Journal Pre-proofs

Perspective

Temporally or spatially? Causation inference in earth system sciences

Bingbo Gao, Manchun Li, Jinfeng Wang, Ziyue Chen

PII: S2095-9273(21)00642-3
DOI: https://doi.org/10.1016/j.scib.2021.10.002
Reference: SCIB 1591

To appear in: *Science Bulletin*

Please cite this article as: B. Gao, M. Li, J. Wang, Z. Chen, Temporally or spatially? Causation inference in earth system sciences, *Science Bulletin* (2021), doi: https://doi.org/10.1016/j.scib.2021.10.002

Perspective

Temporally or spatially? Causation inference in Earth System Sciences

Bingbo Gao[a], Manchun Li[b], Jinfeng Wang[c], Ziyue Chen[d*]

[a] College of Land Science and Technology, China Agricultural University, Beijing 100083, China;

[b] School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China;

[c] LREIS, Institute of Geographic Science & Nature Resources Research, Chinese Academy of Sciences, Beijing 100101, China;

[d] College of Global and Earth System Sciences, Beijing Normal University, Beijing 100875, China.

*Corresponding author: zychen@bnu.edu.cn (Chen Z.-Y.).

The discovery of cause–effect relationships helps to understand the natural or physical mechanism[1]. Causation inference is a key issue in many disciplines and has a long study history, especially in statistics, social, and biomedical sciences[2]. In Earth System Sciences, the cause–effect relationship also plays a fundamental role and has drawn increasing interests. However, for spatially large-scale research, it is not feasible to design and conduct controlled experiments to reveal the cause–effect relationships. Therefore, causation inference from time series data has been frequently employed, under the assumption that the cause precedes the effect [3]. While the temporal inference works effectively to identify most causation between variables, limitations remain. If the time series is not long enough to catch significant changes of causes and effects, some important cause–effect relationships may be neglected. This limitation is highlighted in Earth System Sciences, as the evolution of global changes may take an extreme long period to present discernible variations. For instance, the annually mean temperature in one area demonstrates very limited variations in decades, which is already a long period for Earth observation. On one hand, the causation between temperature and plant growth in this area can hardly be identified using temporal causation models, which mainly detect causation between two variables by examining the successive (or simultaneous) variations of one variable

induced by the variation of the other. On the other hand, the causal influence of temperature on plant growth has been well accepted [4,5] . So clearly, the temporal causality models are not a panacea to all causation inference scenarios.

Given the research objects in Earth System Sciences, characterized with large-scale spatial distribution and the usual lack of complete time-series data, causation inference may be conducted from an alternative perspective to fully utilize the spatial differences. Specifically, while the variation of one variable may not be detected temporally, the wide distribution of this variable makes its variations easily recognized spatially. The general principle of causation inference from time series data is based on temporal change-response mechanisms. Likewise, spatial variations (the change of variables across spatial locations) and corresponding responses may also be employed for causation inference. In fact, there are some classic examples of causation inference according to spatial variations. The variations of animals in Galapagos Islands inspired Charles Darwin to develop the theory of evolution[6]. The latitudinal zonality reveals that different climate caused the variation of soil on the Earth, and longitudinal zonality reflects the influence of water on land cover and agriculture. In addition to these well-known cases, there have been massive studies in Earth System Sciences to quantify the coupling between two variables according to their spatial variations, yet most of them did not name the discovered relationship as causation[7, 8]. In summary, as illustrated in Fig. 1, the causation in Earth System Sciences could be inferred from the observed data, where the temporal fluctuation is one perspective, and the spatial variation is another perspective. If the observed time series is not sufficient to possess significant changes, causation inference from spatial variations is an alternative solution.

Fig.1 Causation inference in Earth System Sciences form the spatial and temporal perspective

To illustrate the prospect of spatial causation inference, we attempt to use the NPP (net primary production)-climate relationship, a clearly existing causation [5], as

an instance to compare the effect of causation inference from a temporal and spatial perspective respectively. To reduce influence of different vegetation structures on NPP, we solely examined the NPP-climate causation in farmlands. Previous studies proved that water and temperature were the essential conditions for plant growth[4,5]. Specifically, crops can only grow at above-10 °C condition and the most suitable temperature for most crops is above 20 °C[4]. Meanwhile, water is necessary for photosynthesis and evapotranspiration, and is one main component for plants[5].

The annually average NPP data from 2000 to 2015, MOD17A3V055, with a spatial resolution of 1 km (http://files.ntsg.umt.edu/data/NTSG_Products/MOD17/MOD17_A3/) were used as the effect variable. Corresponding annually average temperature (TEM) and precipitation (PRE) data with a spatial resolution of 1 km were used as the cause variables. The land use datasets of China in four periods (2000, 2005, 2010 and 2015) with 1 km spatial resolution were downloaded from Resource and Environment Data Cloud Platform (http://www.resdc.cn/). To reduce the influence of land use change, only those pixels of stable farmlands, which kept unchanged in all four periods, were employed as mask files to extract NPP, PRE and TEM in farmland. In addition to 1 km grids, counties and cities were also used as spatial units to extract the causation between NPP and climate factors at larger spatial scales. The average NPP, PRE and TEM of each county and city were used as effect variable and cause variables, respectively. Meanwhile, the nationally averaged NPP, PRE and TEM were used for temporal causation inference.

Three widely employed temporal causation models, Convergent Cross Mapping (CCM), Granger Causality Test (GCT) and Correlation Analysis (CA) with Pearson coefficients were used to infer causation of climate variables on NPP from a temporal perspective. Meanwhile, four widely employed spatial causation models, Structural Equation Modelling (SEM), CA, Partial Correlation Analysis (P-CA) and Geographical Detector (GD) were used to infer causation of climate data on NPP from a spatial perspective.

Since the time lag may exert an influence on the extracted causation, we

experimented different time lags for these spatial and temporal models to identify the potential largest causation. For temporal models, we experimented a series of time lags. For spatial models, we also analyzed the lagging effects by setting a time lag between the PRE (TEM) and NPP, and thus establishing "lagged" cross sectional data. The general principle and algorithms of different models are introduced as follows:

(i) Convergent Cross Mapping. CCM infers reliable causality between two variables by effectively removing the influence from other variables [9, 10]. It first constructs the shadow manifolds for variable *x* and *y* as formula (1) and (2):

$$M_{x,t} = [x_t, x_{t-\tau}, x_{t-2\tau}, \ldots, x_{t-(E-1)\tau}], \quad (1)$$

where $M_{x,t}$ is the shadow manifold of *x* at time *t*, τ is the time lag, *E* is the number of dimensions, and x_t is the observed value of *x* at time *t*.

$$M_{y,t} = [y_t, y_{t-\tau}, y_{t-2\tau}, \ldots, x_{t-(E-1)\tau}], \quad (2)$$

where $M_{y,t}$ is the shadow manifold of *y* at time *t*.

And then uses the shadow manifolds of *x* to predict the state of *y* and *vice versa* based on formula (3):

$$\hat{M}_{y,k_0}|M_x = \sum_{i=1}^{E+1} \frac{d(M_{x,k_i}, M_{x,k_0})}{\sum_{j=1}^{E+1} d(M_{x,k_j}, M_{x,k_0})} M_{y,k_i}, \quad (3)$$

where $\hat{M}_{y,k_0}|M_x$ is the predicted sate of *y* at time k_0, $d(M_{x,k_i}, M_{x,k_0})$ is the distance between two shadow manifold of *x* at time k_i and k_0 and is calculated in formula (4):

$$d(M_{x,k_i}, M_{x,k_0}) = \exp\left(-\frac{\|M_{x,k_i} - M_{x,k_0}\|}{\|M_{x,k_1} - M_{x,k_0}\|}\right), \quad (4)$$

where exp is the exponential function.

Finally, the correlation coefficient between the predicted states and the observed states after convergence is used to measure the causation effect as formula (5):

$$\rho_{x \to y} = \lim_{L \to +\infty} \mathrm{cor}(M_y, \hat{M}_y|M_x), \quad (5)$$

where $\rho_{x \to y}$ is the causation effect of y on x, cor is correlation function, and L is the size of sample. In this research, the optimal value of E was set as two according to the forecast skill. The library set and prediction set were set to be the same as the length of the time series. Finally, a leave-one-out strategy was employed for cross-validation.

(ii) Granger Causality Test. GCT tests the significance of causality between variables according to the null hypothesis in formula (6), which reduces the full model in formula (7) to formula (8) [11]. If the null hypothesis cannot be denied, x_t can be stated as the Granger cause of y_t.

$$H_0: \omega_i = \omega_2 = ... = \omega_p = 0, \tag{6}$$

where ω_i is the coefficient in the full model used to predict state of y at time t using the observations of previous y and x.

$$y_t = \varphi_0 + \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{i=1}^{p} \omega_i x_{t-i} + \epsilon_t, \tag{7}$$

where p is the number of time lags considered, φ_i and ω_i are coefficients, ϵ_t is the error.

$$y_t = \varphi_0 + \sum_{i=1}^{p} \varphi_i y_{t-i} + \epsilon_t', \tag{8}$$

where ϵ_t' is the new error. Through repeated experiments, the order of lags is set as one in this research for the optimal effects.

(iii) Structural Equation Modelling. SEM can test the causality represented by the graphical and mathematical models[12]. The measurement models are formula (9) and (10), and structural model is formula (11):

$$x = \Lambda\xi + \sigma, \tag{9}$$

$$y = K\eta + \epsilon, \tag{10}$$

$$\eta = B\eta + \Gamma\xi + \zeta, \tag{11}$$

where x and y are observed exogenous variables and endogenous variables respectively. ξ and η are corresponding latent variables. Λ, K, B and Γ are their coefficients, while σ, ϵ and ζ are corresponding errors. In this research, NPP was treated as the endogenous variable while PRE and TEM were exogenous variables. Measurement errors were neglected.

(iv) Geographical Detector. GD assumes that if the explanatory variable has a similar spatial distribution pattern with a target variable, we can suspect there exists cause–effect relationship. The q-value of the factor detector in formula (12) measures the similarity of spatial distribution pattern and its significance can be tested using the non-central F distribution [13]:

$$q = 1 - \frac{\sum_{h=1}^{H} n_h \delta_h^2}{N\delta^2}, \tag{12}$$

where H is the number of strata of explanatory variable, n_h is the size of the strata. δ_h^2 is the variance of target variable within the strata, δ^2 is the global variance and N is the total size. In this research, K-Means was adopted to stratify PRE and TEM, and Elbow Method was used to get the optimal number of strata, which was set as three.

Based on different models, the spatial distribution of temporal causation inference of NPP-climate is presented in Fig. S1 (online). It is seen that the causation detected by CCM, GCT and CA was neither significant nor consistent in most areas of China. For CCM, only a very small proportion of sparse spatial units presented significant causation between NPP and PRE (TEM). Although GCT and CA detected more spatial units with significant causation, the detected NPP-climate causation was relatively weak. In addition to causation inference at 1 km grid level, county level and city level, the causation of nationally averaged PRE and TEM on NPP was also examined using these temporal models (ρ for CCM, correlation coefficients r for CA respectively). As shown in Table S1 (online), except for the linear correlation between TEM and NPP, the NPP-climate causation inferred by other models was not significant. The spatial causation of NPP-climate inferred using different models (q value for GD, correlation coefficients r for CA, r_p for P-CA, and causation effect e of SEM respectively) is presented in Fig. S2 (online) and Table S2 (online). All these models detected relatively strong and significant causation between NPP and PRE (TEM) from a spatial perspective. Specifically, q and r were notably larger than corresponding e and r_p. Meanwhile, e and r_p were calculated by removing the inner interactions between influencing factors and presented a direct causation between

NPP and PRE (TEM). Despite the notable differences between these model outputs, it is clear that NPP-climate causation, which is difficult to infer from a temporal perspective, can be effectively detected from a spatial perspective.

Since Aristotle built the causality framework in 350 B.C., the disputes on whether causality can be comprehended have never stopped[14]. Causation models, which can quantify and compare the influence of multiple individual variables on specific environmental processes (e.g., atmospheric pollution and crop yields), have been increasingly employed in relevant studies. In most implementations, causation inference has been interpreted as temporal causation inference using such classic temporal models as CCM and GCT, while causation inference from a spatial perspective is rarely mentioned. However, the evolution of some subjects in the Earth System Sciences has a time span too long to be measured and recorded, or is too expensive to conduct continuous observations. For instance, the Earth's climate has experienced long-term and dramatic changes during billions of years, yet climate monitoring started until the 18th century[15]. Therefore, notable climate changes in a specific area may not be observed in the available time series, and thus cannot present a casual influence on specific ecological issues. Thus under certain circumstances, it is difficult for causation inference from a temporal perspective. Meanwhile, the large area in geographical research possesses a significant variation of environmental processes across regions, which provides useful spatial information for potential causation inference.

In the case study, the clearly existing NPP-Climate causation, which cannot be inferred by multiple temporal models, was detected effectively by spatial models. In other words, if the time series data are not sufficient, it is highly difficult to infer causation from a temporal perspective, yet it may be feasible for spatial causation inference. In addition to the casual influence of climate change on NPP, it is also challenging for establishing complete time series for a diversity of Earth system variables such as soil heavy metal pollution, vegetation succession, land deterioration and natural resource recovery. In this case, causation inference from a spatial perspective can be an important complement to temporal causation inference.

For both spatial and temporal causation inference, it is crucial to reduce the influence of confounding variables. Generally, confounding variables for time series data are usually less than those confounding variables in spatial cross-section data. Accordingly, spatial causation inference based on cross-section data should be conducted with extra cautiousness. Identification and removal of the influence of confounding variables can be realized through well-designed analysis approaches, including the proper control of suspected confounding variables and the use of models that can measure influence of cause variable independently. For this research, precipitation and temperature interact with each other, and are thus confounding variables when calculating their causal influence on NPP. Among these employed models, CCM, GCT, SEM and P-CA are designed specifically to eliminate the influence from other variables. GD and CA mainly focus on the overall effects of one variable on the others. Despite the model differences, the consistent trend of multi-scale outputs from these models provided a robust cross-verification, suggesting the inferred strong causation between NPP and climate factors was generally reliable.

Strictly speaking, till now, there are neither temporal nor spatial statistical models that can directly infer causation based on the temporal or spatial variations of observed data. An alternative way is to firstly assume the causal relationship according to our existing scientific knowledge (e.g., biological or chemical experiments), and then leverage statistical methods and observed data to verify and measure the causal relationship[1]. In this research, although the diversity of temporal and spatial models holds different capability of suggesting and quantifying causation, the comprehensive outputs of multiple models demonstrated the feasibility and prospect of causation inference from a spatial perspective.

Given the increasing demand of causation inference in Earth System Sciences and the limitation of existing spatial models, more emphasis should be placed on the development of more theoretically robust causation models. Firstly, not only the impact of multiple influencing factors on the target variable, sometimes the impact of the target variable in the neighborhood should also be comprehensively considered. In this research, NPP in the neighborhood exerted very limited influence on NPP in the

target locations. However, for some spatial processes, such as the infectious diseases and airborne pollutants, their spatial spread can have a strong influence on the infectious diseases and airborne pollutants in the neighborhood. For causation inference in these complicated systems where strong self-interactions occur, spatial spillover effects should be added to existing causation models. Furthermore, extended spatiotemporal causation models, which can comprehensively consider the spatial distribution and temporal variations of variables, may be explored for better utilizing limited data sources in complicated causation inference.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2021YFE0102300), the Natural Science Foundation of Beijing (8282031) and the Fundamental Research Funds for the Central Universities. Maps in this article were reviewed by Ministry of Natural Resources of the People's Republic of China (GS(2021)6330)).

Appendix A. Supplementary materials

Supplementary materials to this article can be found in the "Supplementary Materials.docx"

References:

[1] Pearl, J., & Mackenzie, D. The book of why: the new science of cause and effect. New York: Basic Books, 2018

[2] Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge: Cambridge University Press, 2015

[3] Runge J, Bathiany S, Bollt E, et al. Inferring causation from time series in Earth System Sciences. Nat Commun, 2019, 10: 2553

[4] Eck MA, Murray AR, Ward AR, et al. Influence of growing season temperature and precipitation anomalies on crop yield in the southeastern United States. Agric For Meteorol, 2020, 291: 108053

[5] Pattison PM, Tsao JY, Brainard GC, et al. Leds for photons, physiology and food. Nature, 2018, 563: 493-500

[6] Darwin, C. R. On the origin of species by means of natural selection, or the preservation of

favoured races in the struggle for life. London: John Murray, 1859.

[7] Stallins JA. Scale, causality, and the new organism—Environment interaction. Geoforum, 2012, 43: 427-441

[8] Knapp AK, Ciais P, Smith MD. Reconciling inconsistencies in precipitation–productivity relationships: implications for climate change. New Phytol, 2017, 214: 41-47

[9] Sugihara G, May R, Ye H, et al. Detecting causality in complex ecosystems. Science, 2012, 338: 496-500

[10] Tsonis AA, Deyle ER, May RM, et al. Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. Proc Natl Acad Sci USA, 2015, 112: 3253-3256

[11] Gelper S, Croux C. Multivariate out-of-sample tests for granger causality. Comput Stat Data Anal, 2007, 51: 3319-3329

[12] Angelini ME, Kempen B, Heuvelink GBM, et al. Extrapolation of a structural equation model for digital soil mapping. Geoderma, 2020, 367: 114226

[13] Wang J-F, Zhang T-L, Fu B-J. A measure of spatial stratified heterogeneity. Ecol Indic, 2016, 67: 250-256

[14] Zheng Z, Pavlou P. Research note—Toward a causal interpretation from observational data: a new bayesian networks method for structural models with latent variables. Inf Syst Res, 2010, 21: 365–391

[15] Shichi K, Kawamuro K, Takahara H, et al. Climate and vegetation changes around lake baikal during the last 350,000 years. Palaeogeogr Palaeoclimatol Palaeoecol, 2007, 248: 357-375

Bingbo Gao is currently an associate professor at China Agricultural University. He got his B.S. degree from Nanjing University, M.S. and Ph.D. degree from University of Chinese Academy of Sciences. His research interest includes spatial statistics, statistical learning and causation inference.

Ziyue Chen is currently an associate professor at Beijing Normal University. He got his B.S. and M.S. degree from Nanjing University and Ph.D. degree from University of Cambridge. His research interest includes the applications of spatial data analysis tools and remote sensing sources in a diversity of geographical and environmental fields, especially airborne pollution research.