Imputation multiple de comptages d'oiseaux d'eau à l'aide de covariables prédictives

Geneviève Robin[1,2*] & Mohamed Dakki [3] & Hichem Azafzaf [4] & Khaled Etayeb[5] & Samir M. Sayoud[6] & Nadjiba Bendjedda[7] & Wed A.I. Abdou[8] & Pierre Defos du Rau[9] & Marie Suet[10] & Jean-Yves Mondain-Monval[9] & Clémence Deschamps[10] & Elie Gaget[10,11] & Laura Dami[10]

[1] *Centre de Mathématiques Appliquées, École Polytechnique, Palaiseau, France*
[2] *Projet XPOP, INRIA Saclay, Palaiseau, France*
[3] *Institut Scientifique, Université Mohammed V de Rabat, Morocco*
[4] *Association "Les Amis des Oiseaux" (AAO / BirdLife Tunisie), Ariana, Tunisia*
[5] *Zoology Dept., Tripoli University, Tripoli, Libya*
[6] *Centre Cynégétique de Réghaia (C.C.R.), Alger,Algeria*
[7] *Direction Générale des Forêts (DGF), Algérie*
[8] *Egyptian Ministry of State For Environmental Affairs, Cairo, Egypt*
[9] *Office National de la Chasse et de la Faune Sauvage, Unité Avifaune Migratrice, Arles, France*
[10] *Institut de Recherche de la Tour du Valat, Arles, France*
[11] *Muséum National d'Histoire Naturelle, Centre d'Ecologie et des Sciences de la Conservation-CESCO, Paris, France*

Résumé. En écologie, et en particulier en suivi de population d'espèces sauvages, nous sommes amenés à analyser des tableaux de comptages contenant une grande proportion de données manquantes. Par ailleurs, des informations complémentaires, telles que des covariables décrivant les lignes et les colonnes du tableau sont souvent disponibles. Nous proposons une nouvelle méthode d'imputation multiple pour les données de comptage avec données manquantes MAR, qui incorpore des covariables complémentaires dans le processus d'inférence. Nous montrons empiriquement que la méthode surpasse les techniques classiquement utilisées. Nous appliquons la méthode à l'analyse de données d'abondance d'oiseaux d'eau d'Afrique du Nord.

Mots-clés. Données manquantes, imputation multiple, modèle log-linéaire, tableau de contingence.

Abstract. In ecology, and particularly in species monitoring, one encounters count tables with many missing values. Furthermore, side information, such as covariates describing the rows (environmental sites) and columns (times stamps, species, etc.) of the table are often available. In this study, we propose a new method for the multiple imputation of count data, which allows to incorporate such complementary covariates. We show empirically that the imputation method performs better than existing techniques, in particular when the percentage of missing values is large. We apply the method to the analysis of a North African waterbird abundance data set.

Keywords. Contingency table, log-linear model, missing data, multiple imputation.

*Corresponding author: genevieve.robin@polytechnique.edu

1 Introduction

Les oiseaux font partie des espèces les mieux étudiées, et le suivi de leurs populations sert à renseigner sur l'état de la biodiversité à l'échelle mondiale. Les populations d'oiseaux d'eau sont suivies depuis les années 1960, et sont dénombrées sur plus de 25 000 sites partout dans le monde (Amano et al. 2018). À partir de ces données d'abondances, sont calculées la tendance des population, et leur changement de distribution sous l'influence des facteurs anthropiques. Ces informations, et leur fiabilité, sont capitales pour orienter les stratégies politiques des conventions internationales pour la conservation de la biodiversité.

1.1 Données manquantes et covariables prédictives

À l'échelle régionale, l'estimation des tendances temporelles des nombres d'oiseaux est souvent mise en défaut par le manque de données. En effet, dans certaines région telles que l'Afrique du Nord, la couverture spatiale des comptages est variable pour des raisons politiques et financières : cela donne lieu a une grande proportion de données manquantes. À titre d'exemple, nous nous intéressons ici à un jeu de données rassemblant les comptages de 12 espèces dans 785 sites d'Afrique du Nord. Selon les espèces, le pourcentage de données manquantes varie entre 40% et 60%. Cela correspond à des régimes où les méthodes classiquement utilisées par les écologues, comme l'imputation par Analyse des Correspondences (Correspondance Analysis, CA), décrite dans Josse et Husson (2016), ou par Tendances et indices des données de surveillance (TRIM), décrite dans Pannekoek et van Strien (2001), sont instables, voire inutilisables. Dans certains cas, cette grande proportion de données manquantes peut être compensée grâce à des données complémentaires, telles que des covariables géographiques, météorologiques ou économiques concernant les sites et les années où les oiseaux ont été recensés. En effet, plusieurs études ont déjà démontré que de tels facteurs sont de bons prédicteurs des abondances d'oiseaux d'eau (Amano et al. (2018)). Cependant, les méthodes existantes d'imputation de données de comptage ne sont pas satisfaisantes à cet égard, puisqu'elles ne permettent pas (CA), ou seulement partiellement (TRIM), de modéliser l'effet de covariables.

1.2 Contributions

L'objectif de cet article est de développer une nouvelle méthode d'imputation multiple de données de comptages, qui tire profit des covariables explicatives à disposition. La méthode repose sur un modèle log-linéaire Poissonien incorporant des effets principaux des sites, des années et des covariables, ainsi que des interactions sites-années. Les paramètres du modèle sont estimés en minimisant la log-vraisemblance négative pénalisée par deux termes de régularisation; l'un encourageant des solutions parcimonieuses pour le vecteur des effets principaux, l'autre encourageant des solutions de rang faible pour la matrice d'interactions. D'un point de vue écologique, la méthode permet d'analyser des

données de recensement d'oiseaux d'eau d'Afrique du Nord collectées entre 1990 et 2017. En particulier, nous obtenons des estimations de l'évolution du nombre total (tous sites confondus) d'oiseaux au cours du temps.

2 Imputation multiple à l'aide d'un modèle log-linéaire Poissonien

2.1 Modèle log-linéaire

Notons $Y = (Y_{ij})$ le tableau de taille $(n \times p)$ contenant les abondances d'une espèce donnée, mesurées dans n sites au cours de p années. Pour chaque entrée Y_{ij}, nous définissons l'indicateur d'observation $\Omega_{ij} = 1$ si Y_{ij} est observé, et $\Omega_{ij} = 0$ sinon. Nous supposons le modèle Poissonien suivant:

$$Y_{ij} \sim \mathcal{P}(\mu_{ij}),\ 1 \leq i \leq n, \text{ and } 1 \leq j \leq p. \tag{1}$$

Pour chaque cellule (i, j), nous disposons d'un vecteur de covariables $X_{ij} \in \mathbb{R}^K$, dont nous notons $X_{ij}(k)$ le k-ième élément. Par exemple, X_{ij} peut contenir la surface du site i, des indices météorologiques concernant l'année j, ainsi que des indices économiques du pays où se trouve le site i, calculés pour l'année j. La matrice de taille $(np) \times K$ dont les lignes sont définies par les vecteurs (X_{ij}), est notée X. Nous supposons enfin que l'intensité Poissonienne définie dans (1) peut être décomposée par un modèle log-linéaire:

$$\log(\mu_{ij}) = \alpha_i + \beta_j + \sum_{k=1}^{K} \epsilon_k X_{ij}(k) + \Theta_{ij},\ 1 \leq i \leq n, \text{ et } 1 \leq j \leq p. \tag{2}$$

Dans le modèle (2), α_i (resp. β_j) correspond à l'effet du i-ième site (resp. de la j-ième année). Un effet positif $\alpha_i > 0$ (resp. $\beta_j > 0$) indique que, toutes choses égales par ailleurs, les comptes du i-ième site (resp. de la j-ième année) ont tendance à être plus élevés que ceux des autres sites (resp. des autres années). Pour $1 \leq k \leq K$, le coefficient ϵ_k indique l'effet de la k-ième covariable sur les comptes. Par exemple, supposons que k indique la covariable correspondant à la distance d'un site au centre urbain le plus proche. Alors, $\epsilon_k < 0$ signifie que, toutes choses égales par ailleurs, les zones situées loin des villes ont tendance à contenir plus d'oiseaux que les zones situées près des villes. Enfin, Θ_{ij} correspond à un terme d'interaction entre le i-ième site et la j-ième année.

2.2 Estimation

Le modèle décrit ci-dessus est sur-paramétré. Nous estimons ses paramètres par minimisation de la log-vraisemblance négative Poissonienne associée aux modèles (1) et (2), pénalisée par deux termes de régularisation induisant des solutions parcimonieuses pour

les vecteurs d'effets principaux $(\alpha, \beta, \epsilon)$, et des solutions de faible rang pour la matrice d'interactions Θ:

$$(\hat{\alpha}, \hat{\beta}, \hat{\epsilon}, \hat{\Theta}) \in \text{argmin} \quad \mathcal{L}(Y; \alpha, \beta, \epsilon, \Theta) + \lambda_1 \|\Theta\|_* + \lambda_2 \left(\|\alpha\|_1 + \|\beta\|_1 + \|\epsilon\|_1\right). \quad (3)$$

Dans le problème (3), $\mathcal{L}$ est la log-vraisemblance négative Poissonienne:

$$\mathcal{L}(Y; \alpha, \beta, \epsilon, \Theta) = \sum_{(i,j)} \Omega_{ij}[-Y_{ij}(\alpha_i + \beta_j + \sum_{k=1}^{K} \epsilon_k X_{ij}(k) + \Theta_{ij}) + \exp(\alpha_i + \beta_j + \sum_{k=1}^{K} \epsilon_k X_{ij}(k) + \Theta_{ij})].$$

Le terme $\|\Theta\|_*$ correspond à la norme nucléaire de la matrice Θ (la somme de ses valeurs singulières), et $\|\alpha\|_1$ à la norme ℓ_1 du vecteur α (la somme de ses entrées en valeur absolue); ce dernier terme de régularisation est une pénalité de type LASSO (Tibshirani (1996)). Les propriétés statistiques de l'estimateur (3) ont été étudiées précédemment dans Robin et al. (2017), et la méthode est implémentée dans le package R `lori`. Les données manquantes peuvent être imputées en utilisant l'estimateur (3):

$$\hat{Y}_{ij} = \exp(\hat{\alpha}_i + \hat{\beta}_j + \sum_{k=1}^{K} \hat{\epsilon}_k X_{ij}(k) + \hat{\Theta}_{ij}). \quad (4)$$

En pratique, les paramètres de régularisation λ_1 et λ_2 sont choisis par validation croisée.

2.3 Imputation multiple

Afin de définir une méthode d'imputation multiple, nous identifions deux sources de variabilité dans les données d'abondance d'oiseaux d'eau considérées. Tout d'abord, la variabilité issue de l'échantillonnage des sites. Ensuite, la variabilité des observations, qui sont bruitées, puisque les ornithologues se reposent sur des méthodes d'approximation pour compter les oiseaux. Nous modélisons ce processus par un ré-échantillonnage du tableau Y en deux temps. Premièrement, nous ré-échantillonnons les entrées de Y uniformément complètement aléatoirement: chaque entrée a une probabilité $1 - \pi$ d'être ré-échantillonnée. Deuxièmement, nous modélisons le bruit des mesures avec un tirage Poissonien de moyenne Y_{ij}. Ces deux étapes sont répétées M fois, de sorte que l'on obtient M nouveaux jeux de données, qui seront chacun imputés à l'aide de la méthode d'imputation simple décrite en Section 2.2. Finalement, les M tableaux imputés $\hat{Y}^1, \ldots, \hat{Y}^M$ sont agrégés afin d'obtenir un tableau imputé, ainsi que des estimateurs de la variance de l'imputation de chaque entrée :

$$\begin{array}{rl} \hat{Y}_{ij} & = \frac{1}{M} \sum_{k=1}^{M} \hat{Y}_{ij}^k, \\ \hat{V}_{ij} & = \frac{1}{M} \sum_{k=1}^{M} \hat{V}_{ij}^k + \frac{M+1}{M(M-1)} \sum_{k=1}^{M} (\hat{Y}_{ij}^k - \hat{Y}_{ij})^2. \end{array} \quad (5)$$

De-même, nous obtenons des estimations, ainsi que des intervalles de variabilité pour les comptes annuels totaux.

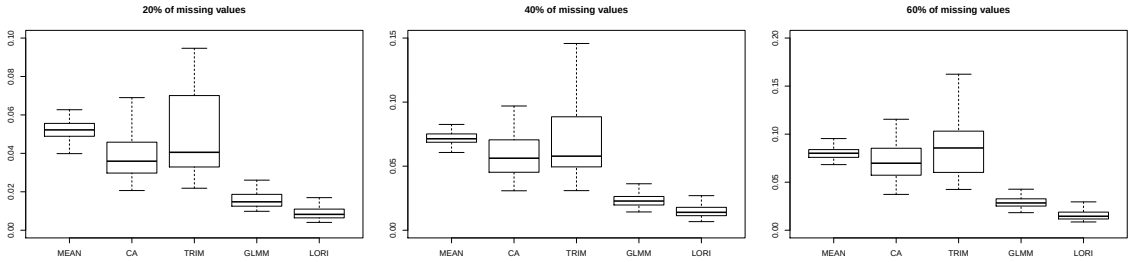

Figure 1: Erreur d'imputation moyenne (100 réplications) pour des pourcentages croissants de données manquantes (20%, 40%, 60%). Méthodes comparées: imputation par la moyenne (MEAN), Analyse des Correspondances (CA), Trends and Indices in Monitoring data (TRIM), Modèle linéaire mixte généralisé (GLMM), Low-rank Interactions (LORI).

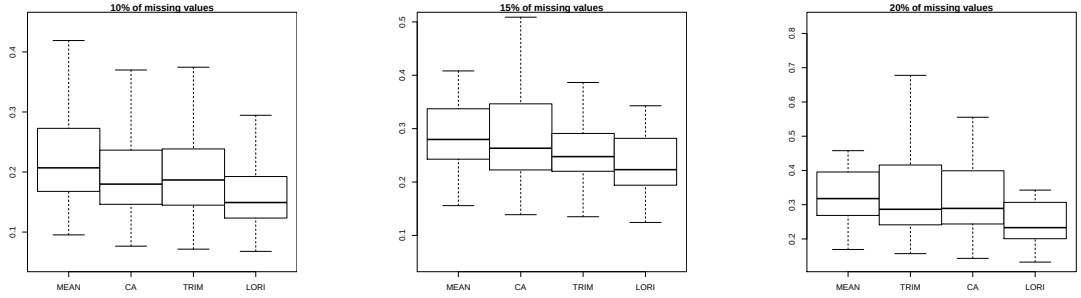

Figure 2: Erreur d'imputation moyenne (100 réplications). Méthodes comparées: imputation par la moyenne (MEAN), Analyse des Correspondances (CA), Trends and Indices in Monitoring data (TRIM), Modèle linéaire mixte généralisé (GLMM), Low-rank Interactions (LORI), pour des pourcentages de données manquantes croissants (10%, 15%, 20%). Le pourcentage indiqué s'ajoute aux 30% de données manquant initialement.

3 Résultats

3.1 Simulations

Nous évaluons ici la méthode décrite, LORI (LOw-Rank Interactions), en terme d'imputation de données de comptage avec covariables. Les figures 1 et 2 présente une comparaison avec des méthodes existantes utilisées en écologie. La Figure 1 présente les résultats sur des données simulées, et la Figure 2 sur un extrait de données d'abondance du canard souchet, contenant initialement 30% de données non-observées. Sur ce dernier jeu de données, nous évaluons les performances en ajoutant des données manquantes artificiellement. Dans les deux cas, nous observons que la LORI donne de plus petites erreurs d'imputation.

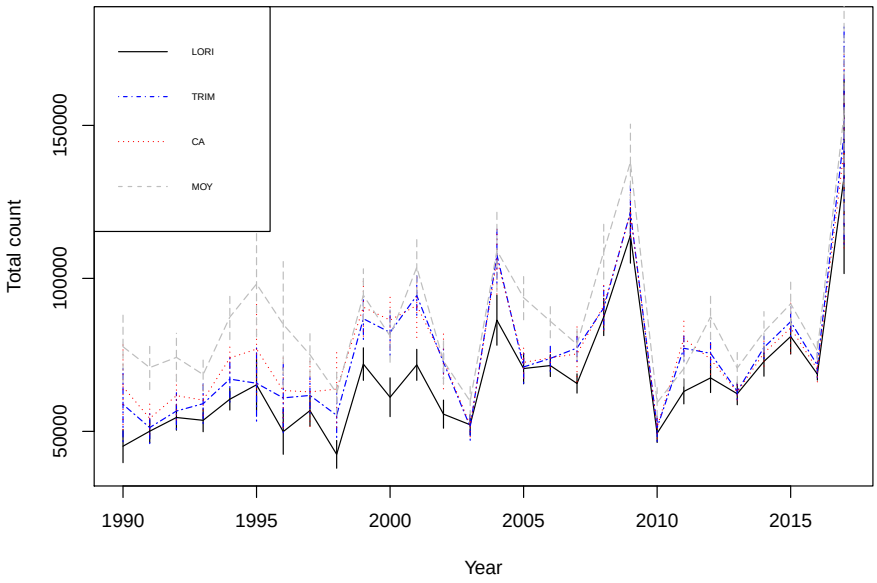

Figure 3: Comptages annuels estimés et intervalles de variabilité.

3.2 Imputation multiple de comptages du canard souchet

Enfin, nous appliquons la méthode à l'imputation multiple des abondances du canard souchet, à partir du jeu de données initial (dont l'exemple de la section précédente est extrait) contenant 55% de données manquantes. Nous obtenons ainsi une estimation du nombre de canards total par année, ainsi que des intervalles de variabilité. La Figure 3 montre que LORI conduit à une révision à la baisse des comptages annuels, en particulier pour les années 1998-2002, par rapport à TRIM et CA qui donnent, quant à elles, des résultats très proches.

Bibliographie

Amano T., Székely T., Sandel B., Nagy S., Mundkur T., Langendoen T., Blanco D., Soykan C. U. and Sutherland W. J. (2018), Successful conservation of global waterbird populations depends on effective governance, *Nature*, 553, 199–202.

Amat, J.A., Green, A.J. (2010), Waterbirds as bioindicators of environmental conditions. In: Hurford, C., Schneider, M., Cowx, I. (Eds.), Conservation Monitoring in Freshwater Habitats. Springer, Netherlands.

Josse, J. and Husson, F. (2016), missMDA: A Package for Handling Missing Values in Multivariate Data Analysis, *Journal of Statistical Software*, 70(1).

Pannekoek, J. and van Strien, A. (2001). TRIM 3 Manual (Trends & Indices for Monitoring Data), *Statistics Netherlands*.

Robin, G., Josse, J., Moulines, É. et Sardy, S. (2017). Low-rank model with covariates for count data analysis, *arXiv e-prints*, `https://ui.adsabs.harvard.edu/#abs/2017arXiv170302296R`.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288. JSTOR, `www.jstor.org/stable/2346178`.