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

ÉCOLE
POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

11.12.2018
MAP573 - R pour les statistiques
Simon Klotz, Thibault de Rycke, Seongbin Lim, Lucas Elbert

Agenda

- Introduction
- Dataset
- Methods
- Results
- Conclusion

Eurasian Coot [5]

Introduction

Estimation of Bird Abundance

Important for ecologists

Helps to understand what is causing decline or increase of population

Supports conservation of birds

Introduction

Estimation of Bird Abundance

Important for ecologists

Helps to understand what is causing decline or increase of population

Supports conservation of birds

Problems

Birds are counted by volunteers

Inaccurate data

Lots of missing values

⇒ Necessary to impute missing values for accurate estimation

Introduction

Estimation of Bird Abundance

Important for ecologists

Helps to understand what is causing decline or increase of population

Supports conservation of birds

Problems

Birds are counted by volunteers

Inaccurate data

Lots of missing values

⇒ Necessary to impute missing values for accurate estimation

Research Questions

How do methods compare for count imputation?

How do external factors affect bird population?

What does temporal trend of population size show?

Dataset

Contingency table

	1990	1991	1992	...	2017
Site 6	100	0	n.a.	...	500
Site 10	n.a.	n.a.	59	...	96

Data availability

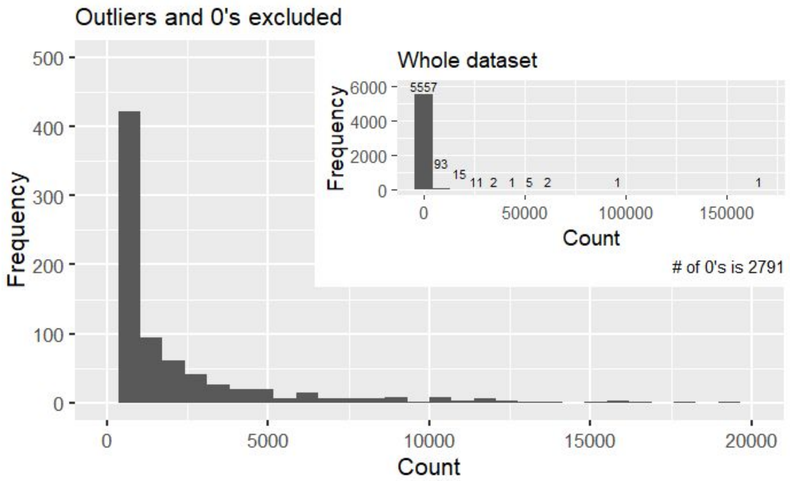
41% available

59% not available

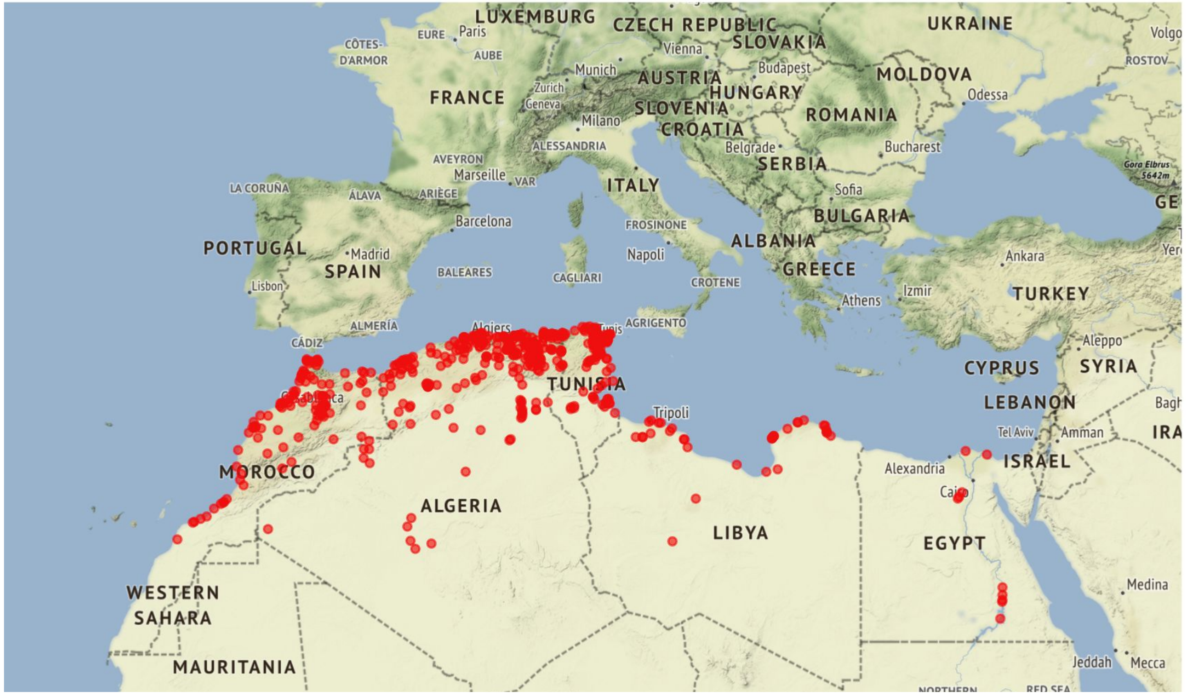
Covariates

- Covariate that depend on the site e.g.:
  - Longitude and latitude
  - Area
  - Distance to town and coasts
- Covariate that depend on the year
  - Temperature anomalies
- Covariates depending on both e.g.:
  - Rainfall
  - Agricultural indicators

Distribution & outliers



Sites



Map of all sites where Eurasian Coot is observed

Methods

- GLM
- Trim
- CA
- Lori

GLM - Generalized Linear Model

Linear model

Response Covariates

$$\mathbb{E}[Y] = X\beta$$

$$(Y|X) \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

Learnable parameters

GLM – Generalized Linear Model

Linear model

Response — Covariates

$$\mathbb{E}[Y] = X\beta$$

$$(Y|X) \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

Learnable parameters

Generalized linear model

Response — Link function

$$\mathbb{E}[Y] = g^{-1}(X\beta)$$

$$(Y|X) \sim \mathcal{D}(X, \beta)$$

Covariates — Distribution

$$\Rightarrow \quad \mathcal{D} = \mathrm{Pois}(g^{-1}(X\beta)) \qquad g = \ln$$

$\Rightarrow \quad \beta$ maximizes likelihood

GLM - Generalized Linear Model

- R packages: glm, glmnet

- Example call:
  - Model fitting
  - Predictions

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

Trim - TRends and Indices for Monitoring data

- Method specifically created to analyze count data from monitoring wildlife
- Produces estimates of annual indices, and trends between these indices

Model 1	**Model 2**	**Model 3**
No time-effect	Linear trend	Effects for each time-point
$ln(\mu_{ij}) = \alpha_i$	$ln(\mu_{ij}) = \alpha_i + \beta * (j - 1)$	$ln(\mu_{ij}) = \alpha_i + \gamma_j$
With α_i the effect for site i	Implies a constant increase	With γ_j the effect for time j

- Time parameters: same for each site
- Covariates: create clusters of sites to improve our model

Trim - TRends and Indices for Monitoring data

- Which model to use:
  - Model 1 oversimplifies the problem
  - Model 3 needs one value for each cluster for each year → not feasible for our data

  ⇒ Model 2 with clusters

- R package: RTrim

- Application example:

```
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)
```

CA - Correspondence Analysis

- Data with 2 categorical variables
- Derive expected contingency table:
  - Calculate marginal sum
  - $f'_{ij} = F_{.i} \cdot F_{j.}$
- SVD (Singular Value Decomposition)

Expected Contingency table

	...	j	...	J	Sum
⋮	⋱	⋮		⋮	⋮
i	...	f_{ij}'	...	f_{iJ}'	$F_{i.}$
⋮		⋮	⋱	⋮	⋮
I	...	f_{Ij}'	...	f_{IJ}'	$F_{I.}$
Sum	...	$F_{.j}$	...	$F_{.J}$	1

Expected prob.

Distance

Correlation

Observed prob.

$$\chi^2_{dist} = \sum_{i}^{I} \sum_{j}^{J} \frac{(Nf_{ij} - Nf'_{ij})^2}{Nf_{ij}} = N\Phi^2$$

Total sum

$$argmax_{u_1, u_2, \ldots} \Phi^2$$

Eigenvectors

CA - Correspondence Analysis

- 2 categorical variables **‘Site’ & ‘Year’**

	1990	1991	1992	...	2017
Site 6	100	0	n.a.	...	500
Site 10	n.a.	n.a.	59	...	96

- R package ‘missMDA’
- Hyperparameter ***ncp***
  - K-Fold cross-validation
  - Better be small

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

Lori - Low-Rank Interaction Contingency

- Specifically for imputation of contingency matrices
- Incorporates additional knowledge using covariates

$$\underbrace{X_{ij}^*}_{\text{Estimation}} = \underbrace{\mu^*}_{\textit{Offset}} + \sum_{k=1}^{K_1} \underbrace{R_{ik}}_{\text{Row covariates}} \underbrace{\alpha_k^*}_{\textit{Site effects}} + \sum_{l=1}^{K_2} \underbrace{C_{il}}_{\text{Column covariates}} \underbrace{\beta_l^*}_{\textit{Year effects}} + \underbrace{\Theta_{ij}^*}_{\textit{Interaction Matrix}} \quad rank(\Theta^*) \leq min(n-1, p-1)$$

$$X_{ij}^* = argmin_{x \in \mathbb{R}}\{-\mathbb{E}[Y_{ij}]x + exp(x)\}$$

- Regularization:
  - Nuclear norm for interaction matrix
  - L1 norm for site and year effects

Lori - Low-Rank Interaction Contingency

- R package: lori

- Application example:
  - Find regularization parameters by cross validation
  - Predictions

```
reg <- cv.lori(Y, cov=covariates, N=10, thresh=1e-05, maxit=100,
rank.max=5)

result <- lori(Y, cov=covariates, lambda1 = reg$lambda1,
lambda2 = reg$lambda2, reff=TRUE, ceff=TRUE)
```

Results

Imputation Quality

Feature Importance

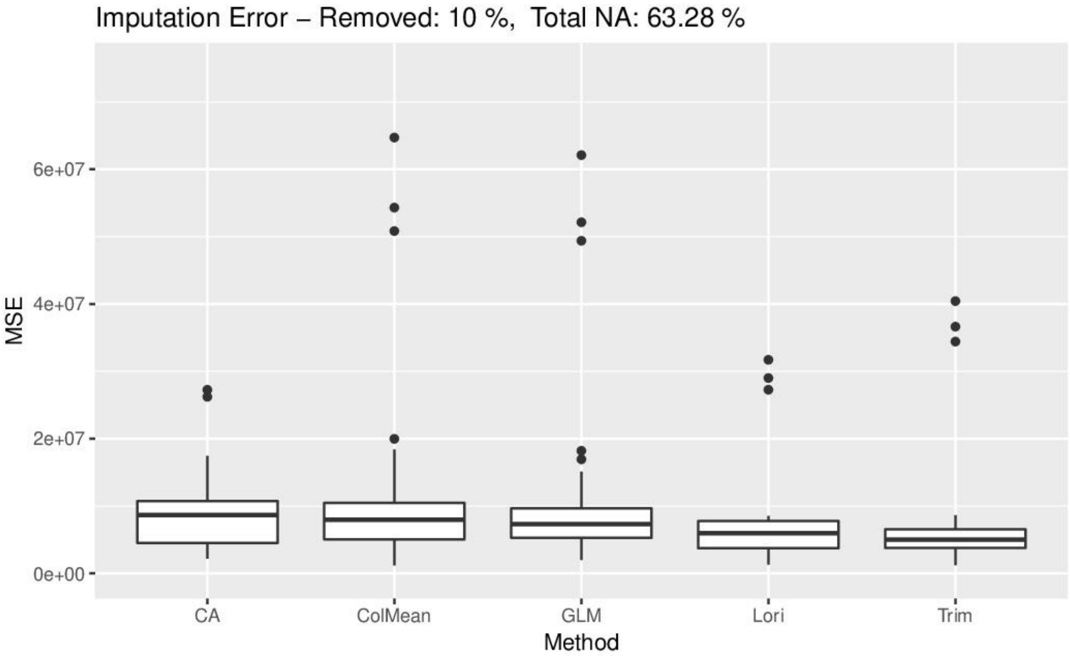
Temporal Trend

Imputation Quality

1. Sample multiple subsets by removing certain percentage of available data
2. Fit methods on remaining data
3. Predict bird count for removed data
4. Calculate error metrics on removed data
5. Compare to baseline model

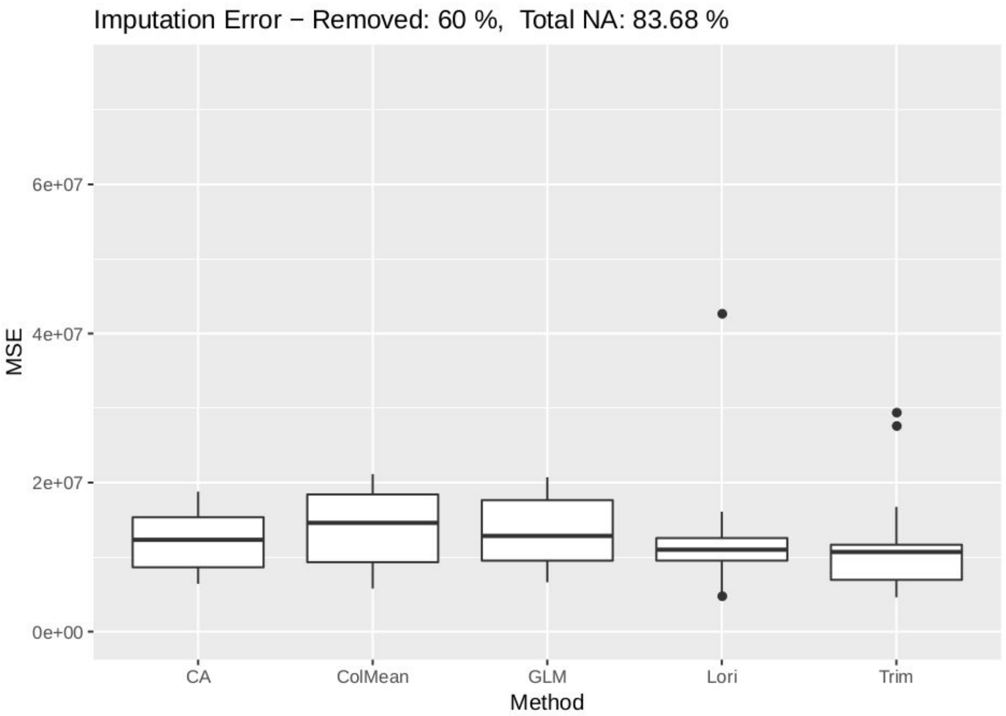
Imputation Quality

- Lori and Trim best performance
- Column mean performs rather good
- Several outliers

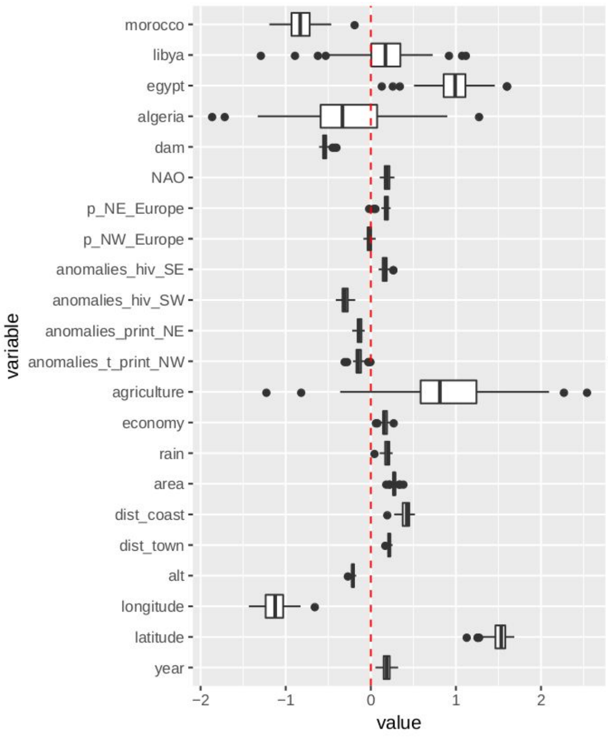


Imputation Quality

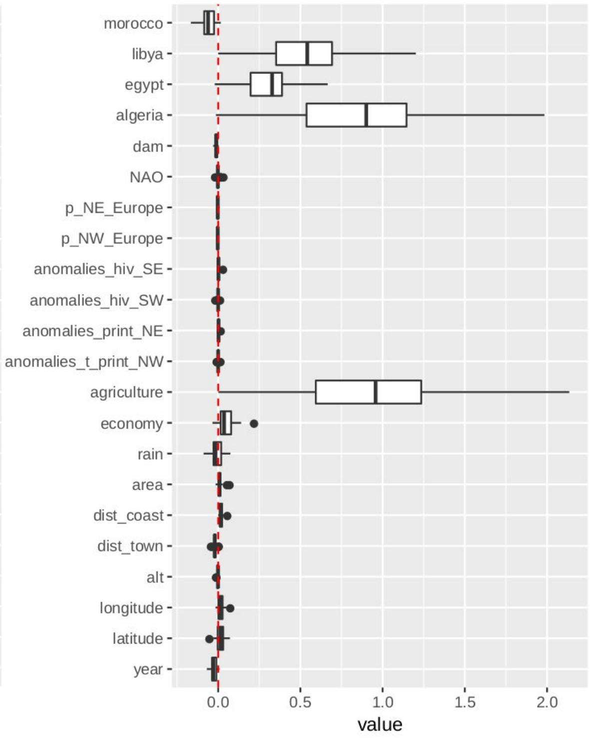
- Lori and Trim best performance
- Worse performance than 10% missing data
- GLM has great outliers



Feature Importance

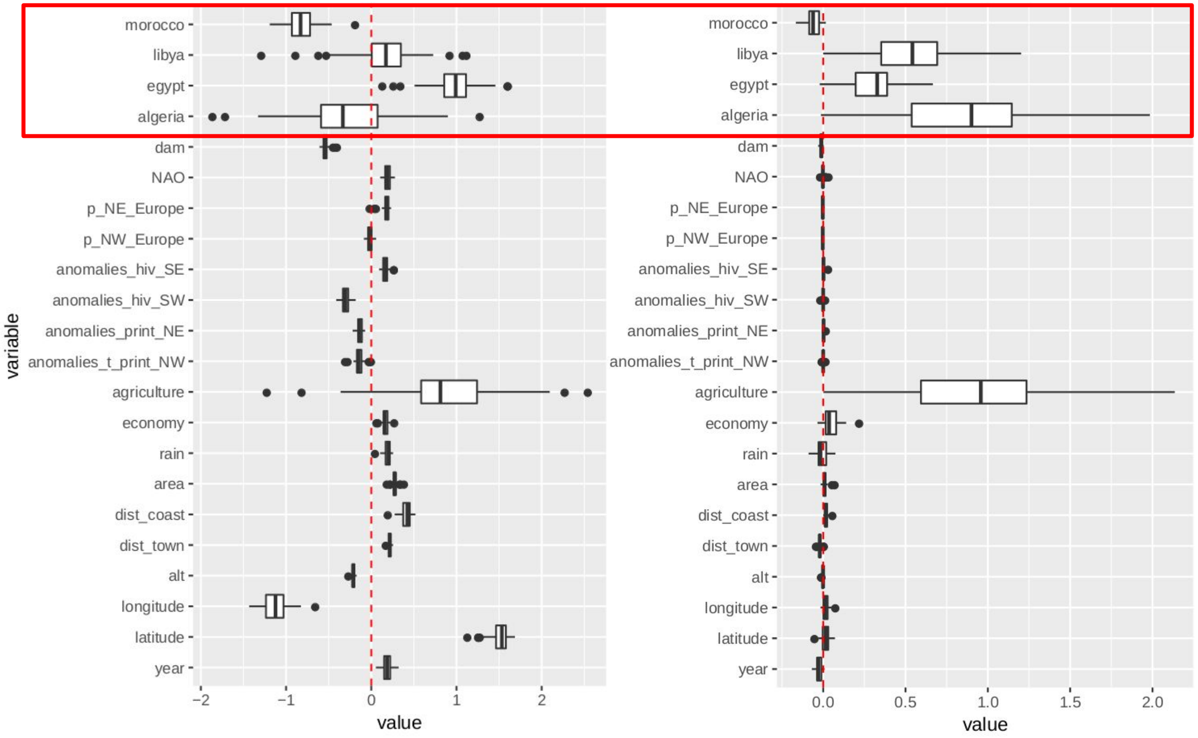


GLM



Lori

Feature Importance

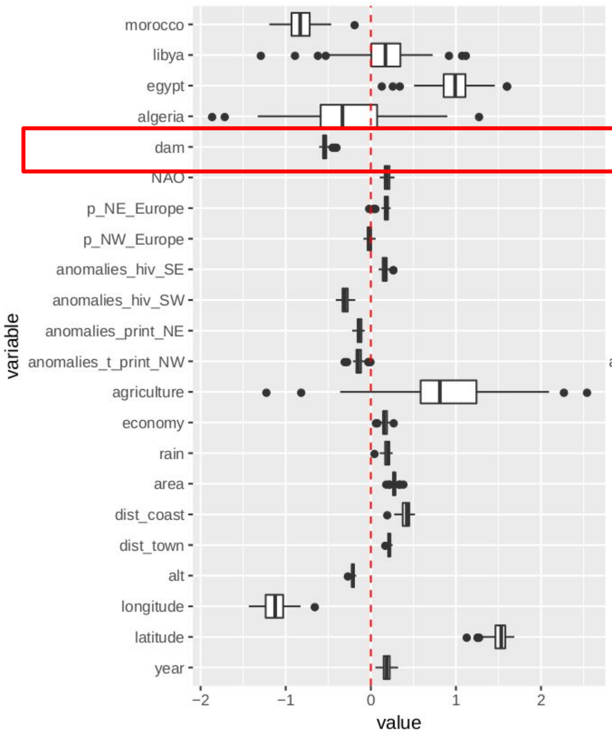


GLM

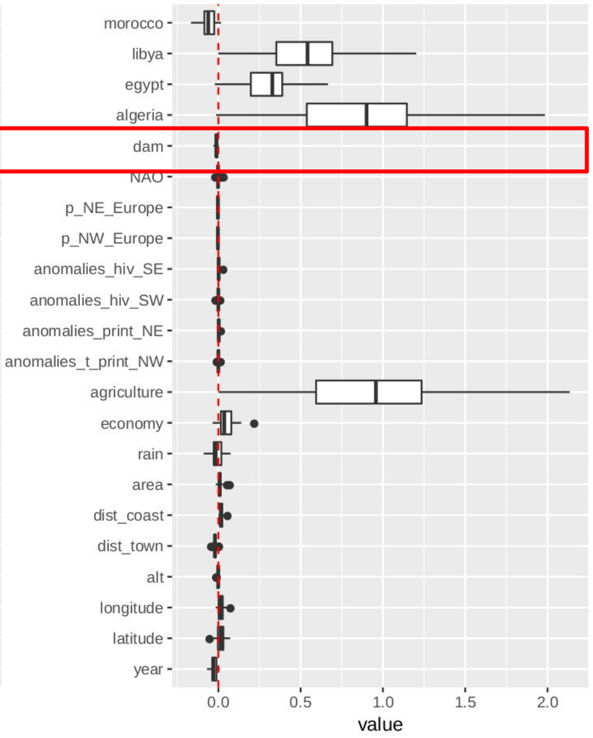
Lori

- Great influence of country on bird count

Feature Importance



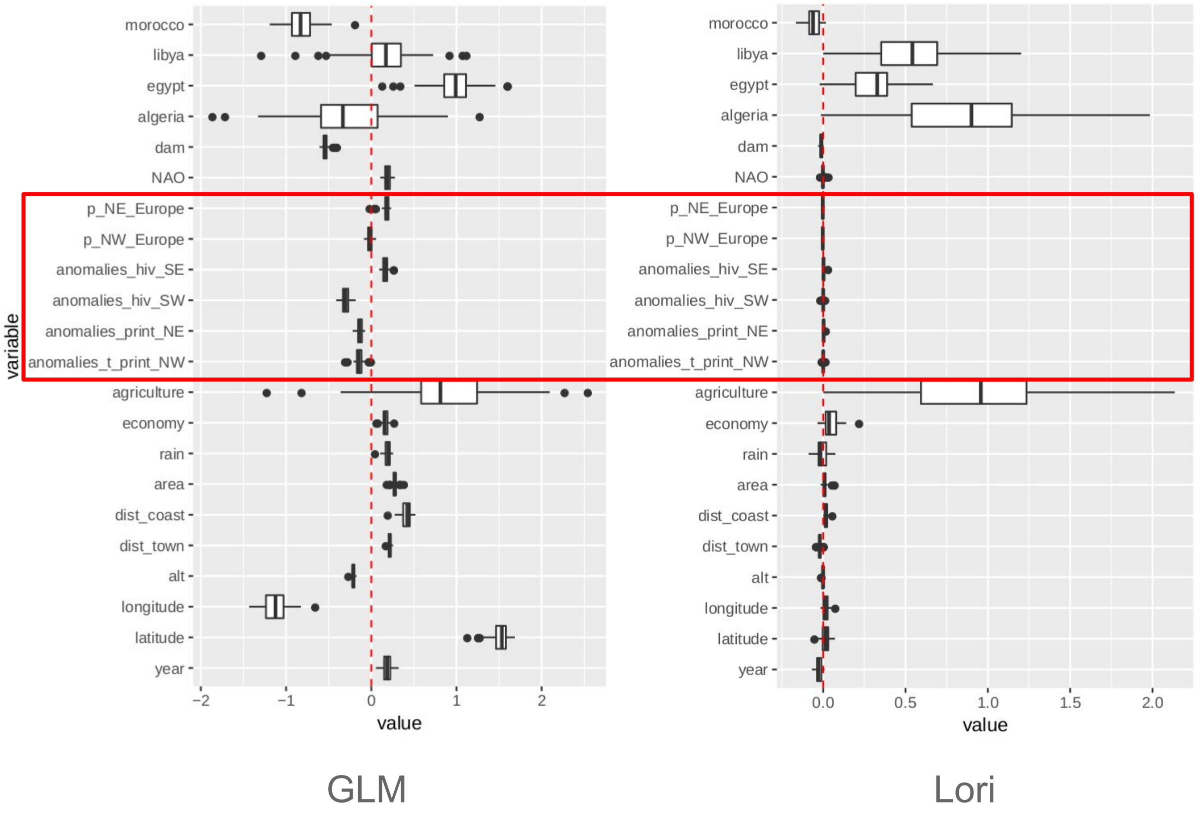
GLM



Lori

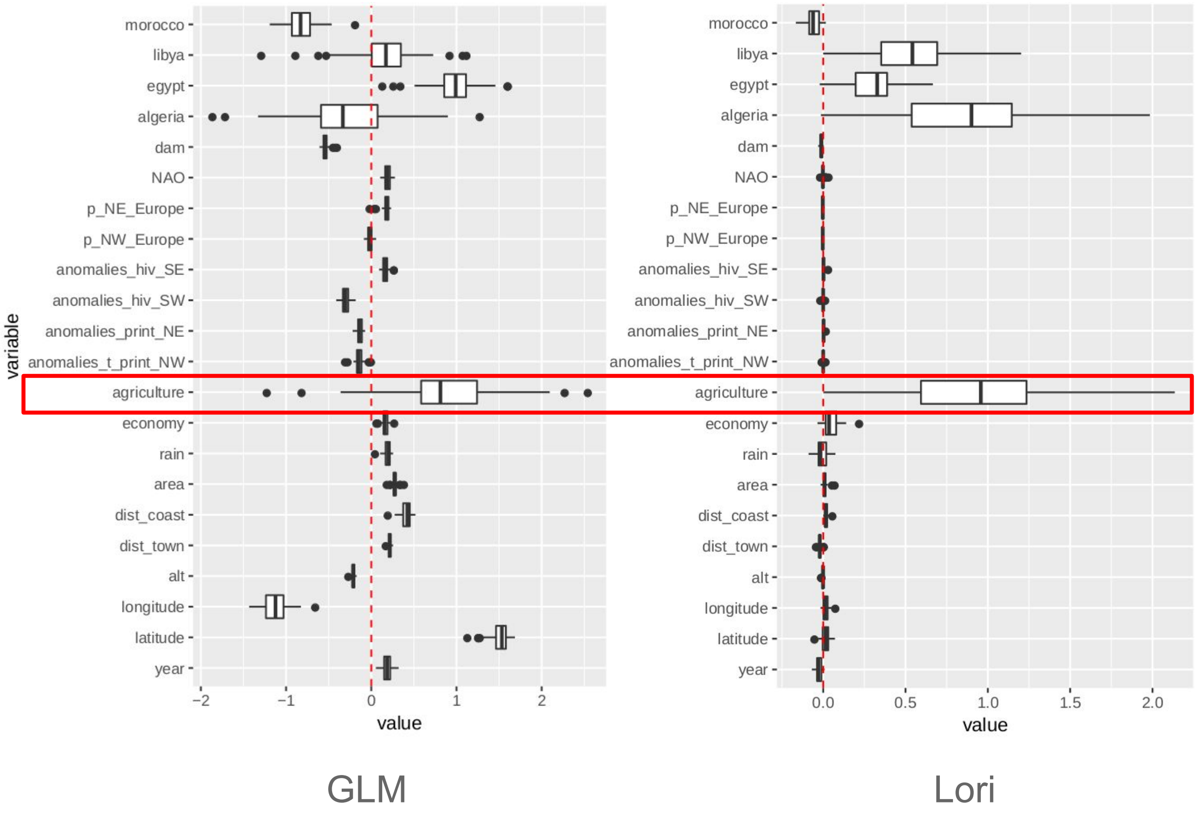
- Great influence of country on bird count
- Negative impact of dam covariate
  → Prefer natural wetlands

Feature Importance



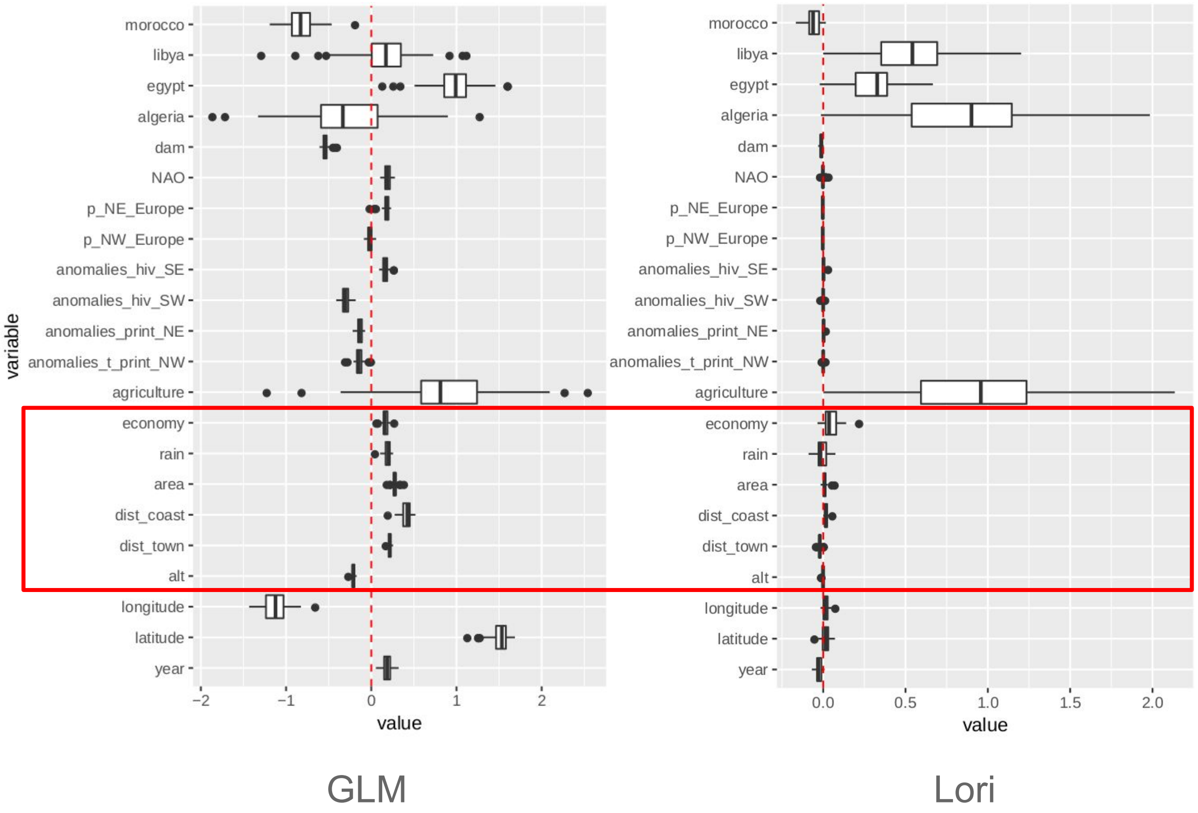
- Great influence of country on bird count
- Negative impact of dam covariate
  → Prefer natural wetlands
- Weather anomalies have almost no influence

Feature Importance



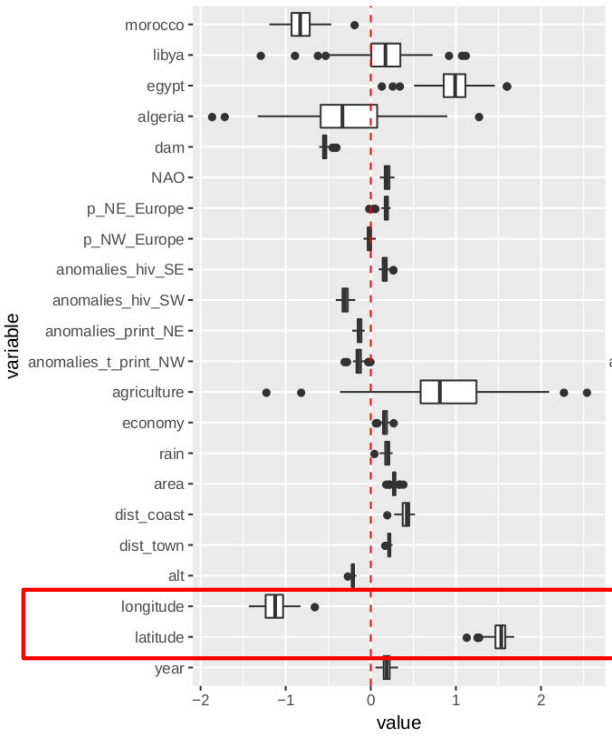
- Great influence of country on bird count
- Negative impact of dam covariate
  → Prefer natural wetlands
- Weather anomalies have almost no influence
- Positive impact of agriculture
  → More food available for birds

Feature Importance

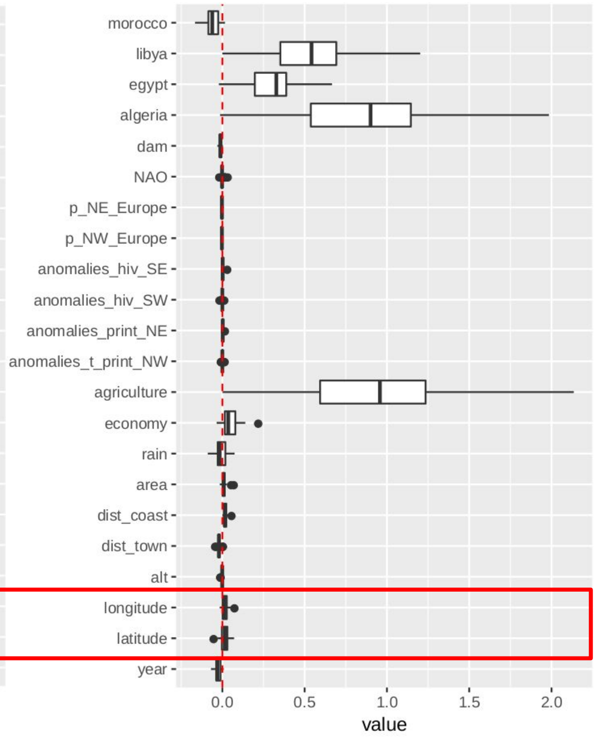


- Great influence of country on bird count
- Negative impact of dam covariate
  → Prefer natural wetlands
- Weather anomalies have almost no influence
- Positive impact of agriculture
  → More food available for birds
- Other covariates have small influence

Feature Importance



GLM

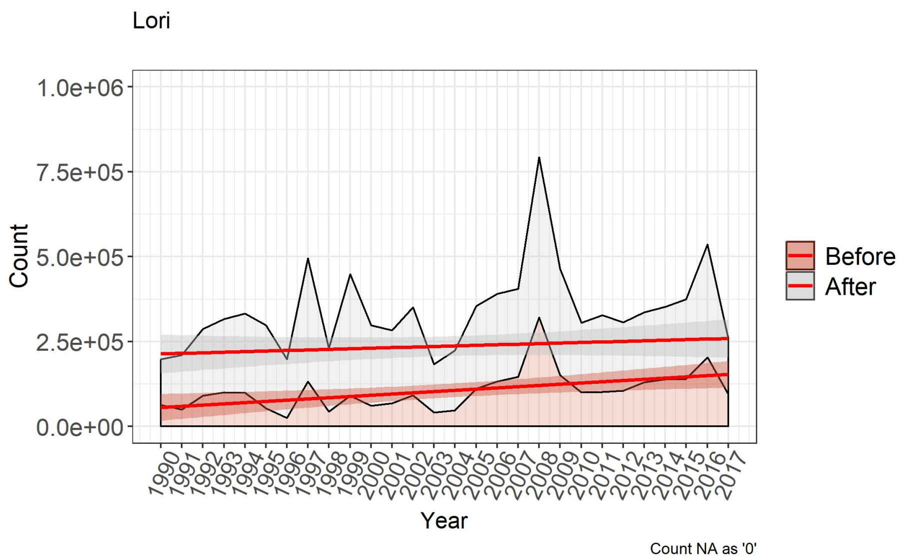


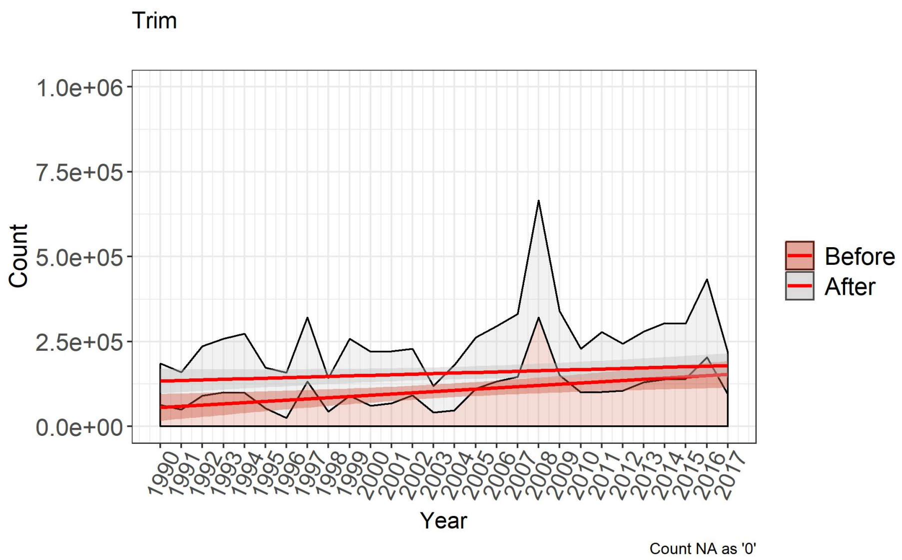
Lori

- Great influence of country on bird count
- Negative impact of dam covariate
  → Prefer natural wetlands
- Weather anomalies have almost no influence
- Positive impact of agriculture
  → More food available for birds

  Other covariates have small influence
- Location has great impact
  → Many possible explanations

Temporal Trend





- Matches shape of original data
- Difference in intercepts

⇒ Population of Eurasian Coot increasing

	Lori	Trim
Intercept	213,688	133,793
Slope	1,683	1,688

Conclusion

Eurasian Coot [6]

- Successfully imputed data better than baseline model
- Lori and Trim obtained best results
- Singled out important covariates which is useful for ecologists
- Determined that population of Eurasian Coot is increasing

Q&A

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

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