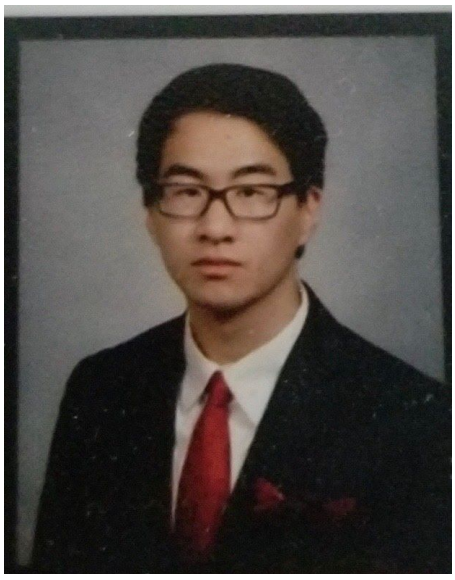
Predictive Modeling for Infectious Disease Dynamics in Influenza

Team 2

Fan Lu
Georgia Institute of Technology
flu40@gatech.edu

Hope Mumme
Georgia Institute of Technology
hopemumme@gatech.edu



INTRODUCTION

Compartmental models of infectious disease are extremely useful to the healthcare field today; they have been used by many prominent research groups to understand the mathematics behind the way influenza infects and spreads. Every year there is a different widespread influenza strain that causes damage to communities around the world. These mathematical models are an effective tool to help healthcare officials better prevent and manage disease outbreaks. Because of the unstructured and complex nature of medical data today, many research groups have trouble dealing with the large volume and variation of data as well as the amount of missing or incomplete entries in these datasets. Although compartmental models are the standard mathematical model for epidemiology, machine learning models have shown better results in predicting infectious diseases because of their ability to handle large amounts of data and interpolate missing entries. Our goals were to create a machine learning model to predict the dynamics of influenza, and to compare the results of this method against predictions made by the standard compartmental models. We also include alternative data sources such as climate and internet search data to factor in different spatio-temporal aspects of the spread of influenza.

After doing an extensive literature search, we decided on using a Random Forest (RF) Regression as our machine learning model and SIR (Susceptible, Infected, Recovered) and SEIR (Susceptible, Exposed, Infected, Recovered) for our compartmental model. We hypothesize that by incorporating climate data, internet search trends, and influenza infection counts into a Random Forest Regression model, we will be able to predict better than the standard SIR and SEIR compartmental models. We will also be comparing the prediction capabilities of the models with two countries, Norway and Australia. We chose these two countries because they experience different flu seasons and also have fairly constant climate within themselves. Therefore, our null hypothesis is that there will be no significant difference when comparing model results in country or model type.

LITERATURE CRITIQUE

In order to determine how standard SIR and SEIR studies are performed and what alternative methods exist to the standard SIR model that would be best for predicting influenza, we performed a literature search. By studying the articles, we learned about new variables to be included in analysis, different machine learning models we could use, and how to compare our model with the compartmental models.

A research group from IMATI in Spain attempted to predict influenza spread using climate data in 2018 [1]. The purpose of their study is to help healthcare professionals predict where outbreaks are more likely to happen in order to better allocate resources. They find that there is improved virus survival in lower temperatures and claim that by using climate data and flu incidence data they are able to accurately and efficiently predict influenza spread in Galicia, Spain. They used an iterative generalized least square model and applied it to a functional regression model. Although their results prove that temperature is a good predictor of influenza incidence, they are somewhat misleading in their title and initial claim; they make it seem like only meteorological data is being used to predict influenza spread, but further down the paper they specify that they are using climate data along with influenza counts to predict. From this study, we know that using climate data along with influenza counts could produce decent results in our study. We, however, compared our new model with the standard SIR and SEIR to see how much it improves, which is something the Spain study did not do.

Researchers Kane, Price, Scotch, and Rabinowitz from Yale University in 2014 compared Autoregressive Integrated Moving Average (ARIMA) and Random Forest (RF)

models in their prediction of H5N1 (avian influenza) outbreaks in Egypt [2]. The purpose of their study is to help healthcare officials in Egypt control the spread of H5N1 by using bird outbreaks to predict human outbreaks. They compare the retrospective model, in which they use all the data from 2007-2014, and simulated prospective model, in which they take a moving window and iteratively add a week on as they predict further ahead. They found that the RF prospective model was the best predictor and it used a 25 week moving window. The researchers concluded that since ARIMA assumes a linear relationship, it did not perform well with the non-linear H5N1 data. The researchers did not clarify how they chose the 25 week moving window and they did not compare their final model's predictive ability with models from other studies, so it was difficult to know how the model does compare to other ARIMA or RF models. This study helped us decide to use the RF model in our project and to use a prospective window, not a retrospective. We also chose the 8 week moving window, compared to their 25 week window, in our study because we had less data than the Kane et al group; they used 7 years of data and we used 2, so we decided that an 8 week window would be sufficient and proportional to their window/data ratio.

Another study by Sebastian Gonclaves and Marcelo F. C. Gomes from the Instituto de Fisica in Brazil in 2009 discusses the ability to model oscillating disease incidence rate using a SIRS model [3] . This model differs from the standard SIR by allowing the recovered population to lose immunity and become reinfected. They also assumed in their model that infected and recovered populations were required to stay in their group for a set amount of time, which they referred to as distributed delay. Using a bifurcation diagram, the researchers found that there was an oscillating region of solutions to the model based on the ratio of initial recovered population to the length of set time that the recovered and infected populations remain in their group. They found there was a lower and higher initial R that bounded these solutions and within this region the system would oscillate similar to the actual trends of flu data we have. This model would be the most accurate to try to fit to our data but it also has with being too sensitive to slight changes as they have a tendency to break to oscillating patterns.

In a paper by Angulo et al from the University of Grenada published in 2013, a spatiotemporal model was combined with a SIR model to predict the spread of Hand Foot and Mouth disease in eastern China [4]. They assumed an even initial distribution in a 20 by 20 grid edited to represent the geography and demographics or the target region. Then, by using a monte carlo simulation with the SIR model then simulated the spread of the disease. They found that in the first couple spans of time the model was accurate, but lacked long term accuracy when compared with real world data. In terms of using this model to represent our influenza data, it also only had two outcomes just like the standard SIR model, an endemic and a disease free solution. It would also be difficult to accurately model the regions of interest in our data so this method, while unique in its ability to not assume homogeneity of its population and outbreak location, would be only applicable to non-seasonal diseases and not influenza.

METHODS

Our project includes two different paths of analysis that converge at the final step. The first is the compartmental modeling that consists of the SIR and SEIR model. The second is the Random Forest (RF) Regression model. We decided to use RF for our machine learning model because it has shown to be more accurate than spatio-temporal models such as ARIMA in predicting types of influenza and because RF does not require the data to be linear. The guidelines for this project also require us to use a SIR and SEIR model, since they are among the most common ways to model disease. Our overall analysis pipeline is shown in Figure 1 below. We

will go over the specific details of each step in the following sections.

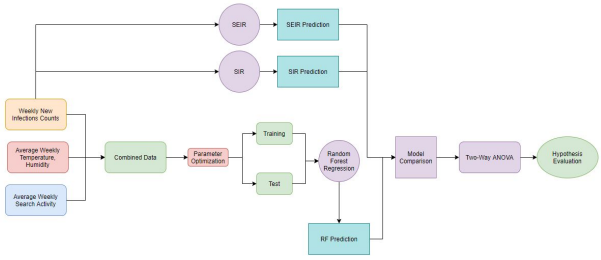


Figure 1. The workflow diagram for our analysis. We started with data preprocessing, then did parameter optimization for the RF model. Once we got the results from the 3 models we compared them using an ANOVA and evaluated our null hypothesis.

SEIR and SIR Compartmental Models:

For our assumptions, for both models, we will say that there are no births and no deaths during the course of our time span as the populations of the two countries are fairly stagnant. We will also assume that influenza is nonlethal since the website doesn't give us data on flu deaths and the death rate is negligible anyways.

Specifically for the SIR model, it was assumed that each person who was infected by the flu would be contagious for two weeks from the moment the virus entered their system [5]. For the SEIR model, the infected population would first enter the exposed state when they would eventually become infected but were not yet contagious. After a lag time or incubation period of two days, the entirety of the population would become infected. For both the SIR and SEIR models, it was also assumed that the recovered population would eventually lose immunity and become part of the susceptible population again. This was done to account for mutations in the flu over the course of different flu seasons.

Since the data from FluNet only included the number of positive tests by week, the S and R values had to be extrapolated from the data. The current. The I values by week were simply the sum of the current and previous weeks worth of positive tests from before two weeks prior to a given date as the model assumed that there were no deaths natural or otherwise. After 52 weeks, a recovered person would lose immunity and become part of the S group once again.The S was the total population, N, minus the current I and R values.

$$dS = -SI/N + \xi R$$
$$dI = SI/N - I$$
$$dR = I - \xi R$$
$$N = S + I + R$$

Equations 1. ODEs (Ordinary Differential Equations) used to represent the SIR model. β,ξ, and □ represent constants to be determined by the model while N is the total population of the study group.

Once the SIR values by week were extracted from the dataset, a set of ODEs (Equations 1) were written in Matlab using the ode15s function, which is used to solve low order non-stiff differential equations. Compared to a classic SIR model, this model made two changes. The β parameter, which can be thought about as the infection rate constant, was multiplied by 1/N to account for the large discrepancy of populations between different countries while also improving the performance of the curve fitting function to be described later. Secondly, ξ was used to describe the rate of loss of immunity and multiplied by the current population of R. γ represents the recovery rate and was unchanged from the standard SIR model.

The next step was to use the lsqcurvefit function to estimate the three parameters. The function requires initial parameters to be inputted so they were estimated in matlab using the ODEs and the first two weeks of data. As the function is highly sensitive to initial conditions, these initial estimations were randomized and the curve fitting function was repeated 100 times, with only the set of constants with the lowest MSE (mean squared error) saved and outputted.

Using these final parameters, the SIR values were predicted using the initial conditions and the ODEs. Finally, a MSE analysis was done between just the predicted and actual I values

since the infected population is the only shared group amongst all three models.

The SEIR model used the same methods to extract the S, I, and R values from the data. The E population was interpolated by multiplying the new positive tests per week by 2/7 since the data was by week and the incubation period of influenza was assumed to be two days.

$$\begin{aligned} dS &= -SI/N + \xi R \\ dE &= SI/N - E \\ dI &= E - I \\ dR &= I - \xi R \\ N &= S + E + I + R \end{aligned}$$

Equations 2. ODEs used to represent the SEIR model. β,ξ, σ,and □ represent constants to be determined by the model.

The ODEs (Equations 2) were again implemented into Matlab using ode15s with the same changes to the SIR model as before except with the E population included. In these set of ODEs, β, ξ, and γ represent the same concepts as in the SIR model while σ represents the rate of the exposed population becoming infected, or the constant of the lag time. The same process was repeated with the lsqcurvefit function except with the four parameters of the SEIR function and a MSE of the I values only was calculated.

Both the SIR and SEIR models were run with influenza data from Norway and Australia from January 1st 2018 to March 1st 2020, or 118 weeks worth of data.

Random Forest (RF) Regression Model:

RF Regression is a supervised learning algorithm that uses the bagging technique to reduce variation between decision trees and prevent overfitting. The final model, g(x), is an additive model that combines predictions from multiple decision trees $f_n(x)$. The number of trees created is represented by B. Equation 3 below shows how the decision trees make predictions, Equation 4 shows how the predictions are combined to make the final prediction.

$$f_0(x) = \frac{1}{B}\sum_{b}^{B} T_b(x)$$

$$g(x) = f_0(x) + f_1(x) + \ldots + f_n(x)$$

Equations 3 and 4. $f_0(x)$ represents the prediction from each decision tree, and they are combined in Equation 4 to make the final prediction from the forest, g(x).

Each decision tree uses training vector X and label vector y to partition the trees so that samples with similar labels are grouped together. A node represents the decision at each tree. If the split at node m is represented by Θ, then $\Theta = (j,t_m)$, with j representing the feature and t_m representing the threshold. The model then partitions the data into two subsets, $Q_{left}(\Theta)$ and $Q_{right}(\Theta)$, shown in equation 5. The impurity at node m is calculated using the impurity function H, shown in equation 6. The model finds the parameters that minimize the impurity, or the variance within a partition, equation 7. The impurity function, H, for RF regression is mean square error (MSE), X_m is the training data for node m, shown in equations 8 and 9.

$$\begin{bmatrix} Q_{left}(\theta) = (x,y)|x_j <= t_m \\ Q_{right}(\theta) = Q / Q_{left}(\theta) \end{bmatrix}$$

$$G(Q,\theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))$$

$$\theta^* = argmin_\theta G(Q,\theta)$$

$$\bar{y}_m = \frac{1}{N_m}\sum_{N_m}^{i} y_i$$

$$H(X_m) = \frac{1}{N_m}\sum_{N_m}^{i} (y_i - \bar{y}_m)^2$$

Equations 5-9. Equation 5 represents the calculation of the subsets, Q_{left} and Q_{right}. The impurity function, H, is shown by equation 6. In equations 7-9, the impurity function is optimized to reduce the variance between partitions.

Our data for the RF Regression model consists of weekly infection counts from FluNet, published by the World Health Organization [6], average weekly temperature from both Norway and Australia, published by climate data online [7-8], and average weekly search activity of the terms “influenza” and “flu” from Google Trends [9]. We used the weekly data from January 1st, 2018 until March 1st, 2020 to predict. For the RF model, we combined the data into a matrix in excel shown below to prepare for parameter optimization and training-testing split. Each week’s true infection counts are being predicted using 24 features: 8 previous weeks of infection counts, temperature, and search activity data. So for example, Week 9 influenza counts are predicted using Weeks 1-9 data. Once the excel sheets are loaded into python using the pandas package, they are converted into pandas data frames for processing.

Flu Total	Week Num	1 Wk Flu	2 Wk Flu	3 Wk Flu	4 Wk Flu	5 Wk Flu	6 Wk Flu	7 Wk Flu	8 Wk Flu	1 Wk Temp	2 Wk Temp	3 Wk Temp	4 Wk Temp	5 Wk Temp	6 Wk Temp	7 Wk Temp	8 Wk Temp	1 Wk Goo	2 Wk Goo	3 Wk Goo	4 Wk Goo	5 Wk Goo	6 Wk Goo	7 Wk Goo	8 Wk Goo
428	201053	421	381	461	607	662	867	1002	891	12.8	13.6	16	14	13.4	15	15.9	13.7	107	124	102	107	105	145	146	132
476	201054	428	421	381	461	607	662	867	1002	12.7	12.9	13.6	16	14	13.4	15	15.9	113	107	124	102	107	105	143	146
359	201055	476	428	421	381	461	607	662	867	13.2	12.7	12.9	13.6	16	14	13.4	15	105	113	107	124	102	107	105	145
294	201056	359	476	428	421	381	461	607	662	13.2	13.2	12.7	12.9	13.6	16	14	13.4	91	103	113	107	124	102	107	105
224	201057	294	359	476	428	421	381	461	607	11.7	13.2	13.2	12.7	12.9	13.6	16	14	75	91	103	113	107	124	102	107
181	201058	224	294	359	476	428	421	381	461	16.3	11.7	13.2	13.2	12.7	12.9	13.6	16	61	75	91	103	113	107	124	102
159	201059	181	224	294	359	476	428	421	381	11.3	16.3	11.7	13.2	13.2	12.7	12.9	13.6	51	61	75	91	103	113	107	124
107	201060	159	181	224	294	359	476	428	421	16.9	11.3	16.3	11.7	13.2	13.2	12.7	12.9	40	51	61	75	91	103	113	107
100	201061	107	159	181	224	294	359	476	428	21.1	16.9	11.3	16.3	11.7	13.2	13.2	12.7	34	40	51	61	75	91	103	113
59	201062	100	107	159	181	224	294	359	476	18.7	21.1	16.9	11.3	16.3	11.7	13.2	13.2	30	34	40	51	61	75	91	103
45	201063	59	100	107	159	181	224	294	359	28	18.7	21.1	16.9	11.3	16.3	11.7	13.2	28	30	34	40	51	61	75	91
34	201064	45	59	100	107	159	181	224	294	28.5	28	18.7	21.1	16.9	11.3	16.3	11.7	25	28	30	34	40	51	61	75
22	201065	34	45	59	100	107	159	181	224	27.4	28.5	28	18.7	21.1	16.9	11.3	16.3	26	25	28	30	34	40	51	61
23	201066	22	34	45	59	100	107	159	181	19	27.4	28.5	28	18.7	21.1	16.9	11.3	27	26	25	28	30	34	40	51
28	201067	23	22	34	45	59	100	107	159	26.8	19	27.4	28.5	28	18.7	21.1	16.9	22	27	26	25	28	30	34	40
40	201068	28	23	22	34	45	59	100	107	32.5	26.8	19	27.4	28.5	28	18.7	21.1	18	22	27	26	25	28	30	34
94	201069	40	28	23	22	34	45	59	100	27.3	32.5	26.8	19	27.4	28.5	28	18.7	21	18	22	27	26	25	28	30
42	201070	94	40	28	23	22	34	45	59	20.8	27.3	32.5	26.8	19	27.4	28.5	28	20	21	18	22	27	26	25	28
38	201071	42	94	40	28	23	22	34	45	38.1	20.8	27.3	32.5	26.8	19	27.4	28.5	21	20	21	18	22	27	26	25
28	201072	38	42	94	40	28	23	22	34	32.5	38.1	20.8	27.3	32.5	26.8	19	27.4	19	21	20	21	18	22	27	26
42	201001	28	38	42	94	40	28	23	22	28.8	32.5	38.1	20.8	27.3	32.5	26.8	19	18	19	21	20	21	18	22	27
60	201002	42	28	38	42	94	40	28	23	33.2	28.8	32.5	38.1	20.8	27.3	32.5	26.8	24	18	19	21	20	21	18	22
59	201003	60	42	28	38	42	94	40	28	34.9	33.2	28.8	32.5	38.1	20.8	27.3	32.5	24	24	18	19	21	20	21	18
113	201004	59	60	42	28	38	42	94	40	33.2	34.9	33.2	28.8	32.5	38.1	20.8	27.3	43	24	24	18	19	21	20	21
84	201005	113	59	60	42	28	38	42	94	34.3	33.2	34.9	33.2	28.8	32.5	38.1	20.8	100	43	24	24	18	19	21	20
80	201006	84	113	59	60	42	28	38	42	32	34.3	33.2	34.9	33.2	28.8	32.5	38.1	89	100	43	24	24	18	19	21
87	201007	80	84	113	59	60	42	28	38	29.1	32	34.3	33.2	34.9	33.2	28.8	32.5	55	89	100	43	24	24	18	19
82	201008	87	80	84	113	59	60	42	28	28.2	29.1	32	34.3	33.2	34.9	33.2	28.8	48	55	89	100	43	24	24	18
72	201009	82	87	80	84	113	59	60	42	23.1	28.2	29.1	32	34.3	33.2	34.9	33.2	107	48	55	89	100	43	24	24

Table 1. Processed data for Australia used in RF Regression Model. Each week’s true counts shown in red are being predicted using 24 features: 8 previous weeks of influenza counts, average temperature, and google search activity.

After importing and processing the data, we performed parameter optimization to find the optimum number of decision trees, represented by the variable n_estimators, for each country. We used the sklearn train_test_split package to split the data into training and testing sets. Then, we used the sklearn package randomForestRegressor to optimize the n_estimator value. In order to optimize n_estimator, we tested a range of values from 10 to 10,000 and trained a specific RF model just for optimization. We found the Out-Of-Bag (OOB) error of each iteration and used that to determine the best n_estimator value; the n_estimator value that produced the highest OOB score was decided as the optimum value. OOB error is a method that is used to validate and optimize random forest models; it is defined as the number of correctly predicted rows from the out of bag sample [10]. The OOB score only uses a subset of decision trees in the model to prevent overfitting and data leakage.

After determining the optimum number of decision trees to use in the RF Regression model, we then trained and tested a separate, new model for each country. We trained the model using the randomForestRegressor function using the n_estimator value found in our parameter optimization. After training, we tested the new model using the testing set and calculated the mean square error as well as the R^2 using the predicted values and the actual values to evaluate our model. To visualize our results, we also create a plot showing the actual versus predicted values of the test data for each country’s model using the matplotlib package from python.

Evaluation and Comparison:

Once the predictions for each window are calculated for the SIR/SEIR models and the RF model, we will compare them. We calculate the mean square error (MSE) between the actual and predicted counts for each model and average them for all the weeks. Using a two-way ANOVA test in excel, we compare the predictive abilities of the models based on two factors, country and model type. Our null hypothesis is that there will be no significant difference between the MSE of the RF and compartmental models. If the p value is less than 0.05, then we will reject the null hypothesis. We will also compare the prediction abilities of our model in Australia versus in Norway. Because there are different strains of influenza in the northern and southern hemispheres, we expect the models to have different prediction accuracies in each country. Our second null hypothesis is that there will be no significant difference in the models based on country. Again, if the p-value from the ANOVA test is less than 0.05, then we will reject the null hypothesis.

RESULTS

SIR and SEIR Compartmental Models:

The SIR model calculated the constants displayed in Table 2 below. Norway had a higher infectivity rate per person relative to Australia, represented by β. Norway's loss of immunity rate, ξ, is also much higher than Australia's, which is basically zero. The recovery rate for Norway, □, is also higher than the same for Australia, by about the same ratio as the differences in infectivity rates. The MSE for Norway was higher than that of Australia by a factor of about 100, despite its lower population.

	Norway	**Australia**
β	3.119	2.434
ξ	.0722	$2.366*10^{-14}$
γ	3.095	2.417
MSE	$1.73*10^{6}$	$2.123*10^{4}$

Table 2. Calculated parameter and MSE values of the SIR function. The values are all dimensionless as they are constants.

In Figures 2 and 3 below, the given and predicted SIR data are graphed for Norway and Australia respectively. While the predicted data fits the curvature of the S and R data reasonably well, the predicted I values fail to capture the seasonal spikes in infections seen in the given data for both countries.

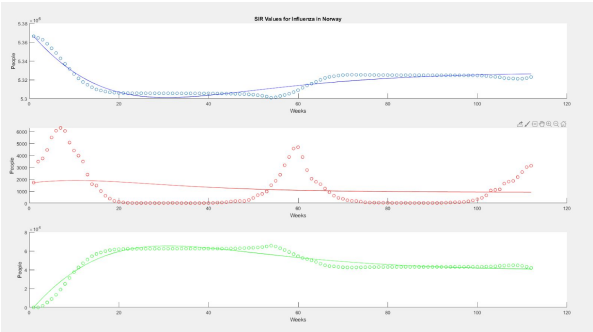


Figure 2. Graph of the given S (blue), I (red), and R (green) values, represented by dots, and the predicted values, represented by smooth curves of the same color. The given data is from Norway from January 1st 2018 to March 1st 2020 and the separation of the graph into three sections is due to the high difference in y values between the S, I and R.

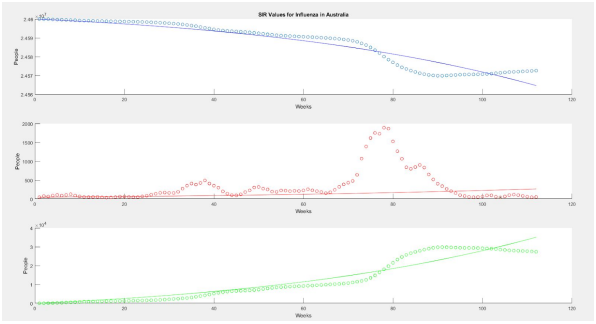


Figure 3. Graph of the given S, I and R values for Australia from January 1st 2018 to March 1st 2020, as well as the predicted values represented by the smooth curves.

The SEIR model calculated the constants displayed in Table 3 below. Norway had a lower infectivity rate per person relative to Australia in this model. This is opposite to the trend found in the SIR model. The lag period constant, σ, was much higher in Norway. Again, Norway's loss of immunity rate is much higher than Australia's, which is higher than the SIR model but still about zero. The recovery rate for Norway is lower than the same for Australia, by about the same ratio as the differences in infectivity rates. The MSE for Norway was higher than that of Australia by a factor of about 100, which matches the difference found in the SIR model.

	Norway	**Australia**
β	1.514	1.814
σ	$1.5*10^{3}$	566.62
ξ	.0189	$5.26*10^{-5}$
γ	1.525	1.796

MSE	$1.008*10^6$	$4.06*10^4$

Table 3. Calculated parameter and MSE values of the SEIR function.

In Figures 4 and 5 below, the given and predicted data SEIR are graphed for Norway and Australia respectively. Just like in the SIR model, the S and R fitted curves matched the fitted data fairly well. The E and I curves failed to model the seasonal spike in infection and exposed population counts in both countries.

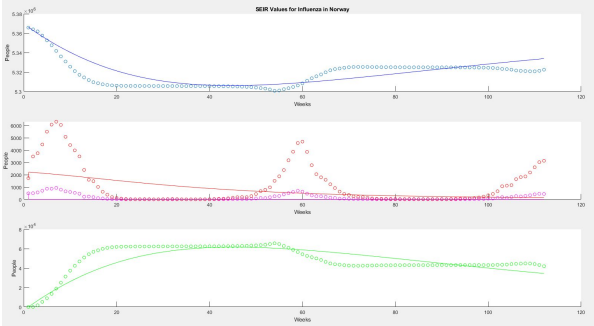


Figure 4. Graph of the given S (blue), E (magenta), I (red), and R (green) values, represented by dots, and the predicted values, represented by smooth curves of the same color. The given data is from Norway from January 1st 2018 to March 1st 2020 and the separation of the graph into three sections is due to the high difference in y values between the S, E, I and R, with E and I on the same axis.

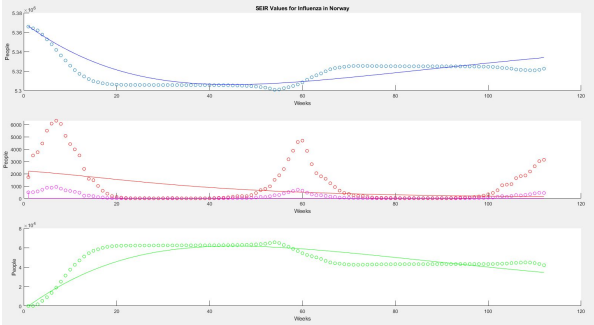


Figure 5. Graph of the given S,E, I, and R values for Australia from January 1st 2018 to March 1st 2020, as well as the predicted values represented by the smooth curves.

Random Forest (RF) Regression Model:

The results of the parameter estimation step for the RF Regression model are shown below in Table 4. The best n_estimator value for Norway was 3000 and that produced an OOB score of 0.8741; for Australia, the best n_estimator was 1000 and that produced an OOB score of 0.9159.

	Norway	**Australia**
Best n_estimator	3000	1000
OOB Error	0.8741	0.9159
MSE	25368	6853

Table 4. Parameter Optimization results for the RF Regression model for Norway and Australia. N_estimator refers to the optimum number of decision trees used based on the OOB score. The Mean Square Error (MSE) is also given.

The optimized RF Regression model for Norway produced an MSE of 27,380 , an R^2 of 0.8731 and an OOB Score of 0.9192. The model for Australia produced an MSE of 2153, an R^2 of 0.9083 and an OOB Score of 0.9233. The actual versus predicted influenza count values for each country's testing data are shown below in Figures 6 and 7.

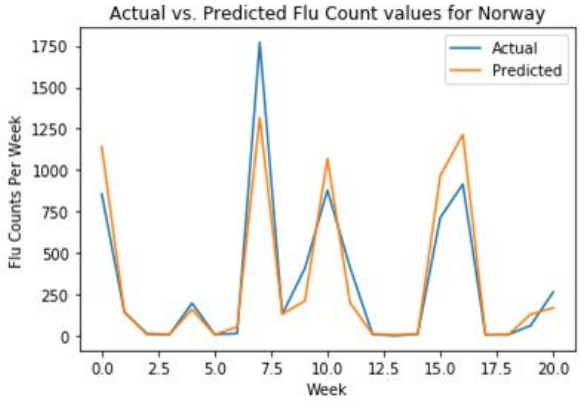


Figure 6. Actual vs. Predicted Flu Count values for Australia's optimized RF Regression model. The x-axis

is the week being predicted and the y-axis is the Flu Counts for that week.

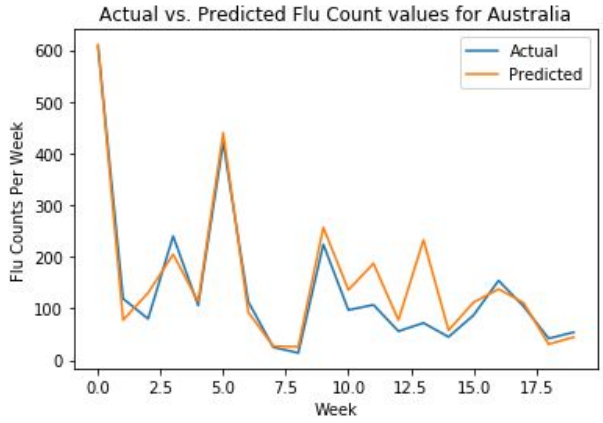


Figure 7. Actual vs. Predicted Flu Count values for Norway's optimized RF Regression model. The x-axis is the week being predicted and the y-axis is the Flu Counts for that week.

Evaluation and Comparison:

The Two-Way ANOVA comparing the compartmental models and the RF model as well as the countries is shown in Table 5 below. The p-value for country as the source of variation was 0.213, and the p-value for model type was 0.494.

SUMMARY	*Count*	*Sum*	*Average*	*Variance*		
Norway	3	2751474	917158	7.42805E+11		
Australia	3	74586	24862	203462332		
SIR	2	1751200	875600	1.46E+12		
SEIR	2	1048600	524300	4.67931E+11		
RF	2	26260	13130	236672		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1.19429E+12	1	1.19429E+12	3.255778839	0.212938	18.51282
Columns	7.52374E+11	2	3.76187E+11	1.0255327	0.493697	19
Error	7.33642E+11	2	3.66821E+11			
Total	2.6803E+12	5				

Table 5. Two-Way ANOVA test with country and model type as factors.

DISCUSSION

The SIR and SEIR models lack the ability to take into account the cyclical nature of the flu. They are much better suited to predicting the nature of acute outbreaks over short periods of time as opposed to long term recurring diseases that affect a relatively small portion of populations. Without the incorporation of a changing infectivity rate, the models are destined to either end in either a runaway number of infections or a elimination of the disease

The high MSEs between the predicted and actual infection counts were also due to the nature of curve fitting function. Due to the higher number of recovered and susceptible populations compared to infected and exposed populations, the fit prioritizes making the parameters create curves that fit the S and R data better as it leads to solutions with lower MSEs. This leads to the function sacrificing the accuracy of the predicted E and I values in favor of better S and R predictions.

For the RF Regression model, we found that the MSE values were higher than in literature for both countries. For instance, Cobrun et al had an average MSE of 21.38 for their LSTM model and 44.69 for their ARIMA model [11]. We believe that when compared with the literature models, our infection counts were relatively higher so that could contribute to the higher MSE; however, our model has shown it needs improvement in many aspects. Norway with an MSE value of 27,380 was much higher than the MSE for Australia, 2,153. We believe that this could be due to Norway having consistently larger infection counts each day. The RF model also had trouble predicting sudden changes in infection counts, which could be due to not having a large enough training set.

Since the p-values for both variance in country and variance in model type were greater than 0.05, we fail to reject the null hypothesis that there is no significant difference. We were surprised to see that there was no significant difference found by the ANOVA since the average MSE values for each country and each model type were fairly different from each other; as shown in Table 3, the average MSE value for the RF model was 13,130 whereas the average values for SIR and SEIR were 876,600 and

524,300 respectively, and the Norway average MSE as 917,158 but the Australia average was much lower at 24,862. We believe the ANOVA test found no significant difference because the sample size was very low, there were only 6 models total.

Overall, we found our approach was not as successful as previous studies but it was novel in its methods of predictive modeling of influenza using random forest regression as well as in using compartmental models and machine learning to compare the predictive capabilities of data from the northern versus the southern hemisphere. We learned that when predicting influenza spread it is important to incorporate and adjust for the seasonal aspect of the disease. In the future we plan to incorporate temporal calculations into our compartmental models. Also, we are going to use our RF Regression machine learning model to compliment our compartmental models to try and accommodate for the seasonal aspect of influenza. In order to better predict using our machine learning model, we also plan to narrow down which feature best predicts the influenza counts for each week; this information is important for our future models.

REFERENCES

[1] Oviedo de la Fuente M, Febrero-Bande M, Muñoz MP, Domínguez À. Predicting seasonal influenza transmission using functional regression models with temporal dependence. *PLOS ONE,* 13(4), April, 2018.

[2] Kane, M.J., Price, N., Scotch, M. et al. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics,* 15(276), August 2014.

[3] Gonçalves, Sebastian & Abramson, Guillermo & Ferreira da Costa Gomes, Marcelo. Oscillations in SIRS model with distributed delays. *The European Physical Journal B*, 81, 2009

[4] Angulo J, Langousis A, Kolovos A, Wang J, Madrid AE, Christakos G. Spatiotemporal Infectious Disease Modeling: A BME-SIR Approach. *PLOS*, September 2013.

[5] Moghadami M. A Narrative Review of Influenza: A Seasonal and Pandemic Disease. *Iranian journal of medical sciences*, 42(1), January 2017.

[6] World Health Organization. FluNet. https://www.who.int/influenza/gisrs_laboratory/flunet/en/

[7] National Climatic Data Center. Climate Data Online. https://www7.ncdc.noaa.gov/CDO/dataproduct

[8] Australian Government. Climate Data Online. http://www.bom.gov.au/climate/data/

[9] Google. Google Trends. https://trends.google.com/trends

[10] scikit -learn.org. OOB Errors for Random Forests. https://scikit-learn.org/stable/auto_examples/ensemble/plot_ensemble_oob.html

[11] Coburn, B.J., Wagner, B.G. & Blower, S. Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1). *BMC Med,* 7(30), 2009.