An introduction to model-based data integration for biodiversity assessments

F. Mancini, P.H. Boersch-Supan, R.A. Robinson, M. Harris & M.J.O. Pocock

JNCC
BTO
Inspired by birds, informed by science
UK Centre for Ecology & Hydrology

Contents

This publication should be cited as: Mancini, F., Boersch-Supan, P.H., Robinson, R.A., Harris, M. and Pocock, M.J.O. 2022. An introduction to model-based data integration for biodiversity assessments. JNCC, Peterborough.

Take home messages

- Model-based data integration is a statistical framework to combine the analysis of data from multiple sources to create a firmer evidence base on which to base decisions.
- Biodiversity data are usually fragmented in multiple datasets collected using a variety of different methods, which are difficult to combine without loss of information (e.g. count data and presence-only data) and which differ in their potential bias. Model-based data integration provides a solution to make the most of these multiple sources of data to produce robust metrics of biodiversity change.
- Model-based data integration has a number of analytical advantages, including increasing the quantity of data available to be included in analysis, deriving more precise metrics, extending the spatial and temporal extent of inference, and better correcting for biases in the data.
- By using model-based data integration we can make better inferences at smaller spatial scales and produce trends for scarce species. Data integration also creates a shared evidence base amongst conservation stakeholders, informs more efficient and flexible monitoring and can lead to a more diverse and inclusive recording community.
- There continue to be challenges and questions about best practices for model-based data integration and implementation still requires considerable technical skills and statistical knowledge. However, as the availability of novel data sources grows, model-based data integration will become more widespread amongst ecologists and user-friendly implementations are likely to become available.

Why should I be interested in model-based data integration?

Almost everywhere nature is under pressure. Thus, the need to monitor the state of nature and identify the many pressures affecting biodiversity has never been greater. However, the current range of biodiversity monitoring activities is varied and complex.

Technological advances have made it possible to collect new types of data on species distributions and abundance (e.g. acoustic devices, environmental DNA - eDNA).

An increasing number of biodiversity records are collected ‘opportunistically’ (i.e. in an unstructured manner, from wherever and whenever a recorder chooses) enhancing engagement with the natural world.

Data from standardised (structured) monitoring schemes are very useful but tend to be limited in their spatial, temporal and taxonomic coverage.

Additionally, some datasets cannot be clearly categorised as structured or unstructured, but fall somewhere along a “structure” gradient, for example surveys where recorders can choose where to look for species but follow a standardised protocol. Information on biodiversity (and other natural resources, such as habitat) is therefore often fragmented in different datasets and it can sometimes be hard to reconcile their different outputs. Model-based data integration is a statistical tool to combine these different sources of data to produce robust biodiversity assessments based on more of the available evidence.

How can model-based data integration help me?

Model-based data integration is an emerging statistical tool that enables multiple data sources to be combined. This provides a firmer and broader evidence base on which to base decisions.

In particular, different sources of species records can be brought together to produce more cohesive summaries of species' distributions in space and time. This type of model-based data integration uses integrated distribution models (IDMs). IDMs make the most of all available data by retaining the strengths of each data source and accounting for the differences between them in the analysis. This is valuable: for example, it can enable better tracking of scarce species (which might appear too rarely in a single dataset to provide a robust trend or map) and/or allow estimation of trends in smaller geographic areas than would otherwise be possible.

This guide is a non-technical introduction to model-based data integration and IDMs. We focus on models that integrate structured and unstructured biological records (Table 1) to derive temporal trends of species abundance or distribution, although this approach can work with other types of data too, such as museum records or novel data sources (e.g. eDNA, acoustic sensors). The aim of this guide is to present the concept of model-based data integration, give an introduction to the modelling framework and present the advantages derived from it through a series of case studies. Therefore, this is not a tutorial nor a practical guide but rather a general introduction to the opportunities of model-based data integration aimed at a non-technical audience. The reference list provides resources for the reader that wishes to learn more about developing and implementing these models.

Table 1. Definitions and benefits/disadvantages of the main types of data and modelling frameworks discussed in this guide.

	What is it?	Benefits	Disadvantages
Structured data	Data collected **with a standardised repeated protocol** (e.g. transects or quadrats) **and a sampling design**. For example, the UK Butterfly Monitoring Scheme[9,10] or Breeding Bird Survey[9,10]. This term is not used to describe the database format: the data are usually stored in a structured way in a database	• Observations are comparable in time and space • Site selection is usually randomised, so representative • All species observed are recorded as well as non-detections	• More demanding for volunteers • Generally lower spatial coverage and/or sampling intensity • Usually covers only one taxonomic group
Unstructured data	Data collected **without a standardised protocol or sampling design** or where the protocols are unknown. For example, species recording in iRecord[9,10],incidental records in BirdTrack[9,10] or museum records. This term is not used to describe the database format: the data are usually stored in a structured way in a database	• Produces large datasets • Broad geographic coverage • High taxonomic coverage • Long time series • Usually less demanding or time-consuming for volunteers/recorders	• Unknown whether all species observed are recorded • Non-detections are not recorded • Survey effort less standardised and/or unknown • Usually contain spatial and temporal bias
Semi-structured data	• Data collected **without a standardised protocol or sampling design, but where some information regarding the observation process is recorded**; for example complete lists in BirdTrack[9,10] • Data collected **with a standardised protocol but without a sampling design**, for example Flower Insect Timed Counts from the UK Pollinator Monitoring Scheme[9,10] This term is not used to describe the database format: the data are usually stored in a structured way in a database	• Effort is standardised by the protocol or quantifiable by the extra information (metadata) reported with the data • Observations are comparable in space and time	• Slightly more effort required by observers compared to unstructured recording • Requires specific data recording tools to capture the additional information on recording • Spatial and temporal sampling biases may be present

Table 1 *cont*. Definitions and benefits/disadvantages of the main types of data and modelling frameworks discussed in this guide.

	What is it?	Benefits	Disadvantages
Data merging	Merging different data types (e.g. unstructured and structured) in a single dataset by bringing all the data to a lowest common denominator (e.g. abundance data were degraded to presence/absence) and assuming the observation processes that generated the different data types are the same	• The size of the dataset is increased • More straightforward way to combine datasets that have a similar observation process (e.g. multiple unstructured datasets)	• Loss of information (i.e. structured sampling is treated as if it was unstructured recording) • Differences in the datasets are disregarded
Model-based data integration	**Combining** different data sources (e.g. unstructured and structured data) i**n a single statistical model by explicitly describing the differences** between the two datasets. When the model is focussed on data related to species distributions, we call it an i**ntegrated distribution model**	• The size of the dataset is increased • Combines the benefits of both unstructured and structured recording • Produces more precise estimates • Corrects for biases	• Technically challenging • Computationally intensive • Very new, therefore non-expert users would rely on contracting work as no user-friendly implementations are yet available

In the **first section** of this guide we describe the advantages of using model-based data integration and some of the potential impacts on biodiversity monitoring and conservation. In the **second section** we explain the modelling framework. We then present **three case studies** that show some of the advantages of using IDMs with different types of data. Finally we talk about different options for **implementation** of IDMs and present some **stakeholder views** on the value and opportunities of data integration.

Table 2. Glossary.

Bias	Systematic error causing a loss of accuracy, which is a persistent difference between an estimate and the underlying true parameter value.
Citizen science	A popular term for volunteer participation in scientific research and monitoring activities, sometimes also called community science. Levels of expertise may vary considerably between citizen scientists.
Credible interval	An interval within which a parameter value falls with a particular probability. Credible intervals in Bayesian statistics are analogous to confidence intervals in frequentist statistics, but relate to the data at hand rather than a notional population of means.
Detection probability	The probability of observing a species that is present at the site being surveyed. Detection is rarely perfect, so detection probability is usually <1
Hierarchical model	A hierarchical model is made of two sub-models that are linked, such as an observation and state sub-model. IDMs are a special type of hierarchical model.
Integrated distribution model (IDM)	A species distribution model where multiple data sources containing shared location information are combined using model-based data integration to account for their different data generation processes.
Integrated population model (IPM)	A model where multiple data sources containing information about population size and life-cycle stages are combined using model-based data integration to understand population changes; we do not focus on this type of integrated model in this guide but we include it in the glossary for completeness.
Latent state	An unobserved, and often unobservable, property of the ecological system that is of interest (e.g. the true species distribution or the actual number of individuals present).
Model-based data integration	The process of combining multiple datasets in a single statistical model while respecting and accounting for differences in data collection among datasets.
Observation sub-model	A statistical description of the way in which the data were collected, which can include known biases created by the observation process (e.g. observer behaviour) and ecological processes influencing detectability (e.g. seasonality).
Observation process	The way in which the dataset was collected.
Precision	A description of random error or statistical variability of an estimate. An estimate is precise if repeated measurements under unchanged conditions show very similar results.
Species distribution	The aggregated spatial locations of all individuals of a species across a geographic space.
Species distribution model	A statistical model that describes how the density of the locations of individuals of a species varies. A species distribution model can produce a static representation of a species distribution (the distribution of individuals at a single point in time), or a dynamic one 6 (the distribution of individuals as it changes across time); species distribution models can be applied to presence-only, presence/absence or abundance data
State sub-model	A statistical description of the underlying ecological processes governing the latent state

Advantages of model-based data integration

Model-based data integration offers several advantages over the analysis of structured or unstructured data separately and data merging.

- By using more sources of data, **it increases the quantity of data available**, thus making the most of all the information available.
- **It can 'borrow strengths' from multiple sources and/or types of data** that contribute to it, for example gaining from both the large spatial and temporal scale of large unstructured datasets and the robust design from the structured monitoring data, while at the same time correcting, to some extent, for the biases present in the data. Or an IDM can combine data with different modes of detection, e.g. from visual surveys and eDNA or audio recorders.
- **This can lead to reduced uncertainty and more precise estimates of species distributions**[7,8] **and trends**[9,10] and of the impact of potential drivers on species status or change[11].
- **Combining data can also allow better inference on 'hidden' parameters of interest** for example rates of site colonisation and extinction.
- **Integrating multiple sources of data can increase the extent of inference**, both in space, by using large scale unstructured datasets, and in time, for example by using museum records[12].
- **We can translate between ecological 'currencies'**, for example, combining presence-absence data with count data to make wide-scale inferences on population abundance[9,13].

Impacts of model-based data integration for biodiversity monitoring and assessment

More widespread use of model-based data integration methods has the potential to achieve many positive outcomes, including:

- **Local scale assessments of biodiversity trends may become possible**, thanks to the increase in the quantity of data available by integrating multiple sources.
- **Assessments of scarce and/or hard-to-detect species become possible** because more information is available.
- **Using data from multiple providers allows different stakeholders to work from a shared evidence base**, increasing trust in the model outputs and accelerating the translation of scientific evidence into conservation action.
- **All contributions by volunteers are valued and used in analysis**, because we no longer exclude opportunistic recording nor degrade structured sampling to the lowest common denominator.
- **We can design more flexible monitoring schemes for volunteers** because more people's contributions can be included, whether they can commit to contributing to structured monitoring or not.
- **The integrated modelling framework can inform the efficient planning and implementation of future monitoring** by answering questions such as: what is the minimum number of sites that need to be surveyed through a standardised monitoring protocol to add value to the unstructured data?; how much structured data is enough?; which sites should be prioritised for structured monitoring?

Box 1 - Imperfect ecological observations

A species record in a dataset is the product of two processes: firstly, the species needs to be present at the site and time of sampling, secondly the species needs to be detected and identified/recorded. Detection is rarely perfect, regardless of the data collection protocol (or lack thereof). In almost all cases not all individuals across sites will be counted, or all species seen, so detection probability is usually less than 1.

Ignoring imperfect detection can lead to inaccurate estimates of the latent state. For example, imagine that we could sample all the sites within the region of interest. If a species is actually present in 70% of the sites, but its detection probability is 60%, then ignoring the imperfect nature of detection will lead us to believe that the species is only present in 42% of the sites. Moreover, a species detection probability can be influenced by many factors: how visible a species is (a large, colourful or very vocal species is easier to detect), observer behaviour (personal preferences and experience, or time spent searching) or ecological variables (season or time of day). Ignoring these processes affecting detection may therefore bias our estimates about the true biological state.

In unstructured recording schemes, volunteer recorders choose which sites to visit – maybe preferring nature reserves, or places close to home. This may create a systematic overestimation of species occupancy or abundance by over-sampling sites where the species of interest is easier to record. Structured sampling can reduce this bias by selecting survey sites at random, so that both sites where the species is easy to detect and sites where occupancy or detectability are low are surveyed. Structured sampling also standardises effort by employing consistent sampling protocols. Because the probability of detecting a species increases with time spent searching, by using a standardised sampling protocol a species detection probability is kept constant across visits making them comparable.

Structured sampling can be more demanding for volunteers, who might be asked to survey a site that is difficult to reach, or where the taxon of interest is unlikely to be seen. Structured sampling might also require specialist equipment and/or ID skills. For these reasons structured sampling usually produces relatively small datasets that are limited in taxonomic, temporal and spatial coverage. On the other hand, unstructured monitoring tends to produce very large, but heterogeneous, datasets with higher spatial coverage, greater sampling intensity and longer time series. Model-based data integration makes it possible to benefit from the strengths of each data source, while controlling for their respective weaknesses.

What is model-based data integration?

Combining data from unstructured recording and structured monitoring schemes has, in the past, been very challenging, because most conventional modelling methods are designed to work with a single specific type of data (e.g. the MaxEnt procedure for presence-only data[14], or TRIM[15], for structured abundance data). Therefore, when faced with multiple datasets, ecologists typically had a choice between selecting only a single dataset for their analysis, analysing each dataset separately, or ignoring any differences between datasets and merging them into one single analysis.

- Dataset selection means discarding other datasets that are available, so the effort in collecting them and the information they can contribute are both wasted.
- Modelling each data source separately can often lead to different results and trying to reconcile these differences can be difficult.
- Data merging (Figure 1A) comes with a major caveat: it completely disregards any differences in the way the data were generated. If the datasets were very similar in the way they were collected, then data merging could be an acceptable option, however, with datasets that were generated through very different observation processes (e.g. structured and unstructured sampling) data merging would be a poor choice (Box 2). Similarly, data-merging is not well-suited when data are collected on different parts of the population (e.g. different life-stages).

The most effective way to combine datasets from different sources is through model-based data integration (Figure 1B). In this case, the two datasets are kept separate and the differences in the way the data were collected are explicitly described in two separate observation models (Box 2).

Box 2 - Problems with data merging and the benefit of IDMs

Data merging is a straightforward way to make use of all the available information on the ecological process of interest (e.g. a species distribution or abundance). It can, however, be problematic. For example, simply merging a dataset of structured sampling of amphibian counts, and structured sampling via environmental DNA sampling, will not take account of the very different observation processes and detectability in each dataset. Similarly, if we wanted to combine data from the National Biodiversity Network (NBN), a database composed largely of unstructured biological records, with data collected by a structured monitoring scheme (e.g. the UK Butterfly Monitoring Scheme - UKBMS), data merging would either ignore significant biases present in the NBN data, or assume that these biases are also present in the structured data. In both cases, the risk with data merging is that signals from the more representative structured data are masked by the biases in the larger unstructured dataset thus compromising the interpretability of the analysis output.

An IDM for the NBN + UKBMS data would include two observation models describing the observation processes that generated the two datasets. The observation model for the NBN data would account for its biases, for example it could include a spatial term to describe spatial bias in recorder density and a proxy variable for effort to account for the lack of a standardised protocol. On the other hand, the observation model for the UKBMS data would be much simpler, only needing to account for changes in detectability due to, for example, the ecology of the species (in this case seasonal flight time). These two observation models would be used together to get a better estimate the latent state of interest.

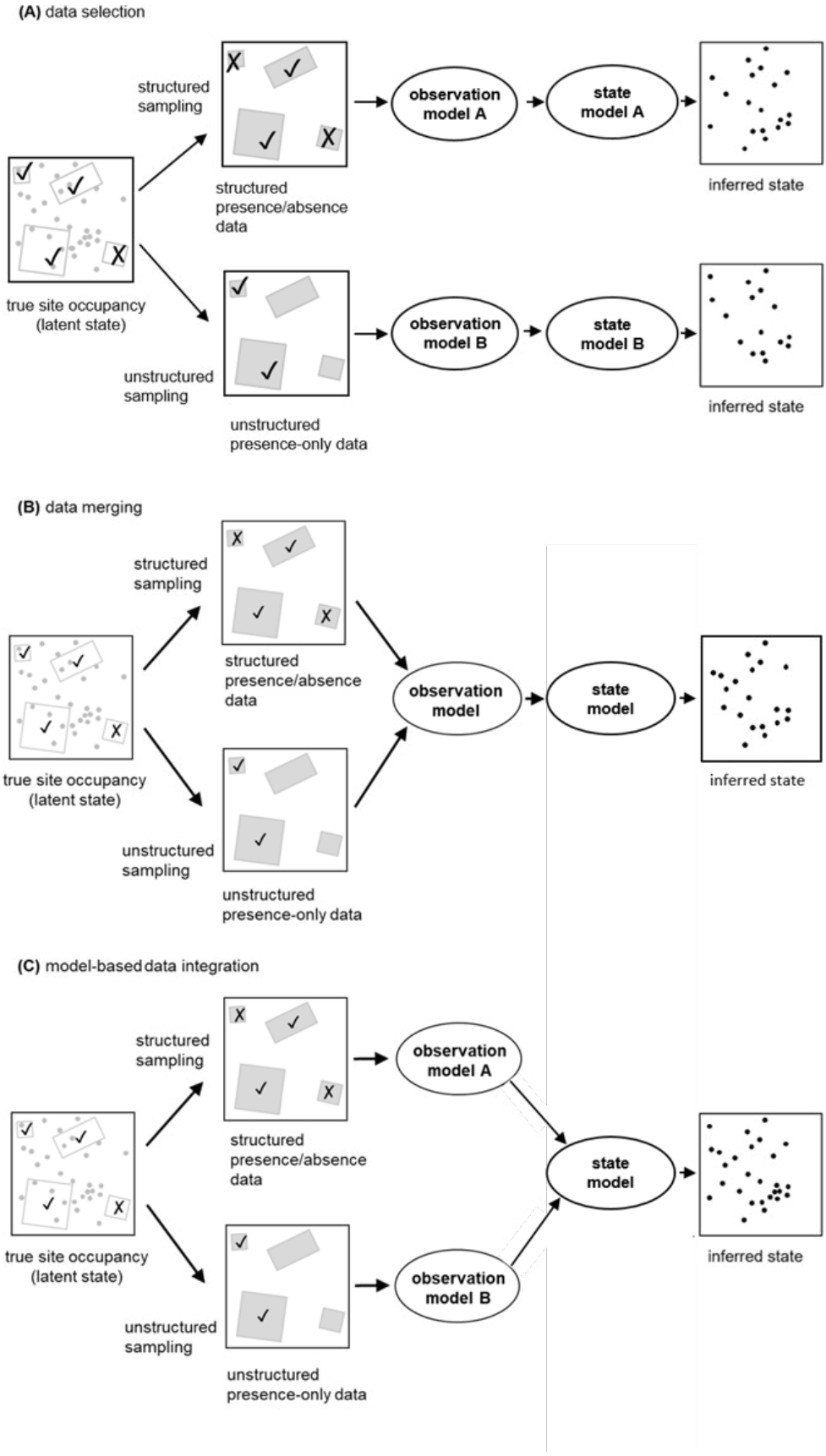


Figure 1. Data selection (A) and merging (B) vs. model-based data integration (C). The structured and unstructured datasets are independent realisations of the same underlying state, the species' true site occupancy. Data selection models one or both datasets separately, often leading to different inferences of the latent state. Data merging attempts to infer this latent state by merging the datasets and assuming the same observation process, while model-based data integration employs two separate observation models to explicitly describe the differences in the two sampling processes.

The modelling framework

Here, we explain in outline how IDMs are constructed. Firstly, we consider the analysis of a single dataset with what is called a 'hierarchical model'. This type of model produces estimates of the imperfectly observed latent state (i.e. the species' true abundance or occupancy – this is the thing we are really interested in), while taking account of the observation process (i.e. biases that are inevitable when making records of the natural world), which we are usually less interested in. The model is hierarchical because the observation process is layered on top of the underlying latent state and we estimate parameters in both sub-models simultaneously within the hierarchical model. In this section, we then show how to extend such a model to integrate multiple data sources to provide more robust inference about our population of interest.

A hierarchical model is made of two sub-models that are linked. Latent states are characterised by the state sub-model, which describes the ecological process that we believe to govern the true state of the target species (e.g. a species-habitat association). The latent state is linked to the observed data by a statistical description of the observation process that generated the data, the observation sub-model (Figure 2). This observation sub-model is designed to account for the fact that a species may be present but not seen and for different forms of bias in the data (Box 1). These hierarchical models have a long history in analysis of ecological survey data, with more recent applications to citizen science data, e.g. in the form of occupancy-detection models for presence-only data[16,17].

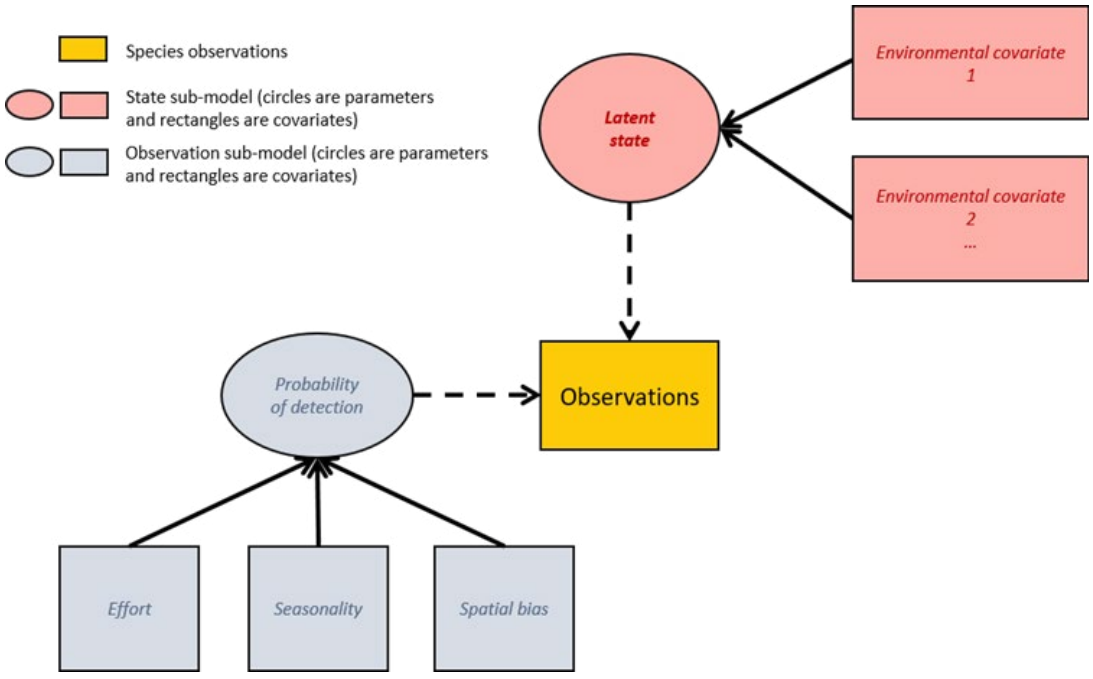


Figure 2. Schematic of an occupancy-detection model. The latent state, described by the latent sub-model in red, is the true presence/absence or abundance of the species at a site. The observation sub-model (in blue) describes the probability of detecting the species as a function of a variable explaining the spatial distribution of recording intensity, a proxy variable for effort and a covariate describing seasonality (e.g. Julian date or month).

Because this type of hierarchical model separates the observation and state process they are ideal for integrating different datasets. We can simply extend our basic model (Figure 2) from one to two data sources by adding a second observation sub-model, while the underlying state sub-model will be shared based on the assumption that the same population is being sampled. Each observation sub-model now describes the observation process that generated each dataset (Figure 3). The multiple datasets represent independent observations of the same latent state, for example a species presence record and a non-zero count are both conditional on the species being actually present at the site at the time in which it was surveyed. Therefore the state sub-model is shared and both observation sub-models are linked to it. Because the observation processes that generated the two datasets are different, their statistical descriptions (the observation sub-models) are different too. While the observation sub-model for the unstructured data (in blue in Figure 2 and Figure 3) includes terms to correct for spatial bias and uneven survey effort, the observation sub-model for the structured data (in green in Figure 3) does not contain these terms as these data were collected with a standardised protocol and, in this example, using a randomised sampling design that does not introduce spatial bias. One term that is common in both observation sub-models in this example is seasonality because the hypothetical species have seasonal behaviours that will influence their detectability regardless of the mode of data collection (e.g. bees flying in spring and summer or birds becoming more vocal during the breeding season). This is an example of how we can use the observation sub-model to better account for ecological processes that we think might influence or bias the observation process, which may be shared by both data sources.

In the next section we present three case studies that show some of the advantages of using the integrated modelling framework described above to analyse biological records data from multiple sources.

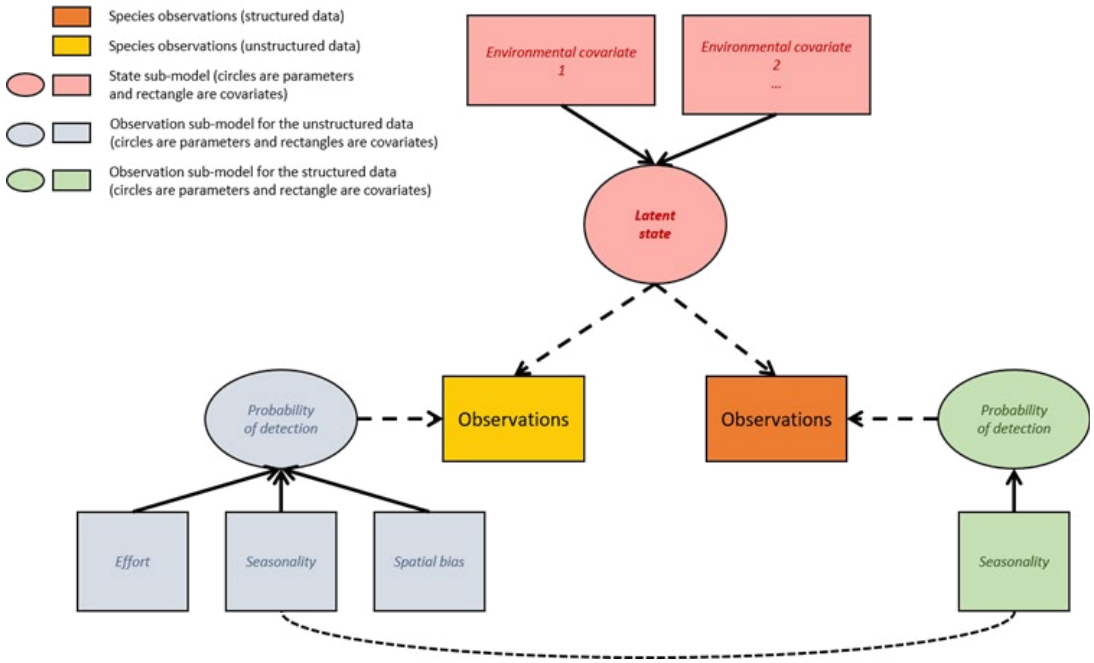


Figure 3. Schematic of an integrated model. The state sub-model in red and the observation sub-model in blue are the same as those in Figure 1. The shapes in green describe the observation sub-model for the second dataset. As in this example the second dataset (in orange) is collected through a standardised protocol, this observation sub-model estimates the probability of detecting the species as a function only of species phenology. The parameters describing this seasonality in the probability of detecting the species are shared between the two observation sub-models.

Case study 1: Improving small-area trends for an endangered farmland bird

National scale biodiversity monitoring schemes are designed to evaluate overall trends for a range of common species. They are often closely linked to conservation legislation and policy. However, the legislative and executive implementation of conservation policy is increasingly devolved within nations and the need to measure the effectiveness of landscape-scale conservation projects (which would be prohibitively expensive using traditional survey techniques) means there is increasingly desire to evaluate trends at smaller spatial scales.

In the UK, comprehensive structured bird monitoring is undertaken through the Breeding Bird Survey[2] (BBS). The BBS uses a strict observation protocol and a randomized sampling design which is stratified to account for different levels of volunteer availability across the UK. It provides national population trends for about 120 common and widespread bird species. Trends are not available for c. 100 rare and cryptic breeding species, and are not available for many species on regional spatial scales. The leading UK scheme for opportunistic citizen science bird recording is BirdTrack[4]. It provides greater coverage in space and time, but lacks the structured protocols and formal sampling design of the BBS. Observers choose when and where to record birds, but they can record effort metadata, such as the time spent compiling a list, or whether all detected species were recorded.

Based on both data sources, we show how population trends for the Corn Bunting *Miliaria calandra*, an endangered lowland farmland bird, can be derived at spatial scales smaller than a BBS stratum. To do this we make use of the overlap of BBS surveys and BirdTrack complete lists and use data from two areas to contrast different levels of recording coverage: the South Downs National Character Area (NCA) of southern England which has an expanse of c. 1,000 km^2, and a similar sized area largely dominated by arable farmland in North East Scotland. The South Downs are close to major population centres and are well covered by BirdTrack records from recreational birdwatchers; recording in North East Scotland occurs at much lower rates.

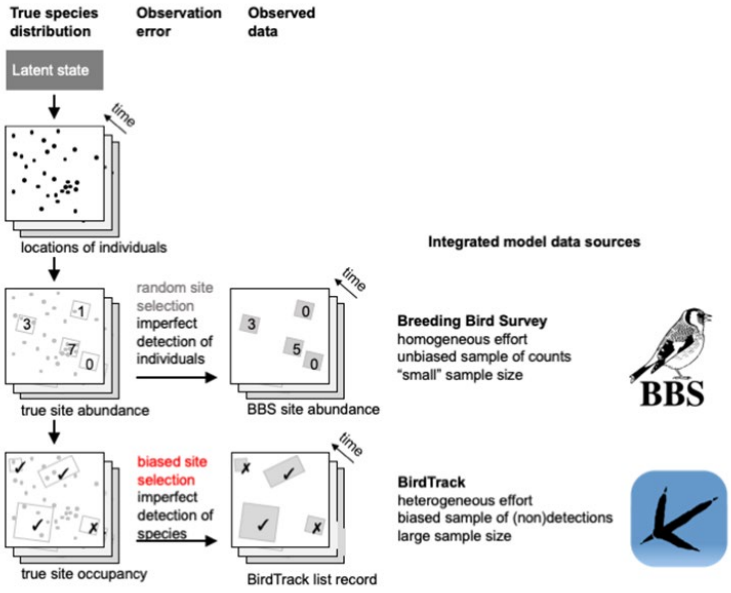


Figure 4. BBS surveys and BirdTrack lists are both observations of the true spatio-temporal distribution of birds. Observations from each scheme differ in their information quality and quantity. BBS counts are collected with known effort and spatially unbiased, but comparably sparse. BirdTrack lists are more numerous, but come from non-random locations and effort is heterogeneous. The IDM account for this by using a separate observation model for each source.

We produced standard BBS trend models and an integrated trend model (BBS + BirdTrack; Figure 4) for both study areas. However, because of the lower density of records in both schemes in North East Scotland, both trend models required the use of records from a larger area spanning Elgin to Peterhead, approximately three times the size of the South Downs. For both areas, the trend estimates from the integrated model did not differ substantially from abundance changes derived from BBS alone. All models for the South Downs NCA showed a decline between 2005 and 2011 followed by a period of relative stability (Figure 5), and the models for North East Scotland yielded highly uncertain abundance trends, which did not provide statistically significant evidence of change since 2005 (Figure 5).

The trends from the integrated model were more negative compared to BBS trends at both locations, but fell within the credible intervals of the BBS trend. In the South Downs the credible intervals for the integrated model were about five times more precise than those of the BBS trend (Figure 5). The integrated model suggests a significant decline of Corn Bunting abundance in the study area between 2005 and 2011. In contrast, models based on either dataset alone did not allow inferences about population change given the large uncertainty about annual index values. In Scotland the knowledge gains from data integration were more modest. The credible intervals of the integrated abundance trend were about half as wide as those for the BBS standard model (Figure 5).

Aside from the inferences about range and abundance changes the integrated model also provides estimates of detection parameters such as the influence of time spent recording on the probability of detecting a given species. This is a useful feature to assess the properties of opportunistic observations, and could e.g. inform minimum effort requirements to detect particular target species.

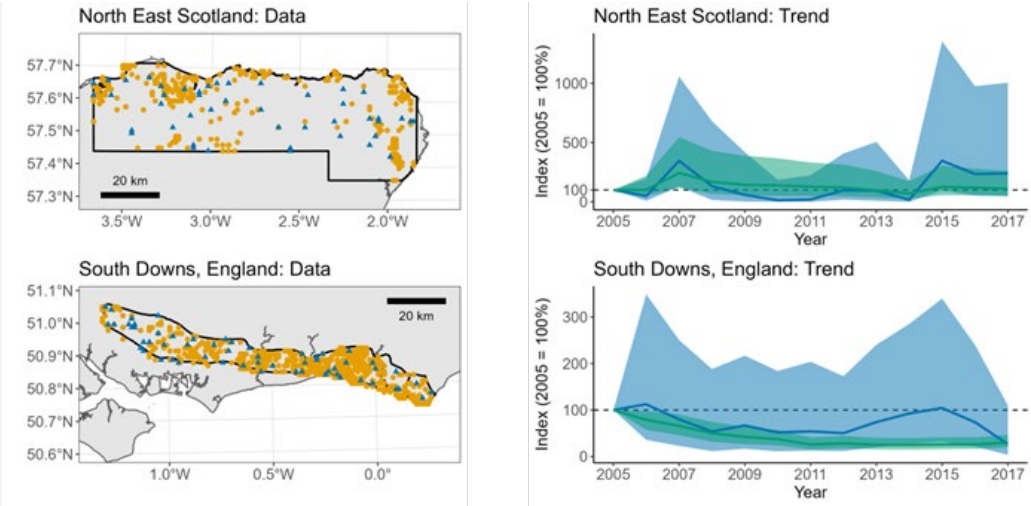


Figure 5. Data sources (blue = BBS locations; orange BirdTrack locations) and model results for the integrated abundance trend (green) for Corn Bunting in the South Downs and in North East Scotland, as compared to the BBS trend (blue). Dashed line shows relative abundance in 2005. Ribbons show posterior 95% credible intervals.

In this case study, we found that model-based data integration of structured and semi-structured bird data is feasible and offers potential to improve regional bird trend estimates, although a need to independently validate trends remains. Smaller gains are achieved in areas where uptake of recording is low. The integration of opportunistic records from volunteer-selected locations alone may therefore not adequately address monitoring gaps for management and policy applications. To achieve the latter, scheme organisers should consider providing incentives for achieving representative coverage of target areas in both structured and less structured recording schemes.

Modelling approach

We used a hierarchical model to integrate count data from BBS surveys and detection / non-detection data from BirdTrack complete lists. The state sub- model in this case describes species-specific abundances at a site in every year. We link this sub-model to the count and detection / non-detection data with a separate observation sub-model for each data source[11]. Standard BBS trends were also calculated for each study area as a comparison. All analyses were conducted in a Bayesian framework and a full description of the models and example code for parameter estimation are provided in[9].

Case study 2: When can model-based data integration provide the most benefits? A case study integrating two citizen science datasets on bumblebees in the UK

In this case study we compared the performance of data selection and data merging with the performance of an IDM in estimating trends in occupancy for bumblebee species in the UK. We used data from two citizen science schemes: the Bees Wasps and Ants Recording Society[18] (BWARS) and the BeeWalk[19] survey scheme from Bumblebee Conservation Trust (BCT). In the case of the BWARS data, details of the observation process, such as survey effort and whether all species observed were recorded, are not reported as standard practice so we would categorise these data as unstructured. On the other hand, the BeeWalk data are collected by volunteers using a standardised protocol: a fixed transect is walked at least four times a year and all individuals of each species of bumblebee observed are counted. The transect location, however, is chosen by the volunteer and therefore it does not follow any sampling design. Because of these characteristics, we would consider this dataset as semi-structured. We implemented four models: two single-dataset models (a BWARS-only and BeeWalks-only model), one merged model and one integrated model. All our models are variants on the occupancy-detection model described in Figure 2. The state (i.e. true bumblebee distribution) sub-model is identical for all model variants, but the observation sub-models differ to account for the different observation processes. We evaluate model performance by looking at the precision of the estimated state parameters. Precision is a measure of uncertainty with which the parameter is estimated, so that the higher the precision the lower the uncertainty.

Estimates of occupancy (proportion of sites occupied by each species) were generally similar across all model variants (Figure 6). The estimates from the integrated model were most similar to those from the merged variant (Figure 6, right), which is not surprising because all of the study sites and records are shared. The biggest difference in estimated values of occupancy is found in the comparison between the integrated model (the IDM) and the BeeWalk-only model (Figure 6, centre) reflecting the small number of sites and records in the BeeWalk data, which hence contribute less to the integrated model. The width of the 95% credible intervals (the uncertainty), shown in Figure 6 as vertical and horizontal lines, also show that measures from the BeeWalk-only model are generally less precise than those from the integrated model.

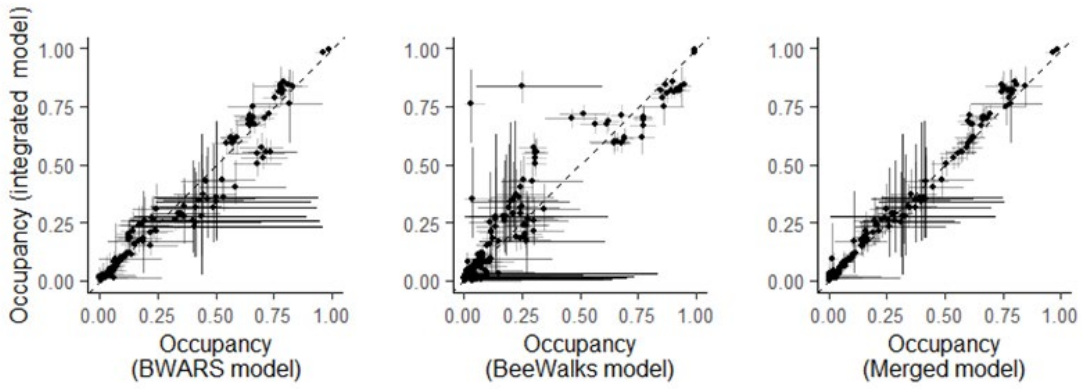


Figure 6. Comparison of estimates of annual occupancy between the four model variants. In each panel, the annual occupancy estimates for every bumblebee species from the integrated model (the IDM) are plotted on the y-axis; the x-axis shows annual occupancy estimates for the same species in the BWARS-only model (left), the BeeWalk-only model (centre) and the merged model (right). Each point represents a species:year combination. The vertical and horizontal lines are the 95% credible interval and the dashed line is the 1:1 line. The further away from the dashed line a point is, the more different the estimates of occupancy are from the integrated and non-integrated model.

The precision of the annual variability in the probability of occupancy (one of the state model parameters) was generally higher in the integrated model than in the BWARS-only and the BeeWalk-only models (reflecting larger sample size), but no substantial difference was found between the integrated and merged model (Figure 7).

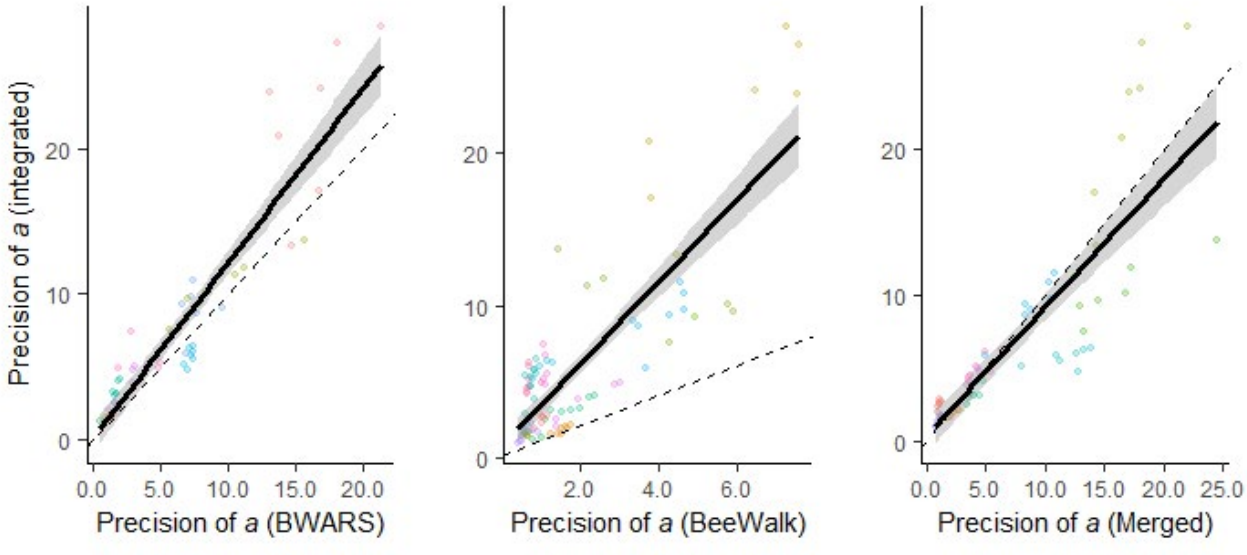


Figure 7. Comparison of precision of one of the state sub-model parameters from integrated model vs BWARS-only model (left), BeeWalk-only model (centre) and merged model (right). Each point represents the precision of an estimate of the temporal effect on probability of occupancy. The different colours represent different species and the black line and grey shaded area are the linear regression line and confidence intervals. The dashed line is the 1:1 line: every point above that line indicates that the parameter is estimated more precisely by the integrated model.

In this case study, we found that combining the two individual datasets in a single analysis had benefits for the precision of the parameter estimates, i.e. using both the BWARS and BeeWalk data together led to more precise estimates of trends in bumblebees.

As we have described, model-based data integration was expected to have benefits compared to data merging, but this was not the case here: the annual occupancy estimates of the merged model were just as precise, on average, as the IDM. So why did the IDM not provide additional benefits compared to data merging? One explanation for this result is that we were estimating occupancy, so treated the BeeWalk data as presence/absence data, thus discarding the BeeWalk information on abundance. This is the first iteration in developing an IDM for these datasets and the decision to degrade the BeeWalk data to presence/absence was made to keep the model simple in this first model development stage. If we had undertaken an analysis of abundance rather than occupancy, then the benefit of IDMs would have been clearer as we would have been able to estimate species abundance. Another explanation is that the selection of site locations did not follow a randomised sampling design for both the BeeWalk (semi-structured) and BWARS (unstructured) data (Figure 8), meaning that the datasets are quite similar in their observation processes, and so the merged model could estimate model parameters adequately. And thirdly, the datasets were very different in size: there were a huge number of records in the unstructured data (32594) compared to the much smaller number (6381) of semi-structured records from BeeWalks. It has been shown that large discrepancies in the size of datasets could lead to domination of the results by a single source and reduce any meaningful gain from integration[20]. In this case study, the information from the larger unstructured BWARS dataset may be swamping the information provided by the BeeWalk data. Weighting has been proposed as an approach to avoid this problem[21], however weighting is going to require some arbitrary choices and guidelines or best practices are not yet available.

This case study shows that model-based data integration may sometimes not be worth the effort it takes to design and implement the models, especially when the datasets are very similar in the way they were generated. Clear and extensive guidelines on when IDMs may provide the greatest benefits are not yet available, but improving our understanding of the limitations of IDMs is an active topic of research. A recent simulation study[20] has shown that integration alone was unable to correct for spatial bias in presence-only data, resulting in single-dataset models performing better than IDMs, while including a covariate to explain spatial bias and having larger structured datasets improved IDM performance.

Our results therefore provide a baseline assessment on the value of integration, rather than the final answer. These results also pose further questions:

- Can we use a better metric to evaluate model performance? A high precision makes a model useful as it can allow us to detect a trend where there is one. However, the model could be very precisely wrong. Model evaluation and metrics for occupancy-detection models are still an active area of research so this remains unknown for now.
- How much structured data is enough to add value to an unstructured dataset? Simulation studies or random sampling of real data followed by re-evaluation of the models would provide some insights. Answering this question will be important to provide guidelines on when model-based data integration can provide benefits over simple data merging, but it can also have more practical implications. It can guide the design and implementation of new monitoring schemes because we would be able to assess the minimum number of sites that need to be monitored using a structured protocol in order to add value to the many large unstructured datasets that are already available.

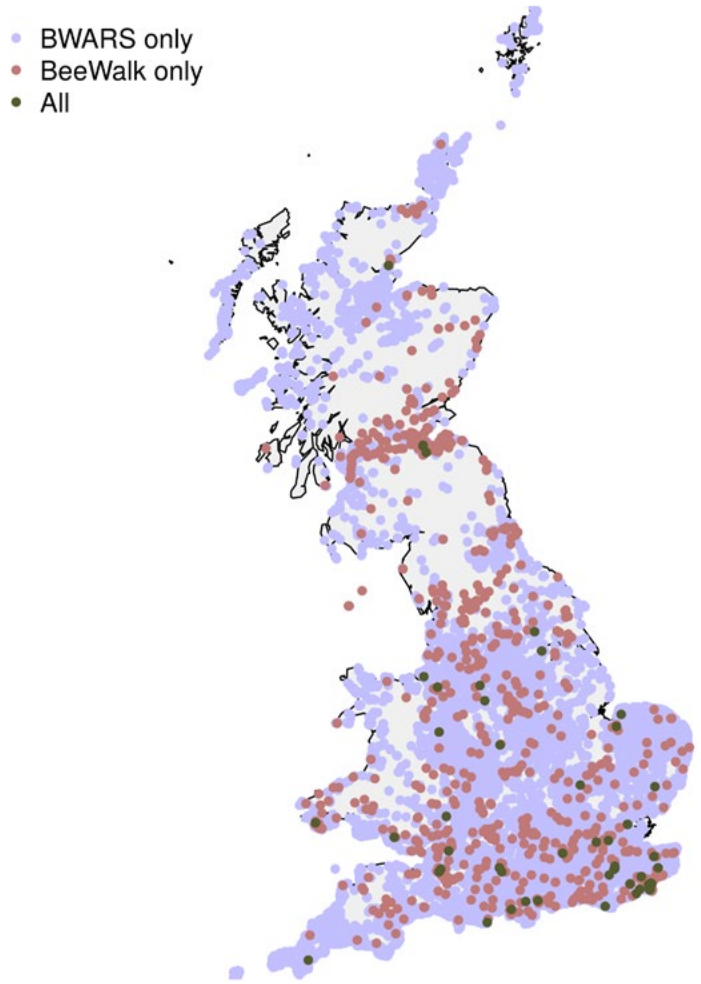


Figure 8. Spatial overlap between BWARS and BeeWalk bumblebee datasets. Points are 1 km sites surveyed by recorders.

Modelling approach

From both datasets we selected observations from the years 2010-2016, as both schemes were active during those years and data were available. In order to keep the model simple, we degraded the BeeWalk counts to presence-absence data (a full analysis would also make use of this extra information). Because the BWARS records are presence-only data, we inferred non-detections from records of other bee species in the same visit (combination of 1 km^2 site and date). We excluded records with a date precision lower than a day, a site precision of less than 1 km^2 and any species that did not have records in both datasets, resulting in 21 species being modelled. The BWARS data accounted for more than 90% of all the sites visited (11315) and more than 80% of the total records (38975). The BWARS data have a greater spatial coverage compared to the BeeWalk data, but there is some spatial overlap between the two sets of records (Figure 8) with 51 sites being surveyed by both schemes.

The four model variants shared the same state sub-model: the species is present (1) or absent (0) from a site with a probability (occupancy probability) that varies in time (yearly) and space (site). The observation sub-models differed between the model variants. The single-dataset variant for the BWARS data has to take into account the potential biases in the unstructured observation process, for example uneven recording effort across visits; therefore, the observation sub-model for this variant includes the number of species recorded in each visit (the list length) as a proxy for recorder effort. We additionally include a term to model the seasonal variation in detection as bumblebees are more likely to be observed in spring/summer than in autumn/winter. Our single-dataset model for the BeeWalk data is simpler, reflecting the fact that transect walks are a standardised protocol, such that effort can be assumed to be constant for all visits and so the observation sub-model for the BeeWalk data models detection as a function only of season. The integrated model variant includes two observation sub-models, one for each data type (Figure 3). The two observation sub-models are taken from the respective single-dataset variants. Note that the seasonal detection part of the two observation sub-models is identical and the parameters describing this seasonality are shared between the BWARS and BeeWalk observation sub-models. This feature allows information from one dataset to be shared with the other, and is one of the key features behind the power of model-based data integration. The final variant is the merged model, in which data from BWARS and BeeWalk are analysed together but ignoring the different data generation processes. In this model, we simply merged the BWARS and BeeWalk datasets and modelled the data with a single observation sub-model taken from the BWARS-only model. This is because in order to merge the two datasets together we had to degrade the BeeWalk data to the lowest common denominator and assume the same biases found in the BWARS data to be present in the semi-structured dataset. More detail can be found in[22].

Case study 3: Designing new monitoring schemes with data integration in mind

Structured monitoring schemes have been a great tool to provide excellent data for rigorous assessment of biodiversity trends. However, collecting these data requires a major commitment by volunteers. For example the UK Butterfly Monitoring Scheme asks recorders to visit a fixed transect weekly between April and September and count butterflies. However, this type of approach only works well for 'popular' taxa (i.e. those with many recorders) because only a small proportion of all the recorders will be willing or able to take on this task (e.g. over 100,000 people take part in the Big Butterfly Count[23], whereas only a couple of thousand people take part in the UK Butterfly Monitoring Scheme[1]), and we need a minimum number of sites to be monitored to provide statistically robust evidence of trends.

Many taxa have fewer dedicated recorders than 'popular' taxa like birds and butterflies, and their recording is overseen by a wide range of societies and recording schemes. Some of these societies have explored the possibility of developing structured monitoring schemes. Many people record dragonflies and damselflies and the records are verified and collated by the British Dragonfly Society (BDS). The BDS have recently published a State of Dragonflies in Britain and Ireland 2021[24], which assesses trends in these species based on occupancy analysis of unstructured data.

Several years ago, BDS trialled a structured recording scheme to assess changes in the abundance (not just the presence) of dragonflies and damselflies. This required regular visits to predetermined sites in suitable weather conditions to identify and count dragonflies. The results for individual sites were very valuable, but members of the BDS concluded that, at that time, this would not be a feasible monitoring scheme: there were not enough volunteers who were able and willing to undertake the monitoring at enough sites across the country, to provide data for statistically robust national-level results.

However, with IDMs, it would be possible to combine data from a small number of well-monitored sites with the larger amount of unstructured data, to provide a rigorous assessment of changes in the abundance of dragonflies. This evidence would enhance the conservation of these species and their use as bioindicators of water quality and climate change. Further investigation will help the BDS, and societies supporting the recording of other taxa, to assess the number of well-monitored sites that are required to provide data for effective IDM analysis. This approach to developing new monitoring schemes with data integration in mind has the potential to lead to a more efficient and effective use of new and existing datasets to understand how biodiversity is changing.

What do I need for model-based data integration?

Model-based data integration requires two or more biodiversity datasets that cover the species and region of interest. The datasets might have been generated through very different observation processes, but as long as the analyst can describe each dataset with its own observation sub-model, it does not matter how the data are categorised in terms of sampling structure (Table 1).

So an IDM could be used to integrate two or more differently-structured datasets, two or more unstructured datasets or a mixture of structured and unstructured datasets, but each dataset will have its own observation sub-model. Usually we want the datasets to overlap in space and/or time to ensure the assumption of modelling the same population holds, thus allowing us to share information within the model.

It is also important to carefully consider what each dataset can contribute to integration. For example, when combining a structured and unstructured dataset that do overlap in space and time we can use detectability information from the structured dataset to estimate detection biases in the unstructured dataset. Additionally, auxiliary data[25] or data filtering[26] can be used to further reduce bias in individual unstructured datasets so that they can be integrated in a model.

The less datasets overlap, the more important it is to have as much information about the recording process as possible. For example data from a lowland bird survey can likely be combined with an upland bird survey if both are reasonably structured and collect detectability information. If only one of the surveys collects detectability information and there is little overlap between surveys, then it will be impossible for the model to distinguish whether any differences in the estimated latent states will be due to habitat type or differences in the surveys.

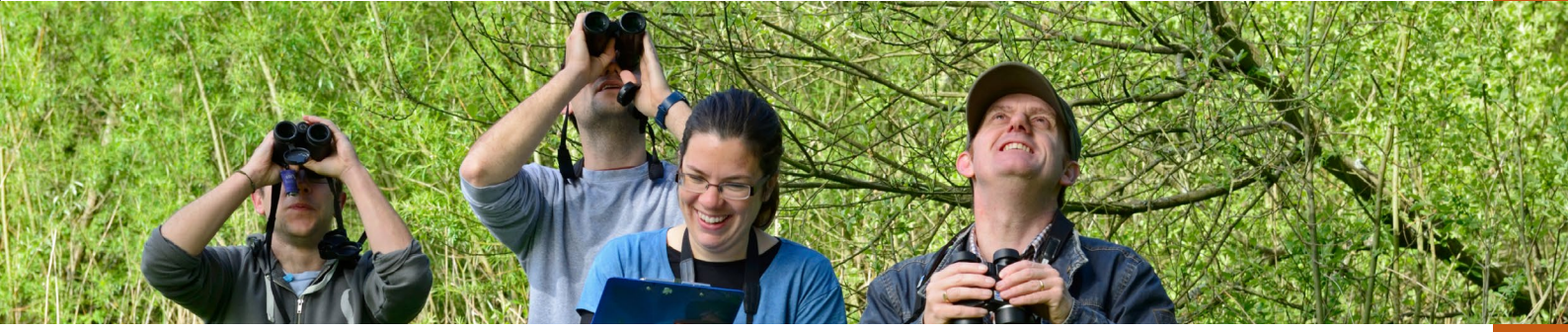
It is important to notice that, although the occupancy-detection models have a hierarchical structure and the observation sub-model provides a mechanism to model the biases that might arise from the observation process, this not in itself a guarantee that those biases will be accounted for[17]. When the observation sub-models are specified, they need to follow careful consideration of the potential biases present in the datasets. This should lead to the inclusion of variables that are relevant to model the ecological and observation processes that might affect the detectability of a species, or to limiting the inference to the extent (spatial, temporal, taxonomic or environmental) at which the datasets can be considered a representative sample of the population of interest. There are now tools available to both explore potential biases in the data[27] and transparently assess the risks from these biases when drawing inferences on species distributions[28].

Implementation: technical considerations

The IDMs described here are typically implemented in a Bayesian statistical framework simply because of its flexibility. Models can be developed using the BUGS language, which is flexible enough to express the hierarchical structures needed and facilitate the inclusion of one latent state linked to multiple observation models. Markov Chain Monte Carlo (MCMC) software can be used to fit these models, such as WinBUGS[29], JAGS[30], or Nimble[31], all of which can be used within the statistical software R[32]. New MCMC software such as Stan[33] and greta[34] are also an option and they can run much faster.

All of these implementations are best suited when space can be represented in discrete units, for example on the British National Grid. If data needs to be modelled in continuous space then software such as INLA[35,36] can be used, which makes it possible to fit many complex ecological models efficiently using point processes[7]. There is however a trade-off with flexibility in terms of the range of models that can be implemented.

Regardless of the software chosen, it is challenging to set the models up, requiring programming skills and statistical expertise, and they often require high performance computing to run. User-friendly implementations have not yet been developed as the framework is still in active development. As this is a new approach, there are still a lot of questions to be answered and best practices are not necessarily clear. We are still discovering how to run models efficiently and how to best evaluate model output. As a consequence, caution (and input from a statistician) is required when developing these models and interpreting the results.



Implementation: stakeholder views

We ran a workshop as part of the UK Terrestrial Evidence Partnership of Partnerships (UKTEPoP) Festival 2021. The aim of the workshop was to introduce the concept of model-based data integration, present the case studies and collect feedback and perspectives from potential users of IDMs. This section will summarise some of the stakeholders' views and reflections from the discussion sessions of the workshop.

Thirty-nine people attended the workshop, the majority of whom indicated working for a Government Agency (44%) or a conservation charity (33%). The rest of the attendees indicated working for a recording scheme or society, a data holding organisation or a research institute (8% each). The majority of attendees described themselves as a data analyst (67%), but many were also subject specialists (38%), project managers (36%), scheme organisers (26%) or team leaders (10%). These responses suggest that we had a diverse audience of potential end users of IDMs, as well as an audience who was mostly knowledgeable and skilled in data analysis.

The most used type of biodiversity data among the attendees was volunteer collected and structured data, although professionally collected, semi-structured and unstructured data were also used by many of the attendees. Participants indicated that the data are mostly used to model temporal trends and the effects of environmental drivers. We also asked what types of data, if any, are underutilised, which revealed that "unstructured data" is the least used data type, followed by citizen science, habitat and marine data. The most common words used to describe why the data is not used were: "complexity", "subjective" and "interpretation". This suggests that the analytical complexities of dealing with data that does not contain any information on the data collection procedure result in the loss of potentially valuable information from further analysis.

We asked the stakeholders whether they could see the value of applying the IDM framework to their data. Respondents highlighted how IDMs could help make the most of available data sources that are currently underused, especially ad-hoc and historical data, and to produce a coherent narrative on the state of biodiversity from all the available information, especially when this information is fragmented into different and usually small datasets. Another advantage of using IDMs highlighted by the stakeholders was that it would be able to combine surveys at different scales, for example national and local scale surveys, to derive better inferences at smaller scales.

Stakeholders also recognised the value of IDMs in informing the development of new monitoring schemes, for example by understanding gaps in available datasets that can be filled by opportunistic data or new structured monitoring, especially for identifying areas to survey for rare species. It was also mentioned that the IDM framework could help redesigning and improving existing surveys to reach a good balance of quality versus quantity of data.

Stakeholders also commented on the potential for IDMs to contribute to diversifying recording and making it more inclusive: different survey types may appeal to different audiences with different abilities or commitment levels, so designing monitoring with data integration in mind can help reach a bigger and more diverse audience. Analysing citizen science data within an integrated modelling framework can therefore increase volunteer motivation as all the data that is collected will be valued and used to produce outputs. Finally, stakeholders highlighted the collaborative aspect of bringing different data sources together in an IDM framework. This would result in the co-production of knowledge and, in turn, facilitate collaboration and reduce conflicts between different stakeholders. This can be particularly helpful in cases of conservation conflicts, where different surveys can be set up to be conducted by the different actors involved in the conflict and then the data brought together to generated evidence that is more likely to be accepted by all parties.

Conclusions

IDMs are a novel statistical tool that allows the use of multiple datasets to estimate biodiversity change from multiple sources of evidence. The power of this modelling framework is best realised when combining datasets where the observation models are too different to permit simple data merging without losing substantial information, for example count and presence-only data. There have been considerable advances in model-based data integration recently, both conceptually and in implementation[7], and as the availability of new data types increases, IDMs will become a valuable approach for many ecological modelling applications.

There are still many challenges to overcome before this framework becomes common practice amongst ecologists. One challenge is to quantify the information gained by data integration and the information that each data source can contribute to the inference. Progress in this area will inform guidelines or rules-of-thumb on when data integration is most useful. Another challenge is the validation of IDMs. Evaluating the performance of IDMs when using real data is an active topic of research and a challenge that, when overcome, will help guide the practice and implementation of IDMs.

Given the novelty of IDMs it is not surprising that the learning curve to develop and implement these models is still quite steep. But as opportunities for data integration grow and more research is done to overcome some of the challenges described above, more training, documentation and user-friendly interfaces will be developed, lowering the barriers to the use of IDMs amongst ecologists.

Model-based data integration also provides new opportunities to shape the future monitoring of biodiversity. Despite the critical need for evidence in the face of an ongoing biodiversity crisis, large-scale biodiversity monitoring is costly and very few countries are able to implement unified standardised biodiversity monitoring programmes of different taxa at a national scale. On the other hand, there are a wide range of grass-root initiatives to monitor biodiversity at different scales and a culture of integration can help to address our large scale monitoring needs. Designing new monitoring schemes with data integration in mind means that we will be better able to cost-effectively fill the gaps in current biodiversity recording, produce high quality datasets that can add value to those already available and engage a more diverse recording community. Integrated monitoring will create a network of stakeholders around biodiversity monitoring, who will work from a shared evidence base, creating a feeling of ownership and trust, which accelerates the translation of conservation evidence into action.

References

1. UK Butterfly Monitoring Scheme. UK Butterfly Monitoring Scheme. https://ukbms.org/.
2. British Trust for Ornithology. Breeding Bird Survey. https://www.bto.org/our-science/projects/bbs.
3. Biological Records Centre. iRecord. https://www.brc.ac.uk/irecord/.
4. British Trust for Ornithology. BirdTrack. https://bto.org/our-science/projects/birdtrack.
5. UK Pollinator Monitoring Scheme. UK Pollinator Monitoring Scheme. https://ukpoms.org.uk/.
6. Milanesi, P., Della Rocca, F. & Robinson, R. A. Integrating dynamic environmental predictors and species occurrences: Toward true dynamic species distribution models. Ecol. Evol. 10, 1087–1092 (2020).
7. Isaac, N. J. B. *et al*. Data Integration for Large-Scale Models of Species Distributions. Trends Ecol. Evol. 35, 56–67 (2020).
8. Bowler, D. E. *et al*. Integrating data from different survey types for population monitoring of an endangered species: the case of the Eld's deer. Sci. Reports 2019 91 9, 1–14 (2019).
9. Boersch-Supan, P. H. & Robinson, R. A. Integrating structured and unstructured citizen science data to improve wildlife population monitoring. bioRxiv 2021.03.03.431294 (2021) doi:10.1101/2021.03.03.431294.
10. Hertzog, L. R. *et al*. Model-based integration of citizen science data from disparate sources increases the precision of bird population trends. Divers. Distrib. 27, 1106–1119 (2021).
11. Zipkin, E. F. *et al*. Integrating count and detection–nondetection data to model population dynamics. Ecology 98, 1640–1650 (2017).
12. Jönsson, G. M., Broad, G. R., Sumner, S. & Isaac, N. J. B. A century of social wasp occupancy trends from natural history collections: spatiotemporal resolutions have little effect on model performance. Insect Conserv. Divers. 14, 543–555 (2021).
13. Farr, M. T., Green, D. S., Holekamp, K. E. & Zipkin, E. F. Integrating distance sampling and presence-only data to estimate species abundance. Ecology 102, e03204 (2021).
14. Elith, J. *et al*. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17, 43–57 (2011).
15. Pannekoek, J. & van Strien, A. J. Trends & Indices for Monitoring data. (2005).
16. MacKenzie, D. I. Occupancy estimation and modeling : inferring patterns and dynamics of species occurrence. (Academic Press, Burlington, Massachusetts,USA, 2006).
17. Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P. & Roy, D. B. Statistics for citizen science: Extracting signals of change from noisy ecological data. Methods Ecol. Evol. 5, 1052–1060 (2014).
18. Bees Wasps and Ants Recording Society. BWARS. https://www.bwars.com/.
19. Bumblebee Conservation Trust. BeeWalk. https://www.bumblebeeconservation.org/beewalk/.
20. Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B. & O'Hara, R. B. Is more data always better? A simulation study of benefits and limitations of integrated distribution models. Ecography (Cop.). 43, 1413–1422 (2020).
21. Fletcher, R. J. *et al*. A practical guide for combining data to model species distributions. Ecology 100, e02710 (2019).
22. Outhwaite, C. L. *et al*. Prior specification in Bayesian occupancy modelling improves analysis of species occurrence data. Ecol. Indic. 93, 333–343 (2018).

23. Big Butterfly Count 2020: The Results. https://butterfly-conservation.org/news-and-blog/big-butterfly-count-2020-the-results.

24. Taylor, P. *et al*. *State of Dragonflies 2021*. (2021).

25. Johnston, A., Moran, N., Musgrove, A., Fink, D. & Baillie, S. R. Estimating species distributions from spatially biased citizen science data. *Ecol. Modell*. **422**, 108927 (2020).

26. Johnston, A. *et al*. Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Divers. Distrib*. **27**, 1265–1277 (2021).

27. Boyd, R. J., Powney, G. D., Carvell, C., Pescott, O. L. & Robin Boyd, C. J. occAssess: An R package for assessing potential biases in species occurrence data. *Ecol. Evol*. **11**, 16177–16187 (2021).

28. Boyd, R. *et al*. ROBITT: a tool for assessing the risk-of-bias in studies of temporal trends in ecology. doi:10.32942/OSF.IO/RHVEY.

29. Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput. 2000 104* **10**, 325–337 (2000).

30. Plummer, M. *DSC 2003 Working Papers JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd International Conference on Distributed Statistical Computing* http://www.ci.tuwien.ac.at/Conferences/DSC-2003/ (2003).

31. Ponisio, L. C., de Valpine, P., Michaud, N. & Turek, D. One size does not fit all: Customizing MCMC methods for hierarchical models using NIMBLE. *Ecol. Evol*. **10**, 2385–2416 (2020).

32. R Core Team. R: A language and environment for statistical computing. (2018).

33. Carpenter, B. *et al*. Stan: A Probabilistic Programming Language. *J. Stat. Softw*. **76**, 1–32 (2017).

34. Golding, N. greta: simple and scalable statistical modelling in *R. J. Open Source Softw*. **4**, 1601 (2019).

35. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B (Statistical Methodol*. **71**, 319–392 (2009).

36. Martins, T. G., Simpson, D., Lindgren, F. & Rue, H. Bayesian computing with INLA: New features. *Comput. Stat. Data Anal*. **67**, 68–83 (2013).

Acknowledgements

This work was supported by the Terrestrial Surveillance Development and Analysis partnership of the UK Centre for Ecology & Hydrology, British Trust for Ornithology and the Joint Nature Conservation Committee, and by the Natural Environment Research Council award number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability. We thank the thousands of volunteers who contribute records to BirdTrack, the Breeding Bird Survey, Bees Wasps and Ants Recording Society, Bumblebee Conservation Trust, British Dragonfly Society and past and present survey organizers and staff. The BTO/JNCC/RSPB Breeding Bird Survey is a partnership jointly funded by the BTO, RSPB, and JNCC, and BirdTrack is supported by the RSPB, BirdWatch Ireland, Scottish Ornithologists' Club, and the Welsh Ornithological Society. We also thank the UK Pollinator Monitoring and Research Partnership (PMRP) for their support in the development of the modelling framework for the bumblebee case study. The PMRP comprises UK Centre for Ecology & Hydrology (UKCEH), Bumblebee Conservation Trust, Butterfly Conservation, British Trust for Ornithology, Hymettus, University of Reading, University of Leeds and Natural History Museum, and is jointly funded by Defra, the Welsh and Scottish Governments, Daera, JNCC and project partners. Computations for the case studies used JASMIN, the UK's collaborative data analysis environment (http://jasmin.ac.uk).

Photo credits
Volunteer surveying © BBS/WCBS volunteers, by David Tipling / BTO
Bumblebees © Maddy Long
Volunteer surveying © BBS/WCBS volunteers, by David Tipling / BTO
Peacock Butterfly © Kie Ker / Pixabay
Volunteers surveying © BBS/WCBS volunteers, by David Tipling / BTO
Hoverfly © Wendy Dalton
Blue tit © Burkard Meyendriesch / Pixabay
Four-spotted chaser © James Williams
South Downs Way signpost © Creative Commons License
Bumblebee on Viper's Bugloss © Natural England / Allan Drewitt
Banded demoiselle © Ian Lindsay / Pixabay
Robin © Evgeni Tcherkasski / Pixabay
Chaffinch © Pixabay
BBS/WCBS volunteers © Tipling / BTO
Large skipper © Maddy Long
Wildflower meadow © Wendy Dalton
Bumblebee © Maddy Long

JNCC is the public body that advises the UK Government and devolved administrations on UK-wide and international nature conservation. As a public body we also work in partnership with business and society. Our people are dedicated to providing high-quality evidence and advice on the natural environment for the benefit of current and future generations.

jncc.gov.uk
twitter.com/JNCC_UK
linkedin.com/company/jncc
facebook.com/JNCCUK
youtube.com/JNCC_UKvideo

ISBN: 978-1-86107-639-7

JNCC