



# Modern causal inference approaches to investigate biodiversity-ecosystem functioning relationships

Jakob Runge



Detecting and quantifying the causal relations of ecosystem functioning is a challenging endeavor. A global study on grasslands illustrates how reasoning about underlying assumptions, from confounding and nonlinearity to fundamental questions of determinism, is key to unlocking the potential of modern causal inference approaches in ecology.

How does biodiversity effect ecosystem productivity? How much diversity does an ecosystem need? Beyond their academic interest, these questions are at the heart of understanding the consequences of largely human-induced changes in biodiversity.

At its core, this is a causal question –predicting the effect of an intervention in species richness on ecosystem functioning–, as opposed to passively observing both: Some measure of correlation would indeed partly measure the causal effect of richness on productivity, but both are also caused by many other factors, for instance, precipitation, and such confounders would also contribute to their correlation.

There are two approaches to answer causal questions: The first one is to actually intervene via the experimental design, where one would randomly select plots to manipulate them with different levels of species richness and then measure how productivity changes at the treated versus the untreated plots. This reduces confounding by generating plots that are comparable with respect to known and unknown confounding factors. However, experiments are generally costly, potentially unethical, and in this case a crucial question is how to intervene: If the causal effect of richness on productivity depends on the specific changes of species, experimental inferences may not generalize to natural ecosystems<sup>1</sup>.

These challenges have led to the development of a second approach in recent decades: To utilize domain assumptions about the underlying system to infer causal effects from observational data within the rich frameworks of causal inference. Two main frameworks are the graphical-causal-model framework<sup>2,3</sup> and the potential-outcome framework<sup>4,5</sup>. Both frameworks are equivalent from a theoretical point of view, but they differ in how assumptions are stated: The potential outcome framework uses an algebra of counterfactuals, while the graphical-causal-model framework uses, well, graphical causal models. These are directed acyclic graphs where the nodes indicate the relevant processes of the system and the arrows indicate direct causal relations. For an introduction to causal inference on time series data see<sup>6</sup>.

A typical observational causal study would start with building a statistical model of species richness and ecosystem productivity, as the naive correlation approach mentioned above, but then add qualitative domain knowledge, for example, in the form of a causal graph, to read off which confounding variables to include in the model such that these block confounding mechanisms (Fig. 1a). However, to fully eliminate confounding effects, observational causal study designs in ecology face the daunting task of identifying, measuring, and statistically controlling for a myriad of confounding variables, from local weather to land-use history<sup>7</sup>.

This can be overcome if multiple time-series (or longitudinal) datasets of the system, for example, at different sites, are available. Then one can jointly leverage these to circumvent observed, as well as unobserved confounding. Multiple datasets are often available in social sciences, where they are called panel data, and the trick then is to include dummy parameters in the statistical regression model, termed fixed-effects panel regression.

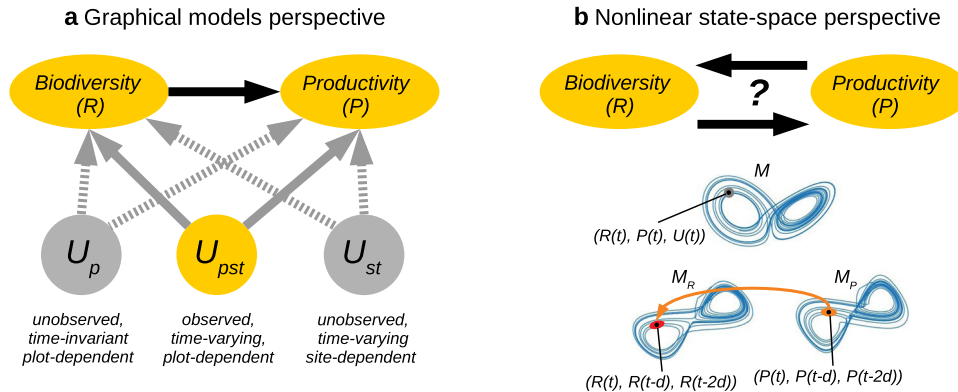
Fixed-effects panel regression was mostly developed within the potential outcome framework<sup>4,5</sup>, but the graphical models framework is catching up<sup>8,9</sup>. On the other hand, explicating assumptions in the form a causal graph is arguably more intuitive than an algebra of counterfactuals, but we will not enter the partially tense discussions between the two factions here<sup>10</sup>.

## Integrating causal graphs in fixed-effects panel regression for a global grassland study

Enter the new study by Dee et al.<sup>11</sup>, which aptly utilized both causal graphs and panel regression to systematically study the effect of richness on productivity from longitudinal data (2007-2017) of 151 unmanipulated plots in 43 grassland sites across 11 countries (Fig. 1a). The authors' goal was to quantify the causal effect of plant richness  $R$  on biomass productivity  $P$  given panel time series data (indexed with  $t$ ) at different sites  $s$  and plots  $p$ . The fixed-effects regression model then is

$$\ln(P_{pst}) = \beta \cdot \ln(R_{pst}) + \delta_p + \mu_{st} + \epsilon_{pst}. \quad (1)$$

The In-In specification is used because the effect of richness on productivity is assumed to be non-linear, but more on nonlinearity later. The regression coefficient  $\beta$  now quantifies the causal effect of interest, the expected percent change in productivity given a one percent change in richness.  $\beta$  is assumed to be independent of time  $t$ , sites  $s$ , and plots  $p$ , and the fixed-effects trick is to add the plot-dummy parameter  $\delta_p$  and the site and time-dummy parameter  $\mu_{st}$  to model any time- or plot- or site-specific effect and apply the model on the pooled data (all time series concatenated) to estimate  $\beta$ ,  $\delta_p$ , and  $\mu_{st}$ . The plot-dummy  $\delta_p$  has a similar effect as de-meaning the plot time series and,



**Fig. 1 | Two perspectives on causal inference in ecology.** **a** Graphical models perspective, where a researcher would assume qualitative causal relationships in the form of a directed acyclic graph among the variables of interest, in Dee et al.<sup>11</sup> biodiversity richness  $R$  and ecosystem productivity  $P$ , as well as confounding variables  $U$ . In the specific fixed-effects panel regression setting of Dee et al.<sup>11</sup> the causal effect of  $R$  on  $P$  can be estimated with a statistical model such as eq. (1) that includes dummy variables to account for unobserved confounders. These can either be time-invariant and only vary across plots ( $U_p$ ), or time- and site-varying, but approximately the same for different plots ( $U_{st}$ ). However, confounders that vary across plots, sites, and time ( $U_{pst}$ ) can only be accounted for if they are

observed and explicitly included in the model (not used in eq. (1)). **b** Nonlinear state-space perspective where the primary goal is to detect whether and in which direction causation exists. A method like convergent cross-mapping (CCM) assumes an underlying deterministic nonlinear system with attractor  $M$  that can be reconstructed from the variables' state spaces ( $M_R, M_P$ ) using time-lagged coordinate embedding. CCM concludes on  $R \rightarrow P$  if points on  $M_R$  can be predicted using nearest neighbors in  $M_P$  (orange ellipse) and the prediction improves the more points on the attractor are sampled. Implicitly, CCM makes the assumption that a successfully reconstructed state-space includes the influence of observed or unobserved confounders  $U$ .

hence, removing any time-invariant confounding effect specific to each plot, while the site and time-dummy  $\mu_{st}$  removes confounding in the site and time-dimension. Any remaining variance is then captured in the noise term  $\epsilon_{pst}$ .

Contrary to expectations (see<sup>12,13</sup>), the authors find that higher plot-level richness causes productivity to decline (that is,  $\beta < 0$ ) rather than increase. The authors attribute the differences to observational studies to prior work not controlling for enough confounding factors, and the difference to experimental studies to the challenge that these had planted fewer rare and non-native species than exist in nature (the challenge to design “natural” interventions). While increases in native species enhance productivity, increases in rare species reduce productivity, which explains the difference.

Panel regression is not a silver bullet. Besides more technical statistical assumptions<sup>5</sup>, the method only accounts for two types of confounders: Those that are essentially invariant over the study period, in Dee et al.<sup>11</sup>, for example, topography or soil type, and those that vary by both site and year, but are largely invariant over the different plots at a site, such as the surrounding land-use change. This leaves open potential confounding effects due to processes that are plot-level specific and time-varying.

To address this issue, Dee et al.<sup>11</sup> went one step further and investigated other model designs to see whether their conclusions are robust, from an instrumental variable design and mediation analysis to accounting for time-lagged dependencies. This is an important step for transparent causal inference<sup>6</sup>: Rather than avoiding explicit causal language, it encourages the researcher to explicitly lay out assumptions that enable more robust conclusions and to discuss conclusions under alternative sets of assumptions. This explicit interpretability is also one of the reasons why causal inference is now becoming a pillar of modern AI<sup>10</sup>.

**Alternative assumptions towards inferring causation in ecology**

The above discussion was focused around overcoming confounding in estimating a specific causal effect by utilizing qualitative assumptions about how the cause and effect in the system are related to potential confounding processes. In the graphical causal models framework these assumptions are explicated in a causal graph whose nodes are random variables of an underlying structural causal model. Each variable's next observation is determined by an assignment function (or autonomous causal mechanism) of the node's parents and an exogenous noise term that subsumes the effects of all factors that are not part of the model and unique to that node. Hence, these noise terms are mutually independent, an assumption called the Causal Markov Condition<sup>2</sup>.

This underlying systemic view then leads to the “intervention-based” notion of causality, which assumes that, in principle, the causal variable can be manipulated, either by humans or nature, by “replacing” its causal mechanism by the intervention value (for example, a certain plant species richness) leading to the causal *do*-calculus<sup>2</sup>.

But there is yet another framework of causality: the prediction-oriented convergent-cross mapping (CCM) approach assumes an underlying deterministic dynamical system<sup>14</sup> (Fig. 1b). CCM is mainly used for what is called causal discovery, also a hallmark of the causal graphical model framework<sup>15</sup>, that is, to detect causation, rather than assume it qualitatively exists and quantify it, as discussed above. However,<sup>16</sup> derived the empirical dynamic modeling (EDM) framework from CCM.

Both CCM and the graphical models (or potential outcome) framework have their distinct sets of assumptions and current capacities of dealing with various challenges. In the original paper<sup>14</sup>, CCM was contrasted against linear Granger causality<sup>17</sup>. CCM utilizes nonparametric methods that make few assumptions about the underlying

functional dependencies and can well model highly nonlinear dynamics. However, CCM also crucially assumes the existence of an (ideally low-dimensional) attractor, exemplified on the Lorenz attractor in ref. 14, and the delicate task is then to reconstruct it using delay embedding<sup>18</sup>. CCM makes the somewhat optimistic assumption that, by successfully reconstructing the attractor, the whole state space, including the influence of observed or unobserved confounders, is implicitly reconstructed. If such a low-dimensional (and ideally not very noisy) attractor is not reconstructable, then the CCM framework cannot well distinguish confounding from causation, as several studies have shown<sup>19,20</sup>.

These findings illustrate that the theory of CCM on dealing with confounding of the various forms treated in this article requires further development. But there have been advances in integrating ideas from causal inference into CCM. Extensions of CCM include dealing with time-delayed causal interactions<sup>21</sup>, making it more robust<sup>22</sup>, (partially) accounting for confounders<sup>23</sup>, but also handling multiple datasets as in panel regression<sup>24</sup>. However, the spatial CCM variant for the latter task cannot well handle heterogeneous dynamics among sites and relatively short time series.

While the causal graphical model and potential outcome frameworks are, in principle, non-parametric and can be combined with machine learning for nonlinear causal effect estimation<sup>25</sup>, the field, starting out from social sciences, has yet to fully embrace the complexity of nonlinear dynamical systems and time series<sup>6</sup>.

In summary, for method developers there is ample room and opportunity to join ideas from different perspectives. More workshops and joint papers from scientists of the different communities could pave the way to such integration. For those wishing to apply causal inference methods to ecology, Dee et al.<sup>11</sup> impressively demonstrate on complex ecosystem interactions how to make assumptions transparent and integrate causal reasoning into data-driven science.

Jakob Runge <sup>1,2</sup> 

<sup>1</sup>Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Datenwissenschaften, Mälzerstr. 3-5, Jena 07745, Germany.

<sup>2</sup>Technische Universität Berlin, Institute of Computer Engineering and Microelectronics, Straße des 17. Juni 135, Berlin 10623, Germany.

 e-mail: [jakob.runge@dlr.de](mailto:jakob.runge@dlr.de)

Received: 23 February 2023; Accepted: 17 March 2023;

Published online: 06 April 2023

## References

- Smith, M. D. & Knapp, A. K. Dominant species maintain ecosystem function with non-random species loss. *Ecol. Lett.* **6**, 509–517 (2003).
- Pearl, J. *Causality: Models, Reasoning, and Inference* 2nd edn. (Cambridge University Press, Cambridge, UK, 2009).
- Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction, and Search* (MIT Press, Boston, 2000).
- Angrist, J. D. & Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion* (Princeton university press, 2009).
- Imbens, G. W. & Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press, 2015).
- Runge, J., Gerhardus, A., Varando, G., Eyring, V. & Camps-Valls, G. Causal inference for time series. *Nat. Rev. Earth Environ.* **10**, 2553 (2023).
- Ferraro, P. J., Sanchirico, J. N. & Smith, M. D. Causal inference in coupled human and natural systems. *Proc. Natl. Acad. Sci.* **116**, 5311–5318 (2019).
- Bareinboim, E. & Pearl, J. Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci.* **113**, 7345–7352 (2016).
- Mooij, J. M., Magliacane, S. & Claassen, T. Joint causal inference from multiple contexts. *J. Mach. Learn. Res.* **21**, 1–108 (2020).
- Pearl, J. & Mackenzie, D. *The Book of Why: The New Science of Cause and Effect* (Basic books, New York, 2018).
- Dee, L. E. et al. Clarifying the effect of biodiversity on productivity in natural ecosystems with longitudinal data and methods for causal inference. *Nat. Commun.* **14**, 2607 (2023).
- Grace, J. B. et al. Integrative modelling reveals mechanisms linking productivity and plant species richness. *Nature* **529**, 390–393 (2016).
- Hooper, D. U. et al. Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecol. Monogr.* **75**, 3–35 (2005).
- Sugihara, G. et al. Detecting causality in complex ecosystems. *Science* **338**, 496–500 (2012).
- Runge, J. et al. Inferring causation from time series in earth system sciences. *Nat. Commun.* **10**, 1–13 (2019).
- Deyle, E. R., May, R. M., Munch, S. B. & Sugihara, G. Tracking and forecasting ecosystem interactions in real time. *Proc. Royal Soc. B: Biol. Sci.* **283**, 20152258 (2016).
- Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometr. J. Econometr. Soc.* **37**, 424–438 (1969).
- Takens, F. Detecting Strange Attractors in Turbulence. In Rand, D. & Young, L.-S. (eds.) *Dynamical Systems and Turbulence, Warwick 1980*, vol. 898 of *Lecture Notes in Mathematics*, chap. 21, 366–381 (Springer, Berlin, 1981).
- Krakovská, A. et al. Comparison of six methods for the detection of causality in a bivariate time series. *Phys. Rev. E* **97**, 042207 (2018).
- Runge, J. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: Int. J. Nonlinear Sci.* **28**, 075310 (2018).
- Ye, H., Deyle, E. R., Gilarranz, L. J. & Sugihara, G. Distinguishing time-delayed causal interactions using convergent cross mapping. *Sci. Rep.* **5**, 1–9 (2015).
- Diaz, E., Adsuara, J., Moreno-Martinez, A., Piles, M. & Camps-Valls, G. Inferring causal relations from observational long-term carbon and water fluxes records. *Sci. Rep.* **12**, 1610 (2022).
- Leng, S. et al. Partial cross mapping eliminates indirect causal influences. *Nat. Commun.* **11**, 2632 (2020).
- Clark, A. T. et al. Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology* **96**, 1174–1181 (2015).
- Chernozhukov, V. et al. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *Econometr. J.* **21**, C1–C68 (2018).

## Acknowledgements

J.R. has received funding from the European Research Council (ERC) Starting Grant CausalEarth under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 948112).

## Competing interests

The author declares no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to Jakob Runge.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023, corrected publication 2023