Received: 13 April 2018 | Revised: 12 September 2018 | Accepted: 14 September 2018

DOI: 10.1111/1755-0998.12949

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Conditionally autoregressive models improve occupancy analyses of autocorrelated data: An example with environmental DNA

Wentao Chen[1] | Gentile Francesco Ficetola[1,2]

[1]Laboratoire d'Écologie Alpine (LECA), CNRS, Univ. Grenoble Alpes, Grenoble, France

[2]Department of Environmental Science and Policy, Università degli Studi di Milano, Milano, Italy

Correspondence
Wentao Chen, Laboratoire d'Écologie Alpine (LECA), CNRS, Univ. Grenoble Alpes, Grenoble, France.
Email: wentao.chen.ecol@gmail.com

Funding information
Labex OSUG@2020, Grant/Award Number: ANR10 LABX56; Horizon 2020 Programme, Grant/Award Number: 772284

Abstract

Site occupancy-detection models (SODMs) are statistical models widely used for biodiversity surveys where imperfect detection of species occurs. For instance, SODMs are increasingly used to analyse environmental DNA (eDNA) data, taking into account the occurrence of both false-positive and false-negative errors. However, species occurrence data are often characterized by spatial and temporal autocorrelation, which might challenge the use of standard SODMs. Here we reviewed the literature of eDNA biodiversity surveys and found that most of studies do not take into account spatial or temporal autocorrelation. We then demonstrated how the analysis of data with spatial or temporal autocorrelation can be improved by using a conditionally autoregressive SODM, and show its application to environmental DNA data. We tested the autoregressive model on both simulated and real data sets, including chronosequences with different degrees of autocorrelation, and a spatial data set on a virtual landscape. Analyses of simulated data showed that autoregressive SODMs perform better than traditional SODMs in the estimation of key parameters such as true-/false-positive rates and show a better discrimination capacity (e.g., higher true skill statistics). The usefulness of autoregressive SODMs was particularly high in data sets with strong autocorrelation. When applied to real eDNA data sets (eDNA from lake sediment cores and freshwater), autoregressive SODM provided more precise estimation of true-/false-positive rates, resulting in more reasonable inference of occupancy states. Our results suggest that analyses of occurrence data, such as many applications of eDNA, can be largely improved by applying conditionally autoregressive specifications to SODMs.

KEYWORDS
conditionally autoregressive model, sedimentary DNA, spatial autocorrelation, species occupancy-detection model, temporal autocorrelation, true skill statistics

1 | INTRODUCTION

Inferring species occurrence is fundamental in biodiversity studies. However, methods for species detection are often imperfect, as they can fail to detect present species and can even be prone to false positives (Guillera-Arroita, 2017; Royle & Link, 2006). Failing to take into account these errors can lead to biased inference; therefore, in the last years, there was a development of site occupancy-detection models (SODMs) to deal with such issues. SODMs are able to estimate occupancy rates with imperfect detection (Mackenzie et al.,

2002) and have been generalized to deal with false positives (Chambert, Miller, & Nichols, 2015; Miller et al., 2011; Royle & Link, 2006). In short, SODMs estimate the probability that a sample is occupied by a species given its detections on that sample, by considering the occupancy probability and true-positive (TP) and false-positive (FP) probabilities over the whole sample set. SODMs were originally developed to deal with the issues of traditional field surveys, but in the last years, they are increasingly applied to a wide range of environmental data, including environmental DNA data (eDNA; Dorazio & Erickson, 2018; Ficetola et al., 2015; Lahoz-Monfort, Guillera-Arroita, & Tingley, 2016; Schmidt, Kéry, Ursenbacher, Hyman, & Collins, 2013).

As with all the biodiversity data, eDNA generally does not detect all the present taxa (false negatives [FNs]), but is also subjected to FPs, which can arise through multiple processes such as contamination and PCR/sequencing errors (Ficetola et al., 2015). In many studies, scientific inference relies on the accurate assessment of occupancy state (presence or absence) of a given species in a particular sample, and both false presences and false absences can lead to severe issues (Ficetola, Taberlet, & Coissac, 2016). For example, in a study aiming at dating the introduction of alien species (e.g., Ficetola, Poulenard, et al., 2018; Sjögren et al., 2017), FPs of the target species may increase the uncertainty of date estimates and even lead to misleading conclusions.

With traditional field survey data, SODMs can integrate calibration design (Chambert et al., 2015) or unbiased detections (Chambert et al., 2015; Miller et al., 2011) to account for detection errors. However, these approaches demand extra sampling efforts and may even be unfeasible for some eDNA applications, such as when eDNA is used to assess biodiversity in ancient samples (Capo, Debroas, Arnaud, & Domaizon, 2015; Epp et al., 2015; Giguet-Covex et al., 2014; Pansu et al., 2015). In such cases, FPs must be estimated from the detection data set itself, even though negative and positive can provide some prior information on the frequency of these errors (Parducci et al., 2017).

Site occupancy-detection models are increasingly used to analyse ecological data, including eDNA (Pansu et al., 2015). These models generally assume independence of observations and thus do not take into account the spatial or temporal autocorrelation among samples. However, ecological variables commonly have some form of internal dependence, such as autocorrelation in space or time (Dormann, 2007; Roberts et al., 2017). When autocorrelation is present, samples that are nearby in space or in time often have similar occupancy because of both intrinsic processes (e.g., dispersion) and extrinsic mechanisms which in turn show autocorrelation (e.g., climate forcing; Beale, Lennon, Yearsley, Brewer, & Elston, 2010; Ficetola, Manenti, De Bernardi, & Padoa-Schioppa, 2012; Wagner & Fortin, 2005). Autocorrelation is particularly important in ecological time series (Legendre & Gauthier, 2014; Legendre & Legendre, 2012), and ignoring the correlation between nearby samples violates a central assumption of many statistical methods and may result in biased inferences.

Approaches exist to integrate autocorrelation into SODMs, which have been sometimes applied to traditional biodiversity surveys. To name a few, Sargeant, Sovada, Slivinski, and Johnson (2011) showed that Markov random fields (multidimensional extensions of Markov chains) can allow modelling spatial dependencies in estimating species distribution. Royle and Dorazio (2008) proposed the autologistic structure to account for temporal autocorrelation; Hines et al. (2010) used Markov process to model the probability of presence of tiger along consecutive trail segments; while Aing, Halls, Oken, Dobrow, and Fieberg (2011) modelled the occupancy of river otter on snow track with an intrinsic conditional autoregressive (iCAR) term. However, despite the wide awareness of the potential biases introduced by the autocorrelation in spatially or temporally structured data, most eDNA surveys using SODMs to analyse such data did not explicitly consider autocorrelation.

Here, we demonstrate the needs and benefits to consider spatial and temporal autocorrelation in eDNA biodiversity surveys. We begin with an analysis of the literature to identify the frequency of SODM analyses in the eDNA literature and to assess to what extent conditionally autoregressive SODM could improve analyses. Then, we show how SODMs with explicit autocorrelation structure can improve occupancy analyses on spatially or temporally structured eDNA data. To this end, we added a conditionally autoregressive (CAR) component into the Royle and Link (2006) SODM to take into account that occupancy can be affected by autocorrelation. The Royle and Link (2006) model can be applied to survey data without calibration designs or unbiased detection data, and also takes into account the probability of FPs, which makes it suitable for general eDNA-based surveys (Ficetola et al., 2015; Lahoz-Monfort et al., 2016). CAR models had been demonstrated to be useful to model spatial autocorrelation in species distribution data, are easy to implement and interpret (Dormann, 2007) and, using CAR or closely related restricted spatial regression models to account for autocorrelation in occupancy models, can yield satisfying results (Aing et al., 2011; Johnson, Conn, Hooten, Ray, & Pond, 2013). Our approach can be particularly important in eDNA studies, where contamination can occur (Parducci et al., 2017), and at the same time, TP rates can be low (Epp et al., 2015), as a result of low DNA content in samples; nevertheless, similar issues also apply to other approaches to biodiversity assessment. We assessed the capacity of a CAR approach to improve the performance of SODM models and to better estimate true occupancy. We tested and compared the CAR-SODM with the original Royle and Link (2006) SODM by applying them to (a) simulated chronosequences of occupancy data, (b) simulated spatial samples and (c) true eDNA data sets from the literature. Our study identifies the cases in which the CAR-SODMs perform better than the original SODMs, and provides easy-to-follow instructions to implement CAR-SODMs.

2 | MATERIALS AND METHODS

2.1 | Analysis of the literature

To elucidate the issues of data autocorrelation in current literature of eDNA biodiversity studies that involved species occupancy

modelling, we collected relevant studies from the Web of Science database (June 2018), using search keywords "DNA occupancy NEAR/2 model" and "DNA detection probability." Each resulting study was screened based on these criteria: (a) whether it reported empirical eDNA data; (b) whether the data showed a temporal or spatial structure; (c) whether studies applied SODMs on those data; and (d) whether the SODMs used took autocorrelation into account if the data were temporally or spatially structured.

2.2 | CAR-SODM

We adopted here an extended SODM based on the Royle and Link (2006) model, which can be specified in a Bayesian framework as follows:

$$z_i \sim \text{Bernoulli}(\psi_0) \tag{1}$$

$$y_i \sim \text{binomial}(z_i p_{11} + (1 - z_i) p_{10}, K) \tag{2}$$

where i = 1, 2, …, n, n being the total number of samples; ψ_0 is the overall occupancy probability in all samples; z_i is the state of occupancy in sample i (z_i = 1 when the site is occupied, otherwise z_i = 0); y_i is the number of positive detections in sample i; p_{11} and p_{10} are the TP and FP probabilities, respectively; and K is the number of replicated observations. For instance, in eDNA studies, K can be the number of PCRs or of replicated DNA extractions performed on a given sample (Ficetola et al., 2015). We note that the likelihood function of this model would have two equally high peaks, because of the symmetry between $z_i \times p_{11}$ and $(1 - z_i) \times p_{10}$ in (Equation 2). In other words, for a given data set, a parameter set with $z_i = z'_i$, $p_{11} = p'_{11}$ and $p_{10} = p'_{10}$ has the same likelihood as another set with $z_i = 1 - z'_i$, $p_{11} = p'_{10}$ and $p_{10} = p'_{11}$. Therefore, there are several solutions for a given data set of y_i, so that additional data or restrictions on parameters are needed to get unambiguous estimation. As a result, for models with a TP/FP setting, additional data are needed to break such symmetry and obtain an unambiguous solution (Guillera-Arroita, Lahoz-Monfort, van Rooyen, Weeks, & Tingley, 2017; Lahoz-Monfort et al., 2016; Royle & Link, 2006).

In the CAR-SODM here proposed, instead of treating ψ_0 as a constant parameter over the whole sample set, we note the occupancy probability at sample i as ψ_i, which can take different values for different is. To model the autocorrelation of ψ_i, we assume that ψ_i can be expressed by a baseline constant ψ_b and an autoregressive term φ_i drawn from a multivariate normal distribution (Jin, Carlin, & Banerjee, 2005):

$$\psi_i = \text{logistic}(\psi_b + \varphi_i) \tag{3}$$

$$\varphi_i \sim \text{multinormal}\left(0,\ \sigma^2[\text{Diag}(\mathbf{m}) - \alpha\mathbf{W}]^{-1}\right) \tag{4}$$

The term $\sigma^2[\text{Diag}(\mathbf{m}) - \alpha\mathbf{W}]^{-1}$ is the covariance matrix of the multivariate normal distribution. Diag(**m**) is an n by n diagonal matrix with $\mathbf{m} = m_i$, that is, the number of neighbours for sample i as its diagonal elements. **W** is the adjacency matrix: W_{ij} = 1 if $i \neq j$ and samples i and j are neighbours, otherwise W_{ij} = 0. The α parameter measures the degree of dependence among samples. α = 0 corresponds to a model without temporal/spatial dependence. σ^2 is the variance parameter to the multivariate normal distribution (Jin et al., 2005; Wall, 2004).

2.3 | Parameter estimation

One of our aims was using CAR-SODM to better estimate the true occupancy state on the basis of detection data, for instance by assessing the probability that a sample i is actually occupied, given the number of positive detections in that sample. This can be calculated from other parameters specified above (Lahoz-Monfort et al., 2016):

$$\begin{aligned} P_i(\text{occupied}|y) &= P(z_i = 1|y_i = y) \\ &= \frac{\psi_i p_{11}^y (1 - p_{11})^{K-y}}{\psi_i p_{11}^y (1 - p_{11})^{K-y} + (1 - \psi_i) p_{10}^y (1 - p_{10})^{K-y}} \end{aligned} \tag{5}$$

In this study, we compare the CAR-SODM with the Royle and Link (2006) model by observing their behaviour and performance. We estimated all parameters involved in the CAR-SODM (p_{11}, p_{10}, φ_i, ψ_b, τ and α) and in the Royle and Link model (p_{11}, p_{10} and ψ_0) in a Bayesian framework. For both models, uniform priors were used for most parameters with boundaries adjusted to their proper definitions ((0, 1) for probabilities and (−1, 1) for the autocorrelation parameter α). For p_{10}, we used a uniform prior between 0 and 0.1 to reflect prior information obtained from positive controls (Pansu et al., 2015) and to break the symmetry between p_{11} and p_{10} in the likelihood function (Lahoz-Monfort et al., 2016). For σ^2, we used a γ(2, 1) prior. Sampling was performed by using Markov chain Monte Carlo (MCMC) sampling provided by the Stan statistical computation platform from the *RStan* interface (Stan Development Team, 2016). Each sampling run used three MCMC chains for 10,000 iterations, the first of which 5,000 were discarded as burn in. The convergence of chains was checked by the Gelman–Rubin statistic (Gelman, Rubin, Gelman, & Rubin, 1992). In all models, input data included number of replicates for each sample K, the observed number of positive detections y, the adjacency matrix **W** and the corresponding m_i. In fact, **W** can be specified according to the sampling scheme for each application.

2.4 | Simulations: Chronosequences

A first set of simulated data mimicked the occupancy data obtained by the analysis of temporal series, such as eDNA data obtained from the analysis of lake sediments. We generated the chronosequence simulation data based on a combination of the autoregressive process (AR1) and the binomial process. In each simulation replicate, the sample numbers corresponded to their chronological order. First, a series of $X_t = -0.005 + \alpha_c X_{t-1} + \varepsilon_t$ was generated, where α_c is the autoregressive parameter and ε_t is an error term drawn from the standard normal distribution. The probability of occupancy ψ_t was obtained by applying the inverse-logit function to X_t. The detection count y_t was then generated as in (Equation 2). We additionally restricted the proportion of positive z_i between 5% and 95% to avoid unrealistic extreme cases. Different values of TP probability

p_{11} (0.15, 0.25 and 0.4) and FP probability p_{10} (0.005, 0.015 and 0.03) were applied to the simulations. We kept p_{11} at relatively low levels because these are the most challenging cases of in real-world studies, and large p_{11}s (≥0.5) allow to obtain satisfying results even by applying the Royle and Link (2006) model (Ficetola et al., 2015; Lahoz-Monfort et al., 2016). Besides, small p_{11}s are common in both classical monitoring and in eDNA studies, especially in ancient DNA as a result of degradation (Ficetola, Romano, et al., 2018; Giguet-Covex et al., 2014; Pansu et al., 2015; Parducci et al., 2017). For each parameter scenario (p_{11}, p_{10}, K, α_c), 100 simulation replicates were performed. We chose three K levels (4, 8 and 12) to reflect common practices in eDNA analyses (Ficetola et al., 2015). Three levels of α_c (0.0001, 0.5, 0.95) were used to represent realistic levels of temporal autocorrelation (see Results), ranging from practically non-correlated (α_c = 0.0001) to highly correlated data (α_c = 0.95). Sample size S was set to 100 to represent typical ancient DNA data sets. Both the CAR-SORM and the Royle and Link (2006) models were applied to the simulation data set. To build the adjacency matrix of the CAR-SODM, **W** was specified as such that consecutive samples were considered to be chronologically adjacent.

2.5 | Simulations: Spatially autocorrelated data set

The data sets representing spatial data with autocorrelation were generated in two steps. First, we generated artificial distribution data for a virtual island using the SNOUTER data set (Dormann et al., 2007). The original data set consists of the *X*-*Y* coordinates in integers of more than 1000 sites and two uncorrelated environmental variables for each site: "rain" and "djungle." To reflect real-world applications, we used only a subset of sites for simulations (i.e., only the sites with both coordinates that were multiples of 3 were retained in the simulations, resulting in 120 sites). Following the authors' instructions, we calculated the presence probability p_i for a site i as

$$p_i = q_i + \varepsilon_i\sqrt{q_i(1-q_i)},\ \text{where}\ q_i = \frac{e^{3-0.004\times \text{rain}}}{1+e^{3-0.004\times \text{rain}}}.$$

To produce the presence/absence data, we set z_i = 1 if $p_i \geq 0.5$ and z_i = 0 if $p_i < 0.5$. The resulting data are spatially correlated because "rain" is essentially determined by altitude with a rain shadow in the east (the SNOUTER data set itself is based on the digital elevation model of a volcano in New Zealand). Second, detection data y_i were generated by applying (Equation 2) to z_i. The same sets of parameter scenarios for p_{11}, p_{10} and K as in the chronosequence simulation section were used to generate the data sets. In the CAR-SODM, the adjacency matrix **W** was defined with a rook scheme.

2.6 | Model performance comparison

We first assessed the ability of models to correctly infer p_{11} and p_{10}. Furthermore, in many applications of SODMs, users are interested in obtaining better information on the actual occupancy state of a sample (e.g., what is the probability that a given site is occupied, if I have only one detection over eight sampling replicates?). We thus obtained model estimations of the conditional probability of occupancy of samples p_i (Equation 5). To evaluate model performances in assessing the actual states of the simulation samples, we then calculated the area under the curve of the receiver operating plot (AUC) and the maximum true skill statistic (TSS). AUC is a threshold-independent measure for score-ranking models (Bradley, 1997) that reflects a model's overall classification accuracy. AUC values range from 0 to 1. Usually, AUC values of 0.5–0.7 indicate low accuracy, values of 0.7–0.9 indicate useful applications, and values of >0.9 indicate high discrimination capacity (Swets, 1988). However, AUC has some limitations (Lobo, Jiménez-valverde, & Real, 2008); thus, we used the maximum TSS as a complementary measure of performance. TSS is a simple and intuitive, threshold-dependent measure of accuracy in predicting presence–absence (Allouche, Tsoar, & Kadmon, 2006). TSS values range from −1 to 1. TSS values of 1 indicate a perfect discrimination, while TSS values equal to or less than zero indicate a performance no better than a random model (Allouche et al., 2006). Because different statistical models can have different optimal prediction thresholds (Bradley, 1997), we maximized the TSS by varying the prediction threshold on the receiver operating characteristic (ROC) curves for each model, in a similar manner to a previous study aimed at finding the optimal detection threshold for eDNA qPCR assays (Serrao, Reid, & Wilson, 2017).

2.7 | Analysis of empirical data

To demonstrate the usefulness of conditionally autoregressive models, we applied the model to published eDNA data available from the literature search. Specifically, we analysed a subset of chronosequences in the Lake Anterne sedimentary ancient DNA data set (Giguet-Covex et al., 2014; Pansu et al., 2015), as well as spatially structured survey data of animal species based on modern eDNA collected from water samples. We considered the studies obtained from the literature search that provided all the data required to apply the CAR-SODM model (i.e., geographic coordinates of sampling sites and detection histories at sites; Dougherty et al., 2016; Ostberg, Chase, Hayes, & Duda, 2018; Vörös, Márton, Schmidt, Gál, & Jelić, 2017). For the ancient DNA data set, we chose six molecular operative taxonomic units (MOTUs) representing two animal (*Bos* and *Ovis*) and four plant taxa (*Achillea macrophylla*, *Alchemilla* MOTU, *Hypericum* MOTU and *Pinus* MOTU) from the data set to evaluate the behaviour of both models for taxa with different features. The Dougherty et al. (2016) data set contains eDNA of rusty crayfish (*Orconectes rusticus*) collected from water samples from 12 lakes in Wisconsin, USA. We analysed the whole eDNA data set, considering multiple collection sites in the same lake as adjacent. The Vörös et al. (2017) data set contains eDNA data of olm salamander (*Proteus anguinus*) from water samples collected from 15 cave systems in Croatia. We analysed the eDNA data obtained with precipitation method, using the Gabriel graph criterion (Legendre & Legendre, 2012) to determine site connectivity because it is an approximation for an irregular spatial distribution to the rook scheme for a regular

grid (Bini et al., 2009). The Ostberg et al. (2018) data set contains eDNA of pacific lamprey (*Entosphenus tridentatus*) and *Lampetra* spp from water samples collected from watersheds in Puget sound, USA. We analysed the eDNA data of the 2015 spring survey, considering consecutive sites on the Puget sound watershed line as adjacent. Settings of the data sets (spatial/temporal structure, sample size S and number of PCR replicates K) can be found in Table 2.

All simulations and analyses were performed in the R statistical programming environment (version 3.3.3; R Core Team, 2017). ROC curves and AUC values were generated by using the *pROC* package (Robin et al., 2011). Spatial weights matrices used in generating the spatially autocorrelated data set were created using the *spdep* package (Bivand, Hauke, & Kossowski, 2013; Bivand & Piras, 2015).

3 | RESULTS

3.1 | Literature analysis

The Web of Science database returned 134 journal articles as search results. We discarded all articles that did not directly discuss eDNA and those that did not report eDNA data from field surveys, including pure simulation studies, theoretical analyses and those that focused on experimental tests on DNA detectability under controlled conditions. The remaining 51 articles, dating from 2008 to 2018 were analysed in detail (Table 1).

Spatial or temporal structures were frequent, as they were present in 35 of 51 studies (69%). Twenty seven of 51 studies (52%) used SODM to analyse eDNA data. However, just four of 27 (15%) of them took into account FPs in their SODMs, and only two of 23 (9.5%) of studies using SODMs on spatially or temporally structured data considered the autocorrelation issue (Table 1, see also Supporting Information Table for the list of the studies considered).

TABLE 1 Analysis of the literature using eDNA for biodiversity assessment. The table shows the number of studies that analysed eDNA survey data with different typologies of autocorrelation (spatial/temporal). We report whether papers analysed data using site occupancy-detection models (SODM), considering autocorrelation and taking into account false positives. Numbers in parentheses indicate the number of studies using SODMs to model the effects of covariates on occupancy or detection probability

		SODM			
Structure	No SODM	Autocorrelation only	False positive only	Both	Neither
Spatially structured	11	0	0	1 (1)	18 (17)
Temporally structured	1	0	0	1 (0)	0
Both	0	0	0	0	3 (2)
Neither	12	0	2 (1)	0	2 (0)
Total	24	0	2 (1)	2 (1)	23 (19)

3.2 | Simulations: Chronosequences

With the non-autoregressive models, the values of TP probability p_{11} were slightly overestimated when the number of replications K was small and detection probability was low, while estimates of the autoregressive model were less biased and more precise (Figure 1). FP probability p_{10} and autocorrelation did not severely affect the estimation of p_{11} with both models. In scenarios with large K ($\geq$8) or large p_{11} ($\geq$0.25), both models estimated TP probability with high accuracy. The autoregressive model also showed lower bias in the estimation of FP probability p_{10}, even though both approaches tended to overestimate p_{10}, unless a large number of replicates per sample was available and TP probability was high (Figure 2).

The model performance in assessing the actual state (presence/absence) of the samples was generally fair to good, with AUC values generally >0.7, and TSS values often >0.5 (Figure 3, Supporting Information Figure S1). With both metrics, performance was better when TP probability was high and when p_{10} was low (Figure 3, Supporting Information Figure S1). In most scenarios, the autoregressive model outperformed the non-autoregressive one. CAR-SODM provided a particularly high improvement compared to the Royle and Link (2006) model when autocorrelation was strong, when TP probability was low and when a few replicates were performed. In scenarios with high autocorrelation and low replication level, the autoregressive model outperformed its non-autoregressive counterpart by more than 10% in terms of AUC, even with relatively high TP probability (p_{11} = 0.4). Both models showed excellent performance (AUC and TSS close to one) if many replicates were performed and TP probability was high (Figure 3, Supporting Information Figure S1).

3.3 | Spatial data simulations

The analysis of data with spatial autocorrelation yielded results consistent with the ones of the chronosequence simulations. In comparison with the Royle and Link (2006) model, the CAR-SODM provided more accurate estimates of p_{11} and p_{10}, particularly in scenarios of low TP probability and with a few replicates (Figure 4). In terms of TSS and AUC, the autoregressive model outperformed the non-regressive one in most scenarios, except in scenarios of high p_{11}, K and low p_{10}, where both models reached excellent performance (Figure 4, Supporting Information Figure S2).

3.4 | Analysis of empirical eDNA data

We then applied both the CAR-SODM and the Royle and Link (2006) model to multiple real-world data sets. When we analysed the eDNA data from two mammal and four plant MOTUs from 44 samples of lake sediment (Giguet-Covex et al., 2014; Pansu et al., 2015), the CAR-SODM detected strong positive autocorrelation for all taxa (estimated $\alpha \geq 0.78$). The CAR-SODM tended to estimate lower values of both p_{10} and p_{11} than the Royle and Link (2006) model. Furthermore, highest probability density intervals (HPDIs) for

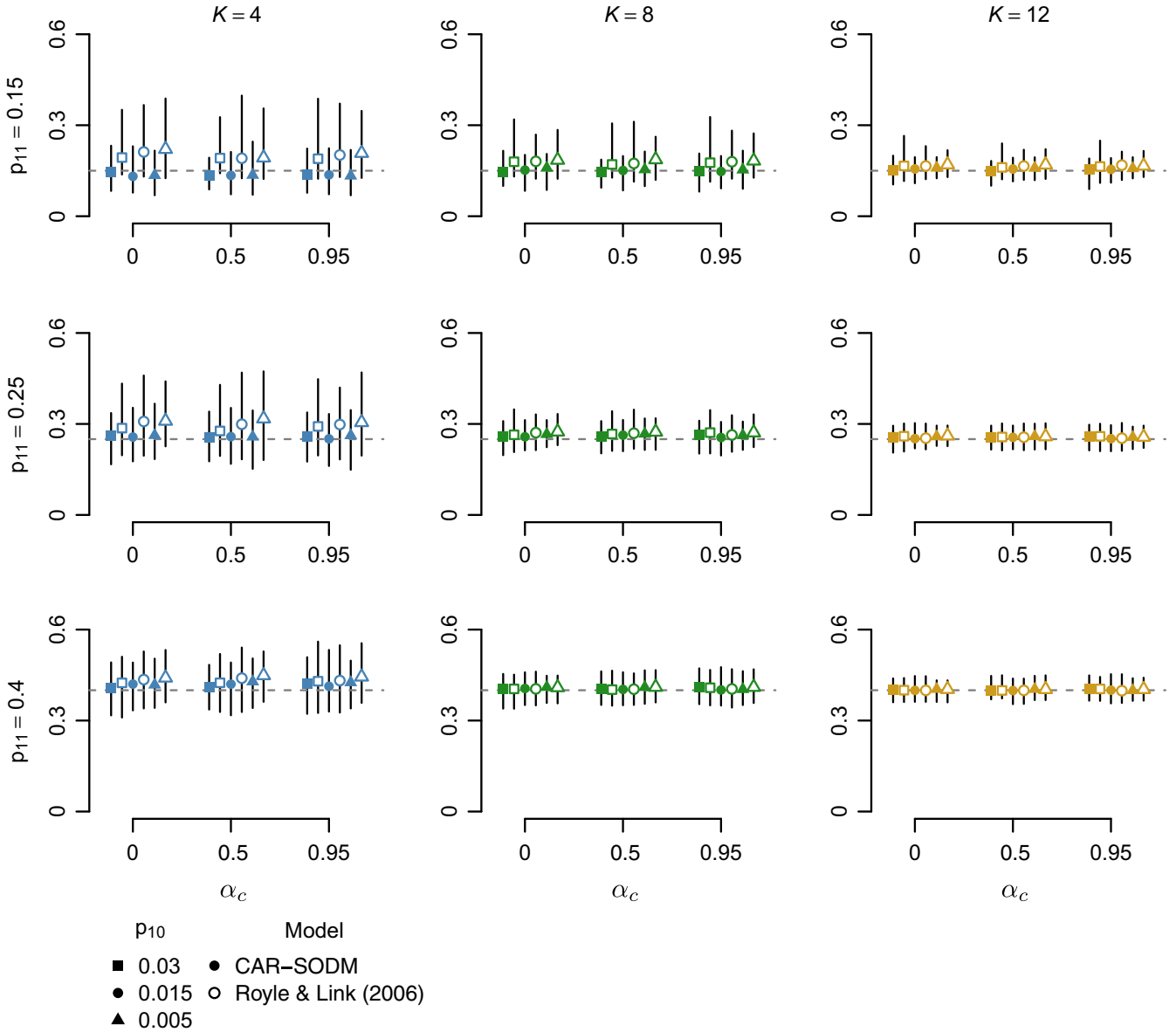

FIGURE 1 Estimations of true-positive (TP) probabilities p_{11} for the eDNA chronosequence simulated data sets by the CAR-SODM (full symbols) and by the Royle and Link (2006) model (open symbols). Results are arranged by scenarios of p_{11}, autocorrelation level α_c and false-positive (FP) probability p_{10}. Plots show medians from 100 simulations. Vertical lines cover the 2.5%–97.5% quantiles. Dashed lines indicate the actual p_{11} levels at 0.15 (top), 0.25 (middle) and 0.4 (bottom) [Colour figure can be viewed at wileyonlinelibrary.com]

corresponding parameters tended to be narrower for CAR-SODM, suggesting lower incertitude of estimates. CAR-SODM models were also applied to spatially structured data sets of modern eDNA collected from water samples. When we analysed the data set on crayfish eDNA (Dougherty et al., 2016), we obtained results analogous to those of the ancient DNA, with strong positive autocorrelation (α = 0.932) and lower p_{10} and p_{11} estimated by the CAR-SODM than by the Royle and Link (2006) model. Conversely, the CAR-SODM did not detect significant autocorrelation in the olm salamander (Vörös et al., 2017) and in the lamprey (Ostberg et al., 2018) eDNA data sets (HPDIs of α included zero). In these two data sets, both models yielded nearly identical results in the estimates of p_{10} and p_{11} (Table 1).

Posterior conditional probabilities of occupancy were not identical between CAR-SODM and the Royle and Link (2006) models. For a given number of positive amplifications, CAR-SODM tended to give more support to the samples nearby to other positive detections than to the ones in isolation (Figure 5), while the Royle and Link (2006) model assigned the same probability of occurrence to all samples sharing the same number of positive amplifications. This discrepancy between the two models was most evident for samples with a single positive amplification. For example, in the lake sediment data set, only 1/8 PCR detected *Bos* (i.e., cattle) eDNA within the sample at 343 cal. years BP. The CAR-SODM assigned to this sample a high (>0.9) posterior conditional probability of occupancy, as multiple positive detections occurred in the same period, while

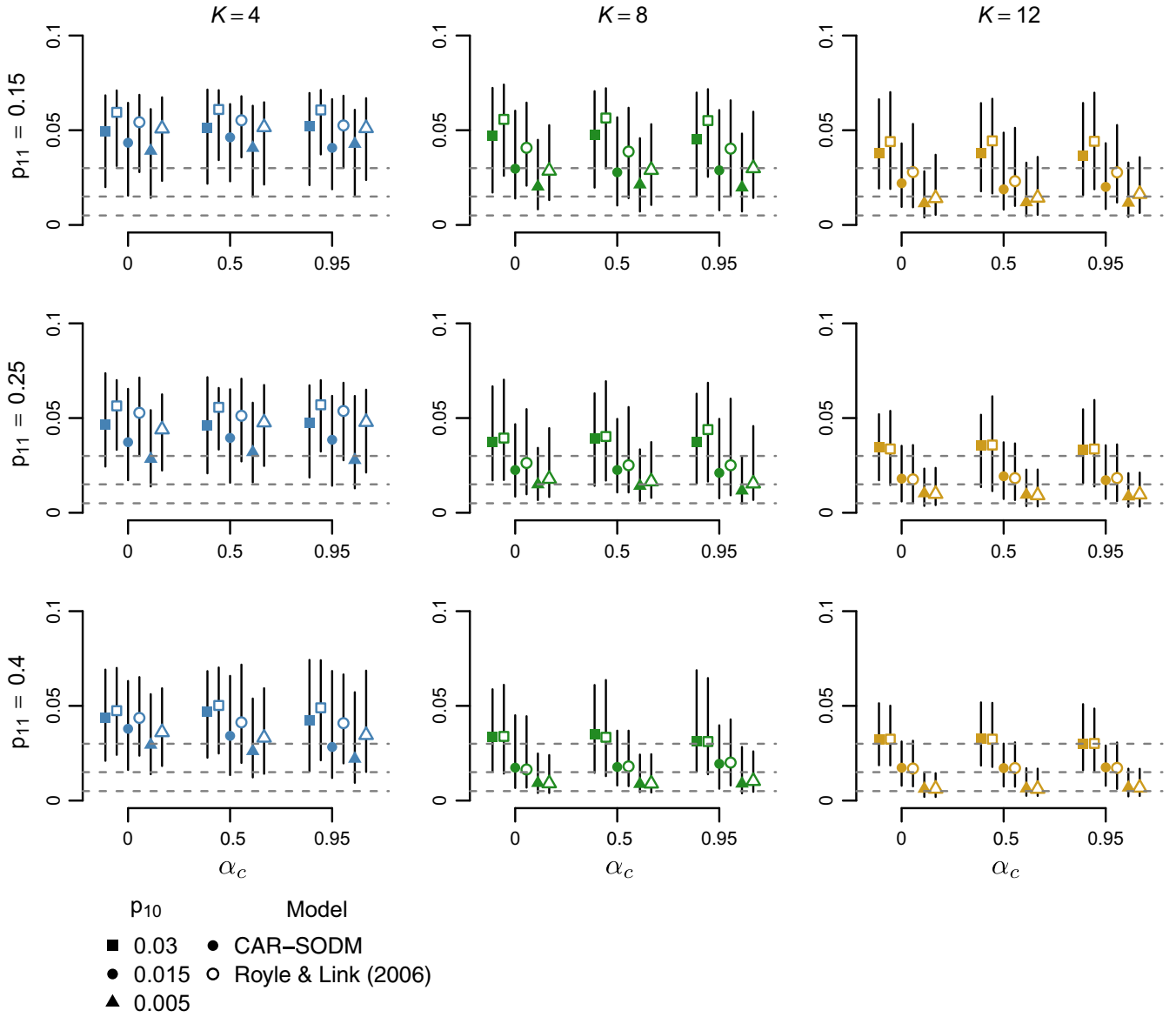

FIGURE 2 Estimations of false-positive (FP) probabilities p_{10} for the eDNA chronosequence simulated data sets by the CAR-SODM (full symbols) and by the Royle and Link (2006) model (open symbols). Results are arranged by scenarios of true-positive (TP) probability p_{11}, autocorrelation level α_c and (FP) probability p_{10}. Plots show medians from 100 simulations. Vertical lines cover the 2.5%–97.5% quantiles. Dashed lines indicate the actual p_{10} levels at 0.005, 0.015 and 0.03 [Colour figure can be viewed at wileyonlinelibrary.com]

the Royle and Link (2006) model assigned a lower probability of occupancy (0.56) to the same sample (Figure 5c).

4 | DISCUSSION

It is well known that both spatial and temporal autocorrelation can introduce biases to ecological data analyses (Beale et al., 2010; Brown et al., 2011; Lennon, 2000), and specifically, to species occupancy modelling (Johnson et al., 2013). Despite repeated calls (Ficetola et al., 2015; Lahoz-Monfort et al., 2016; Schmidt et al., 2013), only half of studies used SODM to analyse eDNA data, and few of them considered FPs in data analysis, though such issue might have been accounted for in their laboratory processing. Furthermore, until now very few studies using SODMs accounted for autocorrelation in spatially or temporally structured eDNA data, while the vast majority of papers did not report tests of autocorrelation before or after applying SODMs. Our study allows elucidating how biodiversity studies based on eDNA can benefit from SODMs taking both FPs and autocorrelation into account.

Detection errors are barely avoidable in biodiversity surveys, occur even with sessile species and can be particularly problematic in environmental DNA studies (Guillera-Arroita, 2017). In cases where unambiguous detections are not available, the Royle and Link (2006) model is usually recommended to account for FPs (Ficetola et al., 2015, 2016; Lahoz-Monfort et al., 2016; Lopes et al., 2017). However, the fact that ecological data sets often show temporal or spatial autocorrelation requires the application of appropriate approaches. Here, we show that a conditionally autoregressive

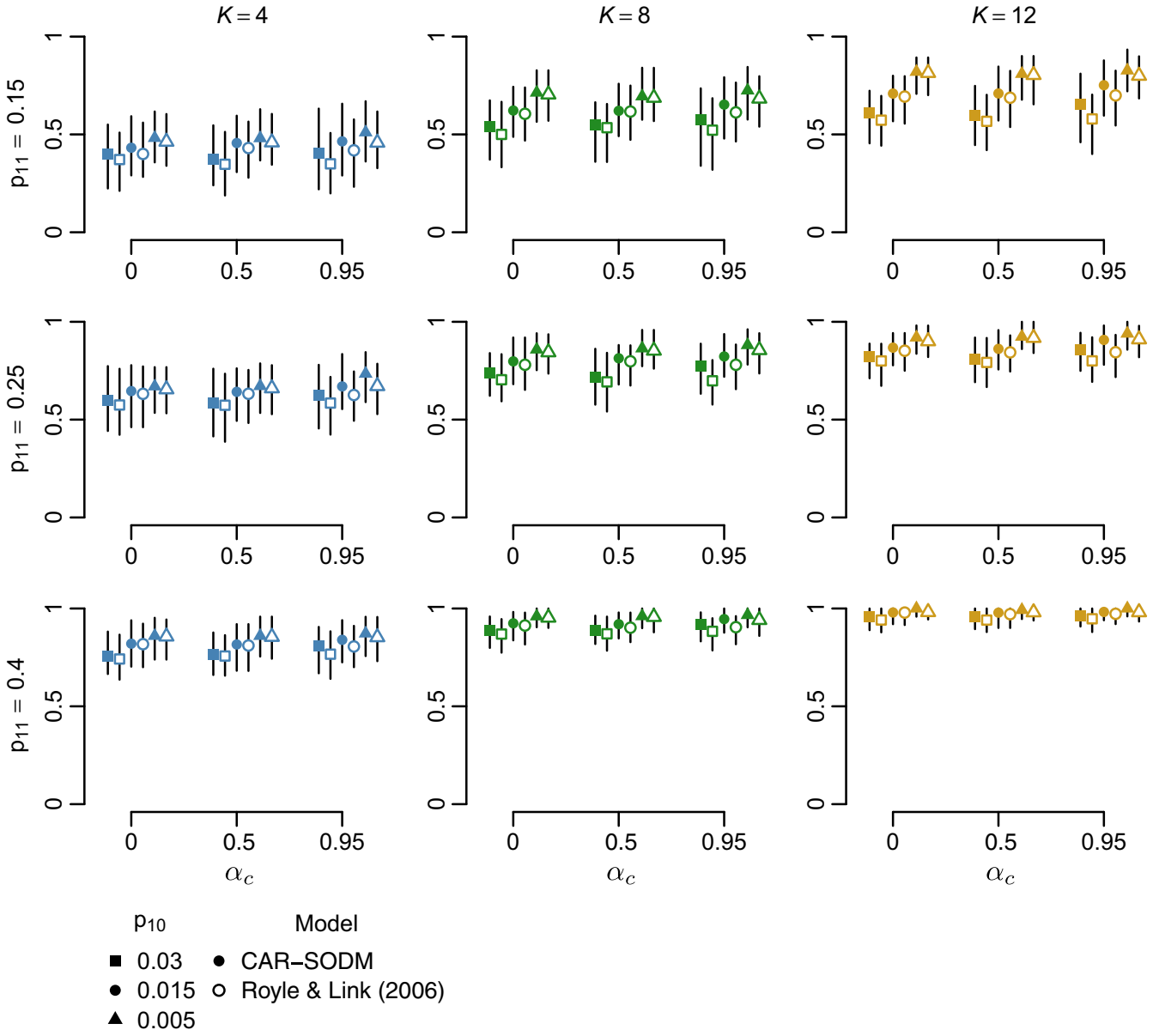

FIGURE 3 Maximum true skill statistic (TSS) for the CAR-SODM and the Royle and Link (2006) model, applied to the chronosequence simulated data sets. Results are arranged by scenarios of true-positive (TP) probability p_{11}, autocorrelation level α_c and false-positive (FP) probability p_{10}. Full symbols: CAR-SODM; open symbols: Royle and Link (2006). Plots show values calculated over 100 simulations. Vertical lines cover the 2.5%–97.5% quantiles [Colour figure can be viewed at wileyonlinelibrary.com]

model allows to successfully deal with autocorrelated occupancy data including both FPs and FNs and that considering the autocorrelation among samples provides an improved statistical inference that can be extremely helpful in biodiversity studies.

In eDNA studies, the estimation of model parameters such as TP/FP probabilities is important not only for correctly predicting site occupancy but also for evaluating the appropriateness of molecular protocols (e.g., Lopes et al., 2017) and for measuring data reliability (Ficetola et al., 2015). In the occupancy models previously proposed for eDNA studies, the occupancy probability ψ_0 is one of such parameters (Guillera-Arroita et al., 2017; Mackenzie et al., 2002; Miller et al., 2011; Royle & Link, 2006). However, if autocorrelation occurs, this parameter is not constant for all samples. In the CAR-SODM, a fixed latent occupancy is replaced by the combination between a fixed baseline occupancy term ψ_0 and a varying autoregressive term (Equation 3). As a result, the CAR-SODM was more flexible to fit autocorrelated data than its non-autoregressive counterpart and therefore less biased in estimating p_{11} and p_{10} (Figures 1 and 2). Both approaches tended to overestimate p_{10} in most scenarios, especially when K was small and p_{11} was low (Figure 2). The overestimation of p_{10} might be partially caused by the fixed uniform prior (0, 0.1), which may not be appropriate if the real p_{10} value is much smaller than the prior's upper limit 0.1 and if there is not enough information to determine p_{10} (e.g., when K is small). This

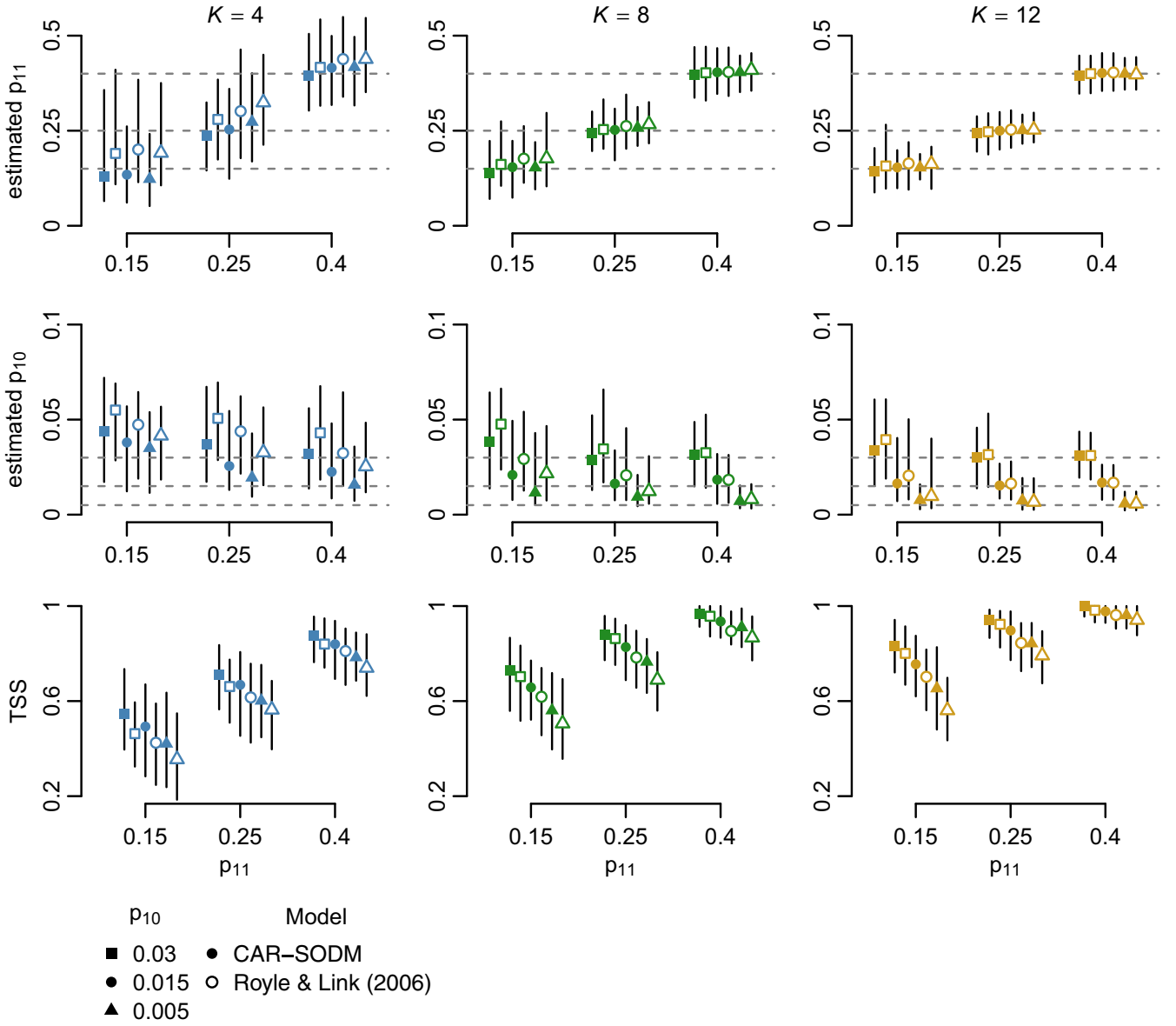

FIGURE 4 Estimations of true-positive (TP) probabilities p_{11} (top row) and of false-positive (FP) probability p_{10} (middle row), applied to the simulated data set with spatial structure, analysed with the CAR-SODM (full symbols) and by the Royle and Link (2006) model (open symbols). Plots show medians from 100 simulations. Vertical lines cover the 2.5%–97.5% quantiles. Bottom row: maximum TSS for the CAR-SODM and the Royle and Link (2006) model. All results are arranged by scenarios of p_{11} and p_{10}. Dashed lines indicate p_{11} levels at 0.15, 0.25 and 0.4 (top row), and p_{10} levels at 0.005, 0.015 and 0.03 (middle row) [Colour figure can be viewed at wileyonlinelibrary.com]

would suggest that the p_{10} values might have been also overestimated for the empirical data sets (Table 2). However, this is probably not a major issue, given that in the empirical data all of the FPs are extremely low or have intervals that include zero. In real-world applications, one can adjust the prior based on knowledge about the FP rate in question, and information obtained from negative controls can help to provide realistic boundaries to p_{10}. A different pair of boundaries can be set to the uniform prior; alternatively, a beta prior can be applied instead (Gelman, Carlin, Stern, & Rubin, 2004), with parameters as the numbers of FP replicates and of true-negative replicates in the control samples, when such information is available. Similarly, one can apply informative prior to other unknown parameters in the SODMs, potentially improving the accuracy and precision of their estimation. Additional analyses are required on the effects of prior specification for important model parameters, such as p_{11} and p_{10}. On the other hand, the baseline latent occupancy was poorly determined by the CAR-SODM (Table 2; note the large HPDIs), suggesting that the determination of this parameter demands a sample size much larger than here applied. This should not pose serious issues in most eDNA studies even with typical sample size (less than a hundred), if the estimation of this parameter is not the focus of the study. Conversely, if autocorrelation is weak (e.g., the salamander and lamprey data sets; Table 2), the CAR-SODM and the Royle and Link (2006) models provide highly

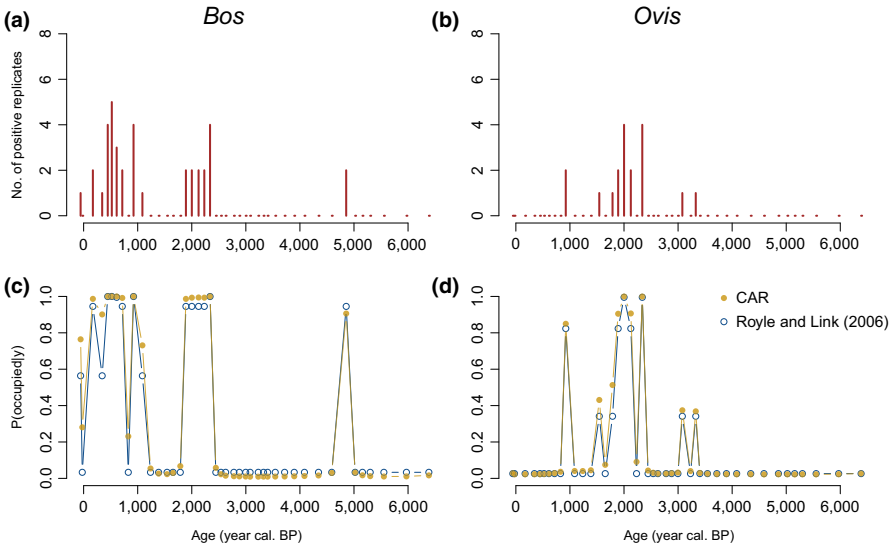

FIGURE 5 Applying SODMs to a temporally structured ancient mammal DNA data set (Giguet-Covex et al., 2014). (a and b) Numbers of positive PCR replicates of *Bos* and *Ovis* eDNA, respectively. (c and d) Corresponding estimated posterior conditional probabilities of occupancy, estimated by the CAR-SODM (full circles) and by the Royle and Link (2006) model (open circles) [Colour figure can be viewed at wile yonlinelibrary.com]

consistent results. In this case, the Royle and Link (2006) model can have advantages, as its convergence is faster and it does not have the issue of the estimation of baseline latent occupancy.

It is no surprise that the CAR-SODM performs best when the samples are highly autocorrelated ($\alpha_c \geq 0.5$, Figures 3 and 4). Despite being high, these autocorrelation levels are frequent in real-world time data sets (e.g., Table 2); thus, the application of CAR-SODM can be useful for many studies dealing with species occupancy data. CAR-SODM provided a particularly high improvement of performance when TP probability was low and only a few replicates were available for detection. Actually, these are the conditions under which inference is more challenging, and the use of SODM is essential to obtain unbiased biodiversity data. The performance of all the approaches clearly increased when more replicates are available, as shown by multiple studies on occupancy modelling (Ficetola et al., 2015; Lahoz-Monfort et al., 2016; Schmidt et al., 2013). For instance, with 12 replicates, the maximum TSS of CAR-SODM can pass over 0.9 even if the TP rate was relatively low (p_{11} = 0.25) and the FP probability was as high as 0.015 (Figure 3). Given that scenarios of very low p_{11}, high p_{10} and autocorrelated sample are common in ancient DNA studies, the application of the CAR-SODM to such cases can be therefore promising. On the other hand, autoregressive models can be also helpful with moderate p_{11} (e.g., p_{11} = 0.4), particularly when only a few replicates are available (K = 4). The simulation study presented here also lends insights to the design of eDNA sampling and PCR schemes. For instance, when dealing with highly correlated chronosequence data (α_c = 0.95) with low TP probability (p_{11} = 0.25), analysing eight PCR replicates with the autoregressive model (K = 8) provided nearly the same discriminating power provided by the non-autoregressive model with 12 PCR replicates (K = 12; Figure 3), thus allowing to improve the performance of studies. On the other hand, it should be remarked that the same improvement could not be reached by the autoregressive model on data with just four replicates, compared with the non-autoregressive model on data with eight replicates (Figure 3).

The CAR-SODM is an extremely flexible approach that can be applicable not only to time series data, but also to spatially explicit data. Actually, the performance of the model when dealing with spatially correlated data was coherent with the results of chronosequence analyses (compare Figures 1, 2 and 4a,b; see also Figures 3 and 4c). The only structural difference between these two applications is in the adjacency matrix **W**, which can be easily modified on the basis of researcher's expectations on how error is autocorrelated through space and time (Legendre & Gauthier, 2014). For instance, in our spatial example, we used a rook scheme of adjacency, but different connection schemes can be used, for instance on the basis of known dispersal distance of study species.

Given that latent occupancy was not strictly constant, in CAR-SODM, the probability of occupancy given a number of positive replicates can be variable, depending on whether the positive is nearby to other positives or not. For instance, one single detection of cattle eDNA is more likely to be considered a true presence if it occurs during periods in which cattle are frequently detected (Figure 5c). Considering the temporal coherence of observations has been frequently considered a good criterion to assess the validity of biodiversity data obtained through eDNA, particularly in ancient DNA studies (Giguet-Covex et al., 2014; Parducci et al., 2017). For instance, Sjögren et al. (2017) considered positive PCR replicates of sedimentary plant DNA in two stratigraphically adjacent samples to be reliable, even if either or both of the two samples were amplified in just one PCR replicate. The outcome of CAR-SODMs is somehow analogous, given that they inherently take into account the temporal or spatial coherence of data, still CAR-SODMs provide a more objective approach to assess the status of samples for which target species have been detected only one or very few times. However, it should be remarked that the conditionally autoregressive model considered here assumed that autocorrelation is homogeneous through space and time, that is, that the same autocorrelation values hold through the entire data set. Such an assumption is common in spatially and temporally explicit analyses, but autocorrelation can be

TABLE 2 Parameters estimated by occupancy models applied to eDNA of ancient livestock (Giguet-Covex et al., 2014), ancient plants (Pansu et al., 2015), crayfish (Dougherty et al., 2016), salamanders (Vörös et al., 2017) and lampreys (Ostberg et al., 2018). *S*, sample size; *K*, number of PCRs; ψ_b, estimated baseline occupancy in the CAR-SODM; ψ_0, estimated occupancy probability in the Royle and Link (2006) model; p_{11}, estimated TP probability; p_{10}, estimated FP probability; α, estimated autoregressive parameter. CAR: CAR-SODM; RL: Royle and Link (2006) model. 95% highest probability density intervals (HPDIs) are presented in parentheses under each estimate

Paper	Structure	*S*	*K*	Taxon	αCAR	ψ_bCAR	ψ_0 RL	p_{11} CAR	p_{11} RL	p_{10} CAR	p_{10} RL
Giguet-Covex et al. (2014)	Temporal	44	8	*Bos* (cattle)	0.864 (0.585, 1.000)	0.395 (0.001, 0.887)	0.333 (0.180, 0.519)	0.302 (0.212, 0.398)	0.313 (0.208, 0.428)	0.008 (0.000, 0.024)	0.012 (0.000, 0.033)
	Temporal	44	8	*Ovis* (sheep)	0.780 (0.034, 1.000)	0.396 (0.000, 0.905)	0.167 (0.031, 0.328)	0.261 (0.098, 0.434)	0.338 (0.118, 0.510)	0.012 (0.000, 0.028)	0.015 (0.000, 0.035)
Pansu et al. (2015)	Temporal	44	8	*Achilleamacrophylla*	0.986 (0.951, 1.000)	0.401 (0.000,0.903)	0.155 (0.042, 0.272)	0.410 (0.281, 0.54)	0.491 (0.315, 0.681)	0.010 (0.001, 0.023)	0.016 (0.001, 0.035)
	Temporal	44	8	*Alchemilla* MOTU	0.948 (0.817, 1.000)	0.415 (0.000, 0.913)	0.285 (0.153, 0.424)	0.820 (0.700, 0.93)	0.871 (0.784, 0.958)	0.039 (0.002, 0.072)	0.053 (0.021, 0.088)
	Temporal	44	8	*Hypericum* MOTU	0.928 (0.736, 1.000)	0.310 (0.000, 0.857)	0.225 (0.093, 0.380)	0.513 (0.376, 0.656)	0.518 (0.352, 0.671)	0.060 (0.031, 0.092)	0.061 (0.030, 0.095)
	Temporal	44	8	*Pinus* MOTU	0.787 (0.230, 1.000)	0.307 (0.000, 0.829)	0.218 (0.062, 0.398)	0.38 (0.232, 0.529)	0.438 (0.264, 0.633)	0.042 (0.011, 0.077)	0.051 (0.017, 0.087)
Dougherty et al. (2016)	Spatial	262	4	*Orconectesrusticus* (rusty crayfish)	0.932 (0.845, 0.996)	0.338 (0.000, 0.791)	0.331 (0.222, 0.433)	0.739 (0.647, 0.830)	0.829 (0.730, 0.924)	0.043 (0.010, 0.079)	0.073 (0.042, 0.100)
Vörös et al. (2017)	Spatial	15	20	*Proteus anguinus* (olm salamander)	0.440 (-0.989, 0.978)	0.118 (0.003, 0.264)	0.319 (0.000, 0.876)	0.681 (0.491, 0.866)	0.682 (0.492, 0.86)	0.039 (0.018, 0.063)	0.039 (0.020, 0.067)
Ostberg et al. (2018)	Spatial	18	12	*Entosphenustridentatus* (pacific lamprey)	0.212 (-0.739, 0.999)	0.272 (0.000, 0.802)	0.209 (0.051, 0.387)	0.731 (0.529, 0.885)	0.740 (0.564, 0.891)	0.028 (0.000, 0.054)	0.030 (0.003, 0.058)
	Spatial	18	12	*Lampetraspp*	0.305 (-0.529, 0.998)	0.691 (0.204, 1.000)	0.7 (0.510, 0.896)	0.842 (0.783, 0.898)	0.842 (0.783, 0.897)	0.045 (0.008, 0.088)	0.045 (0.007, 0.089)

non-homogeneous in real-world data (non-stationarity; Beale et al., 2010). Non-stationary of autocorrelation can greatly increase the bias of model outcomes, but unfortunately, this issue remains challenging to address (Beale et al., 2010) and requires attention in future methodological developments.

The focus of the present work was on occupancy and detection probability estimation; therefore, we compared the two-level Royle and Link (2006) model and its conditionally autoregressive counterpart, without considering any covariates to the model parameters. Nevertheless, the CAR-SODM can be easily modified, in order to meet specific needs. First, covariates can be incorporated to identify the drivers of temporal or spatial variations of TP/FP rates and to take them into account. For example, DNA degradation may decrease detection probability of older samples (Olajos et al., 2017). Second, FPs and FNs can be generated by processes acting at different stages. For example, an FN may be caused by contamination during sampling, as well as at the PCR amplification stage. Therefore, multiple levels of latent states and corresponding parameters could be modelled instead of the singe level of the present work, especially when additional information is available (e.g., unambiguous detections or laboratory calibration data; Guillera-Arroita et al., 2017).

5 | CONCLUSIONS

Current occupancy modelling analysing eDNA from spatially or temporally structured data ignore autocorrelation, despite many studies stressing the importance of autocorrelation in ecological data analysis. Using a conditionally autoregressive SODM, we showed when and how occupancy modelling in eDNA studies can benefit from considering the autocorrelation among samples. In comparison with the non-autoregressive Royle and Link (2006) model, the conditionally autoregressive model can better estimate important parameters such as TP/FP rates and more accurately predict the actual occupancy of taxon, via discriminating detections according to their neighbours' states. The improvement was particularly high when the autocorrelation among samples was strong. We thus recommend using the autoregressive model in the frequent situations in which researchers expect autocorrelation among samples according to temporal or spatial structure, and when the TP rate is low.

ACKNOWLEDGEMENTS

We thank A. Kinziger and three anonymous reviewers for constructive comments on a previous version of this manuscript. Most of the computations presented in this paper were performed using the Froggy platform of the CIMENT infrastructure (https://ciment.ujf-grenoble.fr), which is supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA), the OSUG@2020 labex (reference ANR10 LABX56) and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the programme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche. GFF has received funding from the European Research Council under the European Community Horizon 2020 Programme, Grant Agreement no. 772284

(IceCommunities). The Laboratoire d'Écologie Alpine is part of Labex OSUG@2020 (ANR10 LABX56).

DATA ACCESSIBILITY

R code for simulations, stan source code and commented examples of analyses of real-world data sets: https://gitlab.com/wtchen/DNA_CAR_model.git.

AUTHOR CONTRIBUTIONS

W.C. and G.F.F. jointly designed this study. W.C. performed literature analysis, simulations and data analysis. W.C. wrote the first version of the manuscript, with subsequent contribution by G.F.F.

ORCID

Wentao Chen http://orcid.org/0000-0002-2665-581X
Gentile Francesco Ficetola http://orcid.org/0000-0003-3414-5155

REFERENCES

Aing, C., Halls, S., Oken, K., Dobrow, R., & Fieberg, J. (2011). A Bayesian hierarchical occupancy model for track surveys conducted in a series of linear, spatially correlated, sites. *Journal of Applied Ecology*, *48*(6), 1508–1517. https://doi.org/10.1111/j.1365-2664.2011.02037.x

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*(6), 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., & Elston, D. A. (2010). Regression analysis of spatial data. *Ecology Letters*, *13*(2), 246–264. https://doi.org/10.1111/j.1461-0248.2009.01422.x

Bini, L. M., Diniz-Filho, J. A. F., Rangel, T. F. L. V. B., Akre, T. S. B., Albaladejo, R. G., Albuquerque, F. S., … Hawkins, B. A. (2009). Coefficient shifts in geographical ecology: An empirical evaluation of spatial and non-spatial regression. *Ecography*, *32*(2), 193–204. https://doi.org/10.1111/j.1600-0587.2009.05717.x

Bivand, R., Hauke, J., & Kossowski, T. (2013). Computing the Jacobian in Guassian spatial autoregressive modles: An illustrated comparison of available methods. *Geographical Analysis*, *45*(2), 150–179. https://doi.org/10.1111/gean.12008

Bivand, R., & Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, *63*(18), 1–36. https://doi.org/10.18637/jss.v063.i18

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Brown, C. J., Schoeman, D. S., Sydeman, W. J., Brander, K., Buckley, L. B., Burrows, M., … Richardson, A. J. (2011). Quantitative approaches in climate change ecology. *Global Change Biology*, *17*(12), 3697–3713. https://doi.org/10.1111/j.1365-2486.2011.02531.x

Capo, E., Debroas, D., Arnaud, F., & Domaizon, I. (2015). Is planktonic diversity well recorded in sedimentary DNA? Toward the reconstruction of past protistan diversity. *Microbial Ecology*, *70*(4), 865–875. https://doi.org/10.1007/s00248-015-0627-2

Chambert, T., Miller, D. A., & Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, *96*(2), 332–339. https://doi.org/10.1890/14-1507.1

Dorazio, R. M., & Erickson, R. A. (2018). eDNAoccupancy: An r package for multiscale occupancy modelling of environmental DNA data. *Molecular Ecology Resources*, *18*(2), 368–380. https://doi.org/10.1111/1755-0998.12735

Dormann, C. F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, *16*(2), 129–138. https://doi.org/10.1111/j.1466-8238.2006.00279.x

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., … Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, *30*(5), 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Dougherty, M. M., Larson, E. R., Renshaw, M. A., Gantz, C. A., Egan, S. P., Erickson, D. M., & Lodge, D. M. (2016). Environmental DNA (eDNA) detects the invasive rusty crayfish *Orconectes rusticus* at low abundances. *Journal of Applied Ecology*, *53*(3), 722–732. https://doi.org/10.1111/1365-2664.12621

Epp, L. S., Gussarova, G., Boessenkool, S., Olsen, J., Haile, J., Schrøder-Nielsen, A., … Brochmann, C. (2015). Lake sediment multi-taxon DNA from North Greenland records early post-glacial appearance of vascular plants and accurately tracks environmental changes. *Quaternary Science Reviews*, *117*(0318), 152–163. https://doi.org/10.1016/j.quascirev.2015.03.027

Ficetola, G. F., Manenti, R., De Bernardi, F., & Padoa-Schioppa, E. (2012). Can patterns of spatial autocorrelation reveal population processes? An analysis with the fire salamander. *Ecography*, *35*(8), 693–703. https://doi.org/10.1111/j.1600-0587.2011.06483.x

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., … Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *15*(3), 543–556. https://doi.org/10.1111/1755-0998.12338

Ficetola, G. F., Poulenard, J., Sabatier, D., Messager, E., Gielly, L., Leloup, A., … Arnaud, F. (2018). DNA from lake sediments reveals long-term ecosystem changes after a biological invasion. *Science Advances*. *4*(5), Eaar4292. https://doi.org/10.1126/sciadv.aar4292

Ficetola, G. F., Romano, A., Salvidio, S., & Sindaco, R. (2018). Optimizing monitoring schemes to detect trends in abundance over broad scales. *Animal Conservation*, *21*(3), 221–231. https://doi.org/10.1111/acv.12356

Ficetola, G. F., Taberlet, P., & Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Molecular Ecology Resources*, *16*(3), 604–607. https://doi.org/10.1111/1755-0998.12508

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). Bayesian data analysis. *Chapman Texts in Statistical Science Series*. ISBN 1-58488-388-X

Gelman, A., Rubin, D. B., Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177013604

Giguet-Covex, C., Pansu, J., Arnaud, F., Rey, P.-J., Griggo, C., Gielly, L., … Taberlet, P. (2014). Long livestock farming history and human landscape shaping revealed by lake sediment DNA. *Nature Communications*, *5*, 3211. https://doi.org/10.1038/ncomms4211

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: Advances, challenges and opportunities. *Ecography*, *40*(2), 281–295. https://doi.org/10.1111/ecog.02445

Guillera-Arroita, G., Lahoz-Monfort, J. J., van Rooyen, A. R., Weeks, A. R., & Tingley, R. (2017). Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution*, *8*(9), 1081–1091. https://doi.org/10.1111/2041-210X.12743

Hines, J. E., Nichols, J. D., Royle, J. A., MacKenzie, D. I., Gopalaswamy, A. M., Samba Kumar, N., & Karanth, K. U. (2010). Tigers on trails:

Occupancy modelling for cluster sampling. *Ecological Applications*, *20*(5), 1456–1466. https://doi.org/10.1890/09-0321.1

Jin, X., Carlin, B. P., & Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, *61*(4), 950–961. https://doi.org/10.1111/j.1541-0420.2005.00359.x

Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C., & Pond, B. A. (2013). Spatial occupancy models for large data sets. *Ecology*, *94*(4), 801–808. https://doi.org/10.1890/12-0564.1

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Molecular Ecology Resources*, *16*(3), 673–685. https://doi.org/10.1111/1755-0998.12486

Legendre, P., & Gauthier, O. (2014). Statistical methods for temporal and space – time analysis of community composition data. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1778), 20132728. https://doi.org/10.1098/rspb.2013.2728

Legendre, P., & Legendre, L. F. J. (2012). *Numerical ecology*, 3rd ed. Amsterdam, The Netherlands: Elsevier.

Lennon, J. J. (2000). Red-shifts and red herrings in geographical ecology. *Ecography*, *23*(1), 101–113. https://doi.org/10.1111/j.1600-0587.2000.tb00265.x

Lobo, J. M., Jiménez-valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Lopes, C. M., Sasso, T., Valentini, A., Dejean, T., Martins, M., Zamudio, K. R., & Haddad, C. F. B. (2017). eDNA metabarcoding: A promising method for anuran surveys in highly diverse tropical forests. *Molecular Ecology Resources*, *17*(5), 904–914. https://doi.org/10.1111/1755-0998.12643

Mackenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew, J., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*(8), 2248–2255. https://doi.org/10.1890/0012-9658(2002) 083[2248:ESORWD]2.0.CO;2

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, *92*(7), 1422–1428. https://doi.org/10.1890/10-1396.1

Olajos, F., Bokma, F., Bartels, P., Myrstener, E., Rydberg, J., Öhlund, G., … Englund, G. (2017). Estimating species colonization dates using DNA in lake sediment. *Methods in Ecology and Evolution*, *9*(3), 535–543. https://doi.org/10.1111/2041-210X.12890

Ostberg, C. O., Chase, D. M., Hayes, M. C., & Duda, J. J. (2018). Distribution and seasonal differences in Pacific Lamprey and *Lampetra* spp eDNA across 18 Puget Sound watersheds. *PeerJ*, *6*, e4496. https://doi.org/10.7717/peerj.4496

Pansu, J., Giguet-Covex, C., Ficetola, G. F., Gielly, L., Boyer, F., Zinger, L., … Choler, P. (2015). Reconstructing long-term human impacts on plant communities: An ecological approach based on lake sediment DNA. *Molecular Ecology*, *24*(7), 1485–1498. https://doi.org/10.1111/mec.13136

Parducci, L., Bennett, K. D., Ficetola, G. F., Alsos, I. G., Suyama, Y., Wood, J. R., … Pedersen, M. W. (2017). Ancient plant DNA in lake sediments. *New Phytologist*, *214*(3), 924–942. https://doi.org/10.1111/NPH.14470

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77.

Royle, J. A., & Dorazio, R. M. (2008). *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Elsevier.

Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for false positives and false negative errors. *Ecology*, *87*(4), 835–841. https://doi.org/10.1890/0012-9658(2006) 87[835:GSOMAF]2.0.CO;2

Sargeant, G., Sovada, M., Slivinski, C., & Johnson, D. (2011). Markov Chain Monte Carlo estimation of species distributions: A case study of the swift fox in western Kansas. *Journal of Wildlife Management*, *69*(2), 483–497. https://doi.org/10.2193/0022-541X(2005)069[0483:MCMCEO]2.0.CO;2

Schmidt, B. R., Kéry, M., Ursenbacher, S., Hyman, O. J., & Collins, J. P. (2013). Site occupancy models in the analysis of environmental DNA presence/absence surveys: A case study of an emerging amphibian pathogen. *Methods in Ecology and Evolution*, *4*(7), 646–653. https://doi.org/10.1111/2041-210X.12052

Serrao, N. R., Reid, S. M., & Wilson, C. C. (2017). Establishing detection thresholds for environmental DNA using receiver operator characteristic (ROC) curves. *Conservation Genetics Resources*, *10*, 555–562. https://doi.org/10.1007/s12686-017-0817-y

Sjögren, P., Edwards, M. E., Gielly, L., Langdon, C. T., Croudace, I. W., Merkel, M. K. F., … Alsos, I. G. (2017). Lake sedimentary DNA accurately records 20th Century introductions of exotic conifers in Scotland. *New Phytologist*, *213*(2), 929–941. https://doi.org/10.1111/nph.14199

Stan Development Team (2016). {RStan}: the {R} interface to {Stan}. Retrieved from http://mc-stan.org/

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Vörös, J., Márton, O., Schmidt, B. R., Gál, J. T., & Jelić, D. (2017). Surveying Europe's only cave-dwelling chordate species (*Proteus anguinus*) using environmental DNA. *PLoS One*, *12*(1), 12–14. https://doi.org/10.1371/journal.pone.0170945

Wagner, H. H., & Fortin, M. J. (2005). Spatial analysis of landscapes: Concepts and statistics. *Ecology*, *86*(8), 1975–1987. https://doi.org/10.1890/04-0914

Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, *121*(2), 311–324. https://doi.org/10.1016/S0378-3758(03)00111-3

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Chen W, Ficetola GF. Conditionally autoregressive models improve occupancy analyses of autocorrelated data: An example with environmental DNA. *Mol Ecol Resour*. 2019;19:163–175. https://doi.org/10.1111/1755-0998.12949