JNCC Report 826

Informing biodiversity monitoring with integrated distribution models

Mancini, F., Pocock, M.J.O., Comont, R., Carvell, C. and Isaac, N.J.B.

May 2026

ISSN 0963 8091

JNCC's report series serves as a record of the work undertaken or commissioned by JNCC. The series also helps us to share, and promote the use of, our work and to develop future collaborations.

For further information please contact:
JNCC, Quay House, 2 East Station Road, Fletton Quays, Peterborough PE2 8YY.
https://jncc.gov.uk/
Communications@jncc.gov.uk

This document should be cited as:
Mancini, F., Pocock, M.J.O., Comont, R., Carvell, C. & Isaac, N.J.B. (2026) Informing biodiversity monitoring with integrated distribution models. *JNCC Report 826*. JNCC, Peterborough, ISSN 0963-8091.
https://jncc.gov.uk/resources/5a5bf0d3-2722-42cc-a1b9-7f6bc3a81279

Authors' affiliation:
UK Centre for Ecology & Hydrology, Wallingford OX10 8BB.

Acknowledgments:
This work was supported by the Terrestrial Surveillance Development and Analysis Partnership, comprised of the UK Centre for Ecology & Hydrology, British Trust for Ornithology, and the Joint Nature Conservation Committee. We thank the BeeWalk and BWARS volunteers for their invaluable contributions to data collection, without which this study would not have been possible.

Evidence Quality Assurance:
This document is compliant with JNCC's Evidence Quality Assurance Policy.

Summary

- Model-based data integration offers the opportunity to analyse structured and unstructured monitoring data in a unified analytical framework.
- We investigate whether adding different amounts of structured presence/absence data to a large presence-only dataset produces improved estimates of model parameters. We use data collected through standardised transect walks and opportunistic records of bumblebee species occurrence to answer this question. Model estimates might be improved either through increased precision or by reduction of the biases inherent in opportunistic data.
- We compare integrated models including an increasing number of sites surveyed through the standardised transect walks with an occupancy model that only includes the unstructured presence-only records. We use precision of model parameters to compare the performance of the different models.
- We find that the integrated models produced more precise estimates of some but not all parameters; when the two datasets provided contrasting information the integrated model tried to reconcile the different signals but produced more uncertain parameter estimates.
- Increasing the amount of information shared between the two datasets by including larger numbers of shared sites resulted in an increase in precision for some parameters but we found that precision did not plateau within the range of sites tested (10 to 192), suggesting that the addition of more sites with structured monitoring would lead to further improvements. Our results do not directly address the question of how much of an improvement is sufficient for any specific application.
- Estimates of the phenology of the species improved with the addition of the structured data: both the length and the mean of the flight period were more accurately estimated by the integrated model although we find that only the mean was estimated more precisely.
- Overall, we have shown that the integrated models provide an advantage over modelling the unstructured presence-only data alone. The magnitude of the benefit provided by the integrated model depends on the amount of information shared between the dataset and whether the signals provided by the two data sources are complementary or contrasting.

Contents

1. Introduction

Systematic monitoring is costly and, consequently, structured data from systematic surveys are only available for a few taxonomic groups and spatial and temporal coverage is usually limited. In contrast, unstructured occurrence records are available for many taxa. In the UK, many recording schemes have been active for many years (Pocock *et al.* 2015), producing long-term datasets with good spatial coverage. Although opportunistic data contain biases arising from the unstructured recording process, occupancy-detection models have been shown to produce reliable estimates of species trends from unstructured occurrence records (Isaac *et al.* 2014). As a result, these data are becoming widely used to report on trends in biodiversity for groups that lack systematic monitoring schemes (Powney *et al*. 2019; Outhwaite *et al.* 2020).

Until recently, structured and unstructured data were seen as alternate sources of information. This has changed with the development of tools to analyse data from multiple sources in a single "integrated" model (Miller *et al.* 2019; Isaac *et al.* 2020; Zipkin *et al.* 2021), in which the data generation processes underlying each dataset are captured explicitly. Bringing multiple datasets together has several advantages. First, well-constructed integrated models inherit the strengths of the datasets that contribute to them (e.g. the large sample size offered by citizen science data and robust design of a systematic survey) and, to some extent, overcome their weaknesses (e.g. the biases in unstructured records and the small sample size of structured datasets; Fithian *et al.* 2015; Simmonds *et al.* 2020). Second, combining information from multiple data providers allows stakeholders from different sectors (e.g. government agencies and NGOs) to work from a common evidence base. Where analysing different datasets in isolation presents conflicting messages about biodiversity change, an integrated framework has the potential to reveal where the key uncertainties lie. Third, integration of data with different properties makes it possible to capture processes operating at different spatial scales (Zipkin *et al*. 2021), or to estimate parameters that would not be identifiable using a single dataset. Thus, an integrated framework has the potential to produce cohesive and precise estimates of status and change in key metrics as well as effects of potential drivers, whilst making the most of all available data sources. In addition, model-based data integration offers the opportunity to improve biodiversity monitoring by informing the design of new schemes that specifically add value to the data already available. Planning and implementing new monitoring with data integration in mind can avoid duplication of monitoring effort, fill in knowledge and data gaps, diversify the recording community and achieve more efficient monitoring. Conceived in an integrated modelling framework, new monitoring schemes can be designed in a way that optimises the value added to existing datasets whilst considering the resources available.

There are several approaches to integrate multiple data sources, but the most common is via the "joint-likelihood" method in a hierarchical Bayesian modelling framework (Miller *et al.* 2019). With joint-likelihood, multiple datasets are conceptualised as independent realisations of a common state variable (e.g. site occupancy or abundance). The canonical case involves two datasets: one large "unstructured" dataset consisting of presence-only records, the second being smaller but "structured" by a formal survey design and/or data collection protocol (Pagel *et al*. 2014; Fithian *et al*. 2015; Simmonds *et al.* 2020; Suhaimi *et al.* 2021). The success of the joint-likelihood method is dependent on some way to share information between the data generation models for each dataset, either by sampling some locations in common, by sharing parameters, or both.

Most knowledge about the statistical properties of integrated models has come from computer simulations. For example, (Simmonds *et al*. 2020) explored how much structured data was required to overcome biases in unstructured data under a range of scenarios. Suhaimi *et al.* (2021) compared joint likelihood with other integration approaches.

Simulation studies are useful to establish general principles, but they do not tell us about the optimum design for any specific application. The degree to which integration is successful will depend on the research question being addressed, structure of the datasets, and the specification of the model. There have been few attempts to quantify the added value of data integration using real datasets.

We have previously investigated whether integrating a structured and unstructured dataset could produce more precise estimates of trends in bumblebee species distributions. We found a small advantage of data integration over modelling the datasets separately. In this study, we explore the success of model-based integration of structured and unstructured datasets via joint-likelihood for estimating short-term changes in distribution, using an occupancy-detection model of a common British bumblebee. Specifically, we test the relationship between parameter precision and the amount of information that is shared between the datasets. Following Simmonds *et al.* (2020), we vary the amount of structured data whilst holding constant the size of the unstructured dataset. We expect that the precision of the shared parameters will increase with the number of sites with structured data, due to the increased amount of data available to the model to estimate these parameters. Our goal is to provide a new perspective on the number of sites that would need to be monitored using a structured or semi-structured protocol to add value to a large unstructured dataset. Given that structured monitoring is resource intensive, answering this question will provide guidelines for recording schemes and societies to implement new surveys that will add value to their current datasets while using their resources effectively.

2. Methods

2.1. Data sources

Our unstructured dataset comprises presence-only occurrence records from the Bees, Wasps and Ants Recording Society (BWARS: http://www.bwars.com), which have previously been used to describe multi-decadal trends in the distribution of British bees (Powney *et al.* 2019). BWARS has been active since the 1970s and has collected hundreds of thousands of records of bees, wasps and ants. The data are typical of presence-only occurrence records, in that no information has been retained about survey protocol or sampling effort.

The structured data derive from the Bumblebee Conservation Trust's BeeWalks scheme, which has been active since 2010. BeeWalks has a spatially replicated sampling design with coverage across the UK. The data collection protocol involves fixed transect routes, which are walked approximately monthly, during the main flight period of British bumblebees (March to October). The number of transects walked each year varies, ranging between 45 in 2012 and 630 in 2019. Transects are divided into sections: counts of each species are recorded within each section (see https://www.bumblebeeconservation.org/beewalk/ for more information on the BeeWalks methodology).

Both datasets involve volunteer site-selection, which creates spatial and other biases (Fournier *et al*. 2019). Addressing these issues would be important for the results to be considered representative of the UK. Addressing such biases is possible but adds substantial complexity. Given our primary focus on assessing the value added from data integration, we have not attempted to correct for spatial and other biases in this analysis.

2.2. Data manipulation

We first selected the set of observations from the years 2010–2016, which were common to both datasets. To keep the model simple, we degraded the BeeWalks counts to presence-absence data, whilst retaining information about replication within seasons (repeat surveys of the same transect route within a year). For the BWARS data we only retained observations of bees (excluding ants and wasps) and we inferred non-detections from records of other bee species in the same visit (combination of 1 km grid cell and date).

We coerced each dataset to a common spatial and temporal resolution. For BWARS, we converted the grid reference of all records to 1 km grid cells (hereafter called "sites"), discarding those with coarser resolution. We also excluded records with a date precision coarser than one day. For BeeWalks, we converted the transect locations to a resolution of 1 km using the grid reference of the transect centre point. This means that we do not use the separate transect sections as spatial replicates, but we retain the temporal replication from the different surveys to the same transect within a year. The aggregation reduced the size of the BeeWalks dataset from 838 transects down to 723 grid (1 km) sites. When multiple transects were aggregated into the same site, they were treated as replicates if the survey dates were different.

Following this data cleaning procedure, the BWARS dataset contained 32,594 visits to 10,592 sites; BeeWalks contained 6,703 visits to 723 sites. There are 192 sites shared between datasets. Note that the BeeWalks dataset is much smaller than BWARS in terms of number of visits and sites, but the replication (visits per site) is much higher (~8 vs ~3). This reflects the structured nature of the BeeWalks data (i.e. surveys are collected via a protocol that requires repeat surveys within years at fixed transect locations).

2.3. Modelling approach

We chose the species *Bombus pratorum* as it is a widespread species reported in 10% of BWARS visits and 30% of BeeWalks visits. We ran an occupancy-detection model using the BWARS data only and one using the BeeWalks data only. We also fit an integrated model with separate detection processes for the two sets of observations (described below). In addition, we fitted a "merged" model, in which both datasets are assumed to have the same detection process (i.e. we treated the BeeWalks data as if they were opportunistic presence-only records).

We then created 40 subsets of the BeeWalks dataset of different sizes: 50, 100, 200 and 400 randomly selected sites (10 replicates each). For each of these 40 datasets, we refitted the integrated model (including the full BWARS dataset) for *B. pratorum.* By comparing between the 41 integrated models, we explore how the parameters of the integrated model change as a function of the number of sites that are shared between datasets.

2.4. Model structure

All our models are variants on a Bayesian hierarchical occupancy-detection model used previously to analyse BWARS records (Powney *et al.* 2019). Such models have been shown to perform well with unstructured data (Isaac *et al.* 2014). Occupancy-detection models use two hierarchically coupled sub-models to simultaneously estimate and account for variation in species detectability while estimating the probability that the species of interest is present at a site in a certain year. A state sub-model describes the processes that we believe to govern the species' true state (i.e. its abundance or, as in this case, its distribution). The observation sub-model is a statistical description of the observation process (the data collection) and it explicitly models a probability of detection as a function of covariates. Both single-dataset variants have one observation sub-model, which are different from each other. The merged model also has one observation sub-model only (the observation sub-model for the BWARS data), as the datasets have been merged disregarding the differences in their data collection protocols. The integrated model has two observation sub-models, each describing one dataset. In the following sections we will describe each model component in detail. Information is shared between datasets in two ways: one is by sharing sites in common, the other is via a common relationship between detectability and date.

The state sub-model is identical for all model variants:

$$z_{it} \sim Bernoulli(\psi_{it});\ logit(\psi_{it}) = a_t + u_i \qquad \text{(Equation 1)}$$

where ψ_{it} is probability of occupancy for site i in year t. a_t is the year effect and u_i is a site effect. The year effect is modelled using a random walk prior with half-cauchy hyperprior (following Outhwaite *et al.* 2018), which allows the sharing of information between years. The site effect is modelled using a normal distribution with mean zero. Occupancy $psi.fs_t$ for year t is calculated as the proportion of sites that are occupied by the species from the predicted presences (z_{it}).

Whether a species is detected (y_{iv}=1) or not (y_{iv}=0) on visit v is dependent on its actual presence or absence at the site in that year (z_{it}), its dataset-specific probability of detection ($p_{BWARS,itv}$, $p_{BeeWalks,itv}$,) and the day of the year the site was visited ($f_x(JulDate_v)$). The latter term is included to account for the seasonal colony cycle of bumblebees, and is common to the data generation models for both datasets (Equations 2 and 3), thus permitting information to be shared between datasets:

$$y_{BWARS,iv} | z_{it} \sim Bernoulli(z_{it} * p_{BWARS,itv} * beta3 * f_x(JulDate_v));\qquad \text{(Equation 2)}$$

$$y_{BeeWalks,iv} | z_{it} \sim Bernoulli(z_{it} * p_{BeeWalks,itv} * beta3 * f_x(JulDate_v)); \quad \text{(Equation 3)}$$

where *beta3* is a scaling parameter equal to $1/max(f_x)$. We modelled the seasonal variation in detection as a Gaussian distribution with two parameters estimating the mean (*beta1*) and standard deviation (*beta2*) of the flight dates (Equation 4). We found this formulation to be more numerically stable than using a polynomial function (van Strien *et al.* 2010). In this way, $p_{BWARS,itv}$ and $p_{BeeWalks,itv}$ are the conditional probabilities of detection on the mean flight date.

$$f_x(JulDate_v) = 1/(\sqrt{(2*\pi)}*beta2)*e^{(-(JulDate_v - beta1)^2/(2*beta2^2))}; \quad \text{(Equation 4)}$$

The observation sub-model used to describe the BWARS data follows previous studies (Van Strien *et al*. 2013; Powney *et al*. 2019) in using the number of species recorded (the list length) as a proxy for recorder effort (Equation 5). Specifically, visits were categorised into three Data Types based on whether the species list was a single species (*DT1*), a short list (i.e. 2 or 3 species, *DT2*), or a long list (i.e. 4 or more species, *DT3*).

$$logit(p_{BWARS,itv}) = dtype1.p_t + dtype2.p * DT2_{itv} + dtype3.p * DT3_{itv}; \quad \text{(Equation 5)}$$

In Equation 5, *dtype1.p* is the log odds (i.e. logit probability) that a single species list, made on the median date of the flight period, is a record of the species being modelled. *dtype1.p* also varies between years, such that some years have high detection and others low, reflecting that bumblebee activity is higher in some years than others (e.g. when the weather is relatively warm and dry). Parameters *dtype2.p* and *dtype3.p* are the differences in the log odds for short and long lists, respectively.

The observation sub-model that describes the BeeWalks 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 within a year (Equation 6); however, we still model annual variation in probability of detection with a year effect ($beewalks.p_t$):

$$logit(p_{BeeWalks,itv}) = beewalks.p_t; \quad \text{(Equation 6)}$$

2.5. Fitting, evaluation and inference

We used uninformative priors for all parameters and hyperparameters within the model, except for the year effect (parameter *a* in Equation 1, as noted above). All models were fitted using R2jags (Wood 2016; R Core Team 2018; Su & Yajima 2020) with three chains, 200,000 iterations, a burnin of 150,000 and a thinning rate of 3.

We checked for model convergence by looking at the Rhat values of parameter estimates (Rhat < 1.1) and by visual inspection of the chains. We evaluated the models by looking at the precision of occupancy and both state and observation model parameters. Precision is typically reported as the inverse of the variance of the posterior distribution (for any given parameters), however this retains the units of the parameter in question, making it difficult to compare across parameters. We therefore use the coefficient of variation (CV: standard deviation divided by the mean - its absolute value) as a measure of parameter uncertainty (or imprecision) that is comparable across parameters and models. Differences in CV between models provide a measure of how much uncertainty has been reduced, and in which parameters. Our key question is about the additional information gained by integrating additional sites from the structured surveys, so our inferences are primarily based on comparisons of the precision of various parameters from the models. We restrict these comparisons to those models where convergence was achieved.

3. Results

Figure 1 shows the annual occupancy for *Bombus pratorum* from the BWARS-only, BeeWalks-only, merged and integrated model using the full dataset.

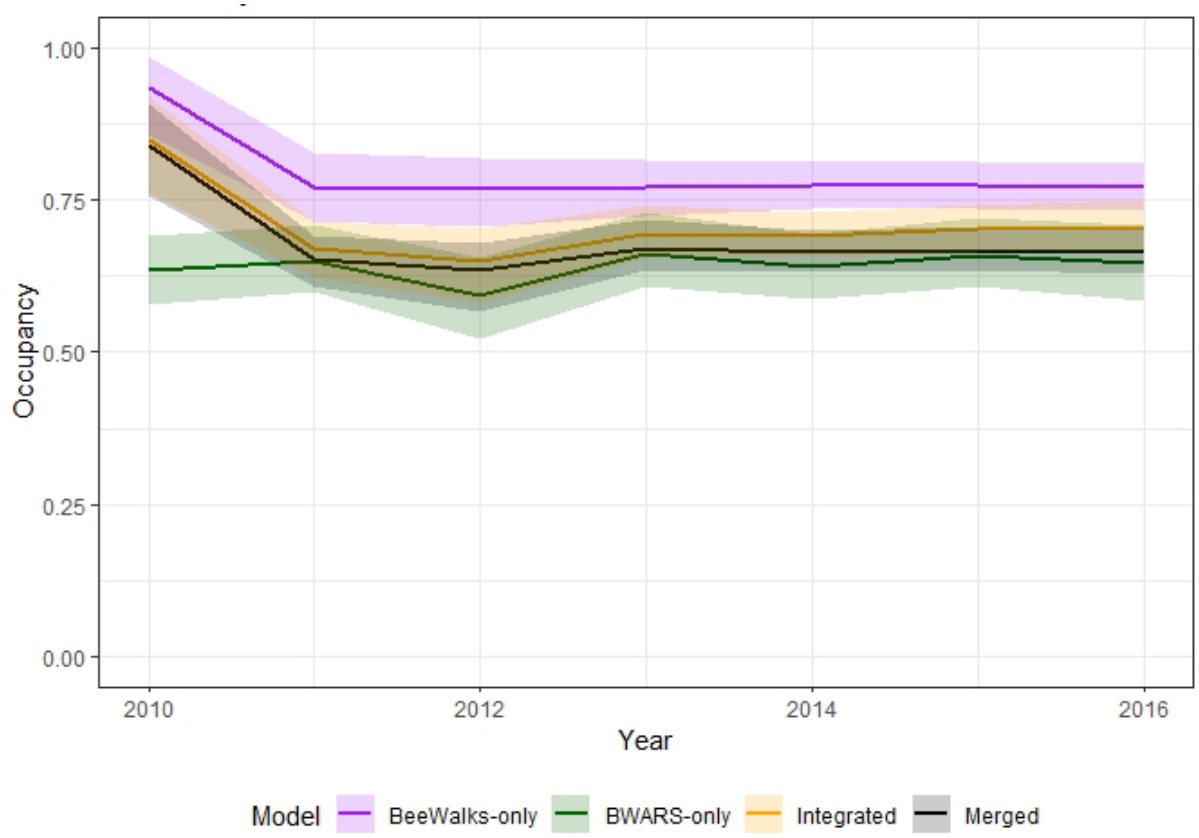


Figure 1. Annual occupancy for *Bombus pratorum* from the BWARS-only, BeeWalks-only, merged and integrated model using the full dataset.

3.1. Models based on the full dataset

Convergence was achieved for all parameters in the single dataset, merged and integrated model variants using the entire BeeWalks dataset. The most notable feature of the occupancy trends is that the models containing BeeWalks data all show a big drop in occupancy between 2010 and 2011. All models show approximately flat trends for the years 2011–2016, but there are differences in the average occupancy of each model. The BeeWalks-only model has the highest estimated occupancy, and the BWARS-only model has the lowest. Interestingly, the integrated model produces an estimate that is intermediate between these two, whereas the merged model, which has identical data but different parameters, produces an estimate that is much closer to the BWARS-only model. This distinction between merged and integrated models suggests that the additional structure in the integrated model (accounting for the systematic protocol of BeeWalks) delivers tangible benefits in terms of reduced bias of parameter estimates.

3.2. Integrated models using data subsets

Five of the 40 models did not reach satisfactory convergence (Rhat > 1.1), so we exclude them from the results below.

Precision of state model parameters (*a*, *psi.fs*) was generally higher (lower Coefficient of Variation; CV) from the integrated models compared to the BWARS-only model (Figure 2). The points above the line in Figure 2 (see also Figure 3 top panel) all refer to the first year (2010, marked as year 1 in Figure 3), for which the BeeWalks data implies a much higher occupancy than other years (Figure 1), so this reduction in precision can be attributed to contradictory information, rather than any failure of the model. The year with the biggest gain in precision (year 3) is the one with the greatest uncertainty (highest CV) in the BWARS-only model (Figure 3, middle panel). There is also a clear advantage to including more grid cells from the structured dataset: the gain in precision increases with the number of BeeWalks sites (Figure 2), especially for the year effect, *a*, and more subtly in the estimated occupancy (*psi.fs*).

The width of the flight period, *beta2,* has substantially greater precision (lower CV) in the integrated model (Figure 2). The remaining observation model parameters (*beta1* and the *dtype* parameters) all exhibit lower precision (increased CV) in the integrated model (Figure 2), although the effect is very slight and for *beta1* it approaches 0 when the full BeeWalks dataset is included.

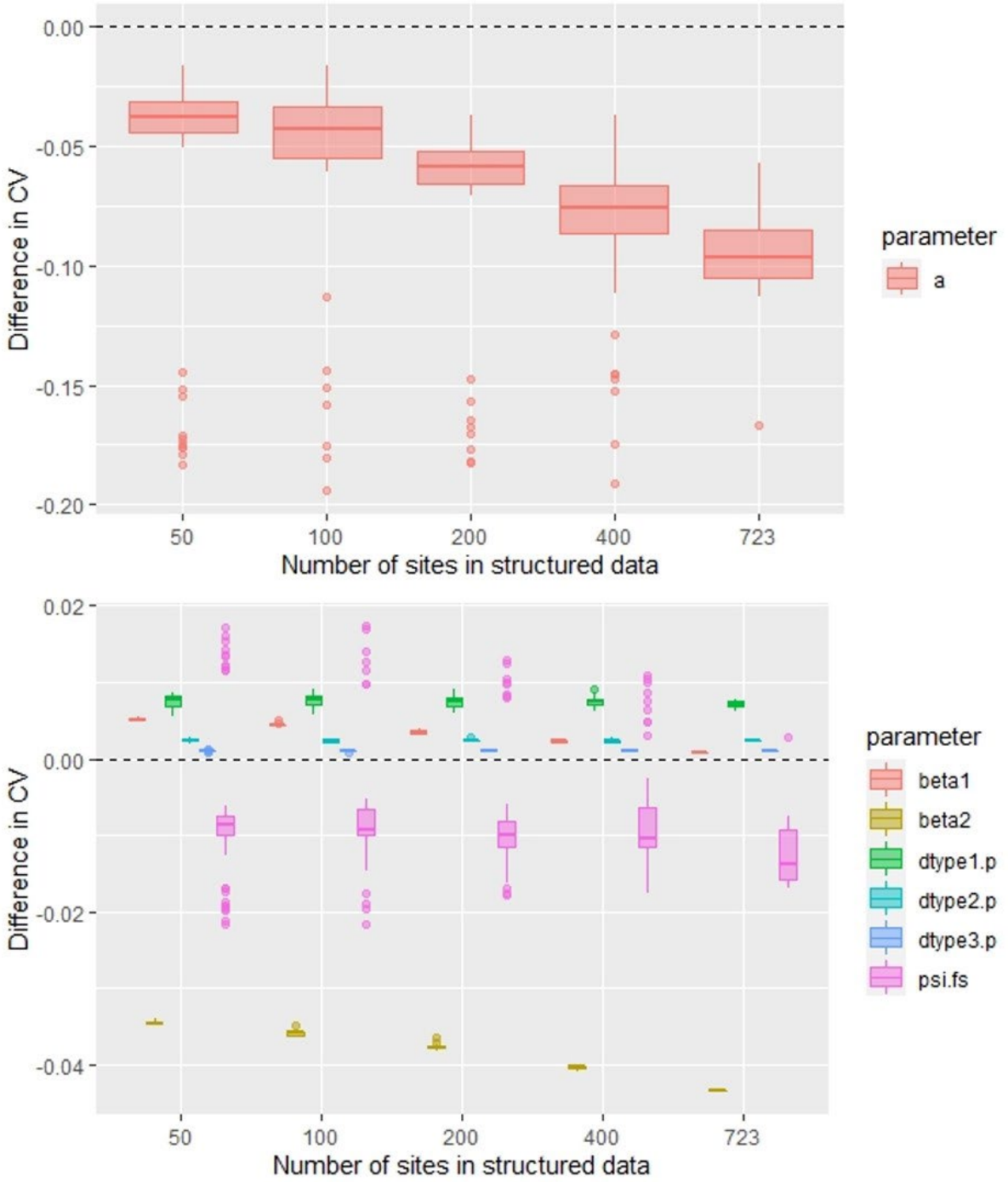
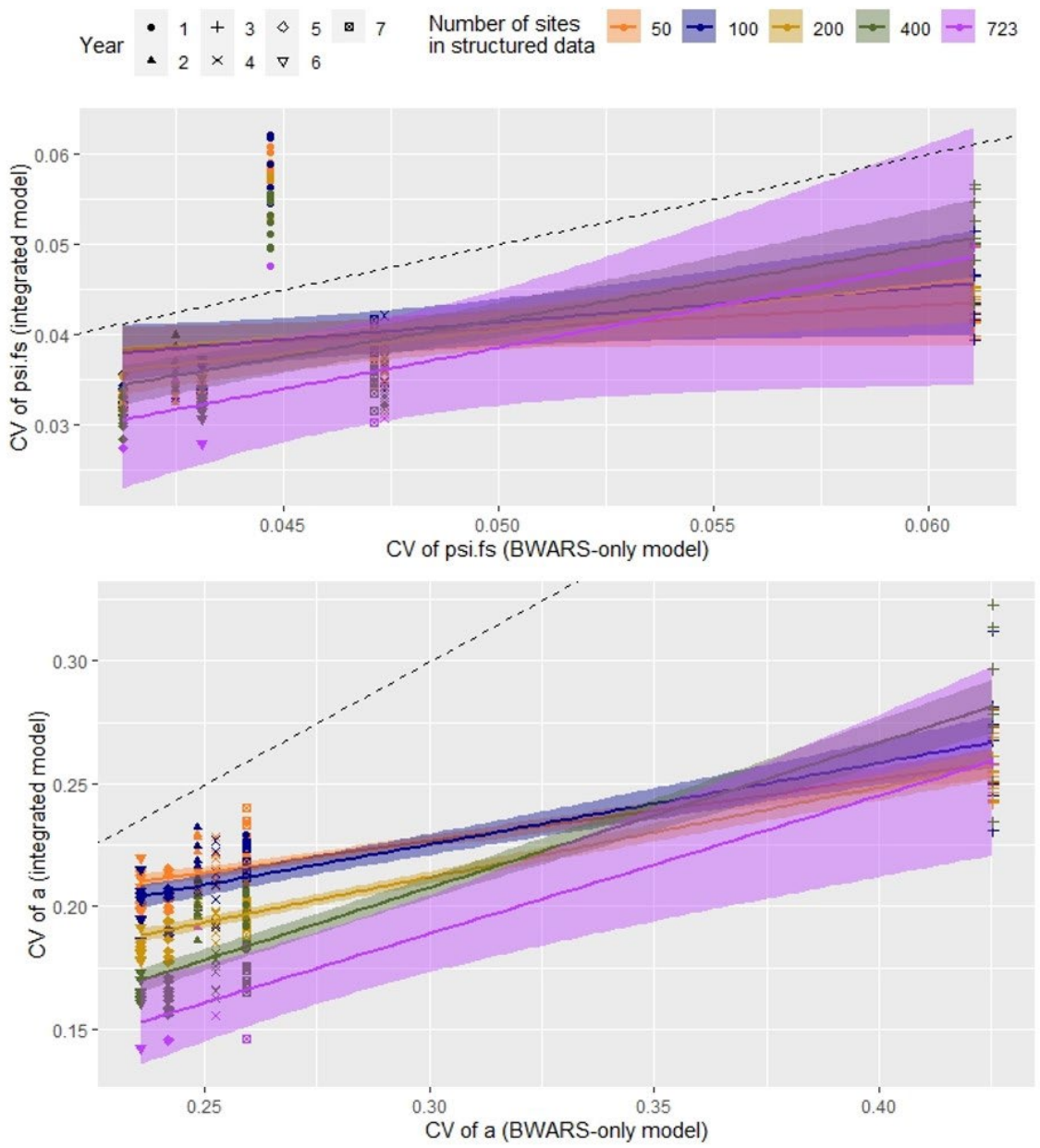


Figure 2. Difference in precision between integrated models and BWARS model for *a,* the year effect on probability of occupancy (top) and all other parameters (bottom), for every subset of BeeWalks data, including the original integrated model with the entire dataset (723 grid cells). The y axis measures the difference in coefficient of variation between the integrated model and the BWARS-only model: negative numbers indicate higher precision in the integrated model; positive numbers indicate higher precision in the BWARS-only model. Box and whiskers show the median, lower and upper quartile, and range of the CV across model runs, with circles representing outliers. The x-axis in the lower plot is scattered to ease visualisation.


Year
1
2
3
4
5
6
7
Number of sites in structured data
50
100
200
400
723
CV of psi.fs (integrated model)
0.06
0.05
0.04
0.03
0.045
0.050
0.055
0.060
CV of psi.fs (BWARS-only model)
CV of a (integrated model)
0.30
0.25
0.20
0.15
0.25
0.30
0.35
0.40
CV of a (BWARS-only model)

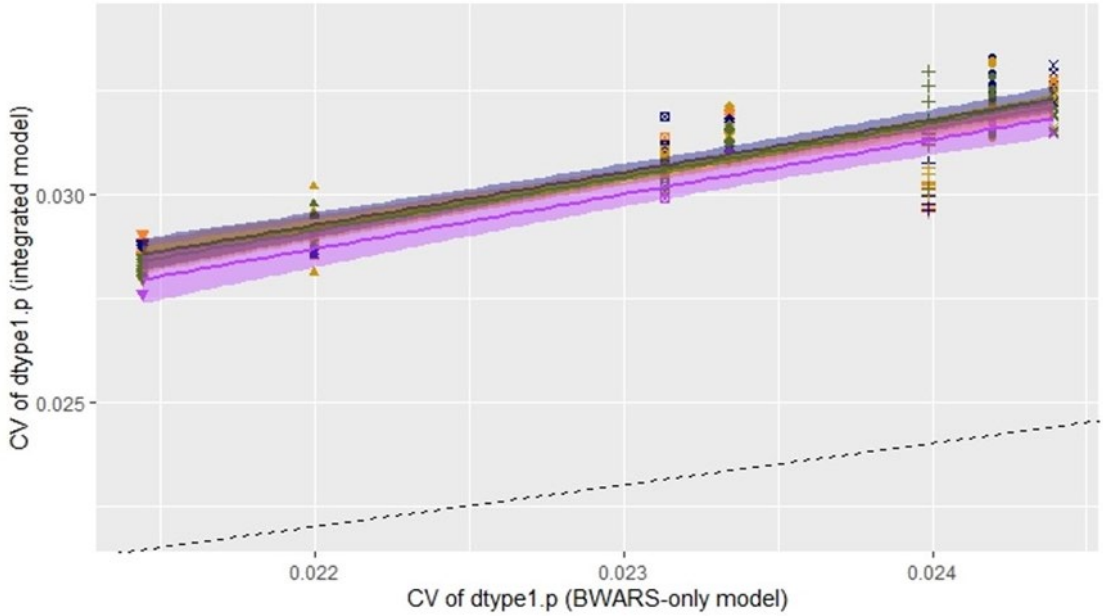


Figure 3. Correlation between coefficient of variation in *psi.fs* (top), *a* (middle) and *dtype1.p* (bottom) estimated by the integrated models and the BWARS-only model. Every shape is a parameter estimate per year and replicate, except for the original integrated model with 723 grid cells, which only has one replicate. Different colours indicate the different data subsets that contributed to the integrated models, different shapes indicate different years (1 = 2010, 7 = 2016). The dashed line is the 1:1 line; all shapes and regression lines below the dashed line indicate the integrated model estimated the parameter more precisely than the BWARS-only model.

Our replicate dataset models were selected to have a set number of sites with BeeWalks data, but they differ in the number of sites that are shared between datasets. Plotting the parameters CV as a function of the number of shared sites brings out the gain in precision more clearly. As expected, there are gains in precision (reduction of CV) in all the shared parameters (*a, psi.fs, beta1* and *beta2)* and the BeeWalks-specific parameter (*beewalks.p*) but not the BWARS-specific parameters (*dtype1.p, dtype2.p, dtype3.p*).

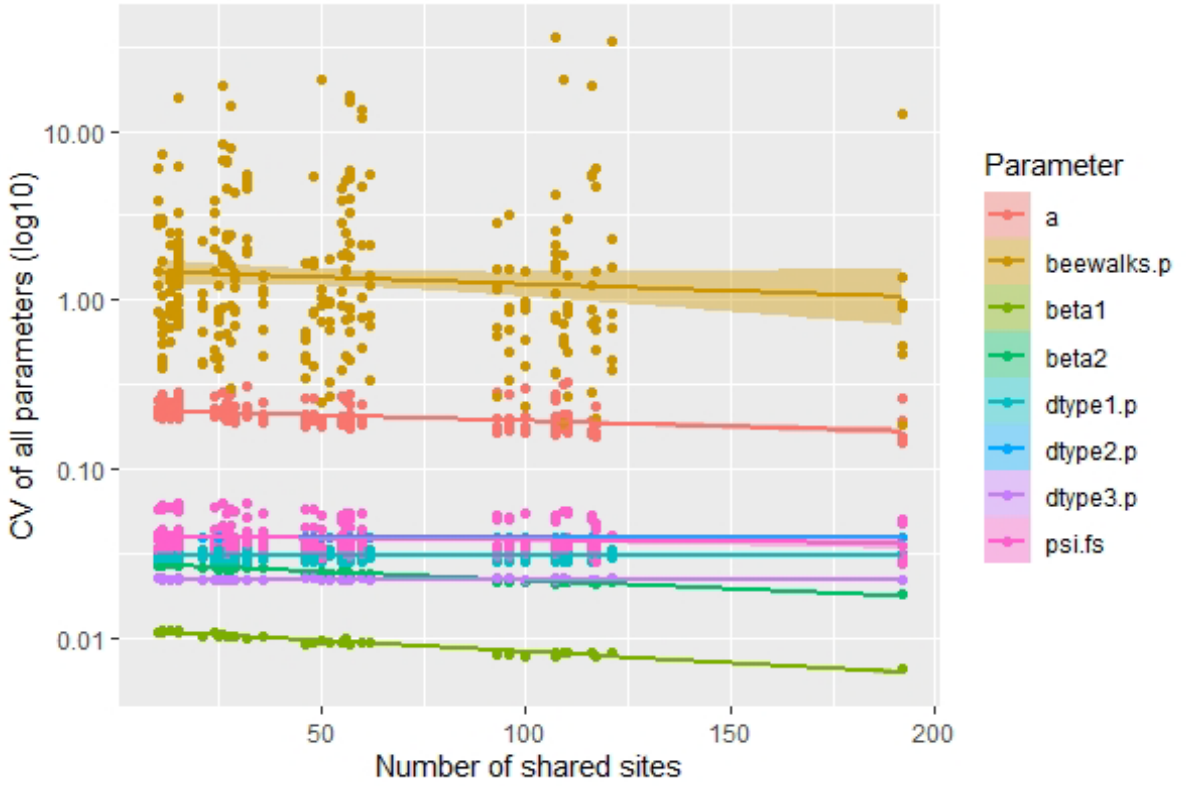


Figure 4. Precision of all parameters estimated by the integrated models by number of sites shared between the datasets. The number of shared sites ranges from 10 (from one of the BeeWalks subsets with 50 sites) and 192 (from the whole BeeWalks dataset). Parameters *a, beewalks.p, dtype1.p* and *psi.fs* have one value of CV per year per replicate, the others have one value per replicate.

4. Discussion

In this analysis we asked whether integrated models produce more precise estimates than simpler models, and whether the added value (in terms of increased precision, measured as a reduction in the CV) increases with the amount of information that is shared between datasets. Our results provide good evidence for such an improvement, although some parameters are estimated with greater uncertainty than in the simple BWARS-only model.

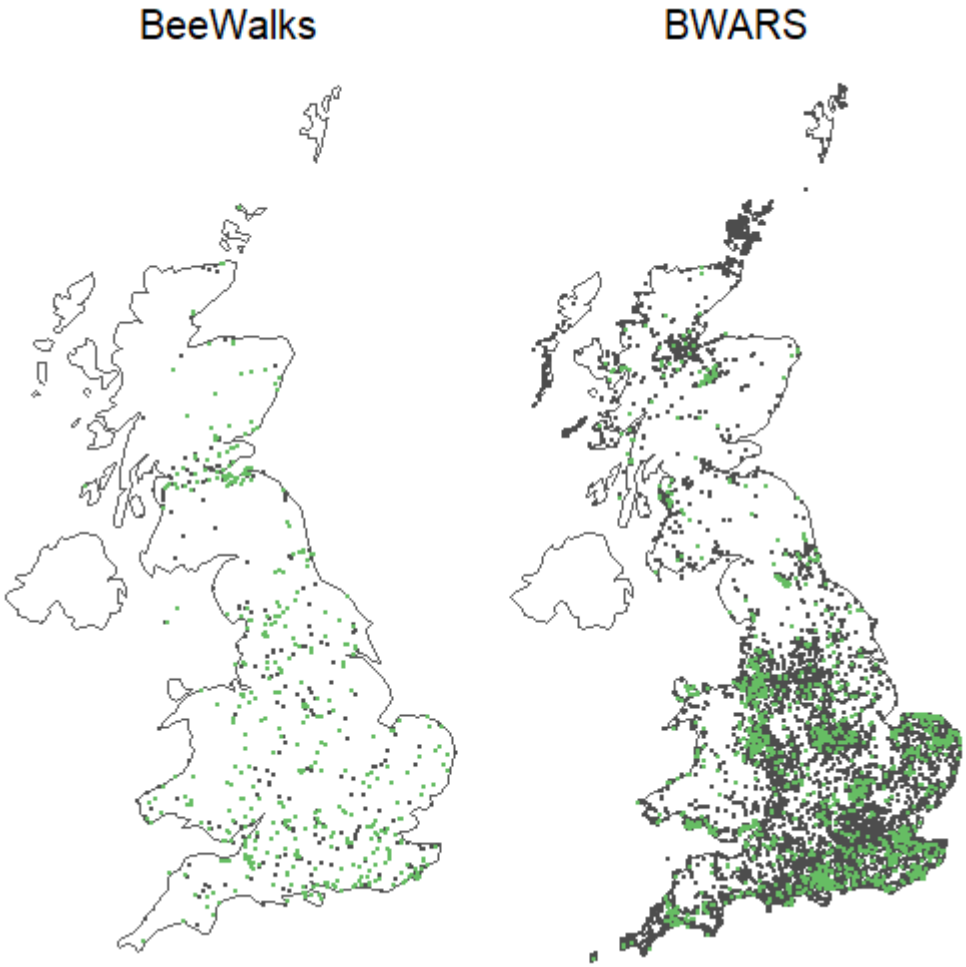


Figure 5. Locations of sites surveyed in BeeWalks and BWARS data. Points in green indicate sites where *Bombus pratorum* was detected.

We focussed our analysis on precision, but it is important to recognise that precision is only part of the story in terms of information content. We would expect integration to produce more precise estimates if both datasets are providing information that is mutually reinforcing, (i.e. converging on the same parameter value). If the two datasets are providing contradictory evidence, we might expect the integrated model parameters to be less precise than in either of the constituent datasets. The occupancy parameters (*a* and *psi.fs*) provide examples of both phenomena. For most years in the dataset, occupancy in the BeeWalks-only model is about 20% higher than in the BWARS-only model (Figure 1), suggesting that the two datasets sample the distribution of *B. pratorum* differently. This is supported by the map of sample locations (Figure 5), which reveals that BeeWalks transects are much less aggregated than the BWARS data. We did not control for spatial effects in this analysis, and the difference between models suggests that addressing this issue should be a priority.

For most years, the difference in occupancy between models is relatively small compared with the uncertainty in those estimates. However, there is something odd about the first year in the dataset (2010): here the difference in estimated occupancy of the separate models is about twice as much as in other years (Figure 1). We suspect this does not reflect a big decline in occupancy but rather is an artefact of turnover among BeeWalks surveyors. Nearly half (67/146) the BeeWalks transect routes that were sampled in 2010 then disappeared from the BeeWalks network. Including sites that are visited in just one year can make it difficult for the model to distinguish between turnover (of sites) and genuine change in occupancy (Isaac *et al*. 2014), so perhaps these should have been excluded. Having included them, the integrated model then must reconcile the different signals from the two datasets, such that occupancy precision for 2010 was lower in the integrated model than for the BWARS-only model (Figure 3).

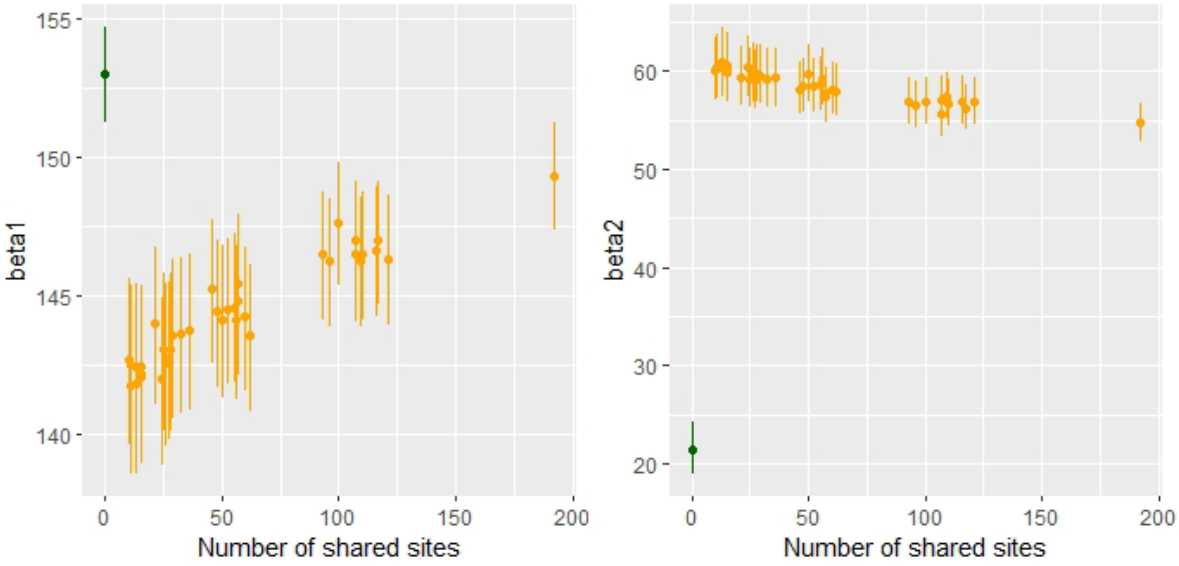


Figure 6. Estimated value of (left) *beta1*, the mean flight date, and (right) *beta2*, the length of the flight period, from 41 models plotted against the number of shared sites. Each dot shows the mean estimate from one model; the bars delimit one standard deviation from the mean. Green points and bars are the estimates from the BWARS-only model (0 shared sites), points and bars in orange are estimates from integrated models.

The most interesting pattern occurred with the phenology parameters, *beta1* and *beta2,* which are shared between the observation models describing the two datasets. Precision of both parameters increases with the number of BeeWalks sites included in the model (Figure 4), but *beta1* (the mean flight date) is estimated with less precision in all the integrated models than in the BWARS-only model (Figure 2). As with the occupancy estimates in the first year, reduction in precision reflects contradictory information from the two datasets. In the BWARS-only model, the mean flight date is estimated at 153 (2 June), but in the integrated model with all BeeWalks data, the best estimate is approximately four days earlier (Figure 6). For integrated models with few shared sites, the estimate is even lower, at around day 142 (21 May). Whilst *beta2* has higher precision in all integrated models than the BWARS-only model (Figure 2), the addition of structured BeeWalks data clearly contradicts the signal coming from the unstructured BWARS data. In this case, however, this contradictory information does not lead to a decrease in precision as with the parameter *beta1*. The BWARS-only model estimates the width (the standard deviation) of flight days at just 22 days, compared with 55 or more in the integrated models (Figure 6).

Bombus pratorum (common name early bumblebee), is one of the earliest bumblebees to emerge. Queens emerge as early as March-April, while males can be often seen by the end of May-June. The BWARS-only model does not capture this early phenology well (Figure 7), estimating a narrow period of high detectability between mid-April and late July. This means that the BWARS-only model would likely over-estimate occupancy for sites that were surveyed outside of this period. Interestingly, the BeeWalks-only model estimates a similar mean flight date to the BWARS-only, but a longer flight period, whereas the mean flight date estimated by the integrated model is earlier than either of the single-dataset models (Figure 6, Figure 7). This suggests that the two datasets are providing complementary information about the phenology of this species. In fact, *B. pratorum* is bivoltine in southern England, with a late summer generation, so our Gaussian curve is a simplification of the true phenology. However, this issue affects all our models equally, so is not pertinent to the research questions at hand. Thus, like occupancy for the year 2010, the integrated model must reconcile contradictory messages from the two datasets.

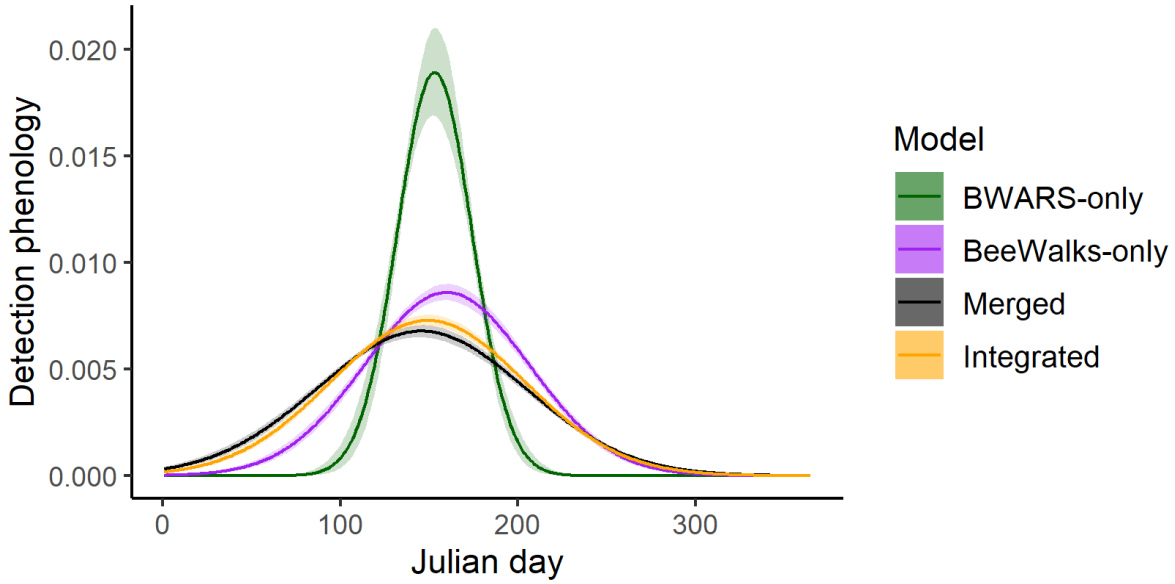


Figure 7. Detection phenology for *Bombus pratorum* estimated by the different model variants. Solid lines are the mean of $f_x(JulDate_v)$ by Julian day. Shaded areas are credible intervals from the posterior distribution. N.B. in most years, the 100th Julian day falls on 10 April.

For the *dtype* parameters, we found that the integrated models all had lower precision than the BWARS-only model. Closer inspection reveals that the largest change in these parameters is a higher estimated mean of *dtype1.p* in the integrated model compared with the BWARS-only model (Figure 8). Again, this implies that the BeeWalks data has provided information that contradicts the BWARS-only model. Specifically, it suggests that there were some shared sites where *B. pratorum* was observed on BeeWalks transects, but not in the BWARS data (such that the detectability in BWARS was under-estimated). However, given this pattern is it perhaps surprising that we found no evidence of increased precision as the number of shared sites increased (Figure 4).

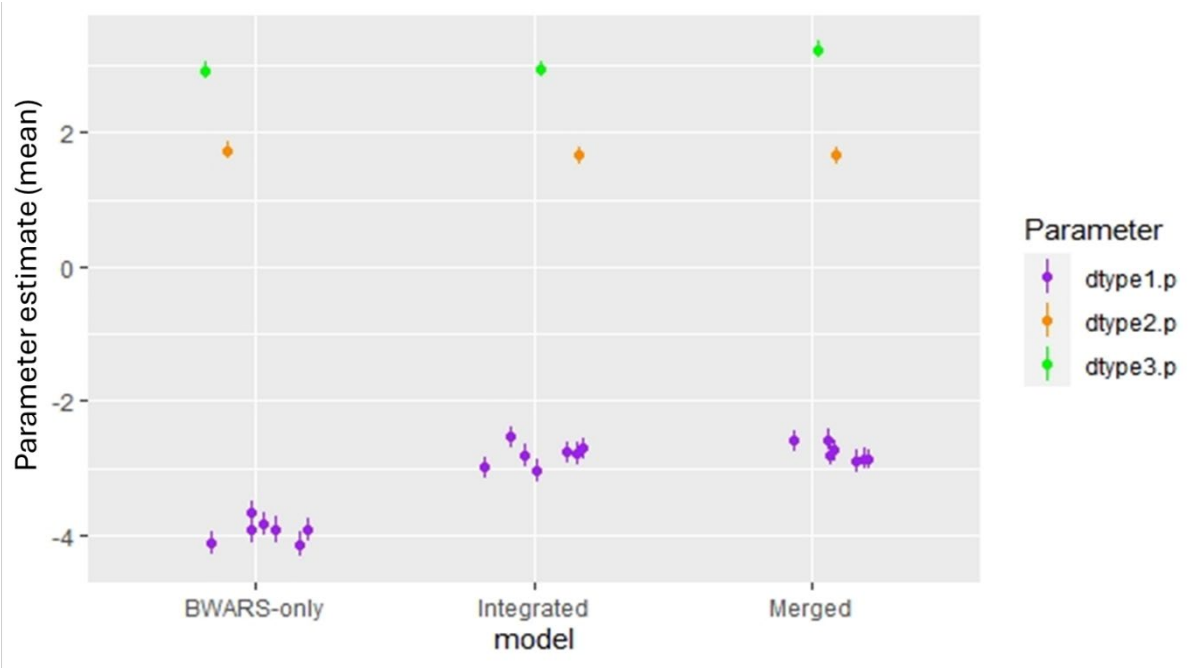


Figure 8. Parameter estimates (mean and standard deviation) for the *dtype* parameters in the BWARS-only, Integrated model (full dataset) and the merged model. Note that *dtype1.p* has one estimate per year, whereas the other two are fixed. The x-axis is scattered to ease visualisation.

beewalks.p is estimated entirely from the BeeWalks data, so we would expect the estimates of this parameter to improve with more data. Precision of *beewalks.p* does increase with the number of BeeWalks sites included in the integrated model (Figure 4), but it does not seem to plateau within the range of shared sites tested here. This suggests that more structured data are required for this parameter to be estimated more precisely.

Through simulations, Dorazio (2014) showed that integrating opportunistic presence-only data with as little as 50 sites from planned surveys (out of 10000 total sites in the study area) produced unbiased and precise estimates of the state variable. We generally find an advantage of integrating as little as 50 sites from structured data to a large unstructured dataset on the precision of some parameter estimates, but the magnitude of this effect is not as dramatic as in Dorazio (2014). Simulated datasets contain a lot of simplifying assumptions due to the need to be able to identify general rules, but these assumptions are not often true in real datasets. For example, the BeeWalk data, although collected through a standardised protocol, cannot be considered unbiased as volunteers choose where to set up their transect and missing surveys are possible. This is why it is important to investigate the advantages of model-based data integration using real world data so that we can provide guidelines for best practices in biodiversity monitoring and analysis of biodiversity data.

Overall, our results demonstrate that integrated models do deliver benefits over simpler models. The magnitude of these benefits depends on the amount of information shared between the datasets and on whether this information is consistent between the two sources. This conclusion makes it difficult to provide a definite answer to the question of how much structured data is enough to add value to existing unstructured data. We do see improvements in precision from integrating as little as 50 sites surveyed using the structured protocol, ten of which are shared between the two datasets; however, these improvements are small. As we included more sites from the BeeWalks data up to including the full dataset, we do not find that precision for any of the model parameters plateaus, which means that there is still room for improvement. In addition, there are other properties of a dataset other than its sample size that will influence its information content and the value

added, one of which is spatial bias. As an example, the UK Pollinator Monitoring Scheme's (PoMS) pan trap surveys have been sampling 75 sites across GB for the last five years. This is like the lower end of the range of sample sizes tested in this study. This suggests that integrating PoMS data with opportunistic records of bees might produce similar benefits as our BeeWalks subsets of 50–100 sites. However, this threshold of 50–100 sites would likely have been lower if the number of sites in the BWARS data was smaller. In addition, while BeeWalks transects contain some unquantified spatial bias, the PoMS surveys follow a stratified random sampling design and would have other sampling differences such as being less aggregated. This may influence the final results. We have shown that integrating information from structured data can improve model performance, especially when enough information is shared between the data sources being integrated. How much "enough information" is will depend on several factors, including the sample size of both datasets and their biases. Understanding how these factors affect the performance of integrated models should be a priority for future research.

References

Dorazio, R.M. (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Glob. Ecol. Biogeogr.* **23**, 1472–1484. doi:10.1111/geb.12216.

Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2015) Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods Ecol. Evol.* **6**, 424–438. doi:10.1111/2041-210X.12242.

Fournier, A.M.V., White, E.R. & Heard, S.B. (2019) Site-selection bias and apparent population declines in long-term studies. *Conserv. Biol.* **33**, 1370–1379. doi:10.1111/cobi.13371.

Isaac, N.J.B., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., Freeman, S.N., Golding, N., Guillera-Arroita, G., Henrys, P.A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O.L., Schmucki, R., Simmonds, E.G. & O'Hara, R.B. (2020) Data Integration for Large-Scale Models of Species Distributions. *Trends Ecol. Evol.* **35**, 56–67. doi:10.1016/j.tree.2019.08.006.

Isaac, N.J.B., van Strien, A.J., August, T.A., de Zeeuw, M.P. & Roy, D.B. (2014) Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods Ecol. Evol.* **5**, 1052–1060. doi:10.1111/2041-210X.12254.

Miller, D., Pacifici, K., Sanderlin, J.S. & Reich, B. (2019) The recent past and promising future for data integration methods to estimate species' distributions. *Methods Ecol. Evol.* **10**, 22–37. doi:10.1111/2041-210X.13110.

Outhwaite, C.L., Gregory, R.D., Chandler, R.E., Collen, B. & Isaac, N.J.B. (2020) Complex long-term biodiversity change among invertebrates, bryophytes and lichens. *Nat. Ecol. Evol.*, 1–9. doi:10.1038/s41559-020-1111-z.

Pagel, J., Anderson, B. J., O'Hara, R. B., Cramer, W., Fox, R., Jeltsch, F., Roy, D.B., Thomas, C.D. & Schurr, F.M. (2014) Quantifying range-wide variation in population trends from local abundance surveys and widespread opportunistic occurrence records. *Methods Ecol. Evol.* **5**, 751–760. doi:10.1111/2041-210X.12221.

Pocock, M.J.O., Roy, H.E., Preston, C.D. & Roy, D.B. (2015) The Biological Records Centre: A pioneer of citizen science. *Biol. J. Linn. Soc.* **115**, 475–493. doi:10.1111/bij.12548.

Powney, G.D., Carvell, C., Edwards, M., Morris, R.K.A., Roy, H.E., Woodcock, B.A. & Isaac, N.J.B. (2019) Widespread losses of pollinating insects in Britain. *Nat. Commun.* **10**. doi:10.1038/s41467-019-08974-9.

R Core Team. (2018) R: A language and environment for statistical computing. Available from: https://www.r-project.org/.

Simmonds, E.G., Jarvis, S.G., Henrys, P.A., Isaac, N.J.B., & O'Hara, R. B. (2020) Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography (Cop.).* **43**, 1413–1422. doi:10.1111/ecog.05146.

Su, Y.-S. & Yajima, M. (2020) R2jags: Using R to Run "JAGS." Available from: https://cran.r-project.org/package=R2jags.

Suhaimi, S.S.A., Blair, G.S. & Jarvis, S.G. (2021) Integrated species distribution models: A comparison of approaches under different data quality scenarios. *Divers. Distrib.* **27**, 1066–1075. doi:10.1111/DDI.13255.

Van Strien, A.J., Van Swaay, C.A.M. & Termaat, T. (2013) Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *J. Appl. Ecol.* **50**, 1450–1458. doi:10.1111/1365-2664.12158.

Wood, S.N. (2016) Just another Gibbs additive Modeler: Interfacing JAGS and mgcv. *J. Stat. Softw.* **75**, 1–15. doi:10.18637/jss.v075.i07.

Zipkin, E.F., Zylstra, E.R., Wright, A.D., Saunders, S.P., Finley, A.O., Dietze, M.C., Itter, M.S & Tingley, M.W. (2021) Addressing data integration challenges to link ecological processes across scales. *Front. Ecol. Environ.* **19**, 30–38. doi:10.1002/fee.2290.

Weblinks

Table 1. Full URLs for weblinks used in the text.

Weblink text	Full URL
Bees, Wasps and Ants Recording Society (BWARS)	http://www.bwars.com
BeeWalks method	https://www.bumblebeeconservation.org/beewalk/