CI 2018
8th International Workshop on Climate Informatics
September 19-21, 2018
Hosted by the National Center for Atmospheric Research in Boulder, CO

TETRAD - A TOOLBOX FOR CAUSAL DISCOVERY

Joseph D. Ramsey[1], Kun Zhang[1], Madelyn Glymour[1], Ruben Sanchez Romero[1], Biwei Huang[1], Imme Ebert-Uphoff[2], Savini Samarasinghe[2], Elizabeth A. Barnes[3], Clark Glymour[1]

***Abstract*—Climate and Earth research seeks to identify causal relationships without the advantage of experimental controls. Algorithmic methods for that purpose have a long history and have developed rapidly in the last quarter century, so that new methods appear almost monthly. Unsurprisingly, these products vary in accuracy, informativeness, quality of implementation, and necessary assumptions. Researchers need a guide and implementation of well-vetted causal search methods and a means to test and compare methods on real and simulated data. We review the TETRAD suite of programs for these purposes, available from the Pittsburgh/Carnegie Mellon Center for Causal (http://www.phil.cmu.edu/tetrad/).**

I. MOTIVATION

Causal inference without experimental control is indispensable in climate science and has a distinguished pedigree in other sciences: Newtonian gravitational theory and Darwinian evolution were established almost without experiment. Early in the 20th century, Gilbert Walker [1] demonstrated the practical possibility of establishing causal connections from correlations of climate phenomena. But even earlier, George Udny Yule [2], [3] described some of the foibles of causal inference from time series. The ambition to develop large-scale causal modeling of climate faces a number of difficulties. For example, the processes of interest are often undersampled; there may be unmeasured variables that create associations among measured variables; time series may be non-stationary and non-linearities may exist; correlation is insufficient, etc. Similar problems arise in neuropsychological measurements (e.g., fMRI), in economics, and elsewhere.

Statistical and computational research since the 1990s has considerably expanded the scope of circumstances in which causal inferences based on background knowledge and observational data are feasible. The result is a plethora of competing methods claiming to extract causal information from equilibrium data or time series, and applications, with many papers attempting to infer dynamic relations of climate indices, or local and global fields of temperature, pressure, precipitation and other variables, from geospatial time series measurements.

An investigator wishing to apply a search method to identify causal climate processes is faced with a very large number of alternative procedures distributed over many websites and programming languages, only some of which have proofs – largely confined to technical papers – of large sample (asymptotic) correctness and many of which do not. These many methods may give very different results on finite samples.

Many of these methods have relied on inferences to conditional independence and dependence by adaptation of correlation and partial correlation to time series. Granger's method [4], for example, tests for partial regression or partial correlation of Y_t on X_{t-n}, controlling for Y_{t-n}, Z_{t-n}, where Z variables are covariates and n indexes all lags up to some number, n. Elaborations of Granger's method have attempted to identify ”contemporaneous” causal connections (i.e., those occurring faster than the sampling rate). Subsequent ”Bayes net search” methods [5], [6], [7], [8], [9], [10], [11] allow the possibility of latent confounding and sample selection bias. Other recent methods have used penalized regression, such as LASSO [12]. Procedures can rely on explicit hypothesis tests of vanishing partial correlations or via methods that score models implicitly on conditional independence properties evidenced in the data [13], [14].

Still more recently, flexible methods have used non-Gaussianity of the signals to infer causal connections from time series [15], [16], [17], [9]. These procedures can give causal information when conditional independence methods based only on second moments of distributions are uninformative. Procedures are available for

Corresponding author: C. Glymour, cg09@andrew.cmu.edu. [1]Philosophy, Carnegie Mellon University, Pittsburgh, PA, USA. [2]Electrical & Computer Engineering, Colorado State University, Fort Collins, CO, USA. [3]Atmospheric Sciences, Colorado State University, Fort Collins, CO, USA.

identifying changes in distributions over time, the number of component distributions in mixed distributions, and sorting cases among component distributions.

Among these many model search methods are those that do and those that do not tolerate unmeasured common causes of measured variables (”latent” variables); non-stationarity; non-linearities or multiple data types (continuous and discrete). There are dozens of proposed algorithms, and for some of them there are multiple implementations of differing quality [18]. Some, but not all, of these methods have large sample correctness proofs, but analytic finite sample error probabilities are in general not possible. Accuracies of search methods must therefore be assessed by the agreement of their data driven results with known relationships. Published rigorous comparisons of search algorithms are few.

Altogether, a climate researcher has good reason to be perplexed as to which methods to use for a problem. Perhaps the best way to choose a model search procedure is to test multiple methods on sample data with known cause-effect relations, either with empirical data or simulated climate data. That would require considerable literature searching, programming and result analysis. What is needed is a facility that collects the best available algorithms, allows users to add algorithms as desired, to simulate time series or equilibrium data from user specified models, and to compare accuracies of multiple algorithms on any appropriately formatted data set for which a true model is assumed. To that end we review the TETRAD suite hosted by the Pittsburgh/Carnegie Mellon Center for Causal Discovery.

II. METHOD

The TETRAD program, developed over 20 years, is a drag and drop suite of procedures for analyzing data for causal relations that allows: uploading raw continuous or covariance or correlation or categorical data, or datasets having both continuous and categorical variables, including time series data; searching for structural relations with more than a dozen well-tested algorithms, including some for time series; specifying prior knowledge to constrain the searches; manipulating data by imputing missing values, logging, discretizing, merging, etc.; creating a statistical model step by step from graph to parametric family to values of the model parameters, including a wide range of linking functions and a variety of probability distributions, and allowing unmeasured confounders; simulating data from a statistical model; estimating parameters from real or simulated data; predicting the effects on other variables of interventions or perturbations on one or more variables; and computing the probability distribution of any variable conditional on specified values of any other set of variables. Some routine descriptive statistics are available, for example histograms, normality tests and correlations or covariances. TETRAD has already been used extensively for causal discovery in climate science, e.g., the results in [19], [20], [21], [22], [23] were obtained using TETRAD.

TETRAD is a general causal discovery software which so far does not contain functionality specific to the earth sciences. Thus, the typical overall workflow for the use of TETRAD is to 1) pre-process data outside of TETRAD (e.g., detrending, spatial aggregation, season selection); 2) read the preprocessed data into TETRAD using its GUI; 3) perform additional data manipulation and causal model searches in TETRAD; 4) get a first graphical representation in TETRAD that is extremely useful for trouble shooting; 5) export either the graph image or the graph edges as a text file from TETRAD; and 6) perform post-processing outside of TETRAD, e.g., to visualize connections with underlying geographic information.

Once inside TETRAD, the design of the program produces a flow chart between boxes that the user inserts, see Fig. 1 for an example. Each box contains a list of pertinent functions from which the user can select. Information produced in an ancestor box flows to its descendant boxes.

Fig. 1 shows a session from a recent exploratory analysis of the interactions between midlatitude jet-stream speed (N1), Arctic temperature (N2) and jet-stream latitude (N3), based on observed time series of $N1$, $N2$ and $N3$. (See [24] for a concise definition of the variables.) **The session in Fig. 1 is included here solely for demonstration purposes, not to draw scientific conclusions for this application.** To be able to capture time-delayed relationships, we use the original variables and 10 lagged copies, denoted by their start times times, $T0$ to $T10$, and resulting in a total of 33 variables in the model.

The *data box*, shown in yellow in Fig. 1, contains the data, which are read from a plain text input file. The *knowledge box*, also in yellow and below the *data box*, establishes the temporal order of the lagged variables, and enforces that effect can never occur before their causes in the models. (The knowledge box can also be used to add any other prior knowledge.) With the data read and the temporal order established, we can then feed these two boxes directly into *causal model search boxes*, which represent different search algorithms, such

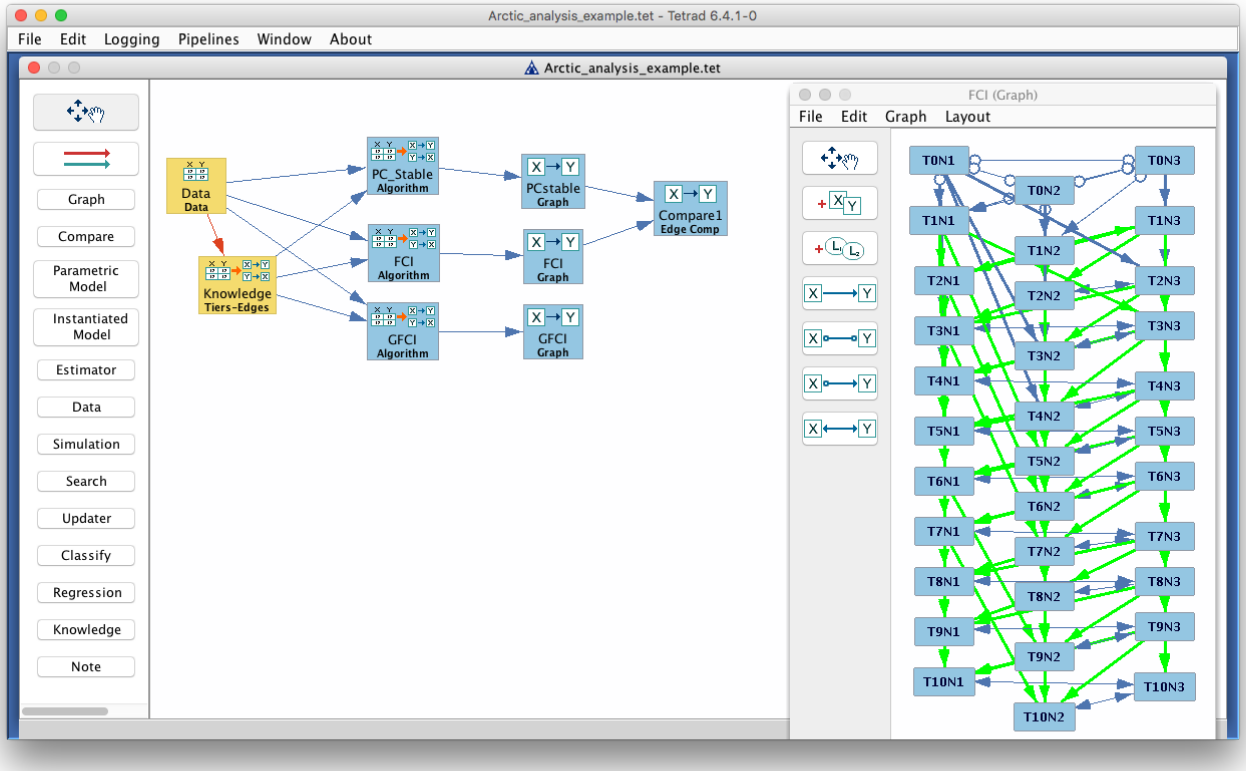

Fig. 1. A sample TETRAD session analyzing the relationships between three time series variables using different model search algorithms. In the graph on the right a green arrow indicates a direct connection; a blue arrow indicates that the connection may be direct or through some other variables; a double-headed arrow indicates that the association is confounded by one or more unmeasured common causes. The first few times slices (e.g., T0,T1) are often discarded, because they may contain erroneous edges due to convergence issues [20].

as *PC stable*, *FCI* and *GFCI*. (*PC stable* performs causal discovery assuming that there are no hidden common causes, while *FCI* and *GFCI* search for causal structures that include latent variables.) Following the workflow further to the right, each *causal model search box* is followed by a *graph box*, which automatically generates a graphical representation for the result of each model search. Double clicking on any box opens it, to view and modify its content, e.g., parameters for a model search or resulting graphs. The result of double clicking on the *FCI graph box* is shown on the right in Fig. 1, which shows the graph structure obtained by *FCI* for this case.

As demonstrated here, the interface allows branching at any point, so multiple models and data sets can be generated or analyzed in a single session, and the user has a visual trace of the steps taken. Sessions can be saved at any stage and shared with others. This fact, along with the graphical nature of the toolbox, makes it very easy for collaborators to follow all analysis steps and to modify or add to these steps, making this an **excellent tool for collaborations and for promoting reproducibility of any analysis**.

Some of the search algorithms are very fast and can be run on very high dimensional datasets with small or large sample sizes. The *FGES* algorithm, for example, has been run (with very accurate results) on data from a sparse structure with one million variables, requiring about 12 hours on the Pittsburgh Supercomputer [25]. Other procedures (for example, some searches with non-Gaussian distributions) are much slower and limited in practice to smaller sample sizes (e.g, not much more than a thousand) or numbers of variables (again, not much more than a thousand). *FASK* is a recent non-Gaussian procedure which is very fast and robust against measurement error.

The TETRAD program is continuously being updated with new, well-tested algorithms, additional facilities and, of course, bug corrections. Help by email is available, as well as an online manual. Video tutorials are in production. Issues, problems and bugs can be posted on the GitHub issues link for TETRAD. Investigators who wish to introduce new algorithms into their own GitHub branch of the software and who need guidance should email Joseph Ramsey, jdramsey@andrew.cmu.edu.

Current work aims to allow flexible comparison of new algorithms and easy introduction of new simulations or real data into the comparisons. Procedures for further automating algorithm comparisons on real or simulated data are being built. Ramsey and Malinsky have designed a procedure for comparing simulation results, and an accompanying interface has been implemented (http://www.pitt.edu/~bja43/causal/#!/load). The facility has been used to study a range of time series methods, finding that several vector autoregression methods perform poorly when processes are undersampled, that non-Gaussian methods work best when there are even small non-Gaussian components to the distributions and are more robust to measurement errors.

All of the TETRAD software is open source JAVA code and is available on GitHub. It can be modified or tailored to the user's wishes and code improvements and suggestions shared. TETRAD is available directly from the GitHub repository, or from the Tools on the Center for Causal Discovery website, and can be called from R or from within Python. The only other software required is up-to-date Java.

III. Conclusions

The TETRAD suite is usable now for various climate study problems, but it is not complete, and with continuing algorithmic developments, is inevitably a work in progress. **The authors welcome and solicit guidance from climate scientists concerning algorithmic needs, interface requirements, data sets and other relevant issues.** Please send your feedback to Clark Glymour (cg09@andrew.cmu.edu) and Joseph Ramsey (jdramsey@andrew.cmu.edu).

Acknowledgments

Research reported in this publication was supported by grants: Integrative Graphical Models for Large Multi-Modal Biomedical Data. 5RO1LM012087; Effective Connectivity in Brain Networks: Discovering Latent Structure, Network Complexity and Recurrence. 1RO1EB022858-01; U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative; and NIH Prime Award N5R01LM12087-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] G. Walker, "World weather," *Quarterly Journal of the Royal Meteorological Society*, vol. 54, no. 226, pp. 79–87, 1928.

[2] G. U. Yule, "Notes on the theory of association of attributes in statistics," *Biometrika*, vol. 2, no. 2, pp. 121–134, 1903.

[3] G. U. Yule, "Selected papers of George Udny Yule," *Griffin, High Wycombe*, 1971. A. Stuart and M.G. Kendall (eds.).

[4] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[5] R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson, "The TETRAD project: Constraint based aids to causal model specification," *Multivariate Behavioral Research*, vol. 33, no. 1, pp. 65–117, 1998.

[6] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social science computer review*, vol. 9, no. 1, pp. 62–72, 1991.

[7] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search.* MIT Press, Cambridge, 2001.

[8] D. Colombo and M. H. Maathuis, "Order-independent constraint-based causal structure learning," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3741–3782, 2014.

[9] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery," in *27th Conference on Uncertainty in Artificial Intelli1gence (UAI 2011)*, pp. 804–813, 2011.

[10] M. Eichler, "Graphical modelling of multivariate time series," *Probability Theory and Related Fields*, vol. 153, no. 1-2, pp. 233–268, 2012.

[11] M. Eichler, "Causal inference with multiple time series: principles and problems," *Phil. Trans. R. Soc. A*, vol. 371, no. 1997, p. 20110613, 2013.

[12] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[13] Y. Hirata, J. M. Amigó, Y. Matsuzaka, R. Yokota, H. Mushiake, and K. Aihara, "Detecting causality by combined use of multiple methods: Climate and brain examples," *PloS one*, vol. 11, no. 7, p. e0158572, 2016.

[14] J. Runge, V. Petoukhov, and J. Kurths, "Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models," *Journal of Climate*, vol. 27, no. 2, pp. 720–739, 2014.

[15] P. Geiger, K. Zhang, M. Gong, D. Janzing, and B. Schölkopf, "Causal inference by identification of vector autoregressive processes with hidden components," *arXiv preprint arXiv:1411.3972*, 2014.

[16] M. Gong, K. Zhang, B. Schoelkopf, D. Tao, and P. Geiger, "Discovering temporal causal relations from subsampled data," in *International Conference on Machine Learning*, pp. 1898–1906, 2015.

[17] K. Zhang and A. Hyvärinen, "On the identifiability of the post-nonlinear causal model," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pp. 647–655, AUAI Press, 2009.

[18] J. D. Ramsey and D. Malinsky, "Comparing the performance of graphical structure learning algorithms with TETRAD," *arXiv preprint arXiv:1607.08110*, 2016.

[19] T. Chu, D. Danks, and C. Glymour, "Data driven methods for nonlinear Granger causality: Climate teleconnection mechanisms," tech. rep., Carnegie Mellon University, Department of Philosophy, 2005. CMU-PHIL-171.

[20] I. Ebert-Uphoff and Y. Deng, "Causal discovery for climate research using graphical models," *Journal of Climate*, vol. 25, no. 17, pp. 5648–5665, 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, no. 19, 2012.

[22] Y. Deng and I. Ebert-Uphoff, "Weakening of atmospheric information flow in a warming climate in the community climate system model," *Geophysical Research Letters*, vol. 41, no. 1, pp. 193–200, 2014.

[23] D. Niyogi, C. Kishtawal, S. Tripathi, and R. S. Govindaraju, "Observational evidence that agricultural intensification and land use change may be reducing the Indian summer monsoon rainfall," *Water Resources Research*, vol. 46, no. 3, 2010.

[24] S. Samarasinghe, M. C. McGraw, E. A. Barnes, and I. Ebert-Uphoff, "A study of links between the Arctic and the midlatitude jet-stream using Granger and Pearl causality," *Environmetrics*, 2018 (in press).

[25] J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour, "A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images," *International journal of data science and analytics*, vol. 3, no. 2, pp. 121–129, 2017.