

Review

LIES of omission: complex observation processes in ecology

Fergus J. Chadwick ^{1,2,*} Daniel T. Haydon,¹ Dirk Husmeier,³ Otso Ovaskainen,⁴ and Jason Matthiopoulos¹

Advances in statistics mean that it is now possible to tackle increasingly sophisticated observation processes. The intricacies and ambitious scale of modern data collection techniques mean that this is now essential. Methodological research to make inference about the biological process while accounting for the observation process has expanded dramatically, but solutions are often presented in field-specific terms, limiting our ability to identify commonalities between methods. We suggest a typology of observation processes that could improve translation between fields and aid methodological synthesis. We propose the LIES framework (defining observation processes in terms of issues of Latency, Identifiability, Effort and Scale) and illustrate its use with both simple examples and more complex case studies.

Increasing complexity of observation processes in ecology

Modern ecologists are called upon to tackle crises in the environment, as well as deal with ongoing scientific tasks of data collection and analysis. Technological advances in our ability to collect and analyse observations should give us unparalleled capacity to address emerging crises, but, instead, we are frequently stymied by the overwhelming scope and complexity of analysing our ever-more complex and multifaceted data. Techniques for collecting data have become almost as complex as the underlying **biological processes** (see [Glossary](#)) we are trying to understand. Environmental DNA [1], remote sensing [2], biologging [3], and citizen science [4] all help get us closer to the spatial, temporal, and taxonomic coverage we need to meet contemporary ecological challenges. However, they also introduce complexities which need to be addressed through sophisticated statistical analyses that are often devised as dedicated solutions to particular data sets. Therefore, a counterpart analytical crisis results from the fact that statistical methods that pay proper attention to these difficulties can appear disconnected, and overly specialist. As a result, advanced methods are rarely shared between fields, leading to duplication of solutions and inhibiting us from identifying methodological gaps that could benefit many fields.

Current solutions to these crises are thin on the ground. As ecology transforms itself into a hard science [5], part of the solution is to encourage ecologists to become more quantitative [6,7]. Although statistical literacy is arguably higher than ever amongst applied ecologists, we must still rely on close collaborations between ecologists and statisticians for method development. Alternatively, the analytical crisis can be circumvented by relying more heavily on experimental design. Many classical statistical techniques were developed for **designed experiments**, involving careful controls of confounders, high numbers of replicates and unbiased measurements. Unfortunately, the nature and scale of ecological questions in the 21st century are not always amenable to experimental design. GPS-tagged animals do not remain within predefined study areas, citizen scientists reconcile their observation efforts with their day jobs and, crucially, there is no Latin square for climate change. The focus on experimental design and user-friendly statistical methods can lead

Highlights

In ecology, the observation process (how we collect data) can be as complex as the biological process we are investigating.

Failure to account for complex observation processes leads to uncertainty, biased inference and poor predictions, resulting in misleading research results.

Often, field scientists are best placed to describe observation problems that occur but are excluded from discussions about how to tackle these problems statistically.

Statisticians are often unaware of the nuances of observation processes leading to the problems being ignored, or tackled on a case-by-case basis.

We propose a typology of observation problems and inferential solutions, hence facilitating the linkages between field protocols and statistical treatments.

¹School of Biodiversity, One Health and Veterinary Medicine, University of Glasgow, Glasgow, G12 8QQ, UK

²Centre for Research Into Ecological and Environmental Monitoring, School of Mathematics and Statistics, University of St Andrews, St. Andrews, Scotland, UK

³School of Mathematics and Statistics, University of Glasgow, Glasgow, G12 8TA, UK

⁴Department of Biological and Environmental Science, P.O. Box 35 FI-40014, University of Jyväskylä, Jyväskylä, Finland

*Correspondence: fergusjchadwick@gmail.com (F.J. Chadwick).



researchers to make strong simplifying assumptions rather than rigorously tackle the more challenging features of their data, to analyse them as if they were gathered in a designed experiment, yielding conclusions that are neither robust nor reproducible (McElreath's statistical golems [8]).

Realistically, therefore, we generally cannot simplify the methods or the data needed. However, we believe that we can simplify **observation process** modelling by developing a shared **typology** of associated problems. Our typology will aim to: (i) aid communication between field scientists and statisticians; (ii) make it easier to navigate the complex literature on closely related problems and their solutions; and (iii) help identify methodological gaps for further research. It will achieve Aim 1 by providing a basis for discussion between the two disciplines, allowing problems to be elicited in a comprehensive way using a shared language. By its nature, a typology creates a set of axes onto which problems and their methods can be placed. These conceptual axes (Box 1), make it easier to identify closely related problems and alternative solutions (Aim 2), and thus, to explore different model types and make methodological synthesis easier. Methods occupying the same problem space can even be unified. Unification often leads to rapid progress as previously disjointed efforts become focused, techniques are shared, and crucial gaps identified. These leaps forward have been seen in the unification of biodiversity metrics [9], the illustration of Poisson point processes [10] as the underlying method in MaxEnt [11] and presence-only modelling [12], and the rebranding of a huge number of methods under the banner of 'hidden Markov models' [13]. Conversely, sparse areas in problem space show areas where new techniques are desperately needed (Aim 3). A successful typology, therefore, helps identify the observation processes at play, navigate the possible solutions, and direct methods development to where it will be most productive.

The LIES Framework

A shared typology for observation problems needs to meet the following criteria. It must be sufficient to describe all observation problems in ecology (Table 1). To make the typology efficient, the problem types must exist independently and be useable in combination to describe more complex problems; be understandable to field scientists and statisticians; and rooted in the existing methodological literature where possible. Finally, the framework will be most effective if it is widely adopted which requires friendly packaging.

Later, we define each of the concepts in non-technical language. We illustrate each with pure form motivating examples (Figure 1, Key figure) rooted in the statistical literature. We make these canonical examples simple but realistic and present the concepts using the moniker LIES (**latency, identifiability, effort, scaling**) of omission, reminding us that failure to model observation processes correctly is to risk dishonesty. Finally, we draw from publications across the literature to demonstrate that the framework can describe real-world observation problems as one or a combination of these four problem types (Box 1).

Latency – relevance of data collected

Motivation

Biological phenomena are often hard to observe directly. Sometimes this is due to practical constraints. It may be possible to weigh the dry biomass of an organism, however, it is often more feasible (and less destructive) to measure a related variable such as the dimensions of the organisms [14, 15]. In other cases, the phenomenon we are interested in is not directly observable, perhaps because it is conceptual in nature (e.g., ecosystem equilibrium or autocorrelation) or has ceased to be observable (e.g., historical species abundance). In such cases, we refer to the quantity of interest as a **latent variable**. To infer the unobservable, we need to infer it using the impacts of the latent variable on **observable variables**. For example, we cannot directly observe

Glossary

Biological process: target of inference or prediction for the ecologist, encompassing all topics of ecological study.

Designed experiment: experiment focusing on a particular relationship between response and explanatory variables, where as many as possible of the confounding (or nuisance) variables are kept constant.

Effort (issues of): amount and distribution of observations of the phenomenon of interest and the amount of information contained within (recording an organism to the genus-level represents a lower amount of effort than the species level).

Functional form: mathematical relationship between two variables. These can be simple transformations (such as a link-function in a generalised linear model) or a more complex modelled relationship that incorporate an element of randomness (e.g., a model that identifies clusters in data).

Generative model: a model that is meaningfully decomposed into interpretable parameters and submodels, and from which data can be simulated.

Identifiability (issues of): inability of a model to make unambiguous estimates of its constituent parameters and thus make precise and accurate inferences about the relationships between its components (due to parameter redundancy or insufficient signal in the data).

Latency (issues of): where some or all parts of the biological process are not directly observed, and thus inference must be made indirectly through its impact on observable parts of the system.

Latent state/variable: state or variable that is not directly observed but must be inferred using observable parts of the system (i.e., observed variables).

Observable state/variable: state or variable that is directly observed and measured, generally to substitute for or help infer a latent variable which may be hard to observe or is truly unobservable.

Non-transferability: situation where a model fits observations well in one context but predicts poorly in novel contexts. Common causes of non-transferability include under- or overfitting.

Observation processes: methods by which an ecological phenomenon is recorded as data and represented during data analysis.

Box 1. The LIES workflow in practice

Johnston *et al.* identified four key challenges in analysing citizen science data caused by observer behaviour (Table I) [89]. In this box, we show how the LIES framework can be used to identify whether any are promising targets for conceptual and methodological synthesis and how this might be done.

Categorise observation processes in terms of LIES and synthesise within problem

Spatial bias and reporting preferences

Using the LIES framework, we found commonalities between spatial bias and reporting preferences. Both are issues of heterogeneous effort (across space and taxonomy, respectively) and latency (treating frequency of observations as related to underlying abundance). Citizen scientists are motivated to record by convenience (site accessibility and ease-of-identification) and ecological interest (site biodiversity and species interest, e.g., rarity status). Convenience can sometimes be predicted using covariates as effort proxies. While this is often effective for spatial bias, reporting preferences are less predictable. Targeting ecologically interesting sites and species leads to identifiability problems in distinguishing between observation and biological processes.

Observer differences

Citizen scientists vary in skill so their effective effort in terms of information gathered differs. Observer-level random effects and skill-scores can be used to estimate effective effort but these methods also need to account for skill improving with experience. A common solution is to use time as a proxy for effort changing within an individual but records are increasingly anonymised to prevent mathematical identifiability of individual observers. In these cases, a latent variable of effort can be used instead to estimate the combined effective effort.

False-positive error

Species misclassifications, where Species A is observed but recorded as Species B, are common in citizen science data and can lead to practical identifiability problems when estimating species-habitat associations. If the species' habitats overlap, then the degree of association may be overstated for Species B. If the record is false positive and the species do not overlap at that location, the habitat association for Species B will be incorrect. Many methods have been developed for dealing with false-positive errors, but they often have mathematical identifiability issues due to equal likelihood support for the species being present-and-correctly-identified or absent-and-falsely-reported. Alternatively, we could frame the species' identity as a latent variable and infer the correct classification by, for example, linking with habitat data from systematic studies.

Find analogous problems in other areas

We have identified the most promising areas for joint method development (spatial bias and reporting preferences) and the existing methods for tackling them (using covariate proxies and latent states to estimate effective effort). The next step is to seek out analogous situations in other areas (both ecological and non-ecological). One way to do this would be to distil the identified problems into search terms for a literature review. If the LIES framework were widely used, this step would be simplified as methods would be pre-categorised. Instead, we need to think how to translate the types of problems we have found into useful search terms (Table II). Literature search protocols should be applied to ensure the search is exhaustive. We advocate refining the searches by promising other fields to achieve a more in-depth assessment of the methods being used (commonly used methods are not always the most promising).

Synthesise methods from analogous problems

Excluding ecological problems, a quick literature search suggests survey design in public health and economics suffer from similar problems, motivating a more refined literature search to identify methods used to solve survey-design problems. The suggested methods are data integration and adaptive survey design. These approaches are effective at tackling a wide range of observation problems and are discussed later.

Using the aforementioned search terms without excluding ecological problems, the dominant scenario to appear is distance sampling. In distance sampling, effort is split into two components: the area covered (i.e., the length or number of transects, confusingly, known within this field as effort) and the detectability function (how the ability of observers to see an animal decays with distance). These structures naturally map onto the citizen science problems we are unifying. Biases can be used in the estimation of either component. Area covered could be modelled as a function of spatial gradients and recent species sightings, and detectability could change with habitat or species.

Naturally, distance sampling data deviate from citizen science data. In citizen science data, we often do not know the transect taken by the observer. Distance sampling also makes core assumptions which may be violated for citizen science data. In Table III, we propose potential (untested) adaptations to these problems.

Proxies: observed state/variable that has a well-defined functional relationship with a latent state/variable.

Scaling (issues of): any discrepancy between the resolution or extent (of, e.g., space, time, or taxonomy) at which the data are collected, and the process of inferential interest occurs.

Sensitivity analysis: exploring the relationship between perturbations to a model's inputs and the consequent changes in its outputs.

Typology (of observation processes): in general, refers to the classification of observations (both at the level of data collection and data analysis) according to their characteristics. Here, it refers to a minimal set of characteristics that can be used to describe any observation problem.

The viability of this approach will depend heavily on how effectively the proposed solutions in Table III are. Adaptations to and violations of these assumptions mean the uncertainty in these parameters may be large and some bias will remain but because the distance sampling framework was designed for structured surveys integration of unbiased surveys is straightforward.

Use the LIES problem space to identify complementary data

While we often motivate observation problems using examples of already collected data, the LIES framework can also be used in the design of data collection by thinking about the LIES problem spaces.

Imagine that each of the four elements of LIES defines an axis in problem space. The problem space contains all possible observation problems and allows them to be related to each other. Each axis extends indefinitely (indicating ever more extreme problems in that space). The axes are not necessarily continuous and may be summaries of other problems (e.g., a practical identifiability problem and a mathematical identifiability problem may occur at the same point on the identifiability axis if the problems are similarly severe, even though they are qualitatively different).

In a perfect experiment, we happily exist at the origin of this problem space, with zero observation problems. While the perfect experiment is unachievable, we can use the LIES problem spaces to try to get as close to perfect as is achievable. We recommend using the LIES framework as a conversation tool for field scientists and statisticians when designing data collection protocols. The process should be iterative and precautionary.

When the experiment is underway, time can be taken to establish whether the protocol is effective in minimizing observation problems and, if not, whether the protocol can be adjusted to do so. This process is known as adaptive survey design and can follow much the same path as the original experimental design process, with the added benefit of testable data. For example, if a key area has not been surveyed, then effort may be redistributed to ensure it is captured, or if the data are noisier than expected, it may be necessary to increase sample size to achieve identifiability.

Establish a joint modelling framework

A natural extension to this line of thinking is data integration. In data integration, we look for data sources that occupy complementary parts of LIES problem spaces, which together can bring us closer to the origin. For example, using the framework outlined earlier, we might combine citizen science data that has issues of latency, identifiability, and effort, with professional data. The transect data does not have the same latency and identifiability problems, but because it is more expensive to conduct it has effort problems. Fortunately, the effort problem is complementary to the citizen science data. The spatial biases in the transect data are known and reporting preferences are standardized across observers.

Table I. Four key citizen scientist behaviour challenges categorised in terms of the LIES framework

Challenge	Latency	Identifiability	Effort	Scaling
Spatial bias	Moderate	Major	Major	None
Observer differences	None	Moderate	Major	None
Reporting preference	Moderate	Major	Major	None
False-positive errors	Minor	Major	None	None

Table II. Translating problems defined using LIES into more general search terms for a literature review^a

	Problem summary	Proposed Boolean search profile
Latency	Frequency of observations is a function of biological process but relationship is unknown.	('unknown' OR 'hidden' OR 'latent' OR 'confounded' OR 'biased' OR 'nonidentifiable' OR 'identifiability') AND ('observation process' OR 'effort' OR 'effort surface') AND ('econometrics' OR 'public health') NOT ('ecology' OR 'biodiversity')
Identifiability	Observations are confounded with process of interest.	
Effort	Observations are biased and may be predicted using proxies.	
Scale	Not applicable	

^aThis is intended as an illustrative and nonexhaustive way that one could find analogous problems and solutions in another field.

(continued on next page)

Table III. Assumptions of distance sampling, how they may be violated in citizen science data, and potential methods to address or ameliorate these violations

Assumption	Consequence of violation	Citizen science data violations	Potential solution
The transect line is known.	The location of the observer is unknown.	The route taken by the citizen scientist is generally not known.	Incorporate citizen scientist movement model. Beware this is likely to lead to high uncertainty due to identifiability issues between transect location (centroid of kernel) and detectability (kernel decay) function.
All animals on the transect line are detected.	Strong bias in model estimates.	Likely to depend on the skill of the observer.	Observer skill can be included using covariates or random effects. Integration with unbiased data source.
Animals are randomly and evenly distributed within transects.	Strong bias in model estimates if observations are not independent (e.g., if species move in flocks or family groups).	Routes taken by citizen scientists are unlikely to be independent so double counting is possible.	Incorporate citizen scientist movement model that can account for non-independence.
Animals do not move before detection.	Bias is generally negligible.	May depend on skill and practice of observer, (e.g., animals avoid noisy observer or are attracted to food or artificial mating calls).	Observer skill can be included using covariates or random effects. Integration with unbiased data source.
Measurements (angles and distances) to animals are exact.	Bias is negligible if error is random, but systemic error leads to moderate bias.	Linked to first violation, citizen scientists are unlikely to give distances so these must be inferred.	Incorporate citizen scientist movement model. Beware this is likely to lead to high uncertainty due to identifiability issues between transect location (centroid of kernel) and detectability (kernel decay) function.

ecological equilibrium but we might observe the direction and speed at which the ecosystem is moving towards or away from it [16,17]. Similarly, we cannot travel through time to see the abundance of historical species but their impacts might persist into the observable present day [18].

Table 1. A comprehensive typology^a

	Relevance of observation	Reliability of observation
Data collection	Latency – what do the variables collected mean biologically?	Effort – how completely, precisely, and accurately have the observations captured the whole biological process?
Data analysis	Scaling – what does the scale at which the parameter is estimated mean biologically?	Identifiability – how completely, precisely, and accurately has the parameter been captured?

^aObservation problems can be introduced during data collection or analysis. There are two sides to these problems: the relevance of the observation to the biological process and the reliability of the observations made. This motivates a (comprehensive) typology of four core concepts.

Key Figure

These panels illustrate how the four types of observation process can affect the same image

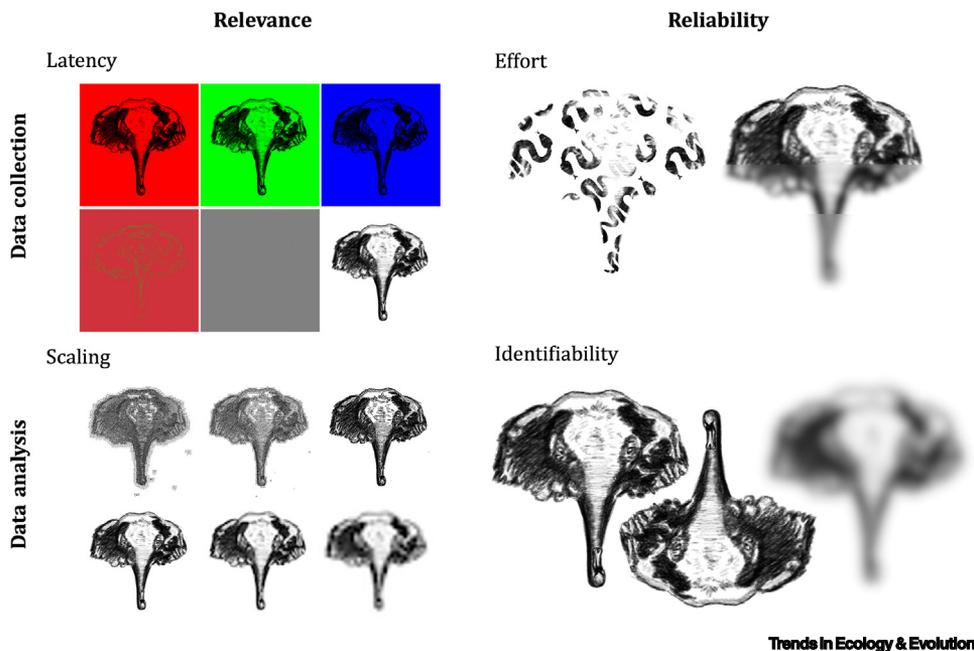


Figure 1. Latency: the image is recorded as six observed variables, namely the red, green, blue, hue, saturation, and light layers of the image. The latent variable, the image, can only be fully reconstructed by combining either the first or second row of images. Effort: the images are observed with heterogeneous effort. In the lefthand picture, the pattern of the effort is so strong it makes it hard to determine what the underlying surface looks like. In the righthand image there is an effort gradient from the top to the bottom of the image. Scaling: the process of data aggregation (to create a coarse scale) or disaggregation (to get a finer scale) is illustrated using pixels. Each pixel has a single value. Aggregating pixels requires averaging the values of the original pixels to create a single larger pixel. The averaging process can homogenise key details making the image hard to parse (bottom right). Disaggregating pixels generates smaller pixels whose average will be that of the original pixel, allowing noise to be introduced (top left). Identifiability: the image can be viewed as an elephant or a swan (first two pictures). When a parameter can take multiple equally plausible values, we have a mathematical identifiability issue. There can also be practical identifiability issues (third image), where uncertainty is so high that the truth may not be discernible.

Latency thus encompasses small to large degrees of discrepancy between the observable and latent variables.

It might be possible to find a closely related observable variable to act as a **proxy** for the latent variable. The key to successfully using a proxy variable is to acknowledge it is different to the latent variable by modelling the relationship between the two. The relationship may be linear and require a simple scale factor adjustment. There may be a known **functional form** that describes the relationship between the two. Functional forms are defined mathematically and can sound complicated, but they are actually often motivated from our biological understanding. For example, the trophic connection between predator and prey depends on their density and the shape of this dependence is informed by ecological experimentation and theory [19,20]. To quantify predator

intake, therefore, we do not simply use raw prey numbers, but we adjust the intake based on the density of the two groups.

Sometimes the situation is even more challenging and multiple observable variables are needed to infer a latent state. For example, to infer ecosystem stability, it is necessary to use the observed abundances from all the species in the ecosystem [21,22], and their lagged effects on one another [23]. Similarly, animal behaviour is often latent from us due to practical constraints (constant direct observation of individuals is extremely resource intensive). Instead, we rely on easily observable variables such as an animal's location through time (i.e., telemetry data) in combination with environmental covariates to infer behavioural states [24].

It can be tempting to assume the observable variables are equivalent to the latent variables to make modelling simpler. Unfortunately, in cases where the relationship between the two is not accounted for generally leads to poorer models that fail to capture the process, often leading to **non-transferability** [25–27].

Existing statistical methods

Entire textbooks are written on latent variable models [28–30]. The aim of latent variable methods in ecology is to map that which is easily observed into a biologically meaningful space. It is, therefore, useful to think of both latency and the models to tackle latency existing on a continuum. Proxies can be accommodated using the linear predictors and link functions in generalised linear models (GLMs). The coefficients can rescale proxies and link functions can approximate functional forms. For example, a quadratic term in the linear predictor can be used to represent intermediate optimum values while a logit-link function can accommodate saturation effects for binary outcomes.

While hidden states often require sophisticated modelling structures, it is useful to start from the simplest form: the generalised linear mixed effects model (GLMM) or hierarchical model. Random effect structures in GLMMs correspond to distributional assumptions about complex latent phenomena for different data groupings [25]. For example, site-level random effects are often used to estimate within-site variability caused by underlying processes such as site history and location. Stepping up in complexity slightly, multilevel hierarchical models (nested GLMMs) use information from different levels of the data to constrain the latent variable estimation while hidden Markov models use autocorrelation to reconstruct stochastic time series of hidden states [13].

The key to effectively tackling latency is to improve our biological understanding of the latent phenomenon [25,31]. Latent variables are often hardest to estimate and interpret when they are only weakly constrained by prior knowledge and model structure. By imposing boundaries informed by, for example, expert prior elicitation [32], we can often improve computation, inference, and model transferability.

Effort – reliability of data collected

Motivation

The aim of data collection is to try to gather information-rich observations of the biological process with the minimum bias and maximum precision [33]. A key tenet of traditional experimental design is to spread observation effort evenly among sampling units so that the observation process does not distort the underlying biological process [34]. Other than in highly controlled conditions, true homogeneous effort is almost impossible, leading to over-recording of some, for example, seasons, years, regions, individuals or population classes, and under-representation of others.

Uneven effort often arises from practical constraints. There are limits to where observers can be sent for safety reasons [35] or due to administrative boundaries [36]. Sometimes unevenness is deliberate. Data collected alongside a rabies vaccination campaign will generally be targeted towards rabies hotspots [37,38]. In these cases, stratified effort is uneven but its distribution is known and can be accounted for in the analysis.

The situation is more complex for opportunistically collected data. The distribution of citizen scientists (Box 1) [39], fisheries bycatch surveys [40], or deer–vehicle collisions [41] are all driven by processes that are rarely measured directly and are often driven by multiple other processes. Sometimes these drivers are spatial (e.g., deer–vehicle collisions depend on traffic flow; fishing boats minimise their travel time to fish stocks; and citizen scientists like to record in attractive locations near to where they live). There are also cultural drivers of what is reported – legal penalties reduce reporting of deer–vehicle collisions and fisheries bycatch, while citizen scientists more likely to report rare or invasive species. As a result, we frequently need to analyse data where the effort distribution is not only uneven but also unknown.

Existing statistical methods

Uneven effort can be accounted for statistically. In general, to retrieve the biological process from our data, we simply need to offset or reweight our observations per unit effort. We can think of this as an offsetting exercise [42]; however, first we need to quantify effort across different sample units. The challenge of this grows with the degree to which effort is unknown. Where effort is fully known, the offset can be incorporated into the model as data.

Where effort is in any way unknown, it must be inferred and the degree to which it is unknown determines the complexity of the modelling required to infer it [39]. Here, effort becomes a latent variable. In the earlier section, we discussed issues of latency between the observation process and the biological process. Here, we have latency between different parts of the observation process. We may be able to use similar modelling techniques to tackle latent effort. However, while we often have a good understanding of biological mechanisms with which to model latent biological processes, modelling latent effort requires an understanding of human behaviour.

In parallel with the methods to address issues of latency, there are three levels of complexity used when inferring effort. The first is to use a proxy variable for effort based on an assumed functional form. For example, in amateur wildlife recording, researchers use the frequency of a focal species or recorder's list length for a given site-visit as a measure of recording effort [39]; however, this makes strong assumptions about how the focal species and biodiversity are distributed. A relaxation of this relationship is to assume a particular functional form linking effort to the variable. Distance sampling (Box 1) is perhaps the most obvious use of this technique, where effort (the detection function) decays with distance from the observer [43].

The most complex method for inferring effort relies on multiple **observable variables** or known relationships. One approach is to use validation data collected with known effort. For the range of the validation data, the biological process is well characterised, meaning that differences in the overlap of the two data types can be attributed to (and used to model) effort. Most effort models use covariates to predict effort but some use properties of how effort is distributed, such as spatial autocorrelation [44] or phylogenetic information-sharing [45–47]. The best approach to modelling effort may need to be case specific and determined through model comparison [48,49] due to dependence on the inferential quantity of interest and amount of data available, as some elements of effort heterogeneity will play more important roles in some questions than others.

Scaling – relevance of inference made

Motivation

Determining the relevant scale for analysis is challenging and often overlooked [50–52]. For statistical models to be ecologically relevant, the signal detected needs to have a biological interpretation. Yet, frequently, our models are designed to look for signal in raw data where the scale is determined by equipment precision and encoding, leading to a discrepancy between the scale of inference and that of the biological process [53]. Indeed, the relevant scale may be variable-specific or a single variable may impact at multiple scales [54]. For example, phylogenetic distance may lead to trait autocorrelation at a large scale (organisms within an order are more similar than those in different orders) but negative correlation at a small scale (closely related species within a genus may be more different than more distantly related species in the genus). To reach the scale relevant to the biological process, our model needs to be able to change how neighbouring regions in the data are grouped together or divided. To do so, we need to understand what proximity means in variables like space, time, and taxonomy, and how these units can be sensibly aggregated (or disaggregated).

Proximity needs to be defined in biologically sensible ways that may be nonlinear and directional. For example, geographic proximity might be defined in terms of landscape resistance to a particular organism [55,56], but also in terms of that organism's mobility [57]. Temporal proximity may be determined by latitude with rapid seasonal changes in weather towards the poles and more smooth transitions in the tropics. Similarly, taxonomic proximity can be defined by a combination of morphometrics, genomics, and functional traits.

Aggregation operations often make an implicit mean-field assumption: that a system's behaviour is defined by the average value (e.g., of a covariate) across the system, so combining small units into larger units will lead to the same inference [58]. However, aggregation of fine-scale processes into coarser scale observations can eliminate our ability to detect signals [59]. A forager can be more efficient if all the prey in its home range is concentrated at one known location, and it may not matter if weather conditions are generally clement if a single storm can ruin a season's breeding [57]. In niche space, aggregating environments into coarse habitat classes might group together distinct habitats that are recognized very differently by a species [60]. Using fine-scale data may lose the signal by obscuring the environmental context within which the important biology is unfolding. Different biological processes may interact with the same covariates at different scales. For example, where a wolf moves in the next minute may be influenced by habitat composition within 200 m, whereas where a wolf establishes its home range may be influenced by habitat composition within 20 km.

Existing statistical methods

Both too much and too little aggregation can lead to discrepancies between our data and the biological process making us vulnerable to over- and underfitting issues [59]. Statistical diagnostics for these issues are common [61], but finding the appropriate scale is more challenging. One option is to fit models at multiple scales and compare using model-selection procedures [62]. A more sophisticated approach is to treat scale (or scales) as a parameter to be estimated [63] or to model scales hierarchically [64].

While conceptually simple, these approaches can be computationally prohibitive or limited by data availability. When aggregating at a particular scale, it is necessary to perform costly numerical integration for each candidate scale (although costs can be reduced using analytical tricks such as fast Fourier transformation algorithms [65]). Another common method is to use a distance decay kernel [66], such that distant observations bear lower importance. The scale

parameter is then the decay coefficient [67], [68]. Estimating nonlinear effects is becoming easier thanks to R packages like INLAbru [69], which extends fast approximate Bayesian methods [70] in a user-friendly way to accommodate more complex models.

Identifiability – reliability of inference made

Motivation

We build statistical models to identify relationships. The richness with which we can describe these relationships by our models will depend on the model definition. If the model is well defined [71] and the data contain sufficient signal, the parameters capture the real relationships and exclude alternative explanations. Advances in statistical computing have removed many constraints on model specification, making specifying interesting, biologically relevant, **generative models** easier. Even then not all relationships are identifiable by all models (mathematical identifiability) or without sufficient data (practical identifiability) [72].

Mathematical identifiability issues can arise in simple situations. A population's growth may be written unambiguously as a balance equation between birth and death rates, but even with unlimited data, it is impossible to estimate birth and death rates if both are unknown as there are infinite plausible combinations that are consistent with any growth rate. As model complexity grows, mathematical identifiability problems can be much more subtle (see discussion of multicollinearity in [73]).

Practical identifiability problems result from trying to make inferences from finite data. Even mathematically identifiable models may be unable to estimate relationships with precision if the noise-to-signal ratio is high, there is strong collinearity between covariates [73], or the model is only weakly identified [71]. Indeed, problems of latency, effort or scaling can contribute to practical identifiability issues. The severity of identifiability issues may depend on the model's purpose. For inference, identifiability is essential. For prediction, an individual parameter's identifiability may not matter so long as the overall effect is identifiable [74]. Similarly, a parameter might only be identified when normalised or transformed. For example, a covariance matrix may not be identifiable but the corresponding correlation matrix is [75].

Existing statistical methods

The relationship between the model definition and the quantity of interest defines both types of identifiability problem. We can think of the models working in two directions. In the forward direction, we simulate from the model. In the inverse direction, we estimate model parameters using data. We can use the forward direction to identify issues of mathematical identifiability by testing whether simulated quantities are affected by the specific model parameters [76]. If changing the model parameter values does not affect the quantities generated, there are mathematical identifiability issues. Once we have ruled out mathematical identifiability issues, we can explore the inverse. Here, we use data on the quantity of interest to estimate the model parameters. If many parameterisations are plausible given the data, we have high uncertainty and practical identifiability issues. Methods to assess these problems have been unified under the topics of **sensitivity analysis** [76] and uncertainty quantification [77,78], respectively.

Sensitivity analysis is solely a function of the model definition (i.e., is not affected by data), and can be conducted using directed acyclic graphs (DAGs; particularly popular in the causal inference literature [79]), inspection of the mathematical definition of a model [71], or simulation based methods [80]. Although mathematical identifiability problems are data-invariant, they are often found when fitting to data, for example, poor sampling in Markov chain Monte Carlo (MCMC)-based algorithms [81] or singularity in the Hessian matrix [71] is often indicative of mathematical identifiability problems in the model.

Uncertainty quantification depends on both the model definition and the noise-to-signal ratio in the data. If the model is over-parameterised or the data are uninformative, then there will be high uncertainty in the parameter estimates. Power analyses, in either Frequentist [82] or Bayesian [83] paradigms, often rely on using simulated data to estimate the nature and amount of data required to identify a relationship to a given precision. It is important to also assess how uncertainty changes under model misspecification (e.g., by using surrogate models for simulation) [84] as this is almost guaranteed. Model selection can also be a useful tool for comparing candidate models [85]. There are two main forms of model selection. Continuous model-space methods carry out variable selection parametrically as part of the model-fitting process, for example, penalised complexity [86] in the Bayesian paradigm and LASSO in the Frequentist [87]. Discrete model-space methods involve fitting candidate models independently and choosing a preferred model based on a separate metric to the fitting process (e.g., information criteria [85,88]). Continuous model-space approaches benefit from internal logical consistency but can be computationally burdensome and challenging to implement.

Concluding remarks

Field scientists and statisticians face an ongoing challenge of how to tackle urgent complex questions with complex data sources (see [Outstanding questions](#)). Eliciting the observation processes requires field science and statistical teams that work closely together and are motivated to understand one another. Where these teams do not exist, observation processes go unaccounted for, and any inference and policies made as a result are compromised. Where these teams succeed, they generate methodological advances, but advances which are often siloed due to field-specific language. Without breaking down these siloes, we stifle our progress. The typology we propose herein is one route through this impasse. However, we believe that it already offers a fresh perspective on observation processes that can lead to methodological synthesis, innovation, and insight as well as provide a mental roadmap through challenging terrain.

Acknowledgements

The manuscript would not have been possible without insights from Alison Johnston, Claire Harris, Crinan Jarrett, Dave Miller, David Pascall, Emilia Johnson, Emily Grace Simmonds, Grant Hopcraft, Halfan Ngowo, Jana Jeglinski, Katie Hampson, Luca Nelli, Megan Laxton, Rita Ribeiro, Simon Babayan, Tom Morrison, and Yacob Haddou. This work was completed as part of F.J.C.'s PhD funded by the Engineering and Physical Sciences Research Council (EPSRC) (EP/R513222/1) and the support of his subsequent employer, Biomathematics and Statistics Scotland (BioSS).

Declaration of interests

No interests are declared.

References

1. Cristescu, M.E. and Hebert, P.D. (2018) Uses and misuses of environmental DNA in biodiversity science and conservation. *Annu. Rev. Ecol. Evol. S* 49, 209–230
2. Cavender-Bares, J. *et al.* (2022) Integrating remote sensing with ecology and evolution to advance biodiversity conservation. *Nat. Ecol. Evol.* 6, 506–519
3. Sequeira, A.M. *et al.* (2021) A standardization framework for bioglogging data to advance ecological research and conservation. *Methods Ecol. Evol.* 12, 996–1007
4. Brown, E.D. and Williams, B.K. (2019) The potential for citizen science to produce reliable and useful information in ecology. *Conserv. Biol.* 33, 561–569
5. Platt, J.R. (1964) Strong inference: certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science* 146, 347–353
6. Clark, J.S. (2005) Why environmental scientists are becoming bayesians. *Ecol. Lett.* 8, 2–14
7. Ellison, A.M. and Dennis, B. (2010) Paths to statistical fluency for ecologists. *Front. Ecol. Environ.* 8, 362–370
8. McElreath, R. (2020) *Statistical Rethinking: a Bayesian Course with Examples in R and Stan*, Chapman and Hall/CRC
9. Chao, A. *et al.* (2014) Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through hill numbers. *Annu. Rev. Ecol. Evol. S* 45, 297–324
10. Aarts, G. *et al.* (2012) Comparative interpretation of count, presence-absence and point methods for species distribution models. *Methods Ecol. Evol.* 3, 177–187
11. Renner, I.W. and Warton, D.I. (2013) Equivalence of MAXENT and poisson point process models for species distribution modeling in ecology. *Biometrics* 69, 274–281
12. Renner, I.W. *et al.* (2015) Point process models for presence-only analysis. *Methods Ecol. Evol.* 6, 366–379
13. McClintock, B.T. *et al.* (2020) Uncovering ecological state dynamics with hidden Markov models. *Ecol. Lett.* 23, 1878–1903
14. Demol, M. *et al.* (2022) Estimating forest above-ground biomass with terrestrial laser scanning: current status and future directions. *Methods Ecol. Evol.* 13, 1628–1639

Outstanding questions

What data integration methods are missing? Many problems can be overcome by integrating complementary data types (e.g., combining fine scale data at a few locations with coarser data across a larger area to overcome issues of scale), however, the key is in identifying them.

How do we incorporate observation process modelling into teaching? Complex observation processes are rarely emphasised in statistics courses but most students will need to tackle them. Would emphasising the observation process guard against defaulting to interpretation of patterns in data as biological signal?

Can we link observation processes to experimental design techniques? How can simulating from models with observation processes improve data collection? Can we think of experimental design as a set of techniques to minimise identifiability issues while focusing effort on a small part of the biological process?

Can the LIES framework be utilised beyond the fields of ecology and evolution?

15. Schneider, S. *et al.* (2022) Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision. *Methods Ecol. Evol.* 13, 346–357
16. Scheffer, M. *et al.* (2001) Catastrophic shifts in ecosystems. *Nature* 413, 591–596
17. Bender, E.A. *et al.* (1984) Perturbation experiments in community ecology: Theory and practice. *Ecology* 65, 1–13
18. Royama, T. (2012) *Analytical Population Dynamics*, 10. Springer Science & Business Media
19. Murdoch, W.W. *et al.* (2013) *Consumer-Resource Dynamics*. (Monographs in Population Biology 36). Princeton University Press
20. Abrams, P.A. and Ginzburg, L.R. (2000) The nature of predation: prey dependent, ratio dependent or neither? *Trends Ecol. Evol.* 15, 337–341
21. Ovaskainen, O. *et al.* (2017) How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* 20, 561–576
22. Niku, J. *et al.* (2019) GLLVM: fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods Ecol. Evol.* 10, 2173–2182
23. Auger-Méthé, M. *et al.* (2021) A guide to state–space modeling of ecological time series. *Ecol. Monogr.* 91, e01470
24. Langrock, R. *et al.* (2012) Flexible and practical modeling of animal telemetry data: Hidden Markov models and extensions. *Ecology* 93, 2336–2342
25. Ives, A.R. (2022) Random errors are neither: on the interpretation of correlated data. *Methods Ecol. Evol.* 13, 2092–2105
26. Torney, C.J. *et al.* (2023) Estimating the abundance of a group-living species using multi-latent spatial models. *Methods Ecol. Evol.* 14, 77–86
27. Büscher, P. *et al.* (2018) Do cryptic reservoirs threaten gambiense-sleeping sickness elimination? *Trends Parasitol.* 34, 197–207
28. Beaujean, A.A. (2014) *Latent Variable Modeling Using R: a Step-by-Step Guide*, Routledge
29. Finch, W.H. and French, B.F. (2015) *Latent Variable Modeling with R*, Routledge
30. Loehlin, J.C. and Beaujean, A.A. (2017) *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis* (5th edn), Routledge
31. Stouffer, D.B. (2022) A critical examination of models of annual-plant population dynamics and density-dependent fecundity. *Methods Ecol. Evol.* 13, 2516–2530
32. Hemming, V. *et al.* (2020) Weighting and aggregating expert ecological judgments. *Ecol. Appl.* 30, e02075
33. Filazzola, A. and Cahill Jr., J.F. (2021) Replication in field ecology: identifying challenges and proposing solutions. *Methods Ecol. Evol.* 12, 1780–1792
34. Williams, B.K. and Brown, E.D. (2019) Sampling and analysis frameworks for inference in ecology. *Methods Ecol. Evol.* 10, 1832–1842
35. Demery, A.-J.C. and Pipkin, M.A. (2021) Safe fieldwork strategies for at-risk individuals, their supervisors and institutions. *Nat. Ecol. Evol.* 5, 5–9
36. Fent, A. *et al.* (2019) Transborder political ecology of mangroves in Senegal and the Gambia. *Global Environ. Chang.* 54, 214–226
37. Hudson, E.G. *et al.* (2019) Modelling targeted rabies vaccination strategies for a domestic dog population with heterogeneous roaming patterns. *PLOS Neglect. Trop. D.* 13, e0007582
38. Lugelo, A. *et al.* (2022) Development of dog vaccination strategies to maintain herd immunity against rabies. *Viruses* 14, 830
39. Isaac, N.J. *et al.* (2014) Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods Ecol. Evol.* 5, 1052–1060
40. Mendo, T. *et al.* (2022) Assessing discards in an illegal small-scale fishery using fisher-led reporting. *Rev. Fish Biol. Fish.* 32, 963–972
41. Nelli, L. *et al.* (2018) Mapping risk: quantifying and predicting the risk of deer-vehicle collisions on major roads in England. *Mamm. Biol.* 91, 71–78
42. Matthiopoulos, J. *et al.* (2022) Integrated modelling of seabird-habitat associations from multi-platform data: a review. *J. Appl. Ecol.* 59, 909–920
43. Buckland, S.T. *et al.* (2015) *Distance Sampling: Methods and Applications*, Vol. 431. Springer
44. Browning, E. *et al.* (2022) Accounting for spatial autocorrelation and environment are important to derive robust bat population trends from citizen science data. *Ecol. Indic.* 136, 108719
45. Kindsvater, H.K. *et al.* (2018) Overcoming the data crisis in biodiversity conservation. *Trends Ecol. Evol.* 33, 676–688
46. Johnson, T.F. *et al.* (2021) Handling missing values in trait data. *Glob. Ecol. Biogeogr.* 30, 51–62
47. Thorson, J.T. *et al.* (2023) Identifying direct and indirect associations among traits by merging phylogenetic comparative methods and structural equation models. *Methods Ecol. Evol.* 14, 1259–1275
48. Johnston, A. *et al.* (2018) Estimates of observer expertise improve species distributions from citizen science data. *Methods Ecol. Evol.* 9, 88–97
49. Johnston, A. *et al.* (2021) Analytical guidelines to increase the value of community science data: an example using eBird data to estimate species distributions. *Divers. Distrib.* 27, 1265–1277
50. Levin, S.A. (1992) The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* 73, 1943–1967
51. Chave, J. (2013) The problem of pattern and scale in ecology: what have we learned in 20 years? *Ecol. Lett.* 16, 4–16
52. Catford, J.A. *et al.* (2022) Addressing context dependence in ecology. *Trends Ecol. Evol.* 37, 158–170
53. Meyer, H. and Pebesma, E. (2021) Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* 12, 1620–1633
54. Glennie, R. *et al.* (2023) Hidden Markov models: pitfalls and opportunities in ecology. *Methods Ecol. Evol.* 14, 43–56
55. Unnithan Kumar, S. *et al.* (2022) Moving beyond landscape resistance: considerations for the future of connectivity modelling and conservation science. *Landscape Ecol.* 37, 2465–2480
56. Peterman, W.E. *et al.* (2019) A comparison of popular approaches to optimize landscape resistance surfaces. *Landscape Ecol.* 34, 2197–2208
57. Matthiopoulos, J. *et al.* (2020) Within reach? Habitat availability as a function of individual mobility and spatial structuring. *Am. Nat.* 195, 1009–1026
58. Wadoux, A.M.J.-C. and Heuvelink, G.B.M. (2023) Uncertainty of spatial averages and totals of natural resource maps. *Methods Ecol. Evol.* 14, 1320–1332
59. Paton, R.S. and Matthiopoulos, J. (2016) Defining the scale of habitat availability for models of habitat selection. *Ecology* 97, 1113–1122
60. Matthiopoulos, J. (2022) Defining, estimating, and understanding the fundamental niches of complex animals in heterogeneous environments. *Ecol. Monogr.* 92, e1545
61. Viana, D.S. *et al.* (2022) Disentangling spatial and environmental effects: Flexible methods for community ecology and macroecology. *Ecosphere* 13, e4028
62. Mancy, R. *et al.* (2022) Rabies shows how scale of transmission can enable acute infections to persist at low prevalence. *Science* 376, 512–516
63. Haddou, Y. *et al.* (2022) Widespread extinction debts and colonization credits in United States breeding bird communities. *Nat. Ecol. Evol.* 6, 324–331
64. Abrego, N. *et al.* (2022) Traits and phylogenies modulate the environmental responses of wood-inhabiting fungal communities across spatial scales. *J. Ecol.* 110, 784–798
65. McLoughlin, M.P. *et al.* (2019) Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. *J. R. Soc. Interface* 16, 20190225
66. Aue, B. *et al.* (2012) Distance weighting avoids erroneous scale effects in species-habitat models. *Methods Ecol. Evol.* 3, 102–111
67. Chandler, R. and Hepinstall-Cyerman, J. (2016) Estimating the spatial scales of landscape effects on abundance. *Landscape Ecol.* 31, 1383–1394
68. Carpentier, F. and Martin, O. (2021) Siland a R package for estimating the spatial influence of landscape. *Sci. Rep. UK* 11, 1–6
69. Bachi, F.E. *et al.* (2019) Inlabru: an R package for bayesian spatial modelling from ecological survey data. *Methods Ecol. Evol.* 10, 760–766

70. Lindgren, F. and Rue, H. (2015) Bayesian spatial modelling with r-INLA. *J. Stat. Softw.* 63, 1–25
71. Cole, D. (2020) *Parameter redundancy and Identifiability*, CRC Press
72. Ogle, K. and Barber, J.J. (2020) Ensuring identifiability in hierarchical mixed effects bayesian models. *Ecol. Appl.* 30, e02159
73. Dormann, C.F. *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27–46
74. Wieland, F.-G. *et al.* (2021) On structural and practical identifiability. *Curr. Opin. Sys. Biol.* 25, 60–69
75. Greene, W.H. (2011) *Econometric Analysis*, Prentice-Hall
76. Razavi, S. *et al.* (2021) The future of sensitivity analysis: an essential discipline for systems modeling and policy support. *Environ. Model. Softw.* 137, 104954
77. Soize, C. (2017) *Uncertainty Quantification*, Springer
78. Reimer, J.R. *et al.* (2022) Uncertainty quantification for ecological models with random parameters. *Ecol. Lett.* 25, 2232–2244
79. Laubach, Z.M. *et al.* (2021) A biologist's guide to model selection and causal inference. *P. R. Soc. B* 288, 20202815
80. DiRenzo, G.V. *et al.* (2023) A practical guide to understanding and validating complex models using data simulations. *Methods Ecol. Evol.* 14, 203–217
81. Gelman, A. *et al.* (2013) *Bayesian Data Analysis* (3rd edn), Chapman & Hall/CRC
82. Johnson, P.C. *et al.* (2015) Power analysis for generalized linear mixed models in ecology and evolution. *Methods Ecol. Evol.* 6, 133–142
83. Kruschke, J.K. and Liddell, T.M. (2018) The bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychon. B. Rev.* 25, 178–206
84. Gramacy, R.B. (2020) *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*, Chapman and Hall/CRC
85. Brewer, M.J. *et al.* (2016) The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods Ecol. Evol.* 7, 679–692
86. Simpson, D. *et al.* (2017) Penalising model component complexity: a principled, practical approach to constructing priors. *Stat. Sci.* 32, 1–28
87. Tredennick, A.T. *et al.* (2021) A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology* 102, e03336
88. Vehtari, A. *et al.* (2017) Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* 27, 1413–1432
89. Johnston, A. *et al.* (2023) Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods Ecol. Evol.* 14, 103–116