



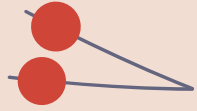
Avian Influenza

The Explorers: Jesus Ibarra, Madison Beckham, and
Tim Deng

ANS128-Intro to Agricultural Data Science

March 13, 2025





Introduction & Problem Statement

Why is this Important? And Relevance to Animal Science or Agricultural Science-

- The dataset analyzes cases of the H5N1 strain of the Avian Influenza in Ireland. Avian Influenza, “The Bird Flu”, is a highly contagious viral disease primarily affecting avian species. It includes multiple strains, however the H5N1 strain, which is analyzed in this dataset, is highly contagious and can affect many mammals including humans. Currently in the United States “The Bird Flu” has picked up within the United States, affecting dozens of commercial poultry and dairies.

Key Research Question or Objective?

- To Analyze past data to help predict future Outcomes and Impacts

What are you trying to solve?

- What variable contributes to the higher probability of Avian Influenza infection.
- Preventative Measure: Use data to create ways to inhibit the spread of Avian Influenza to neighboring countries/ counties



Abstract

- **Used** Multiple Regression Analysis, Cluster Analysis and Decision Tree to have a deeper understanding of the dataset, how accurate it is and how it correlates to real life scenarios
- **Crucial to understand** how to prevent spread of diseases and reduce risk to other countries. Additionally, important to consider various factors such as environmental conditions and animal phenotypic traits which contribute to the likelihood of having the disease
- **Found a High Significance** in Positive Cases between Location and Year in Ireland (Northern Half (specifically Northeastern Quarter of Ireland had more positive cases than Southern Half) .

Literature Review



- **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> .**
 - Focused on the mortality and susceptibility of small terrestrial birds to the H5N1 strain.
 - Small Birds inoculated with H5N1 and housed with non infected birds to see measure transmission
 - Found that “intraspecies transmission in these hosts is very low”

Lit Review Cont...


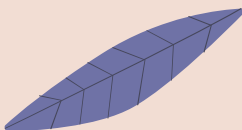
- **Brown, Justin D., et al. “Susceptibility of North American ducks and gulls to H5N1 highly pathogenic avian influenza viruses.” *Emerging Infectious Diseases*, vol. 12, no. 11, Nov. 2006, pp. 1663–1670, <https://doi.org/10.3201/eid1211.060652>.**
 - Study on ducks and gulls that found that H5N1 is very infectious and dangerous in these migratory species
 - Inoculated and then mortality was measured





Lit Review Cont...

How This Project Contributes:

- These studies suggest bird species affects transmission rates and susceptibility (aka more likely to get the disease).
 - Our dataset measures infections across 40 years in 410 different species over 26 different counties in Ireland-> More general overview of the Avian flu's prominence, as well as on-field data.
 - Knowing the variables causing the most infections in one country can help organize a plan to prevent these outbreaks across the world.
- 
- 



Lit Review Cont...

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> .

- **Concluded that “integrating AI, mathematical models, and technological innovations... is essential for improving surveillance, forecasting, and response strategies to mitigate the impacts of the ongoing avian influenza outbreak.”**
- **One study used “spatial regression analysis “for identifying environmental factors and hotspots of migratory bird habitats (Musa et al, 2024).””**
- **Another study used multiple linear regression to “build an early warning system for avian influenza outbreaks based on Google Trends”**



Dataset Description

- Public dataset
 - Collected and tested birds from 1980-2020 in Ireland
 - 16304 observations
 - Species Name, Common Name, State, Month, Year, Latitude, Longitude, County and other identifying components.
 - Focused on Year, State and County.
- 
- 

Dataset As a Whole:

~Easy to see Data Trends ~



- **Trends seen for the occurrence of testing of Avian Bird Influenza**
(includes both positive and negative cases)
- **Most testing Occurred during Warmer Months + Years closer to 2020**
(More symptoms seen in birds during those times which induced the need for testing)



Exploratory Data Analysis

~Highest Mean Cases

- Based off Positive Cases Only

By Year: 2007 (0.29166), **2015** (0.2267), **2008** (.1750)

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

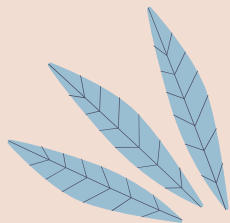
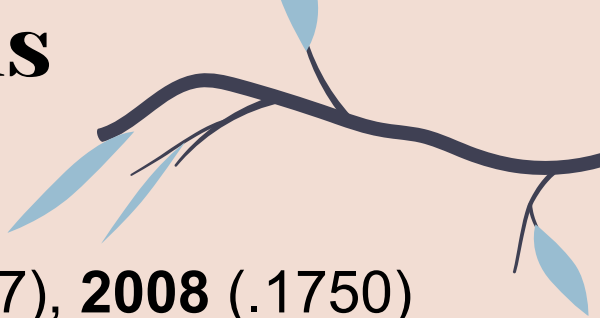
By County: Leitrim (4.0), **Cavan** (3.0) **Roscommon** (2.957) **Dublin** (2.265)

No Reported Cases: Carlow

Summary Statistics:

Response Variable (If individual has the flu or not)

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



Results and interpretations

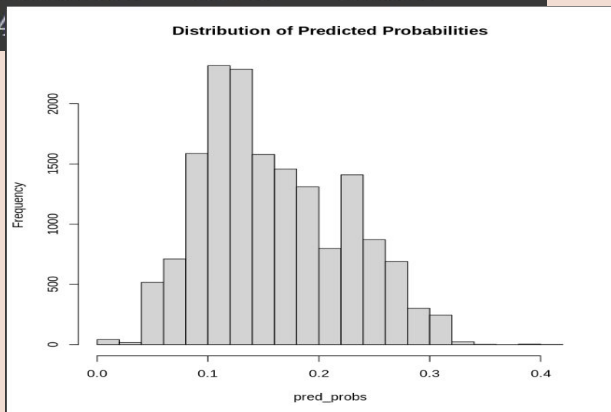
Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-94.848923	245.264436	-0.387	0.699
Year	0.039593	0.010054	3.938	8.21e-05 ***
day_of_year	0.002007	0.000241	8.327	< 2e-16 ***
CountyCavan	13.643175	244.426135	0.056	0.955
CountyClare	12.214893	244.425181	0.050	0.960
CountyCork	12.579469	244.425169	0.051	0.959
CountyDonegal	11.864990	244.425219	0.049	0.961
CountyDublin	13.372037	244.425157	0.055	0.956
CountyGalway	12.940613	244.425161	0.053	0.958
CountyKerry	12.485134	244.425161	0.053	0.958

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	13393	2506
1	226	50

Accuracy : 0.8311
95% CI : (0.8252, 0.8368)
No Information Rate : 0.842
P-Value [Acc > NIR] : 0.9999



After model selection (using BIC), run logistic regression with intercept and year, day of year, and county as predictors. Only year and day of year are significant at size = 0.05.

In R: `glm(formula = target_H5_HPAI ~ 1 + Year + day_of_year + County, family = binomial, data)`

Interpretation example:

Being one year older increases the log-odds of flu occurrence by 0.04.

For predictive use, model accuracy was 83.1% with threshold set to 0.3

- Problems: linear assumption; 84% NIR → struggling to predict true positives → *useless*

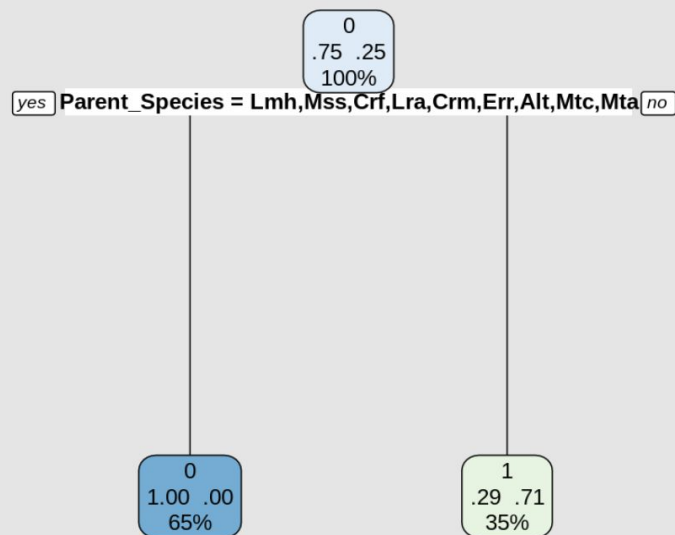
Balanced Accuracy : 0.500

Results and interpretations cont.

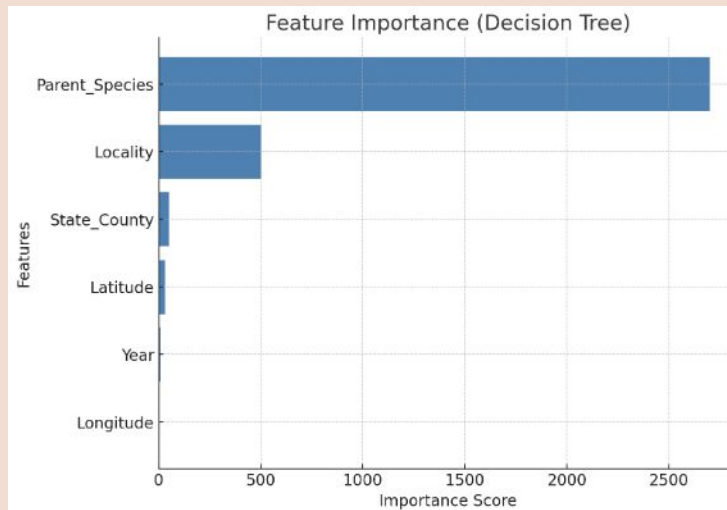
Switch to decision tree classifier?

Example tree trained using the first 20 obs,
actual tree is way too messy to display

Decision Tree for Avian Influenza Prediction (20 Obs)



Full tree visualization & performance metrics



Note: with a somewhat arbitrarily fixed CP (complexity parameter), species is dropped due to collinearity with parent species

```
preds <- predict(tree_model, test_data[, independent_vars], type = "class")
```

Reference	
Prediction	0 1
0	4088 6
1	27 769

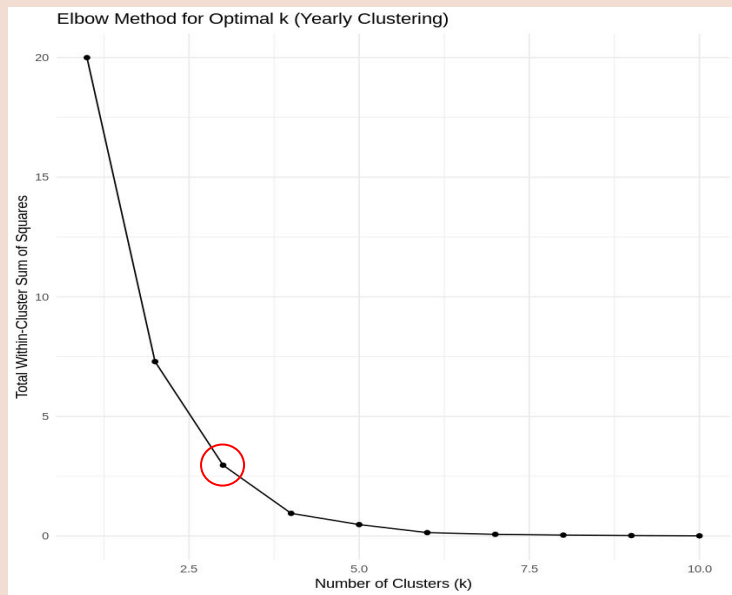
Accuracy : 0.9933
95% CI : (0.9905, 0.9954)
No Information Rate : 0.8415
P-Value [Acc > NIR] : < 2.2e-16

Sensitivity : 0.9927
Specificity : 0.9897
Pos Pred Value : 0.9980
Neg Pred Value : 0.9624
Prevalence : 0.8415
Detection Rate : 0.8354
Detection Prevalence : 0.8370
Balanced Accuracy : 0.9912

Suspiciously high? Given enough data, each possible attribute comb. would have a leaf?

Results and interpretations cont.

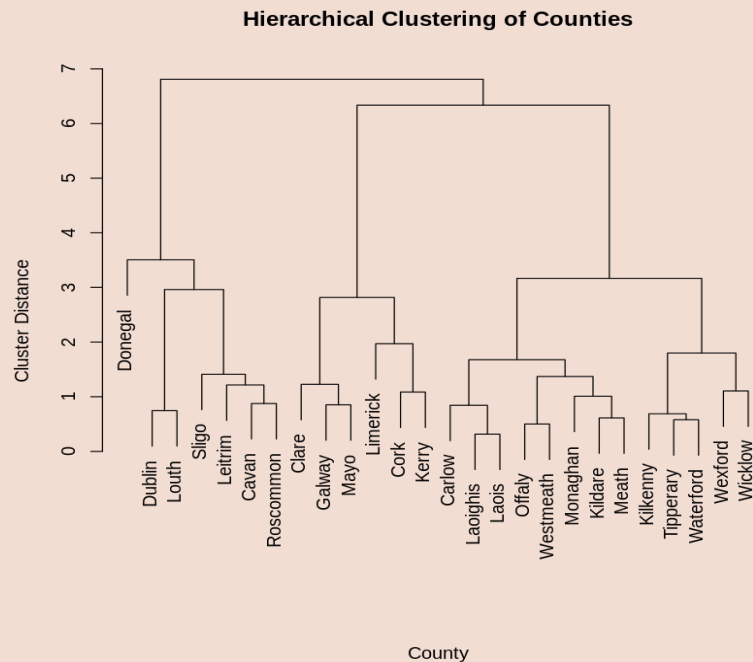
K-means clustering



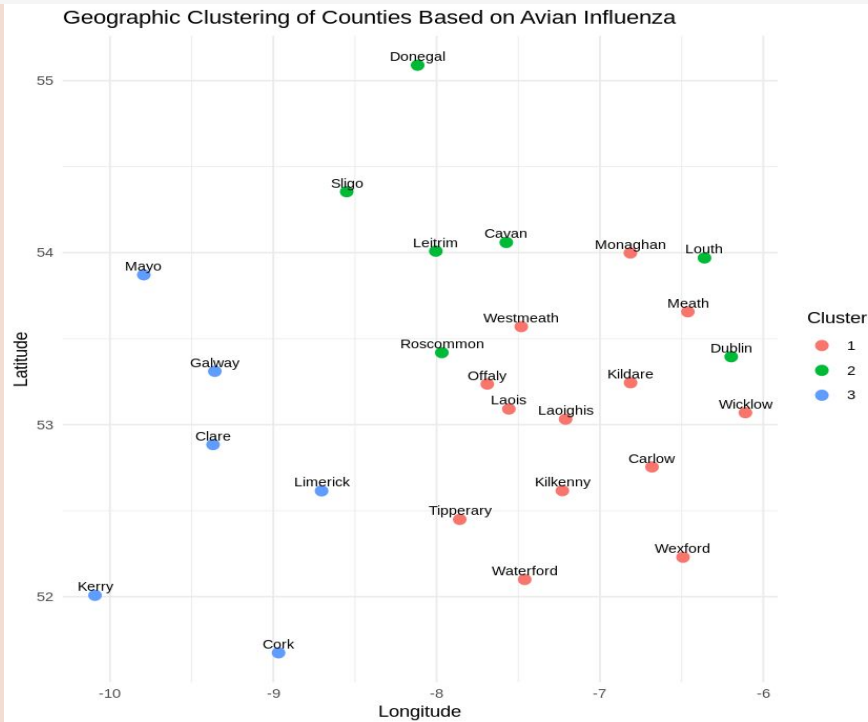
Elbow plot shows $k = 3$ or $k = 4$ reduces within group sum of squares/variances effectively, then WSS starts to flatten out. To simplify the timing of outbreak into periods (presumably pre-outbreak, peak, post-outbreak), we use $k = 3$.

Flu did not break out before 2003 but persisted through 2004 - 2020, with peak infection rates at 2007 and 2015.

Results and interpretation cont.



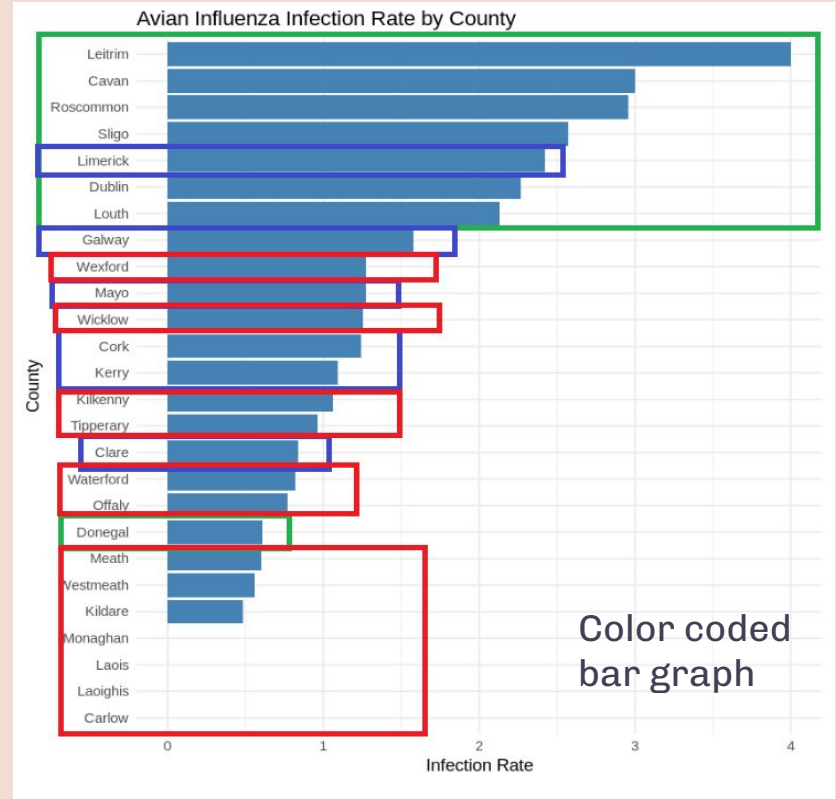
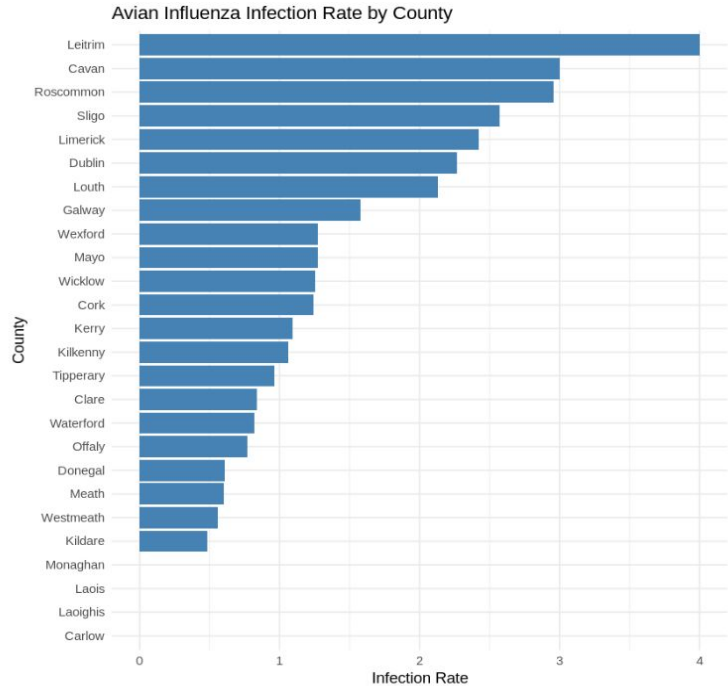
```
geo_summary <- data %>%
  group_by(County) %>%
  summarize(
    count_cases = n(),
    h5_hpai_rate = 10*mean(target_H5_HPAI, na.rm = TRUE), # Mean flu occurence per county
    avg_latitude = mean(Latitude, na.rm = TRUE),
    avg_longitude = mean(Longitude, na.rm = TRUE)
  ) %>%
  drop_na()
```



Consideration of both reported flu outbreak rate (weighted more) and geographical location in clustering.

- Hierarchical clustering shows county similarity in features.
- Geographic clustering shows grouping.

Results and interpretation cont.



- Association can be seen between Location and Infection Rate (Northern and North Eastern Half Of Ireland had more reported positive cases)
 - Can hypothesise that neighboring counties may have shared a similar infection rate (perhaps due to less biosecurity, sanitation practices, and overall animal health)

Summing up: Model Development & Evaluation

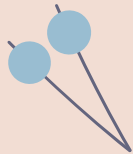


What method was used?

Logistic Regression (1), Decision Tree (2), Cluster Analysis (3)

Why were they chosen?

- (1) Hypothesis testing: estimates, std. errors, p-val
- (1) + (2) Classification
- (3) Unsupervised learning



Model Performance

- How Well Did the Model Perform?

The model performed very well, with some hiccups however it did give clear information which could be easily translated into real world scenarios.

- Limitations & Challenges Faced in Model Performance -

- 1.) Data Took a long time to run (large dataset)
- 2.) Too many categorical variables (1,288 unique localities). Opted to use year, date of year, country, and state in order to have a range which wasn't too large nor too small
- 3.) Decision Tree Results; Data Accuracy



Overall Results, Conclusion & Discussion

Main Takeaways:

We can conclude that 2007 and 2015 were the peak years of Infection. As well as we can infer that the Infection pattern is Geologically related to the rate of infection.

Implications & Applications of Findings:

Knowing this, 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.

What Can Be Done to Improve the Work?

Would be useful to have more statistical variables from the dataset to analyze such as how the birds were raised, phenotypic variables, etc.



Conclusion & Discussion Cont....

Future Directions (what new things we can do)

- Explore how/why location and weather Influence a higher rate of Infection;
Can explore data from other countries to see if it is consistent for this strain of Avian Influenza
- Data can be used as plan on how to take preventative measure; Such as preparing biosecurity and control of disease for peak months

References & Acknowledgments

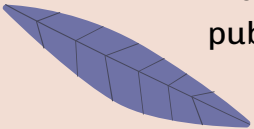


Boon, Adrianus C.M., et al. "Role of terrestrial wild birds in ecology of influenza A virus (H5N1)." *Emerging Infectious Diseases*, vol. 13, no. 11, Nov. 2007, pp. 1720–1724, <https://doi.org/10.3201/eid1311.070114>.

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.

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.





Q&A



Thanks!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**

Please keep this slide for attribution

