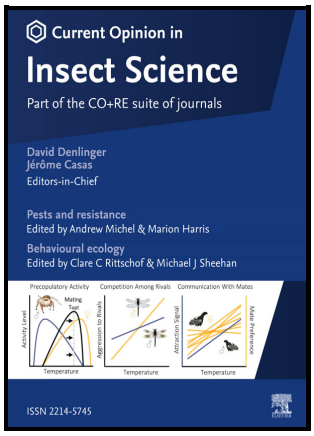
Why your DAG should probably incorporate sampling and measurement processesShort title: Causal diagrams for non-causal questions

Robin J. Boyd, Rob Cooke, Gary D. Powney, Oliver L. Pescott



PII: S2214-5745(26)00094-5

DOI: https://doi.org/10.1016/j.cois.2026.101578

Reference: COIS101578

To appear in: *Current Opinion in Insect Science*

Please cite this article as: Robin J. Boyd, Rob Cooke, Gary D. Powney and Oliver L. Pescott, Why your DAG should probably incorporate sampling and measurement processesShort title: Causal diagrams for non-causal questions, *Current Opinion in Insect Science,* (2026) doi:https://doi.org/10.1016/j.cois.2026.101578

Why your DAG should probably incorporate sampling and measurement processes

[*1,2]Robin J. Boyd, [1,2]Rob Cooke, [1]Gary D. Powney, [1]Oliver L. Pescott

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

[2]Centre for Ecology and Conservation, University of Exeter, Falmouth, Cornwall, UK

*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 Directed Acyclic Graphs (DAGs) 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 DAGs. 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 DAGs are not just a tool for causal inference; assumptions about sampling and measurement are also needed to answer descriptive and predictive questions. Hence, DAGs 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 ºC? 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 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])[1], and assumptions are required to narrow down the range of possibilities (see Box 1 for a worked example).

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 (but see [6,7]).

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 occupancy is not systematically higher or lower across sampled squares than in the wider country. 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 [4,8]. It is therefore important to be explicit about the assumptions that are being maintained so that their plausibility can be evaluated.

Box 1. Numerical example demonstrating why assumptions may be needed to identify estimands from the available data.

[1] Strictly speaking, identification concerns what can be learned from an arbitrarily large sample generated by the same observation process, rather than from the finite sample in hand. Throughout, references to the "available data" should therefore be interpreted as referring to the information contained in the observation process, rather than a particular finite realisation of it. What can be learned from the sample in hand is a matter of statistical inference, which presupposes identification and addresses the additional uncertainty arising from finite sampling. This paper is concerned with identification rather than statistical inference.

Suppose that we want to know the proportion of one-kilometre squares in the United Kingdom that are occupied by the Purple Emperor (*Apatura iris*). This quantity is our estimand. Data are not available for every square, so the estimand cannot be calculated directly.

To formalise the problem, let Y denote per square Purple Emperor occupancy (1 if yes; 0 if no), Z denote habitat quality (say, inside versus outside a nature reserve), and S indicate whether each square was sampled (1 if yes; 0 if no). Then define the national proportion of occupied squares (the estimand) as $P(Y = 1)$, the sample proportion of occupied squares as $P(Y|S = 1)$, the proportion of non-sampled squares that are occupied as $P(Y|S = 0)$, and the proportion of sampled squares as $P(S = 1)$.

By the law of total probability, the national occupancy can be written as a weighted average of occupancy among sampled and non-sampled squares: $P(Y = 1) = P(Y = 1|S = 1)P(S = 1) + P(Y = 1|S = 0)P(S = 0)$. Since $P(Y = 1|S = 0)$ is unknown, the national proportion of occupied squares is not uniquely identified from the available data alone. But what can the available data tell us?

Suppose as an example that 10% of squares are sampled and that occupancy among sampled squares is 0.6. That is, $P(S = 1) = 0.1$, $P(S = 0) = 0.9$, and $P(Y|S = 1) = 0.6$. Plugging these values into the law of total probability gives $P(Y = 1) = 0.6 \times 0.1 + P(Y = 1|S = 0) \times 0.9$. Without making any further assumptions, all we can say about the unobserved term $P(Y = 1|S = 0)$ is that it lies between 0 and 1. Substituting these extreme values for $P(Y = 1|S = 0)$ gives the 'identification region' $P(Y = 1) \in [0.06,0.96]$. Put more simply, the observed data alone tell us that the species occupies somewhere between 6% and 96% of squares, which is not particularly useful! To narrow down the range of possibilities, we have to make assumptions about the data generating process.

One strong assumption is that occupancy among non-sampled squares is asymptotically equivalent to its value among sampled squares and hence in the country as a whole. That is, $P(Y = 1|S = 0) = P(Y = 1|S = 1) = P(Y = 1)$. This assumption 'point identifies' the estimand $P(Y = 1)$, meaning, together with the data, it rules out every value of the estimand except one. Of course, if occupancy differs between sampled and non-sampled squares, then the assumption does not hold and the correct value of the estimand is ruled out. Hence, there is often a trade-off between the strength of an assumption (how many values of the estimand it precludes) and its plausibility.

One way to portray assumptions about the data generating process is in the form of a causal Directed Acyclic Graph (DAG; [9]). DAGs are pictures of how we believe the world works. Variables are depicted as nodes, and arrows are drawn from causes to effects [10]. Feedback loops are not permitted, as this would imply that a variable can cause itself. The structure of the DAG and the assumptions that it encodes provide important information on whether the answer to our question is identified given the available data [3], *regardless of whether the question itself is causal*.

It is interesting to compare the use of DAGs in insect science with other disciplines. In insect science, DAGs are predominantly used to display assumptions about how variables are linked in the world before any data have been collected (e.g. broadleaved woodland → Purple Emperor occupancy). Elsewhere, researchers often go one step further and incorporate observation processes such as sampling and measurement ([11]; e.g. broadleaved woodland → inclusion of a one-kilometre square in the sample). We argue that DAGs in insect science should also incorporate the observation process, since this type of assumption is often needed for identification (whether of descriptive, predictive or causal estimands).

The remainder of the paper develops this argument. We first explain how DAGs encode statistical dependence and why this matters for identification. We then review the use of DAGs in insect science, focusing on inferential goals (descriptive, predictive or causal) and whether the observation process is incorporated. The final two sections demonstrate how to incorporate the observation process and reason about the identification of descriptive, predictive and causal estimands.

A declaration of (statistical) independence

Many primers on DAGs are available [9,10]. Here we introduce the subset of concepts needed to progress our argument.

A DAG 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 DAG, 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; see the variable S in the bottom left panel of Fig. 1). 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. For example, the variables X and C in the bottom left panel of Fig. 1 are 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, which would be true under simple random sampling, 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.

DAGs versus Structural Equation Models

A reviewer rightly pointed out that DAGs are sometimes conflated with Structural Equation Models (SEMs), so it is worth briefly distinguishing the two. Grace [12] defined SEM as "the use of two or more structural equations [essentially regressions] to model multivariate relationships". SEMs may be depicted as DAGs or more general graphical structures [13], but the ultimate goal is parameter estimation. Here, our focus is not on estimating SEMs but on how DAGs encode conditional independence statements and what those statements say about identification.

DAGs in insect science and ecology

In insect science and ecology, DAGs have mostly been applied in causal settings (Table 1). By far the most widely appreciated application is to identify variables that jointly affect both the explanatory and response variables [5,10,12,14–19]. If the DAG indicates that such variables exist, then further assumptions are required to identify whether the observed association between the explanatory and response variables reflects a causal effect. Other applications include reasoning about whether causal effects generalise from one target population to another [20,21] and for depicting SEMs [22,23].

In other disciplines, sampling (the way that units are selected for inclusion in the sample) and measurement (the way that data are collected for sampled units) are often incorporated explicitly in DAGs [11]. The same is not true in insect science and ecology, where DAGs are predominantly used to depict causal links between variables that exist in the world *before* any data have been collected. Table 1 demonstrates this point with a specific focus on applications in insect science.

Table 1. Applications of DAGs 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 (descriptive, predictive or causal), and the final column indicates whether observation processes (nodes for sample selection or measured proxies of relevant variables) were explicitly included in the DAG. The list only includes studies that we are aware of and is therefore unlikely to be exhaustive. Although we did not find any examples of DAGs being applied to predictive questions, the prospect was recently discussed by Priyadarshana and Slade [24].

Reference	Research question	Inferential goal	Observation processes included in the DAG?
Boyd et al. [25]	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. [22]	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. [23]	What are the direct and indirect effects of macrophyte management on aquatic insect abundance?	Causal	No
Gross et al. [26]	What is the effect of early-season insecticide application on secondary pest outbreaks?	Causal	No
Guzman et al. [27]	What is the effect of pesticide application	Causal	No

	(neonicotinoids and pyrethroids), animal-pollinated agriculture and honeybee colonies on wild bee occupancy?		
Proesmans et al. [28]	What is the effect of managed honeybees and a range of environmental drivers on viral prevalence in wild pollinators?	Causal	No
Saavedra et al. [29]	Propose the use of DAGs to increase our causal understanding of ecological systems.	Causal	No
Takeshita et al. [30]	What is the effect of management intervention in nickel concentrations on Ephemeroptera, Plecoptera, and Trichoptera richness?	Causal	No

*invertebrate, not insect

It is hard to justify incorporating assumptions about causal relationships in nature, but not about sampling and measurement (collectively the "observation process"), in DAGs. The former are required to identify causal effects when the available data are observational (i.e. do not come from a randomised experiment; [14]). 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 [31]. (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 DAGs 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 [32–34].

It is also worth noting that assumptions about the observation process are often needed to identify descriptive and predictive estimands. For example, to identify the proportion of one-kilometre squares occupied by the Purple Emperor in the UK, we have to 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, DAGs can be used to reason about whether non-causal estimands are identified [25,35].

Incorporating missing data and measurement error

We have now seen that, in insect science, DAGs are mostly used to reason about whether observed associations can be interpreted causally (Table 1). But sampling and measurement can themselves induce discrepancies between the associations we observe and the causal effect we seek [36]. Since both processes can readily be incorporated in DAGs, 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 DAG [37,38]. 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 [11,31].

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 [39]. 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 ([40]; the bottom left panel in Fig. 1 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 DAG (noting that we are not referring to the latent constructs often depicted in SEM diagrams). Each proxy is affected by the true underlying variable, 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 [41,42]. Fig. 1 demonstrates how one might include sample inclusion and measurement error in a DAG.

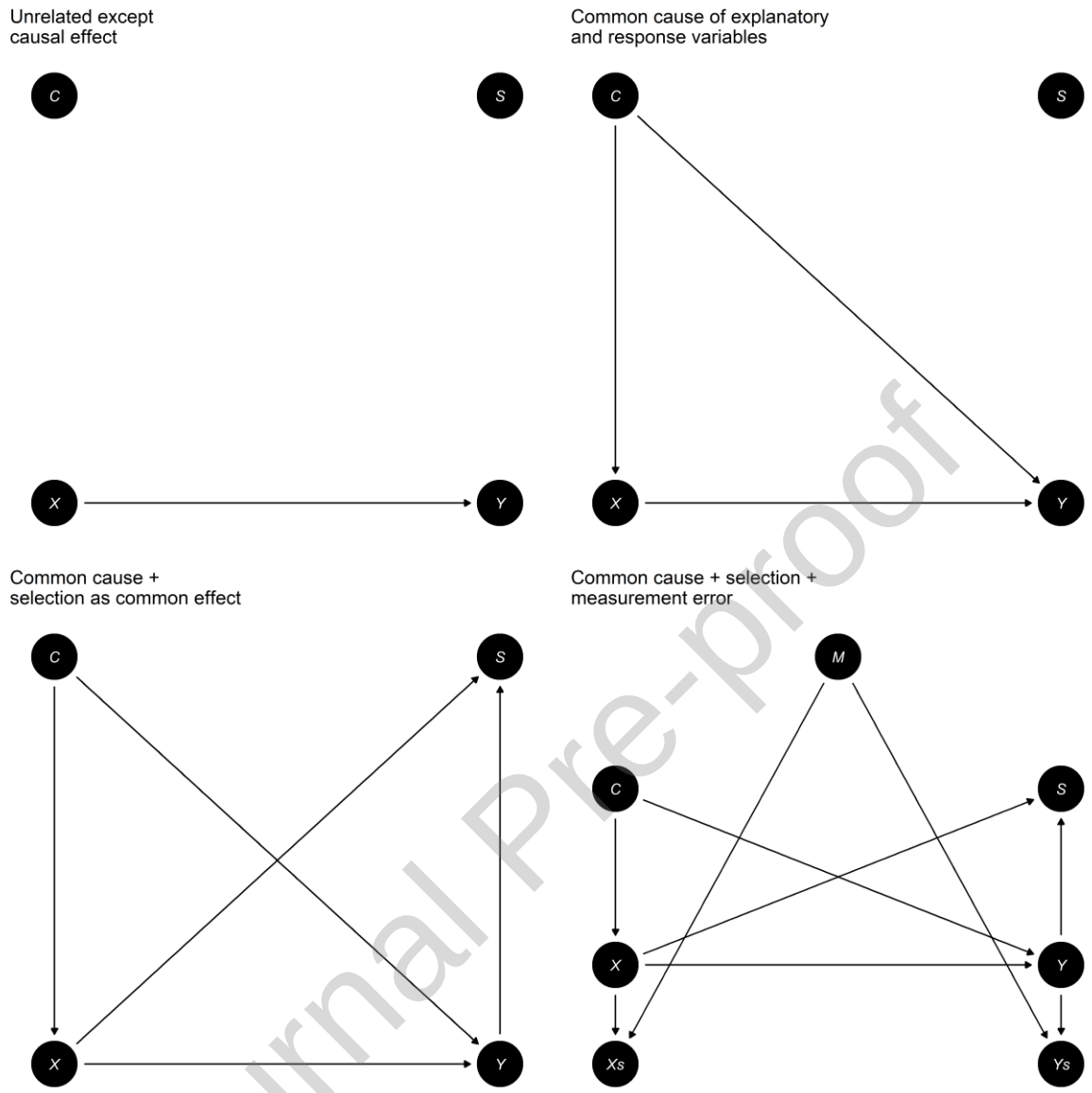


Figure 1. A stylized example depicting assumptions that might be made by a researcher seeking to estimate the average causal effect of plant cover (explanatory variable) on insect abundance (the response) across a target population of plots. s denotes plot inclusion in the sample (yes versus no), x denotes plant cover, y denotes insect abundance, c denotes temperature, X_s denotes the measured proxy for plant cover, Y_s 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 is 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). This measurement error structure renders the causal effect not identified.

DAGs for non-causal questions

Most of the literature on the use of DAGs for descriptive and predictive questions concerns assumptions about missing data *[43,44]. 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 [4,35].

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. 1.

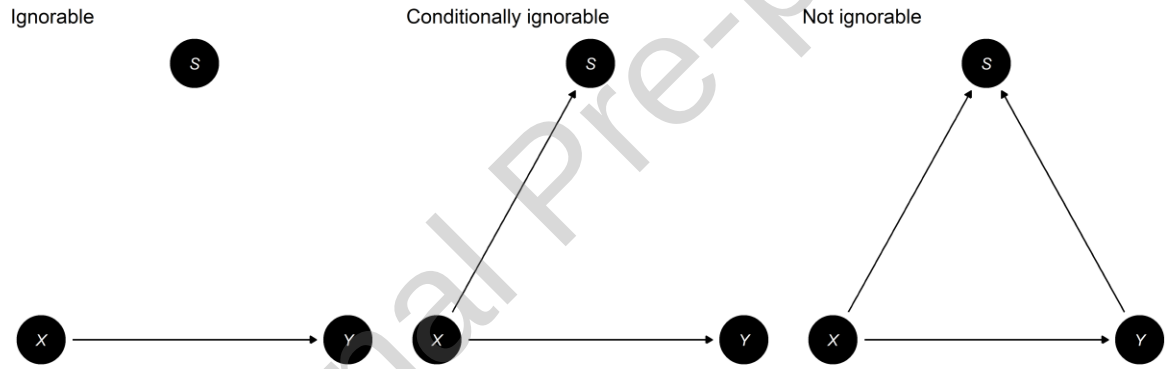


Figure 2. A stylized example depicting assumptions that might be made by a researcher seeking to estimate the mean insect abundance across plots (descriptive estimand) or its mean conditional on the level of plant cover (predictive estimand) in a target population of plots. S denotes plot inclusion in 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

If you are going to draw a DAG, 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 and whether they are plausible. A DAG that incorporates the observation process can also be used to reason about whether descriptive and predictive estimands are identified. In other disciplines, DAGs

form part of official reporting guidelines [45], and we believe that insect science would benefit from following suit.

Acknowledgements

RJB was supported by the NERC Independent Research Fellowship UKRI4865: A new paradigm for biodiversity monitoring. OLP was partially supported by NERC, through the UKCEH National Capability for UK Challenges Programme NE/Y006208/1, and by the Joint Nature Conservation Committee funding to the UK National Plant Monitoring Scheme.

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. Grace JB, Guntenspergen GR, Buffington KJ, Neville JA, Thorne KM, Osland MJ, Martinez M, Carr JA, Willard DA: **Causal interpretations can be based on mechanistic knowledge**. *Journal of Ecology* 2025, **113**:3084–3098.

7. 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.

8. Boyd RJ, Powney GD, Pescott OL: **We need to talk about nonprobability samples**. *Trends Ecol Evol* 2023, **38**:521–531.

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

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

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

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

13. Kunicki ZJ, Smith ML, Murray EJ: **A Primer on Structural Equation Model Diagrams and Directed Acyclic Graphs: When and How to Use Each in Psychological and Epidemiological Research**. *Adv Methods Pract Psychol Sci* 2023, **6**.

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

15. 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.

16. 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**.

17. Arif S, Massey MDB: **Reducing bias in experimental ecology through directed acyclic graphs**. *Ecol Evol* 2023, **13**.

18. 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.

19. 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".

20. 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.

21. 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**.

22. 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.

23. 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**.

24. Priyadarshana TS, Slade EM: **Predictive entomology: A causal framework for detecting and attributing insect population change**. *Curr Opin Insect Sci* 2026, doi:10.1016/j.cois.2026.101538.

25. 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.

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

27. 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.

28. 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**.

29. 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**.

30. 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**.

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

32. 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.

33. 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**.

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

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

36. Spake R, Mcelreath R, Evans LC: **Deterministic by design: causal inference challenges for "biodiversity change" syntheses**. *ecoevorxiv* 2026, doi:https://doi.org/10.32942/X2V38F.

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

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

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

40. 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.

41. 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**.

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

43. 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.

44. 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.

45. 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**.

Declaration of Competing Interest

We have no interests to declare.