Introduction

Fish are essential to the biodiversity of waterways, contributing to the cleanliness and overall health of their ecosystems. Thus, understanding fish migratory patterns provides important ecological insights, and ensuring safe passage of fish during mating season is vital in ensuring the well-being of these environments.

From early March to late May, thousands of fish migrate through canal systems and waterways across Europe. However, in Utrecht, The Netherlands, a lock at the center of the city prevents fish from passing, leaving them vulnerable to predators and wasting time and energy necessary for finding breeding areas. To combat this, the city operates a "fish doorbell", an underwater livestream camera, that allows millions of viewers worldwide to alert the lock keeper that fish are waiting at the lock.

Still, depending on the time of day and season, hours can go by without any fish appearing at the doorbell, making it difficult for viewers to remain engaged. We aim to address this by proposing two models for analyzing fish migratory patterns around the canal. In turn, we hope these can aid in alerting followers when there may be a fish that needs their help, affording fans a higher chance of being able to participate and increasing the likelihood that a fish is noticed.

Because the doorbell operates 24/7 during mating season, there is a large breadth of video data available for tracking fish movement around the lock. Supplemented with historical data from other canal and waterway monitoring systems, we're able to gather the larger picture on fish migratory patterns throughout mating season and across the years. This paper suggests using these data in twofold:

1. To train a convolutional neural net (CNN) to identify when fish or other animals are at the gate.
2. To conduct a time-series analysis to understand trends in fish migration throughout mating season and analyze how these trends are changing.

Literature Review

Multiple research papers show that CNNs are the most accurate and efficient method of detecting animal species in wildlife footage. One study in particular from the University of Zilina tested several different image detection machine learning models with animal identification and found that, among their studies, CNN definitively was the most accurate and efficient (Trnovszky et al, 2017). Multiple different studies have also been done focusing on using CNNs to specifically classify different species of fish, going as far as to attempt to experimentally deduce the most effective CNN architecture for such a model with that purpose. None of these papers focus on the simple detection of a fish being present but they give us a good idea of where to begin our own research.

Another common research topic relevant to our own is that of time-series models being used to track fish migration. These papers have different niche focuses from tracking fish migration in a number of different areas to helping fishermen (Xu et al., 2021), to helping inform hydroelectric power plants when silver eel migration is at its peak in the area so they can shut down the plant and prevent further harm to the species (Trancart et al., 2013). Regardless of the cause, it appears that the consensus is that using either a SARIMA or SARIMAX model is the best method for predicting the peaks of migratory patterns in a specified region.

Methodology

Utrecht Canal Camera Data

Given the hyper specificity of our mission statement, we decided to web-scrape video data of the fish doorbell in question. After scraping live stream footage of roughly 4 hours we scrubbed through the footage to check for any fish, of which there were none. We then found compilation videos on YouTube of times fish have appeared in front of the camera and edited them together so that we had roughly 30 minutes of footage where every single frame had a fish within view. Now that we had these two videos we read in each frame and for each frame we: removed color, flattened it to be a 1D array, resized for dimensional consistency, and added a label (0 if from the video with no fish and 1 for if from the video of all fish). Then we combined all the data into one data frame with the last column being the label and exported it to a feather file. We decided to use feather files so that we could easily share and store large quantities of data in a more compact manner.

With this data collected, we determined that using a CNN would make the most sense to gain relevant insight into our problem of interest. We created a few CNN models all with two convolutional layers to increase runtime. Each convolutional layer had a kernel size of 3 x 3 and stride of one, with the first of the two having an additional padding layer. Our models had 2-3 hidden layers, the first of which had 1000 nodes with the second having 500. If a 3rd layer was used, it had 250 nodes. The models then all used a sigmoid activation function on the output since we're looking for classification (fish or no fish).

Loire River Vichy Counting Station Data

StacomiR is an open-source project developed in France by the French Office for Biodiversity Institute to centralize data obtained by fish pass monitoring. The library acts as an API for connecting to various databases hosted under this project. We specifically pull data from the “r_mig_interannual_vichy” dataset which contains data from 1997 to 2012 on salmon migration at the Vichy counting station in the Loire River in France.

Because the data contains data on daily intervals across several years, it acts as a natural candidate for running time-series analysis to better understand and predict changing fish migration trends. However, the data only contains counts for a small proportion of the days across all years (about 27%). To address this, we adapt a version of forward fill imputation where each missing value is filled with the average value from the previous seven days.

We first use this data to test an Autoregressive (AR) Model with a lag of 365 (since migratory seasons occur annually). Since the data dealt with relatively small values, it varied widely across days and needed to be stabilized to reduce noise. This was done by smoothing calculating the moving average for each value with a window of 10.

We additionally test a Seasonal Autoregressive Integrated Moving Average (SARIMA) Model with a seasonal differencing order of one, seasonal MA order of two, and seasonal AR order of two. SARIMA combines an AR model and Moving Average model while removing the need for stationary processes by applying a differencing procedure before modelling. More importantly, it incorporates both non-seasonal and seasonal components to explicitly account for seasonality, making it apt for our problem since migratory data is highly seasonal. To shorten the seasonal periods fed to the model, data for the SARIMA model was aggregated by summing over weeks. A square root transformation was also performed to stabilize the variance of counts.

Results

Convolutional Neural Network

Figure 1: Example of a frame with a fish present after our 2 convolutional layers

After taking in a total of 5115 different frames, we fit the model with 10 epochs and a batch size of 30. We tested multiple iterations under this framework to determine the optimal model architecture. Specifically, we tested a handful of activation functions for the hidden and output layers, as well as using either two or three hidden layers. After testing our different models we found that the best iteration was one with 3 hidden layers with the activation functions: sigmoid, ReLU, and ReLU in that order. This model returned an average accuracy score of 1 and an average recall score of 1. Additionally, we manually inspected images to make sure they were being correctly classified. Figure 1 shows an example of a "fish frame" after two convolutional layers from our final model.

Forecasting Migration Patterns

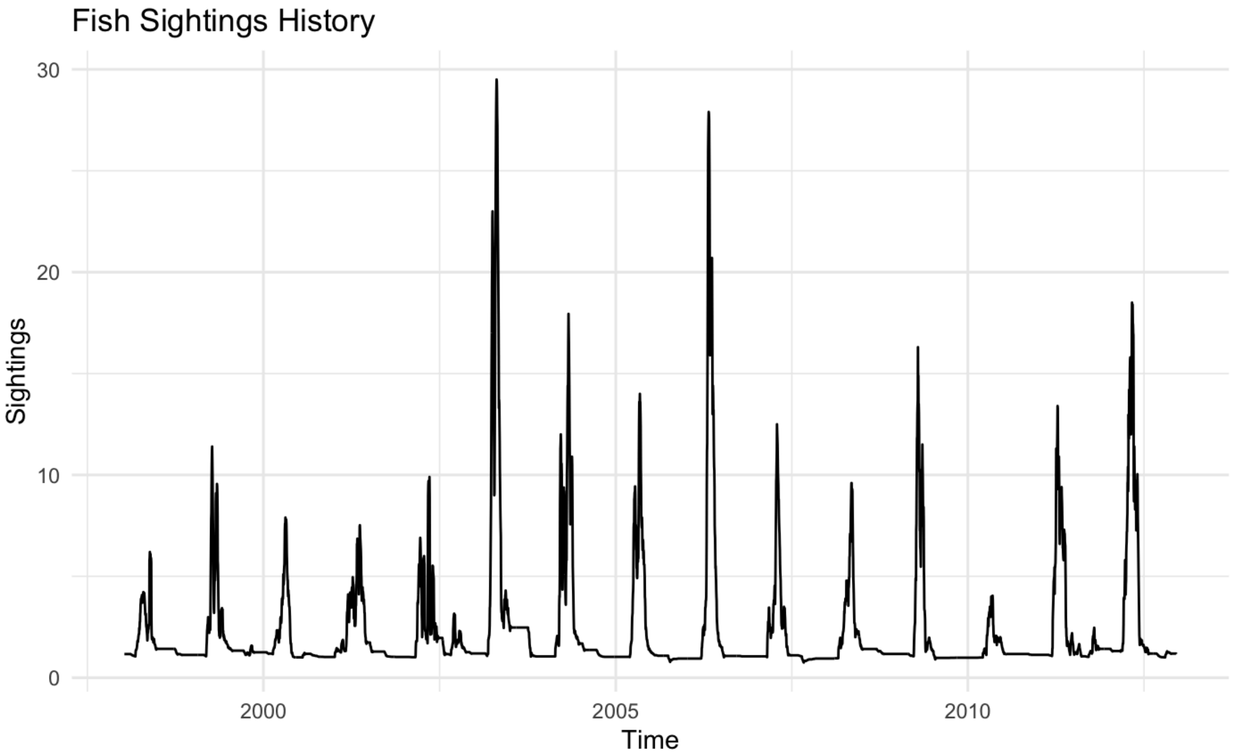

Figure 2: Fish sightings over time at the Loire River Vichy counting station.

After imputing and calculating moving averages, the migration data contained 5,821 observations from January 1st, 1997 to December 12th, 2012. Figure 1 shows the trends in migration after smoothing. You can see that the data is highly seasonal with spikes occurring around the same time each year. Trends across seasons tend to be fairly consistent. Most years saw a maximum of 8-15 sightings in one day, although certain years show outlier counts much higher than average and 2010 shows significantly low sightings.

AR(365)

Before modelling, the Augmented Dickey-Fuller test was run with a lag order of 17 to check the stationary assumption. This resulted in a p-value less than 0.01, indicating stationary data. After fitting the data, the model produced a mean absolute error of 1.18 and a root mean squared error of 2.29. While the model seems to only miscount by one or two fish, other issues suggest this may not be a good fit. We discuss this further below.

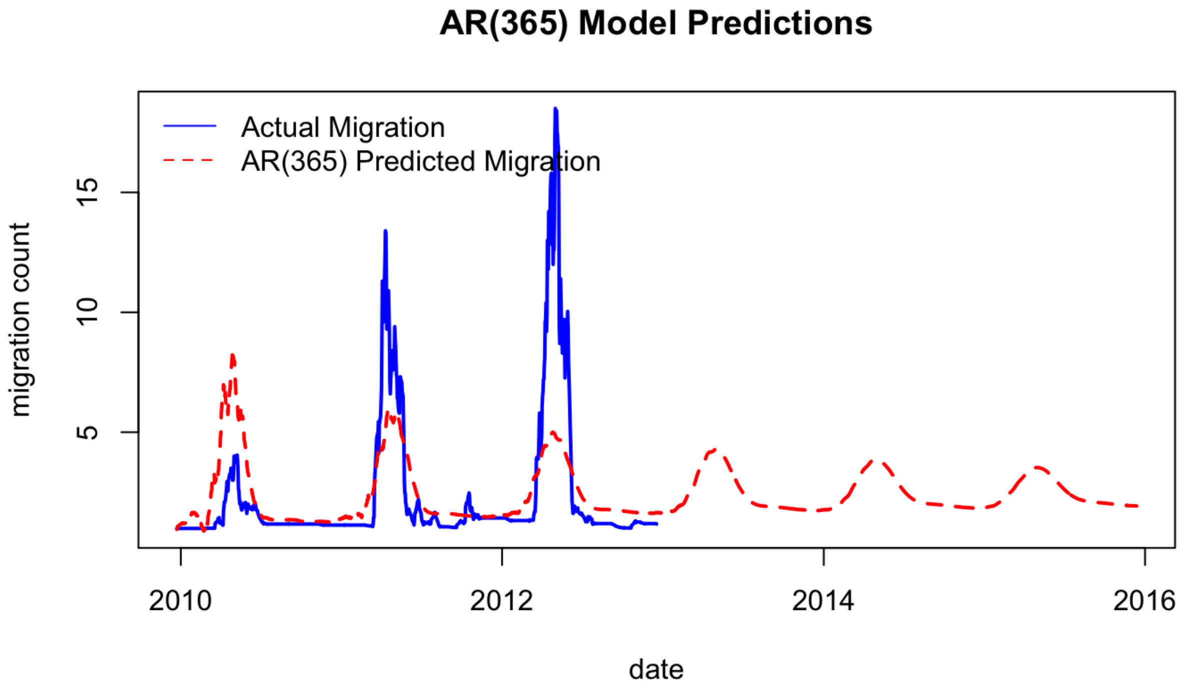

Figure 3: Predictions from the AR(365) model plotted against corresponding Vichy counting station data.

Figure 3 shows the AR(365) model's predicted daily migration counts compared against the true counts from the test set. The model correctly aligns with the seasonal patterns in the data, but incorrectly predicts decreasing migration trends over time which would likely lead to increasing error at the same rate. As noted previously, the historical data show no patterns from season to season, suggesting that the declining predictions indicate an underlying flaw in the model.

Indeed, the model showed several red flags when examining the residuals. Even at high lags, the residuals showed strong autocorrelation. Additionally, the residuals showed right-skewness, indicating the model may not be capturing the full variation of the data. This was further confirmed by the Q-Q plot which had strong high and low tails, indicating non-normality. The residual plot further indicates poor fit, showing strong patterns indicative of non-linearity.

SARIMA

After aggregating by week, the migration data contained 838 observations. Like for the AR model, the Augmented Dickey-Fuller test was run with a lag order of 17 to check the stationary assumption. This, once again, resulted in a p-value less than 0.01, indicating stationary data. After fitting the data, the model had a mean absolute error of 7.08 and root mean squared error of 16.80. This is roughly proportional to the results from the AR model after scaling by day.

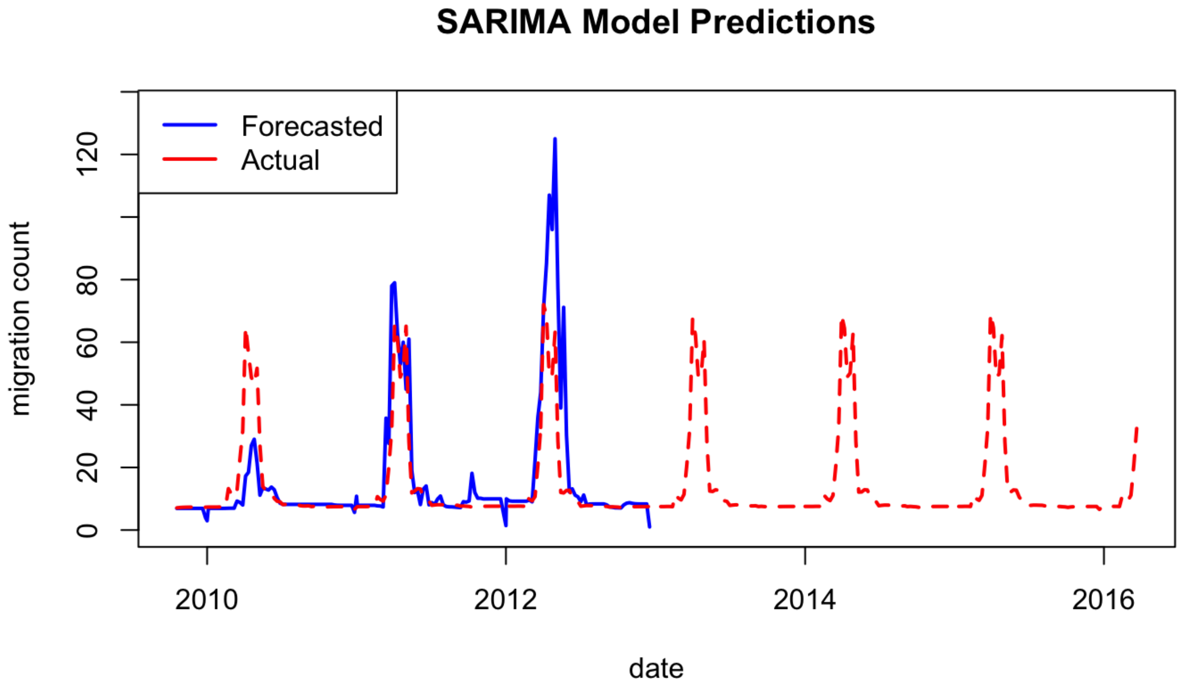

Figure 4: Predictions from the SARIMA model plotted against corresponding Vichy counting station data.

Figure 4 shows the SARIMA model's predictions of weekly fish counts plotted against the testing data. As shown, the model correctly aligns with the seasonality of the data. Unlike the AR model, the predictions show fairly consistent trends across seasons with peak migration count staying steadily around 60 fish a week.

The SARIMA model also seems to be showing less dependency and better pattern recognition than the AR model. With the exception of three early lags showing very small significant autocorrelation, the residuals show no significant autocorrelation. The residuals are also symmetrically distributed about zero, showing no signs of skewness. However, while not as extreme as the AR model, the Q-Q plot for SARIMA still shows strong high and low tails.

Conclusions and Future Work

Final results from our CNN indicate that we are able to detect almost all fish accurately. Future work on this would benefit from continuing to explore different model architectures to optimize these runtimes as we'd want it to be able to make these predictions almost instantly. Another method of analysis we'd consider exploring in future work would be to take the difference in pixel values from one frame to the next to detect movement and train our model to detect movement instead. More computing power may allow for more complex models, which may have better success. Additionally, the CNN required large amounts of internal memory to process the training data. Input data for this model in general requires large amounts of computing power. Further work on this problem would benefit from examining ways of cutting down on this cost.

Both the AR and SARIMA models were able to correctly fit seasonal patterns of the fish migration data. However, both models showed certain signs of poor fit, decreasing our confidence in their predictions. The AR model with a lag of 365, in particular, had trouble properly fitting the data, failing several of the assumptions necessary for producing confident results.

As discussed previously, the Augmented Dickey-Fuller test confirmed the stationarity assumption for AR was met. However, post-hoc analyses of the residuals showed non-linearity and non-normality as well as strong dependence in the error term. Because of this, we cannot be confident in the predictions produced by the AR model. The SARIMA model similarly met assumptions of stationarity but failed to demonstrate normality, potentially because of outliers in the data. Though, it did pass assumptions for both linearity and white noise. Still, continued analysis of this problem would benefit from investigating the normality of the data and potentially down-weighting or removing outliers that may be affecting fit.

Previous research has noted the SARIMA model lacks the ability to account for external influences. This could be factored in by use of a SARIMAX model which certain literature has deemed as the most effective method for predicting fish migratory patterns (Trancart et al., 2013). If we were to continue working on this problem we would certainly look into a SARIMAX model to account for the variety in intensities of the migration spikes.

Lastly, the Loire River data acts as supplemental for data that cannot currently be obtained from the fish doorbell. However, the CNN model we propose could be used to curate similar data and add to our knowledge base. With enough time, similar analyses could be run for the Utrecht Canal, providing a more complete view on the wellness of these ecosystems.

References

Trancart, T., Acou, A., De Oliveira, E., & Feunteun, E. (2013). Forecasting animal migration using SARIMAX: an efficient means of reducing silver eel mortality caused by turbines. *Endang Species Res* 21:181-190. https://doi.org/10.3354/esr00517

Trnovszky, T., Kamencay, P., Orjesek, R., Benco, M., & Sykora, P. (2017). Animal recognition system based on Convolutional Neural Network. *Advances in Electrical and Electronic Engineering*, *15*(3). https://doi.org/10.15598/aeee.v15i3.2202

Xu, F., Du, Y.-A., Chen, H., & Zhu, J.-M. (2021). [retracted] prediction of fish migration caused by Ocean Warming based on Sarima model. *Complexity*, *2021*(1). https://doi.org/10.1155/2021/5553935