2022-2023 H5N1 Bird Flu Modeling and Prediction in the United States

March 12th, 2023

Authors:

UCDAVIS
Cheng, Weilin - wncheng@ucdavis.edu
Liu, Hengyuan - hyliu@ucdavis.edu
Mo, Kathy - kamo@ucdavis.edu

UNIVERSITY OF MICHIGAN
Tian, Sida - startian@umich.edu
Yuan, Li – leeyuan@umich.edu

GitHub Repository:

https://github.com/GitData-GA/iasc2023

Source Data:

[1] United States Counties Database https://simplemaps.com/data/us-counties

[2] H5N1 Bird Flu Detections across the United States (Backyard and Commercial) https://www.cdc.gov/flu/avianflu/data-map-commercial.html

[3] H5N1 Bird Flu Detections across the United States (Wild Birds) https://www.cdc.gov/flu/avianflu/data-map-wild-birds.html

[4] Monthly Average Temperature of each County across the United States https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping

[5] Monthly Average Temperature of each County in Hawaii https://weatherspark.com/map?id=145043

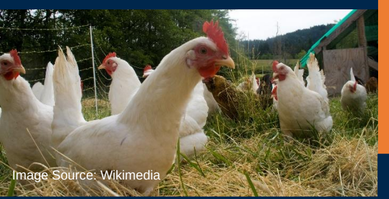


Image Source: Wikimedia

Abstract

This report presents an analysis of the likelihood of H5N1 outbreaks in different counties of the United States in January 2023 using logistic regression, ridge regression, and lasso regression models. The models were trained using historical data from 2022, and the accuracy of the models in predicting H5N1 outbreaks in January 2023 is about 98.4%. The lasso regression model performed the best among the three models, with an AUC of 0.8015. The map generated based on the lasso regression model indicated that counties in the north and west were at a higher risk of having H5N1 outbreaks in January 2023, which matched the actual result. The report concludes that there are limitations to the models, including the consideration of only a limited set of factors affecting the spread of the virus and the use of historical data. Future work could incorporate additional data sources and use more sophisticated machine learning techniques to improve the accuracy of the models. The report also proposes some possible remedies to help control the spread of H5N1.

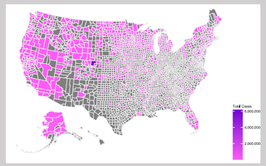


Figure 1: January 2023 Total H5N1 Cases

Background & Objective

Poultry is a popular food in the United States, but it can be infected with viruses such as the H5N1 virus. This virus has caused economic, ecological, environmental, and health consequences.

"Egg prices rose to record highs in December. A dozen large Grade A eggs had more than doubled in price during 2022, on average." – CNBC, Feb. 7 2023

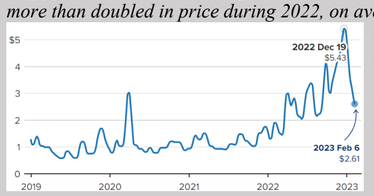


Figure 2: Weekly Price of a Dozen Eggs (From: Urner Barry)

Our research aims to analyze past outbreaks of avian influenza and use mathematical and statistical methods to predict future outbreaks in different regions of the country. By identifying high-risk areas, we hope to contribute to efforts to mitigate the impact of the H5N1 virus and protect public health. Our report will focus on analyzing data from authoritative organizations such as the CDC, USDA, U.S. Census Bureau, and the Bureau of Labor Statistics, and developing and evaluating machine learning models to accurately predict the likelihood of future outbreaks. We will also provide some suggestions to help control and monitor H5N1 outbreaks based on our analyses.

Data Processing & Description

To develop a predictive model for identifying counties that might be at risk of H5N1 infection in the future, we need to understand the structure and content of our data. Our approach involves merging five source datasets to create a single, curated dataset that contains all information on H5N1 cases in each county, during a specific month, from January 2022 to January 2023, and a specific outbreak type. There are 163,436 observations and 10 variables in the curated dataset.

- `fips`: Each FIPS code represents **a unique county** in the United States, so it is a categorical variable with 3,143 unique values, each unique value has 52 entries.
- `state`: **Abbreviation of each state** in the United States, so it is a categorical variable with 51 unique values, each state has its own number of counties.
- `county`: The **names of counties, independent cities, census areas, and same administrative level regions** in the United States. It is a categorical variable with 3,143 unique values with respect to states, each unique value has 52 entries (note that some states have some counties with the same name).
- `lat`: The **latitude** of each county.
- `lng`: The **longitude** of each county.
- `month.index`: The **order of month of avian influenza outbreak** from 1 (January 2022) to 13 (January 2023). Each `month.index` has 12,572 entries.

Every `month.index` has the same number of entries because the cleaned dataset contains all counties' H5N1 situations regardless of how many cases they have, if there is no cases in a county, then the case number is just 0.

- `type`: The **type of outbreak in a specific county and month**. It is a categorical variable with 4 unique values, including `poultry` (40,859 entries), `non-poultry` (40,859 entries), `wild bird` (40,859 entries), and `captive wild bird` (40,859 entries). Every type has the same number of entries because the cleaned dataset contains all counties' H5N1 situations regardless of how many cases they have, if there is no case in a county, then the case number is just 0.
- `avg.temp`: The **average temperature** in a specific county and month in **Fahrenheit degree** (°F).
- `cases`: The number of H5N1 cases detected o**n a specific outbreak type in a specific county and month**.
- `binary.case`: If the case of a type of outbreak in a specific county and month is 0, then it is marked as `uninfected` (160,993 entries). Otherwise, it is marked as `infected` (2,443 entries).

The cleaned dataset will be used to train our classification model to predict which counties might be likely to experience H5N1 infection in the upcoming months.

Visualization

Figure 3 indicates that as the months went by, there seemed more new cases of the H5N1 virus. Most of the new cases were in the West and Midwest regions. There was a fluctuation of new cases in April, May, and from August to the end of the year 2022.

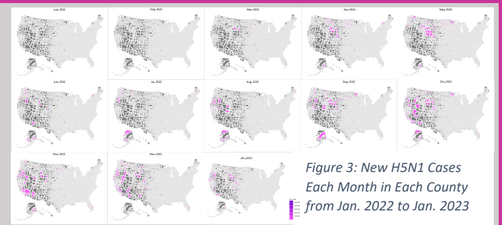
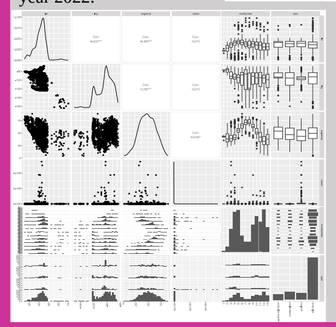


Figure 3: New H5N1 Cases Each Month in Each County from Jan. 2022 to Jan. 2023


<- Figure 4: Scatterplot Matrix of All Existing Outbreaks

In figure 4, the x-axis represents the variables of the rows, and the y-axis represents the variables of the columns. Among the numerical variables, latitude and longitude have the highest correlation. The correlation of -0.423 indicates that these two variables have a moderate negative relationship with each other. Furthermore, the cases variable indicates that there are not a lot of cases in each outbreak, yet there are many outliers. Possible reasons for this are that although wild birds make up most of the dataset, poultry are in large groups while wild birds are not. Since viruses spread more easily through close contact, most of the cases are poultry.

Modeling

We use the outbreaks in 2022 as our training set, which has 150,864 observations. Moreover, we let the outbreaks happened in January 2023 as the testing set, which has 12,572 observations. The testing set will indicate how well our model performs on predicting which county will have H5N1 outbreak on each outbreak types (poultry, non-poultry, wild bird, and captive wild bird).

Our model is

$$Y = \beta_0 + \beta_1 X_{\text{lat}} + \beta_2 X_{\text{lng}} + \beta_3 X_{\text{month.index}} + \beta_4 X_{\text{type(non-poultry)}} + \beta_5 X_{\text{type(poultry)}} + \beta_6 X_{\text{type(wild bird)}} + \beta_7 X_{\text{avg.temp}} + \beta_8 X_{\text{lat * lng}},$$

where β_0 is the intercept, β_1 to β_8 are the coefficients of features X_{lat} to $X_{lat*lng}$.

Moreover, we need to use the sigmoid function

$$p(\boldsymbol{X}) = \frac{1}{1+e^{-\boldsymbol{X}\boldsymbol{\beta}}},$$

where

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{1\,\text{lat}} & X_{1\,\text{lng}} & \dots & X_{1\,\text{lat * lng}} \\ 1 & X_{2\,\text{lat}} & X_{2\,\text{lng}} & \dots & X_{2\,\text{lat * lng}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{150864\,\text{lat}} & X_{150864\,\text{lng}} & \dots & X_{150864\,\text{lat * lng}} \end{bmatrix}_{150864\times 9}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix}_{9\times 1}.$$

The sigmoid function guarantees that the predicted probability is in the range (0,1) and hence allows us to obtain a sensible prediction.

Because this model determines whether a county will have H5N1 case(s) based on each outbreak type, so it is a binary classifier. Since we have 150,864 observations in the training set, the distribution of Y should be a Binomial Distribution

$$Binomial(150864, p).$$

We will perform three models to get the predicted p, which are logistic regression model, ridge regression model for classification, and lasso regression model for classification. We solve following objective functions.

1. Logistic Regression Model

$$\hat{\beta} = \arg\max_{\beta}\Big[\sum_{i=1}^{150864} Y_i \log(p(X_i)) + (1-Y_i)\log(1-p(X_i))\Big].$$

2. Ridge Regression Model for Classification

$$\hat{\beta}^{\text{ridge}} = \arg\max_{\beta}\Big[\ell(\beta) + \lambda\sum_{i=1}^{150864}\beta_i^2\Big],$$

where $\ell(\beta)$ is the loss function of the logistic regression, and $\lambda\Sigma_{i=1}^{150864}\beta_i^2$ is the penalty term. The λ we use here is 0.0014585.

3. Lasso Regression Model for Classification

$$\hat{\beta}^{\text{lasso}} = \arg\max_{\beta}\Big[\ell(\beta) + \lambda\sum_{i=1}^{150864}|\beta_i|\Big],$$

where $\ell(\beta)$ is the loss function of the logistic regression, and $\lambda\Sigma_{i=1}^{150864}|\beta_i|$ is the penalty term. The λ we use here is 1.9733578×10^{-5}.

Note: We use 5-fold Cross Validation to find the λ that produces the smallest deviance $\Sigma_{i=1}^{150864}\text{d}_\text{i}^2 = \Sigma_{i=1}^{150864} 2\left[Y_i\log\left(\frac{Y_i}{p(X)}\right) + (1-Y_i)\log\left(\frac{1-Y_i}{1-p(X)}\right)\right]$. $CV_{(5)} = \frac{1}{5}\Sigma_{i=1}^{150864}\text{d}_\text{i}^2$.

We got the following $\hat{\beta}_i$ for each of these three models (rounded to 5 decimals):

i	Logistic	Ridge	Lasso
0	17.95338	-9.02613	-17.04739
1	-0.28498	0.09103	0.26509
2	0.07552	-0.00266	-0.06755
3	-0.09589	0.08643	0.09541
4	-0.30772	0.01607	0.28437
5	-0.19225	-0.08307	0.16841
6	-2.05547	1.71345	2.0351
7	0.00087	-0.00313	-0.00096
8	-0.00169	6*10^-5	0.00151

Analysis

All models have the same accuracy to determine infection, 98.4%, on the testing set. In order to know which model performs better, we use Receiver Operating Characteristic (ROC) curve, which tests the goodness of fit, and compare the Area Under the Curve (AUC). As we can see in figure 5, the lasso model has the largest AUC (0.8015). We consider it as the best model here. We plot the prediction of the lasso model as figure 6. Counties in the north and west may have more risk to encounter outbreak(s) in January 2023, which matches the last plot in figure 3.

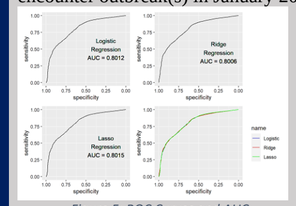


Figure 5: ROC Curves and AUC

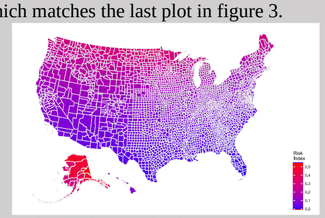


Figure 6: Risk Index Predicted by Lasso Model

Conclusion & Suggestion

- Despite the high accuracy of our models, there are still some limitations that need to be addressed. One limitation is that our models only considered the effects of temperature, outbreak types, time, and density on the likelihood of H5N1 outbreaks, but there may be other factors that also affect the spread of the virus, such as migration patterns of birds, human movement, egg production, breeding size, and so on.
- We suggest the USDA and CDC to make more detailed data collections, including factors that are more relate to bird flu, such as egg production, breeding size, and so on.
- We can do several things to eliminate the spread of H5N1 virus. First, it is essential to implement strict biosecurity measures in poultry farms to prevent the spread of the virus. This includes regular cleaning and disinfection of poultry houses, limiting human and vehicle traffic in and out of the farm, and separating sick birds from healthy ones.
- Second, surveillance programs should be implemented to monitor the spread of the virus in wild birds and poultry farms. This will help identify outbreaks early and prevent further spread of the virus.
- Third, public education campaigns should be launched to raise awareness of the risks of H5N1 and to educate people on how to prevent the spread of the virus.
- Finally, there should be coordinated efforts at the national and international level to track and contain the spread of the virus, including sharing information and resources across different countries.