ANS 128: Group Project

Avian Bird Influenza

Jesus Ibarra Barron
Tim Deng
Madison Beckham

Abstract

This project analyses a dataset of recorded cases on the Avian Influenza (H5N1) strain from Ireland between the years of 1980 and 2020. By applying both a linear regression model and cluster analysis, trends between the occurrence of positive Avian Influenza cases the year they occurred and the region which they occurred in were identified. This analysis highlights a significant relationship between positive cases with the year AND location in Ireland. Specifically, the year 2007 had the highest positive infection rate (.29), and the county of Leitrim (4.0) had the highest occurrences of H5N1 overall. The insights gained from this study could be valuable in further understanding global patterns of Avian Influenza and offer guidance for disease control, prevention, and regulation for Avian Influenza cases seen worldwide.

Introduction & Problem Statement

Avian Influenza, commonly referred to as "The Bird Flu," is a highly contagious viral disease that primarily affects avian species, including waterfowl such as ducks and geese, as well as domestic poultry like chickens and turkeys. The virus comprises multiple strains, including H6, H7, H9, and H10. However, the H5N1 strain is of particular concern due to its high pathogenicity

in birds and its ability to infect a range of mammalian species, including humans. This zoonotic potential has raised significant public health concerns, prompting ongoing surveillance and research efforts to mitigate its spread and impact.

The “Bird Flu Dataset (Avian Influenza)” analyzes cases of the H5N1 strain of Avian Influenza in Ireland from 1980 to 2020, comprising data from 16,304 individuals. This analysis examines independent variables such as the year, locality, county, latitude, longitude, and species to determine their influence on disease prevalence. Using this data, our primary objective was to identify which of these variables listed contributed to a higher probability of positive infection.

To achieve these insights, two analytical approaches were used: Multiple Regression Analysis for identifying which variables correlate with a higher probability of Avian Influenza infection. Cluster Analysis (Unsupervised Learning) was used for examining how cases group based on a combination of location, time, and species to identify patterns and potential hotspots. These methods provided insight to the factors influencing disease transmission and new ways to inhibit the spread of the disease

Literature Review:

Our research aimed to understand what variables in birds contribute to the higher probability of having the avian flu in Ireland.

A similar study conducted in 2007 focused on the susceptibility of small terrestrial birds to the H5N1 strain. In this study, they collected the bird flu from dead infected birds, and they inoculated sparrows and other species with it. After infection, the birds were housed with non-infected birds “to study intraspecies transmission (Boon et al, 2007).” It showed that different species of birds were affected with varying intensity. This study found that “intraspecies transmission in these hosts is very low”, showing that the specific species they studied are not likely to spread the illness (Boon et al, 2007). Another susceptibility study was conducted in 2006, but in North America on ducks and gulls. This study administered inoculations to these birds with the H5N1 virus (which itself was collected from a different infected bird), and symptoms were observed. It essentially found that these birds are extremely susceptible to H5N1 and “ducks can transmit [Avian influenza viruses] over great distances as they migrate, and these viruses can remain infectious for prolonged periods of time in water (Brown et al, 2006).” Both of these studies seem to suggest that the bird species is the variable that is affecting transmission rates and susceptibility (aka more likely to get the disease). Our data set, however, suggests different variables are affecting the number of birds with positive bird flu tests.

For this study, we used a Multiple Regression Model to find which variable from the observations had the most correlation to positive avian flu tests in birds. A review done by Musa et al in 2024 gathered many different studies that used statistical models, such as ours, to predict and follow avian flu. The review concluded that "integrating AI, mathematical models, and technological innovations into a One-Health approach is essential for improving surveillance, forecasting, and response strategies to mitigate the impacts of the ongoing avian influenza outbreak (Musa et al, 2024)." One study presented by this paper used spatial regression analysis "for identifying environmental factors and hotspots of migratory bird habitats (Musa et al, 2024)." It even specifies that another study used "multiple linear regression on a temporal level to build an early warning system for avian influenza outbreaks based on Google Trends (Musa et al, 2024)." We are doing something similar to that study, but with data collected on the field, rather than on the internet. This means that our use of a multiple linear regression model is justified and should even be encouraged.

This data set is becoming more important as time goes on, since avian influenza is becoming more prominent across the world. This includes Ireland, where "...two separate cases of the H5N1 strain of HPAI have recently been identified in buzzards; one in the west of the country in early December [2024] and one in the east of the country in late December [2024]. These are the first wild bird detections in Ireland since September 2023 (Health Protection, 2025) ..." However, bird flu is not just affecting migratory birds, it is also damaging livestock, and we don't "know what the hit is going to be in terms of long-term health for animals that were infected. We know they can survive, but many of them could have impacts to their productivity for years, perhaps the rest of the animal's life (Public Health, 2025)."

Our study will benefit this ongoing research by helping to prevent future outbreaks in Ireland and beyond, since the dataset we use has an abundance of different migratory bird species. Knowing the variables causing the most infections in one country can help others organize a plan to prevent these outbreaks across the world.

Dataset Description

The public dataset we used was "Bird Flu Dataset (Avian Influenza)" and it contains information on the distribution and contagion of migratory birds collected in the country of Ireland. Researchers captured wild birds from the years 1980 to 2020 and recorded whether the bird had the H5N1 strain of avian flu or not. Along with that, they recorded species name, common name, state, month, year, latitude, longitude, and county where the bird was tested. There were 16,304 observations recorded in this dataset, with a majority of them (8429) occurring between the years

2018 and 2020. The most common species of bird caught and tested was the European Robin, which consisted of 3% of all observations.

Exploratory Data Analysis

Positive Infection Rate By Year: 2007 (0.29166), 2015 (0.2267), 2008 (.1750)

No Reported Cases (Mean of zero):1980, 1999, 2002

By County: Leitrim (4.0), Cavan (3.0) Roscommon (2.957) Dublin (2.265)

No Reported Cases: Carlow

Summary Statistics:

Response Variable (If individual has the flu or not)

Mean: (0.1585). Variance: (.1334)

After exploring the dataset, there were several visible trends that could be identified when evaluating the H5N1 strain of Avian Influenza. When evaluating the trend in years, years which were closer to 2020 had more prevalent positive cases, suggesting an increase in infection during that time. In terms of location, the northern half of Ireland can be seen to have more positive cases than the southern half, suggestion a possible association to climate and regional factors. It was also seen that warmer months (spring, summer, and fall) had a higher rate of infection than cooler winter months. When interpreting the mean of 0.1585, it can be inferred that approximately 15.85% of the birds tested were infected, while most weren't (84.15%). The low variance of 0.1334 further supports that majority of the birds were not infected and illustrates the little variability in the infection status. Thus, having a dataset which is skewed towards non-infected birds. Overall, these trends give useful information which can be further used to evaluate how geography, seasons, and temperature can contribute to the overall positive rate of infection.

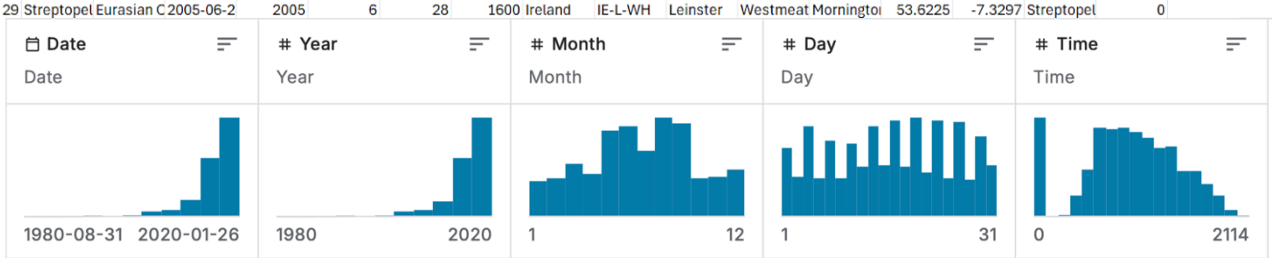
Visualization

The dataset looks at different factors including:

1.) Scientific/Common Name
2.) Year
3.) Month
6.) Country
7.) County
8.) State

4.) Day
5.) Time
9.) Locality
10.) Longitude

_id	Scientific_	Common_	Date	Year	Month	Day	Time	Country	Country_S	State	County	Locality	Latitude	Longitude	Parent_Sp	target_H5_HPAI
1	Acrocepha	Eurasian R	2015-09-1	2015	9	15	1100	Ireland	IE-C-GY	Connaugh	Galway	Inishmore	53.1291	-9.7507	Acrocepha	0
2	Acrocepha	Eurasian R	2015-09-1	2015	9	15	1100	Ireland	IE-C-GY	Connaugh	Galway	Inishmore	53.1291	-9.7507	Acrocepha	0
3	Limosa ha	Hudsonian	2015-09-1	2015	9	15	1100	Ireland	IE-C-GY	Connaugh	Galway	Inishmore	53.1291	-9.7507	Limosa ha	0
4	Limosa ha	Hudsonian	2015-09-1	2015	9	15	1100	Ireland	IE-C-GY	Connaugh	Galway	Inishmore	53.1291	-9.7507	Limosa ha	0
5	Limosa ha	Hudsonian	2015-09-1	2015	9	15	1100	Ireland	IE-C-GY	Connaugh	Galway	Inishmore	53.1291	-9.7507	Limosa ha	0
6	Muscicapa	Spotted Fl	2015-09-1	2015	9	15	1100	Ireland	IE-C-GY	Connaugh	Galway	Inishmore	53.1291	-9.7507	Muscicapa	0
7	Larus can	Mew Gull	2014-08-1	2014	8	16	930	Ireland	IE-M-KY	Munster	Kerry	Ring of Ker	52.0018	-10.1239	Larus can	1
8	Corvus fru	Rook	2014-08-1	2014	8	16	930	Ireland	IE-M-KY	Munster	Kerry	Ring of Ker	52.0018	-10.1239	Corvus fru	0
9	Larus arge	Herring Gu	2014-08-1	2014	8	17	900	Ireland	IE-M-KY	Munster	Kerry	Dingle pen	52.1773	-10.4411	Larus arge	0
10	Corvus fru	Rook	2005-06-2	2005	6	22	1300	Ireland	IE-M-CE	Munster	Clare	Cliffs of Mo	53.0089	-9.3916	Corvus fru	0
11	Corvus mo	Eurasian Ja	2005-06-2	2005	6	22	1300	Ireland	IE-M-CE	Munster	Clare	Cliffs of Mo	53.0089	-9.3916	Corvus mo	0
12	Erithacus r	European I	2005-06-2	2005	6	23	700	Ireland	IE-M-CE	Munster	Clare	Callagan's	53.1336	-9.0747	Erithacus r	0
13	Ardea cine	Gray Hero	2005-06-2	2005	6	23	700	Ireland	IE-M-CE	Munster	Clare	Callagan's	53.1336	-9.0747	Ardea cine	1
14	Corvus mo	Eurasian Ja	2005-06-2	2005	6	23	700	Ireland	IE-M-CE	Munster	Clare	Callagan's	53.1336	-9.0747	Corvus mo	0
15	Cygnus olo	Mute Swar	2005-06-2	2005	6	23	900	Ireland	IE-M-CE	Munster	Clare	Ballyvaugh	53.1192	-9.1532	Cygnus olo	1
16	Ardea cine	Gray Hero	2005-06-2	2005	6	23	900	Ireland	IE-M-CE	Munster	Clare	Ballyvaugh	53.1192	-9.1532	Ardea cine	1
17	Cygnus olo	Mute Swar	2005-06-2	2005	6	23	900	Ireland	IE-M-CE	Munster	Clare	Ballyvaugh	53.1192	-9.1532	Cygnus olo	1
18	Alca torda	Razorbill	2005-06-2	2005	6	23	1430	Ireland	IE-M-CE	Munster	Clare	Cliffs of Mo	53.0089	-9.3916	Alca torda	0
19	Motacilla (	Gray Wagt	2005-06-2	2005	6	23	700	Ireland	IE-M-CE	Munster	Clare	Callagan's	53.1336	-9.0747	Motacilla (	0
20	Motacilla a	White Wag	2005-06-2	2005	6	23	700	Ireland	IE-M-CE	Munster	Clare	Callagan's	53.1336	-9.0747	Motacilla a	0
21	Erithacus r	European I	2005-06-2	2005	6	23	700	Ireland	IE-M-CE	Munster	Clare	Callagan's	53.1336	-9.0747	Erithacus r	0
22	Corvus mo	Eurasian Ja	2005-07-0	2005	7	5	1000	Ireland	IE-M-KY	Munster	Kerry	Tarbert Est	52.578	-9.3668	Corvus mo	0
23	Sterna hiru	Common T	2005-07-0	2005	7	5	1000	Ireland	IE-M-KY	Munster	Kerry	Tarbert Est	52.578	-9.3668	Sterna hiru	0
24	Sterna hiru	Common T	2005-07-0	2005	7	5	1000	Ireland	IE-M-KY	Munster	Kerry	Tarbert Est	52.578	-9.3668	Sterna hiru	0
25	Chroicoce	Black-hea	2005-07-0	2005	7	5	1000	Ireland	IE-M-KY	Munster	Kerry	Tarbert Est	52.578	-9.3668	Chroicoce	1
26	Erithacus r	European I	2005-06-2	2005	6	28	1600	Ireland	IE-L-WH	Leinster	Westmeat	Morningto	53.6225	-7.3297	Erithacus r	0
27	Regulus re	Goldcrest	2005-06-2	2005	6	28	1600	Ireland	IE-L-WH	Leinster	Westmeat	Morningto	53.6225	-7.3297	Regulus re	0
28	Phasianus	Ring-neck	2005-06-2	2005	6	28	1600	Ireland	IE-L-WH	Leinster	Westmeat	Morningto	53.6225	-7.3297	Phasianus	0
29	Streptopel	Eurasian C	2005-06-2	2005	6	28	1600	Ireland	IE-L-WH	Leinster	Westmeat	Morningto	53.6225	-7.3297	Streptopel	0



Reference for Charts: “Avian Influenza” Kaggle Dataset

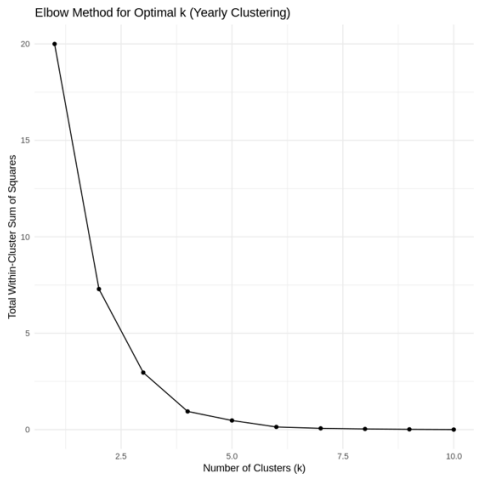
The graphs above show the distribution of five of the variables recorded. The first and second graph show the dates and years, respectively, of the captures, which is why they are very similar. The third graph shows the most common months of observations were taken. Fourth shows what day of the month and fifth graph shows the times.

Model Development & Evaluation

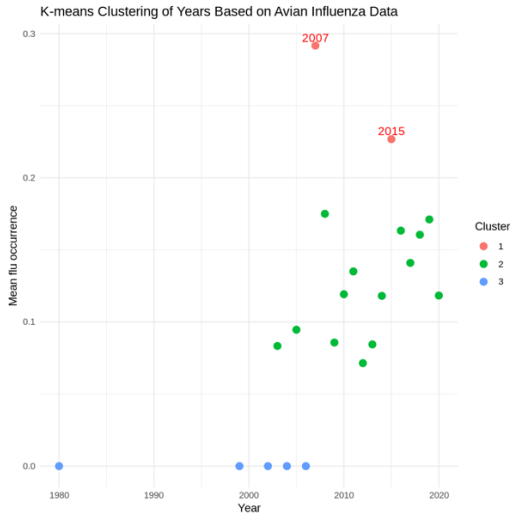
In our dataset analysis a linear regression model in R was used. This model was chosen in order to evaluate the relationship between the data sets’ different variables, allowing for precise and simple evaluation of the data. The year, day, state, counties were all used as significant predictors.

One challenge that we faced in model performance was that that data took a long time to run (due to it being such a large dataset). Another challenge was that there were too many categorical variables (1,288 unique localities). In order to run our model more smoothly, we needed to simplify our variable range, so we opted to use the variables for year, date of year, country, and state. This provided a dataset that was not so small that it lacked depth, but not too large that it would have been too difficult to analyze.

Results and Interpretation

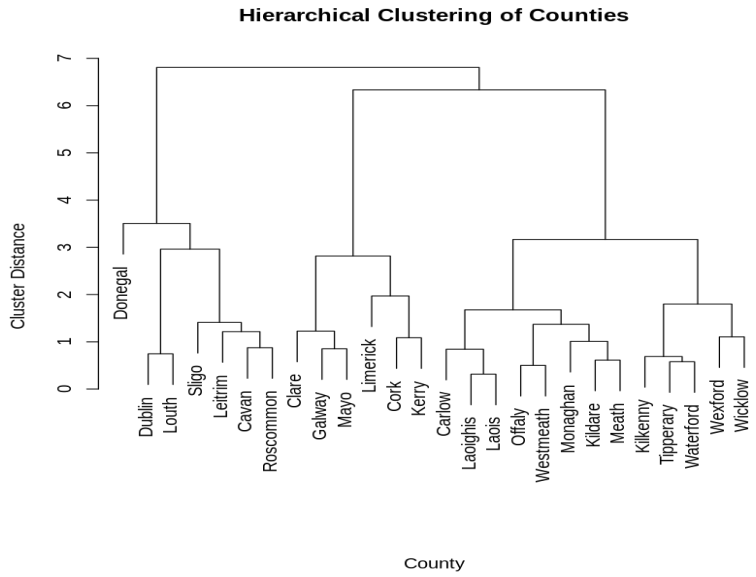


Using an elbow plot was useful in determining the optimal number of clusters to use for K-means clustering on years based on Avian Influenza outbreaks. Good choices of K are accompanied by a sharp drop in total WSS, indicating that the clustering reduces within-group variances effectively. In our case, the curve starts to flatten at K = 3 or K = 4, so these would be the optimal K's before diminishing returns set in. To simplify the timing of the outbreak into periods (presumably pre-outbreak, peak, post-outbreak), we use K = 3.

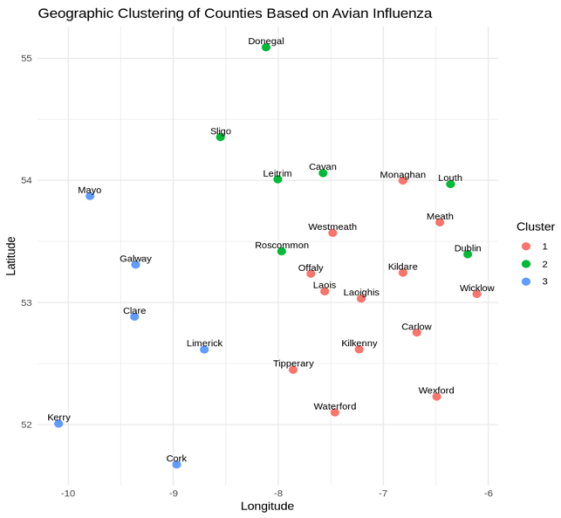


The resulting clusters show that 2007 and 2015 were the peak of the outbreak, as they have large mean flu occurrences per year. Years in cluster 2 have relatively lower mean occurrences; years in cluster 1 have even lower. From this, we could see that the flu did not break out before 2003 but persisted through 2004 - 2020, with peak infection rates at 2007 and 2015.

Then, considering both reported flu outbreak rate (weighted more) and geographical location, hierarchical clustering and geographic clustering of Counties are shown.



The dendrogram above shows the relationship between counties based on their similarity in the considered criteria. Lower branches with closer proximity reflect greater similarity. For example, Wexford and Wicklow are a lot more similar than Wexford with Donegal.



The geographic clustering displays the same information as the hierarchical clustering, except that it attempts to reconstruct the spatial relationship between counties and assign clusters to them. With the effect of flu occurrence rate upscaled, we could see a relationship between spatial proximity and flu infection.

Overall, the model performed well and gave back clear and concise information which could easily be translated into real world scenarios.

Conclusion and Discussion

Avian flu continues to be a danger to Ireland and the rest of the world. Now knowing what variables have the most positive H5N1 tests, we can look back at those years and counties to narrow down the cause of the infections. We can find out specific environmental factors that are contributing to the peak. To improve the work, it would be useful to have more statistical variables from the dataset to analyze. These variables may include how the birds were birds raised and phenotypical variables like age, weight, size, etc.

References

Boon, Adrianus C.M., et al. “Role of terrestrial wild birds in ecology of influenza A virus (H5N1).” *Emerging Infectious Diseases*, vol. 13, no. 11, Nov. 2007, pp. 1720–1724, https://doi.org/10.3201/eid1311.070114.

Brown, Justin D., et al. “Susceptibility of North American ducks and gulls to H5N1 highly pathogenic avian influenza viruses.” Emerging Infectious Diseases, vol. 12, no. 11, Nov. 2006, pp. 1663–1670, https://doi.org/10.3201/eid1211.060652.

Health Protection Surveillance Centre . “Hse Warns of Danger of Wild Birds - Don’t Touch Sick or Dead Wild Birds.” *2025 News Archive: HSE Warns of Danger of Wild Birds - Don’t Touch Sick or Dead Wild Birds - Health Protection Surveillance Centre*, 9 Jan. 2025, www.hpsc.ie/news/newsarchive/2025newsarchive/title-24738-en.html.

Musa, Emmanuel, et al. “Avian Influenza: Lessons from past outbreaks and an inventory of data sources, mathematical and AI models, and Early Warning Systems for forecasting and hotspot detection to tackle ongoing outbreaks.” Healthcare, vol. 12, no. 19, 1 Oct. 2024, p. 1959, https://doi.org/10.3390/healthcare12191959.

Public Health On Call. “Bird Flu Is Raising Red Flags among Health Officials.” Johns Hopkins Bloomberg School of Public Health, 14 Jan. 2025, publichealth.jhu.edu/2025/bird-flu-is-raising-red-flags-among-health-officials.