Influenza Trends for Counties in the State of Montana

Almandub Alanazi; Cassidy Alexander; Baleigh Doyle; Anna Gasner;
Bill Griffin; Michelle Howell; Matthew Malloy; Jakob Oetinger;
Apsara Rodriguez; Evrard Sinarinzi; Craig Stahlberg; Vanna Tran;
Professor Brian Steele

Department of Mathematical Statistics
University of Montana
Missoula, Montana
May 2, 2019

Contents

1. Executive Summary

The graduate and undergraduate students of Professor Brian Steele's Data Science Projects class participated in a project to build a system for forecasting influenza occurrences in Montana. The project was supported by Erin Landguth (in the Montana Department of Public Health and Human Services) with assistance from Professors Jon Graham and Emily Stone. Graduate students Ben Stark and Eli Bayat-Mokhtari also provided assistance.

Professor Landguth's research question targeted assessing the effect of airborne pollutants in determining the incidence of influenza in the state of Montana. The data used in this research centered on flu incident counts that were collected locally and remitted weekly to the Montana Department of Public Health and Human Services (DPHHS), in Helena. These data comprise the foundation data set for this study. Weekly reports, in the form of Montana-County maps (with flu season cumulative maps) are presented on the DPHHS website each week and are the most current publicly available flu incident data. Historical data from flu season 2009-2010 and subsequent to early January 2019 of the 2018-2019 flu season was provided by Stacey Anderson, Communicable Disease Epidemiologist with the Communicable Disease Control and Prevention Bureau.

The Data Science Projects class separated into three groups with responsibilities for gathering data, developing analytical forecasting algorithms, and accuracy assessment. Explanatory variables included utilizing past air quality data and availability of flu vaccination records. The class quickly concluded that the only relevant data available with county-specific significance were the flu incident records coupled with county census estimates for each of the flu seasons. Numerous options for utilizing this data were identified. For example, the use of prior year data to forecast the current season was investigated; the use of contiguous county data to forecast an incidence in a target county, was investigated. Information from prior weeks of the current flu year proved to have the most predictive value. Our efforts focused on building algorithms using variables.

Several methods were developed with these data. These methods will be explained in the body of this writing starting in Section 3.3. The assessment of these methods re-lead to recommendations regarding those with the best forecasting accuracy.

2. Introduction

The world experiences waves of seasonal flu, also known as influenza, each year. The flu occurs during the winter season in outbreaks and epidemics (Zachary, 2019). Symptoms of the flu are unique to each person; however, the symptoms typically include a fever (temperature higher than 100 degrees Fahrenheit or 37.8 degrees Celsius), headache and muscle aches, fatigue and/or a cough and sore throat (Zachary, 2019). These symptoms usually diminish in two to five days, but weakness and fatigue can last several weeks.

The flu can affect individuals differently depending on their health and residence. Residents of nursing homes, pregnant women, individuals with chronic medical conditions, Native Americans and Alaska Natives, and people with extreme obesity are a few categories of populations with high risk complications for influenza (Zachary, 2019). According to the National Center for Health Statistics, around 200,000 people in the United States are hospitalized because of influenza each year (Zachary, 2019). The number of deaths caused by influenza is about 2,905 per 100,000 population annually in the U.S. (Zachary, 2019). There were 79 reported deaths caused by influenza in the 2017-2018 flu season in Montana.

Influenza activity occurs seasonally, beginning in September or October. Flu activity peaks between December and March, and can last as late as May. In Figure 1, the flu trends from July 2003 to March 2008 of Google Flu Trends and CDC Influenza-like Illness are compared to the CDC Virologic Surveillance (Ortiz, 2011). The x-axis represents the month and year of the positive influenza tests. The y-axis on the right side of the figure is the percent of positive influenza tests, and the y-axis on the left side is the percent of the Influenza-like Illnesses. The beginning of the flu season can be seen when the graph begins to increase positively. According to Figure 1, the flu season begins between September and October, and the flu season peaks between December through March depending on the year (Ortiz, 2011).

In the state of Montana, DPHHS produces a report through the influenza season containing information about influenza each week. The report includes newly reported influenza cases, influenza cases by week, influenza hospitalizations and deaths, and information on Respiratory Syncytial Virus (RSV). The University of Montana spring 2019 Data Science Projects course created forecasting models to predict influenza by county in the state of Montana using the reports from DPHHS for the 2018-2019 flu season.

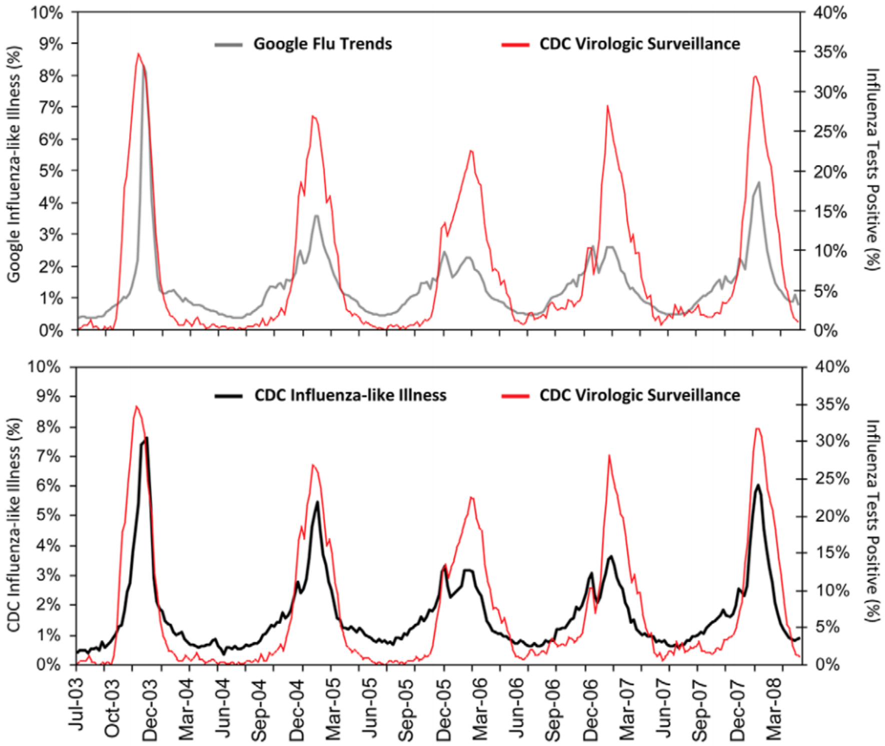


Figure 1. US influenza trends from July 2003 to March 2008 (Ortiz, 2011). Google Flu Trends and CDC Influenza-like Illness (black) are compared to the CDC Virologic Surveillance (red). The x-axis represents the month and year of influenza tests. The y-axis on the right side of the figure is the percent of positive influenza tests and the y-axis on the left side is the percent of the influenza-like illness. The U.S. influenza season begins between September and October and peaks between December and March.

3. Methods

3.1. **Data Sources.** Influenza can be predicted through several data sources. However, a challenge that the class faced over the semester was discovering available data. One of the first data sets the class searched for was vaccination data for Montana counties. Vaccination rates could provide insight into the severity of influenza for each season. If more individuals were vaccinated before the flu season began, the number may be lower for the given flu season. The vaccine data could also provide information on how effective the current vaccine is for preventing or reducing the rates of influenza. Unfortunately, DPHHS does not provide vaccine information for the state of Montana, so the class could not use this information to predict influenza. Research into other means of obtaining this data is ongoing.

Furthermore, influenza can be made worse by several environmental factors. One available data set is on particulate matter (PM). Particulate matter is a combination of solid and liquid droplets found in the air (USEPA, 2019). There are two types of particulate matter pollution: PM10 and PM2.5. PM10 is an inhalable particulate with a diameter of 10 micrometers and smaller, while PM2.5 is a fine inhalable particulate with a diameter of 2.5 micrometers and smaller (USEPA, 2019). Since the particulate matter is so small, it is easily inhaled and can create major health problems. These particles create an air pollution haze in the United States, or a temperature inversion that commonly occurs in valleys such as Missoula. Using the PM2.5 data with influenza counts could provide insight into accurately predicting influenza rates for the upcoming week. The Data Science Projects class had access to environmental factor data, including PM2.5; but limitations on sites relative to the 56 counties, coupled with availability of our dispersion records, made use of this data impractical. The class chose not to use the data to produce a forecasting method for influenza for the state of Montana.

The Data Science Projects class considered using similar spatial units to predict influenza rates. Specifically, the class researched the effect of transportation corridors on the rate of spread of influenza. Montana is filled with several national parks, monuments, and universities. Major Montana cities including Great Falls, Missoula, Billings, Bozeman, Butte, and Helena are connected by an interstate highway system. On average, it takes about two hours to reach one major city from another. The class investigated using the transportation corridors to predict influenza rates based on the traffic rates between counties. For example, Flathead County and Glacier County attract a large population of tourists throughout the year. Visitors can spread influenza even if they are not aware they are contagious with the flu. Missoula, Gallatin, and Beaverhead Counties have several school systems including college campuses. Populations like these travel regularly for sporting events, holidays, special occasions, and breaks. Studying the transportation rates during the flu season with these events also could provide insight for predicting influenza. Due to time restrictions, the Data Science Projects class did not use similar spatial units to predict influenza rates for the 2018-2019 flu season.

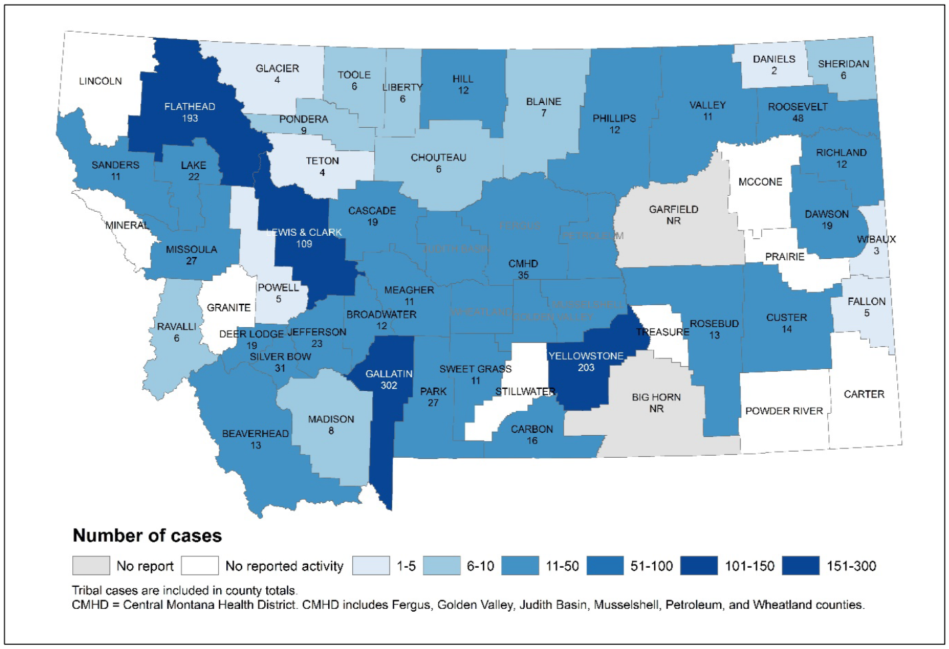
Each year, the Center for Disease Control and Prevention (CDC) measures the severity of influenza seasons to guide public health actions. Flu severity is assessed by considering the percentage of visits to outpatient clinics for influenza-like illnesses, influenza-associated hospitalization rates, and the percentage of deaths resulting from pneumonia or flu (CDC, 2019). By using these criteria, a flu season can be classified as low, moderate, high, or very high depending on how many of the indicators fall within predetermined intensity thresholds. Intensity thresholds are developed using data from past flu seasons to assess the chance that influenza will cross that threshold. By determining whether the indicator values crossed the intensity thresholds during the peak of each flu season, the CDC classifies the severity.

For this project, only the rate of visits to outpatient clinics for influenza-like illnesses was considered. Because of this, the CDC's severity classification cannot be applied to the data. Using only one indicator would, by default, classify all flu seasons in Montana to the low severity classification. The CDC further separates its data into three age groups: children, adults, and older adults (CDC, 2019). Age data was not collected for this project.

Finally, the Data Science Project class tried using past flu season years to predict future flu seasons. Due to the extreme variation of flu season from year to year, the use of this data source did not provide helpful insight to predict the 2018-2019 flu season. The variation in environmental factors, transportation rates, population size, vaccination rates, and severity of the flu each year made it difficult to draw conclusions about the current flu season from past flu data.

3.2. **Data Staging.** A data set was shared among classmates with weekly influenza flu counts for each Montana county from 2009 to early 2018. The data set was stored in an Excel Workbook and included the year, month, flu season, flu week, Morbidity and Mortality Weekly Report (MMWR) week, week start/end, newly reported counts, rates of influenza by counties, and the population for the flu season. The data set was converted to a Comma Separated Values (.csv) file for the forecasting function. The 2009-2010 flu season was removed from the data set due to inconsistencies. Several instances of counts "-9999" in the data set were replaced with interpolated values. For the purposes of the project, non-flu season weeks were removed from the data set.

For the 2018-2019 flu season, newly reported influenza cases for Montana counties were uploaded weekly by DPHHS. An example of the Montana counties with newly reported cases can be seen in Figure 2. The Central Montana Health District (CMHD) consists of the Fergus, Golden Valley, Judith Basin, Musselshell, Petroleum, and Wheatland counties. The counties are colored differently depending on the count of the newly reported influenza cases. A No Report (NR) signifies that there was nothing reported for the week according to DPHHS. An NR is different from zero reported cases. Zero reported cases are represented as white on the Montana counties map and no reported cases are represented as gray. In Figure 2, Lincoln and Mineral counties have 0 newly reported influenza cases, while Bighorn and Garfield have nothing reported for the week.



Tribal cases are included in county totals.
CMHD = Central Montana Health District. CMHD includes Fergus, Golden Valley, Judith Basin, Musselshell, Petroleum, and Wheatland counties.

Figure 2. A map of Montana counties with newly reported influenza cases as of March 2, 2019 (DPHHS, 2019). The counties are colored depending on the count of the newly reported influenza cases. The Central Montana Health District (CMHD) consists of Fergus, Golden Valley, Judith Basin, Musselshell, Petroleum and Wheatland counties. Each week during the flu season, DPHHS uploads this figure in a Montana Influenza Summary report. The newly reported influenza cases are recorded in a spreadsheet and updated weekly.

3.3. **Forecasting Algorithms.** We used several different methods to predict flu counts for the next week using exclusively past flu counts as predictor variables. Some of our methods proved to be successful while others proved fruitless after early investigation and were tabled for the duration of the project. All of the following algorithms were designed to be compatible with both counts and rates of flu by a county. For consistencies sake, we only refer to counts in our language below.

3.4. **Linear Regression.** The linear regression model uses the data in sets of n pairs, where

$$D = \{(y_1, \boldsymbol{x_1}), (y_2, \boldsymbol{x_2}), ...(y_n, \boldsymbol{x_n})\}$$

.

Here, y_i represents the target week, and $\boldsymbol{x_i}$ is a predictor vector of the flu counts of previous p weeks. This model expresses the expected value of Y_i as a linear function of $\boldsymbol{x_i}$ given by

$$E(Y_i|\boldsymbol{x_i}) = \beta_0 + \beta_1 x_i + ... + \beta_p x_p.$$

The β's are trained using the first s weeks of the flu season. The following formula was used:

$$\hat{\beta} = \sum_1^s \boldsymbol{x_i}\boldsymbol{x_i}^T \sum_1^s \boldsymbol{x_i} y_i.$$

3.4.1. ***Conventional Linear Regression***. Since our data had so many zero counts at the beginning of the flu season a sliding window was used to update $\hat{\beta}$, instead of using all of the previous weeks of the flu season. When predicting week $t+1$ the formula for $\hat{\beta}$ is

$$\hat{\beta} = \sum_{t-s}^t \boldsymbol{x_i}\boldsymbol{x_i}^T \sum_{t-s}^t \boldsymbol{x_i} y_i.$$

Thus,

$$\hat{y_i} = \boldsymbol{\hat{\beta}}\boldsymbol{x_i}$$

The option of adding in the flu counts of the surrounding counties was added. A weighted average of flu counts of surrounding counties was calculated based on population. This value was appended onto the predictor vector $\boldsymbol{x_i}$ and $\hat{y_i}$ was calculated as above. Using trial and error we found $p = 2$ and $s = 7$ to yield the most accurate results.

3.4.2. ***Weighted Linear Regression***. In order to improve forecasting results using linear regression an exponential weight, w_i, was added to the model. Where $w_i = (1-\alpha)^i(\alpha)$, giving more weight to the most current week, and less on subsequently previous weeks. Thus $\hat{\beta}$ is calculated

$$\hat{\beta} = \sum_{1}^{t} \boldsymbol{x_i} w_i \boldsymbol{x_i}^T \sum_{1}^{t} \boldsymbol{x_i} w_i y_i.$$

The predicting vector $\boldsymbol{x_i}$ is a vector of the previous p weeks' flu counts, with the option of again adding in the weighted average of the surrounding counties' flu counts. Using trial and error we found using $\alpha = 0.4$ returned the most accurate results.

3.5. **K Nearest Neighbors.** Since a patient usually takes about 1-2 weeks to recover from the flu, we decided to try the k nearest neighbors algorithm assuming the most recent weeks will have a higher effect on the target than the previous weeks. We introduce a weight for each count of k predictor variables (weeks). For example, if we are going to predict week $t+1$ using k weeks, then week t has weight:

$$w_t = \alpha(1-\alpha)^0$$

and the weight of week $t-k$ will be:

$$w_{t-k} = \alpha(1-\alpha)^{k-1}$$

Here,

$$\sum_{i=1}^{k} w = \sum_{i=1}^{k} \alpha(1-\alpha)^{i-1} \approx 1$$

3.5.1. ***K Nearest Neighbors Conventional Functions.*** Instead of using all t prior weeks as predictor variables, we choose k weeks where k ranges from 3 to 6. Thus, the predicted flu counts for week t is

$$\hat{y}_{t+1} = w_t y_t + w_{t-1} y_{t-1} + ... + w_{t-k} y_{t-k}$$

where, y_i denotes the observed flu count at week i, and the choice of α can be varied depends on the choice of k.

3.5.2. ***K Nearest Exponential Function.*** Here we use all t weeks to predict week $t+1$, and thus our predicted flu count for week $t+1$ is

$$\hat{y}_{t+1} = w_t y_t + w_{t-1} y_{t-1} + ... + w_2 y_2 + w_1 y_1$$

We had the most accurate result using $\alpha \approx 0.45$.

3.6. **Holt-Winters.** The Holt-Winters exponential forecast $\hat{\beta}_{t+\tau}$ extends exponential forecasting by introducing a rate of change parameter β_t. Beta is a time-varying slope estimate and is the estimated rate of change in the (exponential forecast) mean level between time step $t-1$ and t. The estimated mean level is updated as

$$\hat{\mu}_t = (1-\alpha_h)(\hat{\mu}_{t-1} + \hat{\beta}_t) + \alpha_h x_t$$

We had the most accurate results using a weighted average of the previous two weeks as our predictor variable x_t.

$\hat{\beta}_t$ is computed after $\hat{\mu}_t - \hat{\mu}_{t-1}$, using

$$\hat{\beta}_t = (1-\alpha_b)(\hat{\beta}_{t-1} + \alpha_b(\hat{\mu}_t - \hat{\mu}_{t-1}).$$

Note that the two equations each have their own α (which may or may not be the same value), where $0 < \alpha_h < 1$ is a smoothing constant. We tested in 0.05 increments from 0.25 to 0.75 for α_h and α_b and had the most accurate results using $\alpha_h = 0.4$ and $\alpha_b = 0.3$.

$$\hat{y}_{t+\tau} = \hat{\mu}_t + \hat{\beta}_t$$

Note that τ is 1 since we are only trying to predict 1 time step, or week in this case.

3.7. **ARIMA Model.** An ARIMA(p,d,q) model combines autoregressive (p), moving average (q) and differencing components (d) to make data stationary. This method is used heavily in economics, the model allows for flexible cyclical patterns. The formula for ARIMA is given below in backshift notation.

$$(1-\phi_1 B)^p (1-B)^d * Y_t = (1-\theta B)^q$$

To someone unfamiliar with backshift notation this formula may be confusing, but essentially backshift notation allows the number of lags in either the moving average or autoregressive terms. The number of terms lagged for the autoregressive components are given by p, moving average by q, and are weighted by constants ϕ and θ, which change with every lag.

Initially, ARIMA modeling seemed like a suitable fit for our project as our flu data moved in a seasonal cycle year after year. However, several difficulties arose when estimating our models using ARIMA. We had to build a separate coding infrastructure in R as opposed to typical Python infrastructure. Our data was inconsistent on a yearly basis. We made the assumption that flu seasons operate independently from one year to the next. As a result, we determined the goal of the project to predict next weeks flu count by county using only the current years flu data. There simply is not enough data yearly to build an effective ARIMA model by evaluating independent flu seasons.

3.8. **Accuracy Assessment.** Accuracy is a measure of the degree of closeness of a measure or calculated value to its actual value. It is rare to make exact predictions in a study. The distance between an observation and an actual value is described by accuracy. Experimental error is defined as the difference between an experimental value and the actual value of a quantity. The difference between these numbers indicates the accuracy of the measurement.

The accuracy assessment goals for our project are below:

- Assess how well a classification worked
- Understand how to interpret the usefulness of someone else's classification

For our project, the following three accuracy assessment methods were used: coefficient of determination (R^2), root-mean square error (RMSE) and mean absolute error (MAE).

3.8.1. *Coefficient of Determination (R^2).* A measure of forecasting accuracy is the coefficient of determination. It compares the sum of squared errors produced by the prediction/forecasting function to the sum of squared errors produced by a baseline prediction function. Often, the baseline prediction function is the sample mean (hence, the sample mean of the target sequence is set to be the prediction function for every instance). The coefficient of determination can be computed as follows:

$$R^2 = \frac{\sum_i (y_i - \bar{y}_i)^2 - \sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \tag{3.1}$$

3.8.2. *Root Mean Square Error (RMSE).* Root-mean-squared error (also known as root-mean-square error) is a frequently used measure of the difference between values (sample or population values) predicted by a model or an estimator and the values observed. The root-mean-square error can be computed as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.2}$$

3.8.3. *Mean Absolute Error (MAE).* The mean absolute error (MAE) is a common measure of forecast error in time series analysis. MAE is a measure of difference between two continuous variables. MAE is also the average horizontal distance between each point and the identity line. The equation can be computed as follows:

$$MAE = \sum_{i=1}^{n} |y_i - \hat{y_i}| \qquad (3.3)$$

For our project Influenza Trends for Counties in the State of Montana, we used accuracy scores to evaluate the best model in comparing the results from the different techniques for each year and each county.

4. Results

4.1. **Forecasting Algorithms.**

4.1.1. *Exponential K Nearest Neighbors.* The exponential K Nearest Neighbors algorithms averaged R^2 value across all counties in Montana was 0.70. The average R^2 value across all large counties in Montana was 0.27. The mean absolute error was 20.45 and the root mean square error was 42.86. See Table 1. In Figure 3, the actual flu counts for the 2018-2019 flu season for Missoula county are compared to the Exponential K Nearest Neighbor forecasts for the flu. The blue line is the KNN forecasting function and the red line is the actual flu counts for Missoula county. The x-axis represents the weeks of the flu season while the y-axis represents the amount of newly reported flu counts for the given week.

4.1.2. *Conventional K Nearest Neighbors.* There were two values for k used for the KNN functions. When k=3, the algorithm used the past three weeks to predict the upcoming week. The average R^2 across all counties was 0.09 and the average R^2 across the large counties 0.57. The mean absolute error was 16.32. See Table 1. The root mean square error was 28.08 when k=3. Then, a conventional KNN with a k value of 7 was used to predict flu. The R^2 across all counties was 0.15 and the average R^2 across the large counties 0.51. The mean absolute error was 17.35. The root mean square error was 32.09. See Table 1. In Figure 5, the actual flu counts for the 2015-2016 flu season for Missoula county are compared to the Conventional K Nearest Neighbor forecasts for the flu. The blue line is the Conventional KNN forecasting function and the red line is the actual flu counts for Missoula county. The x-axis represents the weeks of the flu season while the y-axis represents the amount of newly reported flu counts for the given week. In Figure 4, the actual flu counts for the 2018-2019 flu season for Missoula county are compared to the Conventional K Nearest Neighbor forecasts for the flu. The blue line is the Conventional KNN forecasting function and

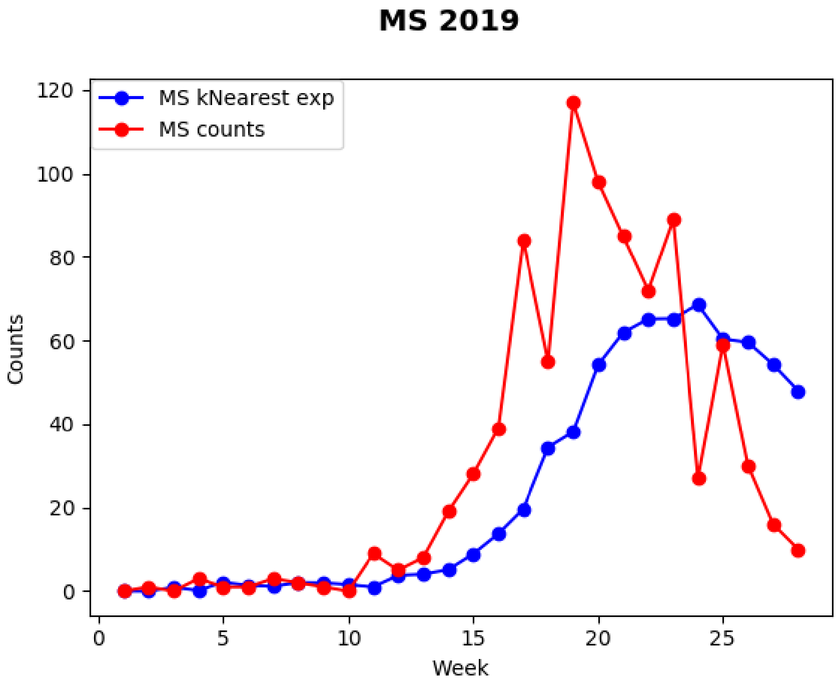


Figure 3. K-Nearest Neighbors forecast for Missoula County for the 2018-2019 flu season. The blue line is the KNN forecasting function. The red line is the actual flu counts for Missoula county. The x-axis represents the weeks of the flu season and the y-axis represents the amount of newly reported flu counts.

the red line is the actual flu counts for Missoula county. The x-axis represents the weeks of the flu season while the y-axis represents the amount of newly reported flu counts for the given week. Figures 4 and 5 were produced with k=3.

4.1.3. *Holt-Winters.* There were two Holt-Winters forecasting algorithms the class used to predict future flu counts. One algorithm had parameters of (0.35, 0.05). The average R^2 value across all counties in Montana was 0.05. The average R^2 value across all large counties in Montana was 0.53. The mean absolute error was 17.03 and the root mean square error was 29.64. See Table 1. Furthermore, the second Holt-Winters algorithm used the parameters (0.70, 0.05). The average R^2 value across all counties in Montana was 0.02. The average R^2 value across all large counties in Montana was 0.55. The mean absolute error was 16.78 and

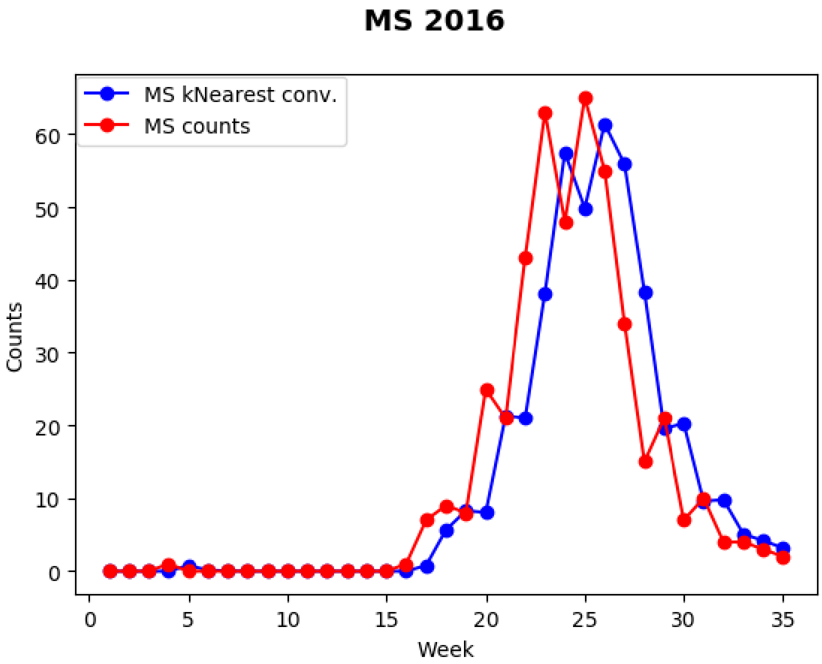


FIGURE 4. Conventional K-Nearest Neighbors forecast for Missoula county for the 2015-2016 flu season. The blue line is the Conventional K-Nearest Neighbors forecasting function. The red line is the actual flu counts for Missoula county. The x-axis represents the weeks of the flu season and the y-axis represents the amount of newly reported flu counts.

the root mean square error was 28.62. See Table 1. In Figure 6, the actual flu counts for the 2018-2019 flu season for Missoula county are compared to the Holt Winters forecasts for the flu. The blue line is the KNN forecasting function and the red line is the actual flu counts for Missoula county. The x-axis represents the weeks of the flu season while the y-axis represents the amount of newly reported flu counts for the given week.

4.2. **Accuracy Assessment.**

4.2.1. *Coefficient of Determination.* The decision, regarding whether or not a given model is acceptable or desirable, must be made on a quantitative basis. A value is generated by one or more of the classification methods (R^2, RMSE, and MAE) as described in the accuracy

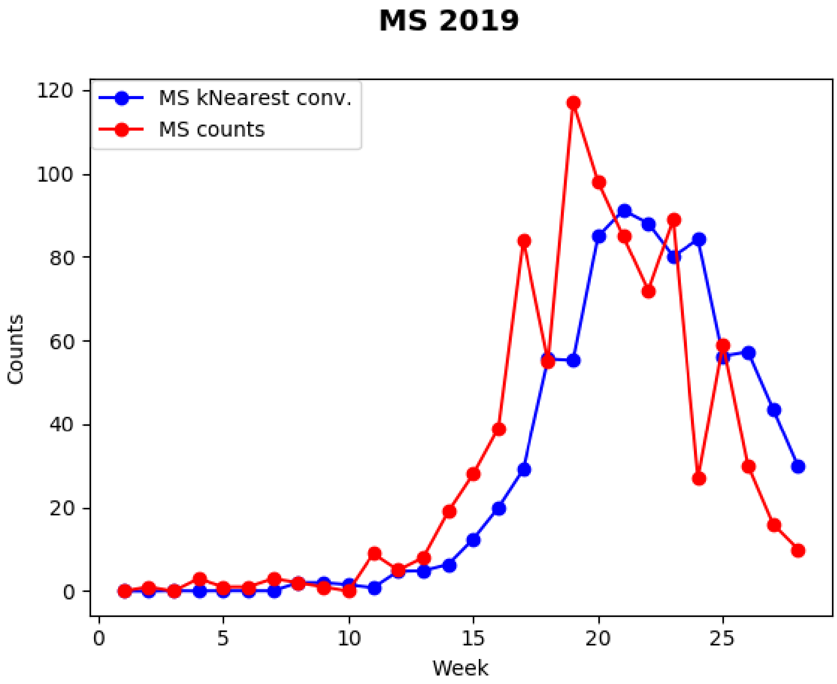


FIGURE 5. Conventional K-Nearest Neighbors forecast for Missoula County for the 2018-2019 flu season. The blue line is the Conventional K-Nearest Neighbors forecasting function. The red line is the actual flu counts for Missoula county. The x-axis represents the weeks of the flu season and the y-axis represents the amount of newly reported flu counts.

assessment section of the report. The classification method used depends on the problem. The minimum value of the metric that is required is established and this value is called the threshold of accuracy. Using the threshold of accuracy, a forecasting model can be classified as good or bad. Our team decided that we would use the R^2 value of 0.80 as our threshold of accuracy.

4.3. **Mapping.** Influenza rates by county were visualized in R using the following Comprehensive R Archive Network (CRAN) packages:

- Simple features (sf) is a package used to read spatial vector data for visualizations. Using this package, spatial data could be read in for mapping.
- The dplyr package is used for data manipulation using functions similar to those in Structured Query Language (SQL). For this

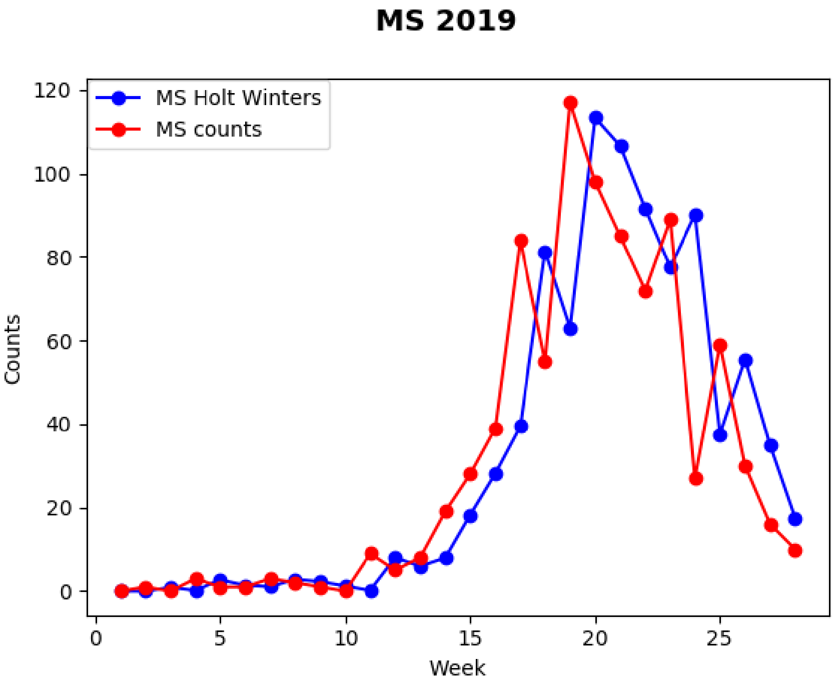


FIGURE 6. Holt Winters forecast for Missoula County for the 2018-2019 flu season. The blue line is the Holt Winters forecasting function. The red line is the actual flu counts for Missoula county. The x-axis represents the weeks of the flu season and the y-axis represents the amount of newly reported flu counts.

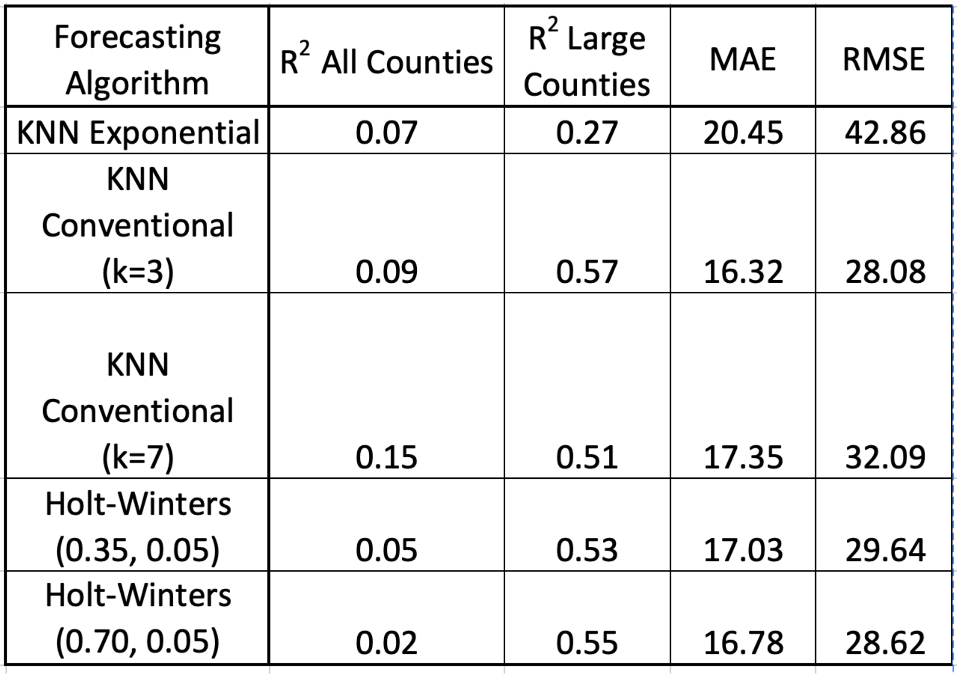
project, the inner join function was used to link influenza rates to their county names. This data could then be appended to a table containing the spatial data for mapping.

- The graphics package ggplot2 is used to map variables to aesthetics and create data visualizations. For this project, the package was used to create a map of Montana counties.

To map the influenza rates by county, shape files, which are an extract of selected geographic and cartographic information from the U.S. Census Bureau's master address file, were used. These files are created to be seamless, with no overlaps or gaps between parts. For this project, spatial data was collected for Montana by county.

After reading in the spatial data, Judith Basin (JB), Petroleum (PE), Golden Valley (GV), Fergus (FE), Musselshell (MU), and Wheatland (WH) counties were merged and renamed as the Central Montana Health District (CHMD). Since each of these counties has populations

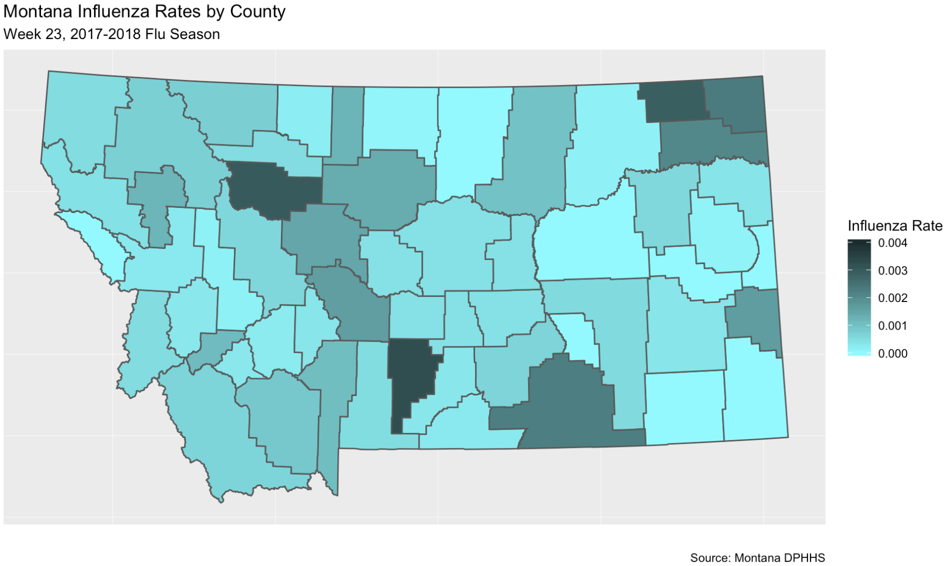
Table 1. The table includes the coefficient of determination (R^2) for all Montana counties, coefficient of determination (R^2) for large Montana counties, mean absolute error (MAE), root mean square error (RMSE), and the difference between the mean absolute error and root mean square error (RMSE-MAE) for each forecasting algorithm.



Forecasting Algorithm	R^2 All Counties	R^2 Large Counties	MAE	RMSE
KNN Exponential	0.07	0.27	20.45	42.86
KNN Conventional (k=3)	0.09	0.57	16.32	28.08
KNN Conventional (k=7)	0.15	0.51	17.35	32.09
Holt-Winters (0.35, 0.05)	0.05	0.53	17.03	29.64
Holt-Winters (0.70, 0.05)	0.02	0.55	16.78	28.62

lower than 1,000, the state of Montana collects influenza data for them collectively as the Central Montana Health District. The influenza rates data was read in, transformed to a data frame, and transposed. The most recent complete flu season, 2017-2018, was explored to display influenza rates from various times in the flu season. Week 12 was selected to represent the beginning of the flu season, week 23 was selected as the approximate peak of the flu season, and week 25 was selected to show a slight decline after the peak. Flu rates were extracted from each of these weeks and combined Montana county abbreviations in their respective data frames. Next, this data was linked to the spatial mapping data using an inner join.

To create static plots, ggplot was used to map each county and fill it based on its influenza rate. The Central Montana Health District was



Montana Influenza Rates by County

Week 23, 2017-2018 Flu Season

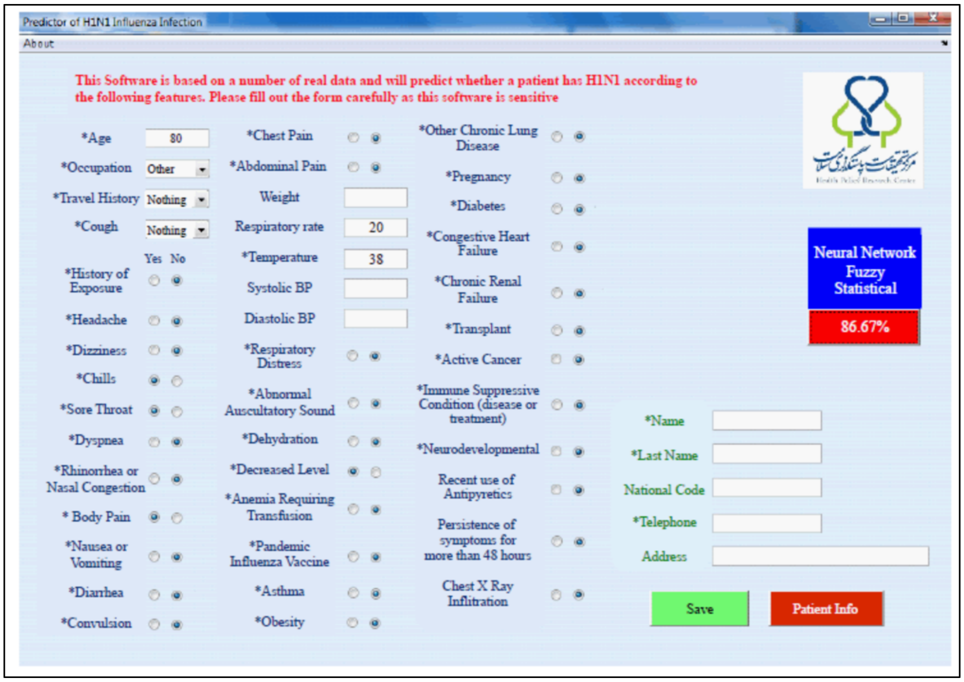
Source: Montana DPHHS

Figure 7. Montana influenza rates by county for the 2017-2018 flu season. The map visualizes the influenza rates for Week 23. Influenza rates are colored in intervals of 0.001, with 0 as the lowest value and the lightest shade of blue and 0.004 as the highest value and darkest shade of blue.

mapped as one cohesive county. The maximum influenza rate in the 12^{th}, 23^{rd}, or 25^{th} week in the 2017-2018 season was approximately .004. Using a gradient color palette of blues, influenza rates were colored in intervals of .001, with 0 as the lowest value and lightest shade of blue and .004 as the highest value and darkest shade of blue. See Figure 7. A map was produced for the 12^{th}, 23^{rd}, and 25^{th} week of the 2017-2018 influenza season using these parameters.

5. Discussion

The statistical models utilized by our class provided varying degrees of statistical value as evidenced in Table 1. Our results do not provide a specific forecasting recommendation for the Montana DPHHS at the present time, not do they indicate a one size fits all approach. A critical assessment of the forecasting power of the variables addressed identifies opportunities to improve forecasts if additional relevant data can be obtained. Examples as identified include vaccination records (inoculation counts and efficacy assessment), volumes and patterns of



This Software is based on a number of real data and will predict whether a patient has H1N1 according to the following features. Please fill out the form carefully as this software is sensitive

*Age 80
*Occupation Other
*Travel History Nothing
*Cough Nothing

	Yes	No
*History of Exposure	○	●
*Headache	○	●
*Dizziness	○	●
*Chills	●	○
*Sore Throat	●	○
*Dyspnea	○	●
*Rhinorrhea or Nasal Congestion	○	●
* Body Pain	●	○
*Nausea or Vomiting	○	●
*Diarrhea	○	●
*Convulsion	○	●

*Chest Pain ○ ●
*Abdominal Pain ○ ●
Weight
Respiratory rate 20
*Temperature 38
Systolic BP
Diastolic BP
*Respiratory Distress ○ ●
*Abnormal Auscultatory Sound ○ ●
*Dehydration ○ ●
*Decreased Level ● ○
*Anemia Requiring Transfusion ○ ●
*Pandemic Influenza Vaccine ○ ●
*Asthma ○ ●
*Obesity ○ ●

*Other Chronic Lung Disease ○ ●
*Pregnancy ○ ●
*Diabetes ○ ●
*Congestive Heart Failure ○ ●
*Chronic Renal Failure ○ ●
*Transplant ○ ●
*Active Cancer ○ ●
*Immune Suppressive Condition (disease or treatment) ○ ●
*Neurodevelopmental ○ ●
Recent use of Antipyretics ○ ●
Persistence of symptoms for more than 48 hours ○ ●
Chest X Ray Inflitration ○ ●

86.67%

*Name
*Last Name
National Code
*Telephone
Address

FIGURE 8. An example of a graphical user interface (GUI) for predicting influenza. The user inputs their qualifications including age, symptoms and personal information and the program will output if the user is predicted to have influenza. In the future, a GUI will be created to forecast influenza in Montana.

air-borne particulates, and human transportation patterns. We believe that working with stakeholders to provide such data will be beneficial to the eventual purpose of providing actionable forecasts which allow for optimal resource management.

5.1. **Future Work.** Although the Data Science Projects course created successful algorithms to predict flu in Montana, there are a few items that were not completed due to time constraints and lack of data. In the future, it would be helpful to create a graphical user interface (GUI) for the user. A GUI would allow a user with minimal coding knowledge to create forecasts of the flu. In Figure 8, a GUI for predictor of H1N1 Influenza Infection is displayed (Boostani, 2012). The user can input their qualifications including their age, symptoms, and personal information and the program will output if the user is predicted to have H1N1. A helpful GUI for our project would allow the user to choose the Montana county, forecasting algorithm, amount of past data, and the amount of future weeks the user would like to predict. The output of the GUI would be plots comparing the counts to the forecasting algorithm similar to Figure 5 and the corresponding accuracy assessment numbers including R^2, MAE, and RMSE.

With more time and funding, we would improve the forecasting algorithms. Improved forecasting algorithms would be defined by the improvement in the accuracy assessment for the algorithms. The forecasting algorithms currently forecast one week in advance. It would be ideal to have forecasting algorithms that could accurately forecast one to four weeks in advance for a flu season. In addition, the forecasting algorithms would produce accuracy assessments and plots so the user could see how successful the forecast had been, for the flu. A GUI for this project would give the user the ability to choose how many weeks in advance they would predict the flu and receive the corresponding information from the algorithms.

Another task to improve the forecasting process is automating the data staging process. Currently, each week the newly reported flu counts are manually extracted from the pdf map seen in Figure 2 into an Excel Spreadsheet. Although it only takes on average 10-20 minutes, it would be helpful to automate this process to save time and money. Creating a function to remove the newly reported flu counts each week, thus decreasing the time spent on staging, would allow for more time to be added to algorithm improvement and disease prevention. Once all these tasks are completed for influenza forecasting, the algorithms and methods could be applied to other diseases. The forecasting algorithms could help prevent or reduce the spread of major diseases across Montana and the U.S.

References

[1] Boostani, R., Rismanchi, M., Khosravani, A., Rashidi, L., Kouchaki, S., et al. (2012). Presenting a Hybrid Method in Order to Predict the 2009 Pandemic Influenza A (H1N1). *Journal of Health & Medical Informatics*, 3(3), 1-6.

[2] Centers for Disease Control and Prevention. (2019). *Weekly U.S. Influenza Surveillance Report*. Retrieved from https://www.cdc.gov/flu/weekly.

[3] Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 30 April 2019.

[4] Montana Department of Public Health and Human Services. (2019). *Seasonal Influenza (Flu)*. Retrieved from https://dphhs.mt.gov/publichealth/cdepi/diseases/influenza.

[5] Ortiz, J. R., Zhou, H., Shay, D. K., Neuzil, K. M., Fowlkes, A. L., et al. (2011). Monitoring Influenza Activity in the United States: A Comparison of Traditional Surveillance Systems with Google Flu Trends. *PLoS ONE, 6*(4), 1-10.

[6] Steele, B., Chandler, J., & Reddy, S. (2016). *Algorithms for Data Science*, Cham, Switzerland: Springer International Publishing.

[7] United States Environmental Protection Agency. (2019). *Particulate Matter (PM) Basics*. Retrieved from https://www.epa.gov/pm-pollution/particulate-matter-pm-basics.

[8] Zachary, Kimon C. (2019). *Treatment of influenza in adults*, Wolters Kluwer (1-33).