Biogeosciences Discussions

Biogeosciences Discuss., referee comment RC1
https://doi.org/10.5194/bg-2021-231-RC1, 2021

Comment on bg-2021-231

Jakob Runge (Referee)

Referee comment on "Technical note: Incorporating expert domain knowledge into causal structure discovery workflows" by Jarmo Mäkelä et al., Biogeosciences Discuss., https://doi.org/10.5194/bg-2021-231-RC1, 2021

The comment by Mäkelä et al. on the paper "Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach" by Krich et al. (including the reviewer) makes the point that the study should not take the outcome of a single causal discovery algorithm (here PCMCI) as an end result, but as a starting point and hypothesis for further study. They further illustrate on synthetic data how different prior expert knowledge affects such algorithms. The authors link to their recent workshop paper in the KDD 2021 conference (Melkas et al., 2021) which expands on the topic of "interactive" structure discovery.

Overall I deem this as a good and valid general point for any causal discovery analysis. However, I am not sure the commenting section is appropriate for this type of content since it does not specifically conduct an in-depth analysis of the paper to elaborate on how results would differ, but here it seems to mainly serve to advertise their workshop paper.

The authors present three different aspects of what they call "user interaction": (1) starting from a user-guided initial state, (2) expert-interactions during the execution of the causal discovery algorithm, and (3) overfitting and concept drift.
These three points are discussed with very minimal examples and a few bits are unclear from the text: Are all numerical analyses conducted with synthetic data? What's the setup? Can point (2) be elaborated on a bit, it is hard to understand how this interaction is meant here.

To put these comments in context with the actual paper: Yes, in the scope of this paper (Krich et al.) no initial prior knowledge (other than the choice of variables and that the type of dependency is linear) was used. However, the resulting graph was discussed from an expert perspective. The problem of overfitting was addressed in so far that the hyper-parameter (pc_alpha) was optimized based on the Akaike Information Criterion, which is aymptotically equivalent to the cross-validation suggested in the comment. Indeed the paper can be viewed as a proof-of-concept and introduction to causality and underlying

problems.

As a remark, including expert knowledge into causal discovery is an interesting and not quite trivial problem. For example, while it may be easy to code-up (PCMCI's software package Tigramite has an option to start from a user-given initial graph), the completeness ("maximal informativeness") of causal discovery algorithms under expert knowledge is an open problem, at least for more complex scenarios such as the presence of hidden variables.

References:
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