UAI'20*, pp. 1388–1397, PMLR, http://proceedings.mlr.press/v124/runge20a.html, 2020.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D.: Detecting and quantifying causal associations in large nonlinear time series datasets, Science advances, 5, https://doi.org/10.1126/sciadv.aau4996, 2019.