SURVIVAL ANALYSIS

Course project : A survival analysis on migratory return dates of white storks to Alsace, France

Datasets, scripts and report are available at:
https://github.com/JoanneAB/survivalAnalysis_StorksMigration

Joanne ADAM
Data ScienceTech Institut
joanne.adam@edu.dsti.institute

1. Introduction

Considered as the symbol of Alsace (South-East region of France), the white stork (figure 1) returns to Alsace every spring from winter migration in Africa. Every spring, the return of the white stork to Alsace, marks an important events of the season.

Figure 1: Picture of the white stork (left, [4]), illustration of Hansi "La Cigogne de notre Village est revenue" (right).

From a scientific perspective, the timing of migratory returns is known to be influenced by a combination of environmental factors, including temperature, wind conditions, and cloud coverage. Understanding and modelling these factors is of growing importance in the context of climate change as seasonal weather patterns may significantly alter bird migration phenology.

This study applies survival analysis to model the time until the first white stork observation of the year in Alsace, using data collected across the region. Survival times are defined as the number of days from the start of the year until the first stork is observed. Weather data as well as lunar phase information are used as covariates to explain the variability in return dates.

A Cox proportional hazards model is fitted to the data, and the proportional hazards assumption is evaluated and addressed through the introduction of time-varying coefficients.

2. Data collection and data processing

The aim of this study is to analyse the time until the first observation of the year of the first white stork in Alsace region, France after their return from migration in Africa. To construct the dataset, the geographical region of interest is divided into several sub-regions (figure 2). Survival times are the number of days, for each year in each sub-region, until the first stork has been observed. Additional data such as weather parameters or moon phase are added to the dataset.

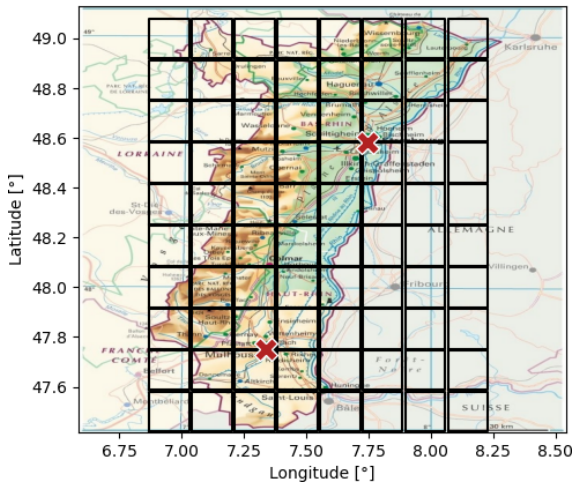

Figure 2: Map of Alsace, France region. Black squares represent the sub-regions in which the data for each first observations of the white storks are collected. Red crosses represent the meteorological stations.

2.1. Data collection

2.1.1 Observations of the white stork

Dates and locations of the observations of white storks in France are collected from the Global Biodiversity Information Facility (GBIF) website using the latin name: Ciconia ciconia (Linnaeus, 1758) [3]. The GBIF website compiles a total of 1990 distincts datasets. This collected dataset of white stork observations consists of 1883 088 observations (latitude, longitude and date) from January 1960 to October 2025.

2.1.2 Weather data

Weather data in Alsace are collected from the "Observation météorologique historiques France" [1]. Data are recorded every three hours from 1996 at two stations : Strasbourg and Mulhouse. Recorded data are temperature, wind speed, pressure, humidity, horizonatal visibility, nebulosity and cloud height.

2.1.3 Lunar phases

The moon is known to be a compass and landmark for the birds migrations. The fraction of illuminated moon could play an important role on the date of observation of the first stork in Alsace. Moon illumination fractions are collected from the Astronomical Applications Department [2].

2.2. Covariates engineering

2.2.1 Time-dependent covariate

Weather and more especially temperatures play an important role on the migratory return dates of the white storks. Indeed, a period of several warm days would lead to an increase in potential stork's food (insects, small mammals...) and an earlier return of the storks.

For each entry of the dataset, a new covariate is introduced to the dataset: 'isWarm'. This boolean covariate indicates whether four consecutive warm days (*i.e.* with a morning temperature greater or equal to 8°C) did occur the same year, before the first observation of the stork in the sub-region (*i.e.* the event). Regarding the time-dependency of this covariate, if warm days have been recorded, the entry of the dataset is split into two entries. Table 1 presents a subset of the dataset where the entry with `id=1` do not show four consecutive warm days (`isWarm=0`) before the event at day 88. The entry with `id=2` is divided into two lines with a first line that represents the time to the occurence of four consecutive warm days (at day 79, `isWarm=1`). The second line is the time between the warm days and the event (observation of the first stork).

id	start	stop	event	isWarm	...
1	0	88	1	0	...
2	0	79	0	1	...
2	79	89	1	1	...

Table 1

A new variable is added to the dataset to stratified by season (listing 1), allowing the survival function to vary freely across seasons (listing 1).

```
data$season <- factor(quarters(as.Date(data$stop)))
```

Listing 1: New stratified variable

2.2.2 Covariate truncations

Truncations on covariates are done to reduce the range of possible values, avoid extrem values and reduce the number of outliers.

White storks typically fly at altitudes between 500 and 1500m. Although some observations report altitudes of up to 4000m, these remain rare. The covariate `cloudHeight` was truncated to 3000 as the effect of higher values might be very limited.

White storks experience difficulties in flying with winds stronger than 3-5 m/s. The covariate `windSpeed` was therefore truncated at 5m/s.

2.2.3 Covariates pre-processing

Each non-boolean covariates (except `id`, `start`, `stop` and `event` covariates) are standardized in order to have a zero-mean and a standard deviation equals to 1.

Missing values, that corresponds to 276 values out of 10 605 total entries, are replaced by the mean of its covariate (*i.e.* 0). Missing values are mostly related to missing weather measurements.

Figure 3 presents the distribution of `temperature/stop` ratio of covariates as a function of `stop` and show some outliers with values greater than 0.1 and smaller than -0.2. These entries are removed from the database.

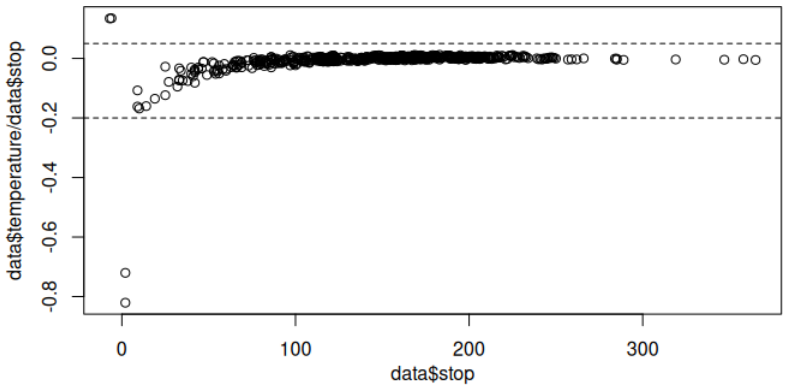

Figure 3: Distribution of `temperature/stop` as a function of `stop`. Horizontal dashed lines outline the presence of outliers.

2.3. The final dataset

A total of 641 observations were collected, each with 16 covariates : `start`, `stop`, `event`, `isWarm`, `latitude`, `longitude`, `temperature`, `windSpeed`, `pressure`, `humidity`, `visibility`, `nebulosity`, `cloudHeight`, `moonPhase` and `season`. The distribution of the standardized covariates are presented in figure 4. More informations and figures are available in the R markdown script.

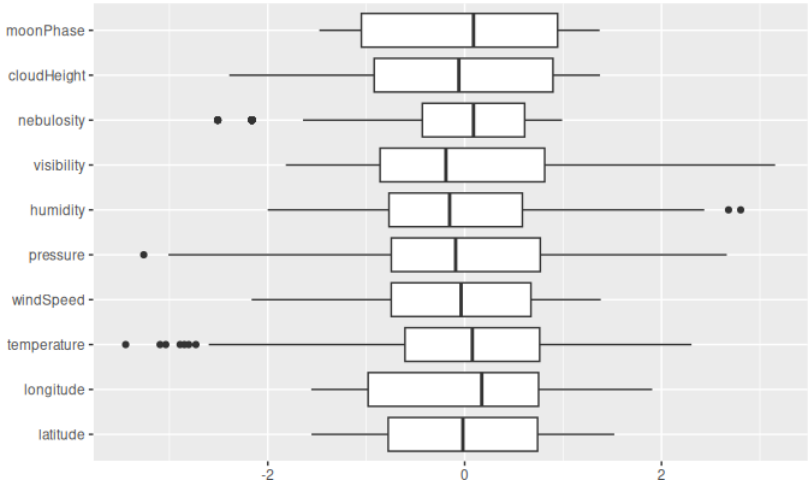

Figure 4: Distribution of the standardized covariates.

3. Survival analysis methods

3.1. Identification of significant covariates

Regarding the correlation matrix presented in figure 5, the `isWarm`, `temperature`, `windSpeed` and `cloudHeight` seem to play an important role in the survival analysis. Cox and ANOVA tests (see listing 2 for extracts of the `R` code) are used to select the significant covariates in order to preceed with analysis. The result of this ANOVA analysis using `isWarm`, `temperature`, `windSpeed` and `cloudHeight` covariates show a p-value smaller than 0.05, which indicates that we rejext the null hypotheses. The difference between the two models are significant and more covariates are necessary to explain the data.

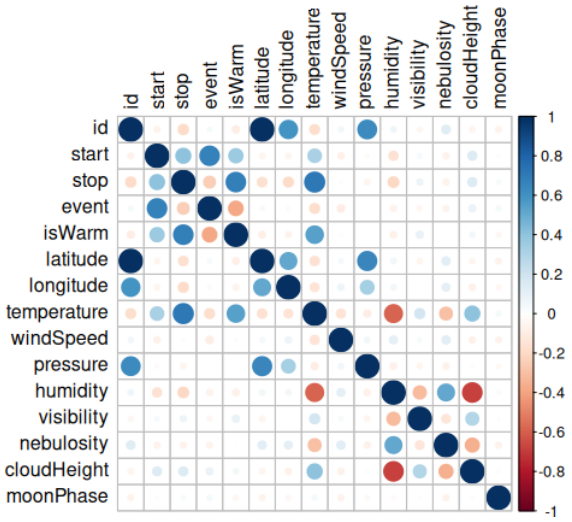

Figure 5: Correlation matrix of the covariates used in this study

```
MtempWarmWindCloud <- coxph(Surv(start, stop, event)~temperature + isWarm + strata(season
    ) + windSpeed + cloudHeight, data=data)

Mall <- coxph(Surv(start, stop, event)~isWarm + temperature + windSpeed + pressure +
    humidity + visibility + nebulosity + cloudHeight + moonPhase + strata(season), data=
    data)

anova(MtempWarmWindCloud, Mall)
```

Listing 2: Cox and ANOVA analysis on a selection of covariates

Stepwise model selection is applied to the dataset using the AIC (Akaike Information Criterion) to select the most informative covariates and find the simplest model that would best fit the data. Results of the R code are presented in the listing 3 and show that the highly significant covariates (with p-values < 0.01) are : `longitude`, `isWarm`, `temperature`, `humidity`, `visibility`, `windSpeed`, and `cloudHeight`. `pressure` is a significant covariate (with p-value < 0.05). `Nebulosity`, `latitude` and `moonPhase` covariates show p-values greater than 0.05 and are not significant covariates. The model considering only the highly significant model is named `Mbest`. ANOVA test is performed to compare `Mbest` model and the model considering all the covariates. Results show a p-value much greater than 0.05 showing that we reject H1 hypothesis. Both models are significantly equals and the covariates of `Mbest` are sufficient to explain the data.

```
coxph(formula = Surv(start, stop, event) ~ latitude + longitude +
    isWarm + temperature + windSpeed + pressure + humidity +
    visibility + nebulosity + cloudHeight + moonPhase, data = data)

  n= 641, number of events= 376

                coef exp(coef) se(coef)       z Pr(>|z|)
latitude     0.04566   1.04671  0.08351   0.547 0.584588
longitude    0.16842   1.18343  0.06528   2.580 0.009880 **
isWarm      -2.07290   0.12582  0.23488  -8.825  < 2e-16 ***
temperature -1.18656   0.30527  0.09937 -11.941  < 2e-16 ***
windSpeed   -0.19953   0.81912  0.05703  -3.499 0.000467 ***
pressure    -0.14994   0.86076  0.07606  -1.971 0.048687 *
humidity    -0.40031   0.67011  0.09945  -4.025 5.69e-05 ***
visibility  -0.22837   0.79583  0.06243  -3.658 0.000254 ***
nebulosity  -0.10825   0.89740  0.06749  -1.604 0.108716
cloudHeight  0.22859   1.25683  0.08634   2.648 0.008106 **
moonPhase    0.04129   1.04216  0.05335   0.774 0.438883
```

Listing 3: Extracts of the output of the R code of the stepwise variable selection.

The estimated Hazard Ratios of each covariates are presented in figure 6 for the highly statistically significant covariates. `CloudHeight and longitude` covariate shows a harzard ratio greater than 1 (HR=1.24 and 1.17 respectively) which indicates that the hazard increases when the cloud height and longitude increases. The thicker the cloud layer, the higher the risk.

Covariate `isWarm` shows the strongest effects with a HR value or 0.11. Event with warm days before the observation of a white stork have greatly reduced hazard.

Relatively strong effect is associated to the covariate `temperature` with a HR value of 0.32. Higher temperature is associated with considerable lower hazard.

With a hazard ratio close the 1 for the `pressure, widSpeed, humidity` and `visibility` covariates, these weather parameters seem to have moderate effect on the modelisation of the event (*i.e.* the observation of the white stork).

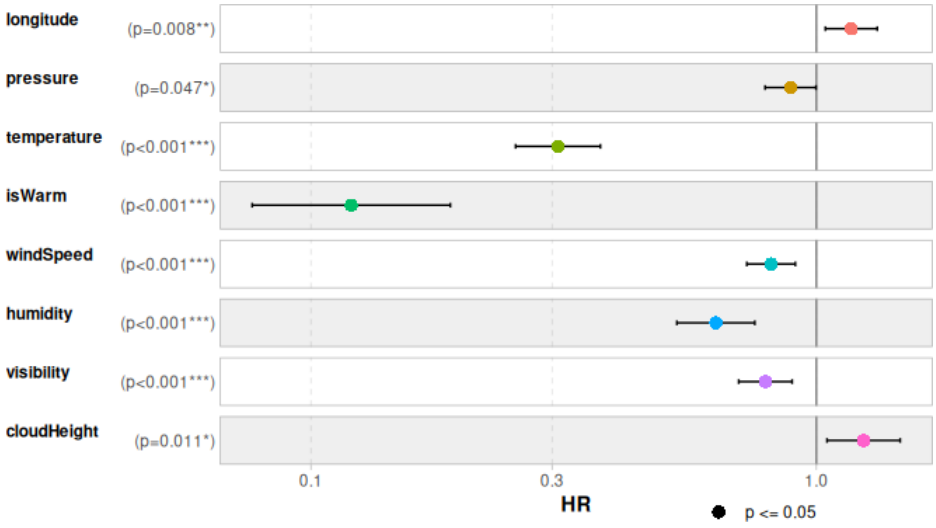

Figure 6: Hazard ratios plots of significant covariates

3.2. Evaluation of time-dependent risks

Schoenfeld residuals test is performed to verify the proportional hazards (PH) assumption that requires that the hazard ratios remain constant over time. A correlation of the scaled Schoenfeld residuals with time (*i.e.* a variation of β over time for each covariate) suggests that the effect of the covariates would change over time.

Results of the Schoenfeld test for the `Mbest` model are presented in listing 4 and figure 7 and show that the PH assumption is violated for all covariates but `visibility`. Indeed, p-values of `longitude`, `latitude`, `pressure`, `humidity`, `temperature`, `windSpeed` and `isWarm` covariates are highly statistically significant (p-value < 0.01). `cloudHeight` covariate show a statistically significant violation of the PH assumption (p-value < 0.05).

```
               chisq  p-value
longitude    18.55   1 1.7e-05 (< 0.01)
latitude     24.64   1 6.9e-07 (< 0.01)
pressure      9.55   1 0.00199 (< 0.01)
temperature 13.98   1 0.00019 (< 0.01)
isWarm        7.46   1 0.00629 (< 0.01)
windSpeed     7.05   1 0.00793 (< 0.01)
humidity     16.21   1 5.7e-05 (< 0.01)
visibility    1.57   1 0.21018 (> 0.05)
cloudHeight   5.49   1 0.01915 (< 0.05)
```

Listing 4: Output of the R code for Schoenfeld test (`cox.zph(Mbest)`).

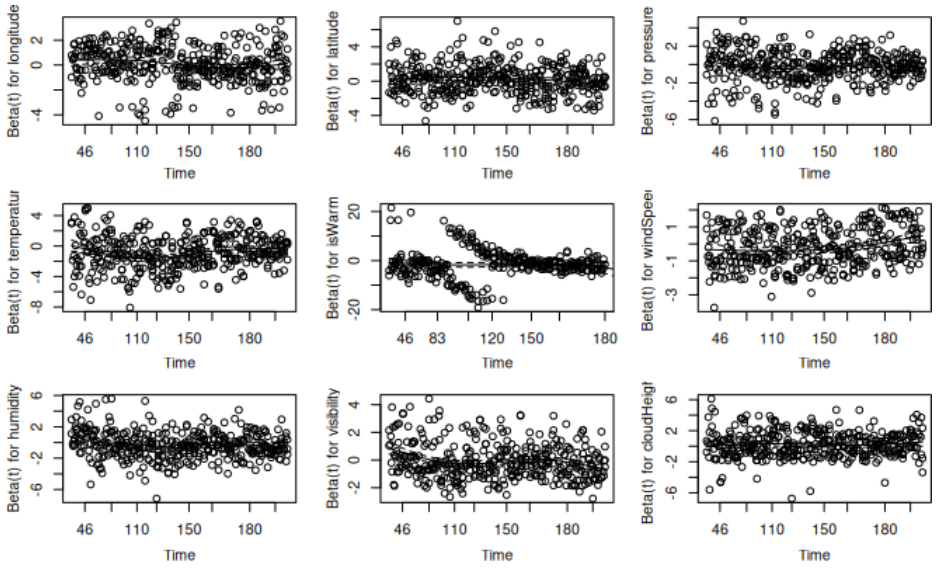

Figure 7: Scaled Schoenfeld residuals over time for each covariate of the `Mbest` model.

Figure 7 shows a clear violation of the PH asusmption for `isWarm` covariate. Indeed, the sharp and strong variation of Beta with time shows a clear time-dependancy of the covariate. Moreover, the distribution of Beta(t) for `temperature` and `windSpeed` shows a slight upward slope while the distrubution for `humidity` and `latitude` show a slight downward slope. Both indicating time-dependancy of the covariates.

3.3. Residuals between observed and expected events

Results of the Schoenfeld test show that the hazard ratios are not constant over time and suggest time variation of many covariates. Time-dependency are added to the cox function as presented in the listing 5. Several functions have been tested and the `x/t` function presents better results which are presented in this report. More details on this analysis are presented in the R code file.

```
Mbest_tt <- coxph(Surv(start, stop, event)~temperature + tt(temperature) +
      isWarm + tt(isWarm) +  windSpeed + tt(windSpeed) + humidity + tt(humidity) +
      visibility + cloudHeight + strata(season), data=data, tt=function(x,t,...) x/t)
```

Listing 5: Modelisation of time variations.

Martingale residuals, showing the difference between the observed number of events and the expected number of events for `Mbest` (left) and `Mbest_tt` (right) models are presented in figure 8. Figure 9 presents their histograms and boxplot distributions. Martingale residuals are partially taken into account in the "Time-varying best model". Indeed, the dispersion of the residuals decreases and most of the residuals are closer to 0. The number of censored observations (over-predicted risk), with residuals lower than -1, are better constraints by the time-varying model. However, the model still fails to predict some measurements and did not fully captured the hazard variation. Indeed, many measurements residuals remain close to 1 which indicates that these events have underestimated risks by the model.

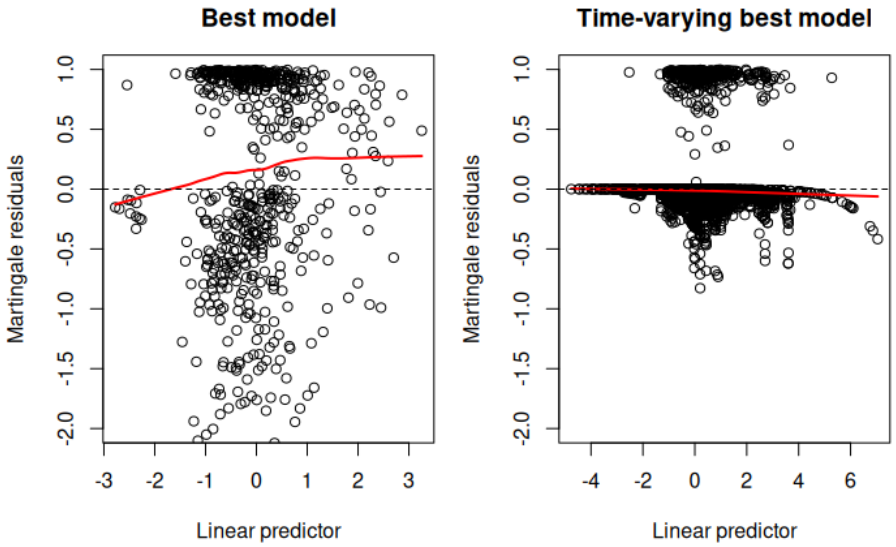

Figure 8: Martingale analysis on `Mbest` (left) and `Mbest_tt` (right) models. Red lines show a line smoothing of the scattered residuals data.

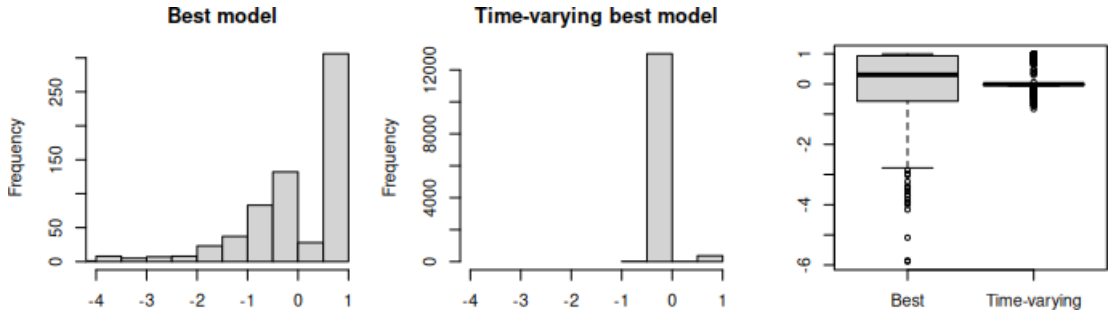

Figure 9: Histograms of the Martingale residuals on `Mbest` (left) and `Mbest_tt` (center) models. Left: boxplot representation of the residuals for both models.

4. Discussion

4.1. Time-dependency of the covariates

Despite the introduction of time-varying coefficients in the `Mbest_tt` model, some temporal variation remains unexplained, particularly for `isWarm` covariate. The Schoenfeld residuals show sharp and non-linear variations in the effect of `isWarm` between $\sim 80 - 120$ days that corresponds to the Spring season that is the key season for warm temperatures and migratory returns. `x/t` transformation is insufficient to fully capture temporal pattern. The binary nature of `isWarm` covariate may be a limitation and

continuous measures of cumulative warm-days might better reflect the mechanisms behind white stork returns.

`isWarm` covariate captures informations on whether four consecutive warm days with a morning temperature greater than 8°C did occur before the event. Changing the number of consecutive days or the temperature from 8°C to 10°C did not meaningfull improved the model fit. This suggests that the threshold matters less than the overall weather conditions of the seasons. Alternative formulations to describe a warmer-than-usual season should be explored.

The stratification variable `season` did not significantly improve the modeling, suggesting that seasonal variation is already partially captured by the weather covariates.

4.2. Limitations

The dataset used in this study show some limitations and could be extended and refined to better analyse the migratory return dates of the white storks in Alsace.

The dataset only relies on weather records at two stations (Strasbourg and Mulhouse), which may not fully represent spatial variability across Alsace's sub-regions. Additional records at weather stations distributed across the region would help capture local meteorological variability more accurately.

Missing values represent ∼2.6% of the dataset which may not be negligeable regarding the small amount of data collected. These missing values were replaced with covariate means, which may introduce bias to the dataset.

5. Conclusion

This study applied survival analysis to model the return dates of white storks (*Ciconia ciconia*) to Alsace, France, using observational data of white storks combined with historical weather and lunar phases records.

The analysis identified six statistically significant covariates: `isWarm`, `temperature`, `humidity`, `visibility`, `windSpeed` and `cloudHeight`. Among these covariates, the occurrence of consecutive warm days (*i.e.* `isWarm` boolean covariate) and ambient temperature showed the strongest effects, with hazard ratios well below 1, confirming that warmer early-season conditions are strongly associated with earlier stork arrivals. In contrast, increased cloud height was associated with a delayed first observation.

The proportional hazards assumption was found to be violated for most covariates, highlighting time-dependency of the weather conditions to stork migration. Introducing time-varying coefficients transformation partially addressed this issue, reducing residual dispersion and improving the fit. However, the model did not fully capture the hazard variation, suggesting that more temporal formulations require further investigation.

Overall, this work demonstrates that weather conditions are key factors of white stork return in Alsace. Going further, the model could be used with up-to-date weather data to forecast the arrival date of the first white stork in Alsace in the future.

6. References

[1] Observation météorologique historiques France (SYNOP). https://www.data.gouv.fr/datasets/archive-synop-omm. Accessed: 2026-01-15.

[2] Astronomical Applications Department, U.S. Navy. Fraction of the Moon illuminated. https://aa.usno.navy.mil/data/MoonFraction. Accessed: 2026-01-15.

[3] GBIF.org. The global biodiversity information facility. URL https://www.gbif.org/occurrence/download/0000635-251009101135966. Accessed: 2026-01-15.

[4] Oiseaux.net. White Stork. https://www.oiseaux.net/birds/white.stork.html. Accessed: 2026-01-15.