Journal of Industrial Pollution and Toxicity
Volume 1 | Issue 1

Research Article Open Access

Machine Learning Estimation of Pollen Abundance Using Landsat and Meteorological Data

Timms CI*[1], Liu X[1], Zewdie GK[1], Levetin E[2], and Lary DJ[1]

[1]Department of Physics, University of Texas at Dallas, USA
[2]Biological Science, University of Tulsa, USA

***Corresponding author:** Timms CI, Department of Physics, University of Texas at Dallas, Richardson TX 75080, USA, Tel: 2144150915, E-mail: cxt130330@utdallas.edu

Citation: Timms CI, Liu X, Zewdie GK, Levetin E, Lary DJ (2018) Machine Learning Estimation of Pollen Abundance Using Landsat and Meteorological Data. J Indu Pollut Toxicity 1(1): 101

Abstract

Allergenic diseases are on the rise in the United States and already affect 50 million Americans. Ambrosia pollen is one of the most allergenic forms of pollen, partially due to the shear volume of pollen that each Ambrosia plant generates. This paper describes a machine learning approach used to forecast the Ambrosia pollen levels in Tulsa, Oklahoma. Machine learning along with comprehensive environmental data (reflectance values derived from Landsat satellite images and parameters from the NASA MERRA meteorological analysis) are utilized to predict Ambrosia pollen levels. The outcome shows an improvement over previous studies and gives new information that could be advantageous for future pollen forecasting. In this study 21 different machine learning approaches for predicting the ragweed pollen levels were tested. These studies highlighted several Landsat variables were useful in forecasting the ragweed pollen levels, including the short wave infrared (SWIR1) and near-infrared (NIR) reflectance of various time lags. When Ragweed blooms and is releasing pollen the flowers are widespread and of a characteristic color. Landsat views the entire visible and near infrared spectrum, so widespread blooming should be manifest in changes in the surface reflectance spectrum seen by Landsat. Of all the machine learning approaches, the Gaussian Processes Regression (GPR) was the most accurate.

Keywords: Pollen; allergies; machine learning; pollen prediction

Introduction

Ambrosia pollen has a strong link with asthma, rhinitis, hay fever and other allergenic diseases [1]. Currently, 18.9 million adults and 7.1 million children suffer from asthma in the United States alone (CDC, 2013) [2]. The estimated total cost of asthma in the United States for 2007 in 2009 dollars was $56 billion. This cost is composed of medical expenses ($50.1 billion), as well as reduced output from missed work and school days and premature death [2]. 75% of hay fever sufferers and about 26.2% of the American population are sensitized to ragweed pollen [3,4].

One of the reasons why ragweed pollen is so allergenic is because of the amount of pollen produced by each ragweed plant [5]. One ragweed plant can yield up to 1 billion pollen grains [6]. Pollen allergens cause problems when they reach the lower airways of the lungs. For the allergens to reach the lower airways they need to be in the size range 0.12-5 μm. For ragweed pollen to be allergenic the pollen grains must fragment, since the typical size of ragweed pollen grains are 15-25 μm [7]. Ragweed allergens are in the form of ne particles that are in the size range 0.2-5.25 [5].

Pollen production and the length of the pollen season have been shown to increase with the levels of atmospheric CO_2 and temperature [4]. The rise in global temperatures and CO_2 abundance may exacerbate pollen allergies in the future (NIEHS, 2010), [8] thus making the ability to predict pollen levels progressively more useful. Accurate pollen forecasts could help those sensitive to pollen to take the appropriate precautions through measures such as staying indoors or taking medications when the pollen levels become too high.

There is already a substantial amount of research done using machine learning to forecast the pollen levels at sites throughout the world [9-12]. Some examples include Voukantsis *et al.* [12], which predicted Urticacea, Poaceae, and Oleaceae pollen levels in Thessaloniki, Greece. Csepe *et al.* [10] showed that the daily total radiation meteorological variable had moderate to high impact on the ability to model ambrosia pollen levels, which suggest that Landsat reflectance variables and other meteorological variables

are helpful in estimating pollen levels. Despite the fact that literature for using machine learning to predict pollen levels is abundant, research using Landsat reflectance data to model pollen levels is sparser. Hjort *et al.* [13] found that the TC greenness (Tasseled Cap Greenness) variable derived from Landsat images is useful in characterizing the grass pollen level in Helsinki, Finland. Other examples of the usefulness of Landsat imagery not related to pollen include applications such as the ability to determine mineral prospectivity [14] and developing a model of evapotranspiration [15].

The aim of this study is to estimate the daily pollen abundance using in-site pollen observations with machine learning and a range of diverse environmental data sets holistically characterizing the environmental state. These data sets include information from Landsat satellite images and NASA MERRA meteorological variables used in a previous paper [16]. In addition, the goal is to use machine learning to help identify the most important environmental parameters.

Methodology

Data Sources

Pollen Data: The pollen data was collected at the University of Tulsa in Oklahoma. A site where data has been collected since December 1986 [1]. Pollen readings were gathered using a Burkard Volumetric Spore Trap placed on the top of Oliphant Hall. Air is drawn into the trap and pollen grains are deposited onto a strip of Melanex tape that is fastened to a rotating drum. Each week the tape is removed and cut into 24-hour strips. The strips are then placed under a microscope at a magnification of 400and manually counted. The pollen counts are then multiplied by a conversion factor to yield the pollen concentration [1]. The final pollen numbers are not given for every day of the year, but rather for the pollen season dates. The data used in this study is formatted in a way that excludes days when no pollen was recorded. The pollen levels used in this study are plotted in Figure 1.

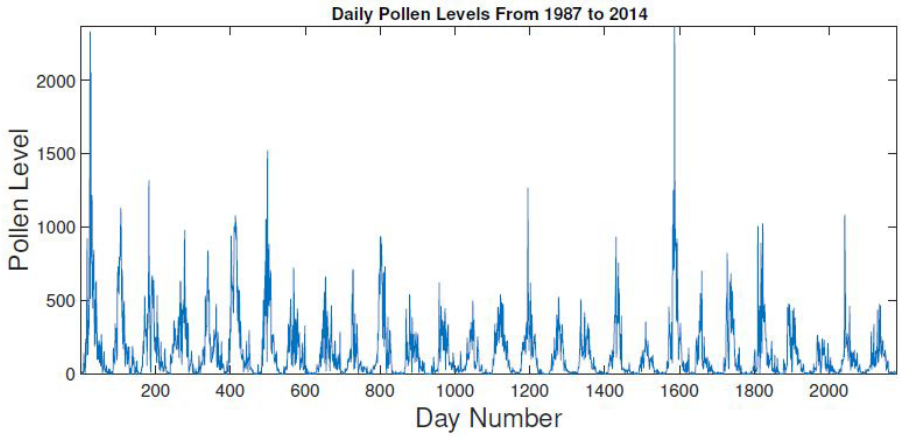


Figure 1: Actual Pollen Levels (Howard and Levetin, 2014) [1]

Contextual Input Data

The contextual environmental input variables were drawn from two data sets. The first data set came from 88 environmental parameters, given in hourly values, taken from the NASA MERRA meteorological analysis. The list of the 88 parameters and their descriptions can be seen in Table 3 in the Appendix. Since these values are hourly values, while only daily values can be used to correspond to the pollen data; the mean, minimum, and maximum were calculated for any given 24 hour period for each of these variables. These values were then time lagged between 1 and 30 days to give a 88x3x30=7,920 variables [16]. In a process that will be discussed later, this number would eventually be reduced to 500.

The second data set was attained from satellite surface reflectance values observed by Landsat 5 and 7. These images were retrieved from the United States Geological Survey website and have a standard size of 170 km north-south and 183 east-west (USGS, 2017) [17]. When Ragweed blooms and is releasing pollen the flowers are widespread and of a characteristic color. Landsat views the entire visible and near infrared spectrum, so widespread blooming should be manifest in changes in the surface reflectance spectrum seen by Landsat. The top of the atmosphere (ToA) reflectance for each pixel was calculated with a Matlab program using the formula given by Landsat (1998). This formula was applied to get the ToA reflectance for the blue, green, red, near infrared (NIR), short wave infrared 1 (SWIR1), and short wave infrared 2 (SWIR2) bands for each Landsat image that had a cloud cover of less than 10% from January 1, 1987 to December 31, 2014. To give more variety to the Landsat data, the average reflectance for all relevant pixels was retrieved for square portions centered at Oliphant Hall with bases of length 1 km, 5 km, 10 km, 15 km, 20 km, 25 km and 30 km as well as for the entire image. Table 1 shows all of the Landsat bands used in this study, along with the relevant wavelengths for Landsat 5 and 7.

One issue was that much of these square scenes from Landsat included the built urban environment and not just vegetation, which was not helpful in determining pollen levels. The built environment including buildings, roads, sports arenas, etc., most of

which can be designated under impervious surfaces. To remove those pixels that contained a significant fraction of impervious surface, a method was used similar to Deng *et al.* [18]. First, a mask detailing the location of the impervious surfaces was manually constructed for a square patch centered at Oliphant Hall with a base length of 1 km and a spatial resolution of 30 m (Landsat, 1998). Next, a set of Landsat images was selected such that this mask could be used without having to worry about changes in the urban environment i.e. construction. Decision trees (discussed in greater detail later) were created for each of the Landsat images that utilized the ToA reflectance values for the six bands of each pixel as input and the output was set as the corresponding mask value. When the decision trees performed predictions, the output of each decision tree was averaged and the average was used for the ultimate determination of impervious surfaces. So, pixels mainly containing impervious surfaces were excluded from the average reflectance calculated for each of the square patches.

Landsat Band	Landsa 5 Wavelength	Landsat 7 Wavelength
Band 1 (Blue)	0.45-0.52 μm	0.45-0.52 μm
Band 2 (Green)	0.52-0.6 μm	0.52-0.6 μm
Band 3 (Red)	0.63-0.69 μm	0.63-0.69 μm
Band 4 (Near Infrared)	0.76-0.9 μm	0.77-0.9 μm
Band 5 (Shortwave Infrared 1)	1.55-1.75 μm	1.55-1.75 μm
Band 7 (Shortwave Infrared 2)	2.08-2.35 μm	2.09-2.35 μm

Table 1: Various Landsat bands used and their corresponding Landsat 5 and 7 wavelengths shows that these wavelengths are almost exactly the same in all cases (USGS, 2017) [17]

Once these average reflectance values were calculated for every date that a good Landsat image existed, the days that did not have Landsat data were estimated using an interpolation strategy. This matrix was then time lagged for up to 31 days to provide (6 Bands)x(8averaged reflectance values)x(31 lagged days) = 1,488 Landsat variables as candidate input variables for the machine learning pollen estimator.

Statistical Methods

Machine learning is a method through which a computer learns the behavior of a system when it is given a data-set of example inputs and outputs, so that it can predict further outputs on its own. A machine learning system has the ability to test a few hundred hypotheses in a fraction of a second, a skill that could take an actual scientist their entire career [16]. Usually, giving the machine learning a larger input data-set will allow the program to draw from more information, automatically selecting the most relevant input features and potentially estimating the variable of interest with greater accuracy. In addition to having the ability to learn linear behavior, machine learning can also use the data provided by each of the input parameters in a non-linear and/or multivariate manner to deliver better results [19]. When employing machine learning methods, it is best to experiment with all of the relevant approaches so that the best approach for a given system can be identified. Machine learning can be harnessed for regression (continuous output), which was used in this study, or for classification.

When any machine learning algorithm is used, one particularly advantageous strategy is to split the original training data set into training and validation data sets. In this study, 10% of the original data set was reserved as an independent validation data set, and the other 90% was used as the training set. The machine learning algorithm will only learn from the training set. Once the algorithm has finished learning, it can estimate the output variable for both the training and validation sets. Different measures of accuracy can be derived from both sets including: root mean squared error (RMSE), mean squared error (MSE), and R^2 values. The accuracy measures from the validation data are a measure of the algorithm's performance on data not previously seen in the training, i.e. how good is the machine learning generalization.

Machine learning has proven to be useful for earth and space sciences for more than two decades [20,21]. Some interesting applications include predicting ambient ozone levels, sulfur dioxide levels, various weather phenomenon, crop disease detection, Indian monsoon rainfall, and solar radiation levels [22-26].

Further subsections will discuss various machine learning approaches used in this study.

Neural Networks

Neural networks are algorithms inspired by the function of neurons found in biological systems and are used for non-linear and non-parametric learning [22,27,28]. There is a weight associated with each node that is characterized as interaction strength between neurons. While the neural network is training, these weights can be adjusted in an effort to give a more accurate approximating function [16]. The output function should be of the form:

$$Q=b^2+\Sigma w_i^2 \;\; f(b_i^1+\Sigma(w_{ij}\, x_j)) \qquad (1)$$

where w, b, and x are the weight, bias, and the inputs respectively. The indices i and j are the neuron indices with w_{ij} being the weight between neuron i and neuron j. The function f is called the activation function and the structure of this function depends

on the type of neural network utilized. One of the more common activation functions is called the sigmoid activation function and is given by the equation:

$$f(y) = \frac{1}{1 + \exp(-y)} \tag{2}$$

In this experiment, output from 100 different neural networks was averaged, to create a more accurate output [29]. A pictorial description of the neural network structure is shown in Figure 2.

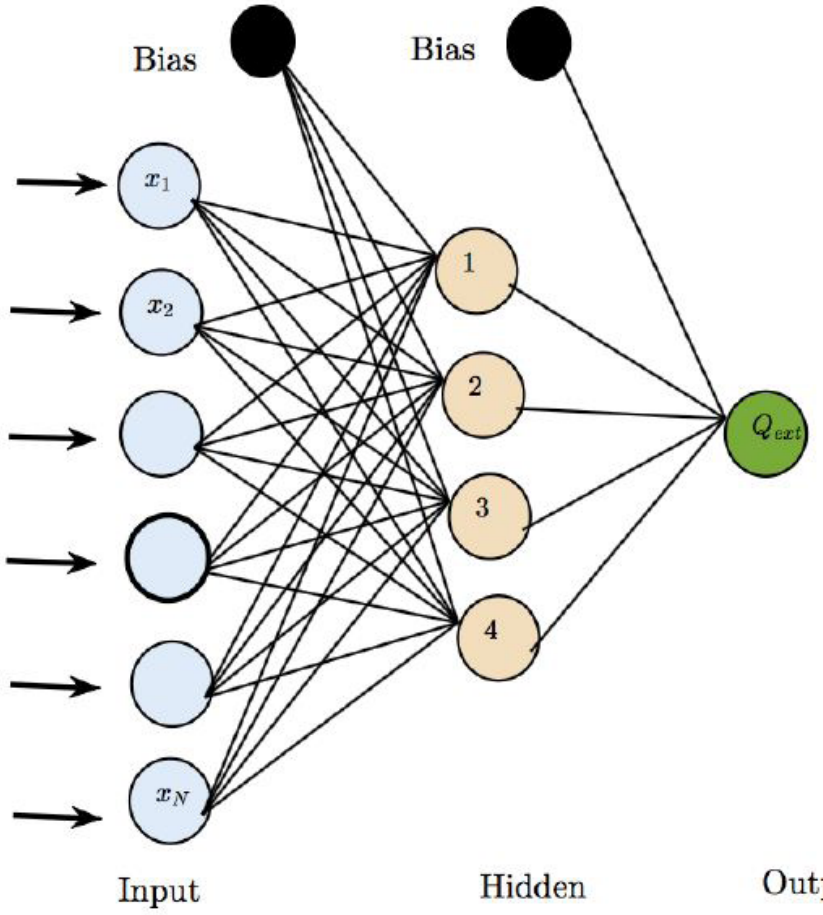


Figure 2: Neural Network Architecture. The arrows show the direction of information flow (Zewdie *et al.*, 2017)

Gaussian Processes Regression (GPR)

Gaussian processes works by defining a continuous function space that is composed of functions defined by a covariant function. The covariant functions exercised in this experiment were Matern 5/2, Exponential, Rational Quadratic, and Squared Exponential functions. These functions can be composed of any number of variables from the input data set and they have additional parameters that can be varied so that the entire function space is covered. As more scalar output is mapped into the function space, the function needs to have the parameters not defined by the input data to fall within a certain range. As a result, only functions that come relatively close to the scalar output points will be relevant for future data. When this process gives a forecast, a distribution of possible answers is comprised, with each answer having a certain probability, from which an educated solution is formulated (Rasmussen and Williams, 2006) [30].

Support Vector Machines (SVM)

The main goal of support vector machine regression is to and a function within the hypothesis space that predicts each of the individual outputs within a certain range of their corresponding targets. The maximum allowed magnitude of the difference between the individual outputs and their targets (maximum error) can be designated as ϵ. So if all of the errors between the individual outputs and their corresponding targets are less than ϵ then the support vector machine can stop learning. However, if this is not the case then the support vector machine must keep learning [31].

Linear Regression

In linear regression the training data is fit to a linear equation of the form:

The linear regression algorithm works to obtain the best fit with a linear equation through such accuracy measures as correlation

$$y=a+bx \quad (3)$$

The linear regression algorithm works to obtain the best fit with a linear equation through such accuracy measures as correlation coefficient and least squares. When using the linear regression method to predict output it is important to have the input variables

Decision Trees

The advantages of decision trees are their ability to simplify problems and turning decision making into simple if-then statements. The first node in the computational structure of the decision tree is called the root. The descendant nodes of the root are called internal nodes. A node with no descendants is no longer an internal node, but rather referred to as a terminal or a leaf. Each node in the decision tree tests an attribute of the decision making process and the leaf designates a classification or decision as output. So when a decision tree is given input, the decision making process flows from the root through any one of the root's descendants until it finally arrives at a leaf and thus produces output [33].

Ensemble Methods

An ensemble method is a way of combining the outputs from several learning algorithms to form an even better predictor than with any of the individual learning algorithms alone [34-36]. We employed various ensemble methods.

Random Forest

A random forest is a bagging method composed of an ensemble of decision trees. These decision trees provide results through the random sampling of the input data. Decision trees arrive at output by going through a series of yes or no decisions. In the case of the random forest, the training dataset will be randomly sampled for each decision tree part of the ensemble such that each tree analyzes a new input data set of the same size as the old one. So the input data that a single tree analyzes can contain multiple instances of same input from the old data set. When a random forest predicts output for an input dataset, the outputs of all of the individual decision trees are averaged together to form a final output that can be used for solving problems [37]. As the number of trees in a forest gets larger, the generalization error gets smaller [38]. A helpful aspect of random forests and decision trees is that it is very easy to retrieve the relative importance of each of the input variables, which can then be used with other machine learning algorithms. Once the relative importance of the input variables is determined, we can use Occam's razor and utilize only the most important inputs to build a more robust regression model. If input variables are highly correlated, their relative importance decreases [39]. A schematic of a random forest can be seen in the figure below:

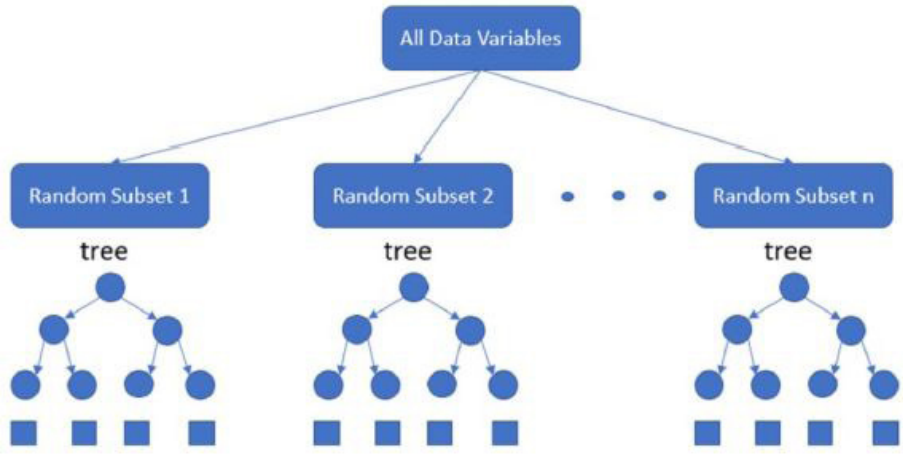


Figure 3: RRandom Forest Structure (Liu *et al*., 2017) [16]

We then also performed an additional step inspired by Newton-Raphson iteration. In the case of the pollen experiment, this procedure works by first having one random forest predict the pollen levels. Then the error of this estimate is calculated with the equation:

$$error = observation - estimation \quad (4)$$

Now, a second random forest learns how to estimate this error. The estimated pollen count from the first random forest is included in the input data for the second random forest. The estimated error can then be used to correct the estimate from the first random forest. This process can be repeated iteratively [16].

Boosting

Boosting is a method in which an ensemble of weak learners is combined to create a strong learner. Boosting works by first having a weak learner obtain information from a random subset of the input data. The accuracy of this weak learner acting on this subset is assessed and then a different weak learner learns new information from a new random subset of input data. Once again the

accuracy is assessed and this process is repeated for a series of weak learners. The boosting algorithm assigns weights to each of the possible examples from the input data such that if a previous weak learner was bad at learning a particular example another weak learner can go back and learn it even better. When making a prediction, the boosting algorithm arrives at an output by giving a weighted vote to each of the weak learners [40].

Relative Importance

One feature that can be very useful is the relative importance feature and can be easily retrieved from several machine learning and ensemble methods including decision trees and random forests. The relative importance of each input feature allows us to identify the most important input features. This can be used to reduce the number of input features to just the most important, usually this improves the model generalization. The three strategies for calculating the relative importance are filter methods, wrapper methods, and embedded methods. The filter method works by ranking individual variables by correlation coefficients. Wrapper methods test subsets of variables based on their accuracy in determining the target. Embedded methods are similar to wrapper methods with the exception that they involve a function that maximizes the goodness of t and minimizes the number of variables [41].

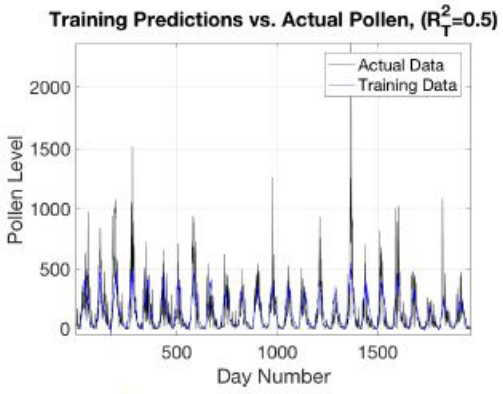
Results

Table 2 shows the performance of various machine learning algorithms used in this study. The approaches are ranked in descending order using their R^2 values. The neural network had a validation R^2 value of 0.59, which is an exciting result because this is a 0.22 increase from only using the NASA MERRA meteorological analysis variables, whose validation R^2 value was $R_v^2 = 0.37$. The data from the neural networks, including the the training prediction, the validation prediction, the t, and the residual plot, can be seen in Figure 4.

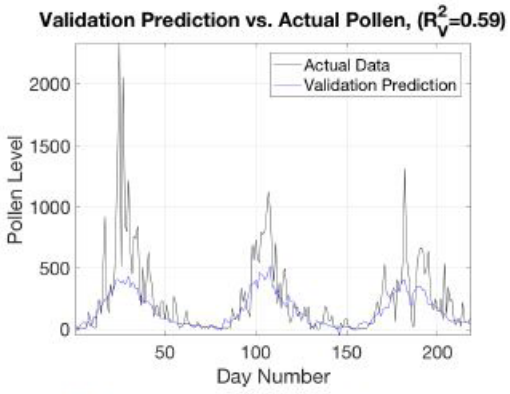
Machine Learning Approach	Validation R^2	RMS Error
GPR Matern 5/2	0.68	122.16
GPR Exponential	0.68	121.47
GPR Rational Quadratic	0.68	121.66
GPR Squared Exponential	0.67	122.96
Bagged Trees	0.63	130.94
Boosted Trees	0.60	136.35
Cubic SVM	0.60	136.45
Neural Network	0.59	244.77
Quadratic SVM	0.53	147.84
Medium Tree	0.50	151.56
Medium Gaussian SVM	0.47	156.11
Random Forest	0.46	412.89
Coarse Tree	0.45	159.61
Fine Tree	0.44	161.4
Ensemble Hyperparameters	0.43	283.84
Fine Gaussian SVM	0.28	182.33
Linear SVM	0.23	188.89
Coarse Gaussian SVM	0.14	199
Linear Regression	0.10	203.75
Interactions Linear Regression	0.08	222.93
Robust Linear Regression	0.03	218.43

Table 2: Performance of the various machine learning approaches

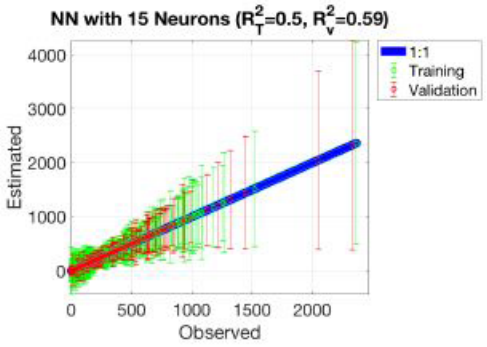
The final step was to perform the Newton-Raphson iterative improvement. The figures below show the results from this method. Figure 5a shows the t and the relevant correlation coefficients from the first iteration. The validation R^2 value for this plot is $R_v^2 = 0.46$. This is a meaningful result due to the fact that this number is 0.05 higher than the equivalent coefficient with only the NASA MERRA meteorological analysis as input, which had $R_v^2 = 0.41$. R_v for 10 iterations is shown in Figure 5b. After 10 iterations, the R_v^2 was exactly the same in this experiment as with the previous experiment, with both being $R_v^2 = 0.96$ [16]. Figure 5c is a plot of both the training and validation correlation coefficients as a function of iteration. Figure 5d is a bar graph of the relative importance of inputs calculated by the Newton-Raphson method after 1 iteration. Figures 5e and 5f show a good fit of the training and validation predictions for the pollen level after 10 iterations.



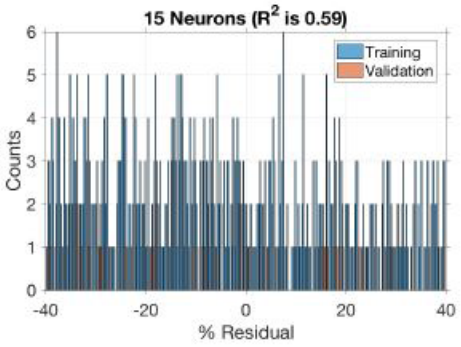
(a) Pollen prediction from training set.



(b) Pollen prediction from validation set.

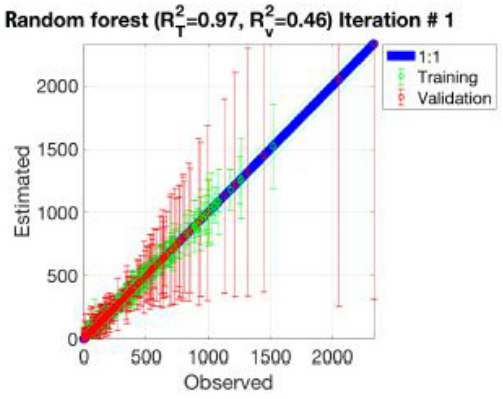


(c) Fit from neural network algorithm.

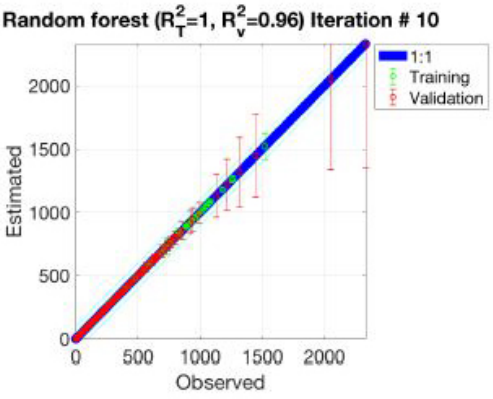


(d) Residual from neural network algorithm.

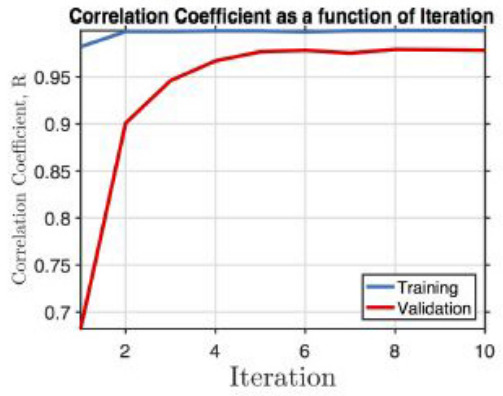
Figure 4: Neural network results using 15 neurons



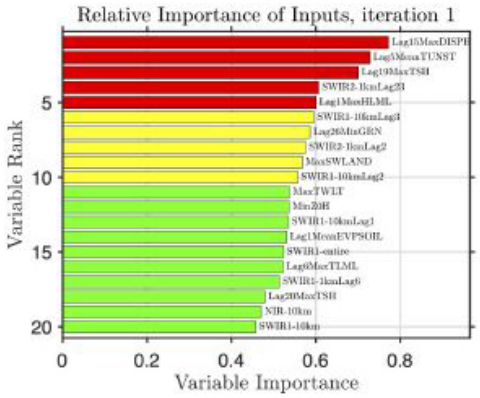
(a) Fit after 1 iteration.



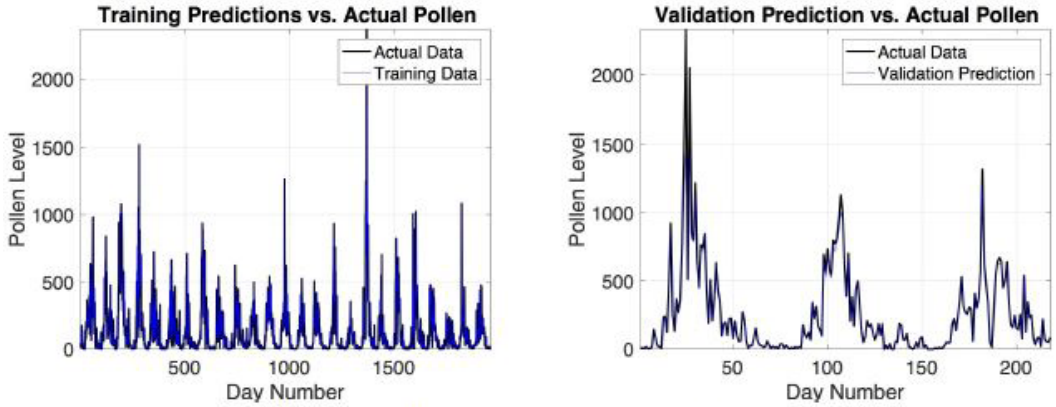
(b) Fit after 10 iterations.



(c) Correlation coefficients as a function of iteration.



(d) Relative Importance after 1 iteration.



(e) Training prediction of pollen. (f) Validation prediction of pollen.

Figure 5: Results from Newton-Raphson method using 10 iterations

Figure 6 shows the top 20 most important inputs from both the NASA MERRA meteorological analysis and the Landsat reflectance values. The relative importance, as stated earlier, can be easily obtained from any decision tree, bag of trees, random forests, etc. In this case it was derived from using the optimized hyper-parameters feature in the Matlab function fitrensemble, which is a function that uses both boosting and bagging ensembles of regression trees and can optimally choose the hyper-parameters that provide the best performance. This method is an effective way to get a good measurement of the relative importance because the relative importance can be easily retrieved and it is supposed to be more accurate than other methods. This chart shows that two bands in particular are taking up more than half of the top 20 positions. These bands are the Near Infrared (NIR) and Short Wave Infrared 1 (SWIR1). The fact that these two bands are so important gives valuable information about the pollen formation process. The SWIR1 reflectance is highly sensitive to the water concentration in plants, which can give information about the plants metabolic activity, since water is one of the main components of photosynthesis. The NIR reflectance has a close association with such variables as green biomass, leaf area index, leaf cover, fraction of radiation intercepted, and chlorophyll per unit ground area [42].

Physical Insights

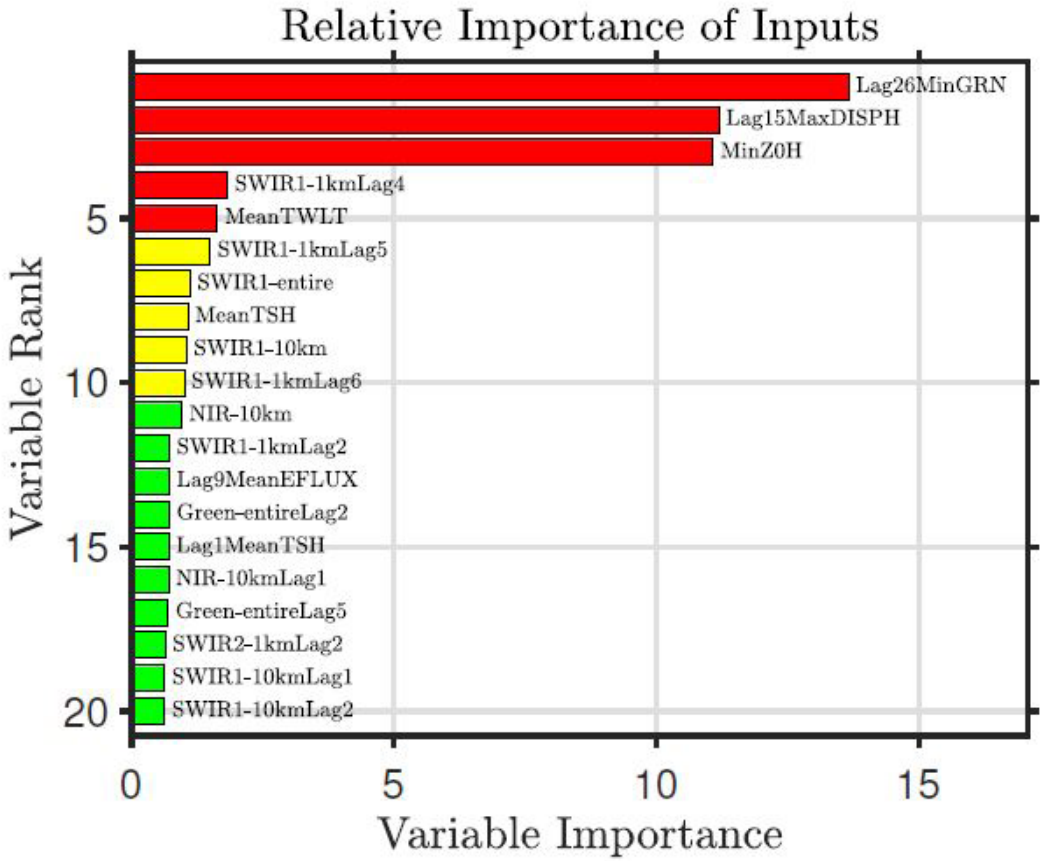


Figure 6: Relative Importance when solving for Hyperparameters

With regard to some of the more important results in Figure 6 from the NASA MERRA variables, MinZOH represents the minimum value over a 24 hour period of the roughness length of sensible heat. This value helps determine the rate at which heat and water vapor transfer from an evaporating surface to a vegetation area, with lower MinZOH values meaning a lower transfer rate [43]. Lag15MaxDISPH is the maximum displacement height over a 24 hour period from 15 days prior. The displacement height is a good indicator of vegetation density with an increase in displacement height correlating to an increase in vegetation density [44]. The Lag26MinGRN variable is the minimum value of the vegetation greenness fraction over a 24 hour period from 26 days prior. These results from both the Landsat reflectance and the NASA MERRA variables along with the fact that high atmospheric CO_2 concentrations mean high pollen production, as stated in the introduction, indicate that healthy, highly metabolic ambrosia ragweed plants cultivate the most pollen. The vegetation greenness fraction from 26 days prior probably demonstrates that the Ambrosia plant has to be healthy for a certain number of days for large amounts of pollen to be generated.

Summary

In this paper, 21 different machine learning approaches for predicting the ragweed pollen levels were tested. These models used reflectance data from several bands of both Landsat 5 and Landsat 7 satellites as well as data from NASA MERRA meteorological analysis variables, which were drawn from a previous paper [16]. These results have added several new variables of high relative importance for forecasting the ragweed pollen levels. The important new parameters include several instances of short wave infrared (SWIR1) and near-infrared (NIR) reflectance of various time lags and areas of interest. There are also two instances of the green reflectance being of high importance.

Of all the machine learning approaches, the Gaussian Processes Regression (GPR) was the most effective in terms of high R2 values, with GPR Matern 5/2 being the best performing. However, it should be stressed that there were also substantial improvements in predictive capability for neural networks, bagged trees and random forests. Further research will go into combining the Landsat reflectance and NASA MERRA meteorological analysis variables; in particular a future improvement will be to include NEXRAD weather radar parameters as input with the hope of further increasing the accuracy of machine learning in predicting pollen levels.

References

1. Howard LE, Levetin E (2014) Ambrosia pollen in Tulsa, Oklahoma: aerobiology, trends, and forecasting model development. Ann Allergy Asthma Immunol 113: 641-6.

2. CDC (2013) Asthma facts|cdc's national asthma control program grantees. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention.

3. Arbes SJ, Gergen PJ, Elliott L, Zeldin DC (2005) Prevalences of positive skin test responses to 10 common allergens in the us population: results from the third national health and nutrition examination survey. J Allergy Clin Immunol 116: 377-83.

4. Kinney PL (2008) Climate change, air quality, and human health. Am J Prev Med 35: 459-67.

5. Bacsi A, Choudhury BK, Dharajiya N, Sur S, Boldogh I (2006) Subpollen particles: carriers of allergenic proteins and oxidases. J Allergy Clin Immunol 118: 844-50.

6. Thompson JL, Thompson JE (2003) The urban jungle and allergy. Immunol Allergy Clin North Am, 23: 371-87.

7. Oswalt ML, Marshall GD (2008) Ragweed as an example of worldwide allergen expansion. Allergy Asthma Clin Immunol 4: 130-5.

8. NIEHS (2010) A human health perspective on climate change: A report outlining the research needs on the human health effects of climate change. In A Human Health Perespective On Climate Change: A Report Outlining the Research Needs on the Human Health Effects of Climate Change. Environmental Health Perspectives (EHP); National Institute of Environmental Health Sciences.

9. Astray G, Fernández-González M, Rodríguez-Rajo FJ, López D, Mejuto JC (2016) Airborne castanea pollen forecasting model for ecological and allergological implementation. Sci Total Environ 548: 110-21.

10. Csépe Z, Makra L, Voukantsis D, Matyasovszky I, Tusnády G, et al. (2014). Predicting daily ragweed pollen concentrations using computational intelligence techniques over two heavily polluted areas in europe. Sci Total Environ 476: 542-52.

11. Rodríguez-Rajo FJ, Astray G, Ferreiro-Lage JA, Aira MJ, Jato-Rodriguez MV, et al. (2010). Evaluation of atmospheric poaceae pollen concentration using a neural network applied to a coastal atlantic climate region. Neural Netw 23: 419-25.

12. Voukantsis D, Niska H, Karatzas K, Riga M, Damialis A, et al. (2010) Forecasting daily pollen concentrations using data-driven modeling methods in thessaloniki, greece. Atmos Environ 44: 5101-11.

13. Hjort J, Hugg TT, Antikainen H, Rusanen J, Sofiev M, et al. (2016) Fine-scale exposure to allergenic pollen in the urban environment: evaluation of land use regression approach. Environ Health Perspect 124: 619-26.

14. Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M (2015) Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol Rev 71: 804-18.

15. Ke Y, Im J, Park S, Gong H (2016) Downscaling of modis one kilometer evapotranspiration using landsat-8 data and machine learning approaches. Remote Sens 8: 215.

16. Liu X, Wu D, Zewdie G K, Wijerante L, Timms CI, et al. (2017) Using machine learning to estimate atmospheric ambrosia pollen concentrations in tulsa, ok. Environ Health Insights 11: 1178630217699399.

17. USGS (2017) What are the band designations for the landsat satellites?

18. Deng Y, Fan F, Chen R (2012) Extraction and analysis of impervious surfaces based on a spectral un-mixing method using pearl river delta of china landsat tm/etm+ imagery from 1998 to 2008. Sensors, 12: 1846-62.

19. Landsat (1998) Landsat science data users handbook. Technical report. Lary, D. J. (2010). Artificial intelligence in geoscience and remote sensing. In Geoscience and Remote Sensing New Achievements. InTech.

20. Gardner MW, Dorling S (1998) Artificial neural networks (the multilayer perceptron)|a review of applications in the atmospheric sciences. Atmos Environ 32: 2627-36.

21. Boznar M, Lesjak M, Mlakar P (1993) A neural network-based method for shortterm predictions of ambient so2 concentrations in highly polluted industrial areas of complex terrain. Atmos Environ Part B Urban Atmos 27: 221-30.

22. Bishop C M (1995) Neural networks for pattern recognition. Oxford university press.

23. Elizondo D, Hoogenboom G, McClendon R (1994) Development of a neural network model to predict daily solar radiation. Agric For Meteoro l 71: 115-32.

24. Lary DJ, Alavi AH, Gandomi AH, Walker AL (2016) Machine learning in geosciences and remote sensing. Geoscience Frontiers 7: 3-10.

25. Navone, H, Ceccatto H (1994) Predicting indian monsoon rainfall: a neural network approach. Climate Dynamics 10: 305-12.

26. Yi J, Prybutok VR (1996) A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environ Pollut, 92: 349-57.

27. Friedman J, Hastie, T, Tibshirani R (2001) The elements of statistical learning,volume 1. Springer series in statistics New York.

28. Werbos PJ (1974) Beyond regression: New tools for prediction and analysis in the behavioral sciences. Doctoral Dissertation, Applied Mathematics, Harvard University, MA.

29. Zewdie GK, Liu X, Wu D, Lary DJ (2017) Applying machine learning to forecast daily ambrosia pollen using environmental and nexrad radar parameters. Environmental Monitoring and Assessment.

30. Rasmussen CE, Williams CK (2006) Gaussian Processes for Machine Learning. MIT press.

31. Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14: 199-222.

32. Yale (1997) Linear regression

33. Yang T (2006) Computational verb decision trees. Int J Comput Cognition 4.

34. Opitz DW, Maclin R (1999) Popular ensemble methods: An empirical study. J Artif Intell Res 11: 169-98.

35. Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits and systems magazine 6: 21-45.

36. Rokach L (2010) Ensemble-based classifiers. Artificial Intelligence Review 33: 1-39.

37. Breiman L (1996) Bagging predictors. Machine learning 24: 123-40.

38. Breiman L (2001) Random forests. Machine learning 45: 5-32.

39. Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recognit Lett, 31: 2225-36.

40. Schapire RE, Freund Y (2012) Boosting: Foundations and algorithms. MIT press.

41. Guyon I, Elissee A (2003) An introduction to variable and feature selection. J machine learning Res, 3: 1157-82.

42. Pe~nuelas J, Filella I (1998) Visible and near-infrared reflectance techniques for diagnosing plant physiological status. Trends in plant sci 3: 151-6.

43. Allen RG, Pereira LS, Raes D, Smith M (1998) Crop evapotranspirationguidelines for computing crop water requirements-fao irrigation and drainage paper 56. FAO Rome 300: D05109.

44. Shaw RH, Pereira A (1982) Aerodynamic roughness of a plant canopy: a numerical experiment. Agricultural Meteorology 26: 51-65.