Why your causal diagram should probably incorporate sampling and measurement processes

*[1]Robin J. Boyd, [1]Rob Cooke, [1]Gary D. Powney, [1]Oliver L. Pescott

[1]UK Centre for Ecology and Hydrology, Benson Lane, Crowmarsh Gifford, OX108BB

*Corresponding author email: robboy@ceh.ac.uk

Prepared for a special issue in Current Opinion in Insect Science

Abstract

Insect scientists are starting to use causal diagrams to display assumptions about causal relationships between variables that exist before any data have been collected. The perception appears to be that these assumptions are sufficient to determine whether observed associations between variables can be interpreted causally. But an observed association implies an observation process (sampling and measurement), and assumptions about that process are also required. We draw on the literature from other disciplines to explain how insect scientists can incorporate assumptions about sampling and measurement in causal diagrams. Making these assumptions explicit allows the investigator to reason more holistically about whether observed associations can be interpreted as causal effects. It also reveals that causal diagrams are not just a tool for causal inference. Assumptions about sampling and measurement are needed to answer descriptive and predictive questions as well. Hence, causal diagrams that incorporate these processes provide a general framework for displaying assumptions regardless of inferential goal.

Introduction

Empirical research questions are often grouped into three categories: *descriptive*, *predictive* and *causal* [1,2]. Imagine that you are studying the geographic distribution of the Purple Emperor (*Apatura iris*), an elusive and relatively scarce species of butterfly, in the United Kingdom (UK). You might ask the following simple questions.

1. **Descriptive question**: What proportion of one-kilometre squares does the species occupy, and has this proportion changed over time? An answer to this question might be useful for national biodiversity monitoring purposes.

2. **Predictive question**: What is the probability that a square within the UK will be occupied given that it has 10-20% broadleaved woodland cover and an average annual temperature of 9-10 ^{o}C? Knowing this conditional probability would allow you to make predictions where there are no data.

3. **Causal question**: On average across squares in the UK, how would the probability of occupancy differ if we were to intervene and increase broadleaved woodland cover by 5%? The answer to this question could inform evidence-based conservation action.

It is natural to begin by expressing a research question in verbal form. To address that question scientifically, however, its answer must be codified as a mathematical quantity whose value can in principle be estimated. In statistical parlance, this quantity is known as the estimand.

An estimand is a summary of a *unit-specific quantity* in a *target population* [3]. The unit-specific quantity is the response variable (e.g. occupancy) or, for causal questions, the change in the response variable under an intervention. The target population is the set of units that is of interest (e.g. all one-kilometre squares in the UK in the above examples). Depending on the question, the population summary might be a mean, a proportion, or some other quantity.

Having defined an estimand, we seek to recover its value from the available data. But what if more than one value is compatible with those data? In this case, we say that the data alone do not uniquely *identify* the estimand (*sensu* Manski; [4]), and assumptions are required.

Causal estimands are never identified from observational (non-experimental) data without assumptions. We observe the association between the explanatory and response variables, but the data themselves cannot tell us whether this association reflects a causal effect (i.e. what would happen if we intervened on the explanatory variable). The association might arise because the independent variable genuinely affects the response, because both are related to something else, or because of some combination of the two [5]. Without additional assumptions, these possibilities cannot be distinguished.

A similar problem arises for descriptive and predictive questions, but for different reasons. Suppose that we want to know the proportion of one-kilometre squares occupied by the Purple Emperor in the UK (the target population), but we do not have data for every square. In this scenario, the proportion of sampled squares that are occupied is known (putting measurement error to one side), but the proportion of non-sampled squares that are occupied could plausibly take a wide range of values. Consequently, many different answers are consistent with the observed data, and the estimand is again not identified [4].

To narrow down the range of possible values that an estimand can take, we have to make assumptions. For our causal question, we might assume that after accounting for measured variables (e.g. soils, climate and land use), there are no remaining common causes of broadleaved woodland and Purple Emperor occupancy. This assumption would permit us to interpret the observed association causally. For our descriptive question, we might assume that same proportion of squares are occupied in the sample as in the target population. In each case, identification is achieved by ruling out alternative values that would otherwise be compatible with the data.

Identifying assumptions are necessary, but they are not a free lunch. If we make an assumption that does not hold, then we risk ruling out not only incorrect values of the estimand but also the correct one [4]. It is therefore important to be explicit about the assumptions that are being maintained so that their plausibility can be evaluated.

A particularly transparent way to portray assumptions about the data generating process is in the form of a causal diagram [6]. Causal diagrams are pictures of how we believe the world works. Variables are depicted as nodes, and arrows are drawn from causes to effects [7]. The structure of the diagram and the assumptions that it encodes can tell us whether the answer to our question is identified given the available data (Fig. 1; [3]).

We are beginning to see the use of causal diagrams in insect science (Table 1), but it is not clear that their full potential is being realised. Existing applications focus almost exclusively on causal questions and rarely incorporate observation processes such as sampling and measurement (more on this below). As a result, many of the assumptions required to identify estimands remain implicit and are not subjected to scientific scrutiny.

The purpose of this paper is therefore to demonstrate how insect scientists can get the most out of causal diagrams. We begin with a brief note on terminology and an introduction to the concept of statistical independence, since it is the bridge between causal diagrams and identification assumptions. We then review the use of causal diagrams in insect science with a particular focus on inferential goals (causal, descriptive or predictive) and whether the observation process is incorporated. The final two sections demonstrate how to incorporate sampling and measurement processes in causal diagrams and how they can be used to display assumptions needed to identify descriptive and predictive estimands.

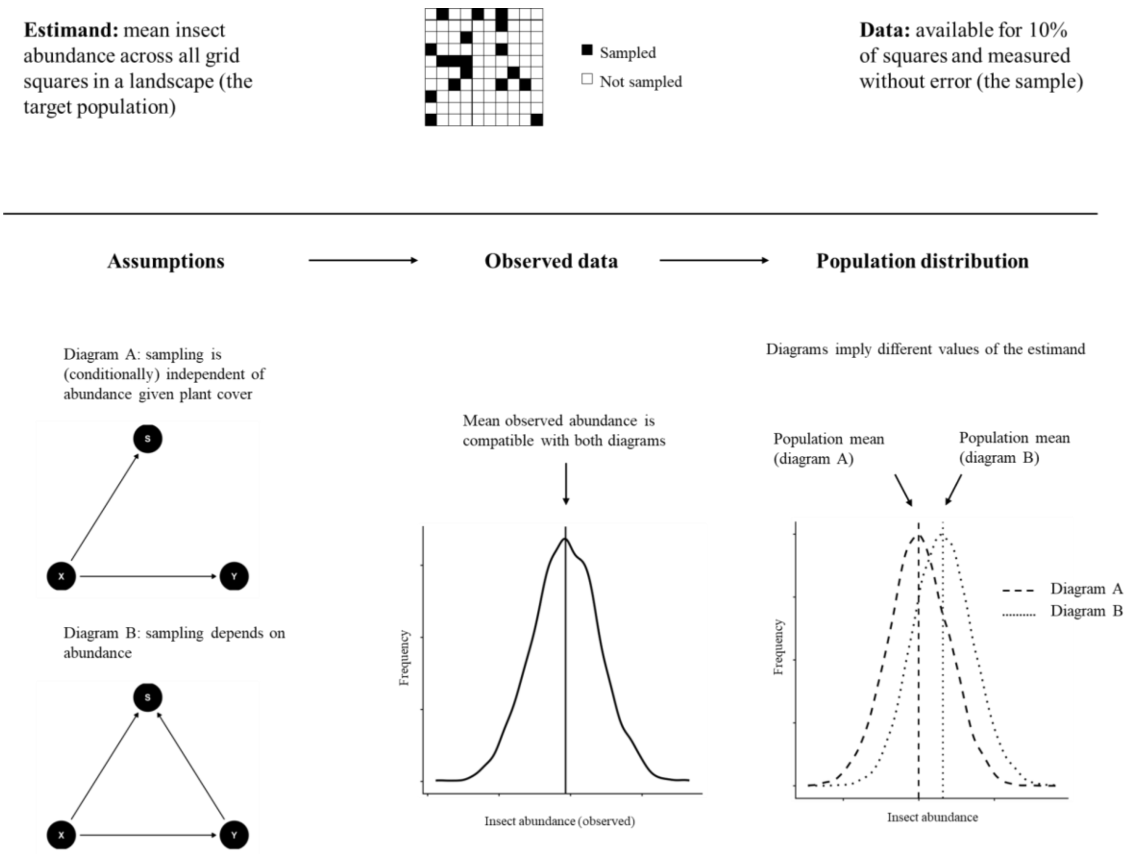

Figure 1. Schematic illustrating why the mean abundance of a species across sites within a landscape (a descriptive estimand) is not identified from a sample without assumptions. The point is that different causal diagrams, which encode assumptions about the data generating process, can be compatible with the same observable data distribution (the distribution of values we could observe) but different values of the estimand. Here, for example, we estimate greater mean abundance following diagram B, implying that sampled locations have greater observed abundance. In this sense the data do not uniquely identify the value of the estimand, and we have to decide which causal diagram is most plausible. The same logic applies to predictive estimands, which are similar to descriptive estimands but conditional on the values of covariates. An example of a predictive estimand is mean insect abundance across sites where plant cover falls within a particular range; this conditional distribution can be used to make site-level predictions.

A note on terminology

We deliberately avoid discipline-specific terminology in this paper. Readers should not expect to see terms like “confounder”, “selection bias”, and “internal validity”. These terms often carry subtly different meanings across fields, which can confuse matters [8]. Instead, we use the language of causal diagrams to describe settings in which estimands pertaining to specified target populations are and are not identified.

A declaration of (statistical) independence

Many primers on causal diagrams are available [6,7]. Here we introduce the subset of concepts needed to progress our argument.

A causal diagram comprises a set of variables with arrows between them pointing from cause to effect. The arrangement of the variables and arrows encodes information about which variables are statistically independent of one another. Many identifying assumptions can be expressed in exactly these terms.

In a causal diagram, any sequence of variables connected by arrows—regardless of their direction—is called a *path*. Whether a path is open or closed determines whether statistical dependence can flow between the variables at either end. An open path transmits dependence; a closed path blocks it.

Certain variables determine whether paths are open or closed. A variable with two arrows pointing towards it is called a *collider* (the arrows “collide” at a variable). By default, a collider closes a path. We will refer to variables on a path that do not have two arrows pointing towards them (e.g. they have one pointing towards them and one away, or two pointing away) as non-colliders. By default, a non-collider opens a path and allows statistical dependence to flow.

Paths can be opened or closed by *conditioning* on variables, which simply means holding them fixed or focusing on particular values they take. (In practice, we condition on a variable when we include it as a covariate in a regression model or adjust for it in some other way.) Conditioning on a non-collider closes a previously open path, whereas conditioning on a collider opens a previously closed one (the opposite of their ‘default’ roles). When all paths linking two variables are closed by conditioning, those variables are said to be conditionally independent given the variables conditioned on.

Many identification assumptions can be expressed as statements of (conditional) independence. Imagine again our descriptive goal of knowing the proportion of one-kilometre squares occupied by the Purple Emperor in the UK. If sample inclusion (e.g. whether data were collected at a square) is independent of occupancy, then the proportion of occupied sites is the same in the sample as it is in the wider population (in expectation). Under this assumption, the proportion of squares that are occupied in the UK is uniquely identified from the sample.

Causal diagrams in insect science and ecology

The assertion that scientists would often like to understand “the causes of things” is an obvious one, but one, nonetheless, that has generated great debate in biology with regards to definitional and operational detail [9–11]. The recent glut of papers on causal inference in insect science and ecology more generally suggests that we are starting to achieve a common understanding of the methods available to us, and what they have (and occasionally have not) achieved elsewhere **[5,7,12–22]. [Others (e.g. [23,24]) might reasonably lament how long it has taken much of the discipline to catch up with them!] Chief among the tools now adopted is the causal diagram.

It is interesting to observe which applications of causal diagrams have been imported to ecology from other disciplines and which have not. What *has* caught on is the use of causal diagrams to represent assumptions about how variables are causally linked in the target population before any data are collected. So too has the fact that causal diagrams can be used to reason about whether causal effects are transportable between target populations [25,26]. What does not appear to have been appreciated is that observation processes such as sampling and measurement can also be incorporated [27]. Table 1 demonstrates this point with a specific focus on applied examples in insect science.

Table 1. Applications of causal diagrams in insect science. Each row pertains to one study. The first column is the reference, the second is the research question addressed, the third is the inferential goal (causal, descriptive or predictive), and the final column indicates whether observation processes (nodes for sample selection or measured proxies of relevant variables) were explicitly included in the causal diagram. The list only includes studies that we are aware of and is therefore unlikely to be exhaustive.

Reference	Research question	Inferential goal	Observation processes included in the causal diagram?
Boyd et al. [28]	How has the mean abundance of two species of butterfly across the UK changed over time?	Descriptive	Selection of sites into the sample
Boyd et al. [29]	What are the relative causal effects of various explanatory variables on the accuracy of species distribution models?	Causal	No
Byrnes and Dee [5]	What is the causal effect of temperature on snail* abundance?	Causal	No
Chen et al. [30]	What are the direct and indirect effects of macrophyte management on aquatic insect abundance?	Causal	No
Gross et al. [31]	What is the effect of early-season insecticide application on secondary pest outbreaks?	Causal	No
Guzman et al. [32]	What is the effect of pesticide application (neonicotinoids and pyrethroids), animal-pollinated agriculture and honeybee colonies on wild bee occupancy?	Causal	No
Proesmans et al. [33]	What is the effect of managed honeybees and a range of environmental drivers on viral prevalence in wild pollinators?	Causal	No

Saavedra et al. [34]	Propose causal inference to increase our causal understanding of ecological systems.	Causal	No
Takeshita et al. [35]	What is the effect of management intervention in nickel concentrations on Ephemeroptera, Plecoptera, and Trichoptera richness?	Casual	No

*invertebrate, not insect

To us it is puzzling that assumptions about causal relationships in nature, but not about sampling and measurement (collectively the "observation process"), are typically represented in causal diagrams. The former are required to identify causal effects when the available data are observational (i.e. do not come from a randomised experiment; [18]). By the same logic, however, assumptions about the sampling process are often required when the available data are not a random sample (or census) of the target population [36]. (Similar reasoning applies to measurement, as we explain in the next section.) It is therefore difficult to see why one class of assumption is routinely represented in causal diagrams while the other is not—especially in insect science, where the available data are seldom a random sample of the target population or measured without error [37–39].

It is also worth noting that the assumptions needed to identify descriptive and predictive estimands often relate to the observation process. For example, to identify the proportion of one-kilometre squares occupied by the Purple Emperor in the UK, we have make assumptions about how the selection of which sites to sample relates to occupancy (see the Introduction). Consequently, by incorporating assumptions about sampling and measurement, causal diagrams can be used to reason about whether non-casual estimands are identified [28,40].

Incorporating missing data and measurement error

We have now seen that, in insect science, causal diagrams are mostly used to reason about non-causal correlations between variables and seldom represent sampling and measurement (Table 1). But sampling and measurement can themselves induce discrepancies between the associations we observe and the causal effect we seek. Since both processes can readily be incorporated in causal diagrams, it makes sense to include them. This section explains how.

Let us first consider sampling from a target population such as the set of one-kilometre squares in the UK. The selection of which units (e.g. squares) to include in the sample is a binary variable (yes versus no) and as such can be included in a causal diagram [41,42]. Data are missing for non-sampled units.

An important feature of sample inclusion is that we are forced to condition on (or focus on one level of) it, because we only have data on sampled units. Hence, if sample inclusion is a collider on a path linking the explanatory and response variables, it induces a dependence between the two (by opening that path) that is not part of the direct causal effect. In this case, the direct effect is not identified on account of sample inclusion [27,36].

It is worth noting that sample selection nodes are not always necessary. If the available data are a census of, or a random sample from, the target population, then sample inclusion is independent of the variables under study and can effectively be ignored. Some have argued that samples need not be representative of the target population because causal effects reflect general mechanisms rather than properties specific to the sampled individuals [43]. At first glance this perspective might appear to undermine the need to represent sample selection. However, even under this interpretation, selection into the sample can still induce a discrepancy between the association observed among sampled units and the causal effect among those same units ([44]; the bottom left panel in Fig. 2 represents one such situation).

As for measurement error, the key is to recognise that we never observe a variable itself but rather a measured proxy. Strictly speaking, these proxies should be included in the causal diagram. Each proxy is affected by its true underlying construct, but it might also be affected by other things. If the measured proxies for the explanatory and response variables share a common cause, then that common cause is a non-collider, and the direct causal effect is not identified without additional assumptions or auxiliary data [45,46]. Fig. 2 demonstrates how one might include sample inclusion and measurement error in a causal diagram.

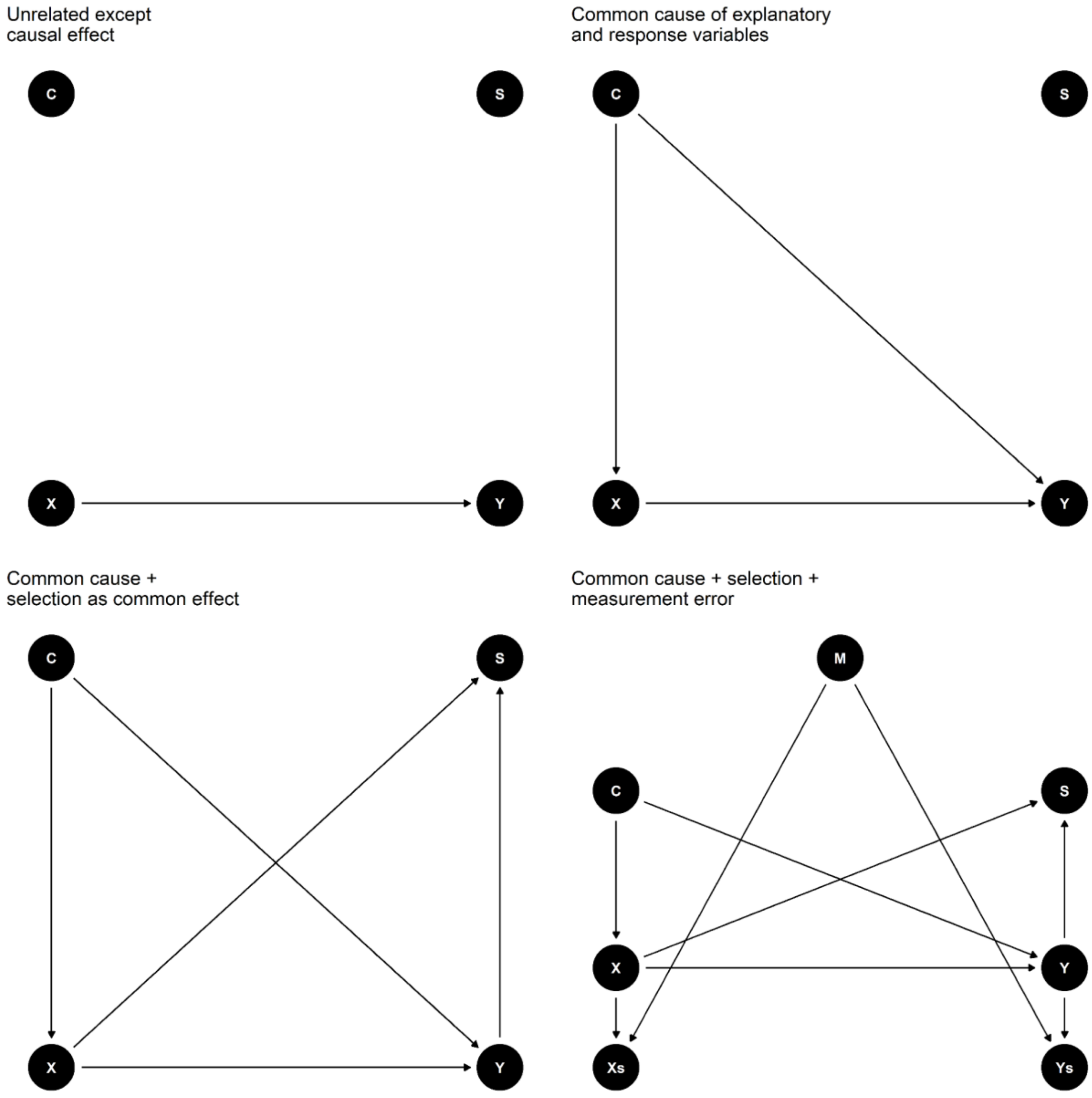

Figure 2. A stylized example depicting assumptions that might be made by a researcher seeking to estimate the average causal effect of plant cover (explanatory variable) on insect abundance (the response) across a target population of plots. s denotes plot inclusion in the sample (yes versus no), x denotes plant cover, y denotes insect abundance, c denotes temperature, xs denotes the measured proxy for plant cover, ys denotes the measured proxy for insect abundance, and m denotes observer expertise. The top left panel depicts the very strong assumption that plant cover and insect abundance are independent apart from the direct causal effect. Under this assumption, the effect is identified. The top right panel assumes that the association between plant cover and insect abundance partly reflects a correlation induced by temperature (this the classical notion of *confounding*). By default, the effect is not identified under this assumption, although it becomes identified if temperature is measured and conditioned on. The bottom left panel assumes that the effect is confounded by temperature and that selection of plots into the sample is a collider on a non-causal path linking the explanatory and response variables. Even if we condition on temperature, the causal effect is not identified under these assumptions, because we are forced to condition on sample inclusion. The bottom right panel contains the same assumptions as the bottom left but also supposes that both plant cover and insect abundance are measured with error. The measured proxies of the explanatory and response variables share a common cause (observe expertise, which determines measurement error). Even if there was no confounding, and sample inclusion was independent of the explanatory and response variables, this measurement error structure would render the effect not identified from the data alone without further assumptions.

Causal diagrams for non-causal questions

Most of the literature on the use of causal diagrams for descriptive and predictive questions concerns assumptions about missing data *[47,48]. In this setting, the main concern is whether sample inclusion is independent of the response variable. If there are no paths connecting the two, then they are unconditionally independent, and the sampling process is said to be *ignorable*. If sample inclusion and the response are connected by a path, but that path can be blocked by conditioning on a non-collider, then the sampling process is said to be *conditionally ignorable*. In both cases, the answers to descriptive and predictive questions about the response variable are identified [4,40].

Sometimes, there are paths connecting sample inclusion and the response variable that are not possible to block. In this case, the sampling process is *not ignorable*, and the answers to causal and descriptive questions about the response variable are not identified. Fig. 2 depicts ignorable, conditionally ignorable and not ignorable sampling processes using the same stylized example as in Fig. 2.

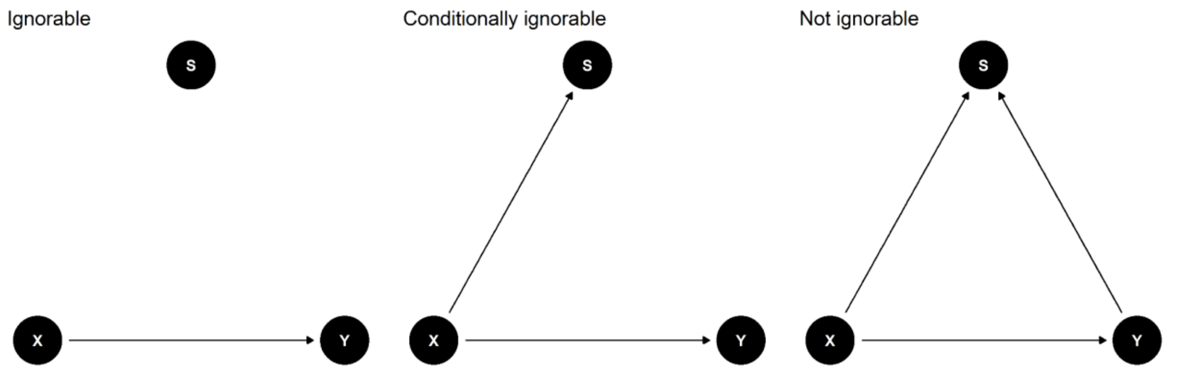

Figure 2. A stylized example depicting assumptions that might be made by a researcher seeking to estimate the mean insect abundance across plots (descriptive estimand) or its mean conditional on the

level of plant cover (predictive estimand) in a target population of plots. s denotes plot inclusion in the sample (yes versus no), x denotes plant cover and y denotes insect abundance. The left-hand panel depicts the very strong assumption that the sampling process is ignorable. That is, sample inclusion is (unconditionally) independent of the insect abundance. This assumption identifies (conditional) mean abundance. The middle panel depicts the weaker assumption that the sampling process is conditionally ignorable if plant cover is measured and conditioned on. Conditional ignorability means that insect abundance is conditionally independent of sample inclusion within levels of plant cover. Under this assumption, the (conditional) mean abundance is again identified. The right-hand panel depicts the assumption that sample inclusion depends directly on insect abundance. Under this assumption, the sampling process is said to be not ignorable, and the (conditional) mean abundance is not identified.

Concluding remarks (100 words)

Our concluding message to insect scientists is this. If you are going to draw a causal diagram, you might as well incorporate the observation process. Doing so will allow you to reason more holistically about the assumptions needed to identify your estimand. A causal diagram that incorporates the observation process can also be used to reason about whether descriptive and predictive estimands are identified. In other disciplines, causal diagrams form part of official reporting guidelines [49], and we believe that insect science would benefit from following suit.

Acknowledgements

RJB was supported by the UK CEH National Capability for UK Challenges programme NE/Y006208/1 and the MAMBO (Modern Approaches to Monitoring Biodiversity) project, which receives funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No.101060639.

References

1. Carlin JB, Moreno-Betancur M, Carlin J: **On the uses and abuses of regression models: a call for reform of statistical practice and teaching**. 2023,

2. Ito C, Al-Hassany L, Kurth T, Glatz T: **Distinguishing Description, Prediction, and Causal Inference A Primer on Improving Congruence Between Research Questions and Methods**. *Neurology* 2025, **104**.

3. Lundberg I, Johnson R, Stewart BM: **What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory**. *Am Sociol Rev* 2021, **86**:532–565.

4. Manski C: *Identification for prediction and decision*. Cambridge, Massachussets; 2007.

5. Byrnes JEK, Dee LE: **Causal Inference With Observational Data and Unobserved Confounding Variables**. *Ecol Lett* 2025, **28**.

6. Pearl J, Glymour M, Jewell N: *Causal inference in statistics: A primer*. Wiley; 2016.

7. Arif S, MacNeil MA: **Applying the structural causal model framework for observational causal inference in ecology**. *Ecol Monogr* 2022, doi:10.1002/ecm.1554.

8. Glymour M, Greenland S: **Causal diagrams**. In *Modern Epidemiology*. . Lippincot Williams & Wilkins; 2008:192–193.

9. Mayr E: *Cause and Effect in Biology*. 1961.

10. Shipley B: *Cause and correlation in biology*. Cambridge University Press; 2000.

11. Laland KN, Sterelny K, Odling-Smee J, Hoppitt W, Uller T: **Cause and Effect in Biology Revisited: Is Mayr's Proximate-Ultimate Dichotomy Still ?Useful**. *New Series* 2011, **334**.

12. Arif S, MacNeil MA: **Predictive models aren't for causal inference**. *Ecol Lett* 2022, **25**:1741–1745.

13. Franks DW, Ruxton GD, Sherratt T: **Ecology needs a causal overhaul**. *Biological Reviews* 2025, **100**:1950–1969. **This paper provides a colourful introduction to the concepts for causal inference for ecologists. The section on causal diagrams is excellent—particularly the response to the criticism that "it's hard".

14. Grace JB, Irvine KM: **Scientist's guide to developing explanatory statistical models using causal analysis principles**. *Ecology* 2020, **101**:1–14.

15. Grace JB, Huntington-Klein N, Schweiger EW, Martinez M, Osland MJ, Feher LC, Guntenspergen GR, Thorne KM: **Causal Effects Versus Causal Mechanisms: Two Traditions With Different Requirements and Contributions Towards Causal Understanding**. *Ecol Lett* 2025, **28**. *The authors distinguish between causal statistics and causal investigation. It is a shame we did not have the space to get into this distinction here. Most of our arguments were developed with causal statistics in mind, but we can see the value of both paradigms.

16. Larsen AE, Meng K, Kendall BE: **Causal analysis in control–impact ecological studies with observational data**. *Methods Ecol Evol* 2019, **10**:924–934.

17. Laubach ZM, Murray EJ, Hoke KL, Safran RJ, Perng W: **A biologist's guide to model selection and causal inference**. *Proceedings of the Royal Society B: Biological Sciences* 2021, **288**.

18. Siegel K, Dee LE: **Foundations and Future Directions for Causal Inference in Ecological Research**. *Ecol Lett* 2025, **28**.

19. Schrodt F, Beck M, Estopinan J, Bowler DE, Fontaine C, Gaüzère P, Goury R, Grenié M, Martins IS, Morueta-Holme N, et al.: **Advancing causal inference in ecology: Pathways for biodiversity change detection and attribution**. *Methods Ecol Evol* 2025, **16**:2276–2304.

20. Correia HE, Dee LE, Byrnes JEK, Fieberg JR, Fortin M-J, Glymour C, Runge J, Shipley B, Shpitser I, Siegel KJ, et al.: **Best practices for moving from correlation to causation in ecological research**. *Nat Commun* 2026, **17**:1981. *This perspective is a nice introduction to concepts in causal inference. We especially like the emphasis on assumptions and identification.

21. Butsic V, Lewis DJ, Radeloff VC, Baumann M, Kuemmerle T: **Quasi-experimental methods enable stronger inferences from observational data in ecology**. *Basic Appl Ecol* 2017, **19**:1–10.

22. Dee LE, Ferraro PJ, Severen CN, Kimmel KA, Borer ET, Byrnes JEK, Clark AT, Hautier Y, Hector A, Raynaud X, et al.: **Clarifying the effect of biodiversity on productivity in natural ecosystems with longitudinal data and methods for causal inference**. *Nat Commun* 2023, **14**.

23. Shipley B: *Testing Causal Explanations in Organismal Biology: Causation, Correlation and Structural Equation Modelling*. 1999.

24. Grace JB: *Structural equation modeling and natural systems*. Cambridge University Press; 2006.

25. Spake R, O'Dea RE, Nakagawa S, Doncaster CP, Ryo M, Callaghan CT, Bullock JM: **Improving quantitative synthesis to achieve generality in ecology**. *Nat Ecol Evol* 2022, **6**.

26. Tabell O, Moser N, Ovaskainen O, Karvanen J: **Transporting Causal Effects in Ecology: Concepts, Models and Software**. 2026, doi:10.64898/2026.03.03.709251.

27. Fox M, Arah O, Swanson S, Viallon V: **Causal diagrams to evaluate sources of bias**. 2024.

28. Boyd RJ, Botham M, Dennis E, Fox R, Harrower C, Middlebrook I, Roy DB, Pescott OL: **Using causal diagrams and superpopulation models to correct geographic biases in biodiversity monitoring data**. *Methods Ecol Evol* 2025, **16**:332–344.

29. Boyd RJ, Harvey M, Roy D, Barber T, Haysom K, ...: **Causal inference and large-scale expert validation shed light on the drivers of SDM accuracy and variance**. *Divers Distrib* 2023, doi:10.1111/ddi.13698.

30. Chen PY, Dong X, Durand J, Yuan HW: **Impact of Emergent Macrophyte Mowing on an Aquatic Insect Community in Urban Ponds: A Case Study in an Artificial Wetland in Southeast Asia**. *Wetlands* 2025, **45**.

31. Gross K, Rosenheim JA: **Quantifying secondary pest outbreaks in cotton and their monetary cost with causal-inference statistics**. *Ecological Applications* 2011, **21**:2770–2780.

32. Guzman LM, Elle E, Morandin LA, Cobb NS, Chesshire PR, McCabe LM, Hughes A, Orr M, M'Gonigle LK: **Impact of pesticide use on wild bee distributions across the United States**. *Nat Sustain* 2024, doi:10.1038/s41893-024-01413-8.

33. Proesmans W, Alaux C, Albrecht M, Bassit K, Cyrille N, Dalmon A, Diévart V, Dominik C, Felten E, Gajda A, et al.: **Drivers of Viral Prevalence in Landscape-Scale Pollinator Networks Across Europe: Honey Bee Viral Density, Niche Overlap With This Reservoir Host and Network Architecture**. *Ecol Lett* 2026, **29**.

34. Saavedra S, Bartomeus I, Godoy O, Rohr RP, Zu P: **Towards a system-level causative knowledge of pollinator communities**. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2022, **377**.

35. Takeshita KM, Hayashi TI, Yokomizo H: **The effect of intervention in nickel concentrations on benthic macroinvertebrates: A case study of statistical causal inference in ecotoxicology**. *Environmental Pollution* 2020, **265**.

36. Hernán MA, Hernández-Díaz S, Robins JM: **A structural approach to selection bias**. *Epidemiology* 2004, **15**:615–625.

37. Cooke R, Outhwaite CL, Bladon AJ, Millard J, Rodger JG, Dong Z, Dyer EE, Edney S, Murphy JF, Dicks L V., et al.: **Integrating multiple evidence streams to understand insect biodiversity change**. *Science* 2025, **388**:eadq2110.

38. Li M, Boyd RJ, Smith C, Fox R, Roy D, Bennie J, ffrench-Constant RH: **Citizen scientists as butterfly predators: using foraging theory to understand individual recorder behaviour**. *Ecol Modell* 2025, **510**.

39. Rossini L, Contarini M, Delfino I, Speranza S: **Does insect trapping truly measure insect populations?** *Agric For Entomol* 2025, **27**:329–339.

40. Schuessler J, Selb P: **Graphical Causal Models for Survey Inference**. *Sociol Methods Res* 2025, **54**:74–105.

41. Bareinboim E, Tian J, Pearl J: *Recovering from Selection Bias in Causal and Statistical Inference*. [date unknown].

42. Thoemmes F, Mohan K: **Graphical Representation of Missing Data Problems**. *Structural Equation Modeling* 2015, **22**:631–642.

43. Rothman KJ, Gallacher JEJ, Hatch EE: **Why representativeness should be avoided**. *Int J Epidemiol* 2013, **42**:1012–1014.

44. Mathur MB, Shpitser I: **Simple graphical rules for assessing selection bias in general-population and selected-sample treatment effects**. *Am J Epidemiol* 2025, **194**:267–277.

45. Wardle MT, Reavis KM, Snowden JM: **Measurement error and information bias in causal diagrams: mapping epidemiological concepts and graphical structures**. *Int J Epidemiol* 2024, **53**.

46. Rudolph JE, Ross RK: **Applying causal diagrams with measurement error: An outline and further considerations**. *Int J Epidemiol* 2025, **54**.

47. Dumelle M, Trangucci R, Nahlik AM, Olsen AR, Irvine KM, Blocksom K, Ver Hoef JM, Fuentes C: **Missing data in ecology: Syntheses, clarifications, and considerations**. *Ecol Monogr* 2025, **95**. *This paper is a nice introduction to missing data theory. It contains a section on representing missing data mechanisms using causal diagrams, which covers a lot of the technical nuance that we did not have the space to incorporate here.

48. Bowler DE, Boyd RJ, Callaghan CT, Robinson RA, Isaac NJB, Pocock MJO: **Treating gaps and biases in biodiversity data as a missing data problem**. *Biological Reviews* 2025, **100**:50–67.

49. McIntyre KJ, Tassiopoulos KN, Jeffrey C, Stranges S, Martin J: **Using causal diagrams within the Grading of Recommendations, Assessment, Development and Evaluation framework to evaluate confounding adjustment in observational studies**. *J Clin Epidemiol* 2024, **175**.