Yale University
EliScholar – A Digital Platform for Scholarly Publishing at Yale

Public Health Theses | School of Public Health

January 2013

A Predictive Model Of Avian Influenza Among Poultry In Egypt

Natalie Price
Yale University, natalie.price@yale.edu

Follow this and additional works at: http://elischolar.library.yale.edu/ysphtdl

Recommended Citation
Price, Natalie, "A Predictive Model Of Avian Influenza Among Poultry In Egypt" (2013). *Public Health Theses*. 1235.
http://elischolar.library.yale.edu/ysphtdl/1235

This Open Access Thesis is brought to you for free and open access by the School of Public Health at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Public Health Theses by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.

A Predictive Model of Avian Influenza Outbreaks Among Poultry in Egypt

Submitted By
Natalie Price
Master of Public Health Candidate
Yale School of Public Health
May 2013

Thesis Readers:
Dr. Maria Diuk-Wasser, Yale School of Public Health
Dr. Peter Rabinowitz, Yale School of Medicine

Acknowledgments

I am very grateful to the guidance and mentorship of Michael Kane and Peter Rabinowitz, without which I couldn't have done this project. I learned so much from you both. Thanks for letting me crash your meetings every week.

I am also grateful to Maria Diuk-Wasser, for fostering my interests in ecoepidemiology and zoonotic diseases and pushing me to be a better public health student. Thank you for all of your guidance.

To Mia Sorkin, Adam VanDeusen, Kate Schedel, Melissa Weiner and my mom, Sandy LaCava, for your support.

Table of Contents

Abstract

Background: Avian Influenza (H5N1) has become entrenched in Egypt since its emergence in 2006. Control measures have failed and surveillance systems remain inadequate. A relatively new method for regression called Random Forests is presented here with the goal of providing accurate and timely predictions of the weekly number of outbreaks in each of the Egyptian governorates.

Methods: Predictions were generated from Random Forests models using outbreak data from the FAO EMPRES-i database, and local weather data from Weather Underground. This data was lagged by one and two weeks in order to make prospective predictions with the current week's data in the future. Model performance was assessed using a variety of methods.

Results: The percent of the variance in observed outbreaks explained by the model in each of the governorates varied greatly, ranging between 20 and 60 percent in governorates with high and medium outbreak activity. The models typically predicted poorly in governorates with low activity. Linear regression of the observed outbreaks on the predicted values provided evidence that while outbreaks were consistently underpredicted across all governorates, predictions in some models tracked observed outbreaks quite accurately.

Discussion: The varying levels of model performance in each of the governorates raises many questions about why this is. While we cannot deduce these reasons from the models themselves, public health officials can use the lessons learned here as a guide to focus future research to better understand what is occurring. Predictive models can be used to evaluate local surveillance systems, and find additional covariates for the model to determine the spatio-temporal risk of avian influenza. As a result of better surveillance data and more complete models, control and prevention measures may be more effectively put in place where and when they are needed most.

Background

Since its emergence globally, there have been 622 confirmed cases of H5N1 in humans, and 371 deaths. Egypt has the second highest number of reported human cases and deaths, at 170 and 61 respectively.[1] Avian influenza also threatens one of Egypt's largest agricultural industries, which is a substantial source of employment, livelihood, and animal protein for Egyptians.[2,3,4] Control efforts including culling, quarantine of suspected cases, vaccination and movement restrictions for poultry have failed. Lack of coordination in central and local veterinary services, inconsistent compensation and vaccination policies, dense populations of humans and poultry living together, and poor biosecurity in smallholder operations are the likely causes of endemicity in Egypt.[5] Recent political instability has decreased both the country's ability to cope with outbreaks, and the willingness of citizens to participate in control activities.[6]

This lack of control and frequent human exposure to H5N1 could have global public health consequences as the virus could gain an increased affinity for human receptors, and thus have greater pandemic potential.[7] The control and prevention of outbreaks in endemic countries like Egypt is necessary to reduce the threat of global emergence. Control may be possible through a greater understanding of disease dynamics and increased surveillance.[8] This paper presents a forecasting model that attempts to explain and predict the spatio-temporal dynamics of avian influenza in Egypt at the governorate level with the eventual goal of providing early warning, to decrease the time to disease detection, and improve the planning of targeted disease control measures.

For influenza forecasting models to be accurate and aid in the planning and rapid implementation of disease control measures, the surveillance data used would ideally be complete and disseminated in a timely manner.[9,10,11] Complete and timely pandemic influenza surveillance data, in reality, is a rarity.[12] Models, however, can be used by public health officials to determine the likely risk of disease over space and time, which may help to target surveillance systems in areas where there is low disease detection, but apparent high risk. Models may also help to determine whether increases or decreases in disease incidence are artifacts of changes in surveillance system functioning, or are a true signal in disease patterns.[13] This creates a positive feedback loop that strengthens surveillance systems, while increasing the models ability to detect risk, and thus further targeting surveillance efforts.[14] With decreased time to detection and better understanding of disease dynamics, control and prevention measures can be put into place where and when they will be most effective.

In order for this goal to be achieved, the predictive model must be able to capture the relationships between the predictors and the occurrence of outbreaks, and be able to use this information to accurately forecast outbreaks in real time. The model in this paper uses a relatively new tool for regression called Random Forests to predict the weekly number of H5N1 outbreaks occurring in poultry at the governorate level. Random Forests is a machine learning technique ideal for use in determining the sometimes subtle and complicated natural relationships between predictors and outbreaks.[15] Random Forests have repeatedly demonstrated their superior ability to make accurate predictions when compared with other regression methods because of their ability to learn non-linear relationships.[16,17] Random Forests deal well with small numbers of observations with large numbers of predictors, and are insensitive to outliers, missing data, and many variable types without the need for transformations. Additionally,

Random Forests are not prone to overfitting nor bias, which are concerns in other modeling methods.[18] These advantages, however, come at the cost of model transparency. Error and the percent of the variance explained are reported for the model, as is relative predictor importance. There are no parameter estimates nor significance values given for predictors, leaving the investigator to find creative methods of testing model performance and design, which will be discussed here further.

Random Forests are used more and more to forecast in a variety of fields, including finance to predict stock index movement, meteorology to predict severe weather events, and criminal justice to predict recidivism, but it has not been used to model outbreaks of pandemic influenza.[19,20,21] Applications in ecology to predict species distribution and habitat suitability could potentially be used in the field of landscape epidemiology.[22,23] However, the only known application in this field was the modeling tick presence or absence, and thus tick-borne disease risk, in Italy.[24] Random Forests in influenza research are exclusively used to study viral genetic sequences because of their efficiency and ability to model the complex genetic causal mechanisms, mutations and viral characteristics.[25] The model presented here would provide a novel approach to pandemic influenza epidemiology modeling and prediction. Specifically, this paper will explore the use Random Forests to predict weekly outbreaks at the governorate level, use new methods to test model performance, and discuss future research for model performance improvement. When assessing model performance, I hypothesize that the models in high activity governorates will be able to predict outbreaks better than chance, and that the magnitude of predictions will be to a lesser degree but will adequately show that predictions and actual outbreaks vary together enough to provide some useful lessons.

Methods

Data

Outbreak data was obtained from the FAO EMPRES-i Global Animal Disease Information System.[26] The EMPRES-i database compiles reports on animal disease outbreaks by location and date. Each event is validated and confirmed with staff in the field or local animal health workers upon entry in the database. Individual outbreak data was aggregated over each week by each governorate. Weather data was incorporated into this dataset using Weather Underground, which provides publicly available historical weather data.[27] Weather data was added to the model because there is evidence that temperature may aid in the transmission of influenza viruses.[28] Daily minimum temperatures were collected from 11 weather stations distributed throughout the country. Each governorate was assigned to the closest weather station, and the temperatures were averaged across each week. Data was collected from February 2006, when avian influenza first emerged in Egypt, to the middle of March 2013. R statistical programming platform was used to aggregate the data and to perform all analyses.[29]

Model Design

Temperature and outbreak variables were each lagged by one and two weeks. This idea was considered due to possible delays in the weather's impact on viral development, or to account for the acceleration of an outbreak's trajectory, or some other possible temporal relationship that remains unknown. Lagged variables were also used to allow for prospective out of sample predictions in the future. Different combinations of these lagged variables were run in a Random Forests model and the percent of the variance in outbreaks that was explained by the model was

recorded. The model that maximized the percent variance explained in the governorates overall was selected. The models were applied to each of the governorates using the freely available randomForest package in R.[30] The final model included the average minimum temperature lagged by one week, and aggregated weekly outbreak information lagged by one week and two weeks.

Random Forests

Random Forests are a nonparametric method of regression that uses an ensemble of decision trees to learn data, estimate the importance of each input variable on the outcome, and make predictions. Five hundred trees were used for each Random Forests model in this paper. Each tree, or learner, was trained on a bootstrapped sample of the original data. This means that the observations in the complete dataset were sampled with replacement, until the training set was the same size as the original, but contained some observations more than once and others not at all.[31] The data was then partitioned in space based on a randomly assigned set of predictor variables in the training set to create nodes and grow the decision tree.[32] Observations not included in the creation of the tree, which are called out of bag samples, are then pushed though that tree, and predictions are generated. These predictions are compared with their observations to obtain an unbiased estimate of that tree's error. Generalization error (the difference in training set error and total error) is not a concern due to the large number of trees grown, meaning that there is no overfitting in Random Forests.[33] Retrospective predictions are determined by the average of each tree's individual predictions for each observation. Prospective predictions are created by pushing new observations through each tree and averaging these values across all trees.[34] The errors and the percent of the variance explained were examined to determine model performance.

Model Performance

The time series predictions generated were plotted over time with actual outbreaks to visually inspect prediction performance. Residuals were examined through residual plots as well as a normal QQ plot. Differences in the week-to-week number of outbreaks were calculated for the actual and the predicted values and were plotted against each other. Visual inspection of these plots was used to estimate the number of predictions that were accurate in both direction and magnitude. To test whether or not magnitude was accurately predicted on average, and that predictions and observed values varied together, observed values were regressed on predictions. Because the residuals are not normally distributed much of the linear regression information cannot be used inferentially, however the slope and its standard error, as well as coefficient of determination were examined to determine how the predictions varied with observed outbreaks. Chi-squared or Fisher's Exact Tests were used to determine whether or not a model predicted week to week increases or decreases, regardless of magnitude, better than chance.

Shrinking Predictions to the Mean

In addition to the prediction of weekly cases, the movement of the virus across space was explored. In Random Forests models it is possible to use additional information after the predictions have been generated to increase their accuracy. This is called "shrinking the predictions towards the mean." Because many governorates become infected at the same time and possibly pass infection back and forth, correlation between each of the governorates was

found. Using this information, predictions for each of the correlated governorates can potentially be "shrunk towards the mean" using the equations:

$$P_i = \tfrac{1}{2}(G_{1i} + G_{2i})$$

$$\hat{G}_{1i} = \alpha P_i - (\alpha - 1)G_{1i}$$

$$\hat{G}_{2i} = \alpha P_i - (\alpha - 1)G_{2i}$$

Where P_i is the average of the predictions from two correlated governorates for the ith observation, G_{ji} is the prediction for governorate j for the ith observation, and α is an arbitrary weight tested and chosen by the investigator. The error is calculated with the new predictions and compared with the original error value to determine if shrinkage increases the accuracy of the predictions.

Results

Percent Variance Explained

An individual model was run for each of the governorates despite activity level, although it was hypothesized that governorates with low outbreak activity would not produce high performing models. In **Table 1** we can see the governorates activity level over time judging by the number of weeks with observed outbreaks. Model error is also reported for each governorate, but as this value is not comparable across governorates the table is sorted by the percent variance explained by the model. The percent of variance explained by the model differed greatly between governorates, ranging between 20 and 60 percent in governorates with medium and high outbreak activity. Quite a few of the models for high activity governorates were outperformed by middle activity governorates. Governorates with negative percent variance explained are very poor predictors and are mostly indicative of low activity governorates. Geographic representations of outbreak activity level and percent variance explained are found in **Figure 1**.

Model Performance Measures

Actual values were regressed on predicted values to determine prediction accuracy in direction and magnitude, and how the predictions varied with the observations. In **Table 1** we see that slopes close to one show an accurate prediction on average of the magnitude of outbreaks weekly. Slopes for all governorates are above one, meaning each model consistently underpredicted the number of outbreaks each week. When examining how much the predictions varied with the number of actual outbreaks, we see R-squared values in some governorates are quite high, with a range between 50 and 80 percent. So while the models consistently underpredict there is strong evidence that for some governorates predictions still accurately track movements in the actual observations. Chi-squared and Fisher's exact test, which were used to test the ability of the model to predict increases and decreases regardless of magnitude showed a range of significance both in high and low performance models, which is relatively inconsistent with the rest of the evidence, and raises questions as to whether this is an appropriate measure of model performance.

Additional Lessons Learned

Visual inspection of the plotted time series predictions and observed outbreaks, like the example seen in **Figure 2**, difference plots seen in **Figure 3,** and QQ plots seen in **Figure 4** showed overarching patterns. These figures are just one example model taken from the governorate of

Menoufia, which has experienced the highest number of weeks with observed outbreaks. These figures demonstrate the typical performance of the models in most medium and high activity governorates. From the time series plots and difference plots we see that the models generally underpredicted outbreaks, and lacked the ability to determine the steepness of initial outbreaks, but could accurately predict peaks and decreases. We can also see that Random Forests has difficulty in predicting zero outbreaks in periods of low activity leading to overpredictions during this time. The QQ plots of the residuals for each model reiterate this fact by showing that the data are heavily skewed to the right. Many governorates experienced high activity years in 2010 and 2011, with sustained transmission outside of the typically active winter months. Activity has decreased since this highly erratic period. It appears the models can detect a seasonal pattern even in governorates with substantial transmission throughout the year. This is apparent in the pervasive increase in predictions during the autumn months regardless of actual activity. In **Figure 5** we can examine the aggregated number of outbreaks and predictions across the whole country of Egypt to determine Random Forests' overall performance. This time series plot demonstrates Random Forests overpredicting in times of low activity and underpredicting large spikes and transmission occurring outside the normal winter season. Attempts to shrink predictions towards the mean of highly correlated governorates were not successful. No additional information caused model errors to be lowered.

Discussion

Each of the models is poorest at predicting the magnitude of sudden large spikes in outbreaks. More information may be needed to improve the accuracy of predictions for these spikes, if at all possible. Regardless of the consistent underpredictions of outbreaks, many of the models were able to predict increases that tracked with observed outbreaks, judging by both percent variance explained from the Random Forests output as well as the R-squared of the linear regression of actual outbreaks on predictions. While sharp increases were difficult for the model to detect, peaks and decreases were better predicted. Most likely the use of outbreak data lagged by one and two weeks aids in the model's ability to catch up to the peak and to predict the decrease more accurately. Low activity governorates have poor performing models because of the large number of weeks with no outbreaks, and thus little information to be learned by the models. Despite these models' "poor performance," predictions for these governorates will often be close to zero, which will typically be accurate, as they are low activity. In these cases, poor performance may not actually mean poor predictions over time.

Using chi-squared and Fisher's exact test to determine whether or not the models could predict week to week increases and decreases, regardless of magnitude, provided some results inconsistent with the rest of the model performance assessments. This result was most likely caused by the level of noise in each model during periods of low activity, leading to many extremely minute increases and decreases in the predictions while the actual observations remained at zero. With those fluctuations, and because we were only looking at increases and decreases of any magnitude, it was difficult to separate the noise from more meaningful increases and decreases in the predictions.

Governorates with a higher proportion of outbreaks detected over summer months, as opposed to the more typical outbreaks in winter months, were not as effective at predicting large spikes in summer outbreaks, the aggregate of which we can see in **Figure 5**. This is one hypothesis for

why the chi-squared and Fisher's exact tests showed sensitivity to noise in some governorates and not others. Predictions in governorates with higher rates of transmission during summer months may have more noise in the summer months that do not have outbreaks. In these cases the model will have learned the weather data differently than in a governorate with primarily winter transmission. This shows the model's ability to detect a natural flu season using the weather data, which becomes less useful as increases in the detection of summer cases occur. Further research will be needed to determine the potential causes for increased detection of influenza in the summer, as well as to determine other variables that can be added to these models to decrease the noise in times of low activity.

The Random Forests models can be used to analyze the trend of the disease over time and judge potential impacts of any known surveillance system improvements or deteriorations, or the impacts of any known interventions. If that information is known they can also be used to learn about potential changes in the nature of the virus over time. The increased levels of activity, specifically activity occurring outside of the typical flu season in the years of 2010 and 2011, may be due to any number of reasons. Since then activity appears to have decreased. Determining whether or not these signals are artifacts of changing surveillance and diagnostics, or whether these are actual changes in the nature of outbreaks, may be important in refining the model over time, and in understanding nature of the virus. Certainly modeling the impacts of political instability over this period could be extremely illuminating. Parsing out the decreasing effectiveness in control efforts that are simultaneous with the potential deterioration of surveillance systems during this time is important in understanding disease dynamics in Egypt.

One interesting outcome that raises questions is why the model works well in some governorates and not in others. Intuitively models may not work as well in low activity governorates, but what are the differences between adjacent high activity governorates? Because we have already seen the weaknesses in using weather data alone, more variables are needed to determine the differences across governorates. This additional information in the model could be used to determine why there is more apparent summer transmission in some places, or to capture a variety of other potential differences that only the discerning Random Forests algorithms can detect. More research into the surveillance and reporting systems, animal health worker presence, animal and human density, and income levels across governorates are all potentially useful inputs. While these are ways to improve model performance, the model as it is can be used to help understand, evaluate, and strengthen the surveillance systems in each governorate. Each model's output should match public health officials' assumptions about perceived H5N1 activity in each governorate and the efficacy of each governorate's surveillance system. If it does not, the models can motivate investigations into why this is.

Lastly, predictions may be improved by incorporating data from other governorates that have concurrent outbreaks or that may be responsible for the repeated spread of infection beyond its borders. While correlations between many governorates were found to be quite high, after many attempts and resulting failures to incorporate this information into better predictions, it is clear that there is a more complicated mechanism of spread between governorates. Further research is needed to determine how the governorates are linked and how this relationship has potentially changed over time. It may be difficult to parse out the relationship between the geographically close high activity governorates in the north, but there are a few interesting cases of

geographically distant governorates being highly correlated. Poultry trade routes may be useful in determining the connectedness of the governorates. Additionally, with the increasing availability of viral genetic sequences that are geographically referenced, the phylogeography of H5N1 can be modeled for a more accurate understanding of how the virus is moving throughout the country.

Tables and Figures

Table 1

Governorate	Weeks with Outbreaks	MSE	Percent Variance Explained	Slope ± SE	R-squared	χ^2 or Fisher Exact P-Value
Kalyoubia	88	1.66	60.26	1.17 ± .02	0.84	0.043
Giza	104	8.4	41.62	1.27 ± .03	0.85	0.690
Gharbia	108	0.84	34.65	1.38 ± .06	0.67	0.097
Fayoum	80	0.61	24.69	1.35 ± .07	0.51	0.291
Menoufia	125	1.94	19.46	1.36 ± .06	0.64	0.016
Sharkia	67	11.37	18.23	1.33 ± .03	0.83	0.024
Dakahlia	96	1.09	17.22	1.36 ± .07	0.56	0.002
Kafr el-Sheikh	50	0.61	14.94	1.53 ± .08	0.53	0.025
Beni Suef	48	0.22	13.22	1.31 ± .09	0.33	0.186
Minya	70	0.76	6.39	1.39 ± .07	0.51	0.150
Behera	55	0.54	4.79	1.63 ± .10	0.40	0.027
Aswan	13	0.09	3.45	1.69 ± .15	0.27	0.013
Cairo	16	0.11	2.37	1.81 ± .19	0.18	0.002
Qina	49	0.24	1.55	1.73 ± .12	0.38	0.029
Alexandria	34	0.26	1.24	1.66 ± .11	0.36	<0.001
Dumyat	45	0.32	-1.19	1.67 ± .11	0.32	0.212
Matruh	3	0.01	-1.36	3.01 ± .78	0.04	0.002
Red Sea	2	0.01	-1.39	2.25 ± .61	0.04	0.013
Asyut	7	0.03	-1.74	3.07 ± .61	0.07	<0.001
New Valley	5	0.03	-1.74	1.60 ± .09	0.48	<0.001
Port Said	8	0.02	-2.09	3.08 ± .46	0.11	0.029
Suez	13	0.05	-2.82	1.93 ± .18	0.22	0.037
Suhaj	20	0.15	-5.02	1.81 ± .08	0.53	<0.001
Ismailia	5	0.05	-9.5	2.13 ± .11	0.50	<0.001

Table 1 shows descriptive statistics and measures of model performance for each governorate. Variables included are: ***a)*** Weeks with observations since 2006 for each governorate, ***b)*** Mean square error and percent variance explained for each model, which are returned in the Random Forest regression output in R, ***c)*** The slope of the linear regression of observed values regressed on predicted values which determines not only whether week to week increases or decreases could be predicted, but also judges the accuracy of the predicted magnitude of outbreaks, ***d)*** R-squared judges how much the variation in the observations is explained by the relationship between the observed and predicted values. ***e)*** Chi-squared or Fisher's exact test p-values used to test whether or not the model could predict increases and decreases (regardless of magnitude) better than chance. *This table shows models for 24 governorates. There have been no outbreaks in the two governorates North and South Sinai, and thus no models were created for them.

Figure 1

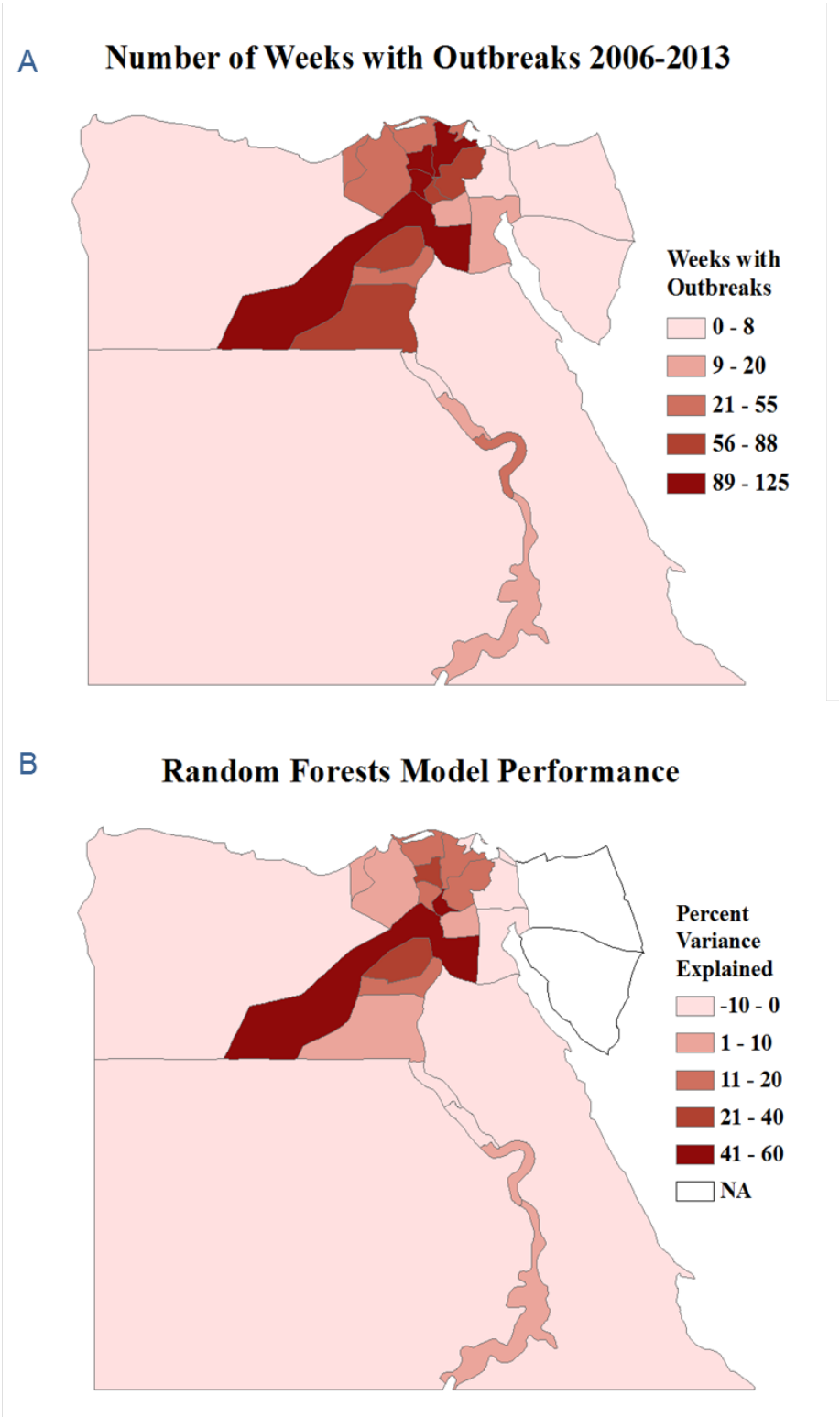

Figure 1 shows a geographical representation of the outbreak activity in each of the governorates (A) as well as Random Forests model performance based on percent of the variance explained (B). There have been no outbreaks detected in the Sinai peninsula at this time.

Figure 2

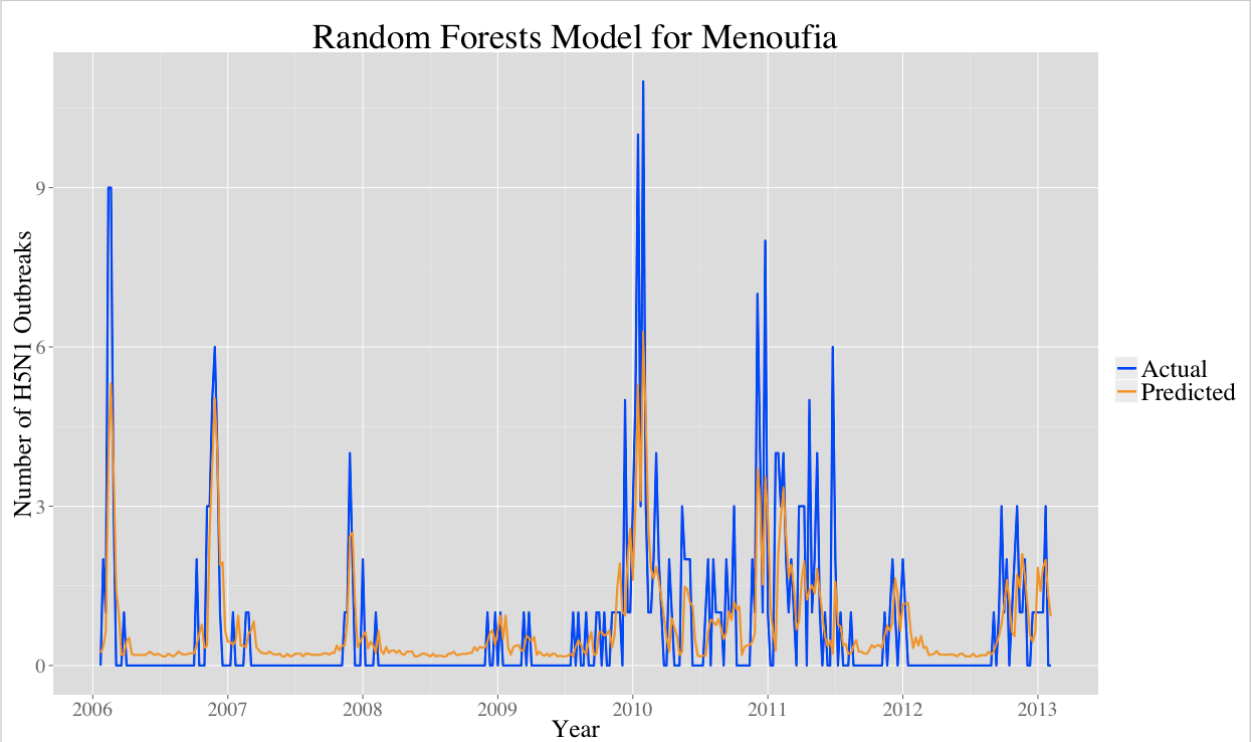

Figure 2 is an example of the time series data of the observed and predicted number of outbreaks given by the Random Forests model in the governorate of Menoufia.

Figure 3

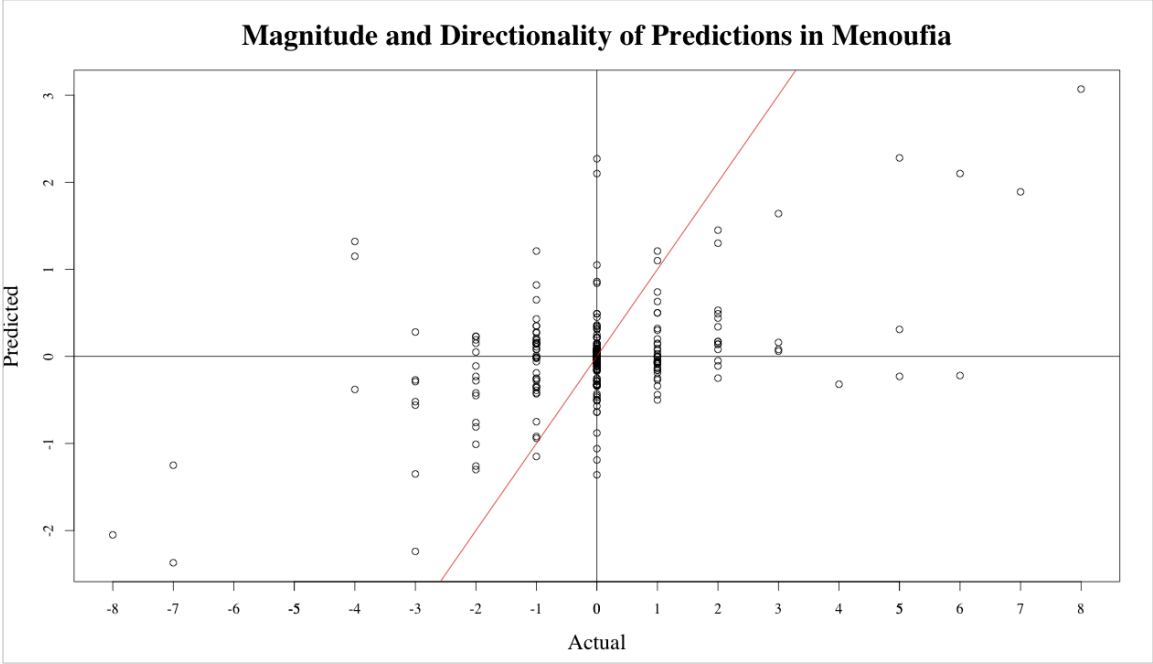

Figure 3 is a plot of the week to week increases and decreases in the actual and predicted observations for an example governorate, Menoufia. Points near the line were accurate predictions in both direction and magnitude. Points in the 1st and 3rd quadrants accurately predicted week by week increases or decreases. Points in the 2nd and 4th quadrants were predicted in the incorrect direction.

Figure 4

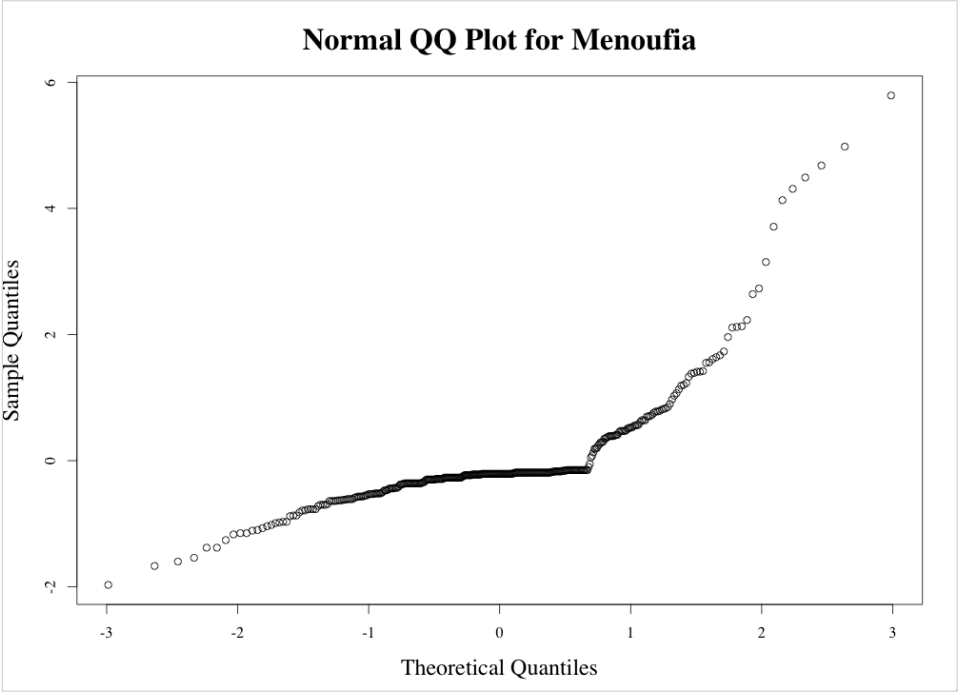

Figure 4 shows the normal QQ plot for the residuals for Menoufia's model. It demonstrates the typical underprediction of outbreaks in each model, as this hints at a residual histogram that is skewed to the right.

Figure 5

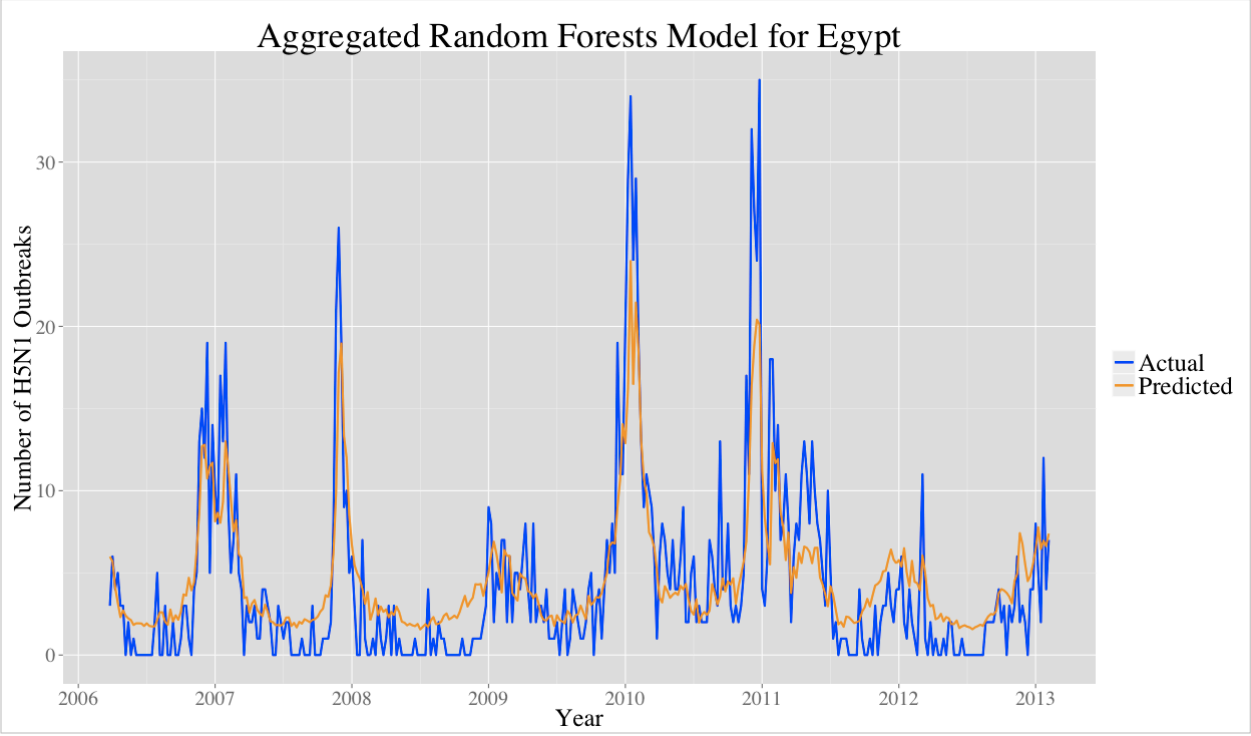

Figure 5 shows the countrywide weekly outbreak data since 2006 and the aggregate predictions from each governorates Random Forests model (although it does not include the initial outbreaks in 2006 as they were on such a larger scale the details in later years are less discernible if graphically represented)

References

[1] WHO. “Cumulative number of confirmed human cases for avian influenza A(H5N1) reported to WHO, 2003-2013.” WHO Influenza at the Human Animal Interface. (2013).
[2] El Nagar, Ahlam, and Ali Ibrahim. "Case study of the Egyptian poultry sector." Proceedings of the International Poultry Conference. (2007).
[3] Hassouneh, Islam, et al. "Food scare crises and developing countries: The impact of avian influenza on vertical price transmission in the Egyptian poultry sector." Food Policy 37.3 (2012): 264-274.
[4] Hosny FA. “Poultry sector country review. FAO animal production and health division.” Emergency Centre for Transboundary Animal Diseases: Socioeconomics, Production and Biodiversity Unit, (2006).
[5] FAO. "Approaches to controlling, preventing and eliminating H5N1 Highly Pathogenic Avian Influenza in endemic countries.” Animal Production and Health Paper. No. 171. Rome. (2011).
[6] Garrett, Laurie, and Steven Cook. "Egypt's Real Crisis: The Dual Epidemics Quietly Ravaging Public Health." Council on Foreign Relations (2012).
[7] Watanabe, Yohei, et al. "Acquisition of human-type receptor binding specificity by new H5N1 influenza virus sublineages during their emergence in birds in Egypt.” PLoS pathogens 7.5 (2011): e1002068.
[8] FAO. "Approaches to controlling, preventing and eliminating H5N1 Highly Pathogenic Avian Influenza in endemic countries.” Animal Production and Health Paper. No. 171. Rome. (2011).
[9] Hall, I. M., et al. "Real-time epidemic forecasting for pandemic influenza." Epidemiology and infection 135.03 (2007): 372-385.
[10] Sebastiani, Paola, et al. "A Bayesian dynamic model for influenza surveillance." Statistics in medicine 25.11 (2006): 1803-1816.
[11] Shaman, Jeffrey, and Alicia Karspeck. "Forecasting seasonal outbreaks of influenza." Proceedings of the National Academy of Sciences 109.50 (2012): 20425-20430.
[12] Zhang, Zhijie, et al. "Spatio-temporal data comparisons for global highly pathogenic avian influenza (HPAI) H5N1 outbreaks." PloS one 5.12 (2010): e15314.
[13] Keeling, Matt J., and Pejman Rohani. Modeling infectious diseases in humans and animals. Princeton University Press, (2011).
[14] Hay, Simon I., et al. "Big Data Opportunities for Global Infectious Disease Surveillance." PLoS medicine 10.4 (2013): e1001413.
[15] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
[16] Prasad, Anantha M., Louis R. Iverson, and Andy Liaw. "Newer classification and regression tree techniques: bagging and random forests for ecological prediction." Ecosystems 9.2 (2006): 181-199.
[17] Cutler, D. Richard, et al. "Random forests for classification in ecology." Ecology 88.11 (2007): 2783-2792.
[18] Robnik-Šikonja, Marko. "Improving random forests." Machine Learning: ECML 2004. Springer Berlin Heidelberg. (2004). 359-370.
[19] Kumar, Manish, and M. Thenmozhi. "Forecasting stock index movement: A comparison of support vector machines and random forest." (2006).
[20] Mcgovern, Amy, et al. "Understanding severe weather processes through spatiotemporal relational random forests." 2010 NASA conference on intelligent data understanding (to appear). (2010).
[21] Barnes, Geoffrey C., and Jordan M. Hyatt. "Classifying Adult Probationers by Forecasting Future Offending." (2012): 64.
[22] Araújo, Miguel B., and Mark New. "Ensemble forecasting of species distributions." Trends in ecology & evolution 22.1 (2007): 42-47.
[23] Prasad, Anantha M., Louis R. Iverson, and Andy Liaw. "Newer classification and regression tree techniques: bagging and random forests for ecological prediction." Ecosystems 9.2 (2006): 181-199.
[24] Furlanello, Cesare, et al. "GIS and the random forest predictor: Integration in R for tick-borne disease risk assessment." Proceedings of DSC. (2003).
[25] Goldstein, Benjamin A., Eric C. Polley, and Farren BS Briggs. "Random forests for genetic association studies." Stat Appl Genet Mol Biol 10.1 (2011): 32.
[26] The United Nations Food and Agricultural Organization: EMPRES-i Global Animal Disease Information System [http://empres-i.fao.org/eipws3g/]
[27] The Weather Underground: The Weather Underground Home Page [http://www.wunderground.com].

[28] Van Kerkhove, Maria D., and Neil M. Ferguson. "Epidemic and intervention modelling: a scientific rationale for policy decisions? Lessons from the 2009 influenza pandemic." Bulletin of the World Health Organization 90.4 (2012): 306-310.
[29] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria 2012, [http://www.R-project.org/].
[30] Liaw, Andy, and Matthew Wiener. "Classification and Regression by randomForest." R news 2.3 (2002): 18-22.
[31] Liaw, Andy, and Matthew Wiener. "Classification and Regression by randomForest." R news 2.3 (2002): 18-22.
[32] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
[33] Prasad, Anantha M., Louis R. Iverson, and Andy Liaw. "Newer classification and regression tree techniques: bagging and random forests for ecological prediction." Ecosystems 9.2 (2006): 181-199.
[34] Biau, Gérard. "Analysis of a random forests model." The Journal of Machine Learning Research 98888 (2012): 1063-1095.