# Atlanta Pollen Counts Prediction

BMED6700 Biostatistics, Final Project Report

Tomáš Brůna*, Dongjo Ban†, Saurabh Gulati‡

May 1, 2018

## Abstract

Allergy is an incurable condition caused by the hypersensitivity of the immune system to something in the environment. Pollen, which is produced by flowering plants is one of the most common known allergen which causes symptoms as severe as impairment of cognitive performance. Pollen counts are expected to be correlated with weather, we thus attempted to predict future pollen levels based on historical pollen counts and weather data using time series models and machine learning algorithms. We are able to predict exact future pollen counts with mean absolute error of ∼3.4 and future pollen levels with ∼90% accuracy.

## 1 Introduction

Allergies are among the most common chronic conditions affecting the global population. The symptoms range from minor inconveniences to life threatening reactions. Among the most common allergens that is known to trigger allergic responses is pollen [1]. Formed from flowering plants, airborne pollen may contain allergens that cause allergic inflammatory response which then could lead to a systemic reaction. Effect of the allergen includes not only nasal and ocular symptoms, but can also impair cognitive performance as well as information processing [2, 3].

Furthermore, as there is no cure for such allergies, it is of importance that people who are allergic make necessary adjustments to avoid contact with pollen. This makes predicting pollen levels an important activity to aid allergic individuals in taking the necessary steps required to stay well. Our project is focusing on such prediction based on historical pollen counts and weather data.

## 2 Data Sources

### 2.1 Pollen Counts Data

Pollen data was sourced form the Atlanta Allergy portal - `www.atlantaallergy.com`. This website contains Atlanta's polled count for almost every day from June 1991 up to today. For any particular day they have the total pollen count and also the individual pollen categories for tree, weed and grass pollen. The categories are either *Extreme (E)*, *High (H)*, *Medium (M)* or *Low (L)*. For some days they also have the top contributor for the different pollen categories. The data resides on the website in the form of a calender showing the above mentioned information, hence we decided to write a web-scraping script to extract the information.

*Tomáš Brůna, 54675, bruna.tomas@gatech.edu
†Dongjo Ban, 42323, dban8@gatech.edu
‡Saurabh Gulati, 24499, saurabhg59@gatech.edu

#### 2.1.1 Scraping

Data for each day reside at a unique URL. For example, pollen counts for May 14, 2002 are located at `http://www.atlantaallergy.com/pollen_counts/index/2002/05/14`. We designed a script which iterates through each day and corresponding URL since June 18, 1991 (first entry) up to today. Next step was to extract pollen counts and pollen levels for different categories from these pages. Several special cases such as days with missing or invalid values needed to be handled. The scripts saves all the results into a single tab separated file with the following columns: *Year, Month, Day, Total Count, Tree Level, Tree Top Contributors, Grass Level, Grass Top Contributors, Weed Level, Weed Top Contributors*. We attached the full script in the `scraper.py` file. To test, just run `./scraper.py` and the result will be saved in `out.tsv` file.

### 2.2 Weather Data

The weather data was sourced from Weather Underground (`https://www.wunderground.com/`). Weather underground is a weather service that provides accurate real-time weather information; the information in Weather Underground comes from the National Weather Service and from 250,000 Personal Weather Stations, making it the most reliable source of weather data available. The National Weather Service, which is part of the National Oceanic and Atmospheric Administration (NOAA), and is considered the standard for most climate and weather data related research. We collected location specific information of weather in Atlanta and matched the time range to the available pollen data.

#### 2.2.1 Scraping

We extracted historical weather information, ranging from 1991 until present day, using Beautiful Soap, a python package for parsing HTML and XML documents. In order to stay consistent with the pollen counts dataset, we set the location to Atlanta, Georgia. Using the scraper, we parsed through the monthly weather data that consisted of temperature, wind speed, precipitation, humidity, and weather events.

## 3 Dataset Description

### 3.1 Merging Datasets

Once the weather and pollen data were scraped from the individual sources, we merged the datasets based on the date. By merging we got a dataset which had both the pollen and weather information for most days starting June 1991 till today.

The merged dataset has all of the described pollen data and weather information: *Temperature, Humidity, Precipitation, Wind Speed, Dew Point, Sea Level and the Weather Events* that occurred on that day. Since the websites were missing the data for some days, the merged dataset had some of the data missing. For more details, see the included `mergedData.csv` file.

### 3.2 Handling Missing Values

To handle missing values in the merged dataset, we used R package *imputeTS* [4] which is a collection of algorithms and tools for univariate time series imputation. Linear interpolation was used to impolate the missing values for Weather columns and total pollen counts. Once the data was interpolated, it was ready for the time series analysis.

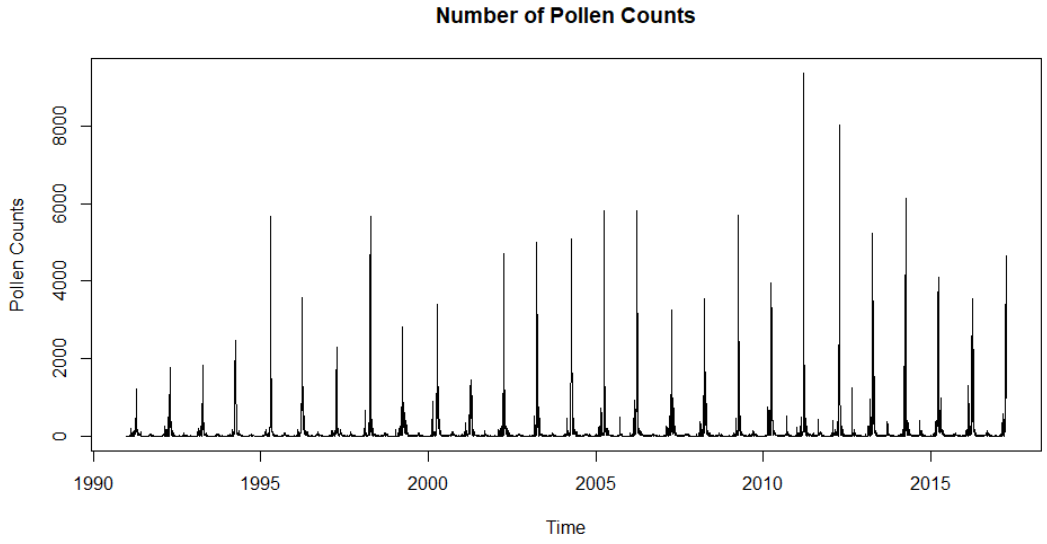


Figure 1: Number of pollen counts from 1991 to 2018.

# 4 Time Series Analysis

Time series consists of sequential numerical data in successive order. Its application range from social setting, such as economics and unemployment, to areas of health where an effect of a certain drug type is measured over time. One of the most rigorous and advanced practice involves environmental sciences.

The pollen data we extracted consists of discrete time data measuring daily pollen counts. In order to derive meaningful statistics and to analyze the dataset, we chose to perform time series analysis. With this method, we will be able to achieve our primary goal of being able to predict the number of pollen counts for a given day of a year by utilizing the seasonal and past data [5].

## 4.1 Exploratory Analysis

Visualizing the time series plots before performing any statistical tests can give an insight as to how to move forward with the analysis. It is considered to be one of the important topics in environmental analysis as it can assist in identifying and understanding association between the data. The visualization is displayed in the Figure 1.

Initial screening of the pollen counts throughout the years does not offer a definite answer as to whether our dataset contains a trend or not. It does, however, show that the counts reached its peak around 2011. A quick search was done to inspect a possible cause for sudden rise of pollen level, but there were no significant findings. In order to examine our time series more in depth, we decomposed our time series into 3 different components: seasonal, trend, and residuals (Figure 2).

After the decomposition, we can observe that seasonality is consistent throughout the years with peaks happening around the same time period. We can also notice consistency in residuals, with largest peak occurring around the same time where our time series was at its highest. The trend seems to be the only component from decomposition that is not consistent and showing a deterministic trend as year progresses.

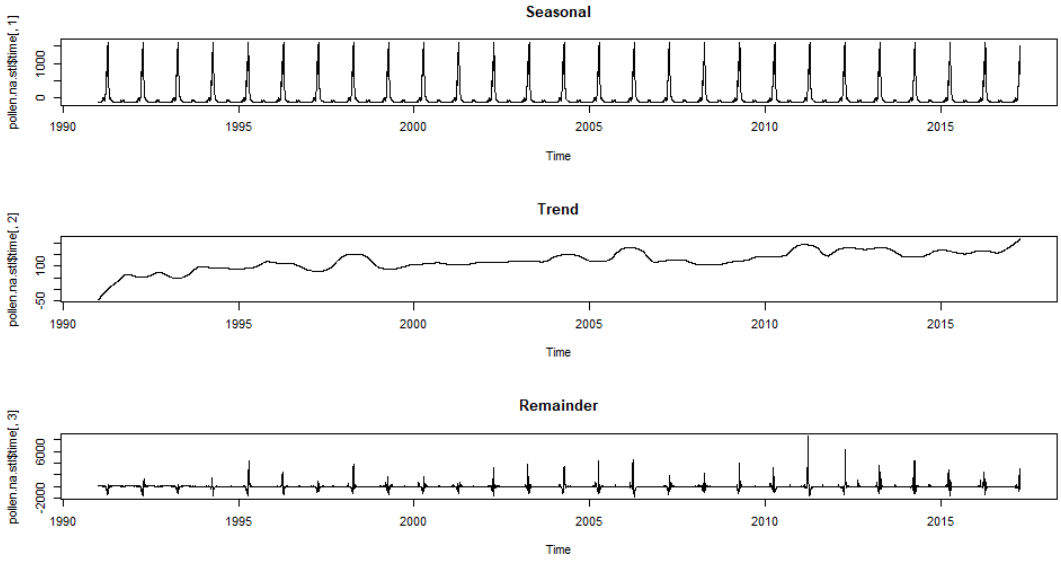


Figure 2: Decomposition of pollen count time series data.

## 4.2 Tests for Stationarity

Stationarity is essential in research where the variables are dependent on time. Thus, it is an important assumption that precedes making any type of predictions or forecasts. In order to determine whether our pollen count data is stationary or not, we employed two statistical tests: augmented Dickey-Fuller and KPSS [6].

Unit root tests are generally used to determine if the trending data needs to be differenced or regressed on some function of time to make it stationary [7]. Developed by Said and Dickey in 1984, augmented Dickey-Fuller expands the autoregressive unit root test to handle ARMA models, allowing it to handle more complicated autoregressive model.

$$H_0 : \textit{time series is nonstationary}$$

$$H_1 : \textit{time series is stationary}$$

We carried out the statistical test using 'tseries' R package and observed a p-value smaller than 0.01. Therefore, we reject the null in this case and can say that our time series is stationary and does not require detrending. It is a common practice to perform more than one test to determine stationarity of the time series of interest. More specifically, Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test is jointly used with augmented Dickey-Fuller test (ADF).

$$H_0 : \textit{time series is stationary}$$

$$H_1 : \textit{time series is nonstationary}$$

We can notice that the hypotheses are different from ADF. Using the same package as ADF, we performed another statistical test and observed a p-value greater 0.1. Thus, we cannot reject the null hypothesis and can say that our time series is stationary.

## 4.3 Model Fitting

The aim of fitting a model for a time series is to be able to incorporate past observations from the series to create a model that best describes the characteristics of the series [8]. One of the most popular and widely used time series model is autoregressive moving average (ARMA)

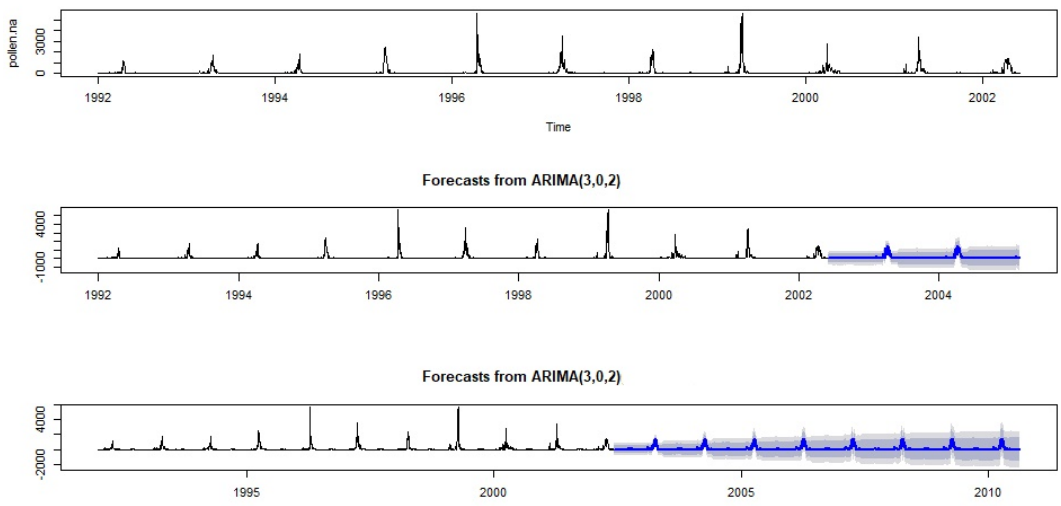


Figure 3: Top: original data; Middle, Bottom: Forecasts based on the predicted ARMA model.

model. Using auto.arima() function in 'forecast' package in R, we are able to predict best parameters for ARMA model using AIC or BIC scores. The resulting model for our pollen counts data was ARMA(3,2) [9].

Using the fitted model, we performed two predictions: 2 years and 8 years into the future. From the predictions, we notice that they depend on the shape and height of the peak near the end. Furthermore, we see the confidence interval for both predictions increasing. See Figure 3 for details. This increase is reasonable considering the level of error and uncertainty would increase as time progresses.

### 4.4 Time Series Conclusion

From the exploratory analysis, we noticed a seasonal pattern where pollen counts experience a dramatic increase. Moreover, we became aware of the trend in our dataset. However, we learned that our pollen counts dataset is stationary and does not require any type of detrending by performing statistical tests for stationarity. Furthermore, the predictions seem to be relatively close to the overlapping original data. They also seems to stay constant for the most part, mimicking the peak at the end of the plot. Lastly, we can also notice confidence interval slowly increasing.

Examining the ACF plot (Figure 4) shows us that pollen counts for the current day depends not only on the past couple of days or weeks, but also on the same day of previous years as well.

## 5 Machine Learning

Apart from using seasonal and past pollen data, we decided to add more explanatory variables to help us predict the future pollen counts. We chose weather data for this purpose because pollen counts are expected to be highly correlated to weather. For example, rain can wash the pollen away and water associated with increased humidity weighs the pollen grains down so they do not travel as far when the wind blows.

Weather data have the advantage of being closely and precisely monitored. Apart from that,

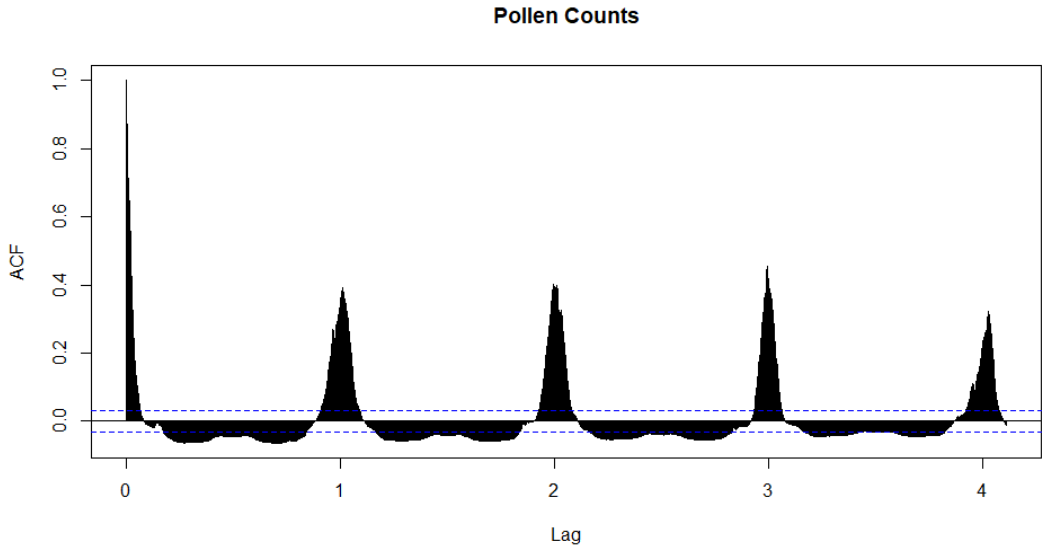


Figure 4: ACF plot with lag 4

reliable weather forecasts are usually available, therefore we can base our future predictions on the weather forecasts.

Due to the high number of explanatory weather variables which can interact with each other and affect pollen counts in a nonlinear fashion, we decided to use supervised machine learning algorithms. We tested neural networks and random forests which are designed to handle such interactions as well as other methods such as linear regression. In general, supervised machine learning maps input data (weather, season and past pollen counts) to output which are future pollen counts.

## 5.1 Weather Data Only

To see how well the weather explains pollen counts (without using the past pollen data), we based our first set of models on the day of the year (representing seasonality) and the weather data only. This approach assumes the availability of a reliable weather forecast on which the prediction for the future pollen counts will be based[1].

This method has several advantages: Since it does not use any past information, the rows are independent and any rows with missing values can be ignored during the model training. Also, predictions can be made even when past pollen counts are not available.

### Algorithms

We compared four different regressors:

- Linear regression [10]
- Neural network [11]
- Support Vector Machine [12]
- Random Forest [13]

[1]This is not an issue when prediction few days into the future (which is what we are trying to do). However, long term predictions using this approach are not possible due to the limitations of weather forecasts.

| Algorithm | Linear Regression | Neural Network | Support Vector Machine | Random Forest |
|---|---|---|---|---|
| $R^2$ | 0.002 | 0.34 | 0.334 | 0.346 |

Table 1: A comparison of $R^2$ scores for different regressors. Weather and seasonal data are used to predict pollen counts. The higher the score, the better the model is. All models were run with a set of best parameters found.

| Measure | Random Forest |
|---|---|
| $R^2$ | 0.346 |
| Mean Absolute Error | 99.817 |
| Median Absolute Error | 5.316 |

Table 2: Error measures for Random forest regressor. Weather and seasonal data are used to predict pollen counts.

**Parameter Estimation**

For a fair comparison, we tried to find a set of optimal parameters for the methods. This was achieved by running each of the algorithms multiple times with changing parameters and measuring the proportion of variance explained ($R^2$) [14]. To account for overfitting and estimate the true performance of each model, we measured $R^2$ using 10-fold cross-validation [15]. See Table 6 in the Appendix for a list of optimal parameters found for each algorithm.

**Results**

The final $R^2$ scores for each of the methods are displayed in the Table 1. The table shows that the performances of Neural network, Support vector machine and Random forest are comparable.

Neural networks are extremely sensitive to parameter setting. We observed that a small change in the number of neurons can bring the $R^2$ performance to almost 0. Support vector machines are more robust to parameter changes, however, the explanatory features need to be normalized prior to model training and the training itself takes a relatively long time. The Random forest is robust with respect to parameter settings and the training time is almost instant compared to SVM. Therefore, Random forest is the model we recommend for pollen count predictions and we will focus on random forests in the following analyses. Table 2 shows additional qualitative measures for the Random forest.

## 5.2 Inclusion of Historical Pollen Counts

The time series analysis showed us that pollen counts are highly correlated to pollen count levels from the past several days (see Figure 3). Therefore, it makes sense to include this information to achieve better predictions – we added values of pollen counts for the past 8 days among our explanatory variables. Because of that, each row depends on the 8 previous rows – dropping a row with missing values would also affect the subsequent 8 rows. Thus, we needed to impute

| Future prediction | 1 day | 2 days | 3 days | 4 days | 5 days |
|---|---|---|---|---|---|
| $R^2$ | 0.755 | 0.617 | 0.49 | 0.422 | 0.429 |
| Mean Absolute Error | 56.244 | 75.296 | 85.911 | 95.218 | 97.924 |
| Median Absolute Error | 3.395 | 5.157 | 6.391 | 7.119 | 7.998 |

Table 3: Performance of Random forest predictions (with weather, seasonal and recent pollen counts explanatory data) with respect to how many days in the future we are predicting.

| Future prediction | 1 day | 2 days | 3 days | 4 days | 5 days |
|---|---|---|---|---|---|
| $R^2$ | 0.732 | 0.597 | 0.484 | 0.419 | 0.387 |
| Mean Absolute Error | 59.352 | 79.090 | 89.529 | 98.030 | 101.293 |
| Median Absolute Error | 3.345 | 5.120 | 6.156 | 7.291 | 7.747 |

Table 4: Performance of Random forest predictions (with seasonal and recent pollen counts explanatory data) with respect to how many days in the future we are predicting. Compared to table 3, this model is not using weather data at all.

the missing values in this case. Again, we used Random forest regressor as our model and cross-validation to evaluate the performance.

**Results**

Inclusion of the recent pollen counts significantly improved the performance of our model. Table 3 summarizes the prediction errors. Analysis of decision trees in the fitted Random forest model shows (not surprisingly) that the most important splitting attribute is the value of pollen counts in the $day - 1$. The weather and seasonality are still used in the lower levels of the tree.
This type of prediction can only predict pollen counts one day into the future (because it needs data from $day - 1$). We also show how the prediction accuracy changes when we predict more than one day into the future. This is achieved by deleting column $day - 1$ when prediction 2 days into the future, deleting $day - 1$ and $day - 2$ when prediction 3 days etc. The behavior of accuracy with increasing future days is shown in the Table 3.

**Removing Weather**

We also measured the performance of this model without any weather data. The result is displayed in the Table 4. This result shows that recent and seasonal pollen data are more important explanatory variables than weather. The importance of weather increases the more days in the future we are trying to predict.

### 5.3 Machine Learning Conclusion

Overall, we recommend to use both weather and historical pollen data for model fitting and predictions. Even though the historical pollen counts are better explanatory variables, weather still improves the predictions and might be crucial when the observations for past pollen count levels are not available.

| Explanatory Variable | Weather & Season | Weather & Season & History |
|---|---|---|
| Accuracy for tree pollen levels | 0.820 | 0.872 |
| Accuracy for weed pollen levels | 0.864 | 0.924 |
| Accuracy for grass pollen levels | 0.897 | 0.925 |

Table 5: Prediction accuracies for pollen levels in each of the pollen categories. Random Forest classifier and cross-validation are used. In case of column where past data are used as explanatory variable, accuracy of 1 day future prediction is displayed.

### 5.4 Pollen Category Level Prediction

So far in our analysis, we focused on predicting the total pollen counts. Apart from overall counts, the data from `http://www.atlantaallergy.com/` also contain pollen levels for three categories: *tree*, *weed* and *grass*. Predicting these individual categories can be important since people can be sensitive only to some of these pollen types.

The values for these categories are not continuous, there are only 4 possible labels: *Extreme (E)*, *High (H)*, *Medium (M)* or *Low (L)*. Therefore, predicting these labels becomes a classification instead of a regression problem.

#### Method and Results

Using the knowledge and observations from the previously described experiments, we use the following method to predict pollen levels in distinct categories: First, we use only weather and seasonal data to train a Random Forest model. Because linear imputation of missing values is not a robust option for filling in missing categorical values [16], we use this model to predict pollen level for rows with missing values.

Next step is to include data from past 8 days (described in Subsection 5.2) and measure the prediction accuracy. The result is summarized in Table 5. The tables show that prediction of pollen levels can be done very accurately (∼90% accuracy) with Random forest classifier for each of the pollen categories. It is also worth noting that the prediction quality remains high even when pollen data from past 5 days are not available.

## 6 Conclusion

In this project, we successfully employed time series analysis and machine learning to predict future pollen numbers. Time series results show that we are able to predict the pollen counts to a certain extent, but the model is not able to capture accurately how our series is behaving. Concerning machine learning, we are able to predict pollen levels much more accurately than the exact pollen counts. This might be sufficient for real-world application because prediction of general pollen concentration for different pollen categories is more important from the medical standpoint than the precise overall pollen count number. This prediction will enable people who are allergic to the specific pollen type to take the necessary precautions.

### Lessons Learned

In the first part of the course, we learned to apply several statistical tests on real biological data. The course was especially useful in showing us how these tests differ from each other and in which context they should be used.

Next part of the class was also very helpful because we covered essential biological methods: how to account for multiple comparisons and time series.

Last but not least, this project taught us key concepts related to time series and we also gained hands-on experience with using supervised machine learning for both classification and regression.

## References


[1] Kiotseridis, H.; Cilio, C. M.; Bjermer, L.; et al. Grass pollen allergy in children and adolescents-symptoms, health related quality of life and the value of pollen prognosis. *Clinical and translational allergy*, volume 3, no. 1, 2013: p. 19.

[2] Church, M. K.; Scadding, G. K. Allergic rhinitis: impact, diagnosis, treatment and management. *Stroke*, volume 13, 2018: p. 57.

[3] Bensnes, S. S. You sneeze, you lose:: The impact of pollen exposure on cognitive performance during high-stakes high school exams. *Journal of health economics*, volume 49, 2016: pp. 1–13.

[4] Moritz, S.; Bartz-Beielstein, T. imputeTS: Time Series Missing Value Imputation in R. *R package version 0.4*, 2015.

[5] Shumway, R. H.; Stoffer, D. S. *Time Series Analysis and Its Applications With R Examples*, volume 3. Springer, 2011.

[6] Kebłowski, P.; Welfe, A. The ADF–KPSS test of the joint confirmation hypothesis of unit autoregressive root. *Economics Letters*, volume 85, no. 2, 2004: pp. 257–263.

[7] Diebold, F. X.; Kilian, L. Unit-root tests are useful for selecting forecasting models. *Journal of Business & Economic Statistics*, volume 18, no. 3, 2000: pp. 265–273.

[8] Adhikari, R.; Agrawal, R. K. An Introductory Study on Time Series Modeling and Forecasting. *CoRR*, volume abs/1302.6613, 2013, `1302.6613`. Available from: `http://arxiv.org/abs/1302.6613`

[9] Hyndman, R. J.; Khandakar, Y.; et al. *Automatic time series for forecasting: the forecast package for R*. 6/07, Monash University, Department of Econometrics and Business Statistics, 2007.

[10] Seber, G. A.; Lee, A. J. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[11] Haykin, S.; Network, N. A comprehensive foundation. *Neural networks*, volume 2, no. 2004, 2004: p. 41.

[12] Suykens, J. A.; Vandewalle, J. Least squares support vector machine classifiers. *Neural processing letters*, volume 9, no. 3, 1999: pp. 293–300.

[13] Liaw, A.; Wiener, M.; et al. Classification and regression by randomForest. *R news*, volume 2, no. 3, 2002: pp. 18–22.

[14] Carpenter, R. Principles and procedures of statistics, with special reference to the biological sciences. *The Eugenics Review*, volume 52, no. 3, 1960: p. 172.

[15] Kohavi, R.; et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, Montreal, Canada, 1995, pp. 1137–1145.

[16] Allison, P. D. Imputation of categorical variables with PROC MI. *SUGI 30 proceedings*, volume 113, no. 30, 2005: pp. 1–14.

# Appendix

## Optimal Machine Learning Parameters

| Algorithm | Important Parameters | | |
|---|---|---|---|
| Neural Network | $hidden\ layers = 2$ | $layer\ 1\ neurons = 10$ | $layer\ 2\ neurons = 10$ |
| Support Vector Machine | $C = 100$ | $Kernel = rbf$ | |
| Random Forest | $max\ tree\ depth = 4$ | $num\ trees = 25$ | $criterion = mse$ |

Table 6: Optimal parameters found for different machine learning algorithms. Only the important parameters affecting performance the most are displayed. Weather and seasonal data were used to predict pollen counts.