Avian Migration on the spread of Influenza A (H7N9) in China

Brian Zhang, 02-251 Final Project

Large-scale poultry factory farming operation (Dreamstime/TNS)

Abstract

Avian influenza A viruses (AIVs) have been of great public health concern in China ever since the first case of H7N9 in humans in early 2013. H7N9 is a subtype of the Avian Influenza A virus that primarily targets bird populations and is very common in poultry farming operations along with natural populations of birds. Therefore, in order to understand how H7N9 is spread, it is necessary to study the large scale movement of birds in seasonal avian migration is a major vector for the wide scale spread of the influenza A virus. This paper seeks to further confirm the role which avian migration plays in the transmission of H7N9 and additionally identify specific key bird species which are most responsible for the spread by conducting a cross correlation analysis (CCF). This was done using the GISAID Database for AIV sequences along with the eBird Dataset (EBD). Furthermore, using these publicly available influenza A genomes along with the metadata about the collection source of the sample, we sought to understand the evolution and geographic spread of the virus. Lastly, combining the previous data along with the FAO geographic data about the density of chickens, we developed a SIR model which simulated the spread of H7N9 during one year. Through these series of analysis, we were able to identify key factors in the spread of influenza A.

1. Background

Avian influenza A viruses (AIVs), as the name suggests, targets primarily populations of birds. In the 1990s and 2000s, there were numerous outbreaks of influenza A in poultry largely due to the growth of large-scale poultry farming and live bird markets in China. Among birds, this virus is "highly pathogenic… and may kill nearly an entire poultry flock within 48 hours" (CDC, 2020). While "birds are the natural host for [AIVs]", there was a first human outbreak of H7N9 in 2017 in eastern China with 144 cases and staggering 46 deaths from respiratory failure (WHO, 2020). The scale of this initial outbreak was shocking as the spread of H7N9 to humans was primarily from human-to-bird contact with human-to-human transmission being much more difficult "without prolonged contact" (CDC, 2020). Since then there have been sporadic outbreaks of H7N9 yearly with 916 lab-confirmed human cases in total (WHO, 2017).

Natural populations of birds serve as a "natural reservoir of AIVs", and their migration is a massive vector in the global spread of AIVs across bird populations (Shi et al., 2018). Eastern China, where almost all of the H7N9 outbreaks are contained, lies in the near the center of the East Asian-Australian migration flyway (EAAFP), servicing millions of birds and over 492 species annually (Birdlife, n.d.). Active surveillance of bird infections along this flyway "could provide early-warning for the introduction of AIVs into new regions" and help countries better predict and prepare for future outbreaks of avian influenza A (Shi et. al., 2018). This paper seeks to further confirm the role which avian migration plays in the transmission of H7N9, identify key species and other factors relevant in the spread of influenza A, along with constructing a rudimentary SIR which serves to more accurately model the yearly spread of AIVs by taking into consideration the movement of migratory birds.

2. Methods and Analysis
Preprocessing

For data on influenza A, we queried the Global Initiative on Sharing all Influenza Data (GISAID) Database for full genomic sequences of H7N9 collected from both birds and humans in China since 2013. Each sequence came with metadata which we stripped, leaving just the genome ID, the province the sample was collected from, collection date, and the genome itself. Overall, we found that 85% of all cases were located in the 5 provinces/municipalities of Jiangsu, Jiangxi, Guangdong, Zhejiang, and Shanghai which are all near the sea in eastern China and at the middle of the Asian-Australian flyway. As such, this study focuses only on these 5 regions. Filtering the GISAID dataset for these regions left us with 683 sequences. As data on the cases of H7N9 were not available, these 683 sequences served as a flawed substitute for week-by-week influenza A cases for the purposes of this study.

We identified 150 migratory bird species active along the Asian-Australian migratory flyway and thus candidates for the spread of AIVs (Shi et al., 2018). Data on these 150 bird species in China was retrieved from the March 2019 release of the Bird Observation Dataset (EOD) hosted by the Global Biodiversity Information Facility. The EOD is a collection of bird observations and sightings which are contributed by an enthusiastic online community of amateur bird watchers. A caveat with this dataset is that the EOD often inaccurately measures the number of birds in more urban areas as it is difficult to determine whether one individual bird has been recorded in checklists by multiple birdwatchers. Nevertheless, the long process of vetting and checks by bird experts continues to allow the EOD to be a valuable resource for scientific research (Sullivan, 2014). After filtering the EOD for data in the 5 regions since 2013, we had 57,797 checklists in total. Each checklist consists of the bird species sighted, geographic coordinates of the sighting, the number of birds counted, date of observation, and the observer's name.

Correlation Analysis

In order to quickly get a picture of the data we were working with, we plotted the cases of influenza A against the number bird sighting for each of the 5 provinces. Both datasets were temporally aggregated by weeks in a year. In Jiangxi, Guangdong, and Zhejiang, there are clearly noticeable spikes of bird observations around the new years (around weeks 47-2) and subsequent spikes in influenza A cases a few weeks after. This annual timing of outbreaks and bird sightings align with previous outbreaks of H7N9 and arrival of seasonal wintering birds in China.
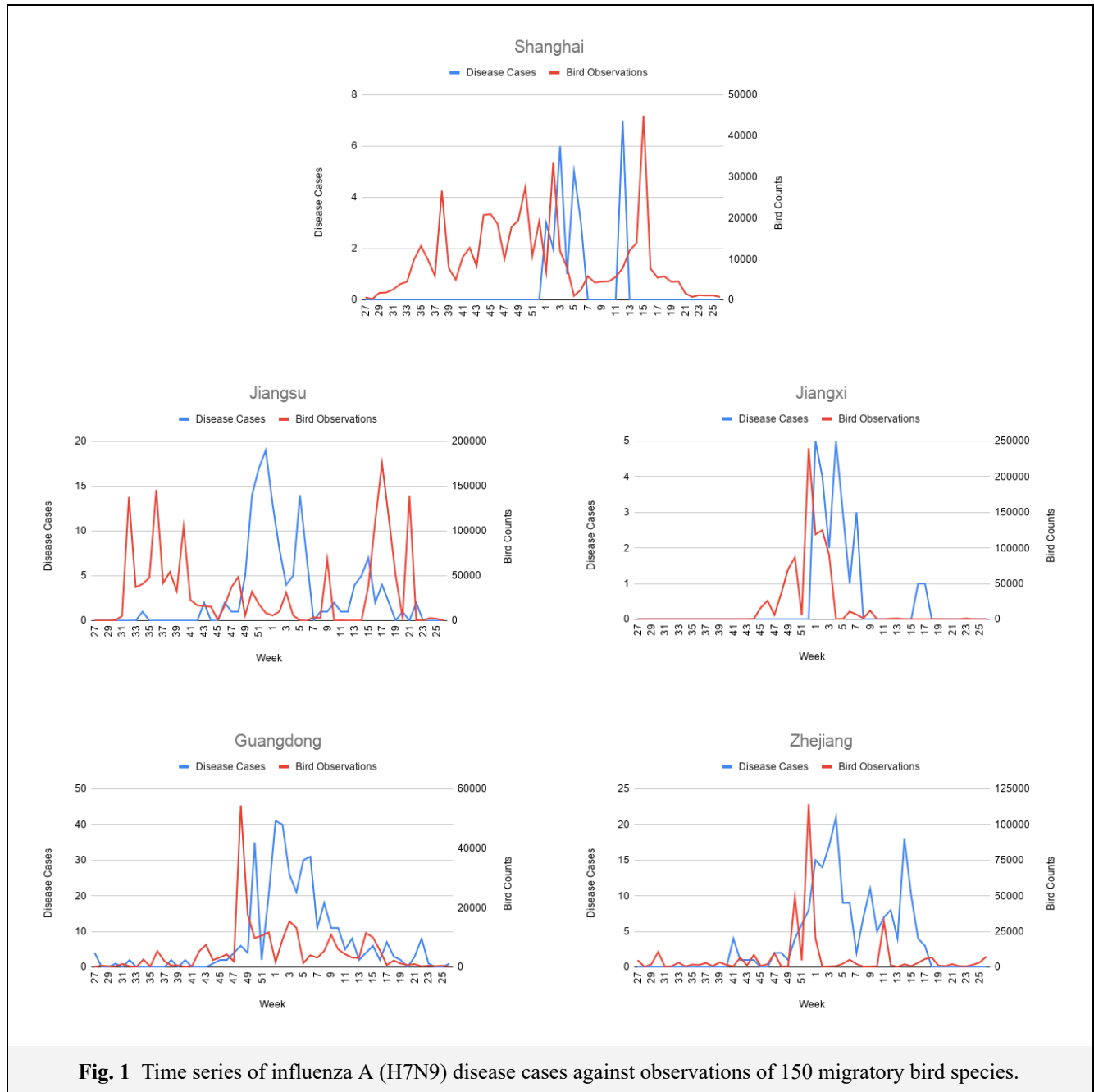

Fig. 1 Time series of influenza A (H7N9) disease cases against observations of 150 migratory bird species.

In order to quantify the previously observed relationship between the two plots, a cross correlation function (CCF) was used in order to compute the correlation between *Bird Counts* and *Disease Cases* as it also takes into consideration a possible *Lag* (in weeks) between the two series. For instance, a positive *Lag* for the CCF is the correlation between *Disease Cases* and the *Bird Counts + Lag,* the time series *Bird Counts* shifted forwards *Lag* weeks. This analysis was conducted for values of *Lag* = -10 to 10 for all 150 bird species and across all 5 regions. Tests which did not score above a correlation of 0.5 for at least one of the values of *Lag* were omitted.
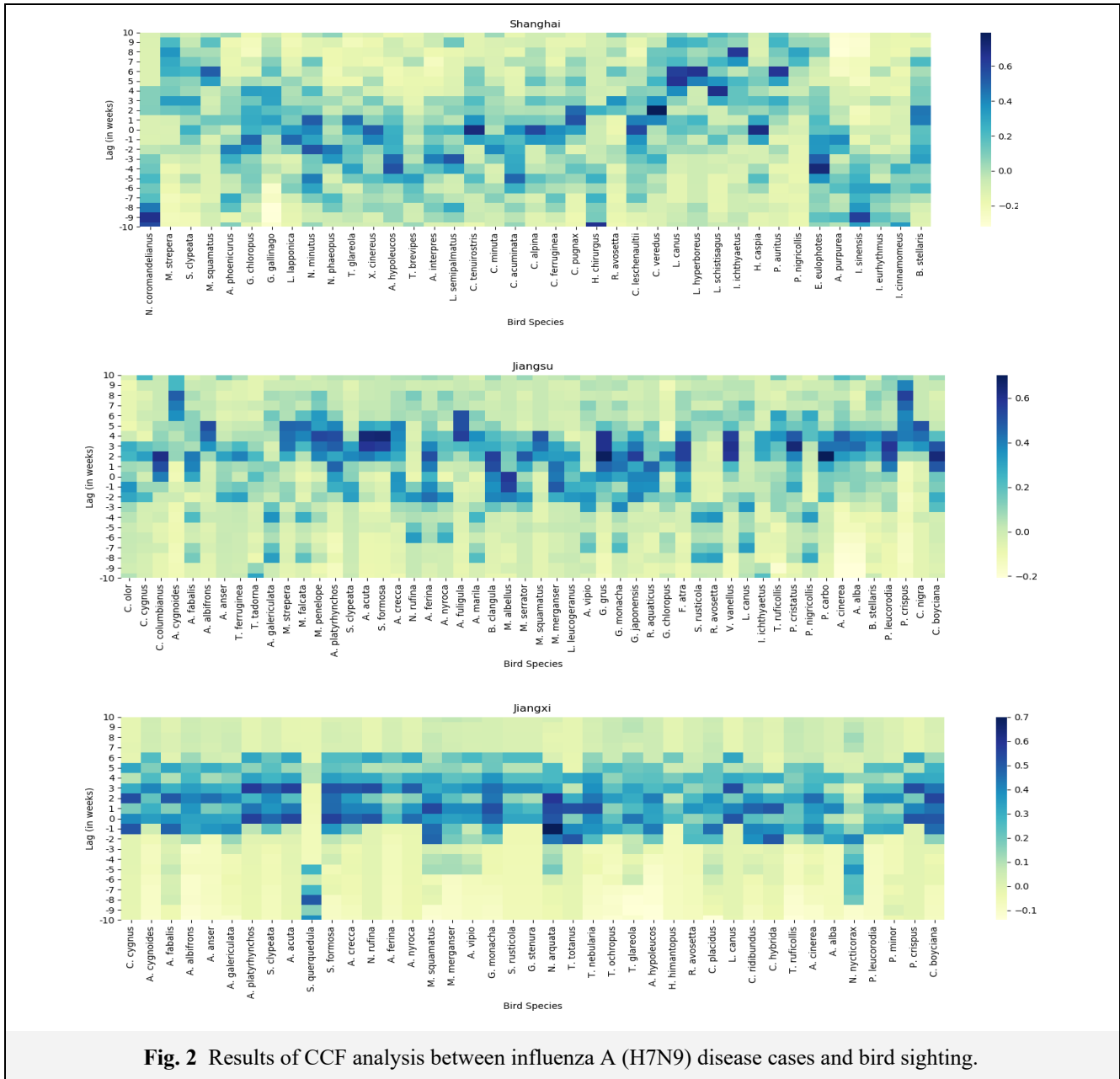
Fig. 2 Results of CCF analysis between influenza A (H7N9) disease cases and bird sighting.
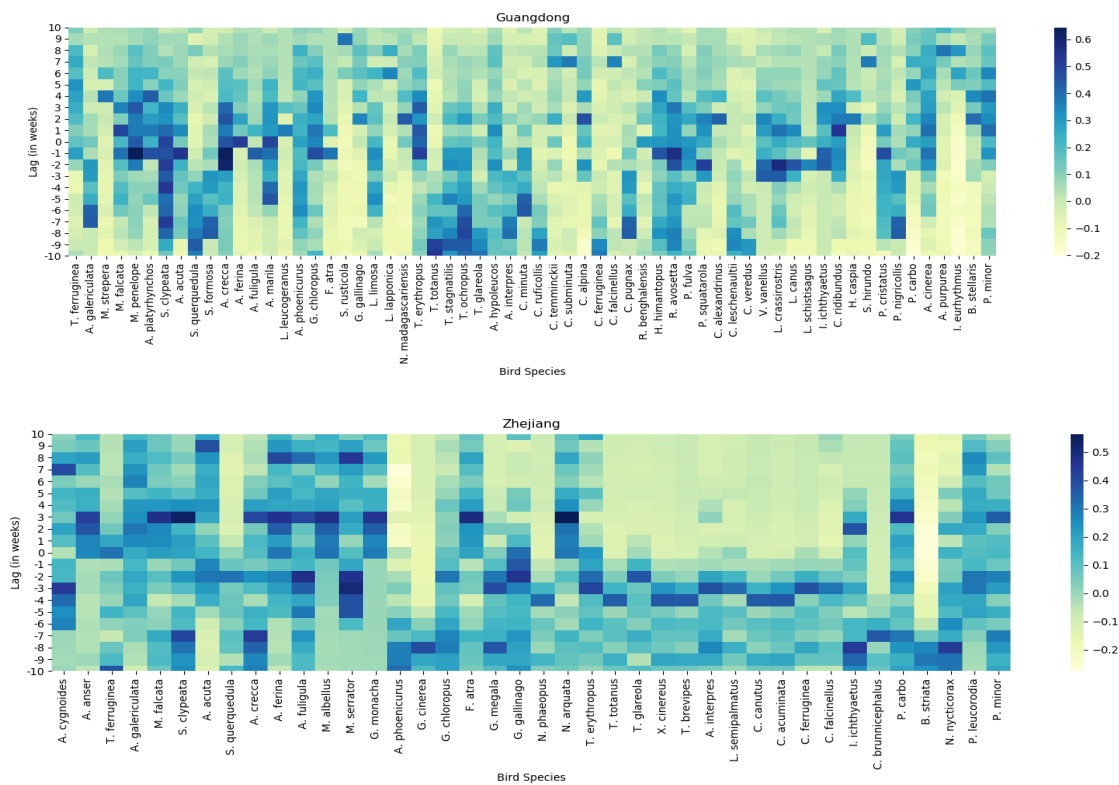
Fig. 2 (cont.) Results of CCF analysis between influenza A (H7N9) disease cases and bird sighting.

Particularly in Shanghai, Jiangsu, and Jiangxi, we noticed a band of high correlation for the 0-5 week values. This observation allows us to conclude that typically 0-5 weeks after birds arrive in a region, the number of influenza A cases tends to rise as well. The CCF plots for Guangdong and Zhejiang are much less clearcut and seem to be much "noisier". In order to determine the species most responsible for the spread of influenza A across all regions, we filtered for species with at least one strong correlation across all lag values in at least 4 of the 5 regions. This method of partitioning our large dataset into 5 regions and seeing how each species performs in each region helps reduce the impact of overfitting as the number of bird observations is substantially larger than the number of influenza A cases.

Aix galericulata (Mandarin Duck), Anas acuta (Northern Pintail), Anas crecca (Eurasian Teal), Aythya ferina (Common Pochard), Gallinula chloropus (Common Moorhen), Recurvirostra avosetta (Pied Avocet), Spatula clypeata (Northern Shoveler), Tringa glareola (Wood Sandpiper)
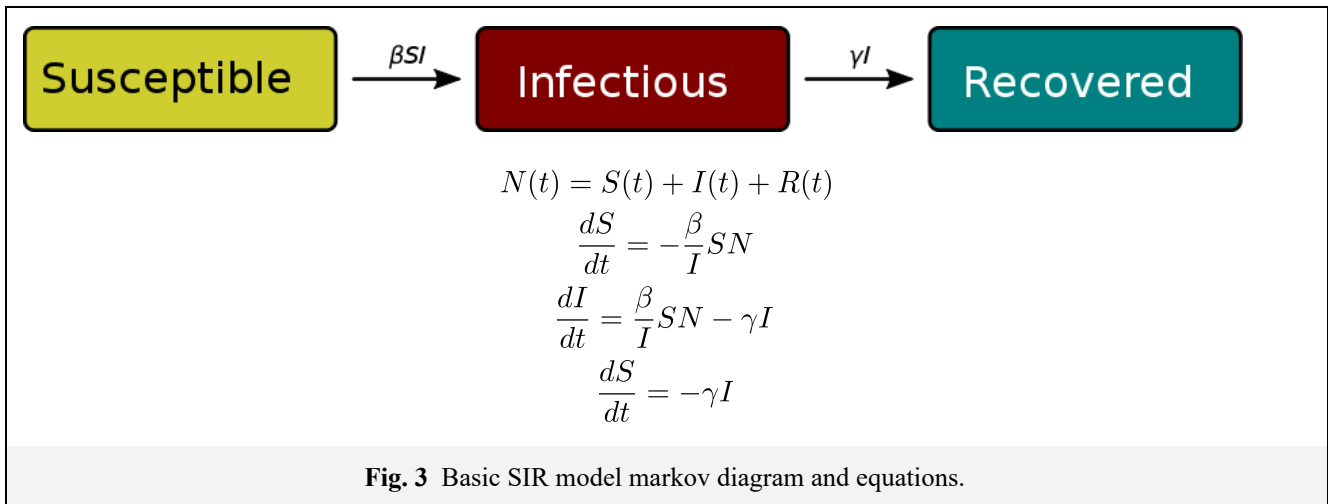
Fig. 3 Key migratory bird species identified to be relevant to the spread of influenza A.

Phylogenetic Analysis

We aligned these sequences using MUSCLE on MEGAX and constructed the phylogenetic tree using the neighbor joining (NJ) method with 1000 bootstrap replicates. After trimming the branches which have low confidence (< 0.4) from these 1000 replicates, we are left with 283 sequences. Using these remaining sequences, we used a maximum likelihood (ML) approach to produce the final tree. Trees constructed using the NJ and maximum parsimony methods differed only slightly and are not shown. Each leaf of the tree corresponds to a given sequence and is labeled with the province of origin and sequence ID.

Fig. 4 ML tree created with 283 influenza A sequences. Each leaf is labeled with the region the sequence was collected from and its corresponding genome ID.

The resulting tree differed from our prediction that all sequences from a specific province would be clustered in one sector of the tree. In reality, this was not the case as sequences from the same region are rather dispersed throughout the entire tree. Additionally, another key feature of the tree is the detached branch in the north-west quadrant of the tree. As this branch is highly disconnected from the rest of the tree, it is likely that these strands were from a relatively isolated or recent outbreak of H7N9.
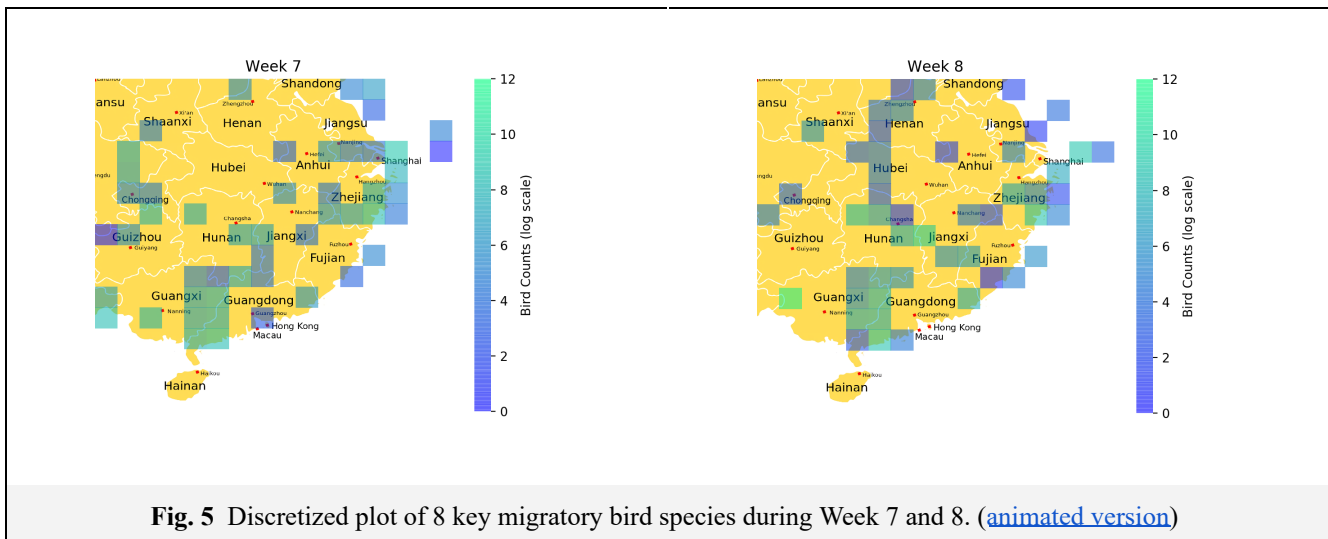
SIR Modeling

The final step of this analysis involved developing a SIR model which takes into account the movement of the identified key migratory birds in order to better simulate the spread of influenza A. A basic SIR model keeps track of the number of susceptible, infectious, and recovered individuals in a population. These values are represented as a function of time and are dynamic with respect to each other by the markov model and equations shown below. The variable β corresponds to the chance that an infected individual spreads the disease to a susceptible individual and γ corresponds to the loss of infectious individuals and the gain of recovered individuals, whether via recovery or mortality, during one time step.

$$N(t) = S(t) + I(t) + R(t)$$
$$\frac{dS}{dt} = -\frac{\beta}{I}SN$$
$$\frac{dI}{dt} = \frac{\beta}{I}SN - \gamma I$$
$$\frac{dS}{dt} = -\gamma I$$

Fig. 3 Basic SIR model markov diagram and equations.

Before we began building the model, we plotted the geographic abundance of birds throughout the year in order to get a better picture of the data we were working with. As our data often had many bird sightings in close proximity, we decided on a discretized plot as that would allow us to better see the distribution of birds in each cell and it would also work well with a compartmentalized SIR model.

Fig. 5 Discretized plot of 8 key migratory bird species during Week 7 and 8. (animated version)

In the snapshots from Week 7 and 8, through just looking at the abundances of bird sighting, we are able to "observe" a great deal of complicated movements of the migratory bird populations. In order to integrate this into our SIR model, we will use the change in abundance as a heuristic for bird movement. In other words, the decrease in abundance of birds from a cell corresponds to the movement of birds from a cell while the increase in abundance corresponds to the movement of birds to a cell. However, even with this assumption it is difficult to answer the question of what cell birds are leaving to or arriving from since a population of migratory birds could potentially cover hundreds or thousands of kilometers in a week. For instance, notice the collection of birds north of Chongqing in week 7. In week 8 this collection of birds has reduced significantly and there is no real precise method of determining exactly where they flew off to given the data we have. To deal with this issue, we will adopt the strong assumption that in a given time step of a week, a given bird can either stay in the same cell or migrate to one of the 8 adjacent cells.

Through this we developed the following model regarding bird migration. M is the total number of migratory birds in a cell and $\Delta M = M(t) - M(t-1)$ is the difference in the number of migratory birds in a cell from the current and previous timestep. If $\Delta M = 0$, then it is a "stable" cell and the number of susceptible (M_S), infectious (M_I), and recovered (M_R) migratory birds stay stable. If $\Delta M < 0$, then there is an emigration of migratory birds from the cell and the proportion of susceptible ($\%M_S$), infectious ($\%M_I$), and recovered ($\%M_R$) birds remain stable. We assume that susceptible, infectious, and recovered individuals have equivalent probabilities of migrating. The last case $\Delta M > 0$ is the most interesting case and shown below. The proportion of incoming birds that are susceptible, infectious, and recovered depends on the states of the neighbors of the cell.

For a given cell C which had a influx of birds:

- Let the following be the set of neighboring cells which experienced an outflux of birds:
$$\text{OUT} = \text{N}_{\text{OUT}}(\text{C})$$

- For a given neighbor C_i, we compute the proportion of birds which enter C from C_i:
$$\%M_{\text{IN}}(C_i) = \frac{\Delta M(C_i)}{\sum_{C_j \in \text{OUT}} \Delta M(C_j)}$$

- Using this we can calculate the proportion and number of incoming birds which are susceptible, infectious, and recovered:

$$\%M_S(C) = \sum_{C_i \in \text{OUT}} \%M_{\text{IN}}(C_i) * \%M_S(C_i) \qquad \Delta M_S(C) = \%M_S(C) * \Delta M(C)$$

$$\%M_I(C) = \sum_{C_i \in \text{OUT}} \%M_{\text{IN}}(C_i) * \%M_I(C_i) \qquad \Delta M_I(C) = \%M_I(C) * \Delta M(C)$$

$$\%M_R(C) = \sum_{C_i \in \text{OUT}} \%M_{\text{IN}}(C_i) * \%M_R(C_i) \qquad \Delta M_R(C) = \%M_R(C) * \Delta M(C)$$

An example of a cell that has seen an influx of migratory birds is shown below. The left cell is responsible for 2/3rds of the incoming birds while the right cell is responsible for 1/3rds as calculated with their respective ΔM values. The other neighbors of the center cell are not depicted as they are either "stable" cells or also "influx" cells and do not have an impact on these calculations. ΔM values. The other neighbors of the center cell are not depicted as they are either "stable" cells or also "influx" cells and do not have an impact on these calculations.
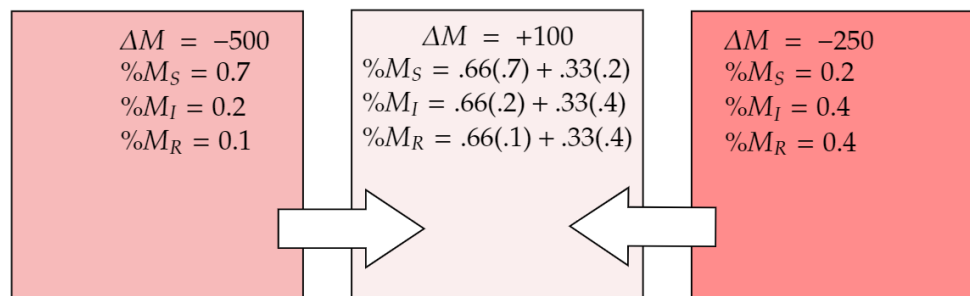
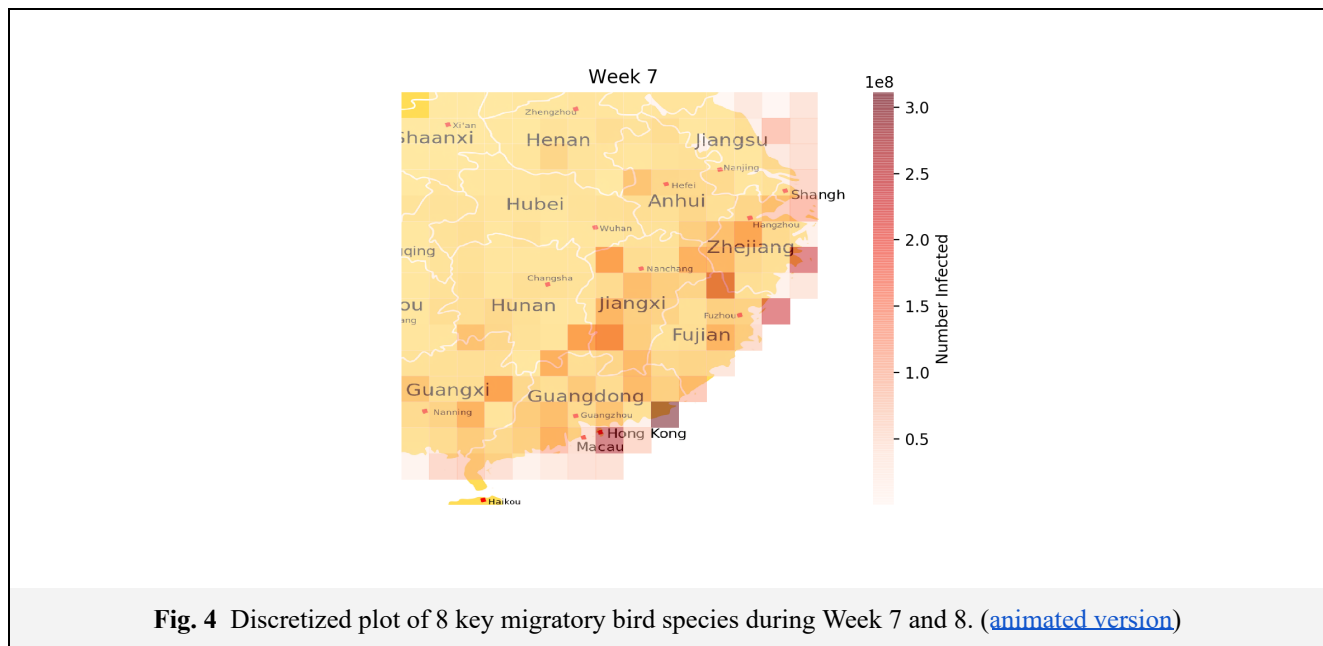Fig. 6 An example of the calculations done to determine the proportion of susceptible, infected, and recovered birds entering a given cell. This diagram only depicts only 2 of the 8 neighbors of the center cell for simplicity.

After figuring out how to model migratory bird movement, we turned towards the FAO dataset on the geographic spread and abundance of chickens. This data was provided in a geotiff format which allows us to determine the population of chickens at each cell in our SIR. A plot of this dataset can be seen below.
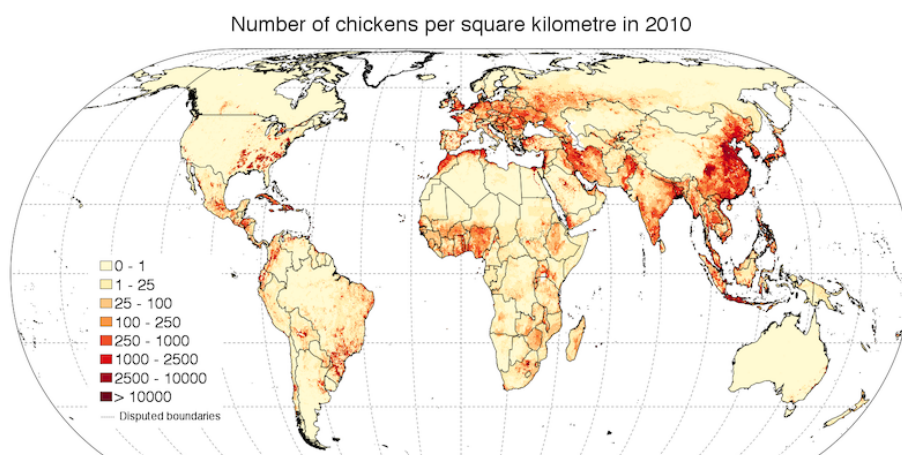
Fig. 7 Food and Agriculture Organization of United Nations (FAO) dataset on geographic distribution of chickens in intensive systems.

Incorporating this information into our SIR model results in a two layer system where each cell has a large, stable, and immobile population of chickens and a smaller, more volatile, and mobile population of migratory birds. We introduce a parameter ζ which represents the interaction between the two populations in a given cell and thus how likely individuals are to infect each other between them. In the end, with more or less arbitrary values of $\beta = .2$, $\gamma = .3$, and $\zeta = .1$ we were able to produce the following visualization.

Fig. 4 Discretized plot of 8 key migratory bird species during Week 7 and 8. ([animated version](#))

We did not get to test and play around with our SIR model as much as we hoped and it has many glaring issues. In the above visualization, we initialized the model with a very high and unrealistic number of influenza A cases during Week 0 in order to get the model to do something interesting given our parameters. Ideally we would have tested various values for β, γ, and ζ and developed a method of testing how realistic the outbreak simulated by the model is when compared to the real-world GISAID data about the spread. Additionally, we would have liked to test the model with the bird migration data with different sets of migratory species and see how that changes the results as well.

3. Works cited

Birdlife. (n.d.). East Asia/Australasia Flyway. Retrieved from
    https://www.birdlife.org/sites/default/files/attachments/8_East_Asia_Australasia_Factsheet.pdf

CDC. (2020). Information on Avian Influenza. Retrieved from https://www.cdc.gov/flu/avianflu/index.htm

Elbe, S., and Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to
    global health. *Global Challenges,* 1:33-46. doi:10.1002/gch2.1018

Gilbert, M., Nicolas, G., Cinardi, G., Van Boeckel, T. P., Vanwambeke, S. O., Wint, G. W., & Robinson, T. P.
    (2018). Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in
    2010. Scientific data, 5(1), 1-11.

Levatich T, Padilla F (2019). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset
    https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2020-05-02.

Shi, B., Zhan, X. M., Zheng, J. X., Qiu, H., Liang, D., Ye, Y. M., ... & Liu, J. (2018). Identifying key bird
    species and geographical hotspots of avian influenza A (H7N9) virus in China. *Infectious diseases of
    poverty*, *7*(1), 97.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., ... & Fink, D. (2014). The
    eBird enterprise: an integrated approach to development and application of citizen science. *Biological
    Conservation*, *169*, 31-40.

WHO. (2020). Avian and other zoonotic influenza. Retrieved from
    https://www.who.int/influenza/human_animal_interface/en/

Shi, B., Zhan, X. M., Zheng, J. X., Qiu, H., Liang, D., Ye, Y. M., ... & Liu, J. (2018). Identifying key bird
    species and geographical hotspots of avian influenza A (H7N9) virus in China. *Infectious diseases of
    poverty*, *7*(1), 97.