# 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, Morocco
[2]Inria - Université Gustave Eiffel, CERMICS (ENPC), F-77455 Marne-la-Vallée, France
[3] Centre de Recherche de la Tour du Valat, Le Sambuc, 13200 Arles, France
[4] Faculté des Sciences Ben M'sik, Univ. Hassan II, Casablanca, Morocco
[5] Association "Les Amis des Oiseaux" (AAO/BirdLife en Tunisie), 14, Rue Ibn El Heni, 2ème étage - Bureau N° 4, 2080 Ariana , Tunisia.
[6] Egyptian Environmental Affairs Agency, 30 Misr/Helwan Road, PO 11728, El Maadi Helwan, Egypt
[7] Zoology Dept., Tripoli University, PO Box: 13227, Tripoli, Libya
[8] Environment General Authority, Ganzor Algheran, PO Box 13793, Tripoli, Libya
[9] Libyan Society for Birds, P.O. Box 81417, Tripoli, Libya
[10] Conservation des Forêts de la Wilaya d'Oran, 31000, Oran, Algeria
[11]Centre Cynégétique de Réghaia, Direction Générale des Forets, BP 54/02 Réghaia 16112 Alger, Algeria
[12]Direction Générale des Forêts, 11 Chemin Doudou Mokhtar, Ben Aknoun, 16000 Alger, Algeria
13 International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
[14] Office Français pour la Biodiversité, Unité Avifaune Migratrice, Le Sambuc, 13200 Arles, France

* corresponding author: pierre.defosdurau@ofb.gouv.fr ; tel: +33 (0)4 90 97 06 71

**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 6 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 suffers 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.

**Résumé**

1. Dans le contexte du suivi de la biodiversité, de très grands jeux de données sont aujourd'hui disponibles et de plus en plus utilisés à l'échelle mondiale pour estimer les

tendances des espèces ainsi que leur état de conservation. Ces jeux de données de grande échelle remettent en question les méthodes d'analyses statistiques existantes, puisque beaucoup ne sont pas adaptées à leur taille, leur hétérogénéité et leur caractère incomplet. Le développement de nouvelles méthodes permettant d'imputer les valeurs manquantes dans les données de suivis de grande échelle est donc essentiel pour équilibrer l'échantillonnage dans le temps ou dans l'espace, et ainsi contribuer à mieux définir les politiques de conservation.

2. Nous avons développé une nouvelle méthode basée sur des modèles de Poisson pénalisés pour imputer et analyser les données manquantes des suivis de grande échelle. Cette méthode permet de paramétrer (a) les effets spatiaux et temporels, (b) les effets principaux des covariables prédictives, ainsi que (c) les interactions spatio-temporelles. Elle présente également une performance de calcul satisfaisante lors de l'estimation de nombreux paramètres.

3. La méthode a été largement testée sur des données simulées et réelles de comptages d'oiseaux d'eau ; les résultats révèlent qu'elle surpasse 6 méthodes existantes en termes d'erreurs d'imputation de données manquantes. En appliquant cette méthode sur 16 espèces d'oiseaux d'eau, nous avons estimé pour la première fois leur tendance sur le long terme à l'échelle de l'Afrique du Nord, une région où les données de suivi souffrent de nombreuses lacunes, spatiales et temporelles.

4. Cette nouvelle approche ouvre des perspectives prometteuses dans l'amélioration de la précision des estimations des tendances d'abondance d'espèces. La méthode est disponible dans une librairie R libre appelée «lori» (https://CRAN.R-project.org/package=lori), et nous recommandons son utilisation pour les données de comptage à grande échelle, en particulier dans les programmes de sciences participatives sur les suivis d'espèces.

**Key words**

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 et al. 2017b). 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 data sets 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 et al. 2015, 2017a). 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 (Nakagawa & Freckleton 2008, Penone et al. 2014, Ellington et al. 2015). 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 (Van Strien et al. 2004, Van Swaay et al. 2008, Lehikoinen et al. 2013). Other commonly used methods rely on chained equations (Van Buuren & Groothuis-Oudshoorn 2011) or Random Forests (Stekhoven and Bühlmann 2012). More recently, the use of multiple imputation procedures has been discussed (Onkelinx et al. 2017a & b, Bogaart 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, i.e. 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, Klopp et al. 2019). Such situations with a large amount of missing data and large predictor sets are frequent, as species count data is 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 (Murray et al. 2010, Stephenson et al. 2015, Amano et al. 2018).

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 (Samraoui et al. 2011, Galewski 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 (El Agbani et al. 1996, Dakki et al. 2001, EGA - RAC/SPA waterbird census team 2012, 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 6 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:

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

In Eq. (1), $\alpha_i$ corresponds to the effect of site $i$, and $\beta_j$ 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:

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

In Eq. (2), the additional coefficients $\gamma_k$ 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 $\theta_{ij}$. Model (2) becomes:

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

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 ($\alpha$, $\beta$ and $\gamma$) 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 $\theta$ to be of low rank. Indeed, if $\theta$ is of rank $r$, then for any site $i$ and year $j$, the corresponding interaction $\theta_{ij}$ 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, $\sigma_l$ 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 $\alpha$, $\beta$ and $\gamma$, and by the number of latent factors driving the interactions. The standard Poisson negative log-likelihood is given by:

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

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 $\lambda_1$ and $\lambda_2$ control the trade-off between fitting the data and imposing low-complexity models: the larger $\lambda_1$ and $\lambda_2$, 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 and Recht 2009, Candès and 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:

$$(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\theta}s.) \in argmin\, L(\alpha, \beta, \gamma, \theta) + \lambda_1 \|\theta\|_* + \lambda_2 (\|\alpha\|_1 + \|\beta\|_1 + \|\gamma\|_1). \quad (8)$$

This estimation problem was initially studied in Robin, Josse et al. (2019) and in Robin, Klopp et al. (2019). In these papers, the authors provide strong theoretical guarantees of the estimation capacities of (8). In particular, the main advantage of (8) is that the estimation error of the parameters $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\theta})$ 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 is known to be prone to over-

dispersion 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 3$^{rd}$ 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: $Y_1, \cdots, Y_N$ denotes the true values of the missing data, and $\hat{Y}_1, \cdots, \hat{Y}_N$ denotes the corresponding imputed values:

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

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

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

*2.3 Comparing imputation methods*

Using the data sets 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. Greenacre 1984, Fithian and Josse 2017). We used the implementation of the R package 'missMDA' (Josse & Husson 2016). The fourth competitor was TRIM, a widely applied imputation model in wildlife monitoring schemes (Van Strien et al. 2004, Van Swaay et al. 2008, Lehikoinen et al. 2013), 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 and 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 (Fig. 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/).

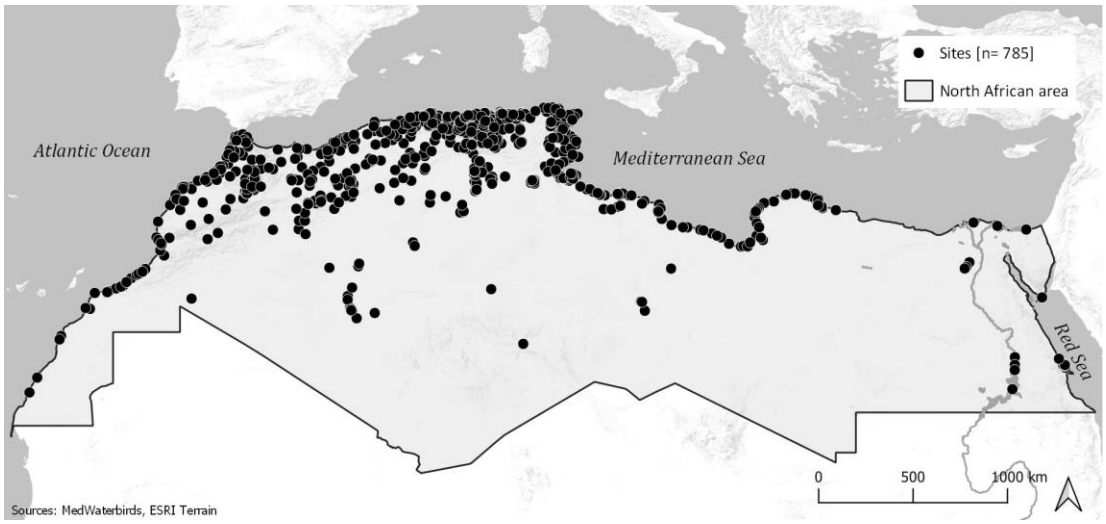



**Figure 1**: The 785 IWC monitoring sites surveyed for at least two years between 1990 and 2017.

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 (Fig. 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.

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 semi-variogram 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 & S4 for a complete description of the covariates and ecological hypotheses.

## 3. Results

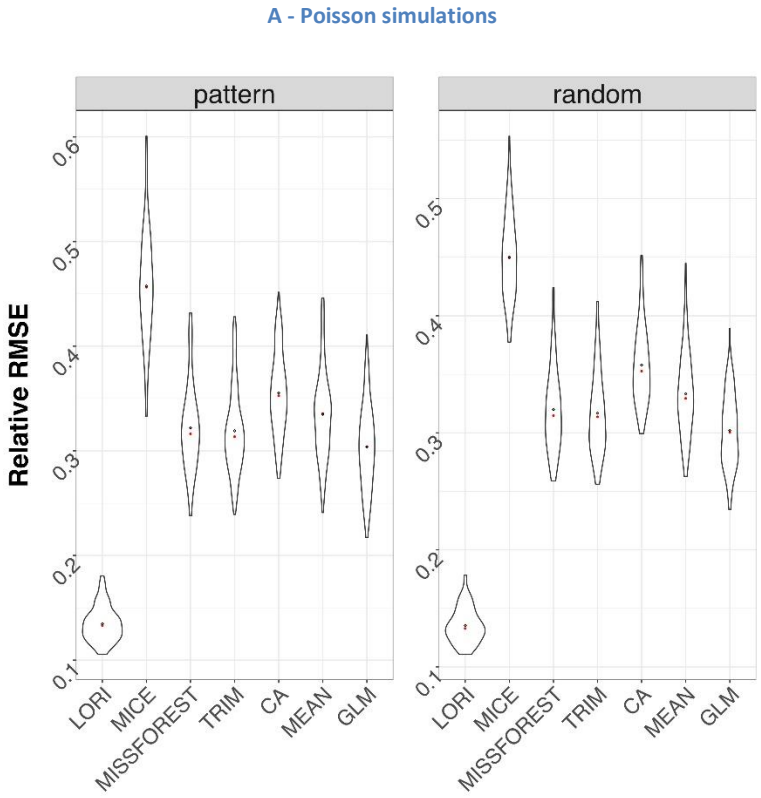
### *3.1 Comparing LORI to existing imputation methods*

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.

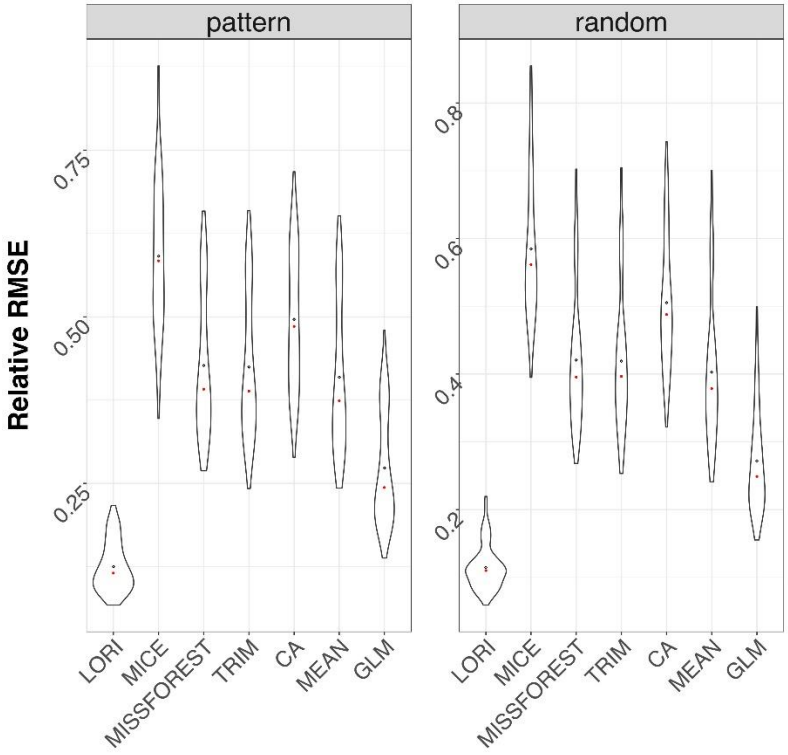
Overall, the LORI method outperformed competitors in both accuracy and precision (Fig. 2). If the dataset had 20% missing data, LORI provided a stable imputation procedure, with around 0.2 relative RMSE for the Poisson and ZINB data, with little variability. The competing methods had larger imputation 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 (10ms), then CA and TRIM (200 ms), then LORI (2s) and finally MICE and MISSFOREST (5-10s); 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 5 times larger than TRIM and CA.

The results of the experiment on ZINB simulated data (Fig.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 (Fig. 2C), LORI improved on its competitors in accuracy, and yielded quite stable imputation results.

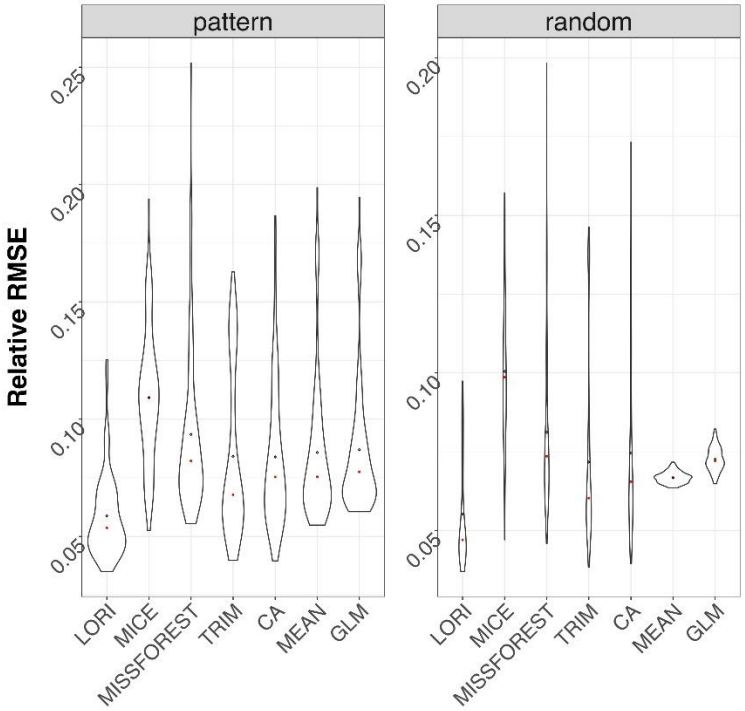
**Figure 2:** Violin plot of relative RMSE for eight imputation methods on (A) simulated Poisson data (**20%** missing data), (B) simulated ZINB data (**20%** missing data) and (C) real-life Northern Shoveler count data (**10%** additional missing data amounting to **30%** overall missing data), for two missing data patterns (see 'Materials & methods'). Mean and median are indicated by points (mean: black point, median: red point)



B - ZINB simulations




C – Northern Shoveler simulations



*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, 2 showed different trends between LORI and TRIM (Great Cormorant and Northern Shoveler, see Fig. 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).

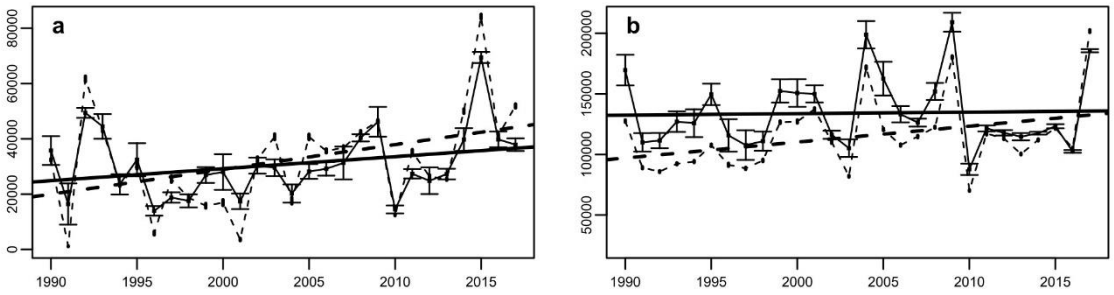



**Figure 3:** Yearly counts over all 785 North African sites for the Great Cormorant (a) and the Northern Shoveler (b) as modelled by LORI (solid lines) and TRIM (dotted lines) with the respective linear time trend.

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 two 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 this data is most needed, it often 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 $\lambda_1$ and $\lambda_2$ 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 (Fig. 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, e.g. a ten-year span, which is the recommended timescale for short-term waterbird trend assessment (Van Roomen et al. 2011, Lougheed et al. 1999). 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 (Wintle & Bardos 2006, Bardos et al. 2015). 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, 3 out 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.

**Conclusion**

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.

Funding: This work, including data collection and analysis, was supported by the French Ministry in charge of Environment (Ministère de la Transition Ecologique et Solidaire) throught 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.

**Authors' contributions**

Pierre Defos du Rau, Geneviève Robin, Marie Suet, Mohamed Dakki, Hichem Azafzaf, Elie Gaget, Jean-Yves Mondain-Monval, Clemence Deschamps and Laura Dami conceived the ideas and designed methodology; Mohamed Dakki, Abdeljebbar Qninba, Mohammed A. El Agbani, Asmâa Ouassou, Rhimou El Hamoumi, Hichem Azafzaf, Sami Rebah, Claudia Feltrup-Azafzaf , Naoufel Hamouda, Wed A.L. Ibrahim, Hosni H. Asran, Air A. Elhady, Haitham Ibrahim, Khaled Etayeb, Essam Bouras, Almokhtar Saied, Ashrof Glidan, Bakar M. Habib, Mohamed S. Sayoud, Nadjiba

Bendjedda collected the data; Pierre Defos du Rau, Geneviève Robin, Marie Suet, Mohamed Dakki analysed the data; Pierre Defos du Rau, Geneviève Robin, Mohamed Dakki, Marie Suet, Elie Gaget, Jean-Yves Mondain-Monval and Laura Dami led the writing of the manuscript. Pierre Defos du Rau, Clemence Deschamps, Jean-Yves Mondain-Monval and Laura Dami ensured the acquisition of funding. All authors contributed critically to the drafts and gave final approval for publication.

**Data availability**

Data deposited in the Figshare Digital Repository: Dakki et al. (2021), https://doi.org/10.6084/m9.figshare.12662360

Codes used for this article deposited in the Figshare Digital Repository : Dakky et al. (2020), https://doi.org/10.6084/m9.figshare.14054732

**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.

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.

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.

Bogaart P, van der Meij T, Pannekoek J, Soldaat L, Van Strien AJ, Underhill LG (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). J Ornithol. doi: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.

Candès, E. J., Tao, T. (2010) The power of convex relaxation: Near-optimal matrix completion. IEEE Transactions on Information Theory 56(5), 2053-2080.

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.

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.

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 Bull., 96, 47-59.

Donoho, David. (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.

EGA - RAC/SPA waterbird census team, 2012. Atlas of wintering waterbirds of Libya, 2005-2010. Imprimerie COTIM, Tunisia.

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, Game Wildl., 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.

Fithian, W. and Josse, J., 2017, Multiple correspondence analysis and the multilogit bilinear model, Journal of Multivariate Analysis, 157 87--102.

Gaget E, Le Viol I, Pavón-Jordán D, Cazalis V, Kerbiriou C, Jiguet F, Popoff N, Dami L, Mondain-Monval JY, Defos du Rau P, Abdou WAI, Bozic L, Dakki M, Encarnação VMF, Erciyas-Yavuz K, Etayeb KS, Molina B, Petkov N, Uzunova D, Zenatello M, Galewski T., 2020. Assessing the effectiveness of the Ramsar Convention in preserving wintering waterbirds in the Mediterranean. Biological Conservation 243 : 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.

Greenacre, M. (1984) Theory and Applications of Correspondence Analysis, Academic Press, Cambridge MA.

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.

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

Harris, I. P. D. J.; Jones, P. D.; Osborn, T. J.; & Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset. International journal of climatology, 34(3), 623-642. doi: 10.1002/joc.3711.

Harris, I., Osborn, T.J., Jones, P. et al. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Sci Data 7, 109 (2020). https://doi.org/10.1038/s41597-020-0453-3

Hughes, B. B., Beas-Luna, R., Barner, A. K., Brewitt, K., Brumbaugh, D. R., Cerny-Chipman, E. B., ... & Figurski, J. D. (2017). Long-term studies contribute disproportionately to ecology and policy. BioScience, 67(3), 271-281.

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.

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.

Little, R. J. A. and Rubin, D. B. (2002). Statistical Analysis with Missing Data. *Population* 43(6), 1174-1174. https://doi.org/

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.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1/2), 17-23.

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.

Nakagawa, S., & Freckleton, R. P. (2008). Missing in action: the dangers of ignoring missing data. *Trends in Ecology & Evolution*, *23*(11), 592-596.

Onkelinx T, Devos K, Quataert P (2017a) Working with population totals in the presence of missing data comparing imputation methods in terms of bias and precision. J Ornithol. doi:10.1007/s10336-016-1404-9

Onkelinx, T., Devos, K., Jansen, I., Van Calster, H. & Quataert, P. (2017b). 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.

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., Karp, D., McGeoch, M., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J., Stuart, S., Turak, E., Walpole, M., Wegmann, M. (2013). Essential biodiversity variables. Science. 339:pp. 277–278. http://dx.doi.org/10.1126/science.12299.

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.

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.

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, 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.

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., Bashaiman, W., Berbash, A., Deschamps, C., Mondain-Monval, J.L., Brochet, A.L., Véran, S. & 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.

Stekhoven, D. J. and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data, *Bioinformatics*, 28(1) 112–118.

Stephenson, P. J., Neil D. Burgess, Laura Jungmann, Jonathan Loh, Sheila O'Connor, Thomasina Oldfield, Will Reidhead & Aurélie Shapiro (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,DOI: 10.1080/14888386.2015.1070373

Stephenson, P. J., Bowles-Newark, N., Regan, E., Stanwell-Smith, D., Diagana, M., Höft, R. & Banki, O. (2017a). Unblocking the flow of biodiversity data for decision-making in Africa. *Biological conservation*, *213*, 335-340.

Stephenson, P. J., Brooks, T., Butchart, S., Fegraus, E., Geller, G.N., Hoft, R., Hutton, J., Kingston, N., Long, B. & McRae, L. (2017b). Priorities for big biodiversity data. *Frontiers in Ecology and the Environment*, *15*(3), 124-125.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1 - 67. doi:http://dx.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, 22p.

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.

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. wpe.wetlands.org (accessed 5.10.12).

Wetlands International, 2019. Waterbird Population Estimates. wpe.wetlands.org (accessed 3.09.19).

White, E. R. (2019). Minimum time required to detect population trends: the need for long-term monitoring programs. BioScience, 69(1), 40-46.

Wintle, B. A., & Bardos, D. C. (2006). Modeling species–habitat relationships with spatially autocorrelated observation data. Ecological Applications, 16(5), 1945-1958.

**Supporting information**

Additional Supporting Information may be found in the online version of this article.

The code of all experiments is available as supporting information.

**Appendix S1.** Additional numerical experiments

**Appendix S2.** List of professional or volunteer field observers involved in the collection of waterbird count data

**Appendix S3.** Spatial and temporal covariates used as predictors

**Appendix S4.** A priori hypothesis governing our choice of covariates

**Appendix S5.** Yearly count totals over all North African sites for the 16 selected species as modelled by LORI

**Appendix S6.** Effect parameters for all spatial and temporal covariates as estimated by LORI for the 16 selected species

## Appendix S1

Here we present additional results of our experiments described in Section 3.1, with increasing proportion of missing values: 20%, 40%, 60% and 80% for the Poisson and ZINB simulations, and 30%, 40%, 50% and 70% for the northern shoveler. Overall, we observe that LORI is quite robust up to 60% of missing values (RMSE less than 0.4). For 80% of missing values (simulations), and 70% (northern shoveler data), the RMSE increases, and goes up to 0.7. However, in this setting, LORI is the only one of the methods tested that has an RMSE smaller than 1 (i.e. does better than simply imputing by 0). We also observe that all methods seem less robust to large proportions of missing values in the "pattern" missing data setting where missing values are aggregated along specific rows.

### *S1.1 Imputation of Poisson simulated data*

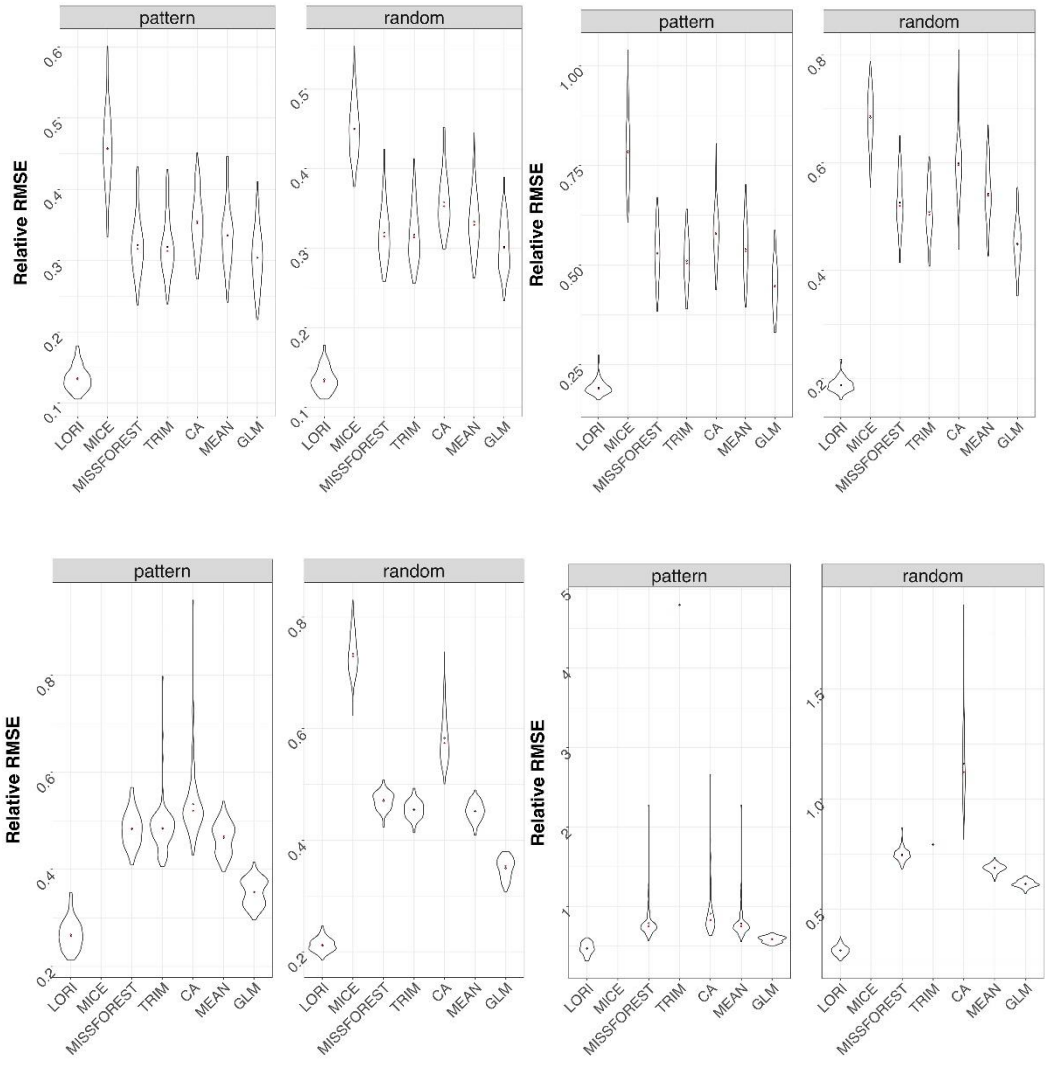



**Figure 6:** Relative RMSE of 5 imputation methods replicated 100 times for a simulated Poisson dataset with **20%** (top left), **40%** (top right), **60%** (bottom left) and **80%** (bottom right) missing data. Mean: black point; median: red point.

### *S1.2 Imputation of zero-inflated negative binomial simulated data*

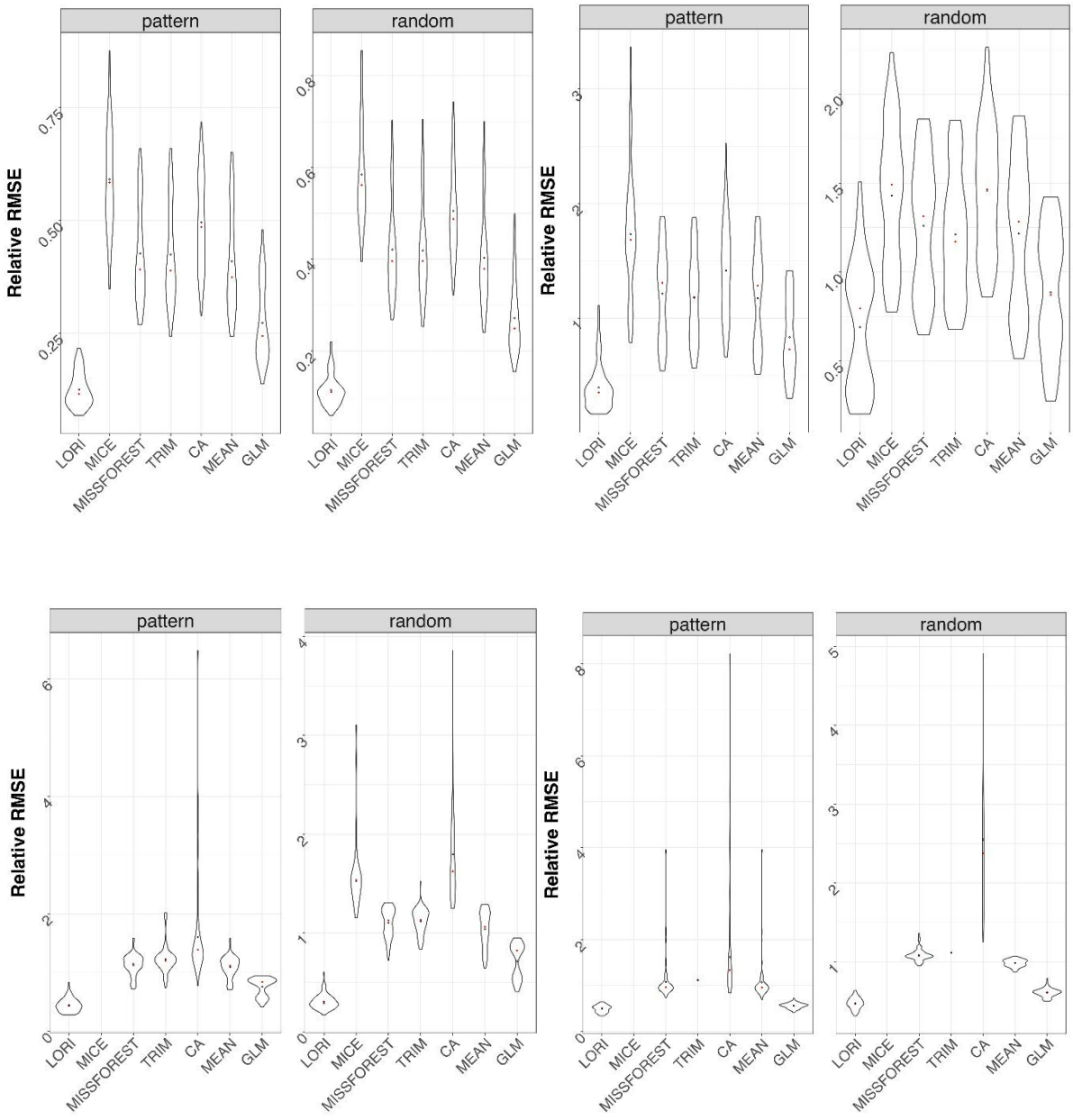



**Figure 7:** Relative RMSE of 5 imputation methods replicated 100 times for a simulated zero-inflated negative binomial dataset with **20%** (top left), **40%** (top right), **60%** (bottom left) and **80%** (bottom right) missing data.

*S1.3 Imputation of Northern Shoveler count data*

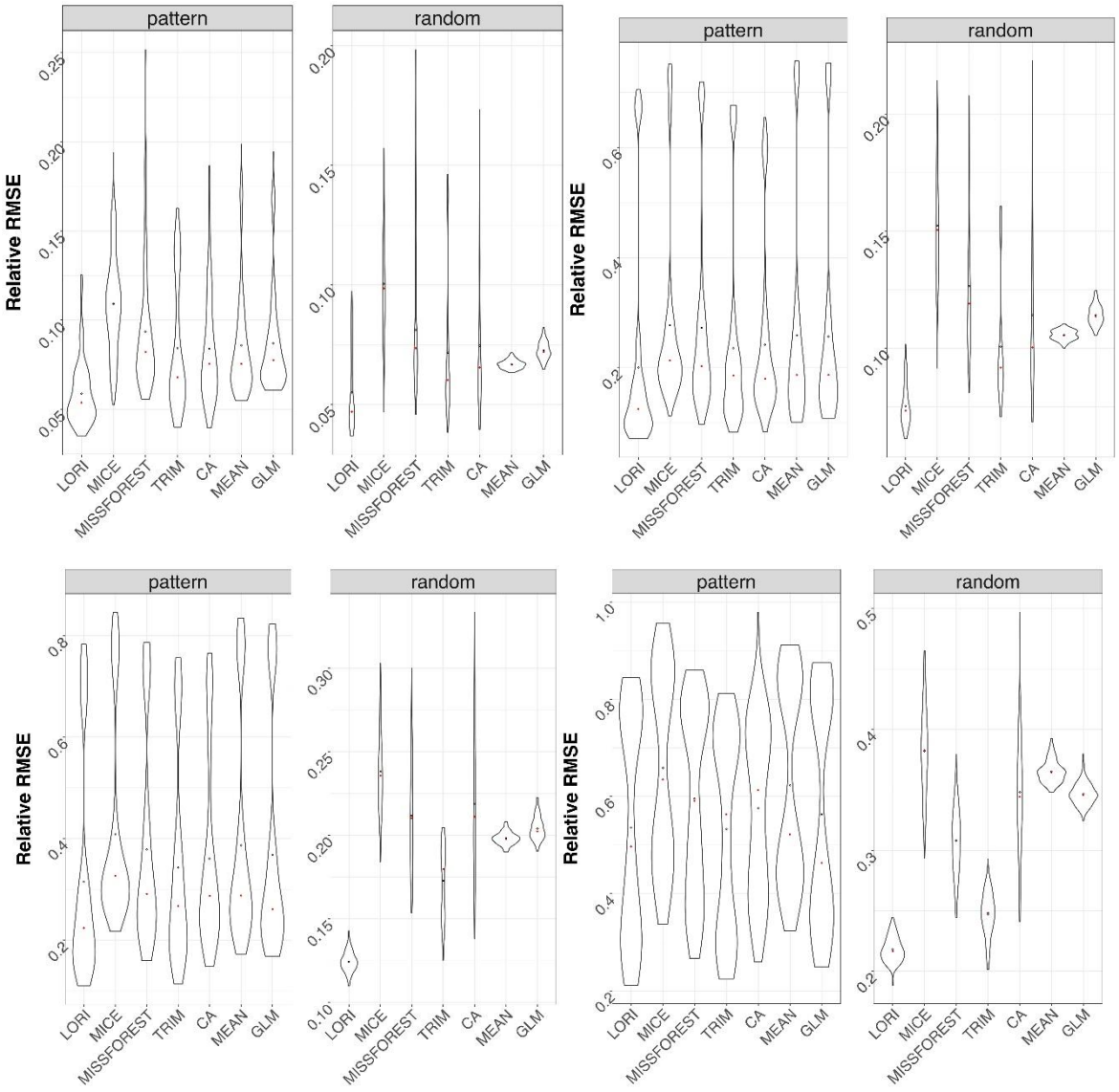



**Figure 8:** Relative RMSE of 5 imputation methods replicated 100 times for the Northern Shoveler dataset with **30%** (top left), **40%** (top right), **50%** (bottom left) and **70%** (bottom right) of additional missing data.

*S1.4 Parameters estimation and computational time*

In addition to the imputed dataset, LORI and the GLM model also output covariate coefficient estimates for the initial Poisson GLM used to generate the first simulated dataset. Both LORI- and GLM-estimated covariate coefficients are compared to the true coefficients in Table 1, showing that the two estimates were equally accurate.

Table 1: Estimated covariate coefficients for LORI and GLM (Poisson dataset, 40% of missing data, 100 replications). Mean value in bold and standard deviation in parentheses.

| | **X1** | **X2** | **X3** | **X4** | **X5** |
|---|---|---|---|---|---|
| **True values** | 0.5 | 0 | -0.5 | 0 | 0.5 |
| **LORI estimates** | **0.50**(0.005) | **0.00**(0.005) | **-0.51**(0.005) | **-0.01**(0.005) | **0.50**(0.005) |
| **GLM estimates** | **0.50**(0.008) | **0.00**(0.007) | **-0.52**(0.01) | **-0.02**(0.01) | **0.54**(0.01) |

Finally, we present in Table 2 the computational time (in seconds) of the competing methods averaged across 100 simulations performed on a machine with Intel 4-core (i7) 2.3GHz with 16GB RAM. The reported computational times correspond to the simulation with synthetic Poisson data and 20% of missing values; we observed comparable times across all simulations.

Table 2: Computational time in seconds of each competing method averaged across 100 simulations performed on a machine with Intel 4-core (i7) 2.3GHz with 16GB RAM. Results are presented for the three types of simulations.

| | **LORI** | **MICE** | **missForest** | **TRIM** | **CA** | **GLM** | **MEAN** |
|---|---|---|---|---|---|---|---|
| **Poisson data** | 1.60 | 9.83 | 5.62 | 0.41 | 0.281 | 0.02 | 0.00 |
| **ZINB data** | 1.13 | 10.15 | 3.87 | 0.50 | 0.65 | 0.01 | 0.00 |
| **Northern shoveler data** | 4.19 | 4.12 | 5.04 | 0.40 | 0.05 | 0.03 | 0.00 |

**Appendix S2**

List of professional or volunteer field observers involved in the collection of waterbird count data between 1990 and 2017.

**Algeria :** Abadou Samir, Abbassia Maatougui, Abd El Fateh Djennas, Abda Loudjane, Abdel Ali Bakir, Abdel Hafid Boudraa, Abdelah Arab, Abdelaziz Abbassi, Abdeldjalil Bouras, Abdelghani Maouche, Abdelhakim Benmokhtar Elalmi, Abdelhamid Hanache, Abdelhamid Malki, Abdelhamid Meddah, Abdelhamid Zaidi, Abdelkader Allali, Abdelkader Dahmoune, Abdelkader Djarri, Abdelkader Hachmaoui, Abdelkader Hadi, Abdelkader Horo, Abdelkader Ketfi, Abdelkader Maseghouni, Abdelkader Moumnin, Abdelkader Sahnoun, Abdelkarim Belahcen, Abdelkarim Guendouz, Abdelkarim Kaci, Abdelkaser Abdelhadi, Abdellah Bay, Abdellah Benamane, Abdellah Boudraa, Abdellah Doufen, Abdellah Tayebi, Abdellah Yousfi, Abdellatif Gasmi, Abdelmalek Raddas, Abdelwahab Chedad, Abdelwaheb Meradi, Abderaouf Sassa, Abderezak Benyekhlef, Abderrahmane Boulahlib, Abderrahmane Hamrani, Abdeselem Boushaba, Abdessalam Bettahar, Abdessalem Grira, Adda Bouich, Adel Bezzela, Adel Ghilani, Adel Hamadi, Adel Soualah, Adlene Madaci, Ahmed Bouzekri, Ahmed Djemai, Ahmed Ferroukhi, Ahmed Guerfi, Ahmed Kessal, Ahmed Mouissi, Ahmed Salem, Ahmed Souissi, Ahmim Djaouida, Aida Saifouni, Aimen Boulaoued, Aissa Essahel, Aissa Fillali, Aissa Medjahdine, Akli Amour, Ali Aichouni, Ali Banne, Ali Benlami, Ali Benzahir, Ali Berboucha, Ali Daif, Ali Ferhouni, Ali Ferradji, Ali Gharbi, Ali Gouacem, Ali Kadi, Ali Kamal Boualem, Ali Nouredine, Ali Zagrour, Amar Belâama, Amar Boumeuzber, Amar Bousba, Amar Fayçal Benali, Amar Gherbi, Amar Koudache, Amar Soltani, Amel Chatti, Amel Mokrani, Ameur Rahmouni, Amine Kahlerasse, Amine Miloud Zelif, Amor Bendjedidi, Amoumen Barrouk, Aoudia Mohamed Oulhadj, Arslane Benseddik, Assem Zarouala, Assia Brahiti, Assia Mattala, Atroune Abdelmalek Lamkoureb, Azzedine Beliacine, Azzedine Gherbi, Baba Arbi Houria, Bachir Mihoub, Bachir Remil, Bachir Sedira, Bagura Razki, Belabes Ben Garaa, Belkacem Chaib, Belkacem Chebrek, Belkacem Mhamdia, Belkhir Ben Bouzid, Ben Azouaou Salah, BenAbderrahmane Abdel Wahab, Benalia Azzeddine Beliacine, Bensalem Naas, Benyattou Araibi Sadek, Bouabdalah Rebah, Bouabdellah Nedjadi, Boualem Agoune, Boualem Kouma, Boubekar Chenafi, Boubeker Talha, Boudjemaa Nouari, Bouelem Khaladi, Bouhadjar Maghari, Boukabacha Makhefi, Boukrouk El Khamsa, Bouzid Chalabi, Bouzid Mesbahi, Brahim Adda Bouziane, Brahim Akkouche, Brahim Omri, Chaabane Meziane, Chahinaz Nezzar, Chawki Zigh, Dahia Imatoukene, Darradji Bâadache, Djahid Boussaha, Djahida Boukhalfa, Djamel Bachiri, Djamel Boularas, Djaouida Ahmim, Djelloul Yakoubene, Djelloul Zaoui, Djillali Belghoul, Doria Chirine Nedjai, Dounia Khelifi, E. Nour, Farah Yessaad, Fares Kessoum, Farid Hachemi, Farid Yachi, Farouk Terki, Fateh Yahiaou, Fatiha Ameur, Fatiha Bakaria, Fatiha Chikhe, Fatiha Ferradji, Fatima Larinouna, Fatma Mellem, Fatma Zohra Remichi, Faycel Gazzouz, Fellah Lahcen, Ferroudja Dokkal, Fethi Toubi, Fodil Touati, Fouad Dernouni, Gaci Aziz, Ghania Chalabi, Ghania Oumari, H. Hassani, H. Ounes, Habib Bakkar, Habib Grin, Hachemi Feghoul, Hadjira Hannachi, Hafed Hafdaoui, Hafedha Hasnaoui, Hafid Badreddine, Hahia Bendrabni, Hakkoum Oum Keltoum, Hamadou Ahmed Faouzi, Hassen Haridi, Hichem Benghazet, Hichem Diab, Hind Samai, Hocine Boudraa, Hocine Sadat, Houari Bouchiba, Houari Medjahdi, Houari Tadj, Houda Betatache, Houria Khemkhoum, Idir Gassi, Idir Sadat, Ikram Bendahmane, Ilias Laghmissi, Imane Boukhaloua, Imed Djemadi, Imen Allout, Ismahane Djendli, Jamel Horri, Jihane Zaafri, Kadour Boudinar, Kamal Bouchama, Kamal Gouaref, Kamal Messaouden, Kamal Oulmane, Kamel Bouchama, Karim Dehlas, Karim Touil, Karima Rakem, Khadidja Dhilis, Khadidja Moulay Meliani, Khaldia Boudraa, Khaled Bakhadra, Khaled Haffar, Khaoues Belaid, Khaoula Belahcini, Kheira Achit, Kheira Sebaa, Khemissi Dekhi, Kouider Benhamida, Kouider Litim, Lahouari Djardini, Laid Tadj, Lakhdar Sehini, Larbi Badaoui, Lotfi Moussouni, Lyamine Ben Kara, Lyna Gouichiche, M. Barhoune, M. Bouregaa, Maamar Doumi, Mabrouk Chergui, Mabrouk Guesmia, Mabrouk Lalmi, Mabrouk Tria, Madjid Kaci, Mahmoud Bouteldja, Mahmoud Malki, Mahmoud Mezzi, Mahmoud Nedjari, Malika Benhadj, Manal Haniche, Mariem Zermane, Mebarek Dehamchi, Meguelati Lokmane, Mehmoud Belmahdi, Merouane Cheliout, Merzouk Belaa, Messali Nedjma, Messaoud Adhimen, Messaoud Benmerad, Messaoud Chaoui, Messaoud Gueddoul, Messaoud Ouarem, Messaoud Rouene, Messaoud Saoudi, Messaouda Belgourte, Messaouda Dada, Meziane Chaabane, M'hamed Arous, M'Hamed Chaddad, Miloud Amine Zahaf, Miloud Athamna, Miloud Mechri, Mme Rouane, Mohamed Abbou, Mohamed Adda Djeffal, Mohamed Amine Sadji, Mohamed Baazizi, Mohamed Babai, Mohamed Belghoul, Mohamed Benchouk, Mohamed Bensassi, Mohamed Boughalia, Mohamed Bouhedja, Mohamed Boukri, Mohamed Boutlélis Chikh, Mohamed Djemal, Mohamed Djermani, Mohamed Faouzi Haou, Mohamed Gacem, Mohamed Gouichiche, Mohamed Hayouni, Mohamed Ishak Boushaki, Mohamed Khelifa Kerfa, Mohamed Khiari, Mohamed Lachachi, Mohamed Laidini, Mohamed Lamine Megharbi, Mohamed Maarouf, Mohamed Mahammedi, Mohamed Mazouzi, Mohamed Mekhloufi, Mohamed Moghladj, Mohamed Nafti, Mohamed Oumari, Mohamed Rabai, Mohamed Riadh Ouchene, Mohamed Saiah, Mohamed Samir Sayoud, Mohammed Kébiche, Mokhtar Bakhti, Mokhtar Boukhallala, Moktar Badji, Moncef Boukalba, Morade Goumiri, Mostepha Belarbi, Mouloud Hassani, Mounia Bazhroun, Mounir Aouissi, Mounir Miklati, Mourad Amalou, Mourad Zerrouki, Moussa

Houhamdi, Mustapha Azibi, Mustapha Lamari, Mustapha Zahar, Nabila Bouras, Nacer Bouhmime, Nacer Boukhmis, Nacer Boukhris, Nacer Bouzrara Zogar, Nacer Mansouria, Nacer Ouradi, Nacerdine Khroufi, Nacira Saibi, Nadhir Taleb, Nadia Azzaoui, Nadia Ramdane, Nadia Sidiouis née Dehl, Nadia Zian, Nadir Boussem, Nadjib Ben Ayad, Nadjib Messah, Nadjiba Bendjedda, Naima Bendali, Naima Bentata, Naima Chiebieb, Nasreddine El Maharrat, Nasreddine Kechida, Nassreddine Laazer, Nedjla Adamou, Nesrine Mohammadi, Nora Merabet, Norddine Malag, Nordine Kellil, Noureddine Khellil, Nourreddine Rahmani, Nourredine Boukteb, Omar Ould Amara, Othman Fardeheb, Ouarda Aouiche, Oussama Bey, P.N. Taza, Rabah Kaoudji, Rabah Megherbi, Rabah Rahal, Rachid Abidat, Rachid Ait Medjber, Rachid Bourayou, Rachid Hamdouche, Rachid Rouag, Rachid Rouagh, Rachid Rouane, Rachid Soualem, Radia Megrirouche, Rafik Rahmoun, Rafik Tabiti, Rahma Ben Bouriche, Raouf Guechi, Razek Bougara, Redha El Ferroudji, Rezki Bagra, Riadh Moulai, Rokia Khedar, Saad Beria, Saad Bouakaz, Saad Naamous, Sabah Boukhobza, Saber Benkaddour, Sabrina Chebieb, Sabrina Choubane, Saci Haddad, Saci Oulmi, Said Chabouni, Said Chegar, Said Fritas, Said Khataoui, Said Noureddine, Salah BenHamdoun, Salah Feradci, Salah Hennouche, Salah Ouarem, Salem Abbassi, Saliha Hamlet, Salim Mihoub, Salim Ourlis, Salma Chetara, Samia Maza, Samiha Bentrad, Samir Chaouchar, Samir Khalessnan, sara Benkacimi, Sara Boudraa, Sarah Belkacemi, Sarah Menaa, Sarah Mokadem, Sebti Derrar, Seddik Garah, Sedik Boubekeur, Sekiou Saleh eddine, Sellam Ben Aissa, Selloua Hassad, Selma Hamadou, Seti Ahcen, Sihem Bakour, Slimane Derbal, Slimane Naas, Sonia Chaibdour, Souad Lourdjane, Tahtah Khikhi Oum El Khir, Taous Blibek, Taous Torch, Tassaâdit Bacha, Toufik Rebouh, Toufik Youcefi, Walid Dahmani, Wassila Lilia Bedouhene, Yacine Boulenouar, Yacine Bourahla, Yacine Essalami, Yacine Khalfallah, Yacine Khettab, Yacine Soltane, Yasmina Djetli, Yasmine Sahbi, Yassad Farah, Yazid Neaidji, Yesaad Mohamed Chareb, Youcef Aifa, Youcef Djebbour, Youcef Hassaine, Youcef Kahoul, Youcef Meribai, Youcef Saradj, Zahir Benallaoua, Zahra Brahmia, Zineb Seboiai, Zohra Belkaoussa, Zohra Hayat Remmas, Zouzou El Hadi, Houari Djardini, Aba Ramzi, Mohamed Chami, Djamel Hadj Aissa, Forestiers des 48 conservations des forêts

**Egypt:** Abd el Halim El Sayed, Ahmed Ebaid, Ahmed Ibrahim, Amr Abd Elhady, Anne-Laure Brochet, Carol Fouque, Dik Hoek, Habib Dlensi, Haitham Ibrahim, Hellin de Wavrin, Hosni Helmy Asran, Jean-Yves Mondain-Monval, Khaled Noby, L. Ben Nakhla, Madg Sad, Maurice Benmergui, Mohamed Ezat, Mohamed Hammad Mohamed, Mohamed Zaki, Naoufel Hammouda, Osama El Gebaly, Pierre Defos du Rau, Wed Abdel Latif Ibrahim

**Libya:** Khaled Etayeb, Abdulmaula Hamza, Mohamed Essghaier, Michael Smart, Nicola Baccetti, Pierre Defos du Rau, Hichem Azafzaf, Habib Dlensi, Almokhtar Saied, Ibrahim Tabouni, Essam Bouras, Ali Berbash,, Marco Zenatello, Wageh Bashemam, Mohamed Bouzainen, , Jaber Yahia, Ashraf Galidan, Ibrahim Madi. Saleh Deryaq, Fuzi Dhan, Aboajela Mansouri, Amer Al-Jamel, Nader Azabi, Saied Azabi, Nader Ghreeba, Rabee Azabi, Shokri Dhan, Abdulati Swyieb, Abdulmonem Albaour.

**Morocco:** Abad D.J.Abi K., Aebischer A Aghnaj A., Aissami L Aït Ben Addi, Ait Salah A., Ait Salem A., Alami Mhamdi M., Alaoui B., Alaoui Mdaghri M., Alaoui M.Y., Alouache E., Amezian M., Amhend M., Aourir M., Arahou M., Arhzaf Z.L., Asdad M., Azaouaghe S., Azizi D., Baha M., Banan H., Baouab R.E., Bayed A., Baytar H., Bazairi H., Beaubrun P.C., Belghyti D., Bendikia H., Bendraou H., Benhamza A., Benhoumman B., Benhoussa A., Benlahcen M., Bennouna M., Bensusan K., Beudels M.O., Borderelle M.C., Bouaissi M., Bouajaja A., Bouariche B., Bouchefra A., Bouhdadi L., Bouhdadi M.L., Boumaaza M., Boumejou L., Bourass K., Bourouah B., Boussadik H., Bowden C., Cabo J., Carp E., Cerezo P., Chafi H., Chafik M., Chahlaoui A.Chakri S.Charaf M.A., Charco J., Chavanon G., Cherkaoui I., Chichi A., Chikri K., Chillasse L., Cole J., Cortes J., Cuzin F., Dakki M., Dalai M., De Bellefroid F., De La Cruz A., De La Perche N., El Agbani M.A., El Aouki A., El Bakkali M., El Banak A., El Bekkay M., El Brini H., El Ghazi A., El Hamoumi R., El Hamzaoui M., El Haoua M.K., El Hassani A., El Idrissi Essougrati A., El Khamlichi R., El Kharrim K., El Malki S., El Mghari A., El Oualidi J., Errati A., Espinar R.R., Essabbani A., Essalai K., Ettalibi A.., Exo K.M., Fahd S., Fahmi A., Faqyhi Y., Fekhaoui M., Fernandez J., Fettah M., Fozzi A., Franchimont J., Fraval C., Fraval E., Fuzzi A., Gambarotta C., Gargallo G., Garza-hernandez J.A., Graf O., Grangier F., Gretton A., Guerinech A., Guillem R., Gullick T., Haâbi L., Hajib S., Hajibi M., Hamidi S., Hammouradia H., Hamza L., Hamza O., Hamzaoui A., Hanane S., Hannane N., Harmas A., Hassani H., Heridia R., Himmi O., Hlal M., Houilat N., Houilat R., Ibariouen M., Ibn Tattou M., Issaouballah A., Jacob H., Jacob J., Jadid A., Jai Fayez A., Jansen J., Jaoudi F., Jaziri H., Jenifer J., Jensen R., Jerez Abbad D., Johnson A., Jouda I., Joulami L., Jubete F., Kacemi M., Kachiche H., Kamps M., Kiiss M., Knauss P., Ksassoua K., Laaouichi S.., Labidi D., Lafontaine D.A.J., Lafontaine R.M.J., Lahrouz S., Lahyane H., Lasfar S., Ledant J.P.H., Lemrabett E., Lemssiah J., Lieron V., Maamri A., Maghnouj M., Magin C., Mahé E., Marraha M., Mars N., Martin B., Mdarhri Alaoui E., Melhaoui M., Mesleard F., Messaoudi N., Meziane, Mirari S., Mokalik M., Mokhlis M., Monchatre T., Monier B., Naciri M., Nahhi H., Naya A., Neves R., Nicolle S., Nouiri H., Ochfy A., Onrubia A., Orueta J.F., Ouassou A., Oubrou W., Oukanou L., Pariselle A., Pascon J., Perennou C., Perez C., Perthuis A., Petit Y., Pilot M., Pina P., Pineau O., Pisu D., Pouteau C., Qninba A., Radi M., Rahhali I.E., Ramdani M., Ramirez R., Rguibi Idrissi H., Rhemali A., Rhonam A., Ribi M., Rihane A., Romanski T., Rougui A., Rousselon P.L.Rufino R., Ruth G., Sahri N., Salathe T., Sara S., Sayad A., Sehhar E., Slimani T., Smit C.J., Taib N., Tejjeni Y., Terrouzi

E., Thompson I., Tilly B., Torralvo C., Touati Malih F.Touzri M., Vangeluwe D.P.H.,Vollum A. Wahnon M., Warham P., Warr S., Werner S., Willefert S., Yahyaoui A., Yome A., Zadan Y., Zaïr I., Zaouia Y., Zerdeb A., Znari M., Zwarts L.

**Tunisia :** Ahmed Kilani, Ahmed Slimane, Claudia Feltrup-Azafzaf, Mohamed Dhahak, Mohamed Sadok Chafra, Naoufel Hammouda, Paul Mahoney, Sami Harbaoui, Sami Rebah, Slaheddine Bessadok, Sofien Turki, Ahmed Slimane, Amir Hakim, Badreddine Jemaa, , Habib Dlensi, , Jamel Tahri, Jean-Yves Mondain-Monval, Khaled Guettari, Marco Zenatello, Michael Smart, Moez Touihri, Mohamed Ali Dahkli, Mohamed Ayed, Mourad Amari, Naoufel Hammouda, Nejib Mokhtar, Pigniczki Csaba, Pierre Defos du Rau, Sami Rebah, Sofien Turki, Thomas Blanchon, Yves Kayser, Nicola Baccetti, Hedi Aissa, Samar Kilani, Mougib Gabous, Hamed Mallat, Ahmed Zadem, Sahbi Dorei, Anis Ben Brahim, Olivier Pineau, Thomas Galewski, Amir Hakim, Abdelnacer Ghlis, Gordon Allison, Malek Grairi, Mabrouk Raggad, Sofiane Mnara, Laura Dami, Zied Gtari, Dhafar Ben Othmen, Omor El Agrbi, Wissem Ben Zakour, Yasmine Azafzaf, Rahma Ben Hmida, Khaled Chaker, Hella Guidara, Nabiha Ben Mbarek, Taieb Housaini, Saba Guellouz, Samia Boufares, Jamel Jrijer, Jamel Zayati, Mariem Bsibes, Hedi Bel Haj Brahim, Ali Zneidi, Hatem Ben Belghasem. Yoldes Kesraoui, Habib Ghazouani, Rachid Haggui, Abdessatar Belkhouja, Ala Maki, Safouen Touati

**Appendix S3**

Spatial and temporal covariates used as predictors in LORI

| | Name | Description | Source | Unit | Scale |
|---|---|---|---|---|---|
| Spatial covariates | latitude | Latitude coordinates of the centroids of the surveyed sites[1] | MedWaterbirds database | Decimal degrees | Site |
| | country | Four covariates: Algeria, Egypt, Libya and Morocco. If all of these have a 0 value for a site, it means this site is in Tunisia. | MedWaterbirds database | Binary value (0 or 1) | Country |
| | alt | Mean of the altitude values located in the polygon of each site | SRTM 1 Arc-Second Global (https://www.usgs.gov/) corrected by MWO[2] with ALOS Global Digital Surface Model "ALOS World 3D - 30m" (AW3D30) | Metres (decimal value) | Site |
| | dist_town | Distance between the site and the nearest city | OpenStreetMap data | Metres (decimal value) | Site |
| | dist_coast | Distance between the site and the coastline | Natural Earth data | Metres (decimal value) | Site |
| | area | Maximum water extent, i.e. surface area of each site ever detected as water between 1984 and 2018 | Global Surface Water (Pekel et al. 2016)[3] | Square kilometres (decimal value) | Site |
| | ecosys_1, 2 and 3 | Raster of the macrogroups of Global Ecological Land Units (ELUs) describing ecosystem units. The macrogroups are the finest vegetation units classified by USGS, regarded as meso-scale (100s to 10,000s of hectares) ecosystems. | USGS Global Ecosystems https://rmgsc.cr.usgs.gov/ecosystems/dataviewer.shtml | 1 to 3 (integer) | Site (median value of the raster pixels located under each site) |
| | dam | Information about whether or not the site is a dam | FAO dam database corrected and completed by national IWC coordinators | Binary value (1 or 0) | Site |

[1] Longitude was not taken into account as this information is provided by the covariate "country".

[2] Mediterranean Wetlands Observatory (Tour du Valat)

[3] Pekel, J. F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. Nature, 540(7633), 418.

| | | | | | |
|---|---|---|---|---|---|
| Temporal covariates | Tspring_NW | Departure from a reference value or long-term temperature average (°C) in northwest Europe, in the spring (April to July) of year n-1[4] | Climate Explorer of the Royal Netherlands Meteorological Institute (KNMI)[5] HadCRUT4.6 SST/T2m anom field | Celsius degrees (decimal values) | European |
| | TspringNE | Same as above but in northeast Europe | KNMI, HadCRUT4.6 SST/T2m anom field explorer | Celsius degrees (decimal values) | European |
| | Twint_SW | Departure from a reference value or long-term temperature (°C) average in northwest Europe, in the winter (August year n-1 to January year n), per year n | KNMI, HadCRUT4.6 SST/T2m anom field | Celsius degrees (decimal values) | European |
| | Twint_SE | Same as above but in southeast Europe | KNMI, HadCRUT4.6 SST/T2m anom field | Celsius degrees (decimal values) | European |
| | p_NW | Average precipitation in northwest Europe in the spring, year n-1 | KNMI, CRU TS4.03 precipitation field (Harris et al., 2020[6]) | Millimetres (decimal value) | European |
| | p_NE | Average precipitation in northeast Europe in the spring, year n-1 | KNMI, CRU TS4.03 precipitation field (Harris et al., 2020) | Millimetres (decimal value) | European |
| | NAO | Irregular fluctuation of atmospheric pressure over the North Atlantic ocean that has a strong effect on winter weather in Europe | National Climatic Data Center (https://www.ncdc.noaa.gov/cdo-web/) | Index (decimal value) | North Atlantic |

[4] For example, April to July 1989 for a survey in winter 1990 that takes place from December 1989 to January 1990.
[5] http://climexp.knmi.nl
[6] Harris, I., Osborn, T.J., Jones, P. et al. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Sci Data 7, 109 (2020). https://doi.org/10.1038/s41597-020-0453-3

| | | | | | |
|---|---|---|---|---|---|
| Spatio-temporal covariates | agriculture | Percentage of farmland per year and per country. This includes croplands used for perennial crops or permanent pasture areas, or temporal or set-aside lands. | The World Bank https://data.worldbank.org/indicator/AG.LND.AGRI.ZS Compiled data from the FAO | Percentage (decimal value) | Country |
| | economy | GDP growth (Gross Domestic Product) per country and per year. The GDP growth rate is the most important indicator of economic performance. When the economy is expanding, the GDP growth rate is positive. If it is growing, so will businesses, jobs and personal income. If the GDP growth rate turns negative, then the country's economy is in https://www.thebalance.com/the-history-of-recessions-in-the-united-states-3306011. | The World Bank https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?page=6 Note: some of the values are missing (mainly for Libya, and for all the countries in 2017): the average of the values of all the countries and all the years is used to complete the missing values | Index (decimal value) | Country |
| | rain | Sum of winter precipitation (August year n-1 to February year n) per site and per year n (1990 to 2017)[7] | http://www.globalclimatemonitor.org/# Rainfall data used and compiled on this site are from the CRU TS 3.21 (Climatic Research Unit, Harris et al. 2014[8]) and the NOAA (National Oceanic and Atmospheric Administration) | Millimetres (decimal value) | Grid of 0.5 degrees in North Africa, values of the grid are applied to each site |

[7] If a site was surveyed in February during a particular year, the sum of precipitation was calculated from August (year n-1) to January (year n) for this site and this year. The same method was applied if the site was surveyed in December: we took into account precipitation from August to December.

[8] Harris, I. P. D. J.; Jones, P. D.; Osborn, T. J.; & Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset. International journal of climatology, 34(3), 623-642. doi: 10.1002/joc.3711

**Appendix S4**

The following a priori hypothesis governed our choice of covariates:

| HYPOTHESIS | REFERENCE |
|---|---|
| The field experience of a country's team of ornithologists, site accessibility, road infrastructure and/or the political situation and governance in the country could affect sampling and detection processes and, as a result, the abundance estimates on a national level. | 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. |
| Gross Domestic Product growth rate per country and per year (i.e. performance of the economy) was used as a proxy for governance, which has been shown to be a major driver of waterbird conservation. | 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. |
| Wetland surface area is a primary predictor of waterbird abundance, a proxy of which was provided by the maximum water extent extracted from the Global Surface Water dataset. | De Goeij, P.J., Van der Have, T.M., Keijl, G.O., Van Roomen, M.W.J., Rutters, P.S. (1992). The network of wetlands for waterbird migration in the eastern Mediterranean, in Finlayson, C.M., Hollis, T., Davis, T. (Eds.), Managing Mediterranean Wetlands and Their Birds, Proceedings of an IWRB International Symposium, Grado, Italy. IWRB special publication 20: 70-72.<br><br>Pekel, J. F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. Nature, 540(7633), 418. |
| Site-specific flooding is a primary predictor of habitat availability and hence abundance of waterbirds. The sum of autumn/winter precipitation per site and per year (rainfall) was used as a proxy of yearly wetland flooding. | Kingsford, R. T., Curtin, A. L., & Porter, J. (1999). Water flows on Cooper Creek in arid Australia determine 'boom'and 'bust'periods for waterbirds. Biological Conservation, 88(2), 231-248. |
| Agriculture is one of the major drivers impacting bird communities at a large scale: for example, through habitat loss and reclamation. Percentage of farmland per year and per country was used to index this impact. | Gaston, K. J., Blackburn, T. M., & Goldewijk, K. K. (2003). Habitat conversion and global avian biodiversity loss. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270(1521), 1293-1300. |

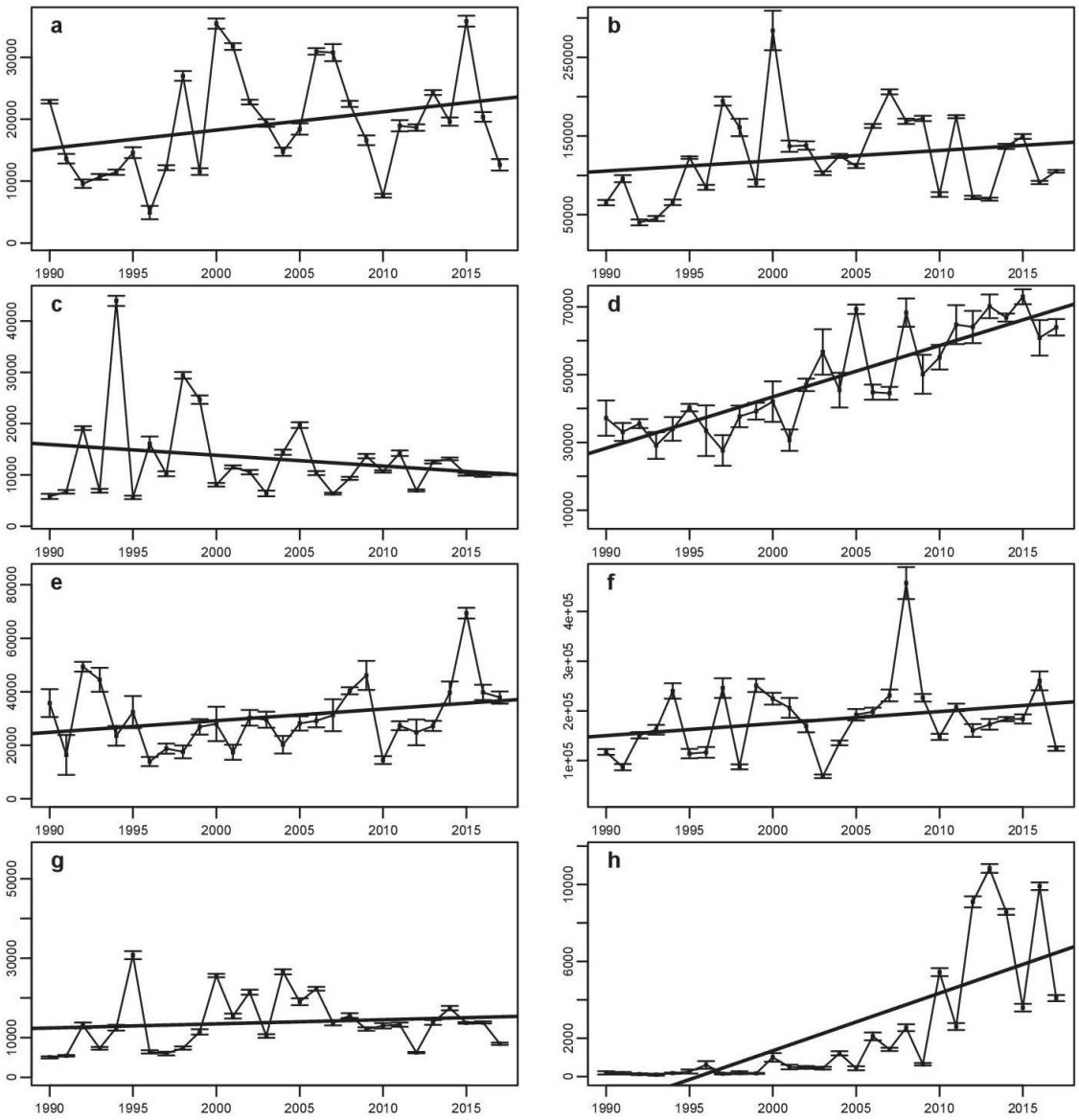
| | |
|---|---|
| | Teyssèdre, A., & Couvet, D. (2007). Expected impact of agriculture expansion on the world avifauna. Comptes Rendus Biologies, 330(3), 247-254.<br><br>Vickery, J. A., Ewing, S. R., Smith, K. W., Pain, D. J., Bairlein, F., Škorpilová, J., & Gregory, R. D. (2014). The decline of Afro-Palaearctic migrants and an assessment of potential causes. Ibis, 156(1), 1-22. |
| Altitude and distance from the coast covariates were used to compare low-altitude coastal lagoons and inland wetlands (e.g. chotts, sabkhas or reservoirs) and mountainous areas or plateaus. The combination of low altitude and proximity to the coast was used as a proxy for the threatened ecosystems of Mediterranean lagoons, a preferred habitat for several wading species. | Ayache, F., Thompson, J.R., Flower, R.J., Boujarra, A., Rouatbi, F., Makina, H. (2009). Environmental characteristics, landscape history and pressures on three coastal lagoons in the Southern Mediterranean Region: Merja Zerga (Morocco), Ghar El Melh (Tunisia) and Lake Manzala (Egypt). Hydrobiologia 622, 15-43.<br><br>Hüttich, C., Reschke, J., Keil, M., Dech, S., Weise, K., Beltrame, C., Fitoka, E., Paganini, M. (2012). Using the Landsat Archive for the Monitoring of Mediterranean Coastal Wetlands: Examples from the Glob-Wetland-II Project. http://www.earth-zine.org/2011/12/20/using-the-landsat-archive-for-the-monitoring-of-mediterranean-coastal-wetlands-examples-from-the-globwetland-ii-project/ (accessed 13.09.15).<br><br>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 Bull., 96, 47-59.<br><br>Qninba A., Dakki M., El Agbani M.A. & Benhoussa A. (2007). Rôle de la côte atlantique marocaine dans l'hivernage des Limicoles (Aves, Charadrii). Ostrich, 78, 2, 489-493. |
| Distance from the nearest urban agglomeration was used as a proxy both for wetland monitoring accessibility and potential impact of disturbance and/or pollution. Hence, this covariate could potentially negatively or positively impact the detection | Zhang, Y., Fox, A. D., Cao, L., Jia, Q., Lu, C., Prins, H. H., & de Boer, W. F. (2019). Effects of ecological and anthropogenic factors on waterbird abundance at a Ramsar Site in the Yangtze River |

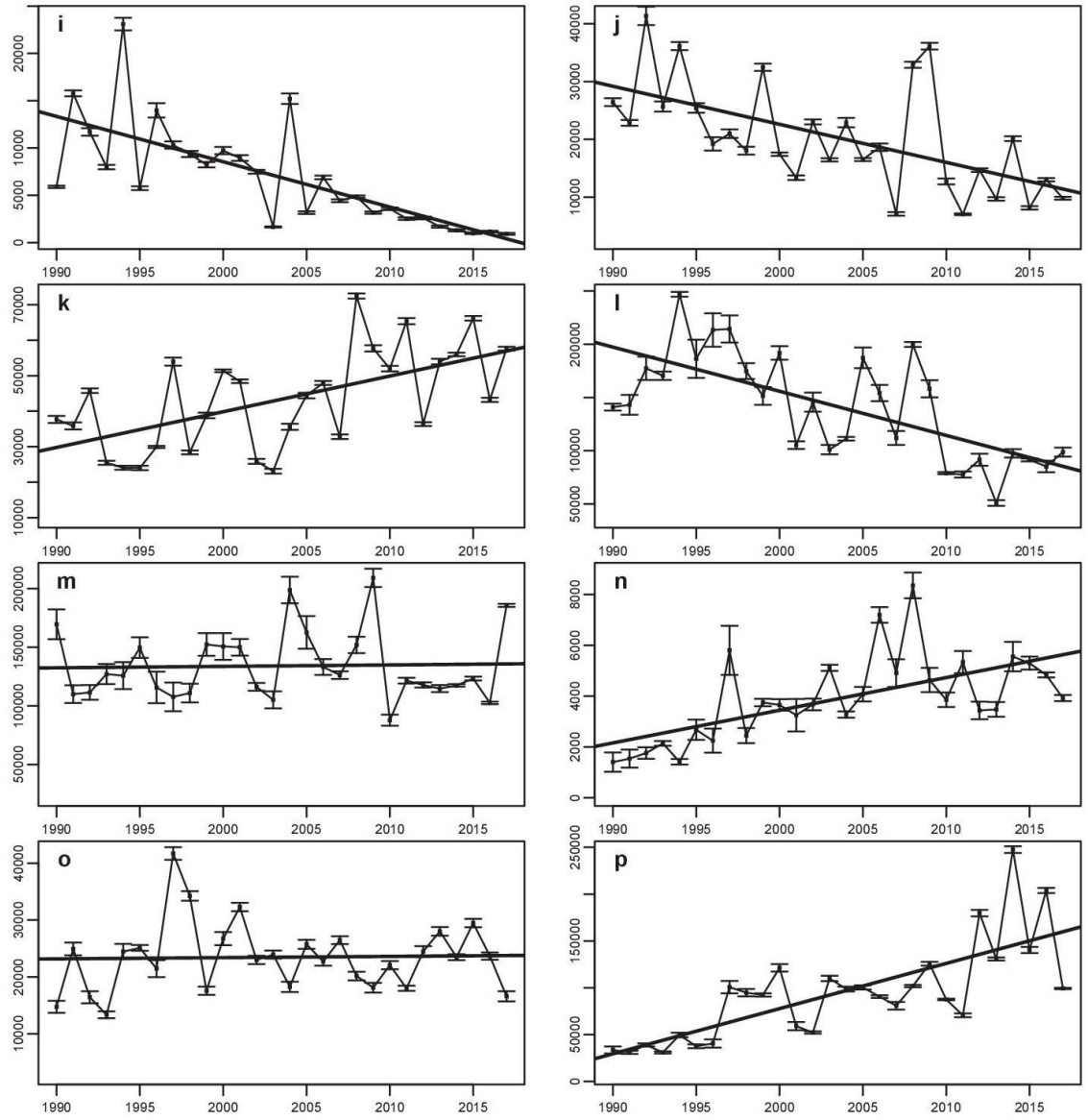
| | |
|---|---|
| of waterbirds. | Floodplain. Ambio, 48(3), 293-303. |
| As for any living organism, habitat is a primary influence on abundance and was thus indexed under 3 ecosystem macrogroups ('Temperate & Boreal Forest', 'Shrubland & Grassland', 'Desert & Semi-Desert'). | Guadagnin, D. L., & Maltchik, L. (2006). Habitat and landscape factors associated with neotropical waterbird occurrence and richness in wetland fragments. In Vertebrate Conservation and Biodiversity (pp. 405-418). Springer, Dordrecht. |
| Dams can positively or negatively affect waterbird abundance through habitat modification; some dams were created in the course of the time-series of counts we used, with a likely ecological impact on the waterbird community. | 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, Game Wildl., 13, 233-249.<br><br>Bergkamp, G., McCartney, M., Dugan, P., McNeely, J., & Acreman, M. (2000). Dams, ecosystem functions and environmental restoration. Thematic review II, 1, 1-187.<br><br>Dakki M., El Agbani M.A. & Qninba A. (Eds), 2011. Zones humides du Maroc inscrites jusqu'en 2005 sur la Liste de la Convention de Ramsar. Trav. Inst. Sci., Rabat, Sér. Générale, 7, 1-238. |
| Spring temperature and precipitation anomalies in breeding areas can affect reproduction and thus the abundance of birds subsequently migrating from northern Europe to North Africa. | Forcey, G. M., Thogmartin, W. E., Linz, G. M., Bleier, W. J., & McKann, P. C. (2011). Land use and climate influences on waterbirds in the Prairie Potholes. Journal of Biogeography, 38(9), 1694-1707.<br><br>Pavón-Jordán, D., Santangeli, A. & Lehikoinen, A. (2017). Effects of flyway-wide weather conditions and breeding habitat on the breeding abundance of migratory boreal waterbirds. Journal of avian biology, 48(7), 988-996. |
| Winter temperature anomalies in wintering areas of southern Europe, especially lower-than-average temperatures, could influence migration numbers to North Africa, notably in the case of cold spells. | Pavón-Jordán, D., Azafzaf, H., Balaž, M., Bino, T., Borg, J. J., Božič, L., ... & Devos, K. (2020). Positive impacts of important bird and biodiversity areas on wintering waterbirds under changing temperatures throughout Europe and North Africa. Biological Conservation, 246, 108549. |
| The North Atlantic Oscillation (NAO) was used as a synthetic proxy for the yearly weather conditions affecting waterbirds | Pavón-Jordán, D., Clausen, P., Dagys, M., Devos, K., Encarnaçao, V., Fox, A. D., ... & Langendoen, T. (2019). Habitat-and species-mediated |

| | |
|---|---|
| in their wintering range and annual displacements between Europe and North Africa. | short-and long-term distributional changes in waterbird abundance linked to variation in European winter weather. Diversity and Distributions, 25(2), 225-239. |

**Appendix S5**

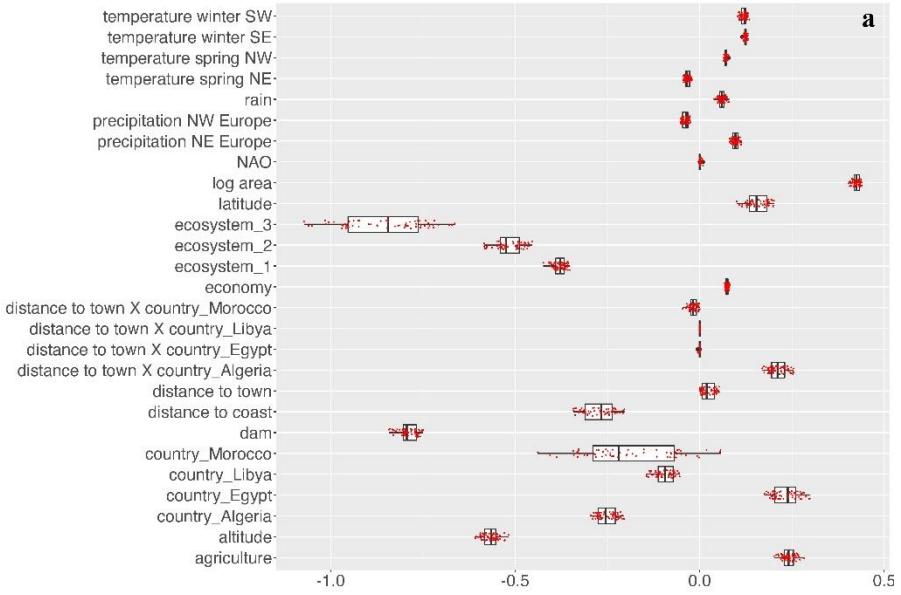
Yearly count totals as estimated by LORI and trends over all North African sites for the Pied Avocet (a), Dunlin (b), Gadwall (c), Mallard (d), Great Cormorant (e), Common Coot (f), Common Crane (g), Glossy Ibis (h), Greylag Goose (i), Northern Pintail (j), Common Teal (k), Wigeon (l), Northern Shoveler (m), Eurasian Spoonbill (n), Ringed Plover (o) and Greater Flamingo (p) as modelled by LORI with the respective linear time trend.

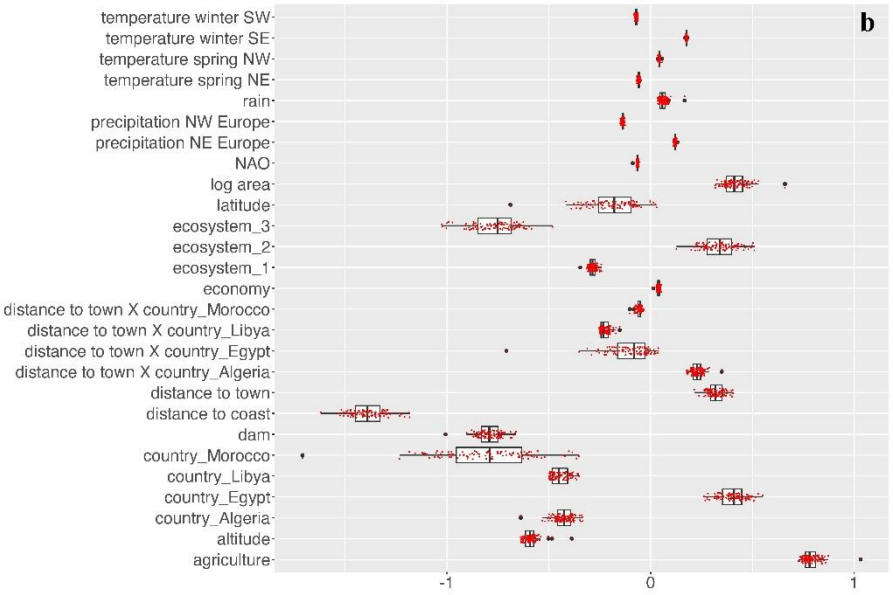


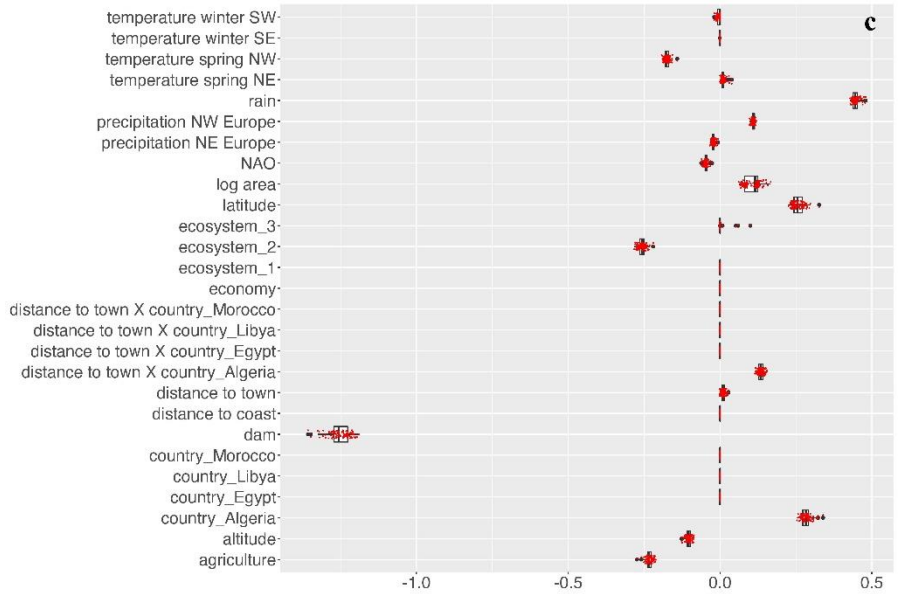

i
j
k
l
m
n
o
p
1990
1995
2000
2005
2010
2015

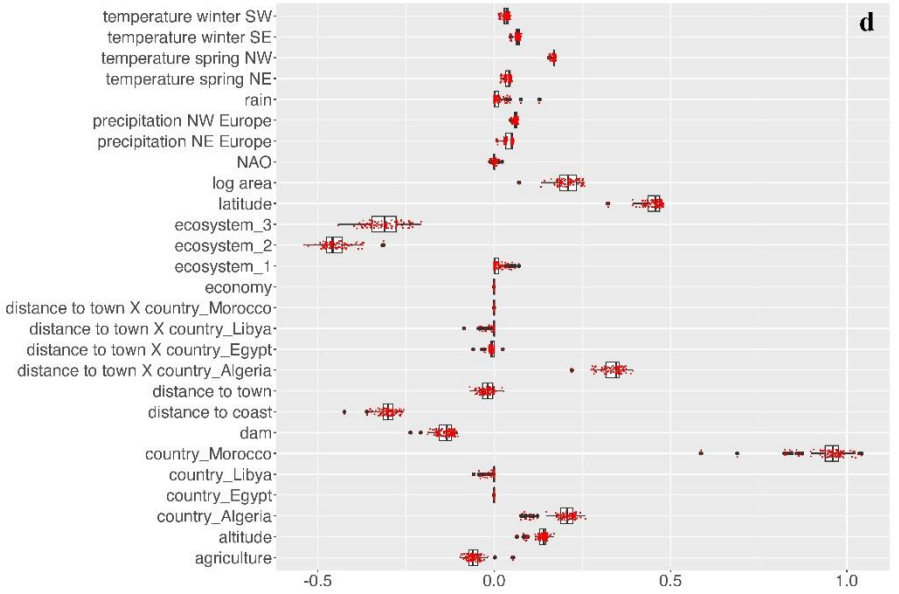
**Appendix S6**

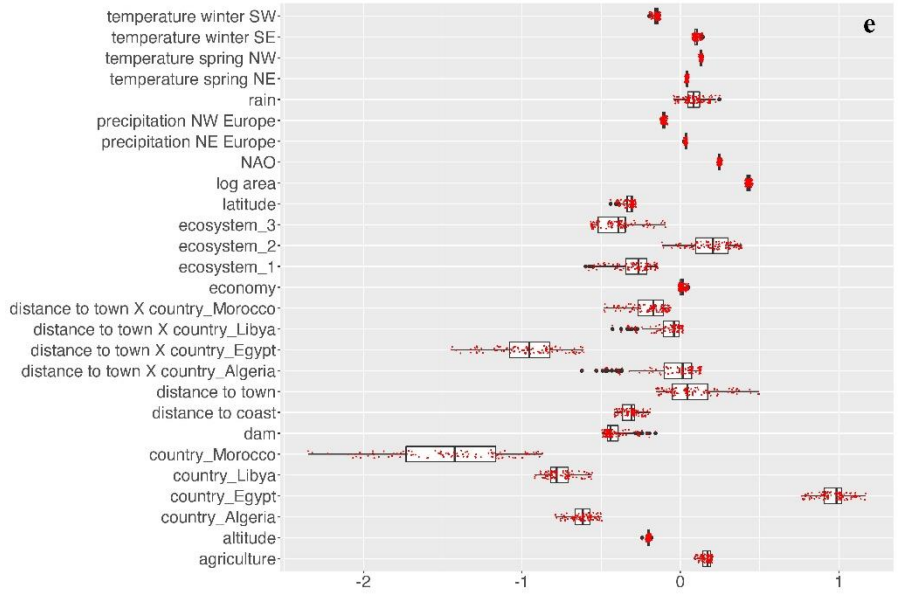
Effect parameters for all spatial and temporal covariates as estimated by LORI for the: Pied Avocet (a), Dunlin (b), Gadwall (c), Mallard (d), Great Cormorant (e), Common Coot (f), Common Crane (g), Glossy Ibis (h), Greylag Goose (i), Northern Pintail (j), Common Teal (k), Wigeon (l), Northern Shoveler (m), Eurasian Spoonbill (n), Ringed Plover (o) and Greater Flamingo (p).

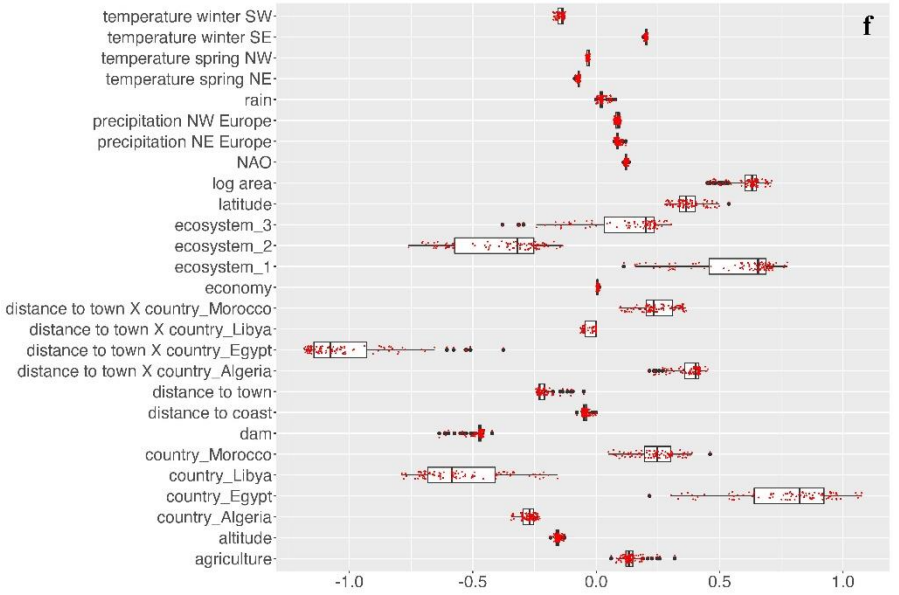


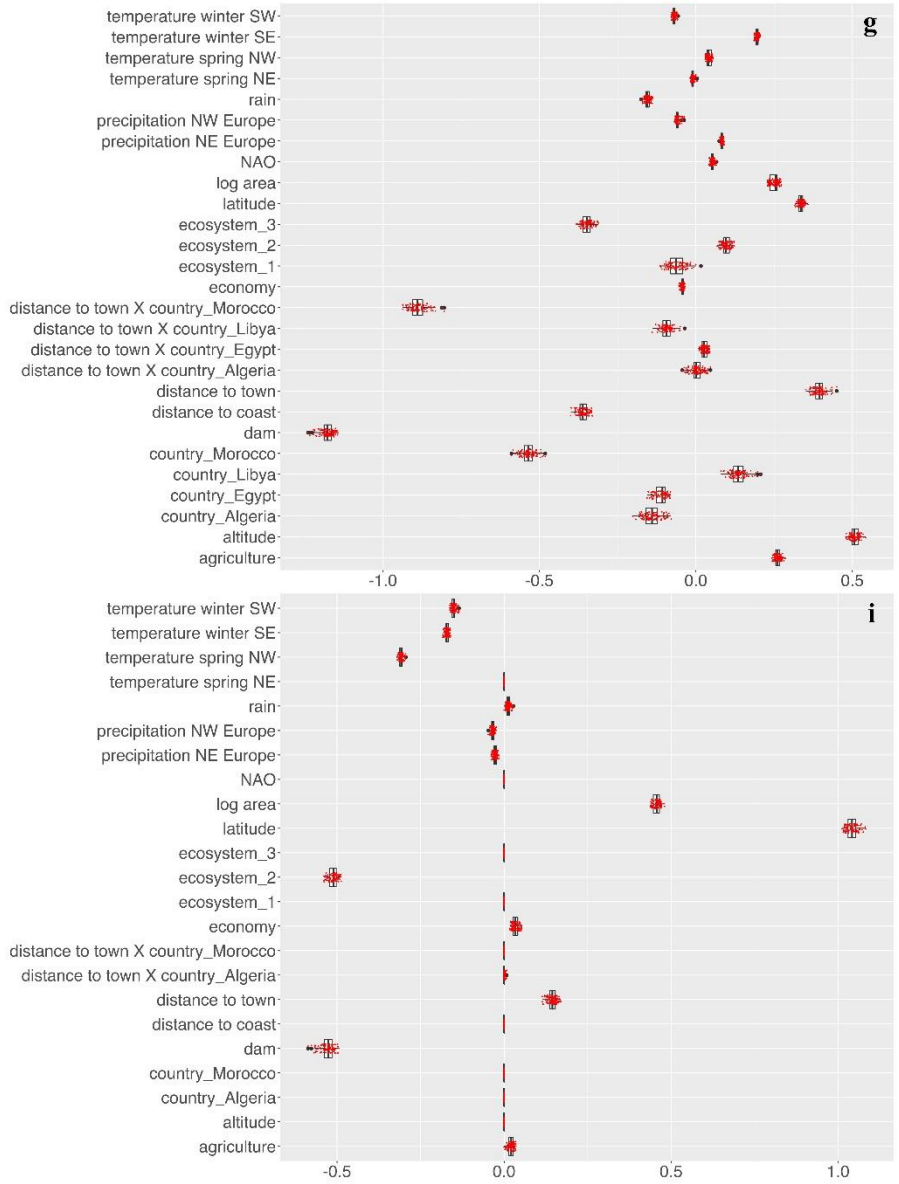



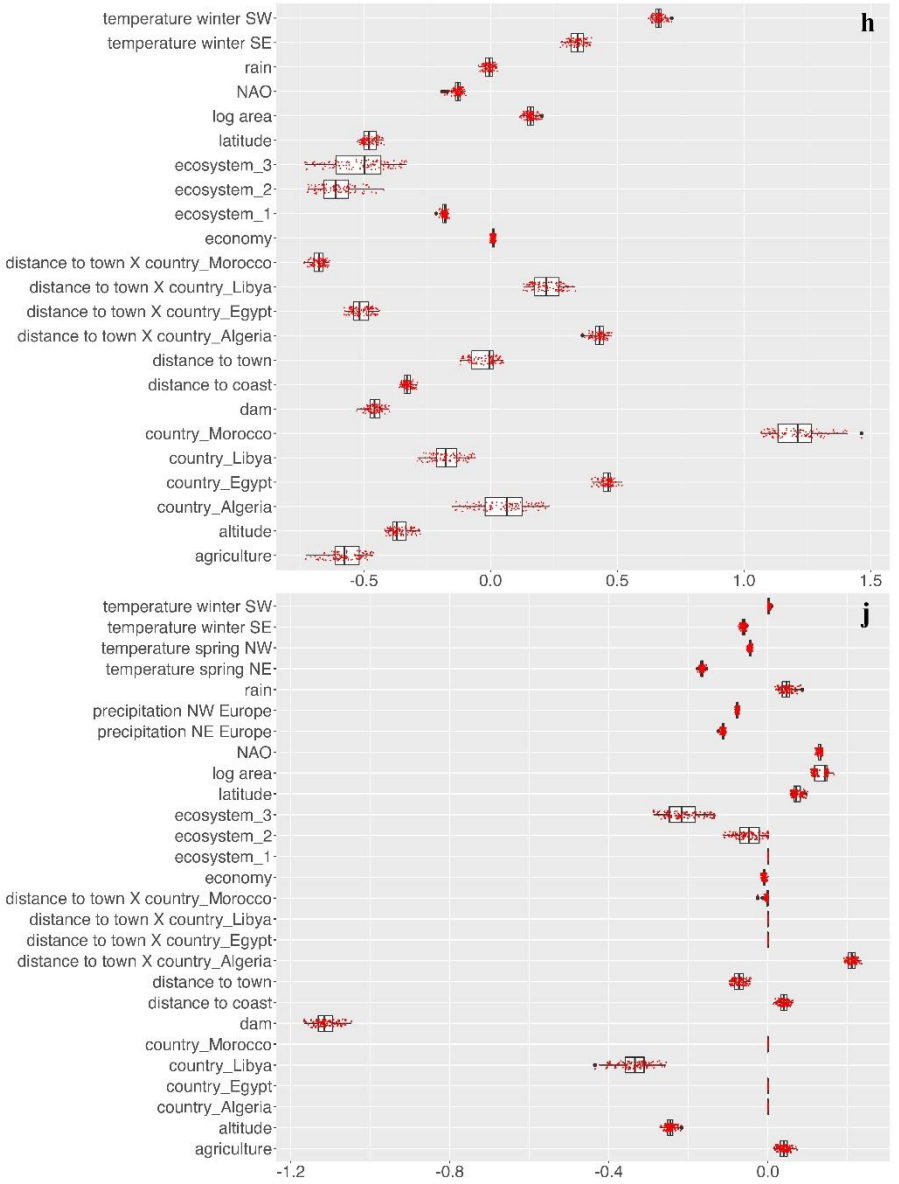

c
temperature winter SW
temperature winter SE
temperature spring NW
temperature spring NE
rain
precipitation NW Europe
precipitation NE Europe
NAO
log area
latitude
ecosystem_3
ecosystem_2
ecosystem_1
economy
distance to town X country_Morocco
distance to town X country_Libya
distance to town X country_Egypt
distance to town X country_Algeria
distance to town
distance to coast
dam
country_Morocco
country_Libya
country_Egypt
country_Algeria
altitude
agriculture
-1.0
-0.5
0.0
0.5


d
temperature winter SW
temperature winter SE
temperature spring NW
temperature spring NE
rain
precipitation NW Europe
precipitation NE Europe
NAO
log area
latitude
ecosystem_3
ecosystem_2
ecosystem_1
economy
distance to town X country_Morocco
distance to town X country_Libya
distance to town X country_Egypt
distance to town X country_Algeria
distance to town
distance to coast
dam
country_Morocco
country_Libya
country_Egypt
country_Algeria
altitude
agriculture
-0.5
0.0
0.5
1.0


e
temperature winter SW
temperature winter SE
temperature spring NW
temperature spring NE
rain
precipitation NW Europe
precipitation NE Europe
NAO
log area
latitude
ecosystem_3
ecosystem_2
ecosystem_1
economy
distance to town X country_Morocco
distance to town X country_Libya
distance to town X country_Egypt
distance to town X country_Algeria
distance to town
distance to coast
dam
country_Morocco
country_Libya
country_Egypt
country_Algeria
altitude
agriculture
-2
-1
0
1


f
temperature winter SW
temperature winter SE
temperature spring NW
temperature spring NE
rain
precipitation NW Europe
precipitation NE Europe
NAO
log area
latitude
ecosystem_3
ecosystem_2
ecosystem_1
economy
distance to town X country_Morocco
distance to town X country_Libya
distance to town X country_Egypt
distance to town X country_Algeria
distance to town
distance to coast
dam
country_Morocco
country_Libya
country_Egypt
country_Algeria
altitude
agriculture
-1.0
-0.5
0.0
0.5
1.0

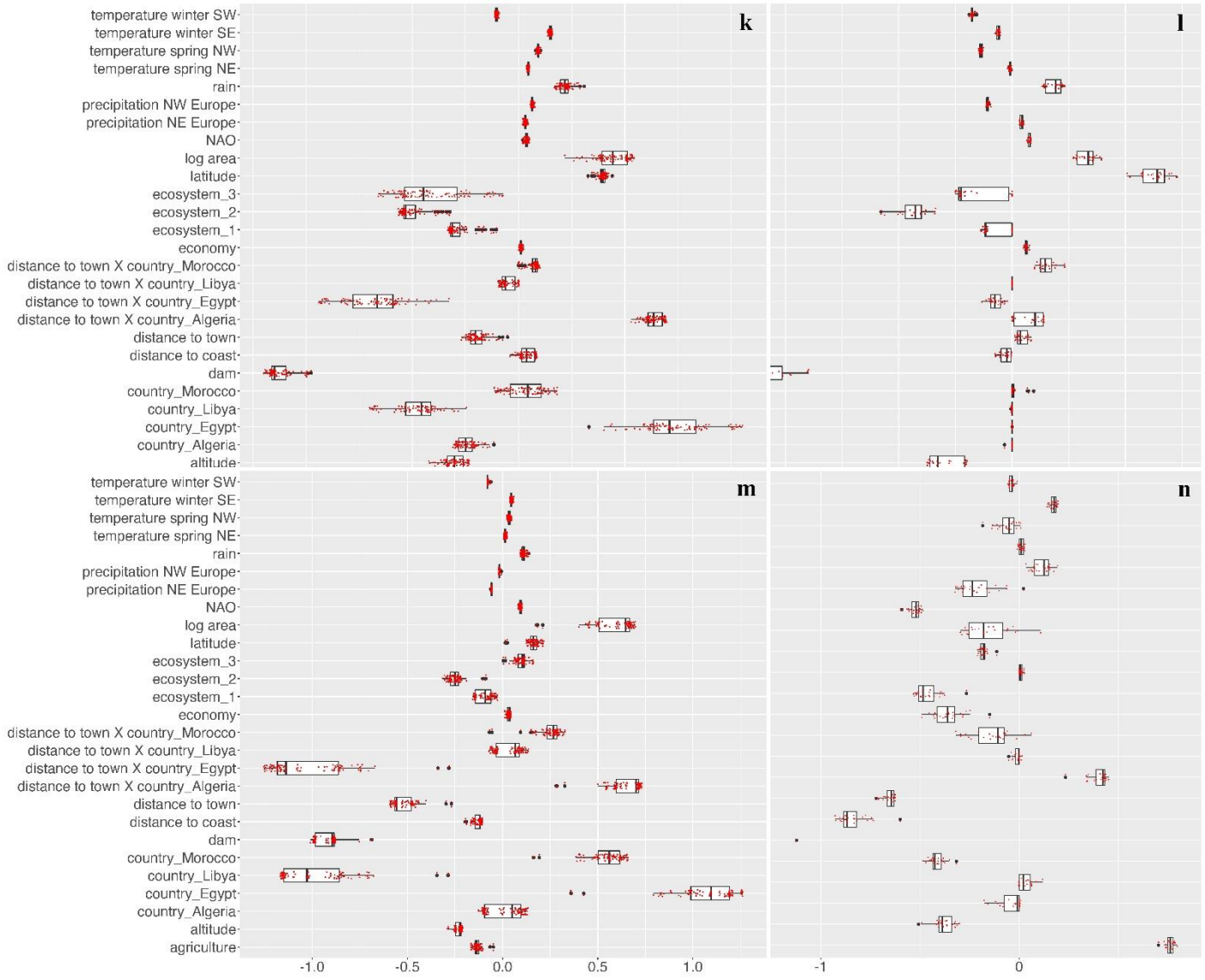

g
temperature winter SW
temperature winter SE
temperature spring NW
temperature spring NE
rain
precipitation NW Europe
precipitation NE Europe
NAO
log area
latitude
ecosystem_3
ecosystem_2
ecosystem_1
economy
distance to town X country_Morocco
distance to town X country_Libya
distance to town X country_Egypt
distance to town X country_Algeria
distance to town
distance to coast
dam
country_Morocco
country_Libya
country_Egypt
country_Algeria
altitude
agriculture
-1.0
-0.5
0.0
0.5

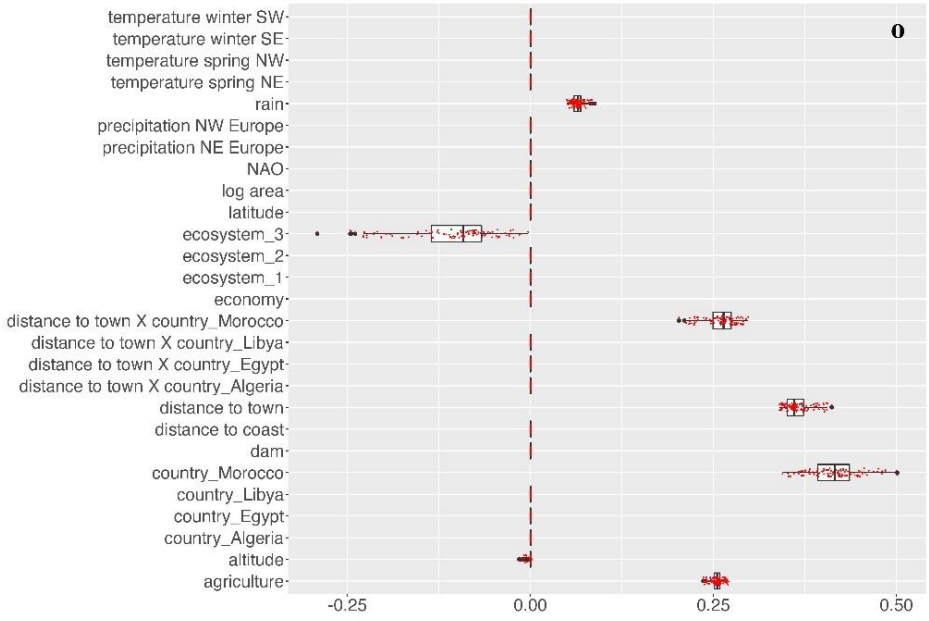

h
temperature winter SW
temperature winter SE
rain
NAO
log area
latitude
ecosystem_3
ecosystem_2
ecosystem_1
economy
distance to town X country_Morocco
distance to town X country_Libya
distance to town X country_Egypt
distance to town X country_Algeria
distance to town
distance to coast
dam
country_Morocco
country_Libya
country_Egypt
country_Algeria
altitude
agriculture
-0.5
0.0
0.5
1.0
1.5


i
temperature winter SW
temperature winter SE
temperature spring NW
temperature spring NE
rain
precipitation NW Europe
precipitation NE Europe
NAO
log area
latitude
ecosystem_3
ecosystem_2
ecosystem_1
economy
distance to town X country_Morocco
distance to town X country_Algeria
distance to town
distance to coast
dam
country_Morocco
country_Algeria
altitude
agriculture
-0.5
0.0
0.5
1.0

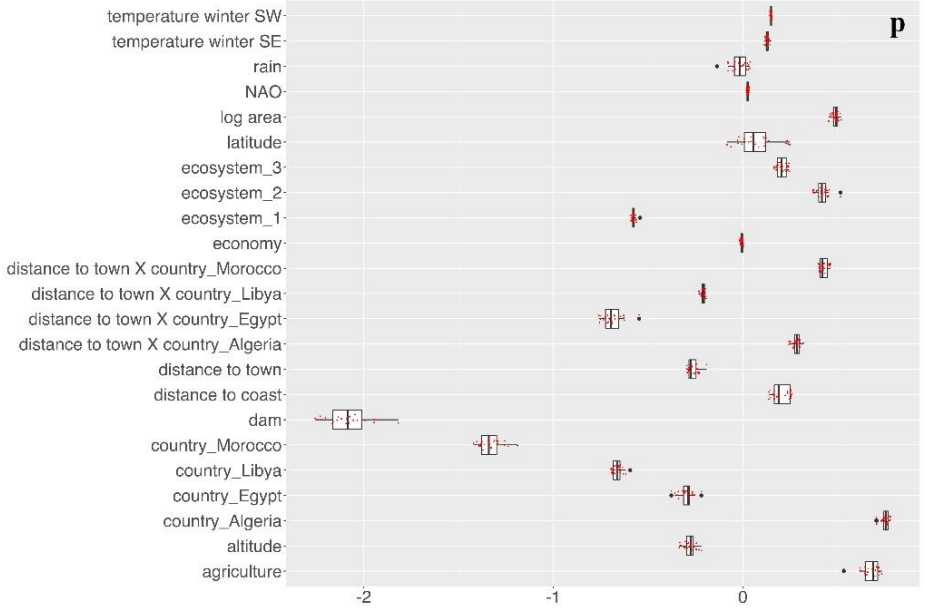

j
temperature winter SW
temperature winter SE
temperature spring NW
temperature spring NE
rain
precipitation NW Europe
precipitation NE Europe
NAO
log area
latitude
ecosystem_3
ecosystem_2
ecosystem_1
economy
distance to town X country_Morocco
distance to town X country_Libya
distance to town X country_Egypt
distance to town X country_Algeria
distance to town
distance to coast
dam
country_Morocco
country_Libya
country_Egypt
country_Algeria
altitude
agriculture
-1.2
-0.8
-0.4
0.0

1


k
l
m
n
temperature winter SW
temperature winter SE
temperature spring NW
temperature spring NE
rain
precipitation NW Europe
precipitation NE Europe
NAO
log area
latitude
ecosystem_3
ecosystem_2
ecosystem_1
economy
distance to town X country_Morocco
distance to town X country_Libya
distance to town X country_Egypt
distance to town X country_Algeria
distance to town
distance to coast
dam
country_Morocco
country_Libya
country_Egypt
country_Algeria
altitude
agriculture
-1.0
-0.5
0.0
0.5
1.0
-1
0


o
temperature winter SW
temperature winter SE
temperature spring NW
temperature spring NE
rain
precipitation NW Europe
precipitation NE Europe
NAO
log area
latitude
ecosystem_3
ecosystem_2
ecosystem_1
economy
distance to town X country_Morocco
distance to town X country_Libya
distance to town X country_Egypt
distance to town X country_Algeria
distance to town
distance to coast
dam
country_Morocco
country_Libya
country_Egypt
country_Algeria
altitude
agriculture
-0.25
0.00
0.25
0.50

2


p
temperature winter SW
temperature winter SE
rain
NAO
log area
latitude
ecosystem_3
ecosystem_2
ecosystem_1
economy
distance to town X country_Morocco
distance to town X country_Libya
distance to town X country_Egypt
distance to town X country_Algeria
distance to town
distance to coast
dam
country_Morocco
country_Libya
country_Egypt
country_Algeria
altitude
agriculture
-2
-1
0

3
4