Supplementary material: description of the variable selection procedure used prior to trend estimation

Introduction

An important step in model construction is the choice of variables to be included. We are interested in temporal variations in abundance for STOC and in occurrences for Vigie-flore. Time must therefore be included as a fixed effect, and corresponds to the variable of interest. The next step is to identify all additional covariates that could impact abundance (resp. occurrence), to avoid possible confounding effects, particularly with time. This would be the case in particular if other variables affecting abundance were to change over time. For example, if the spatial distribution of sampling sites changes over time, a model without spatial information would be unable to differentiate an effect of time on abundance from an effect of environment or location. The aim of this section is therefore to identify, for each monitoring scheme, the covariates to be included in the model. In particular, this involves studying how well the various models fit the data.

Possible covariates

Fixed effects

The list of variables likely to influence bird abundance (resp. plant occurrence) and have a confounding effect with time is listed in Table 1 and has been split into three categories.

The first set of variables is related to **observation conditions**: those variables are simple descriptors of the conditions under which the observations took place. For Vigie-Flore, the observation date is the variable describing observation conditions, which can influence species occurrence. For STOC, the number and timing of recording sessions were included by means of a 3-mode categorical variable (early, late or double recording sessions). The observation date was not directly tested for STOC because of the aggregation between passages.

The second set of variables is related to the **spatial distribution of observations**, and is composed of variables that describe the location of squares in France and synthesize large-scale environmental variations (climate, geology, etc.). The effect of these variables on abundance could be confounded with time if changes in spatial distribution of squares were to happen. These variables include longitude and latitude, as well as transformations of these variables (i.e. 2^{nd} order polynomial and interaction between the two) to test non-linear relationships between geographic coordinates and abundance. Altitude is also used as a spatial variable.

The last set of variables is related to **land-cover** of squares and is based on Corine Land Cover (CLC) data, that may also have a potential impact on mean abundance and be confounded with an effect of time if squares distribution was to change. For STOC, four variables were used to describe land cover, namely the percentage of surface area covered with forests, urban areas, aquatic areas and "other" land use, within a radius of one kilometre around the centre of each square. This buffer size makes it possible to describe land use over the entire square. The percentage of agricultural area was not included in the model, to avoid collinearity between landscape variables (the sum of relative areas being equal to 1). For Vigie-Flore, a single five-category variable corresponding to the main land-cover (i.e., with the largest surface area) was used to describe vegetation cover. This main land-cover is deduced from a buffer zone of five meters around the coordinates of each plot. The size of this buffer zone enables us to describe the local land cover where the plot is located. The choice of a categorical variable stems from the low number of plots (1.4%) with several land-use classes.

All these observations, spatial and land-use variables are treated as fixed effects in the models.

***Table 1** : Description of independent variables tested as fixed effects in the models*

VARIABLE CLASS	VARIABLE NAME	TYPE	VARIABLE DESCRIPTION	STOC	VIGIE-FLORE
	year	Numerical	Observation year	X	X
Land use land cover variables	*prop_forest*	Numerical	Percent forest cover in the square	X	
	prop_urban	Numerical	Percent urban cover in the square		
	prop_water	Numerical	Percent aquatic cover in the square		
	prop_other	Numerical	Percent "other" cover in the square		
	main_cover	Factorial	Majority vegetation type around the plot		X
Spatial variables	*longitude*	Numerical	Longitude coordinates (included as a 2nd order polynomial)	X	X
	latitude	Numerical	Latitude coordinates (included as a 2nd order polynomial)		
	long_lat	Numerical	Interaction between longitude and latitude		
	elevation	Numerical	Square (resp. plot) elevation		
Variables for observation conditions	*date*	Numerical	Observation date (Julian date)		X
	session	Factorial	Type of observation session (early, late, double)	X	

Random effects

STOC (respectively Vigie-flore) data are characterized by their non-independence, due to the repetition of visits to the same square from one year to the next, and to the proximity of different sampling points (resp. plots) within a square. This non-independence was modelled via random effects, which are presented in Table 2.

For STOC, we tested square, either nested or not in administrative department as two random effects on the intercept of abundance. We also included a random effect of square on the slope of the year effect, to let temporal trends in abundance vary across squares. For Vigie-flore, for which we use data at the plot level, we included plots nested in square and nested or not in administrative department as random intercepts. We also tested for a random effect of plots on the slope of the year effect. A random intercept of observer was also introduced for Vigie-flore, since 33% of participants follow at least two squares, corresponding to 72% of sites. This effect was not included for STOC, as the majority of squares are associated with a single observer (82%): the observer effect is therefore almost confounded with the square effect.

***Table 2** : description of independent variables tested as random effects in models*

SCHEME	***VARIABLE NAME***	**HYPOTHESIS**
STOC	*(1|square)*	Dependence between squares regarding mean abundance
	(1|dep/ square)	Spatially structured (administrative department) dependence between squares regarding mean abundance
	(year| square)	Dependence between squares regarding mean year effect on abundance
VIGIE-FLORE	*(1| square /plot)*	Dependence between plots nested in squares regarding mean abundance
	(1|dep/ square / plot)	Spatially structured (administrative department) dependence between plots nested in squares regarding mean abundance
	(year|plot)	Dependence between plots regarding mean year effect on occurrence
	(1|observer)	Dependence between observers regarding mean occurrence

Approach and models

Overall approach

The aim is to identify the covariates that are essential for estimating the trend. To achieve this, several choices were made to simplify the approach, reduce computation times and avoid multiplying models. The covariates were selected in two stages: first by choosing the variables structuring the random effect, then by selecting the variables treated as fixed effects. This approach ensures that only relevant random effects are introduced in step 2, thus avoiding model complexity and reducing computation time. This approach is also consistent with the idea of correcting for data dependency first, before focusing on fixed effects. In the second stage, selection of fixed effects was carried out by group of variables rather than variable by variable to avoid multiplying the models. These groups of variables are listed in Table 4.

Step 1: random effect selection

The tested random effect variables are listed in Table 3. For each of two levels of structure (square or department), we examined:

- a "spatial" model, which allows mean abundance (resp. occurrence) to vary across squares (resp. plots within squares);
- an "observer" model, which also allows mean occurrence to vary between observers (only for Vigie-flore);
- a "full" model, which additionally allows the effect of time on abundance (resp. occurrence) to vary across squares (resp. plots within squares).

***Tableau 3** : Parametrization of models tested in step 1 (random effect selection)*

Spatial structure level	**Scheme** / **Model**	**STOC**	**Vigie-Flore**
Site	***Spatial*** model	(1|square)	(1|square/plot)
	Observer model	-	(1|square/plot) + (1|observer)
	Full model	(1|square) + (year|square)	(1| square/plot) + (1|observer) + (year|plot)
Department	***Spatial*** model	(1|dep/square)	(1|dep/square/plot)
	Observer model	-	(1|dep/square/plot) + (1|observer)
	Full model	(1|dep/square) + (year|sqaure)	(1|dep/square/plot) + (1|observer) + (year|plot)

Step 2: fixed effects selection

In this step, we test five models grouped together in Table 4:

- The “simple” model contains no covariates;
- The “conditions” model contains the covariates of the observation conditions;
- The “spatial” model corrects for the covariates of observation and spatial conditions;
- The “land use” model corrects for the covariates of observation conditions and land use;
- The “full” model corrects for all covariates.

All these models additionally contain the random effects selected in step 1 for structuring the random effect.

***Table 4** : description of variables included in each model*

Variables / Model	**Observation variables**	**Spatial variables**	**Land use variables**
***Simple** model*			
***Condition** model*	X		
***Spatial** model*	X	X	
***Land use** model*	X		X
***Full** model*	X	X	X

Results

Step 1: random effect selection

The goodness of fit of the different models to the data was calculated using the BIC (Bayesian Information Criterion), which penalizes additional parameters more than the AIC (Akaike Information Criterion). For each species, the best model thus corresponds to the one with the lowest BIC. Species for which at least one of the models had a convergence issue were set aside for this analysis. The results are shown in Table 5. We note that:

- The level of complexity of the random effect in models with the best fit varies according to species;
- Structuring by department improves the fit for only 13% of species in Vigie-Flore, compared with 77% of species in STOC;
- For STOC, the full model structured by department is the best model for two-thirds of species;
- For Vigie-Flore, when the random effect is structured by plot, the simplest spatial model is selected for a majority of species (38% of species in total), although more complex models are selected for almost half the species.

***Table 5** : Number of species for which each model obtains the lowest BIC*

Spatial structure level	**Scheme** / **Model**	**STOC**	**Vigie-Flore**
Site	***Spatial*** model	22 (17%)	112 (38%)
	Observer model	-	54 (18%)
	Full model	9 (7%)	87 (30%)
Department	***Spatial*** model	14 (11%)	32 (11%)
	Observer model	-	7 (2%)
	Full model	87 (66%)	0 (0%)

The models selected to structure the random effect vary across schemes:

- → For STOC, it seems appropriate to retain the full model structured by department, which achieves a minimum BIC for most species;
- → For Vigie-Flore, it seems reasonable to keep the full model structured by square, since:
  - o Structuring by department does not seem relevant;
  - o Even if the simple model is chosen for the greatest number of species, a higher level of complexity is selected for almost half of the species

Step 2: Fixed effects selection

As in the previous step, the performance of the different models was measured using the BIC. The results are shown in Table 6. Species for which at least one of the models had a convergence problem were set aside for this analysis. In total, the table shows results for 113 STOC bird species and 146 Vigie-Flore plant species. The large number of species discarded for Vigie-Flore is explained by the spatial variables, which caused convergence problems for 79 species in the spatial model, and 98 species in the full model. In addition, VIFs were used to verify the absence of multi-collinearity (maximum VIF values = 1.24 for STOC and 1.52 for Vigie-Flore, confirming there was no problem in including all covariates in the models).

***Table 6** : Number of species for which each model obtains the lowest BIC*

		Simple *model*	***Condition*** *model*	***Spatial*** *model*	***Land-use*** *model*	***Full*** *model*	*Total*
Number of species with lowest BIC	STOC	26 (23%)	17 (15%)	7 (6%)	26 (23%)	**37 (33%)**	113 (151)
	Vigie-Flore	4 (2%)	**160 (92%)**	0 (0%)	10 (6%)	0 (0%)	174 (300)

For Vigie-flore, the “condition” model was selected as the best model for the majority of species (92%). Adding landscape or spatial variables does not therefore seem to improve model fit. On the other hand, the date on which the squares were visited provides additional information compared with the model without covariates. In view of the large number of species for which one of the “complete” or “spatial” models failed to converge for Vigie-flore (probably due to redundant information between the random plot effect and the spatial/land-use covariates), the approach was replicated by comparing only the “simple”, “conditions” and “land-use” models. This allows the inclusion of more species (284) and confirms the result: the “condition” models are selected as the best models in 90% of species in this case.

For STOC, the results are less clear-cut. Models with fixed covariates are selected for over three-quarters of species, with the full model selected for a third of species. The model without fixed covariates (random effects only) is still selected for almost 25% of species. It should be noted that the geographical variables here seem to be partly confused with the random department effect, resulting in a number of species that do not converge. If we follow the same approach as for Vigie-flore, i.e. compare only models without spatial variables, then the “land use” model is selected for 50% of species, which strongly encourages to retain the most complex model.

Following the same reasoning as above:

- → For Vigie-flore, it seems appropriate to keep the condition model, which is chosen for the majority of species;
- → For STOC, the complete model can be reasonably chosen, as models with at least the observation conditions and landscape or spatial variables as covariates are selected for more than half the species;

Conclusion

Our two-steps procedure allowed us to select which covariates were best suited for most species, in each protocol. This leads to two different choices depending on the monitoring scheme. For STOC, we retained the most complex model regarding both the fixed and the random part. We thus will include a random intercept and a random year slope on squares, structured by administrative department, while also accounting for all covariates related to conditions of observation, spatial distribution and land-cover. For Vigie-flore, we retained a much simpler model: though we decided to keep the full random structure, except that we will not take into account the administrative department structuration. Regarding the fixed part, including only the observation date seems sufficient.

It is worth noting, that for species in which the BIC favours less complex models, the use of a more complex model should not be a problem for several reasons. In our case, most species gather large enough number of observations so that these additional parameters can be estimated without severely impacting estimates precision. Also, adding parameters that carry no relevant information for explaining abundance (resp. occurrence) should not bias estimates, only impact statistical power. Finally, using a more complex model does not necessarily prevent model convergence. If so, the analysis pipeline associated with this methodology includes the use of simpler models in case of non-convergence.