Birds as Environmental Indicators

Līva Freimane, Quang Minh Do and Daniel Zirat

October 2022

1 Introduction and background

Birds are recognized as one of the most important indicators of the environment, because they are sensitive to habitat change and because they are easy to measure. Ecological indicators can carry out ecosystem monitoring to preserve and manage the natural environment. Since it is impractical to monitor all ecosystem components, a few individual species or groups of species can be used as indicators of wider conditions.

Indicator species have been used in practical management at the local, regional, national and international levels. Birds have been utilized as indicator species by government agencies. Several studies show that birds may be used as indicators of conditions encountered in ecosystem, at both local and regional spatial scales. [2]

Changes in bird populations are often the first indication of environmental problems. Whether ecosystems are managed for agricultural production, wildlife, water, or tourism, success can be measured by the health of birds. A decline in bird numbers tells us that we are damaging the environment through habitat fragmentation and destruction, pollution and pesticides, introduced species or they can be reliable indicators of nutrient status and the abundance of other organism [1].

Birds are a part of the balance of nature. Birds provide insect and rodent control, plant pollination, and seed dispersal which result in tangible benefits to people. Birds live in an environment that is subject to both regular and irregular fluctuations, and bird populations respond to these changes in predictable ways. The cause and effect link between an environmental change and birds are direct and simple one.

For example distribution and biomass of the seagrass Zostera japonica decreased sharply, which was when migrating swans arrived [4].

In this paper we will mainly focus on the prediction and the preprocessing of the swan data. Incomplete data in bird count is a common problem and prevalent in basically all data sets. Therefore we will describe in detail how we deal with missing values. Afterwards we will provide a method for predicting new unforeseen data, which can be used for further use.

2 Data

The arrival and departure dates and the daily maximum populations of migrating swans (*Cygnus cygnus*) on the Asadokoro tidal flat, Hiranai town, Aomori Prefecture, Japan, were recorded by elementary school students for more than 50 years between 1956 and 2010. This long history of observation unfortunately came to an end with the closure of the elementary school in 2012. [5]

The data (available at [6]) were acquired by visual observations and counting. We used the raw version of the data of 5899 entries over 10 columns: year, month, day, number of individuals (AM) and (PM), number of adults (AM) and (PM), number of young swans (AM) and (PM) and remarks. We used the first three columns to make a date-time index and predicted the number of individuals (AM), because it had the least amount of missing values. The rest of the columns were not used, because their values were missing or they didn't give any useful information.

For training our models we didn't use the whole data set because of some misreported values in some years between 1966 and 1977.

2.1 Imputation

Data imputation is a challenging part due to the vast number of missing values in our data set. Since our objective was to predict future daily swan population, we need a data set of historical daily swan population. We expected to gather 16,142 data points for daily count of swans from 17/10/1966 to 26/12/2010. However, there are only 5774 entries having recorded number of swan during the morning because the swan counting activity was not performed every day. As a result, we need to devise a proper data imputation method to impute 10,368 missing values, which is approximately twice as much as the number of recorded ones, while retaining the data distribution.

The data imputation process comprises three main steps. Firstly, for each month with recorded swan population in the period (from 10/1996 to 12/2010), we impute daily missing data by interpolation. Next, we need to impute missing data of the months that have no recorded values. We calculate the average number of swan for each month the period(from 10/1966 to 12/2010) and impute the months without any recorded swan population by interpolation. All data points belonging to these months are then imputed with the monthly average values. Finally, since data points from June to September are either not recorded or almost zero, we assume that the swans usually arrive at Asadokoro tidal flat in October and depart in May the next year. Thus, we replace all data points from June to September with zeros, implying that there are no swans at Asadokoro during this period every year.

2.2 EDA

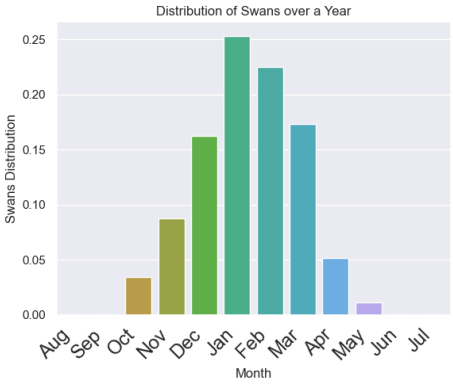


Figure 1: Distribution of swans over a year

The distribution of the swans can slightly vary from one year to another, but overall it is very similar. In Figure 1 we can see the distribution of the swans over a year. We can see that most of the swans can be viewed from December until March. For the period from June until September there are no swans in the area.

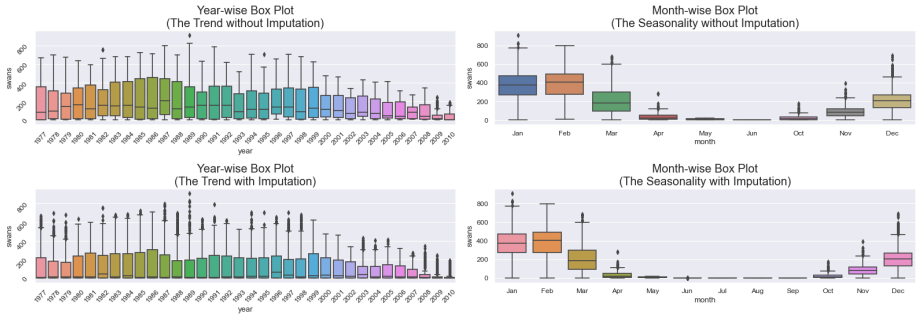


Figure 2: Trend of the birds is unchanged after imputation

The box plots make the year-wise and month-wise distributions evident. It is important to note that the imputation does not change the trend of the distribution and further can be used for the training the model.

3 Models and Experiments

Since our data is time-series data, we had to use a model that would take as an input the date-time index and predict the number of swans as an output.

We experimented with two different time series forecasting models, namely Recursive Auto-regressive and Prophet, and made comparison between them. We will give an overview of the models in this section.

3.1 Recursive Auto-regressive Forecasting

We used Recursive Auto-regressive forecasting. The forecasting process consists of predicting the future value of a time series by modeling the series based on its past behavior. To achieve this we used the `skforecast` library.

Since the value t_{n-1} is required to predict t_n, and t_{n-1} is unknown, a recursive process is applied in which each new prediction is based on the previous one. This process is known as recursive forecasting. [7]

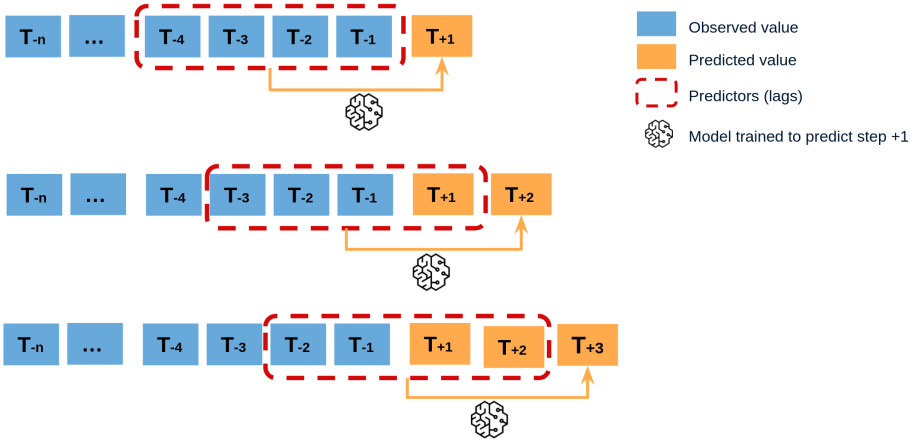
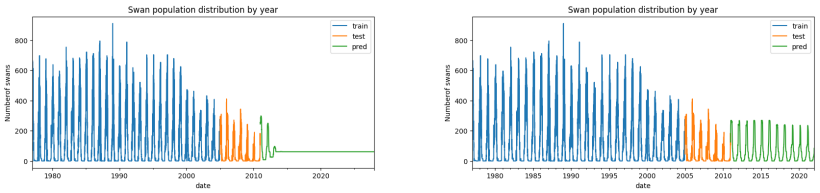


Figure 3: Recursive multi-step prediction process diagram to predict 3 steps into the future using the last 4 lags of the series as predictors. [7]

Together with the imputed values, the total number of entries in the data set is 16142, of which we used 12413 (starting from 1977-01-01) and we wanted the model to take into account a long period of time to better capture the trend. With the `ForecasterAutoreg` class, a model is created and trained from a Random Forest Regressor with a time window of 5000 lags (measured in days). Figure 4 shows how model predictions of 365 and 5000 compare - even though the MSE on the test set is lower for `lags`=365, we see that the model ends up predicting a constant value for all months in the future.



(a) `lags`=365, MSE on test set: 4433.46

(b) `lags`=5000, MSE on the test set: 4722.03

Figure 4: Comparison of the `lags` parameter of Auto-regressive model

We did some hyperparameter tuning using grid search. In order to identify the best combination of lags and hyperparameters, the `skforecast` library provides the `grid_search_forecaster` function. It compares the results obtained with each model configuration. We looped over `lags, max_depth` and `n_estimators`.

The pipeline that gave the best results with respect to mean squared error turned out to be `RandomForestRegressor(max_depth=10, n_estimators=200)`.

3.2 Prophet

We also implemented a time series forecasting model with Prophet, a modular regression model developed by Facebook [3].

Prophet time series model can be decomposed in to three components, namely trend, seasonality and holiday, each of which is configurable. Prophet offers two types of trend functions. Logistic function models time series with decreasing rate of change. By contrast, piece-wise linear function models time series with multiple changepoints and different rates of change. Rate adjustment is determined by variable δ whose prior belongs to Laplace distribution with zero mean and τ scale. The changepoint prior scale τ can be configured to control the model's trend flexibility. Seasonality is periodic changes of time series on different levels such as weekly, monthly and yearly. Seasonality can be modeled as an additive or multiplicative component. Seasonality is added to trend function if it is additive, or multiplied with the trend otherwise. The user can also model the effect of holiday on time series by providing a list of events. Each holiday is assigned with a parameter κ controlling the impact of the corresponding holiday on the time series.

We experimented with different settings, such as logistic trend vs linear trend, different values of changepoint prior scale, and additive vs multiplicative seasonality. We used yearly seasonality because the swan population data shows strong yearly periodic changes. We did not use weekly seasonality and holiday because we assumed that swan population does not have weekly periodic changes and is not affected by holidays. After conducting many experiments, we found a set of hyperparameters that yields the lowest mean squared error: linear trend, changepoint prior scale = 0.001 and multiplicative seasonality. The trend and seasonality captured by the model are illustrated in Figure 5.

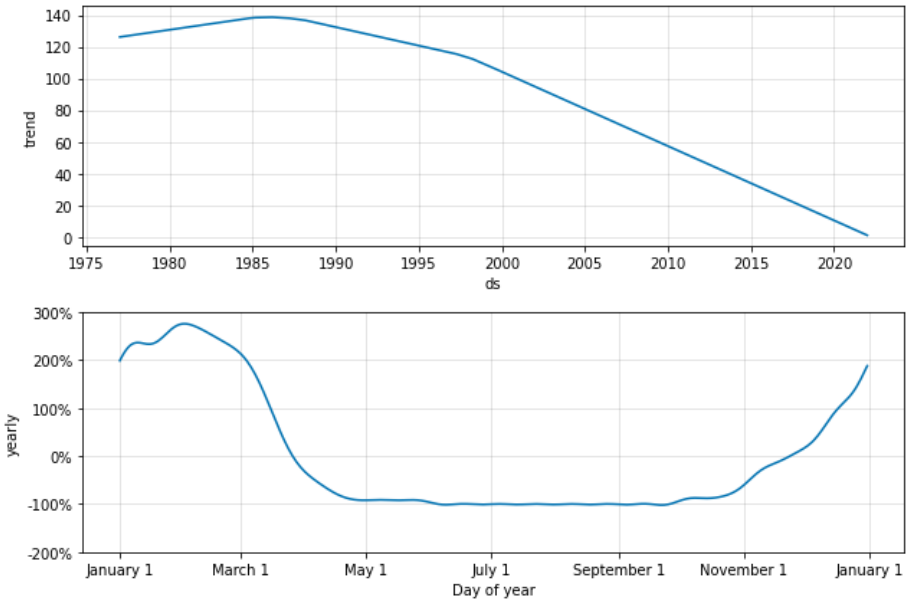


Figure 5: Trend and seasonality of swan population captured by the Prophet model

3.3 Results

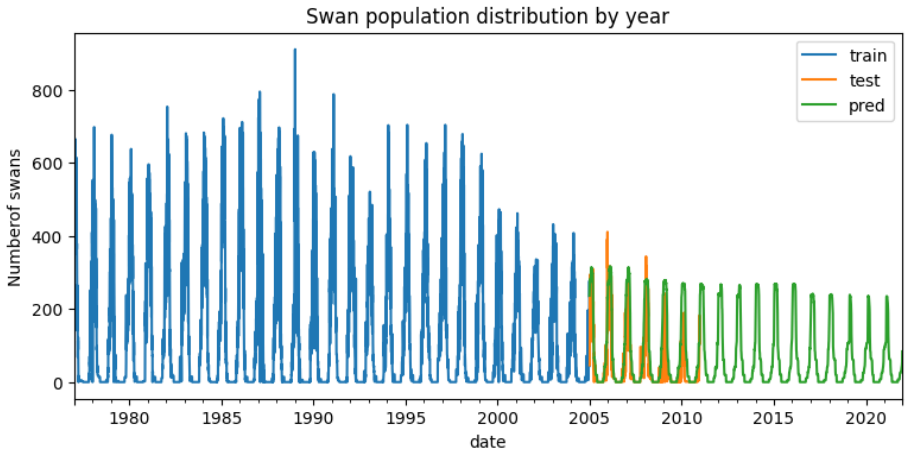


Figure 6: Predictions of Recursive Auto-regressive model on testing set and in the future

The Recursive Auto-regressive Forecasting model (Figure 6) achieves a MSE of 4722.03 on the test set. The model manages to capture the seasonality of the data and predict realistic values for the number of swans till 31/12/2021 factoring in a small decrease.

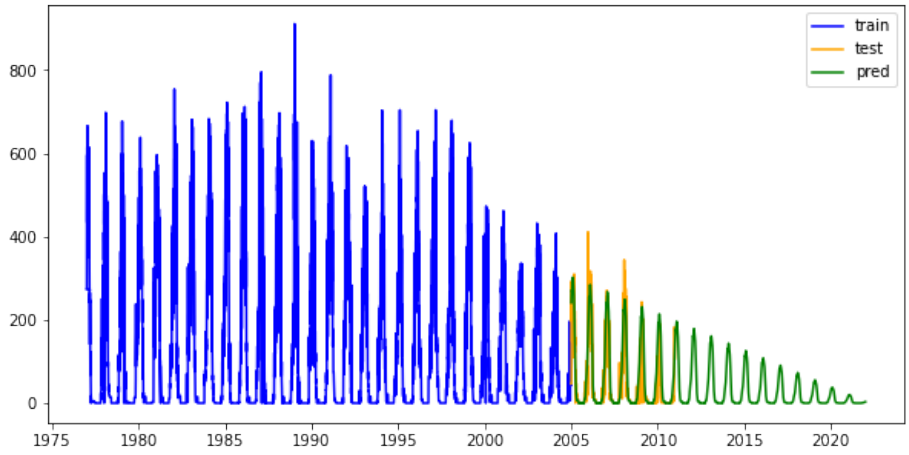


Figure 7: Prophet prediction (green) compared with the training (blue) and testing (orange data)

The Prophet model (Figure 7) achieves a Mean Squared Error of 2453.31 on the test set. It predicts a negative trend of swan population from 01/01/2005 to 31/12/2021.

4 Conclusion

We have the results of two models - Recursive Auto-regressive model and Prophet (Figure 6 and Figure 7) with corresponding MSE values of 4722 and 2453.

Both models capture the seasonality of the distribution of population of swans in Asadokoro, and both predict a negative trend, however Prophet predicts a faster decline than Recursive Auto-regressive model.

If we only consider the MSE, we might say that Prophet is best out of the two models, however, if we plug in some knowledge about the historical distributions of swans in Asadokoro, we might not want to believe that in 20 years there is no migration of the swans in the area, as suggested by Figure 7. This makes us think that the predictions of Recursive Auto-Regressive model might paint a better picture of the real situation.

References

[1] Juan Amat and Andy Green. "Waterbirds as Bioindicators of Environmental Conditions". In: Oct. 2010, pp. 45–52. ISBN: 978-1-4020-9277-0. DOI: 10.1007/978-1-4020-9278-7_5.

[2] Eduardo Alexandrino et al. "Bird sensitivity to disturbance as an indicator of forest patch conditions: An issue in environmental assessments". In: *Ecological Indicators* 66 (July 2016), pp. 369–381. DOI: 10.1016/j.ecolind.2016.02.006.

[3] Sean J. Taylor and Benjamin Letham. "Forecasting at Scale". In: *The American Statistician* 72.1 (2018), pp. 37–45. DOI: 10.1080/00031305.2017.1380080. eprint: https://doi.org/10.1080/00031305.2017.1380080. URL: https://doi.org/10.1080/00031305.2017.1380080.

[4] Fumiyuki Sato et al. "The influence of migratory birds on the distribution of the seagrass Zostera japonica". In: *Botanica Marina* 63.6 (2020), pp. 521–525. DOI: doi:10.1515/bot-2020-0045. URL: https://doi.org/10.1515/bot-2020-0045.

[5] Masaki Ogata, Takeshi Mitsuya, and Yoshiyuki Tanaka. "Data on swan arrival, departure, and population size on the Asadokoro tidal flat, Aomori Prefecture, Japan, from 1956 to 2010". In: *Data in Brief* 35 (2021), p. 106825. ISSN: 2352-3409. DOI: 10.1016/j.dib.2021.106825.

[6] Masaki Ogata, Takeshi Mitsuya, and Yoshiyuki Tanaka. *Raw data on swan arrival, departure, and population size on the Asadokoro tidal flat, Aomori Prefecture, Japan, from 1956 to 2010*. Tech. rep. 2021. DOI: 10.17632/g9tcw92bgy.1.

[7] Javier Escobar Ortiz Joaquín Amat Rodrigo. *Skforecast: time series forecasting with Python and Scikit-learn*. URL: https://www.cienciadedatos.net/documentos/py27-time-series-forecasting-python-scikitlearn.html.