Causal Discovery from Spatio-Temporal Data with Applications to Climate Science

Imme Ebert-Uphoff
School of Electrical and Computer Engineering
Colorado State University
Fort Collins, CO, USA
Email: iebert@engr.colostate.edu

Yi Deng
School of Earth and Atmospheric Sciences
Georgia Institute of Technology
Atlanta, Georgia, USA
Email: yi.deng@eas.gatech.edu

***Abstract*—Causal discovery algorithms have been used to identify potential cause-effect relationships from observational data for decades. Recently more applications are emerging, for example in climate science, that extend over large spatial domains and require temporal models. This paper first reviews how the causal discovery problem can be set up for such spatio-temporal problems using constraint-based structure learning, then discusses pitfalls we encountered and some solutions we developed. In particular, we consider how to handle temporal and spatial boundaries (which often result in causal sufficiency violations) and discuss the effects of temporal resolution and grid irregularities on the resulting model.**

I. Introduction

Causal discovery theory is based on *probabilistic graphical models* and provides algorithms to identify potential cause-effect relationships from observational data [1], [2], [3]. The output of such algorithms is a graph structure showing potential causal connections of the observed variables. Causal discovery has been used routinely in applications in the social sciences and economics for decades [2], [3]. More recently, causal discovery has been used with great success in biology [4] and bioinformatics [5], for example to identify gene regulatory networks [6], [7], identify protein interactions [8], [9] and discover neural connections in the brain [10]. In recent years causal discovery has emerged in many physics-related applications, including applications in earth science, such as studying tele-connections in the atmosphere [11], pollution models [12], precipitation models [13], sea breeze models [14], and studying the impact of climate change [15].

So far the great majority of causal discovery applications uses static models with few variables, but a growing number of applications is emerging that (1) require temporal models and (2) use a large number of variables extending over large spatial domains. In this paper we first summarize the key concepts for using causal discovery for spatio-temporal data. Then we present pitfalls we encountered and some solutions we developed over the years while using causal discovery to learn temporal models from spatio-temporal data.

The remainder of this paper is structured as follows. Section II provides a quick introduction to causal discovery through constraint-based structure learning. Section III reviews a little known extension of standard constraint-based structure algorithms that yields temporal models. Section IV introduces two sample applications that are used throughout this paper to illustrate these concepts. Sections V and VI present pitfalls that arise when using spatio-temporal data and discuss potential solutions and Section VII provides some discussion and conclusions.

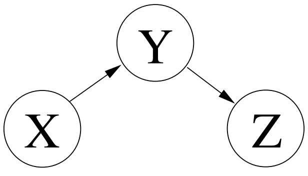
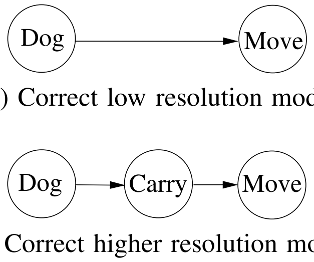


Fig. 1. Sample graph illustrating direct and indirect connections



(a) Correct low resolution model

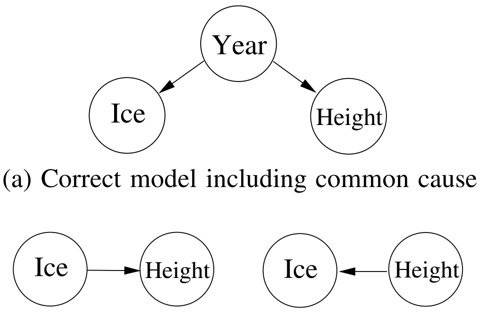
(b) Correct higher resolution model

Fig. 2. Direct connections can change into indirect connections when more variables are included in the model. Both models are correct.

II. Key concepts of Causal Discovery

Fig. 1 shows a simple graph indicating causal relationships between three variables, X, Y, Z. Each variable constitutes a node in the graph and an arrow from one node to another indicates *direct* cause-effect relationships between the variables. For example, for the system in Fig. 1, X is a direct cause of Y and Y is a direct cause of Z. X also has an effect on Z through Y, but as such X is only an *indirect* cause of Z. Thus there is no arrow connecting X directly to Z in the graph.

In this context *directness* is a relative term, because it is always *relative to the nodes included in the model*. Let us consider the example of a dog, a room and a ball. Every time the dog enters the room, it picks up the ball and carries it somewhere else. So if we use only two variables to describe this system, *Dog*, which denotes whether the dog is in the room, and $Move$, which denotes whether the ball is moving,



(a) Correct model including common cause

(b) Incorrect models omitting common cause

Fig. 3. Example showing errors caused by omitting a common cause

then the model simply becomes as shown in Fig. 2(a), where Dog is a direct cause of $Move$. However, if we also introduce the variable $Carry$, which denotes whether the dog carries the ball, then the model becomes as in Fig. 2(b), where Dog is no longer a *direct* cause of $Move$. Thus, by increasing the causal resolution of the model, direct connections can turn into indirect connections. Both of those models are correct and match our intuitive understanding of causality.

Hidden common causes - also known as *latent variables* or *confounding variable* - are another key concept. We illustrate them briefly with a very simple example. Let us say we have three variables, Ice, which describes the amount of ice left on polar caps (generally decreasing in recent years), $Height$, which describes the average height of people on our planet (generally increasing in recent years), and $Year$, which represents the year in which the data is taken. The correct model relating those three variables can be seen in Fig. 3(a), where $Year$ represents the relevant processes taking place over time and is thus the common cause of changes in the amount of ice and the average height. However, if $Year$ is not included in the model, the strong anticorrelation between Ice and $Height$ may be misinterpreted as a direct connection (Fig. 3(b)), leading to the erroneous conclusion that the melting ice makes people taller, or vice versa. We will return to the topic of hidden common causes in the following sections.

A. Constraint-Based Structure Learning

We employ the well known framework of *constraint-based structure learning* of graphical models [1], [2], [16] for causal discovery. The specific algorithms we use are the classic *PC* algorithm [17], [2] and the *PC stable* algorithm [18]. *PC stable* is a variation of *PC* and has several advantages, namely (1) it is order-independent, i.e. the order of variables does not impact the results; (2) it is more robust, i.e. mistakes early on cause less follow-up mistakes in the graphs; (3) it is easy to parallize, thus reducing execution time.

Learning causal structure from data with *PC* or *PC stable* is based on two key facts:

1) We can distinguish between *direct* and *indirect* connections based on observed data using conditional independence tests (CI tests);
2) We *cannot prove* causal connections (primarily due to potential hidden common causes), but we can *disprove* so many connections that only few potential causal relationships are left at the end.

The basic steps of both the *PC* and *PC stable* algorithms are as follows:

1) We establish a graph, where the observed variables form the *nodes* of the graph.
2) First we assume that every variable is a cause of every other variable (fully connected graph).
3) Then we perform CI tests to eliminate as many connections as possible (pruning).
4) Whatever is left at the end are the *potential* causal connections.
5) Arrow *directions* are determined (as far as possible) from additional CI tests, temporal constraints (if available) or any available prior knowledge.

This procedure yields one or more *independence graphs*, which represent the conditional independencies in the data.

B. Conditions for Causal Interpretation

The result of the constraint-based structure learning are independence graphs and we need to consider under which conditions these graphs can be interpreted in a causal way. There are two types of conditions, which are briefly discussed below (for more details see for example [16]).

1) From data to independence graph: Going from probability distribution (data) to independence graph, we have to make sure that the obtained independence graph actually models the data well, i.e. that it is *faithful* to the probability distribution. This condition roughly translates into the following practical guidelines: (a) The independence signal must be strong enough to be picked up by the statistical tests in the presence of noise. (b) No selection bias is allowed. (c) Probability distributions must be identical and independent. (d) If the independence graph is directed, no causal loops are allowed in the system. If causal loops are present, then a temporal graphical model should be used instead.

2) From independence graph to causal interpretation: Going from independence graph to causal interpretation, we have to make sure that there are no hidden common causes or other conditions that could cause the independence graph to misrepresent a system's causal relationships. The primary concern is to ensure that the nodes in the graph are *causally sufficient*, i.e. if any two nodes X, Y of the graph have a common cause, Z, then Z must also be included in the graph. In practice for real-world systems this condition is often impossible to ensure, typically because some common causes may be unknown or may be hard to observe. Some algorithms have been developed that can identify the existence of many hidden common causes ([2]), but are of high computational complexity and currently not feasible for large graphs. Recent advances [19] may change that in the near future.

Our approach is to simply *not* assume causal sufficiency, and to interpret the results accordingly. We accept that then we *cannot prove* causal connections, but we can *disprove* so many connections that only few *potential* causal relationships are left at the end. Each one of those relationships may present a true causal connection, be due to a common cause or a combination of both.

Evaluation step: Thus we include a final evaluation step in our analysis. In the final graph, every link (or group of links) must be checked by a domain expert. If we can find a mechanism that explains it (e.g. from literature), the causal connection is confirmed. Otherwise, the link presents a *new hypothesis* to be investigated.

When seeking to learn new knowledge from data we have found that the optimal scenario is to have most links in the final graph confirmed from literature, thus confirming that the overall approach is correct, but also having a few unconfirmed links. The unconfirmed links are crucial, because they are the only ones that provide *new* hypotheses of causal connections and thus potentially *new knowledge*.

III. Incorporating Spatio-Temporal Data

A. *Incorporating Spatial Dimensions*

Spatial dimensions are incorporated by using different variables, X_i, that represent the quantity of interest at different locations. If a single type of quantity is considered (e.g. temperature or wind), then i simply indicates the location where it is measured. Otherwise some numbering scheme is used, so that the combinations of quantities and locations can be referred to as $X_i, i = 1, \ldots, N$. While setting up the problem seems trivial at first, we will see in Section VI-B that the spacing between the measurement locations can result in artifacts in the resulting graphs, so proper spacing - or at least understanding the potential problems if proper spacing is not possible - is critical.

B. *Incorporating Time*

Both the *PC* and *PC stable* algorithms can be extended to learn *temporal* models from time series data. However, in their standard form they simply disregard the temporal component, as follows. Let us assume the time series data is provided in a matrix with each *column* representing one observed variable and each *row* representing a time step. The standard *PC/PC stable* algorithm treats this data as if all samples are independent, i.e. one can mix up the order of the rows arbitrarily in the matrix without changing the results. Thus the algorithm effectively disregards time and treats the system as a *static* system, so we call the results *static models*. This is sufficient for many applications, e.g. most applications in economics and social sciences, where causal relationships last for very long times, and thus the relationships can be treated as quasi-static.

For our applications in climate science temporal information typically plays a crucial role, especially when dealing with daily data. The climate system is very dynamic, with states at individual locations changing from day to day, interactions between different locations happening within days, and the strength of many signals also decaying within days. Therefore for our applications taking time into account provided much stronger causal signals. In fact, for the applications in climate science we considered, static models were unable to provide robust results, and we had to move to temporal models to be able to identify strong, robust causal signals. We believe the same to hold for many physical systems in which temporal order is important. Another advantage of temporal models is that temporal information helps to establish causal *directions*.

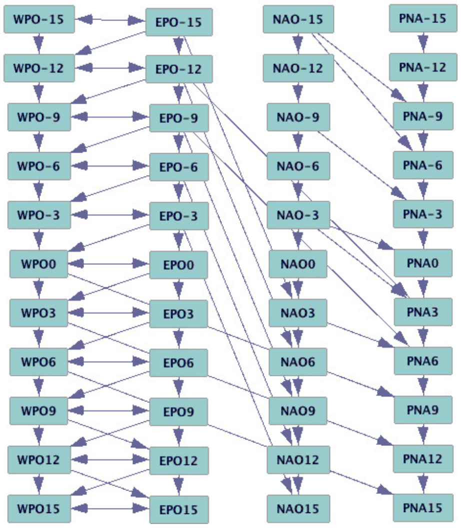


Fig. 4. Temporal graph from *PC* algorithm for $D = 3$ and $\alpha = 0.001$

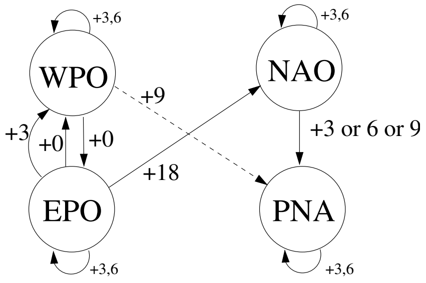
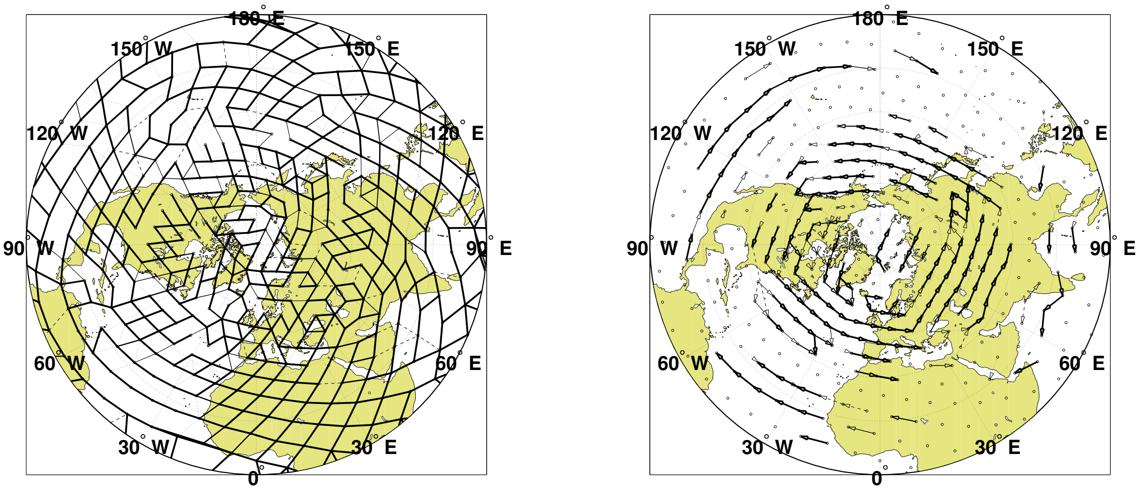


Fig. 5. Summary graph for Fig. 4 summarizing strongest connections

To incorporate time explicitly into the modeling we use the approach first introduced by Chu et al. [11], which adds lagged variables to the model. Since this approach does not seem to be widely known, we briefly outline it below. Given time series data for N variables, $X_1, \ldots, X_N$, the basic approach is as follows:

1) Choose the distance, D, between time slices, e.g. $D = 3$ time steps.
2) Choose number of time slices to include, e.g. $[-M, M]$.
3) Define lagged variables for all $i = 1, \ldots, N$ and $s \in [-M, M]$: $X_i^s = X_i$ lagged by s time slices. This results in a total of $(2M + 1) \cdot N$ variables, which form the *nodes* of the temporal graph.
4) Add temporal constraints: Causes can only occur before or at the same time as their effect, i.e. X_i^s can be a cause of X_j^t only if $s \leq t$.

Using this procedure we can express the temporal model with N time series variables and $S = (2M + 1)$ time slices as a



(a) Travel $<$ 1 day

(b) Travel $\approx$ 1 day

Fig. 6. Network plots for Northern hemisphere from *PC stable* ($D = 1$ day, $\alpha = 0.1$) based on Fekete grid (800 grid points) yield meaningful results.

standard static problem with $(N \cdot S)$ variables, plus temporal constraints. The temporal constraints can be incorporated in *PC/PC stable* as prior knowledge, so that the standard algorithms can now be used to provide a temporal model. The price we pay for this though is much higher computational complexity, since we are now dealing with $N \cdot S$, rather than N variables. While the worst-case complexity for the PC algorithm is exponential in the number of variables, in practice, its complexity depends very much on the application, since its connectivity properties determine which order of conditional independence tests are required to eliminate the majority of the graph edges. Thus in practice the number of calculations required for the PC algorithm is polynomial in the number of nodes (e.g. N^5), but that means that going from a static to a dynamic model, say with $S \geq 10$, increases complexity by several orders of magnitude.

It is a little known fact that use of the lagged variables violates one of the assumptions of constraint-based structure learning discussed in Section II-B, namely that the probability distributions should be independent of each other. So far it seems that this violation does not affect the method at all, since it works well even in spite of this violation. Nevertheless, this issue should be studied further, to understand why it works so well and to ensure that this is always the case. However, that topic is beyond the scope of this paper.

Finally, we will see later that initialization of the first time slices is a critical issue, but one that can be resolved easily, see Section V-A.

IV. Sample Applications

We introduce two applications from climate science to demonstrate the great potential of causal discovery in this area, and to demonstrate some pitfalls in later sections.

Application 1 - Relationships between compound indices: We investigated the potential causal connections between four compound indices that are often used in climate science, namely the Western Pacific Oscillation (WPO), Eastern Pacific Oscillation (EPO), Pacific North America Pattern (PNA) and North Atlantic Oscillation (NAO). Using daily data for those indices, and using time slices that are $D = 3$ days apart and a significance value of $\alpha = 0.001$ for the conditional independence tests (Fisher Z-tests), the *PC* algorithm yielded the temporal model shown in Fig. 4. Fig. 5 is a summary graph with the strongest connections from Fig. 4. The numbers next to the arrows indicate the delay in days from potential cause to effect (they are multiples of 3, because we used $D = 3$). There are a few arrows hidden behind other blocks in Fig. 4. Namely, each compound index, WPO, EPO, PNA and NAO, actually affects itself strongly for 3-6 days, as shown in the self-loops in Fig. 5. For more interpretation, see [20].

Application 2 - Graphs of information flow: One of the most complex applications of causal discovery in climate science is to track the pathways of physical interaction around the globe. In order to do that we define a grid around the globe and evaluate an atmospheric field (such as temperature or geopotential height) at all grid points, which provides time series data at all grid points. Our approach is to then use the temporal version of *PC stable* to identify the strongest *pathways of interactions* around the globe based on the time series data [21]. Gaussian graphical models present an alternative approach for this purpose [22] and [23] investigates this and other approaches. No matter which method is used, the key idea is to interpret large-scale atmospheric dynamical processes as information flow around the globe and to identify the pathways of this information flow (physical interactions) using causal discovery.

Figure 6 shows a sample network plot obtained for geopotential height at 500mb for boreal winter months (Dec-Feb) of years 1950-2000 (using daily NCEP-NCAR reanalysis data [24], [25]) using *PC stable* with $D = 1$ day between time slices and significance level $\alpha = 0.1$. Fig. 6(a) shows the strongest direct connections that take significantly less than

1 day to travel from cause to effect, while Fig. 6(b) shows the strongest direct connections taking about 1 day. These graphs were obtained by first generating temporal graphs (similar to that in Fig. 4 for Application 1), then converting them to summary graphs (similar to Fig. 5 for Application 1) summarizing the strongest connections. It turns out that the interactions captured by this particular graph are *storm tracks*, and the pathways shown in this graph cannot be obtained by traditional climate science methods. Thus these plots reveal new information not available through existing methods. For detailed interpretation of this type of graph, see [21].

In general, which physical processes are tracked in the network depends primarily on two factors: (1) the atmospheric field used and (2) the time scale (e.g. daily data vs. monthly data). Thus using a variety of different atmospheric fields we can track the causal pathways of a variety of different dynamical processes around the globe using this approach.

V. Issues Related to Boundaries

As discussed in Section II-B2 there are incidences where we cannot avoid violating causal sufficiency, have to accept that fact and deal with the consequences in the evaluation step. However, there are also many instances where one may happen to violate causal sufficiency that *can* be avoided. The two primary situations we came across are described below. They both have to do with introducing artificial boundaries - either in time or in space - in the modeling process. Those boundaries cut nodes off from their neighbors in the system - neighbors either in terms of time or space.

A. The Initialization Problem in Temporal Analysis

When using a temporal model, the model usually needs a few time slices to converge to a proper independence model. The reason is an initialization problem; namely, to determine the *causal flow originating in a time slice*, it is crucial to have information on the *causal flow into that time slice*. Since the first few time slices are lacking that information - because no prior time slices are included - they often yield erroneous links. This can be interpreted as a causal sufficiency problem: for the nodes in the first few time slices the common causes in any prior slices are not included, thus violating the causal sufficiency condition to an extend that renders the first few slices useless.

This problem is easily solved by developing the model for more slices than needed and then deleting the first few slices in the results. How many slices should be deleted is usually obvious from the resulting graph because the first (erroneous) slices usually differ significantly from the stable pattern emerging in the later slices. In theory, the number of slices to be deleted depends on the maximal duration of the major connections in the network, i.e. if *direct* connections extend over up to P time slices, then up to the first P time slices *may* be compromised. In practice, the number of slices to be deleted is often lower than P. For example, in Application 1 we had to delete only the first three slices (representing 9 days) to obtain the graph in Fig. 4, and for Application 2 we deleted just the first slice (representing 1 day), although some connections last several days.

Although time is cut off at the beginning *and* the end of the model, we only need to consider the *first* time slices for deletion, never the *last* ones. The reason is that we need to care only about having neglected *incoming* connections (hidden common causes) and causes can only occur *before* effects.

B. Dealing with Spatial Boundaries

When modeling a real-word system one sometimes may want to focus only on a spatial subset of the whole system, e.g. a geographic region such as North America. When doing so one must be aware that the same effect that is happening in the temporal domain also applies in the spatial domain. For example if using only grid points in a selected region, which is a subset of a much larger region, the grid points on or near the boundaries of the selected region are cut off from their interactions outside of it, thus potentially violating causal sufficiency.

In determining which slices on the boundaries have indeed to be deleted it is important to look at the *direction* of causal interactions for the system under consideration. Just like we only had to worry about the *first* time slices potentially being compromised in the temporal initialization problem (Section V-A), a similar effect exists in the spatial domain. For example, if we only want to look at a rectangular region of the earth, and we know that for the variables under consideration we know that causal connections occur generally from SE to NW direction, then we only need to consider deleting slices from and near the Southern and the Eastern boundaries.

Taking this effect into account is particularly helpful in climate science applications where the prevalent direction of atmospheric flow greatly reduces which boundaries may be compromised. In Application 2 the grid spans the whole globe, so we did not have to worry about cutting off neighbors, but we recently extended that model to the third dimension, using data from several geopotential height layers and therefore using a 3D grid [26]. Since we can only include a finite number of height layers, we have to consider that the highest and lowest layers may be compromised. The results often reveal primary directions of cause-effect flow which indicate that only certain layers may be compromised (either the highest *or* the lowest) - or sometimes none at all, if the primary connections near the boundary planes are mainly horizontal.

VI. Issues related to spacing and signal strength

Now we focus on two types of issues that can occur anywhere, not just at the boundaries, and that have nothing to do with causal sufficiency, but instead with signal strength and the fact that causal discovery only picks up the very strongest connections for each node.

A. Effect of Temporal Resolution

Choosing the distance, D, between time slices determines the temporal resolution of the model. If temporal resolution is very high, it can happen that the autocorrelation relationships,

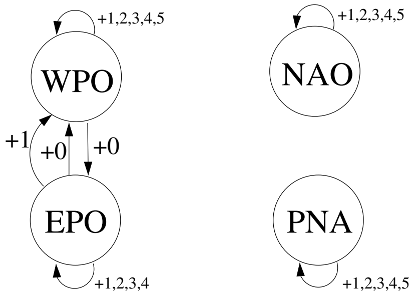


Fig. 7. Summary graph for Application 1 for $D = 1$ day between slices

i.e. connections from a variable in one time step to the same variable in later time steps, dominate the network. The reason is that the causal discovery algorithm for each node only identifies the very strongest connections - and only a limited number of those. Thus if each variable state has a strong influence on the state of the same variable for several time slices *and* the distance D between time slices is chosen very small, then the auto correlation relationships sometimes are the only ones to show in the graph. The problem can usually be solved by choosing a larger value D between time slices, thus reducing how many auto-correlation connections can make it into the strongest connections. Choosing D very large on the other hand causes the model to miss connections that have short duration from cause to effect (shorter than D), so there is usually a trade-off involved when choosing D.

In Application 2 choosing $D = 1$ day worked fine (Fig. 6), since the strongest interactions all happen within 1 day anyway. However, in Application 1, if we increase temporal resolution so that each time slice represents 1 day ($D = 1$), we obtain the summary graph shown in Fig. 7, which is rather boring. While it provides more detailed information about the duration of self-loops (1-5 days for most indices), it does not tell us much about the connections *between* the different variables. Choosing $D = 3$ (Fig. 5) yields much more interesting results. Both models, Fig. 5 and Fig. 7, are correct in the sense that they *show the strongest connections between the set of variables included in the model.* Nevertheless, Fig. 5 is more useful to answer questions about connections between *different* compound indices than Fig. 7. The lesson to learn from this is that to make sure the model answers the right question it is crucial to carefully choose which nodes to include in the model - including temporal resolution.

B. Effect of Irregular Spacing of Grid Points

In this section we consider the impact that irregular grid spacing can have on the resulting graphs. To demonstrate the concepts we use the example of discretizing the globe, but the conclusions drawn apply to other applications as well, e.g. wherever data is available only on irregularly spaced locations.

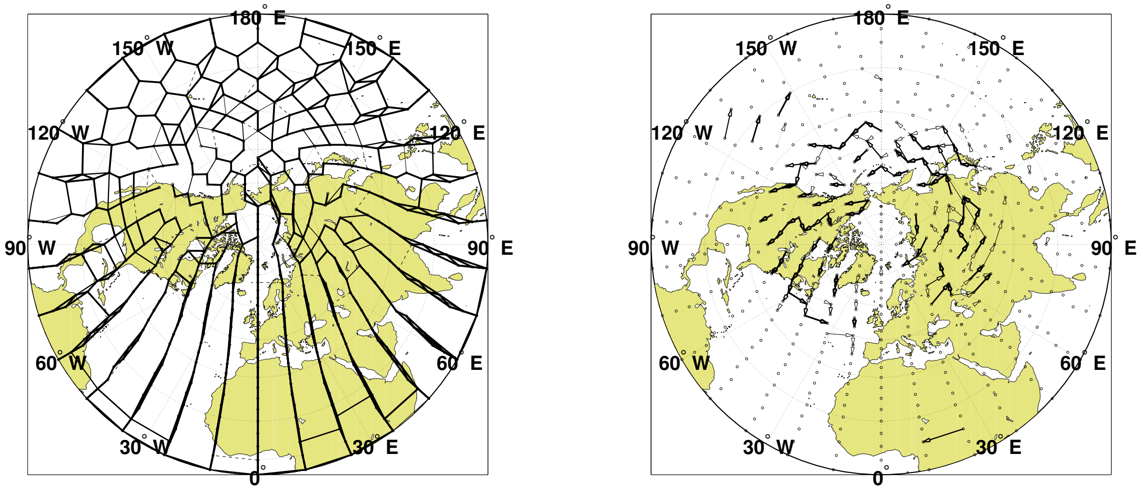
Discretizing the globe - or unit sphere - is particularly tricky, because most parametrizations have a singularity near the poles. Climate scientists thus use a great variety of different grids, depending on the purpose of the model. The simplest - and probably most common - type of grid splits the range of longitude ($[0^o, 360^o]$) and latitude ($[-90^o, 90^o]$) angles into equal increments. This type of discretization works well near the equator, but creates a very dense grid near the poles, which often creates irregularities, e.g. in numerical calculations.

A common solution is to use some type of *equal area grid*, i.e. a grid where the area belonging to each grid cell is about the same. A very simple implementation of an equal area grid, used for demonstrative purposes here, is to use circles of latitude (circles parallel to the equator) at equal distances, and then to place a varying number of grid points on each circle[1]. This type of grid is used in Fig. 8, where the grid points are shown as small black circles in Fig. 8(b). This grid results in a large number of points on circles near the equator and decreasing numbers on circles closer to the poles. While equal area grids guarantee that the *area* of each grid cell is (nearly) identical, the shape of the cells can still differ - e.g. where cells are nearly rectangular and become squished near the poles - which means that the *distance* between neighboring points can vary greatly. Furthermore, in this simple implementation, since the number of grid points on each circle must be an integer, we have to round the number of points and thus create additional geometric irregularities. A strong irregularity in Fig. 8(b) is near the equator, where grid points over Africa (0^o longitude) line up along straight lines toward the pole, while grid points at the opposite site of the globe (180^o longitude) form hexagonal groups instead.

Geodesic grids achieve nearly perfect regularity of cell geometry and area and are sometimes used in climate science. However, geodesic grids only come in few resolutions, namely the number of cells can only be one of the following: 42, 162, 642, 2562, 10242, etc., which limits their use. There is *no* closed form solution to determine a grid on a sphere with any desired number of points and where neighboring grid points have equal distance. However, there are excellent *approximations*, namely Bendito et al. [28] developed an algorithm to calculate any number of *Fekete points*, which are nearly equally spaced around the globe. This method yields grids that are extremely regular, where most cells are nearly regular hexagons (just like in a geodesic grid), and just a few are nearly regular pentagons.

1) Sample results for different grids: Figures 6 and 8 show results obtained using the same data (see Section IV), but Figure 6 uses a Fekete grid with 800 points and Figure 8 uses the simple implementation of the equal area grid with 918 points. The results in Figure 6 are physically meaningful, since they show information flow in form of storm tracks (Section IV). However, what happened in Figure 8? Fig. 8(a) should look similar to Fig. 6(a) and Fig. 8(b) to Fig. 6(b), but clearly they do not. The most striking effect can be observed in 8(a) near the equator, where it seems as if information flow occurs in straight lines over Africa (0^o longitude), but

[1] Climate scientists often use more sophisticated equal area grids, see for example Leopardi [27] which generates much more regular cells. However, the simple implementation is perfect to demonstrate the problems that can arise from irregular grid spacing.



(a) Travel $<$ 1 day

(b) Travel $\approx$ 1 day

Fig. 8. Network plots from *PC stable* ($D = 1$ day, $\alpha = 0.1$) based on simple implementation of equal-area grid (918 grid points) yield strange results.

in hexagonal patterns at the opposite site of the globe (180^o longitude). Obviously, physical processes are unlikely to occur in such patterns, and indeed this pattern matches exactly the irregularities of the equal-area grid discussed above. In fact this is purely an artifact due to grid irregularities, namely any two grid points that are unusually close to each other in this grid are connected in the resulting network!

2) Interpretation of Results: These results raise the following question - why does the proximity pattern in the irregular grid have such an *overwhelming* effect on the network connections? The reason is as follows. Using causal discovery we seek to identify the very strongest connections between grid points. Signal strength of most atmospheric processes decays quickly with increasing distance, so that points that are unusually close to each other are more likely to appear strongly connected. Thus the unequal proximity in this grid creates a *stronger* connection between those grid points than the actual causal pathways we wanted to identify. We believe that this behavior is likely to occur for many physical systems, not just those related to our planet's atmosphere.

One may think that this issue could be solved by increasing the grid resolution. However, unless the spacing is more regular, the problem persists even if grid resolution is increased, since any point pairs that are unusually close always dominate, regardless of scale.

What exactly prevents these effects from happening in the Fekete grid (Figure 6)? Every Fekete point has six (sometimes five) nearly equally distant points to choose from as closest neighbors, so there is *no bias in any of those directions*, because they are all at the same distance. Thus choosing a *regular polygon* as cell structure is the key to avoiding artifacts, by making the grid as **isotropic** as possible.

Just like the models from varying temporal resolution in Section VI-A, the results in Figure 8 are actually *correct in terms of identifying the strongest connections between the nodes included in the model*. So, again, the lesson to learn is that in order to answer the question we are interested in, we need to carefully select the nodes to include in the model. Namely, in order to identify the *pathways* of global information flow, we need to make sure to remove any bias of direction between neighboring points, i.e. use isotropic grids.

VII. Discussion and Conclusions

In this paper we have identified a number of issues that are unique to causal discovery with spatio-temporal data and established the following connections:

1) The initialization problem and the spatial boundary problem are both effects with the same origin, violation of causal sufficiency, and can both be addressed in the same way. Namely, one has to develop the model for more time slices/boundary layers than required, then check causal flow directions to identify which of them may have been compromised and should be deleted.
2) The effect of temporal resolution and irregular spacing of grid points both have the same origin, namely the fact that we can only pick up the few strongest connections for each node and that causal connectivity is a relative concept that changes with spatial/temporal distance.
   Specifically, if we choose D very small autocorrelations may dominate, as seen in Application 1, and hide the relationships between different variable types (here compound indices). Similarly, if a grid with irregular spacing is chosen, then proximity in the grid becomes the dominating signal and hides the actual spatial pathways of causal interaction.

The issues described as Item 2 above (and in Section VI) are *not* due to any errors in the method of causal discovery - in fact the models were all correct given the way the problem was set up. Instead, the issues arose from the fact that (due to unequal spacing or for certain temporal resolution) the problem set-up does not actually address the question we were interested in - namely the *pathways* around the globe or the interaction between *different* compound indices. This implies that the pitfalls in Section VI are inherent to the problem set-up, not the method used, and are thus likely to occur - at least to some extent - also when other methods are used for causal discovery, such as Granger graphical models or Gaussian graphical models.

There are many potential applications of causal discovery in spatio-temporal systems and many of those also have to deal with irregular grids. For example, the locations of sensor networks are often irregular, as are the locations of neurons in the brain. Thus future work should look for ways to compensate for irregular grids, if the irregularities cannot be avoided. In constraint-based learning, for example, one could raise the threshold for the statistical tests for certain variable pairs - e.g. those pairs representing locations unusually close to each other or variable pairs representing autocorrelation - thus effectively providing a handicap for those pairs and thus reducing their dominance in the results. However, that is a tricky task, since it is hard to come up with schemes to adjust the threshold for variable pairs that do not introduce any new artifacts.

In conclusion, causal discovery for temporal models is a powerful framework with many applications. In this paper we have highlighted a number of pitfalls that must be considered when developing such models from temporal data and many kinks are still to be worked out. Nevertheless, it is exciting to see the great potential of these approaches for speeding up the process of discovery in areas ranging from earth science to bioinformatics.

Acknowledgment

This work was supported in part by NSF Climate and Large-Scale Dynamics (CLD) program Grant AGS-1147601 awarded to Yi Deng.

References

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, 2nd ed., Morgan Kaufman Publishers, 552 pp, 1988.

[2] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, Springer Lecture Notes in Statistics. 1st ed. Springer Verlag, 526 pp., 1993.

[3] R.E. Neapolitan, *Learning Bayesian Networks*, Prentice Hall, 647 pp., 2003.

[4] B. Shipley, *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*, 1st ed. Cambridge University Press, 332 p., 2002.

[5] C.J. Needham, J.R. Bradford, A.J. Bulpitt, D.R. Westhead, *A Primer on Learning in Bayesian Networks for Computational Biology*, PLoS Comput Biol 3(8): e129. doi:10.1371/journal.pcbi.0030129, 2007.

[6] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, *Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics, 7 (Suppl.), S7, doi:10.1186/1471-2105-7-S1-S7, 2006.

[7] N. Friedman, M. Linial, I. Nachman and D. Peer, *Using Bayesian networks to analyze expression data*, J. Comput. Biol., 7 (34), 601620, 2000.

[8] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger and G.P. Nolan, *Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data*, Science 22, Vol. 308 no. 5721 pp. 523-529 DOI: 10.1126/science.1105809, April 2005.

[9] X. Chen1, M.M. Hoffman, J.A. Bilmes, J.R. Hesselberth and W.S. Noble, *A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data*, Bioinformatics, Vol. 26 ISMB 2010, pages i334-i342, doi:10.1093/bioinformatics/btq175, 2010.

[10] S.E.-d. El-dawlatly, *Graph-based methods for inferring neuronal connectivity from spike train ensembles*, Ph.D. thesis, Electrical Engineering, Michigan State University, 2011. Available at http://etd.lib.msu.edu/islandora/object/etd%3A357/datastream/OBJ/view

[11] T. Chu, D. Danks, and C. Glymour, *Data Driven Methods for Nonlinear Granger Causality: Climate Teleconnection Mechanisms*, Tech. Rep. CMU-PHIL-171, Dep. of Philos., Carnegie Mellon Univ., Pittsburgh, Pa., 2005.

[12] M. Cossention, F. Raimondi, and M. Vitale, *Bayesian models of the pm 10 atmospheric urban pollution.* Proc. Ninth Int. Conf. on Modeling, Monitoring and Management of Air Pollution: Air Pollution IX, Ancona, Italy, Wessex, 143152, 2001.

[13] R. Cano, C. Sordo, and J. Gutierrez, *Applications of bayesian networks in meteorology. Advances in Bayesian Networks*, J. A. Gamez et al., Eds., Springer, 309327, 2004.

[14] R.J. Kennett, K.B. Korb, and A.E. Nicholson, *Seabreeze prediction using Bayesian networks*, Proc. Fifth Pacific-Asia Conference on Knowledge Discovery and Data Minung (PAKDD01), Hong Kong, China, PAKDD, 148153, 2001.

[15] Y. Deng and I. Ebert-Uphoff, *Weakening of Atmospheric Information Flow in a Warming Climate in the Community Climate System Model*, Geophysical Research Letters, 7 pages, doi: 10.1002/2013GL058646, Jan 2014.

[16] D. Koller, and N. Friedman, *Probabilistic Graphical Models - Principles and Techniques*, 1st ed. MIT Press, 1280 pp., 2009.

[17] P. Spirtes and C. Glymour, *An algorithm for fast recovery of sparse causal graphs*, Social Science Computer Review, 9(1):6772, 1991.

[18] D. Colombo and M.H. Maathuis, *Order-independent constraint-based causal structure learning*, (arXiv:1211.3295v2), 2013.

[19] D. Colombo, M.H. Maathuis, M. Kalisch, and T.S. Richardson, *Learning high-dimensional directed acyclic graphs with latent and selection variables*, Ann. Stat., 40, 294321, 2012.

[20] I. Ebert-Uphoff and Y. Deng, *Causal Discovery for Climate Research Using Graphical Models*, Journal of Climate, Vol. 25, No. 17, doi:10.1175/JCLI-D-11-00387.1, pp. 5648-5665, Sept 2012.

[21] I. Ebert-Uphoff and Y. Deng, *A New Type of Climate Network based on Probabilistic Graphical Models: Results of Boreal Winter versus Summer*, Geophysical Research Letters, vol. 39, L19701, 7 pages, doi:10.1029/2012GL053269, 2012.

[22] T. Zerenner, P. Friedrichs, K. Lehnerts and A. Hense, *A Gaussian graphical model approach to climate networks*, Chaos, 23, 023103, 2014.

[23] J. Runge, *Detecting and Quantifying Causality from Time Series of Complex Systems*, Ph.D. thesis, Humboldt-University Berlin, Germany, Aug. 2014. Available at http://edoc.hu-berlin.de/dissertationen/runge-jakob-2014-08-05/PDF/runge.pdf.

[24] E. Kalnay et al., *The NCEP / NCAR 40-year reanalysis project*, Bull. Am. Meteorol. Soc., 77, 437471, doi:10.1175/1520-0477 (1996)077¡0437:TNYRP¿2.0.CO;2, 1996.

[25] R. Kistler et al., *The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation*, Bull. Am. Meteorol. Soc., 82, 247267, doi:10.1175/1520-0477(2001)082¡0247:TNNYRM¿2.3.CO;2, 2001.

[26] I. Ebert-Uphoff and Y. Deng, *High efficiency implementation of PC and PC stable algorithms yields three-dimensional graphs of information flow for the earth' atmosphere*, Research Report Nr. CSU-ECE-2014-1, Dept. of Electrical and Computer Engineering, Colorado State University, Sept. 3, 2014. Available at http://hdl.handle.net/10217/83709

[27] P. Leopardi, *A Partition of the Unit Sphere into regions of equal area and small diameter*, Electronic Transactions on Numerical Analysis, Volume 25, pp. 309-327, 2006.

[28] E. Bendito, A. Carmona, A.M. Encinas, and J.M. Gesto, *Estimation of Fekete points*, J. Comput. Phys., 225, 23542376, doi:10.1016/j.jcp.2007.03.017, 2007.