Received: 24 July 2020 | Accepted: 29 December 2020

DOI: 10.1111/2041-210X.13594

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Imputation of incomplete large-scale monitoring count data via penalized estimation

Mohamed Dakki[1] | Geneviève Robin[2] | Marie Suet[3] | Abdeljebbar Qninba[1] | Mohammed A. El Agbani[1] | Asmâa Ouassou[1] | Rhimou El Hamoumi[4] | Hichem Azafzaf[5] | Sami Rebah[5] | Claudia Feltrup-Azafzaf[5] | Naoufel Hamouda[5] | Wed A. L. Ibrahim[6] | Hosni H. Asran[6] | Amr A. Elhady[6] | Haitham Ibrahim[6] | Khaled Etayeb[7,9] | Essam Bouras[8,9] | Almokhtar Saied[8,9] | Ashrof Glidan[8,9] | Bakar M. Habib[10] | Mohamed S. Sayoud[11] | Nadjiba Bendjedda[12] | Laura Dami[3] | Clemence Deschamps[3] | Elie Gaget[3,13] | Jean-Yves Mondain-Monval[14] | Pierre Defos du Rau[14]

[1]Institut Scientifique, Université Mohammed V de Rabat, Rabat, Morocco; [2]LaMME, CNRS, Université d'Évry Val d'Essonne, Évry, France; [3]Centre de Recherche de la Tour du Valat, Arles, France; [4]Faculté des Sciences Ben M'sik, Univ. Hassan II, Casablanca, Morocco; [5]Association "Les Amis des Oiseaux" (AAO/BirdLife en Tunisie), Ariana, Tunisia; [6]Egyptian Environmental Affairs Agency, Cairo, Egypt; [7]Zoology Department, Tripoli University, Tripoli, Libya; [8]Environment General Authority, Tripoli, Libya; [9]Libyan Society for Birds, Tripoli, Libya; [10]Conservation des Forêts de la Wilaya d'Oran, Oran, Algeria; [11]Centre Cynégétique de Réghaia, Direction Générale des Forets, Alger, Algeria; [12]Direction Générale des Forêts, Alger, Algeria; [13]International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria and [14]Office Français de la Biodiversité, Unité Avifaune Migratrice, Arles, France

Correspondence
Pierre Defos du Rau
Email: pierre.defosdurau@ofb.gouv.fr

Funding information
This work, including data collection and analysis, was supported by the French Ministry in charge of Environment (Ministère de la Transition Ecologique et Solidaire) through the SPOVAN and Technical Support Unit projects, the TOTAL Foundation, the Critical Ecosystem Partnership Fund, the Albert II de Monaco Foundation, the MAVA Foundation for Nature, as well as the Agence Française de Développement, the Fonds Français pour l'Environnement Mondial and the European Union, respectively, in the framework of the Réseau Oiseaux d'Eau Méditerranée project coordinated by Tour du Valat and the RESSOURCE project coordinated by the Food and Agriculture Organisation.

Handling Editor: Dave Hodgson

Abstract

1. In biodiversity monitoring, large datasets are becoming more and more widely available and are increasingly used globally to estimate species trends and conservation status. These large-scale datasets challenge existing statistical analysis methods, many of which are not adapted to their size, incompleteness and heterogeneity. The development of scalable methods to impute missing data in incomplete large-scale monitoring datasets is crucial to balance sampling in time or space and thus better inform conservation policies.
2. We developed a new method based on penalized Poisson models to impute and analyse incomplete monitoring data in a large-scale framework. The method allows parameterization of (a) space and time factors, (b) the main effects of predictor covariates, as well as (c) space–time interactions. It also benefits from robust statistical and computational capability in large-scale settings.
3. The method was tested extensively on both simulated and real-life waterbird data, with the findings revealing that it outperforms six existing methods in terms of missing data imputation errors. Applying the method to 16 waterbird species, we estimated their long-term trends for the first time at the entire North African scale, a region where monitoring data suffer from many gaps in space and time series.
4. This new approach opens promising perspectives to increase the accuracy of species-abundance trend estimations. We made it freely available in the R package

'LORI' (https://CRAN.R-project.org/package=lori) and recommend its use for large-scale count data, particularly in citizen science monitoring programmes.

KEYWORDS

biodiversity monitoring, high-dimensional statistics, incomplete count data, missing data imputation, penalized estimation, waterbird trends in North Africa

1 | INTRODUCTION

Biodiversity monitoring datasets are becoming more complex and high-dimensional, as the biodiversity crisis urges the collection and analysis of data, particularly at large scales of space and time (Han et al., 2014; Hughes et al., 2017; Kindsvater et al., 2018; White, 2019). The resulting datasets, emerging in particular from citizen science monitoring programmes, contribute to answering many important ecological and conservation questions (Pereira et al., 2013; Stephenson, Brooks, et al., 2017). However, their high-dimensional complexity challenges existing statistical data analysis procedures. Indeed, statistical guarantees for commonly used, state-of-the-art methods for large biodiversity datasets usually assume an asymptotic regime, where the number of observations is large compared to the number of parameters. Yet, one acute issue in biodiversity monitoring schemes is the occurrence of a substantial amount of *missing data* (Harel & Zhou, 2006; Nakagawa & Freckleton, 2008; Wauchope et al., 2019), up to the point where the asymptotic assumption becomes obsolete. This is especially the case in areas where data collection is costly or logistically difficult to undertake, but where biodiversity is no less in need of monitoring (Stephenson, Bowles-Newark, et al., 2017; Tibshirani, 1996). Hence, the development of scalable methods to impute missing data in incomplete large-scale monitoring datasets is crucial to unbiased inference.

In practice, missing data in biodiversity monitoring has often been tackled by case removal or missing value imputation (Ellington et al., 2015; Onkelinx, Devos, & Quataert, 2017; Penone et al., 2014). In particular, using model-based imputation methods dedicated to spatio-temporal count data (e.g. Blanchong et al., 2006). The TRIM (TRends and Indices for Monitoring data) methodology is an important example of such methods, and is frequently used for modelling incomplete wildlife count datasets (Lehikoinen et al., 2013; Van Strien et al., 2004; Van Swaay et al., 2008). Other commonly used methods rely on chained equations (Van Buuren & Groothuis-Oudshoorn, 2011) or Random Forests (Stekhoven & Bühlmann, 2012). More recently, the use of multiple imputation procedures has been discussed (Bogaart et al., 2017; Onkelinx et al., 2017; Onkelinx et al., 2017) for trend modelling of wildlife counts.

Most of these imputation methods are backed up by theoretical results guaranteeing their consistency in asymptotic settings where the sample size is much larger than the number of parameters. However, these do not scale up to high-dimensional, finite sample settings which appear whenever the proportion of missing values is large: This is known as the *curse of dimensionality* (Donoho, 2000).

In this study, we develop a new tool for count data imputation, which is effective in such high-dimensional settings, that is, when the count table, the proportion of missing values and the set of predictor covariates are large. This method is based on *penalized estimation*, using the Lasso penalty (Tibshirani, 1996). We argue that this new tool, implemented in the R package 'LORI' (Low-Rank Interactions), is a competitive option for imputing count datasets, in particular when there is a large proportion of missing data, and when predictor covariates are available. It benefits from statistical guarantees with optimal estimation error in the described high-dimensional settings (Robin et al., 2019). Such situations with a large amount of missing data and large predictor sets are frequent, as species count data are often difficult to collect, but covariates related to sampling sites and time points (e.g. meteorological data) can generally be recovered easily: for example, via web scraping (Amano et al., 2018; Murray et al., 2010; Stephenson et al., 2015).

North Africa (comprising Morocco, Algeria, Tunisia, Libya and Egypt) is of strategic importance for the conservation of waterbirds migrating along the African-Eurasian flyway (Sayoud et al., 2017) when they need to find stopover or wintering habitats between the Mediterranean Sea and the Sahara Desert.

Assessing species population sizes and trends at the North African scale is thus essential (Galewski et al., 2011; Samraoui et al., 2011). However, for several reasons (mainly lack of financial or human resources, or political context), coverage of North African wetlands for the IWC has been highly irregular in both time and space (Dakki et al., 2001; EGA - RAC/SPA Waterbird Census Team, 2012; El Agbani et al., 1996; Sayoud et al., 2017). Thus, a large proportion of counts (up to 60%, depending on the species) are missing.

This study had three objectives. First, we developed a general model for count data imputation using penalized estimation; this method is able to integrate high-dimensional predictor sets. Second, we evaluated the performance of the method compared to six existing imputation methods on simulated and real-life waterbird monitoring data. The LORI method outperformed competitors, in particular when the proportion of missing data was large. Third, we applied the LORI method to recover actual missing data and infer species-specific trends for 16 waterbird species over 785 North African wetlands between 1990 and 2017. The trends identified for North Africa were compared to those proposed at the flyway scale by Wetlands International (2019).

2 | MATERIALS AND METHODS

2.1 | Low-Rank Interaction (LORI) model for incomplete count data

In biodiversity surveys, count datasets are typically organized as a large space × time matrix of site- and period-specific counts. These contingency tables are often analysed using Poisson GLMs with row (site) and column (time) effects (e.g. Van Strien et al., 2004). Consider a count table Y where rows correspond to ecological sites, and columns to different time points, with Y_{ij} the number of individuals observed at site i at time j. A simple example of a Poisson log-linear model is:

$$\log\left[E\left(Y_{ij}\right)\right] = \alpha_i + \beta_j. \quad (1)$$

In Equation (1), α_i corresponds to the effect of site i, and β_j corresponds to the effect of time j. If additional covariates are available, such as meteorological and geographical information, model (1) may be generalized in order to incorporate these as well. For any site i and any year j, X_{ij} is denoted as a vector of p covariates, and $X_{ij}(k)$ as its k-th coefficient, corresponding to the value of site i and year j at the k-th covariate (e.g. the level of precipitation at location i and time j). Model (1) may be extended to:

$$\log\left[E\left(Y_{ij}\right)\right] = \alpha_i + \beta_j + \sum_{k=1}^{p} \gamma_k X_{ij}(k). \quad (2)$$

In Equation (2), the additional coefficients γ_k correspond to the effect of the covariates. In a missing data imputation perspective, incorporating additional covariates is an opportunity to improve the prediction of missing entries, as these could be good predictors of species counts (Amano et al., 2018).

In addition, row–column interaction terms may also be modelled. For any row i and column j, the interaction term is denoted as θ_{ij}. Model (2) becomes:

$$\log\left[E\left(Y_{ij}\right)\right] = \alpha_i + \beta_j + \sum_{k=1}^{p} \gamma_k X_{ij}(k) + \theta_{ij}. \quad (3)$$

Model (3) is over-parameterized; thus, we based our approach on two main assumptions. First, we assumed that not all sites, years and covariates have a non-zero effect on the counts. Thus, the vectors of row, column and covariate effects (α, β, and γ) may contain several zeros. Second, we assumed the existence of a few groups of similar sites and similar years, which can be embedded by constraining the matrix of interactions θ to be of low rank. Indeed, if θ is of rank r, then for any site i and year j, the corresponding interaction θ_{ij} can be decomposed as the sum of multiplicative interactions between r latent factors:

$$\theta_{ij} = \sigma_1 u_{i1} v_{j1} + \sigma_2 u_{i2} v_{j2} + \ldots + \sigma_r u_{ir} v_{jr}. \quad (4)$$

In (4), r is the number of latent factors, σ_l is the strength of the interaction between the l-th site and year latent factors and u_{il}, v_{jl} are the values of the l-th factor for site i and year j.

To estimate the parameters of the model, we used penalized estimation approaches. These methods consist of minimizing the sum of two terms: the first term is the standard negative log-likelihood, and the second is a penalty term designed to increase with the model's complexity. In our case, the model's complexity was specified by the parsimony of the vectors α, β and γ, and by the number of latent factors driving the interactions. The standard Poisson negative log-likelihood is given by:

$$\sum_{(i,j)\in\Omega} \left\{ -Y_{ij}\left(\alpha_i + \beta_j + \sum_{k=1}^{p} \gamma_k X_{ij}(k) + \theta_{ij}\right) + \exp\left(\alpha_i + \beta_j + \sum_{k=1}^{p} \gamma_k X_{ij}(k) + \theta_{ij}\right)\right\}. \quad (5)$$

We defined our penalty term as:

$$\lambda_1 \|\theta\|_* + \lambda_2 \left(\|\alpha\|_1 + \|\beta\|_1 + \|\gamma\|_1\right). \quad (6)$$

In (6), for any vector x, $||x||_1$ is the l_1 norm of x (the sum of entries in absolute value). For any matrix M, $||M||_*$ denotes the nuclear norm (the sum of singular values, also known as trace norm). Finally, the parameters λ_1 and λ_2 control the trade-off between fitting the data and imposing low-complexity models: The larger λ_1 and λ_2, the more coefficients are set to zero. In practice, the choice of these parameters is made using cross-validation. Note that this penalty term is the combination of two well-known and extensively used penalties in high-dimensional statistics. The l_1 norm penalty comes from the Lasso technique, developed by Tibshirani (1996). The nuclear norm penalty comes from *matrix completion* (Candès & Recht, 2009; Candès & Tao, 2010). Both techniques have the advantage of benefiting from sound, non-asymptotic theoretical guarantees.

We fit the parameters of the imputation model by solving the following minimization problem:

$$\left(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\theta}s.\right) \in \arg\min L(\alpha, \beta, \gamma, \theta) + \lambda_1 \|\theta\|_* + \lambda_2 \left(\|\alpha\|_1 + \|\beta\|_1 + \|\lambda\|_1\right). \quad (7)$$

This estimation problem was initially studied in Robin et al. (2019) and in Robin, Klopp, et al. (2019). In these papers, the authors provide strong theoretical guarantees of the estimation capacities of (7). In particular, the main advantage of (7) is that the estimation error of the parameters $\left(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\theta}\right)$ increases linearly with the number of *non-zero parameters in the model*, instead of the *total number of parameters* (including zeros). In high-dimensional settings where the number of parameters is large, this can allow for a drastic reduction in estimation and imputation errors compared to standard estimation procedures.

2.2 | Testing datasets

We first evaluated the imputation capacities of the LORI method on simulated count data. We simulated species counts using GLM for 100 sites, 30 years, 5 covariates and 2 latent factors. The covariates, as well as the latent factors, were generated from multivariate Gaussian distributions. We simulated site, year and covariate

effects, using standard normal distributions. Once these parameters were fixed, we simulated two different species count datasets using two different GLMs. The first model was a Poisson GLM. As for other wildlife, waterbird count data are known to be prone to overdispersion and zero-inflation (Gaget et al., 2020); we thus also simulated a dataset using a zero-inflated negative binomial model (ZINB) with 10% of zero values.

To evaluate the imputation capacities of LORI on real-life waterbird count data, we selected the Northern Shoveler (*Spatula clypeata*) as a most widespread species from the IWC North African dataset in order to artificially introduce missing data. We extracted the 209 most frequently monitored sites for this duck species. As the IWC dataset for North Africa contains a lot of missing data, we could not extract a complete subsample. This real-life waterbird dataset initially had 25% missing data. In this example, the size of the predictor set was 21, and most covariates were quantitative.

For both the simulated and real-life data, we tested two different missing data mechanisms. The first was Missing Completely At Random (MCAR): Each entry is missing with probability $0 < p < 1$; hereafter, we refer to this mechanism as *random*. The second mechanism was Missing At Random (MAR), and the probability of missing data depends on site covariates. Specifically, for each entry Y_{ij}, the probability of missing is equal to p_i, where p_i is a site-based probability that may depend on the site covariates (e.g. the country etc.). This second mechanism aimed to mimic practical cases where remote sites with difficult political, financial or logistical conditions are visited less regularly. Hereafter, we refer to the second mechanism as *structured*.

For the simulated data, the proportion of missing values was set to 20%. For the real-life waterbird data, which already had 25% missing data, we added 10% of additional missing data among the observed ones. These missing values were added to all but 20 sites out of the 209; these 20 correspond to most densely occupied sites (more than 1,000 birds while the third quantile is 130 birds), unlikely to be missed during surveys. We repeated each of the scenarios 100 times to compare the imputation models. We also performed a more thorough simulation study with increasing proportions of missing data: The entire study is presented in Appendix S1.

In all experiments, we evaluated the performance of the different methods in terms of the relative root mean square error (RMSE) of imputation. We defined the relative RMSE as follows: $Y_1, \ldots, Y_N$ denote the true values of the missing data, and $\hat{Y}_1, \ldots, \hat{Y}_N$ denote the corresponding imputed values:

$$\text{RMSE}\left(\hat{Y}\right) = \sqrt{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}.$$

The relative RMSE of the imputation $\hat{Y}$ is defined by:

$$\rho\left(\hat{Y}\right) = \frac{\text{RMSE}\left(\hat{Y}\right)}{\sqrt{\sum_{i=1}^{N} Y_i^2}}.$$

2.3 | Comparing imputation methods

Using the datasets described above, we compared the LORI method to six other existing imputation methods.

The first competitor was the imputation of missing entries by the mean value of each row (hereafter, MEAN). The second competitor was a Poisson GLM (hereafter, GLM); for which we performed model selection prior to the missing values imputation using the AIC criterion; such selection of variables led to smaller imputation errors for GLM compared to using the entire set of covariates. The third competitor was Correspondence Analysis (CA, e.g. Fithian & Josse, 2017; Greenacre, 1984). We used the implementation of the

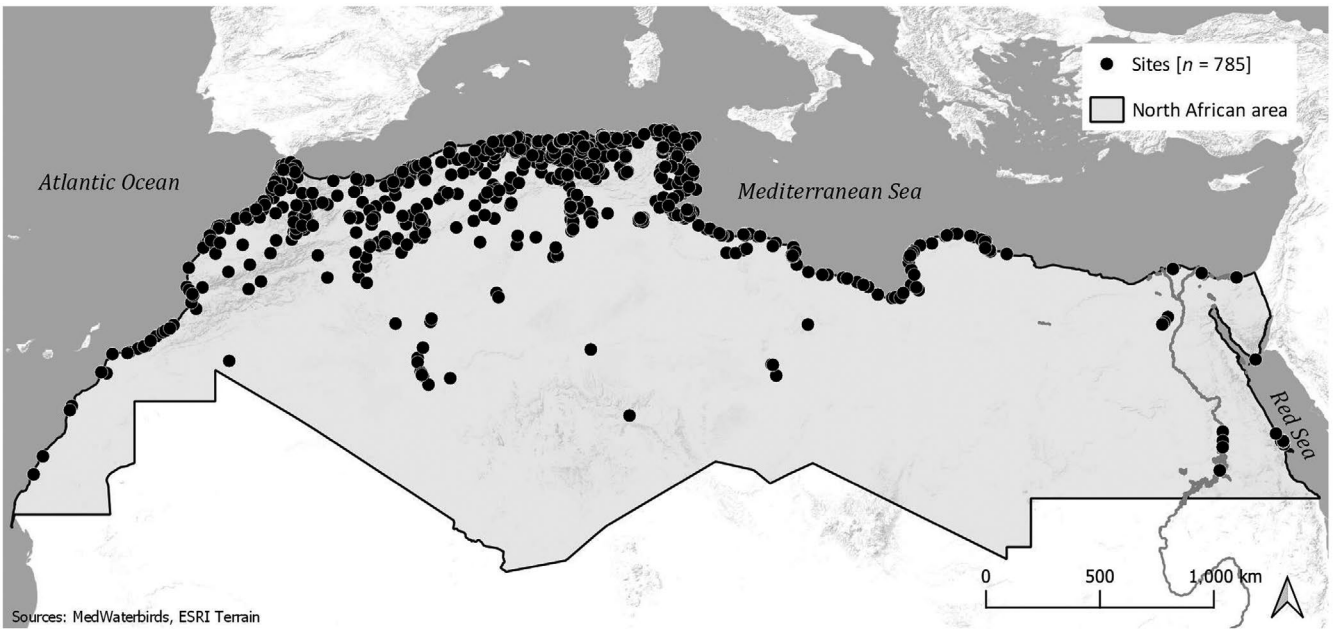

FIGURE 1 The 785 IWC monitoring sites surveyed for at least 2 years between 1990 and 2017

R package 'missMDA' (Josse & Husson, 2016). The fourth competitor was TRIM, a widely applied imputation model in wildlife monitoring schemes (Lehikoinen et al., 2013; Van Strien et al., 2004; Van Swaay et al., 2008), which is based on Poisson regression, implemented in the 'rtrim' R package. TRIM allows the use of categorical covariates and the modelling of over-dispersion in species count data. In the experiment on simulated data, we discretized our quantitative covariates so that they could be incorporated into the TRIM model, which only allows categorical covariates. Furthermore, we set the TRIM 'overdisp' parameter to TRUE whenever it led to better imputation results. In the real-life waterbird data, incorporating some or all of the 21 discretized covariates increased the frequency of failure of TRIM because of the large number of level parameters. Thus, we did not use covariates in TRIM for the real-life waterbird dataset. The fifth competitor was Multivariate Imputation by Chained Equation (MICE, Van Buuren & Groothuis-Oudshoorn, 2011); we used the implementation available in the R package 'mice'. For this method, we included all the covariates in the imputation model, and used the predicted mean matching methodology (method = 'pmm'). The sixth competitor was imputation based on Random Forests (Stekhoven & Bühlmann, 2012), implemented in the 'missForest' R package; we also incorporated all the additional covariates in the imputation model.

2.4 | North African waterbird trends

The final step of the study was the application of the LORI method to the analysis of time series of count data for 16 waterbird species over 785 North African wetlands between 1990 and 2017 (Figure 1): 163 sites (21%) in Morocco, 373 sites (47%) in Algeria, 138 sites (17%) in Tunisia, 91 sites (12%) in Libya and 20 sites (3%) in Egypt. The IWC scheme in North Africa involves teams of experienced observers (Appendix S2), trained specifically for the IWC, who follow the field protocol for waterbird monitoring recommended by Wetlands International (2010). Count results are centralized into the Medwaterbirds database (https://www.medwaterbirds.net/datacounts.php).

We ran the LORI model on 16 species, of conservation or research concern or exploited/game species in need of monitoring: the Gadwall *Mareca strepera*, Mallard *Anas platyrhynchos*, Northern Pintail *Anas acuta*, Northern Shoveler *Spatula clypeata*, Wigeon *Mareca penelope*, Common Coot *Fulica atra*, Great Cormorant *Phalacrocorax carbo*, Glossy Ibis *Plegadis falcinellus*, Dunlin *Calidris alpina*, Pied Avocet *Recurvirostra avosetta*, Greylag Goose *Anser anser*, Common Teal *Anas crecca*, Eurasian Spoonbill *Platalea leucorodia*, Ringed Plover *Charadrius hiaticula*, Common Crane *Grus grus* and Greater Flamingo *Phoenicopterus roseus*. We computed the yearly sum of imputed or observed site- and year-specific counts to obtain a yearly abundance. Based on this yearly abundance, we inferred species-specific linear temporal trends through linear regressions. Because on average TRIM slightly outperformed other imputation methods, LORI excepted (Figure 2) and is by far the most frequent approach currently implemented in waterbird trend modelling (Lehikoinen et al., 2013), we imputed these waterbird trends using both TRIM and LORI.

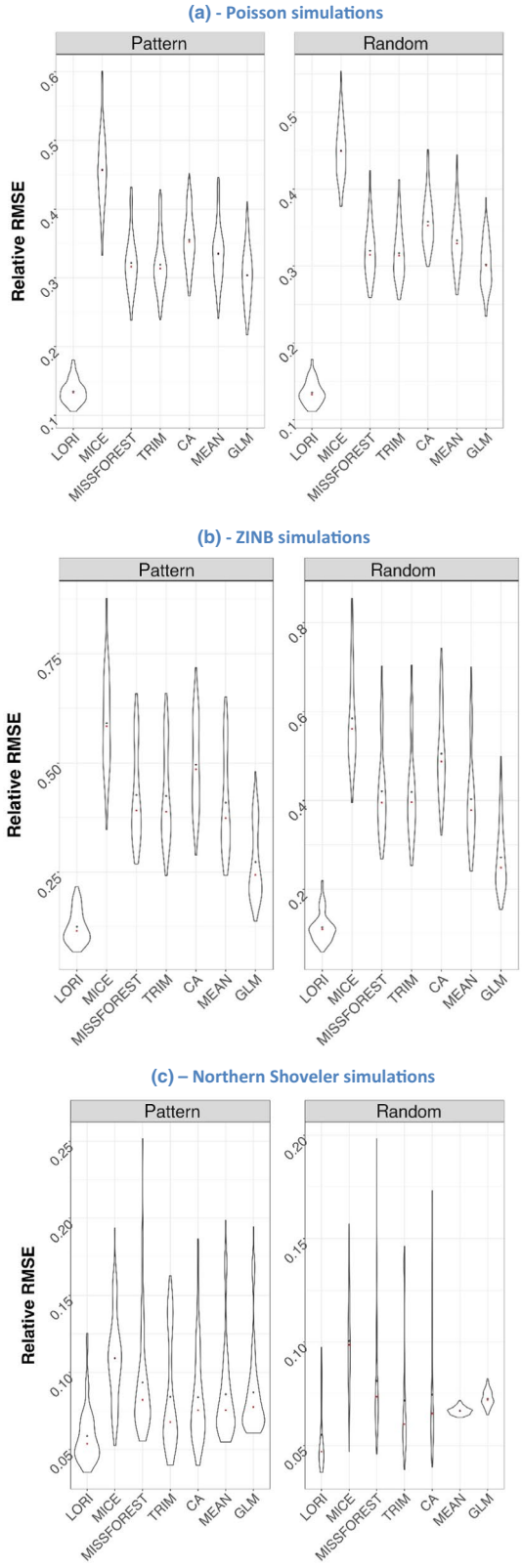

FIGURE 2 Violin plot of relative RMSE for eight imputation methods on (a) simulated Poisson data (20% missing data), (b) simulated ZINB data (20% missing data) and (c) real-life Northern Shoveler count data (10% additional missing data amounting to 30% overall missing data), for two missing data patterns (see Section 2). Mean and median are indicated by points (mean: black point, median: red point)

Temporal and spatial autocorrelation are frequent in species distribution data (Dormann et al., 2007). We assessed spatial autocorrelation by Moran's I (Moran, 1950) over the site-specific LORI residuals averaged over all observed years, using the coordinates of each site. Temporal autocorrelation was assessed by checking the semivariogram of the yearly LORI residuals averaged over all observed sites.

We modelled the time-trend and spatial distribution of waterbird counts in North Africa using 21 covariates. Our choice of each covariate was governed by a priori hypothesis; see Appendices S3 and S4 for a complete description of the covariates and ecological hypotheses.

3 | RESULTS

3.1 | Comparing LORI to existing imputation methods

Our first experiment compared the relative RMSE of imputation between seven competing methods and two different missing data patterns on the two simulated datasets generated by Poisson and ZINB models, and on the real-life waterbird data subset containing abundance data for the Northern Shoveler in North Africa.

Overall, the LORI method outperformed competitors in both accuracy and precision (Figure 2). If the dataset had 20% missing data, LORI provided a stable imputation procedure, with around 0.2 relative RMSE for the Poisson and ZINB data, with little variability. The competing methods had larger imputation errors (between 0.3 and 0.8). Overall, imputation performance had the same behaviour for both missing data mechanisms across the seven methods. However, imputation precision was generally higher for the random missing data mechanism. Similar results for 40%, 60% and 80% missing data are presented in Appendix S1. In addition, computational times differed between methods, the fastest being GLM (10 ms), then CA and TRIM (200 ms), then LORI (2 s) and finally MICE and MISSFOREST (5–10 s); see Appendix S1 for the full results. Overall, the computational time of LORI was of the same order of magnitude as MICE and MISSFOREST, and around five times larger than TRIM and CA.

The results of the experiment on ZINB simulated data (Figure 2b) show that LORI is quite robust to over-dispersion and zero-inflation. This may be explained by the incorporation of year/site interactions, which allows for larger variability between the imputed counts. In the experiment on waterbird data (Figure 2c), LORI improved on its competitors in accuracy, and yielded quite stable imputation results.

3.2 | Regional trends of North African waterbirds

Using the LORI method to impute count matrices for 16 waterbird species, we estimated their long-term trends for the first time at the North African scale over all 785 IWC sites (Appendix S5).

Out of the 16 species trends produced, two showed different trends between LORI and TRIM (Great Cormorant and Northern Shoveler, see Figure 3). For the Northern Shoveler, LORI indicated a stable/fluctuating trend ($F = 0.03$, $df = 26$, $p = 0.865$), whereas TRIM indicated an almost significant increase ($F = 3.97$, $df = 26$, $p = 0.057$). For the Great Cormorant, the TRIM-inferred trend showed a significant increase ($F = 5.10$, $df = 26$, $p = 0.036$), whereas the LORI trend remained inconclusive, hence could also qualify as stable/fluctuating ($F = 2.39$, $df = 26$, $p = 0.134$). LORI also provides parameter estimates for the ecological and anthropic drivers potentially governing the distribution of each species in space and time (Appendix S6).

All 16 species had a Moran's I below 0.05 when modelled with LORI. When modelled with TRIM, two species displayed weak but significant spatial autocorrelation (Moran's $I > 0.07$, $p < 0.05$). Overall, p-values for Moran's I were significantly lower for spatial residuals with TRIM than with LORI (Pairwise Wilcoxon Rank Sum Tests over $n = 16$ species: $Z = 2.02$, $p = 0.044$). Similarly, only one species (Northern Shoveler) showed significant temporal autocorrelation within 2 years when modelled with LORI while three, including the Northern Shoveler, showed significant temporal autocorrelation within the same two time-lags when modelled with TRIM.

4 | DISCUSSION

Large-scale count datasets are essential to biodiversity monitoring and biodiversity management (Hughes et al., 2017; White, 2019). In the remote areas where these data are most needed, it often

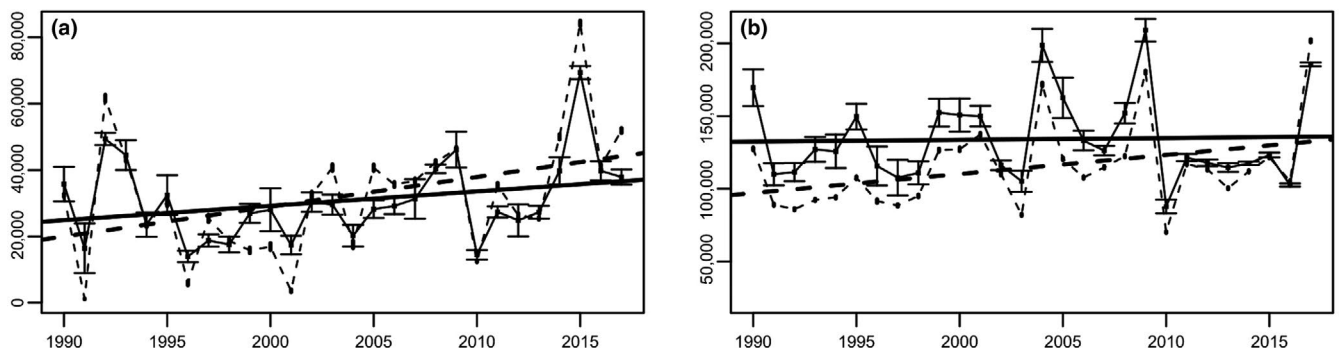

FIGURE 3 Yearly counts over all 785 North African sites for the Great Cormorant (a) and the Northern Shoveler (b) as modelled by LORI (solid lines) and TRIM (dotted lines) with the respective linear time trend

suffers from significant gaps in space or time sampling. This study experimentally demonstrates, using an empirical real-life waterbird dataset and two simulated datasets, that the LORI method is a robust solution to accurately impute missing data. LORI systematically outperformed competing methods in imputation accuracy. The imputation performance of LORI is likely due to its capacity to take into account a large number of covariate effects, as well as the most influential time × space interactions, and to the penalization of the model's coefficients, which tends to reduce their variability. In addition, as shown by the experiment on ZINB data, the method seems to show relative robustness to over-dispersion and zero-inflation.

However, even in the context of penalized maximum likelihood, estimating the effect of several covariates as well as time, space and time × space effects remains demanding in sample size. One limitation of our approach is that the LORI modelling tools preferably apply to relatively large datasets (such as our original 785 sites × 28 years count table). Thus, we recommend investigating the influence of sample size on the performance of LORI. In terms of computation time, the cross-validation approach to select the regularization parameters λ_1 and λ_2 can increase the computational time, but overall, our proposed method remains reasonably fast computer-wise, with computational times of the same order of magnitude as TRIM and MICE.

Given the observed differences in imputation accuracy, LORI can potentially indicate different trends compared to those inferred from existing methods. For instance, out of the 16 species we studied, two showed large differences in trend estimation when analysed using LORI or TRIM (Figure 3). If differences between these methods appear over such a long time span (28 years), they could potentially be even more blatant at a shorter span, for example, a 10-year span, which is the recommended timescale for short-term waterbird trend assessment (Lougheed et al., 1999; Van Roomen et al., 2011). This discrepancy in trend estimation illustrates the promising applications of this new method. As trend estimation is a major diagnostic tool in the conservation and management of wild species, we argue that LORI is a tool well adapted to supporting conservation decisions as it provides good imputation performance, and hence trend estimation that is more likely to be reliable, particularly when predictor covariates are available.

Overall, there was less autocorrelation in the LORI residuals compared to the residuals of other methods. This shows that the use of several covariates may account for most of the autocorrelation that could otherwise penalize subsequent modelling (Bardos et al., 2015; Wintle & Bardos, 2006). In our case, only the Northern Shoveler displayed some significant yet relatively marginal first-order positive temporal autocorrelation (ACF = 0.41*), suggesting either relative site fidelity to the North African wintering areas or our inability to account for the multiplicative effect of reproduction on previous winter counts despite the use of various meteorological indices for the corresponding breeding season (Appendix S3).

Interestingly, three of the studied 16 species show a significantly different trend between the North Africa scale and the (wider) corresponding flyway scale according to Wetlands International (2019). Our results found that the trend for the Mallard in particular shows a highly significant increase in North Africa, whereas Wetlands International assesses it as stable/fluctuating at the wider European/Mediterranean level. Conversely, we found a stable/fluctuating trend for the Common Crane in North Africa, but this is assessed as increasing at the wider European/Mediterranean level. Similarly, we estimated a highly significant decline for the Greylag Goose in North Africa, but this species is assessed as increasing at the wider central European/North African level. In a context of climate change, the winter distribution of migratory birds in general (Visser et al., 2009), including waterbirds (Maclean et al., 2008), is drifting north. This seems to be the main reason for the absence of an increase seen in the latter two species in North Africa, in spite of their increase at the wider European scale (Cusack et al., 2019). For the Mallard, the trend discrepancy between North Africa and the European/Mediterranean is surprising, as southerly migration from northern Europe is shortening and the wintering range is shifting north, possibly as a result of climate change (Guillemain et al., 2015). Total North African counts of the Mallard are mostly driven by counts in Morocco (Appendix S6), where the species is the most widespread breeding Anatidae (Cherkaoui et al., 2017). Breeding may indeed have been enhanced in the last decades by improved, mainly hydrological, conditions in lakes and marshes and an increase in the number of reservoirs.

5 | CONCLUSIONS

The penalized estimation approach applied to missing data imputation opens promising perspectives for analyses of large-scale count data. Taking advantage of the Lasso penalty, the LORI method has the capacity to integrate many environmental covariates, as well as time × space interactions. This brings improvement over standard approaches by incorporating more information, reducing autocorrelation, as well as estimating outliers, including for reasonably over-dispersed or zero-inflated count distributions. As covariate data will become increasingly available, allowing for analyses of large waterbird count datasets, penalized approaches such as LORI may become the recommended option to enhance analyses of incomplete wildlife counts.

ACKNOWLEDGEMENTS

We would like to thank all the field observers who participated and/or participate today in the trans-North African IWC, and who made this study possible; their names are listed in the supplementary materials (Appendix S2). We are also grateful to Elise Bradbury for editing the English.

AUTHORS' CONTRIBUTIONS

P.D.d.R., G.R., M.S., M.D., H.A., E.G., J.-Y.M.-M., C.D. and L.D. conceived the ideas and designed methodology; M.D., A.Q., M.A.E.A., A.O., R.E.H., H.A., S.R., C.F.-A., N.H., W.A.L.I., H.H.A., A.A.E., H.I., K.E., E.B., A.S., A.G., B.M.H., M.S.S., N.B. collected the data; P.D.d.R., G.R., M.S., M.D. analysed the data; P.D.d.R., G.R., M.D., M.S., E.G., J.-Y.M.-M. and L.D. led the writing of the manuscript; P.D.d.R., C.D., J.-Y.M.-M. and L.D. ensured the acquisition of funding. All

authors contributed critically to the drafts and gave final approval for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13594.

DATA AVAILABILITY STATEMENT

Data deposited in the Figshare Digital Repository https://doi.org/10.6084/m9.figshare.12662360. Codes used for this article deposited in the Figshare Digital Repository https://doi.org/10.6084/m9.figshare.14054732.

ORCID

Mohamed Dakki https://orcid.org/0000-0002-8042-3972
Geneviève Robin https://orcid.org/0000-0002-6264-0842
Laura Dami https://orcid.org/0000-0002-8951-9751
Elie Gaget https://orcid.org/0000-0003-3462-9686
Pierre Defos du Rau https://orcid.org/0000-0002-8876-8529

REFERENCES

Amano, T., Székely, T., Sandel, B., Nagy, S., Mundkur, T., Langendoen, T., Blanco, D., Soykan, C. U., & Sutherland, W. J. (2018). Successful conservation of global waterbird populations depends on effective governance. *Nature*, *553*(7687), 199. https://doi.org/10.1038/nature25139

Bardos, D. C., Guillera-Arroita, G., & Wintle, B. A. (2015). Valid auto-models for spatially autocorrelated occupancy and abundance data. *Methods in Ecology and Evolution*, *6*(10), 1137–1149. https://doi.org/10.1111/2041-210X.12402

Blanchong, J. A., Joly, D. O., Samuel, M. D., Langenberg, J. A., Rolley, R. E., & Sausen, J. F. (2006). White-tailed deer harvest from the chronic wasting disease eradication zone in south-central Wisconsin. *Wildlife Society Bulletin*, *34*(3), 725–731. https://doi.org/10.2193/0091-7648(2006)34[725:WDHFTC[2.0.CO;2

Bogaart, P., van der Meij, T., Pannekoek, J., Soldaat, L., Van Strien, A. J., & Underhill, L. G. (2017). Comment on "Working with population totals in the presence of missing data comparing imputation methods in terms of bias and precision" by Onkelinx et al (2016). *Journal of Ornithology*, *158*(3), 887–889. https://doi.org/10.1007/s10336-017-1456-5

Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, *9*(6), 717–772. https://doi.org/10.1007/s10208-009-9045-5

Candès, E. J., & Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, *56*(5), 2053–2080. https://doi.org/10.1109/TIT.2010.2044061

Cherkaoui, S. I., Selmi, S., & Hanane, S. (2017). Ecological factors affecting wetland occupancy by breeding Anatidae in the southwestern Mediterranean. *Ecological Research*, *32*(2), 259–269. https://doi.org/10.1007/s11284-017-1436-5

Cusack, J. J., Duthie, A. B., Rakotonarivo, O. S., Pozo, R. A., Mason, T. H., Månsson, J., Nilsson, L., Tombre, I. M., Eythórsson, E., Madsen, J., Tulloch, A., Hearn, R. D., Redpath, S., & Bunnefeld, N. (2019). Time series analysis reveals synchrony and asynchrony between conflict management effort and increasing large grazing bird populations in northern Europe. *Conservation Letters*, *12*(1), e12450. https://doi.org/10.1111/conl.12450

Dakki, M., Qninba, A., El Agbani, M. A., Benhoussa, A., & Beaubrun, P. C. (2001). Waders wintering in Morocco: National population estimates, trends and site-assessments. *Wader Study Group Bulletin*, *96*, 47–59.

Donoho, D. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math Challenges Lecture, 1–32.

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, *30*, 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

EGA - RAC/SPA Waterbird Census Team. (2012). Atlas of wintering waterbirds of Libya, 2005-2010. Imprimerie COTIM.

El Agbani, M. A., Dakki, M., Beaubrun, P. C., & Thévenot, M. (1996). L'hivernage des anatidés (Anatidae) au Maroc (1990–94): Effectifs et sites d'importance Internationale et Nationale. *Gibier Faune Sauvage*, *13*, 233–249.

Ellington, E. H., Bastille-Rousseau, G., Austin, C., Landolt, K. N., Pond, B. A., Rees, E. E., Robar, N., & Murray, D. L. (2015). Using multiple imputation to estimate missing data in meta-regression. *Methods in Ecology and Evolution*, *6*(2), 153–163. https://doi.org/10.1111/2041-210X.12322

Fithian, W., & Josse, J. (2017). Multiple correspondence analysis and the multilogit bilinear model. *Journal of Multivariate Analysis*, *157*, 87–102. https://doi.org/10.1016/j.jmva.2017.02.009

Gaget, E., Le Viol, I., Pavón-Jordán, D., Cazalis, V., Kerbiriou, C., Jiguet, F., Popoff, N., Dami, L., Mondain-Monval, J. Y., Defos du Rau, P., Abdou, W. A. I., Bozic, L., Dakki, M., Encarnação, V. M. F., Erciyas-Yavuz, K., Etayeb, K. S., Molina, B., Petkov, N., Uzunova, D., … Galewski, T. (2020). Assessing the effectiveness of the Ramsar Convention in preserving wintering waterbirds in the Mediterranean. *Biological Conservation*, *243*, 108485. https://doi.org/10.1016/j.biocon.2020.108485

Galewski, T., Collen, B., McRae, L., Loh, J., Grillas, P., Gauthier-Clerc, M., & Devictor, V. (2011). Long-term trends in the abundance of Mediterranean wetland vertebrates: From global recovery to localized declines. *Biological Conservation*, *144*, 1392–1399. https://doi.org/10.1016/j.biocon.2010.10.030

Greenacre, M. (1984). *Theory and applications of correspondence analysis*. Academic Press.

Guillemain, M., Champagnon, J., Massez, G., Pernollet, C. A., George, T., Momerency, A., & Simon, G. (2015). Becoming more sedentary? Changes in recovery positions of Mallard Anas platyrhynchos ringed in the Camargue, France, over the last 50 years. *Wildfowl*, *65*(65), 51–63.

Han, X., Smyth, R. L., Young, B. E., Brooks, T. M., de Lozada, A. S., Bubb, P., Butchart, S. H. M., Larsen, F. W., Hamilton, H., Hansen, M. C., & Turner, W. R. (2014). A biodiversity indicators dashboard: Addressing challenges to monitoring progress towards the Aichi Biodiversity Targets using disaggregated global data. *PLoS One*, *9*(11), e112046. https://doi.org/10.1371/journal.pone.0112046

Harel, O., & Zhou, X. (2006). *Multiple imputation - Review of theory, implementation and software*. UW Biostatistics Working Paper Series. Working Paper 297. http://biostats.bepress.com/uwbiostat/paper297

Hughes, B. B., Beas-Luna, R., Barner, A. K., Brewitt, K., Brumbaugh, D. R., Cerny-Chipman, E. B., Close, S. L., Coblentz, K. E., de Nesnera, K. L., Drobnitch, S. T., Figurski, J. D., Focht, B., Friedman, M., Freiwald, J., Heady, K. K., Heady, W. N., Hettinger, A., Johnson, A., Karr, K. A., … Carr, M. H. (2017). Long-term studies contribute disproportionately to ecology and policy. *BioScience*, *67*(3), 271–281. https://doi.org/10.1093/biosci/biw185

Josse, J., & Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, *70*(1), 1–31.

Kindsvater, H. K., Dulvy, N. K., Horswill, C., Juan-Jordá, M. J., Mangel, M., & Matthiopoulos, J. (2018). Overcoming the data crisis in biodiversity conservation. *Trends in Ecology & Evolution*, *33*(9), 676–688. https://doi.org/10.1016/j.tree.2018.06.004

Lehikoinen, A., Jaatinen, K., Vähätalo, A., Clausen, P., Crowe, C., Deceuninck, B., Hearn, R., Holt, C. A., Hornman, M., Keller, V.,

Nilsson, L., Langendoen, T., Tománková, I., Wahl, J., & Fox, A. D. (2013). Rapid climate driven shifts in winter distributions of three common waterbird species. *Global Change Biology*, *19*, 2071–2081.

Lougheed, L. W., Breault, A., & Lank, D. B. (1999). Estimating statistical power to evaluate ongoing waterfowl population monitoring. *The Journal of Wildlife Management*, *63*, 1359–1369.

Maclean, I. M., Austin, G. E., Rehfisch, M. M., Blew, J. A. N., Crowe, O., Delany, S., Devos, K., Deceuninck, B., Günther, K., Laursen, K., Van Roomen, M., & Wahl, J. (2008). Climate change causes rapid changes in the distribution and site abundance of birds in winter. *Global Change Biology*, *14*(11), 2489–2500. https://doi.org/10.1111/j.1365-2486.2008.01666.x

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17–23. https://doi.org/10.1093/biomet/37.1-2.17

Murray, D. L., Anderson, M. G., & Steury, T. D. (2010). Temporal shift in density dependence among North American breeding duck populations. *Ecology*, *91*(2), 571–581. https://doi.org/10.1890/MS08-1032.1

Nakagawa, S., & Freckleton, R. P. (2008). Missing in action: The dangers of ignoring missing data. *Trends in Ecology & Evolution*, *23*(11), 592–596. https://doi.org/10.1016/j.tree.2008.06.014

Onkelinx, T., Devos, K., Jansen, I., Van Calster, H., & Quataert, P. (2017). Reply to the comment on 'Working with population totals in the presence of missing data comparing imputation methods in terms of bias and precision' by Bogaart et al *Journal of Ornithology*, *158*(3), 891–893. https://doi.org/10.1007/s10336-017-1457-4

Onkelinx, T., Devos, K., & Quataert, P. (2017). Working with population totals in the presence of missing data comparing imputation methods in terms of bias and precision. *Journal of Ornithology*, *158*(2), 603–615. https://doi.org/10.1007/s10336-016-1404-9

Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H., & Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution*, *5*(9), 961–970. https://doi.org/10.1111/2041-210X.12232

Pereira, H. M., Ferrier, S., Walters, M., Geller, G., Jongman, R. H. G., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H. M., Cardoso, A. C., Coops, N. C., Dulloo, E., Faith, D., Freyhof, J., Gregory, R., Heip, C., Höft, R., Hurtt, G., Jetz, W., … Wegmann, M. (2013). Essential biodiversity variables. *Science*, *339*, 277–278. DOI: 10.1126/science.1229931

Robin, G., Josse, J., Moulines, É., & Sardy, S. (2019). Low-rank model with covariates for count data with missing values. *Journal of Multivariate Analysis*, *173*, 416–434. https://doi.org/10.1016/j.jmva.2019.04.004

Robin, G., Klopp, O., Josse, J., Moulines, É., & Tibshirani, R. (2019). Main effects and interactions in mixed and incomplete data frames. *Journal of the American Statistical Association*, *115*(531), 1-31. https://doi.org/10.1080/01621459.2019.1623041

Samraoui, F., Alfarhan, A. H., Al-Rasheid, K. A., & Samraoui, B. (2011). An appraisal of the status and distribution of waterbirds of Algeria: Indicators of global changes? *Ardeola*, *58*, 137–163. https://doi.org/10.13157/arla.58.1.2011.137

Sayoud, M. S., Salhi, H., Chalabi, B., Allali, A., Dakki, M., Qninba, A., El Agbani, M. A., Azafzaf, H., Feltrup-Azafzaf, C., Dlensi, H., Hamouda, N., Abdel Latif Ibrahim, W., Asran, H., Abu Elnoor, A., Ibrahim, H., Etayeb, K., Bouras, E., Bashaimam, W., Berbash, A., … Defos du Rau, P. (2017). The first coordinated trans-North African mid-winter waterbird census: The contribution of the International Waterbird Census to the conservation of waterbirds and wetlands at a biogeographical level. *Biological Conservation*, *206*, 11–20. https://doi.org/10.1016/j.biocon.2016.12.005

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Stephenson, P. J., Bowles-Newark, N., Regan, E., Stanwell-Smith, D., Diagana, M., Höft, R., Abarchi, H., Abrahamse, T., Akello, C., Allison, H., Banki, O., Batieno, B., Dieme, S., Domingos, A., Galt, R., Githaiga, C. W., Guindo, A. B., Hafashimana, D. L. N., Hirsch, T., … Thiombiano, A. (2017). Unblocking the flow of biodiversity data for decision-making in Africa. *Biological Conservation*, *213*, 335–340. https://doi.org/10.1016/j.biocon.2016.09.003

Stephenson, P. J., Brooks, T., Butchart, S., Fegraus, E., Geller, G. N., Hoft, R., Hutton, J., Kingston, N., Long, B., & McRae, L. (2017). Priorities for big biodiversity data. *Frontiers in Ecology and the Environment*, *15*(3), 124–125. https://doi.org/10.1002/fee.1473

Stephenson, P. J., Burgess, N. D., Jungmann, L., Loh, J., O'Connor, S., Oldfield, T., Reidhead, W., & Shapiro, A. (2015). Overcoming the challenges to conservation monitoring: Integrating data from in-situ reporting and global data sets to measure impact and performance. *Biodiversity*, *16*(2–3), 68–85. https://doi.org/10.1080/14888386.2015.1070373

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Van Roomen, M., van Winden, E., & van Turnhout, C. (2011). Analyzing population trends at the flyway level for bird populations covered by the African Eurasian Waterbird Agreement: Details of a methodology. SOVON-information Report, 5, 22 p.

Van Strien, A., Pannekoek, J., Hagemeijer, W., & Verstrael, T. (2004). A loglinear Poisson regression method to analyse bird monitoring data. *Bird*, *482*, 33–39.

Van Swaay, C. A., Nowicki, P., Settele, J., & Van Strien, A. J. (2008). Butterfly monitoring in Europe: Methods, applications and perspectives. *Biodiversity and Conservation*, *17*(14), 3455–3469. https://doi.org/10.1007/s10531-008-9491-4

Visser, M. E., Perdeck, A. C., van Balen, J. H., & Both, C. (2009). Climate change leads to decreasing bird migration distances. *Global Change Biology*, *15*(8), 1859–1865. https://doi.org/10.1111/j.1365-2486.2009.01865.x

Wauchope, H. S., Johnston, A., Amano, T., & Sutherland, W. J. (2019). When can we trust population trends? A method for quantifying the effects of sampling interval and duration. *bioRxiv*, 498170.

Wetlands International. (2010). *Guidance on waterbird monitoring methodology: Field Protocol for waterbird counting*. https://wpe.wetlands.org

Wetlands International. (2019). *Waterbird population estimates*. https://wpe.wetlands.org

White, E. R. (2019). Minimum time required to detect population trends: The need for long-term monitoring programs. *BioScience*, *69*(1), 40–46. https://doi.org/10.1093/biosci/biy144

Wintle, B. A., & Bardos, D. C. (2006). Modeling species–habitat relationships with spatially autocorrelated observation data. *Ecological Applications*, *16*(5), 1945–1958. https://doi.org/10.1890/1051-0761(2006)016[1945:MSRWSA]2.0.CO;2

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Dakki M, Robin G, Suet M, et al. Imputation of incomplete large-scale monitoring count data via penalized estimation. *Methods Ecol Evol*. 2021;12:1031–1039. https://doi.org/10.1111/2041-210X.13594