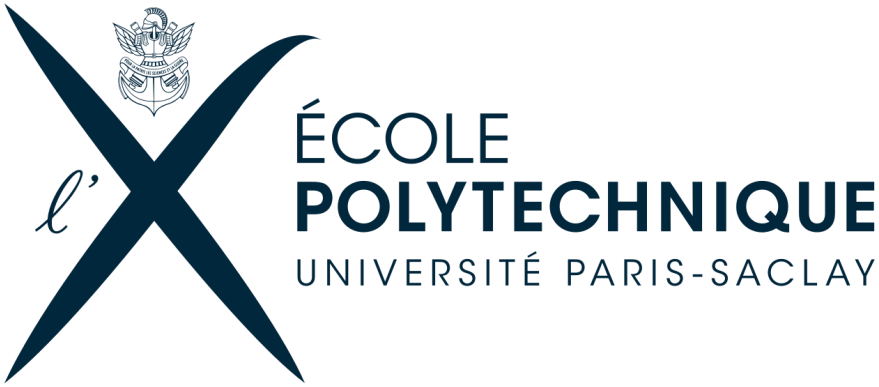
Estimation d'abondances d'espèces à l'aide de modèles log-linéaires

MAP573 - R pour les statistiques



Presentation Date: 11.12.2018

Group Members: Simon Klotz, Thibault de Rycke, Seongbin Lim, Lucas Elbert
Supervisor: Geneviève Robin
Course Coordinator: Julie Josse

Table of Contents

Introduction

The estimation of bird abundance is an important ecological task and is widely used for bird conservation. Knowing the size of bird populations over the years allows to draw several conclusions such as how habitat loss, agriculture or pollution affects the population. This helps to detect whether a species is declining and also how external factors influence this. In order to draw these conclusions, accurate data on bird abundance is necessary. Otherwise conservation resources may be applied in the wrong location or on the wrong issue. However, the bird count data is mainly gathered by volunteers who count the birds by hand. This can lead to inaccuracies and frequent missing data which impedes finding answers to those questions. Methods such as Correspondence Analysis or Generalized Linear Models can be used to estimate these missing values and allow a more accurate analysis of bird abundance.
The objective of this project is to investigate the abundance for the Eurasian Coot, which is mainly observed in the mediterranean part of North-Africa, and its relation to external geographical and meteorological factors. First, different methods are compared in terms of accuracy. Afterwards, external factors and their impact on bird abundance are examined and finally the temporal trend is investigated to determine whether the Eurasian coot is declining or not.

Figure 1: Eurasian Coot [5]

Dataset

This section takes a brief look at the structure of the provided data. The dataset contains a contingency table of bird counts for 498 sites in the years from 1990 to 2017 as well as several covariates describing the sites and years.

	1990	**1991**	**1992**	**...**	**2015**	**2016**	**2017**
Site 6	100	0	n.a.	...	21	51	500
Site 10	n.a.	n.a.	59	...	88	98	96

Table 1: Sample of contingency table

The covariates can be summarized in two categories:

- Row covariates containing fixed geographic information about the site:
    - Longitude and latitude
    - Altitude
    - Distance to towns and coasts
    - Existence of a dam
    - Area
    - Water surface
    - Country
    - ...
- Column covariates containing year-dependent information:
    - Temperature anomalies in europe
    - Rainfall
    - Economic indicators
    - Agricultural indicators
    - ...

One important measure of the dataset is the amount of both available and missing data. There are a total of 13944 data points, whereas 5688 values (41%) were actually counted and 8256 values (59%) are missing.

Another interesting aspect of the dataset are outliers. The distribution of all counts shows that almost all counts are in the hundreds but there are some outliers greater than 140000 which can lead to difficulties for the prediction of missing values.

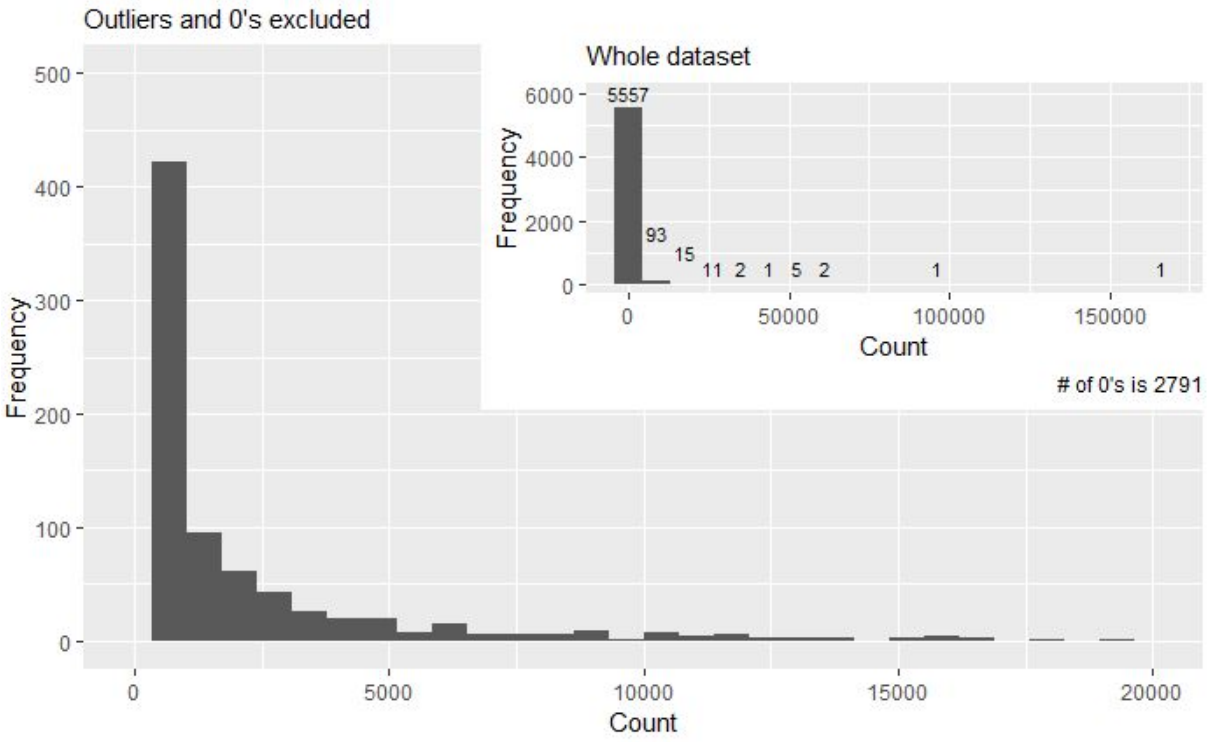


Figure 2: Distribution of bird count

Figure 2 depicts a map with all sites the Eurasian Coot is observed by bird watchers. It is interesting to note that most sites are near the coast or near a river (such as in Egypt). Furthermore, the sites are mainly in Morocco, Algeria, and Tunisia whereas Libya and Egypt only have a few.

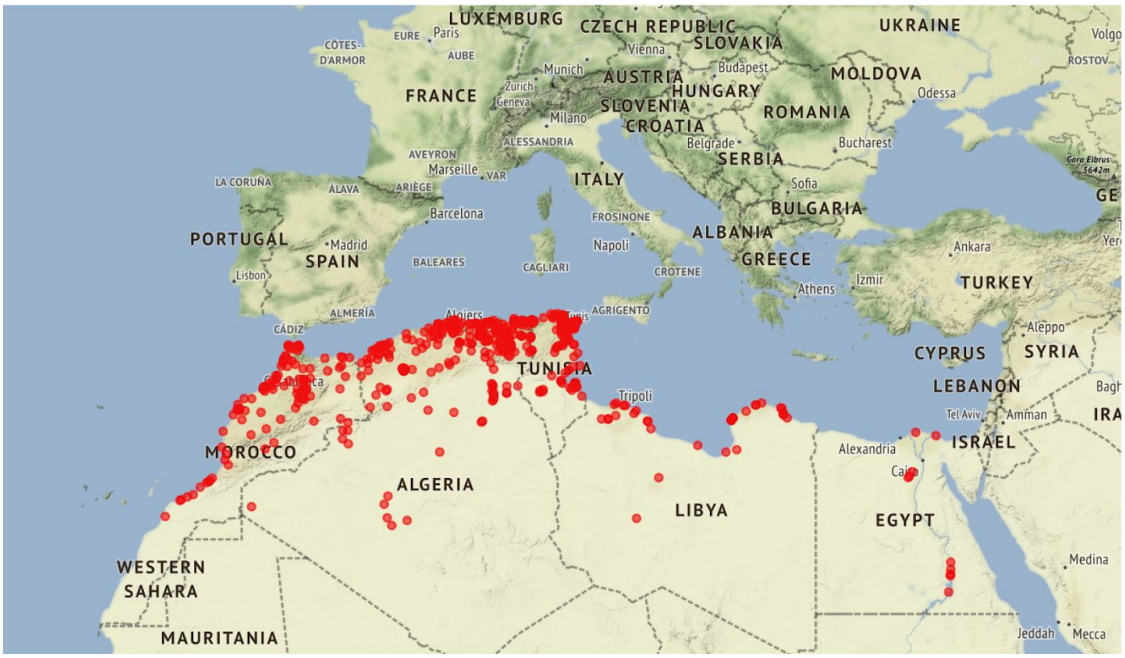


Figure 3: Map of used sites

Methods

This section describes the methods used to impute the missing values from the contingency table and to estimate the effect of the covariates indicating the impact of features on the bird population.

Generalized Linear Model

The dataset contains many explanatory variables in form of covariates and the bird count as one response variable. Since the task is to understand their interaction, regression methods can be applied. The simple linear regression does not match the special requirements of the given dataset for many reasons (e.g. a linear regression would possibly give negative valued predictions, which do not exist in the context of counts). The generalized linear model is a generalization of the linear model which is more capable in adapting to these needs.

Method Description

The generalized linear model [1] is very similar to the linear model. Given n observations of p explanatory variables row-wise in X and the random variable of the respective response as Y then the GLM (generalized linear model) consists of mainly three ingredients

- A probability distribution D from the exponential family
- A (previously unknown) coefficient vector β
- A link function g

The model then states

$$E[Y] = g^{-1}(X\beta)$$
$$(Y|X) \sim D(X, \beta)$$

In the ordinary linear regression, g would be the identity function and D would be the normal distribution $N(X\beta, \sigma^2)$. Because the count data seems to follow a Poisson distribution and contains only positive integer values, it is reasonable to choose a Poisson distribution with logarithmic link function in this case.

The model then searches for a β maximizing the likelihood of the observations.

In this model there is a strong connection between the link function g and the distribution D because with given X, β, g the first equation already makes a statement about the mean of D. When D is chosen from the exponential family, then, by its parametrization, there always exists a natural choice for the link function g which fulfils that restriction.

After optimizing the coefficient vector β, the model can be used to interpret the coefficients and also make predictions of the response variable for unseen explanatory variables. Given a single new observation x with all p covariates, then the prediction can simply be derived as $E[y] = exp(x\beta)$. This also gives some indication about how to interpret the coefficients.

Coefficient	Interpretation
$\beta_i < 0$	Increasing parameter value by 1 divides the prediction by $exp(\|\beta_i\|)$.
$\beta_i = 0$	No effect
$\beta_i > 0$	Increasing parameter value by 1 multiplies the prediction by $exp(\|\beta_i\|)$.

Implementation

In R there are two popular packages that provide access to generalized linear models, namely glm and glmnet. Whilst glm implements exactly the above described method and supports some more link functions, glmnet has the advantage of also handling Lasso and Ridge regularizations as well as tradeoffs between those. Since either of the two packages implement GLM the experiments are conducted with glm.

```
model <- glm(formula="Y~X1+X2+X3", family=poisson(link=log), data=train_data,
na.action=na.omit)
prediction_glm <- predict.glm(model,newdata=test_data,type="response")
```

The first statement defines and fits a GLM to the data and the second statement produces predictions for new data based on the trained model.
It is worth mentioning that the formula parameter at the first line determines which variables are used as input to the model. The mathematical relationship is given by the family parameter.

Rtrim

Trim (TRends & Indices for Monitoring Data) [3] is a package tailored for our problem. It analyses count data obtained by monitoring wildlife. The purpose of this analysis is not only to produce estimates of annual indices, but also to analyse trends between these indices. TRIM estimates a model using the observed counts, and then uses this model to predict the missing count values. The only issue is the use of covariates: it does not focus on predicting which covariate is the most important, and only accepts site clusters as an input covariate.

Method Description

There are 3 different models, and all of them belong to the class of loglinear models.
Let the count in site i at time j be $f_{ij}(i = 1...I, j = 1...J)$, with I the number of sites, and J the number of years. The expected values of the count is expressed as a function of site-effects and time-effects. Simpler models will work even with a substantial amount of missing counts.

The expected count will be denoted by μ_{ij}, and estimated expected count by $\mu\$_{ij}$.
The count after imputation will be :

$$f_{ij}^{+} = \delta_{ij} * f_{ij} + (1 - \delta_{ij}) * \mu\$_{ij}$$

with $\delta_{ij} = 1$ for observed site by time combinations, and $\delta_{ij} = 0$ if the count value is missing.
The three models are:

- **Model 1: No time-effect**
  Model one is a very simple model, and the expected count only varies with the sites:

  $$ln(\mu_{ij}) = \alpha_i$$

  With α_i the effect for site i. This model is not good enough for us, because it considers that the count only varies from one site to another, and not from a year to another. It is therefore useless to study whether a population of birds is disappearing.
- **Model 2: Linear trend**
  This model contains a site-effect and a linear effect of time. It can be written like below:

  $$ln(\mu_{ij}) = \alpha_i + \beta * (j - 1)$$

  Therefore, in this model, the log expected count increases with an amount from one time-point to the next. This implies a constant increase from one time point to the next, which might not be suited for a long period of time.
- **Model 3: Effects for each time-point**
  This model uses separate parameters for each time-point, and therefore can depict more accurately the time trends. It can be expressed as:

  $$ln(\mu_{ij}) = \alpha_i + \gamma_j$$

  With γ_j the effect for time j on the log-expected counts. Both models 2 and 3 are restrictive, since the time-parameters (β and γ_j) are the same for each site. It is therefore useful to covariates to relax this assumption and improve the model. The time-parameters will then be the same for each cluster of sites.

Implementation

In order to improve our implementation, clusters are added as covariates using the package mclust. In order to create a covariate as meaningful as possible, the clusters were created based on the most important features according to the other methods: agriculture, latitude, and countries.
Model 2 was the best fit for the implementation with the rtrim package. Model 1 was eliminated because it oversimplified the problem. Model 3 might have been a good fit, but was ultimately unusable since there must be at least one observation for every value of each covariate, at each time-point, which is not the case for the given data frame.

```
cov_trim$cluster = Mclust(delay[,3:6], verbose=FALSE)$classification
result <- trim(cov_trim, count_col = "value", site_col = "site", year_col =
"year", month_col = NULL, covar_cols="cluster", model=2, autodelete=FALSE)
```

With delay being a data frame with sites, countries, agriculture and latitude, and cov_trim being a data frame with sites, years, count values and clusters.

Lori

Most probabilistic methods for the estimation of contingency tables containing count data only rely on the table itself for predictions. However, in many cases there is additional information available such as row and column variates which may help to describe the contingency table. In contrast to other methods, Lori (Low-Rank Interaction Contingency Tables) [2] incorporates these covariates to increase the prediction accuracy.

Method Description

Lori is an implementation in R using a method that extends the row-column model to describe the structure of a count matrix $Y \in \mathbb{R}^{n \times p}$. In addition to terms depending on the rows and columns of the count matrix which are called main effects, also covariates and interactions between them are used. The matrix $X^* \in \mathbb{R}^{n \times p}$ is modeled as:

$$X^*_{ij} = \mu^* + \sum_{k=1}^{K_1} R_{ik}\alpha^*_k + \sum_{l=1}^{K_2} C_{il}\beta^*_l + \Theta^*_{ij} \qquad rank(\Theta^*) \leq min(n-1, p-1)$$

Where $R \in \mathbb{R}^{n \times K_1}$ is the matrix of row covariates, $C \in \mathbb{R}^{p \times K_2}$ the matrix of column variates, $\mu^* \in \mathbb{R}$ is an offset, $\alpha^* \in \mathbb{R}^{K_1}$ and $\beta^* \in \mathbb{R}^{K_2}$ are called main effects and $\Theta^* \in \mathbb{R}^{n \times p}$ is the matrix of interactions that are unexplained by the covariates called interaction matrix. X^* is then found by minimizing the negative Poisson log-likelihood on $\mu^*, \alpha^*, \beta^*, \Theta^*$:

$$\Phi_Y(X) = -\frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (Y_{ij} X_{ij} - exp(X_{ij}))$$

The log-likelihood is minimized using the alternating direction method of multipliers. Furthermore, regularization is used to regularize the interaction matrix and covariate effects.

Implementation

The R package implementing the method is called lori which provides several methods to prepare data, tune the methods hyperparameters and predict the count table. The method cv.lori can be used to tune the hyperparameters using cross validation. Finally, to do the imputation of the count matrix one uses the lori method. Lori also allows to ignore row or column effects by setting either the reff or ceff parameter to FALSE.

```
reg <- cv.lori(Y, cov=covariates, N=10, thresh=1e-05, maxit=100,
rank.max=5)
result <- lori(Y, cov = NULL, lambda1 = reg$lambda1, lambda2 = reg$lambda2,
maxit=1000, reff=TRUE, ceff=TRUE)
```

Correspondence Analysis

Correspondence analysis (CA) [4] is designed for data with two qualitative variables, which means that it is comparable to other methods because it does not take into account all the explanatory variables like others. At the same time, however, this method could be very powerful since the data has so many missing values that others might not be able to find

correlations properly. CA was used not for analysing the data itself but for the purpose of the imputation of these missing values.

Method Description

CA is based on the contingency table having rows and columns, which represent two quantitative variables ($V_I,\ V_J$). The numbers of rows (I) and columns (J) are determined by each number of categories ($\{1, ..., i, ..., I\} \in V_I,\ \{1, ..., j, ..., J\} \in V_J$) in each variable.
CA can be applicable to the data having two quantitative variables ($V_I,\ V_J$). Then the contingency table is constructed to have categories from each variable ($i \in V_I, j \in V_J$) at rows and at columns, respectively. Each entry should be the probabilities (f_{ij}) with respect to the whole sum of data. At the end of each row and column, it calculates the sum of each line, called the marginal sum ($F_{i.},\ F_{.j}$).

$$f_{ij} = \frac{x_{ij}}{\sum_i^I \sum_j^J x_{ij}} = \frac{x_{ij}}{N},\ F_{i.} = \sum_j^J f_{ij},\ F_{.j} = \sum_i^I f_{ij}$$ (where x_{ij} is a real value.)

As a null hypothesis, CA assumes the estimates independent to each other and calculates the expected probability (f_{ij}'). Then it estimates the association between two variables using the Chi-square distance between the observed and the expected.

$$f_{ij}' = F_{i.} \cdot F_{.j} \text{ and } \frac{f_{ij}'}{F_{i.}} = F_{.j},\ \frac{f_{ij}'}{F_{.j}} = F_{i.}$$

$$\chi^2_{dist} = \sum_i^I \sum_j^J \frac{\left(Nf_{ij} - Nf_{ij}'\right)^2}{Nf_{ij}} = N\Phi^2$$

Lastly, CA applies a factor analysis such as SVD (singular value decomposition) algorithm to quantify how strong the association would be. The solution gives eigenvalues and eigenvectors that maximize the variance of the original data.
For the purpose of the imputation, CA makes use of the decomposed eigenvectors. The number of components is to be necessarily less than the smallest number of categories.

$$ncp < I \wedge ncp < J$$

Implementation

The R package missMDA provides helpful imputation functions using PCA, CA, and even MCA (Multiple CA). Among those, a function called imputeCA was used to perform the imputation. The function employs the iterative algorithm that guesses missing values randomly and regularizes them. It has one main hyperparameter, namely ncp, which refers to the number of components.
Along with the function imputeCA, a simple k-fold cross-validation was added to determine the hyperparameter ncp.

```
imputeCA(X, ncp = KFold(), threshold = 1e-08, maxiter = 1000)
```

Results

This section describes the results obtained during all experiments. First, the imputation quality of different methods was compared. Afterwards, the influence of different factors on the bird count was investigated and finally the temporal trend of the Eurasian Coot abundance was examined.

Imputation Quality

First, the methods were compared in regard to their imputation quality. Different subsets of the count matrix and covariates were sampled to fit a model and then this model was used to predict the remaining counts to measure its performance. Mean squared error was used to compare the methods as it is a common error measure for Poisson distributed data. In order to compare the performance of the models on different amounts of available data, each experiment was executed for 10% and 60% of missing data. A baseline model that imputed the column means for each cell was used as it is common for these kinds of problems.

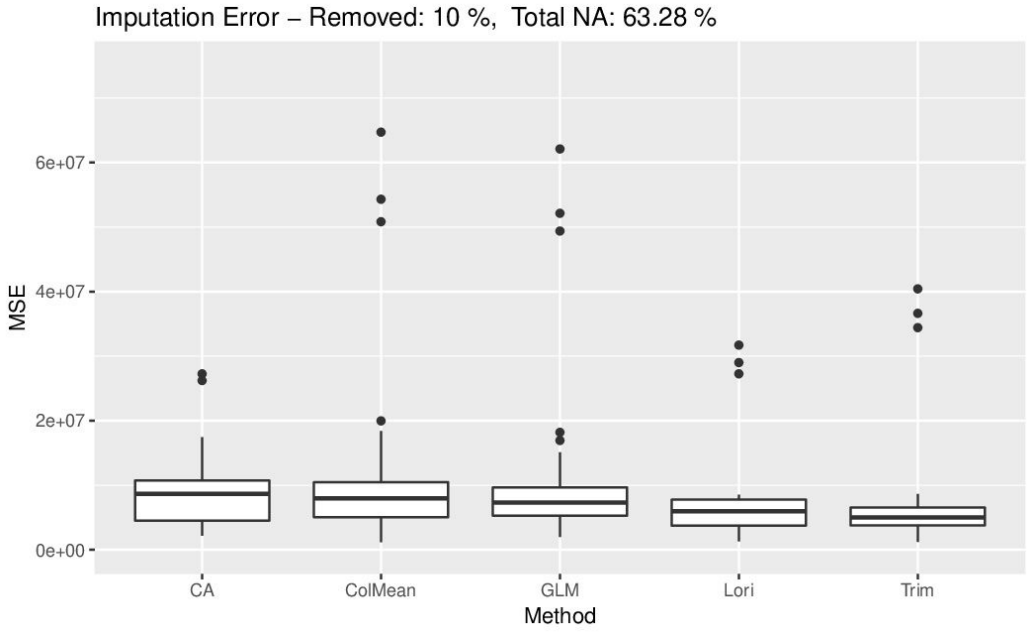


Figure 4: Mean squared error for methods applied on 10% missing values

Figure 4 shows the mean squared error obtained by each method on samples where 10% of the counts are missing. Trim obtained the best results shortly followed by Lori. It is important to notice that the mean of the errors for CA was worse compared to the column wise mean. Furthermore, for every method there were several outliers with significantly higher error which was most likely due to the distribution of all counts having a long tail and the small size of the data tested on which was only 10% of the original counts.

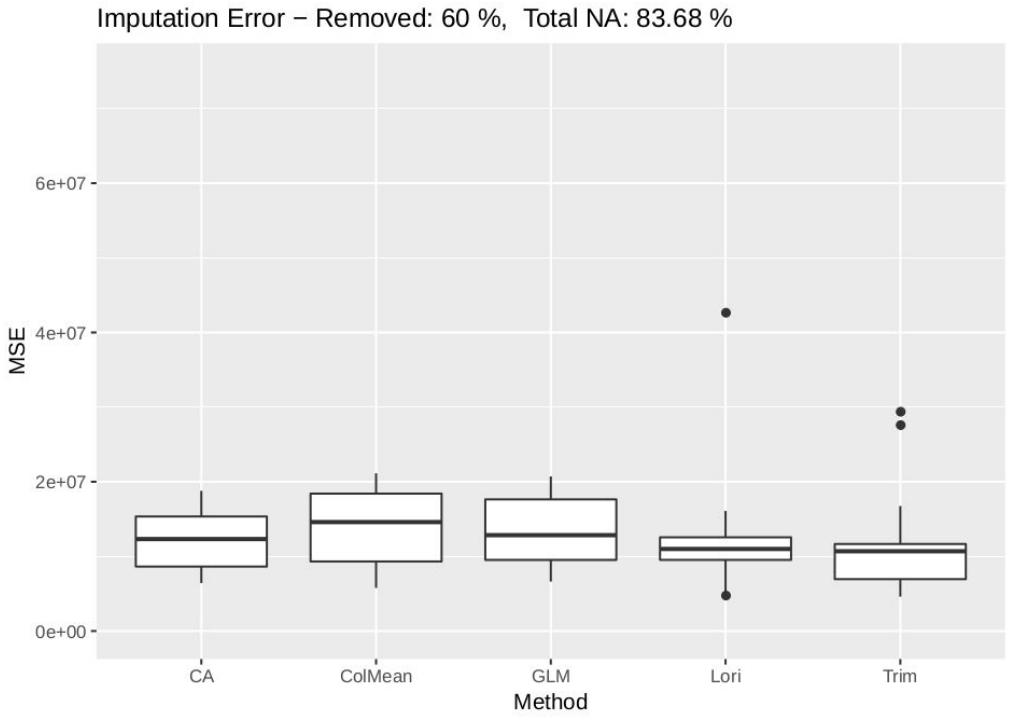


Figure 5: Mean squared error for methods applied on 60% missing values

Figure 5 depicts the errors for 60% removed data. With less data available for fitting the models, Trim and Lori still outperformed the other methods. Also, CA and GLM achieved a slightly better mean error than imputation by column means. Compared to only 10% missing values the methods achieved worse results on average and the variance in error scores was generally higher. Furthermore, there existed some outliers for GLM which where orders of magnitude greater than the errors depicted in the boxplot which were removed in the visualization to allow a better comparison. This only happened if the fraction of missing data was high, which leads to the conclusion that GLM is less robust to these outliers compared to the other methods.

In conclusion, Trim achieved the best results for all experiments shortly followed by Lori. The other methods were only slightly better compared to the column mean imputation and GLM was less robust when only few data was available. As expected, the error generally increased when less data was available for fitting each method.

Feature Importance

The coefficients of both Lori and GLM indicate how a covariate influences the count of birds which helps to understand what is causing an increase or decline in bird populations. To achieve a more robust result, the same method as for the comparison of different methods was applied by sampling multiple subsets of the data with 10% missing values, fitting the models to each subset and then obtaining the coefficients for the covariates. The results are depicted for both Lori and GLM in Figure 6.

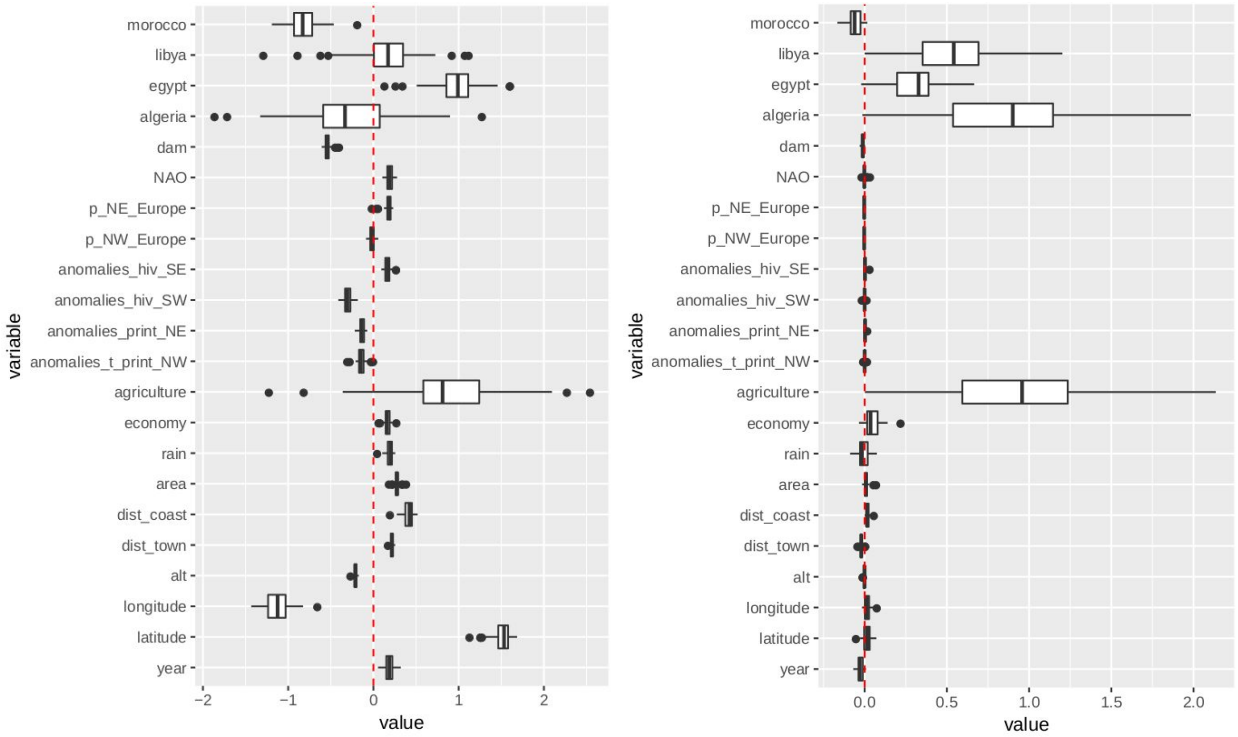


Figure 6: Coefficients for covariates of GLM (left) and Lori (right)

It is important to note that the coefficients with high values are similar for both methods. There seem to be more birds in Libya and Egypt and less birds in morocco. The only difference in methods is that for GLM Algeria seems to have a mostly negative impact while for Lori it has a positive impact on the bird count. Also, the agriculture indicator has a positive impact for both methods which may be explained by the fact that if more land is cultivated there exists more food for the birds. Furthermore, the dam covariate for GLM is negative which indicates that artificial wetlands are less favourable than natural ones. Also, for GLM the location of the site measured by longitude and latitude has great coefficients which proves that the geographic location matters and the positive coefficient for latitude may indicate that birds prefer areas closer to the mediterranean sea. Other smaller coefficients for GLM indicate that birds prefer areas with more rain, greater distance to both towns and coasts and lower altitude. However the results of the GLM should be treated with caution since it showed a big overdispersion on the data. In general, Lori has less coefficients significantly greater than zero due to the regularization.

Temporal Trend

Temporal trend is one of the most important indices. It tells whether the population of a species is increasing or decreasing, and even further whether this species is in danger of extinction so that it should be conserved. For the purpose of clarity, only two time trends, which were obtained from Lori and Trim, were picked because the other methods were proven to be less accurate by earlier experiments.
The result in Figure 7 summed up birds through all sites annually. It shows the change before and after the imputation as well as a tendency by using linear regression.

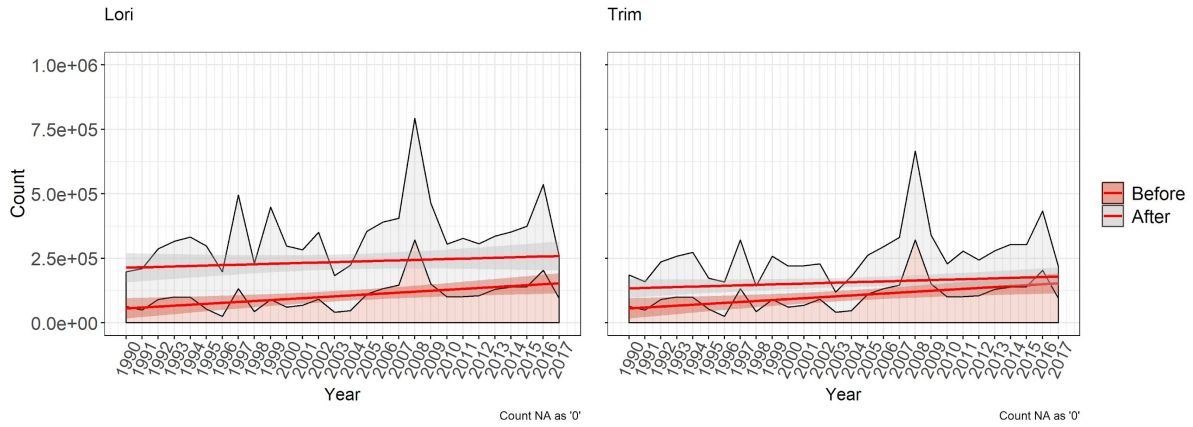


Figure 7: Time trends for Lori and Trim before and after the imputation.

After the imputation, both methods showed distributions similar to the original data. Especially, some distinct peaks at the years of 1997, 2008, and 2016 were still distinguishable from other years. However, the difference occurs at some peaks at the years of 1999 and 2002. Lori suggests that there should have been a big missing number birds, while Trim followed the tendency of the original data.

Even though both predict that the number of birds has been increasing, there was a difference in the intercept of the linear regression. The intercept of Lori was bigger than Trim's, meaning that the number of missing values may be bigger than the expectation of Trim.

Conclusion

The goal of this project was to impute the missing values for the counts of the Eurasian Coot along different years and sites in North Africa. The comparison of different methods showed that Trim and Lori are best suited for this task in terms of accuracy and that all methods outperform the baseline based on the column means. Furthermore, accuracy decreases when less data is available for fitting each model. The investigation of feature coefficients using Lori and GLM indicated that bird counts are higher in Libya and Egypt. It also lead to some interesting conclusions for the conservation of birds: Since agriculture has a positive effect on the bird count, supporting the local farmers will also have an advantage on the bird population. It may also be worth looking into what distinguishes artificial wetlands from real ones to draw conclusions why birds prefer natural ones and how artificial wetlands can be improved to attract more birds. The temporal trend for both imputed and original data showed that the population of the Eurasian Coot is already increasing. In conclusion, both Lori and Trim are well suited to answer these kind of questions and especially Lori can help to understand the causes for a decline or increase of population size through covariate coefficients which provide valuable information for ecologists.

References

[1] Nelder, J. A., & Baker, R. J. (2004). Generalized linear models. Encyclopedia of statistical sciences, 4.

[2] Robin, G., Josse, J., Moulines, E., Sardy, S., & Robin, G. E. (2017). Low-rank Interaction Contingency Tables. arXiv preprint arXiv:1703.02296.

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

[4] Yelland, P. M. (2010). An introduction to correspondence analysis. The Mathematica Journal, 12(1), 86-109.

[5] http://www.oiseaux.net/photos/nathalie.santa.maria/foulque.macroule.2.html#espece