



Applying machine learning to forecast daily *Ambrosia* pollen using environmental and NEXRAD parameters

Gebreab K. Zewdie  · Xun Liu · Daji Wu · David J. Lary · Estelle Levetin

Received: 4 January 2017 / Accepted: 20 March 2019
© Springer Nature Switzerland AG 2019

Abstract Approximately 50 million Americans have allergic diseases. Airborne plant pollen is a significant trigger for several of these allergic diseases. *Ambrosia* (ragweed) is known for its abundant production of pollen and its potent allergic effect in North America. Hence, estimating and predicting the daily atmospheric concentration of pollen (ragweed pollen in particular) is useful for both people with allergies and for the health professionals who care for them. In this study, we show that a suite of variables including meteorological and land surface parameters, as well as next-generation radar (NEXRAD) measurements together with machine learning can be used to esti-

mate successfully the daily pollen concentration. The supervised machine learning approaches we used included random forests, neural networks, and support vector machines. The performance of the training is independently validated using 10% of the data partitioned using the holdout cross-validation method from the original dataset. The random forests ($R=0.61$, $R^2=0.37$), support vector machines ($R=0.51$, $R^2=0.26$), and neural networks ($R=0.46$, $R^2=0.21$) effectively predicted the daily *Ambrosia* pollen, where the correlation coefficient (R) and R -squared (R^2) values are given in brackets. Three independent approaches—the random forests, correlation coefficients, and interaction information—were employed to rank the relative importance of the available predictors.

This article is part of the Topical Collection on *Geospatial Technology in Environmental Health Applications*

G. K. Zewdie (✉) · X. Liu · D. Wu · D. J. Lary
William B. Hanson Center for Space Sciences,
The University of Texas at Dallas, Richardson, TX, USA
e-mail: gebreab.zewdie@utdallas.edu

X. Liu
e-mail: xun.liu@utdallas.edu

D. Wu
e-mail: daji.wu@utdallas.edu

D. J. Lary
e-mail: david.lary@utdallas.edu

E. Levetin
The University of Tulsa, Tulsa, OK 74104, USA
e-mail: estelle-levetin@utulsa.edu

Keywords Pollen · Machine learning · Environmental parameters · NEXRAD measurements

Introduction

Pollen is known to be a trigger for allergic diseases, e.g., asthma, hay fever, and allergic rhinitis (Oswalt and Marshall 2008; Howard and Levetin 2014). It is interesting that a variety of non-respiratory issues such as strokes (Low et al. 2006; Matheson et al. 2008), and surprisingly, even suicide and attempted suicide (Postolache et al. 2005; Stickley et al. 2017)

have an association with the daily concentration of atmospheric pollen and particulates in general. However, so far, there is no defined threshold amount of pollen known to trigger allergy for sensitive individuals (Voukantsis et al. 2010). One of the factors for the lack of knowledge of the threshold amount of pollen is the absence of an accurate estimation on a fine spatial scale of the hourly, bi-hourly, or daily amount of pollen. Individual differences among sensitive people such as gender and age also adversely affect in knowing the threshold amount of pollen in the surroundings (Britton et al. 1994; Ernst et al. 2002).

Of all plants, weeds, and particularly those of the *Ambrosia* species, e.g., *Ambrosia artemisiifolia* (common ragweed) and *Ambrosia trifida* (giant ragweed), are major producers of large amounts of pollen. For example, common ragweed can produce up to about 2.5 billion pollen grains per plant per day (Laaidi et al. 2003). *Ambrosia artemisiifolia* and *Ambrosia trifida* together can produce more allergens than all other plants combined (Lewis et al. 1983). Grasses (e.g., rye grass) are also known to trigger an allergic response. Following *Ambrosia artemisiifolia*, grass pollen are known for their high allergic potency compared with other weeds (Esch et al. 2001; Lewis et al. 1983). Tree pollen can cause an allergic response, but one that is typically less than that of weeds and grasses, although in some regions tree pollen can trigger a significant allergic response. For instance, the airborne concentration of mountain cedar pollen grains can reach tens of thousand of pollen grains per cubic meter and trigger a significant allergic response in central Texas during winter, known as cedar fever (Andrews et al. 2013; Ramirez 1984).

Both global climate change and air pollution affect the abundance of airborne pollen, and consequently, its allergic impact (Kinney 2008; Wayne et al. 2002; Voukantsis et al. 2010). For example, the abundance of pollutants such as CO₂ (Wayne et al. 2002) and NO₂ (Zhao et al. 2016) can affect the extent of growing region of major pollen-producing plants, and thereby also affect the airborne pollen concentration as well as altering the onset and end dates of seasonal allergies. Overall, more people are exposed to pollen and sensitive individuals become exposed to large amount of pollen for longer period of time over larger areas.

Globally, millions of people are affected by seasonal allergies, and the number of people affected is increasing each year. In North America alone, as

of 2008, about 50 million adult Americans and 9% of children aged below 18 have experienced pollen-caused allergies (Howard and Levetin 2014). Similarly, in Europe, about 15 million people are affected by hay fever, asthma, and rhinitis (D'amato and Spieksma 1991). Hence, pollen allergies are becoming an increasingly significant environmental health issue. Thus, just as accurate daily weather forecasts are of significant use, accurate daily pollen forecasts are likely to become increasingly important.

Remote sensing has been employed to study atmospheric pollen concentrations. For example, the polarization of LIDARs has been used to observe the airborne tree pollen abundance at Fairbanks, AK (Sassen 2008). In this case, the pollen produces a depolarization of the LIDAR backscattering signals from the lower atmosphere. The light scattering properties of pollen are also manifested in the shape of the solar corona they create. The shape of the solar corona associated with pollen depends on the shape of the pollen grains and their atmospheric concentration (Tränkle and Mielke 1994). However, this approach can be complicated as atmospheric light scattering is also caused by other airborne particulates.

Common pollen estimation techniques, particularly those made in Europe, stress the importance of meteorologic variables (Kasprzyk 2008). Usually, forecasting the amount of airborne pollen is based on the interaction of atmospheric weather and pollen (Arizmendi et al. 1993). Meteorologic variables such as the daily mean, maximum, and change in temperature, dew point, wind speed, and wind direction show positive correlation with the pollen concentration, whereas atmospheric humidity and rainfall are negatively correlated to the increase in the pollen concentration (Kasprzyk 2008). Other studies show that daily temperature, precipitation, and wind speed are significant meteorologic parameters in estimating pollen concentration (Stark et al. 1997).

Most of these meteorologic variable-based forecasting techniques employed statistical methods such as linear regression, the polynomial method (non-linear regression in which the relationship between the input and response variables is modeled to a degree of polynomial n), and time series analysis (Sánchez-Mesa et al. 2002). Only few studies used advanced machine learning methods such as neural networks (Csépe et al. 2014; Sánchez-Mesa et al. 2002; Rodríguez-Rajo et al. 2010; Puc 2012; Voukantsis

et al. 2010) and random forests (Nowosad 2015) for pollen forecasting and support vector machines are applied for related environmental studies (Voukantsis et al. 2010; Osowski and Garanty 2007). Liu et al. (2017) used 85 meteorological variables along with up to 30 days of lagging and machine learning methods to estimate ambrosia pollen.

To the best of our knowledge, no one has so far estimated atmospheric pollen abundance using machine learning and meteorological data provided by NEXRAD measurements: wind direction and speed vertical profiles, reflectivity, Doppler velocity, and spectral width nor the recently upgraded polarimetric measurements: differential reflectivity, differential phase, and correlation coefficient. Detailed description of the NEXRAD parameters is given by Doviak and Zrnic (2014). A combination of the readily available meteorological and environmental data such as the daily temperature, humidity, rainfall, and land surface greenness fraction and measurements made by the NEXRAD (e.g., reflectivity, wind direction and speed, spectral width) can provide the potential to predict allergenic pollen using advanced regression and machine learning methods.

This study contributes an advance by using both the NEXRAD parameters and advanced machine learning methods for forecasting the daily *Ambrosia* pollen in the atmosphere. The objective of this paper is, therefore, to use machine learning (neural networks, random forests, and support vector machines) together with a suite of NEXRAD and other environmental parameters to forecast the daily pollen concentration.

Materials and methods

Data description

Environmental and NEXRAD parameters

After pollen is produced in plant anthers, its emission, dispersion, and deposition are influenced by meteorological variables such as the temperature, wind speed and direction, and pressure (Kasprzyk 2008; Csépe et al. 2014; Howard and Levetin 2014). Other meteorologic parameters such as dew point, humidity, rainfall, and sunshine duration are also known to affect pollen emission and distribution (Kasprzyk 2008).

In this study, we used a set of environmental and NEXRAD parameters (Table 1) in our machine learning training. Environmental parameters such as vegetation greenness fraction, roughness length (sensible heat), energy stored in all land reservoirs, displacement height, and leaf area index are selected. The other sets of data we used are the NEXRAD measurements which consist of the reflectivity, Doppler velocity, and spectral width which represent, respectively, the amount of backscattered signals from a scattering volume, the velocity of the scatterers along the radar line of sight and the width of the power spectrum. All NEXRAD measurements are taken at the lowest elevation, 0.5° from the surface of the Earth. Additionally, the NEXRAD provides measurements of the vertical profile of the direction and speed of the wind starting from 50 m from the surface of the Earth. The dual polarization measurements—differential reflectivity, differential phase, and correlation coefficient—use the horizontal and vertical polarization signals and are particularly suited for particle identification. However, we do not use the dual

Table 1 Name and type of predictors (input variables) used for our machine learning training. Parameters consist of environmental (Env.) and NEXRAD measurements

| Parameter | Unit | Type |
|---------------------------------|-----------|--------|
| Vegetation greenness fraction | Fraction | Env. |
| Leaf area index | m^2 | Env. |
| Roughness length, sensible heat | m | Env. |
| Displacement height | m | Env. |
| Energy stored in land | Jm^{-2} | Env. |
| Mean reflectivity | dB | NEXRAD |
| Mean doppler velocity | ms^{-1} | NEXRAD |
| Mean spectral width | ms^{-1} | NEXRAD |
| Reflectivity [−10 to 10 dB] | dB | NEXRAD |
| Velocity [−10 to 10 dB] | ms^{-1} | NEXRAD |
| Spectral width [−10 to 10] dB | ms^{-1} | NEXRAD |
| Reflectivity [−20 to 20 dB] | dB | NEXRAD |
| Velocity [−20 to 20 dB] | ms^{-1} | NEXRAD |
| Spectral width [−20 to 20 dB] | ms^{-1} | NEXRAD |
| Reflectivity [−40 to 40 dB] | dB | NEXRAD |
| Velocity [−40 to 40 dB] | ms^{-1} | NEXRAD |
| Spectral width [−40 to 40 dB] | ms^{-1} | NEXRAD |
| Wind direction at altitude 50 m | Degree | NEXRAD |
| Wind speed at altitude 50 m | ms^{-1} | NEXRAD |

polarization (polarimetric) NEXRAD measurements as we have only few days of the measurements in contrary to the ideal high-dimensional data requirement for machine learning.

Daily pollen data

Description of the pollen data used to train the machine learning in this study is given by Howard and Levetin (2014). The pollen grains were collected using a Burkard volumetric spore trap (Hirst 1952) at the University of Tulsa. The spore trap apparatus is placed at the top of a building 12 m from the surface of the ground. The Burkard trap is a classical, manual-intensive method in which air containing pollen is directed into the sampler through a small orifice and deposited on to a sticky slide attached to a rotating drum. The drum rotates at 2 mm per hour. After a weekly sampling period, the tape is changed and the sample carefully cut into strips for each day (24 h) period. After drying and applying the necessary solution, the slides are observed under a microscope at $\times 400$ magnification for counting.

Observation of *Ambrosia* pollen at the University of Tulsa, OK, started in 1987. Howard and Levetin (2014) made an analysis of the first 27 years of the pollen data. They found that the mean annual start and end dates are August 22 and October 20, respectively, and that the mean peak date of *Ambrosia* pollen in Tulsa is September 10. We use the same pollen dataset in this study. Due to lack of NEXRAD data from 1987 to 1994, the pollen data in this period are not included in our study. Figure 2 shows the actual pollen observations made for 20 years from 1995 to 2014 for the high *Ambrosia* pollen season (mid August to end of October).

Machine learning methods

Machine learning is a mathematical approach that allows computers to “learn by example” and extract information from data, often very large amounts of data. It has been applied to various fields in geosciences and remote sensing, agriculture, banking, etc., for code acceleration and detection of diseases in crops (Lary 2010; Lary et al. 2018), and prediction of atmospheric gases such as CO₂ (Gardner and

Dorling 1998) and ozone (Yi and Prybutok 1996; Prybutok et al. 2000). Beyond geosciences, it is used very widely for applications such as for spam filtering (Guzella and Caminhas 2009), credit scores, fraud detection, and image processing.

Machine learning methods can learn the behavior of the system and retrieve the necessary information if they are provided with data spanning as many parameters as possible in the training. It can “learn” the behavior of the system even in the case the relation between the information and the parameters is non-linear and multivariate (Lary 2010). We do not need to know a priori the functional form relating the input variables to the parameter(s) being estimated.

Some commonly used machine learning approaches that include neural networks, support vector machines, decision trees, and random forests (an ensemble of decision trees). The applications of these methods can be put into two broad categories, regression and classification. Some methods, for example, the random forest, neural network, support vector machines, can do both regression and classification. Although there are different types of machine learning algorithms currently used, there is no single method that always will perform better than the rest for all problems. The best machine learning method to apply depends on the problem and the available training data (Kotsiantis 2007). The following subsections briefly describe the various machine learning approaches that we have employed.

Neural networks

Neural networks are “learning” algorithms “inspired by interconnection and information flow among neurons in the human brain (Haykin 1994; Haykin et al. 2001, 2009). A neural network employs a system of interconnected processing units called “artificial neurons.” The basic idea in a neural network is to model the response variable (output) based on a non-linear combination of the input variables (Friedman et al. 2001). Information in a neural network is represented by the interaction strengths of the neurons (the weights). A neuron (node) receives information from another neuron or an external input variable. The weighted linear sum of the input signals represents the body of the neuron. The weight associated with the input can be modified to imitate the synaptic learning. The neuron computes a function f based on

the weighted sum of the inputs. The output of the neural network structure shown in Fig. 1 can be written as:

$$Q = b^2 + \sum_i^4 w_i^2 f \left(b_i^1 + \sum_{j=1}^N w_{ij} x_j \right) \tag{1}$$

The function f is called the activation or transfer function and can be linear or non-linear. x_j , w_{ij} , and b , respectively, represent the inputs, the weight from neuron j to neuron i , and the biases. The activation function uses the input values and determines the output activity of the neuron. Different neural network structures may have different types of activation functions but the basic inherent structure of the neuron, linear sum of the inputs followed by an activation function, is the standard to all networks. Linear, threshold functions and non-linear Gaussian and sigmoid functions are common examples of activation functions. The sigmoid activation function is given by Eq. 2

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

The neural network connection can be feed-forward or feed-backward showing the flow of information. Figure 1 shows a schematic diagram of feed-forward

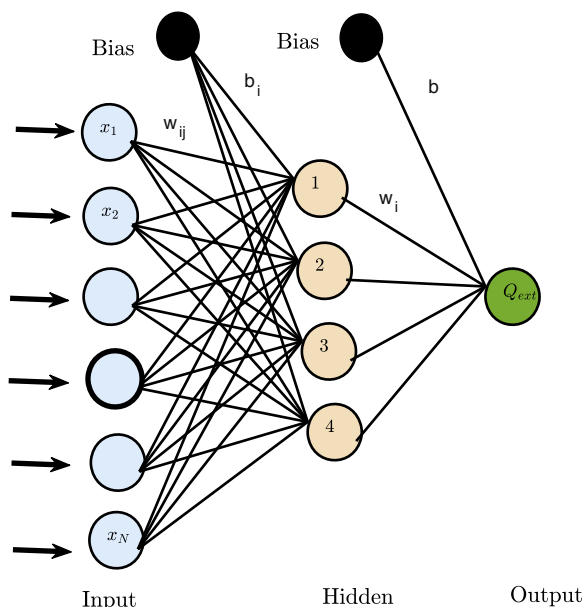


Fig. 1 The neural network architecture consisting of the input, hidden and output layers consisting respectively N, four and one neuron

neural network consisting of the N inputs, 4 hidden and a single output layers. The arrows show the direction of flow of information in the network.

Random forest

Another important ensemble approach machine learning method is the random forest introduced by Breiman (2001). A random forest works based on random sampling of data to form ensemble of decision trees. Each tree will provide its “vote” to make a decision. When the number of trees in the forest gets larger, the generalization error gets smaller (Breiman 2001). After a number of trees are grown, internal estimates are made for regression and to calculate variable importance. Random forests can perform prediction and outlier detection (Friedman et al. 2001; Verikas et al. 2011). Random forests also provide a useful facility to rank the relative importance of the input variables. However, the presence of highly correlated variables in the training results in the reduction of the value of variable importance (Genuer et al. 2010).

Support vector machines

Support vector machines were introduced by Vapnik (2013, 1998). Support vector machines employ hyperplanes that define decision boundaries separating the data into two classes. The best hyperplane is the one that separates the data into two classes with a large marginal distance between the hyperplane and the classes. The simplest example is the linear classifier that separates the data into their respective classes using a line. In the general case, the data cannot be separated by a straight line and complex structures are needed to separate the data leading to a non-linear classifier. For regression, an important non-linear function is learned in a high dimensional space that maps the input variables (Basak et al. 2007). Mathematical analysis of the linear and non-linear support vector machines is given by Smola and Schölkopf (2004).

Procedure

Measurements of different input parameters are made on different scales. Hence, it is a common machine learning process to normalize all parameters to lie

between 0 and 1. The normalization can be carried out using the simple technique of dividing each parameter by its corresponding maximum value. The support vector machines highly depend on normalized data whereas the random forest works well independent of normalization. However, unnormalized output values of the support vector machines can be estimated by applying the normalization parameters after training.

The entire dataset consisting of a combination of environmental parameters, NEXRAD measurements, and the response variable for several days of observation are separated into training and validation sets using the holdout cross-validation partitioning technique (Kohavi and et al. 1995). Only 10% of the data is holdout for independent validation and the remaining 90% is used for the training process. The proportion to split the data into training and validation can be done in many ways depending on the amount of data we have. In ideal situations, we train the machine learning on big data and validate on another big data (Witten and Frank 2005).

Table 1 presents the environmental variables and NEXRAD measurements used as predictors. We processed the NEXRAD data into different signal to noise ratio (in dB) levels to optimize the performance of the machine learning. Spatial and mean values of the scattering in the signal to noise ratio from -10 to 10 dB, -20 to 20 dB, and -40 to 40 dB are used as separate predictors. The separation of the NEXRAD data into different signal to noise ratio levels is important as we do not know a priori the amount of scattering coming from biological scatterers such as pollen, clouds, precipitation, dust, or even insects (Vivekanandan et al. 1999; Hannesen and Weipert 2003; Wilson

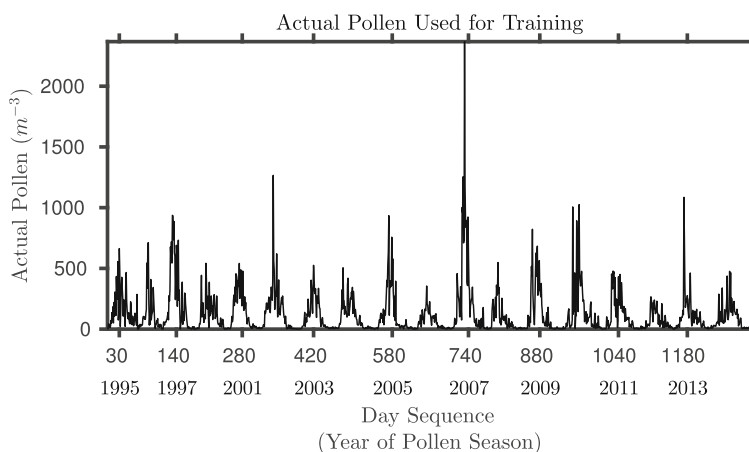
et al. 1994). In clear air mode, the NEXRAD moves slowly and is able to detect small objects such as pollen, dust, and smoke (Gali 2010). Hence, using clear air NEXRAD data in the training will improve the quality of the data for the machine learning.

The daily pollen count (shown in Fig. 2) is used as the target parameter to train the neural network, random forest, and support vector machines. A total number of 19 variables are used for training the three machine learning algorithms. The generalized regression neural network (Specht 1991) in the function approximation sense is applied to train the neural network. The random forest machine learning is trained using 200 decision trees. The support vector machine is trained using a Gaussian kernel function to map the 19 predictor data.

Results

As mentioned in Section 2, in order to evaluate the performance of the machine learning methods independently, we applied the holdout cross-validation partitioning technique (Kohavi and et al. 1995) to split the data into the training (90%) and validation (10%) sets. The three machine learning models are developed using the the 90% training data and are tested using the 10% independent validation dataset. The independent validation dataset roughly corresponds to the last 2 years of the data measured from 1995 to 2014 in the sense that the dataset from 1995 to 2012 is used to develop the machine learning models and predictions are made for the 2013 and 2014 pollen seasons. The results are shown in Fig. 3. Panels (a), (b), and (c)

Fig. 2 Actual pollen data observed from 1995 to 2014 for the peak *Ambrosia* pollen season at Tulsa, OK



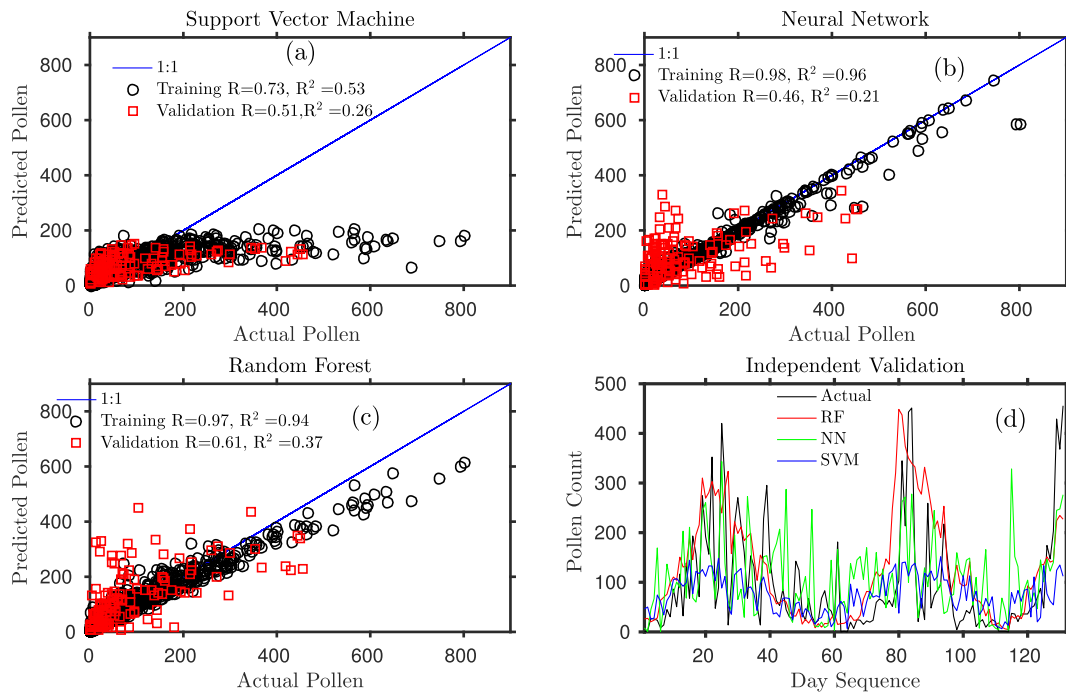


Fig. 3 Scatter plots of actual and predicted pollen for the support vector machine, SVM (panel a), neural network, NN (panel b), random forest, RF (panel c). Panel (d) Plots of independent

validation results using the random forest (red line), neural network (green line), and support vector machine (blue line). The actual pollen shown by the black curve

in Fig. 3 show scatter plots of prediction made by the support vector machine, neural network, and random forest machine learning methods, respectively, using the training data (black circles) and the independent validation data (red squares).

From the top three panels of Fig. 3, we observe that the neural network and random forest methods produced better predictions than the support vector machine. The random forest method produced the best independent validation results ($R=0.61$, $R^2=0.37$) of all the three methods. The high correlation value of neural network found using the training data ($R=0.98$, $R^2=0.96$) is not reproduced in the independent validation test which had an R and R^2 values of only 0.46 and 0.21, respectively.

Panel (d) in Fig. 3 shows comparisons of the predicted pollen using the regression models developed by the training dataset for the three methods. The validation predictions are made using the 10% of the predictors data (test set) that is not employed to develop the model. It consists of about 130 days (roughly corresponding to the 2013 and 2014 pollen seasons) of predictors and target data we withhold before training the model. The black curve shows the actual pollen

data for those number of days and the other curves show predictions made by the random forest (red curve), neural network (green curve), and support vector machines (blue color). The results indicate superior performance of the random forest method followed by the neural network and support vector machines. However, a technique that combines the three methods together is expected to show a robust performance as indicated by Voukantsis et al. (2010).

Another important application of machine learning methods is the selection of the best features (variables) that contribute most to the prediction and ranking them in order of their importance. In this way, we can determine the most important predictor variables and estimate the output leaving features that contribute less. The random forest provides such a ranking based on criteria attributed to the splitting variable in the data sampling to form decision trees (Genuer et al. 2010; Kotsiantis 2007; Friedman et al. 2001).

In addition to the random forest method of variable ranking, we used the correlation coefficient and interaction information methods to sort the input variables in order of their importance. The correlation coefficient method sorts based on the relation between the

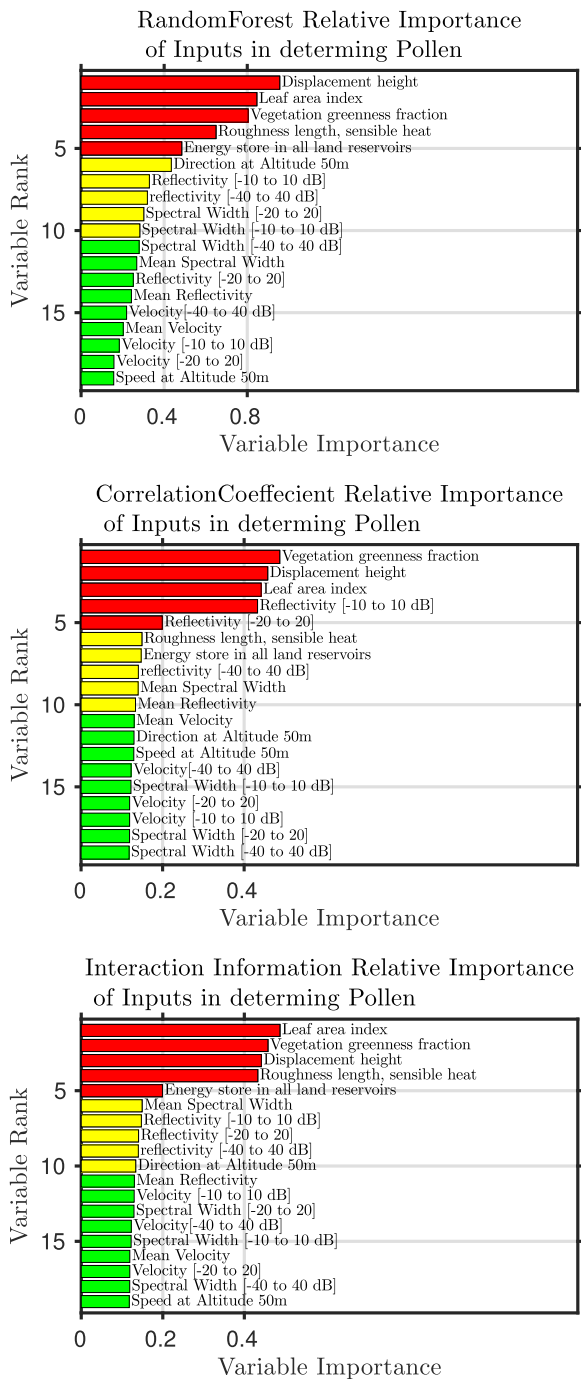


Fig. 4 Variable importance sorted using the random forest (top panel), correlation coefficient (middle panel) and interaction information (bottom panel) methods

predictors and the pollen data, whereas the interaction information method implements a generalization of the mutual information technique (Darbellay et al.

1999) by calculating for each predictor (and sorting based on) a value called information gain (Brown 2009). The results of variable importance selection is given by Fig. 4. The top panel in Fig. 4 shows the variable sorted using the random forest machine learning method. The middle and bottom panels show the rank of variable importance sorted using correlation coefficient and interaction information values, respectively. From the three methods, we found that the environmental parameters—leaf area index, vegetation greenness function, and displacement height—took the top rank and from the NEXRAD predictors, the mean reflectivity in signal to noise ratio from -10 to 10 dB and from -20 to 20 dB constitute among the top predictors as seen in the random forest and correlation coefficient methods. Additionally, the direction of the wind measured by the NEXRAD at the lowest altitude (about 50 m) from the Earth’s surface is the top predictor following the environmental variables. These agree with the finding of Palacios et al. (2000) and Rojo et al. (2015), which showed the direction of the wind highly influences the concentration of pollen in the surrounding.

Discussion

This study employs advanced machine learning methods (random forest, neural network, and support vector machine) regression to predict daily *Ambrosia* pollen concentration at Tulsa, OK (location, 36.1511° N, 95.9446° W). In these advanced machine learning methods, we used a combination of environmental parameters and NEXRAD radar measurements as predictors. The combined parameters are listed in Table 1. Successful application of advanced machine learning methods and meteorologic variables measured in highly allergic pollen polluted areas would help to predict and notify the public in advance. This will help allergic susceptible individuals and health workers to take the necessary precaution.

Previously, the support vector machine, neural network, and random forest machine learning methods are rarely applied for pollen prediction. Over the past decade, the neural network has been applied to study pollen of different species over the European region. For example, Csépe et al. (2014) used different computational intelligence (CI) methods to predict the *Ambrosia* pollen at two different places in Hungary

and France. Castellano-Méndez et al. (2005) and Puc (2012) have employed the neural network to predict *Betula* pollen over Spain and Poland, respectively. Recently, Nowosad (2015) used the random forest method to forecast different tree pollen species.

Of all the three machine learning methods, we found that the random forest method produced better performance ($R = 0.61$, $R^2=0.37$) when tested with independent dataset that is not used to develop the model. The neural network contrarily produced lower correlation when tested with our independent test data ($R = 0.46$, $R^2=0.21$) as shown in Fig. 3b despite its high correlation ($R = 0.98$, $R^2=0.96$) when tested using the training data. The discrepancies can be explained in terms of the robustness of the random forest against overfitting (Breiman 2001; Liaw et al. 2002). However, another version of neural network, the multi-layer perceptron, has been applied by Csépe et al. (2014) to forecast *Ambrosia* pollen in different locations in Europe and has produced robust results compared to other tree-based methods. The support vector machine produced competitive performance but outperformed by the random forest and neural network machine learning methods. This agrees with the finding of Meyer et al. (2003) who compared the support vector machine with other methods including the neural network and random forest.

Most pollen forecasting studies applied environmental parameters as input parameters. For example, Howard and Levetin (2014) used minimum temperature, precipitation, dew point, and phenology as predictors. Csépe et al. (2014) used a total of eight meteorological parameters and different computational intelligence methods to predict the concentration of *Ambrosia* pollen and alarm levels for the future 7 days at two locations in Europe. Our variable importance and ranking using the random forest, correlation coefficient, and interaction information methods show the dominance of these environmental parameters. Among the NEXRAD parameters, the reflectivity and direction of wind are among the top predictors. However, using only environmental parameters alone can affect the spatial resolution of the pollen forecasting region (Prank et al. 2013).

This research applies the NEXRAD weather measurements to forecast allergic pollen for the first time. The NEXRAD has large spatial coverage (Maddox et al. 2002). The use of only NEXRAD measurements and robust machine learning method would lay the

foundation to forecast allergic pollen at a fine spatial scale over the USA (Zewdie et al. 2019).

Conclusion

In this paper, we implemented advanced supervised machine learning methods—random forest, neural network, and support vector machine—to predict the daily *Ambrosia* pollen in the atmosphere of Tulsa. To supervise the learning process, we used pollen data measured using the Burkard's pollen trap apparatus at the University of Tulsa, OK.

We use a combination of environmental parameters and NEXRAD measurements as predictors. We implemented the random forest, interaction information, and correlation coefficient methods to rank these variables in their rank of importance. We observe that the most useful parameters in estimating *Ambrosia* pollen were displacement height, leaf area index, vegetation greenness fraction, and NEXRAD measurements of reflectivity at low signal to noise ratio and direction of wind. These parameters stand out as top predictors in the measure of variable importance.

Among the three machine learning methods, the random forest showed superior performance and also provided a ranked list of the relative importance of the input variables. The neural network and support vector machine methods also provided comparative prediction using independent data.

References

- Andrews, C.P., Ratner, P.H., Ehler, B.R., Brooks, E.G., Pollock, B.H., Ramirez, D.A., Jacobs, R.L. (2013). The mountain cedar model in clinical trials of seasonal allergic rhinoconjunctivitis. *Annals of Allergy, Asthma & Immunology*, 111(1), 9–13.
- Arizmendi, C., Sanchez, J., Ramos, N., Ramos, G. (1993). Time series predictions with neural nets: application to airborne pollen forecasting. *International Journal of Biometeorology*, 37(3), 139–144.
- Basak, D., Pal, S., Patranabis, D.C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10), 203–224.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Britton, J., Pavord, I., Richards, K., Knox, A., Wisniewski, A., Wahedna, I., Kinnear, W., Tattersfield, A., Weiss, S. (1994). Factors influencing the occurrence of airway hyperreactivity in the general population: the importance of atopy and

- airway calibre. *European Respiratory Journal*, 7(5), 881–887.
- Brown, G. (2009). A new perspective for information theoretic feature selection. In: AISTATS, pages 49–56.
- Castellano-Méndez, M., Aira, M., Iglesias, I., Jato, V., González-Manteiga, W. (2005). Artificial neural networks as a useful tool to predict the risk level of *Betula* pollen in the air. *International Journal of Biometeorology*, 49(5), 310–316.
- Csépe, Z., Makra, L., Voukantsis, D., Matyasovszky, I., Tusnády, G., Karatzas, K., Thibaudon, M. (2014). Predicting daily ragweed pollen concentrations using computational intelligence techniques over two heavily polluted areas in Europe. *Science of the Total Environment*, 476, 542–552.
- D'amato, G., & Spiekma, F.T.M. (1991). Allergenic pollen in Europe. *Grana*, 30(1), 67–70.
- Darbellay, G.A., Vajda, I., et al. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4), 1315–1321.
- Doviak, R.J., & Zrnic, D.S. (2014). Doppler radar & weather observations. Academic Press.
- Ernst, P., Ghezzi, H., Becklake, M. (2002). Risk factors for bronchial hyperresponsiveness in late childhood and early adolescence. *European Respiratory Journal*, 20(3), 635–639.
- Esch, R.E., Hartsell, C.J., Crenshaw, R., Jacobson, R.S. (2001). Common allergenic pollens, fungi, animals, and arthropods. *Clinical Reviews in Allergy and Immunology*, 21(2), 261–292.
- Friedman, J., Hastie, T., Tibshirani, R. (2001). *The elements of statistical learning* Vol. 1. Berlin: Springer series in statistics Springer.
- Gali, R.K. (2010). *Assessment of NEXRAD P3 data on stream-flow simulation using SWAT for North Fork Ninescaw watershed, Kansas*. PhD thesis: Kansas State University.
- Gardner, M.W., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14), 2627–2636.
- Genuer, R., Poggi, J.-M., Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14), 2225–2236.
- Guzella, T.S., & Caminhas, W.M. (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7), 10206–10222.
- Hannesen, R., & Weipert, A. (2003). *Detection of dust storms with a C-band doppler radar*. Germany: AMS-Gematronik.
- Haykin, S. (1994). *Neural networks: a comprehensive foundation*. New York: Macmillan College Publishing Company.
- Haykin, S.S., et al. (2001). Kalman filtering and neural networks. Wiley Online Library.
- Haykin, S.S., Haykin, S.S., Haykin, S.S., Haykin, S.S. (2009). *Neural networks and learning machines*, volume 3. Pearson Upper Saddle River, NJ, USA.
- Hirst, J. (1952). An automatic volumetric spore trap. *Annals of Applied Biology*, 39(2), 257–265.
- Howard, L.E., & Levetin, E. (2014). Ambrosia pollen in Tulsa, Oklahoma: aerobiology, trends, and forecasting model development. *Annals of Allergy, Asthma & Immunology*, 113(6), 641–646.
- Kasprzyk, I. (2008). Non-native Ambrosia pollen in the atmosphere of rzeszów (SE Poland); evaluation of the effect of weather conditions on daily concentrations and starting dates of the pollen season. *International Journal of Biometeorology*, 52(5), 341–351.
- Kinney, P.L. (2008). Climate change, air quality, and human health. *American Journal of Preventive Medicine*, 35(5), 459–467.
- Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, vol. 14, pp. 1137–1145. Stanford, CA.
- Kotsiantis, S. (2007). Supervised machine learning: a review of classification techniques. *Informatica*, 31, 249–268.
- Laaidi, M., Laaidi, K., Besancenot, J.-P., Thibaudon, M. (2003). Ragweed in France: an invasive plant and its allergenic pollen. *Annals of Allergy, Asthma & Immunology*, 91(2), 195–201.
- Lary, D.J. (2010). Artificial intelligence in geoscience and remote sensing. INTECH Open Access Publisher.
- Lary, D.J., Zewdie, G.K., Liu, X., Wu, D., Levetin, E., Allee, R.J., Malakar, N., Walker, A., Mussa, H., Mannino, A., et al. (2018). Machine learning applications for earth observation. In: Earth observation open science and innovation, pp. 165–218. Springer.
- Lewis, W.H., Vinay, P., Zenger, V.E. (1983). Airborne and allergenic pollen of North America. Johns Hopkins University Press.
- Liaw, A., Wiener, M., et al. (2002). Classification and regression by random forest. *R news*, 2(3), 18–22.
- Liu, X., Wu, D., Zewdie, G.K., Wijerante, L., Timms, C.I., Riley, A., Levetin, E., Lary, D.J. (2017). Using machine learning to estimate atmospheric ambrosia pollen concentrations in Tulsa, OK. *Environmental Health Insights*, 11, 1178630217699399.
- Low, R.B., Bielory, L., Qureshi, A.I., Dunn, V., Stuhlmiller, D.F., Dickey, D.A. (2006). The relation of stroke admissions to recent weather, airborne allergens, air pollution, seasons, upper respiratory infections, and asthma incidence, September 11, 2001, and day of the week. *Stroke*, 37(4), 951–957.
- Maddox, R.A., Zhang, J., Gourley, J.J., Howard, K.W. (2002). Weather radar coverage over the contiguous United States. *Weather and Forecasting*, 17(4), 927–934.
- Matheson, E.M., Player, M.S., Mainous, A.G., King, D.E., Everett, C.J. (2008). The association between hay fever and stroke in a cohort of middle aged and elderly adults. *The Journal of the American Board of Family Medicine*, 21(3), 179–183.
- Meyer, D., Leisch, F., Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, 55(1), 169–186.
- Nowosad, J. (2015). Spatiotemporal models for predicting high pollen concentration level of *Corylus*, *Alnus*, and *Betula*. *International Journal of Biometeorology*, pp 1–13.
- Osowski, S., & Garanty, K. (2007). Forecasting of the daily meteorological pollution using wavelets and support vector machine. *Engineering Applications of Artificial Intelligence*, 20(6), 745–755.

- Oswalt, M.L., & Marshall, G.D. (2008). Ragweed as an example of worldwide allergen expansion. *Allergy, Asthma & Clinical Immunology*, 4(3), 1.
- Palacios, I.S., Molina, R.T., Rodríguez, A.M. (2000). Influence of wind direction on pollen concentration in the atmosphere. *International Journal of Biometeorology*, 44(3), 128–133.
- Postolache, T., Stiller, J., Herrell, R., Goldstein, M., Shree-ram, S., Zebrak, R., Thrower, C., Volkov, J., No, M., Volkov, I., et al. (2005). Tree pollen peaks are associated with increased nonviolent suicide in women. *Molecular psychiatry*, 10(3), 232–235.
- Prank, M., Chapman, D.S., Bullock, J.M., Belmonte, J., Berger, U., Dahl, A., Jäger, S., Kovtunen, I., Magyar, D., Niemelä, S., et al. (2013). An operational model for forecasting ragweed pollen release and dispersion in Europe. *Agricultural and forest meteorology*, 182, 43–53.
- Prybutok, V.R., Yi, J., Mitchell, D. (2000). Comparison of neural network models with ARIMA and regression models for prediction of houston's daily maximum ozone concentrations. *European Journal of Operational Research*, 122(1), 31–40.
- Puc, M. (2012). Artificial neural network model of the relationship between Betula pollen and meteorological factors in Szczecin (Poland). *International Journal of Biometeorology*, 56(2), 395–401.
- Ramirez, D.A. (1984). The natural history of mountain cedar pollinosis. *Journal of allergy and clinical immunology*, 73(1), 88–93.
- Rodríguez-Rajo, F., Astray, G., Ferreira-Lage, J., Aira, M., Jato-Rodríguez, M., Mejuto, J.C. (2010). Evaluation of atmospheric Poaceae pollen concentration using a neural network applied to a coastal Atlantic climate region. *Neural Networks*, 23(3), 419–425.
- Rojo, J., Rapp, A., Lara, B., Fernández-González, F., Pérez-Badia, R. (2015). Effect of land uses and wind direction on the contribution of local sources to airborne pollen. *Science of the Total Environment*, 538, 672–682.
- Sánchez-Mesa, J., Galán, C., Martínez-Heras, J., Hervás-Martínez, C. (2002). The use of a neural network to forecast daily grass pollen concentration in a Mediterranean region: the southern part of the Iberian Peninsula. *Clinical & Experimental Allergy*, 32(11), 1606–1612.
- Sassen, K. (2008). Boreal tree pollen sensed by polarization lidar: depolarizing biogenic chaff. *Geophysical Research Letters*, 35(18), L18810.
- Smola, A.J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222.
- Specht, D.F. (1991). A general regression neural network. *IEEE transactions on neural networks*, 2(6), 568–576.
- Stark, P.C., Ryan, L.M., McDonald, J.L., Burge, H.A. (1997). Using meteorologic data to predict daily ragweed pollen levels. *Aerobiologia*, 13(3), 177–184.
- Stickley, A., Ng, C.F.S., Konishi, S., Koyanagi, A., Watanabe, C. (2017). Airborne pollen and suicide mortality in Tokyo, 2001–2011. *Environmental research*, 155, 134–140.
- Tränkle, E., & Mielke, B. (1994). Simulation and analysis of pollen coronas. *Applied Optics*, 33(21), 4552–4562.
- Vapnik, V. (2013). The nature of statistical learning theory. Springer science & Business Media.
- Vapnik, V.N., & Vapnik, V. (1998). *Statistical learning theory* Vol. 1. New York: Wiley.
- Verikas, A., Gelzinis, A., Bacauskiene, M. (2011). Mining data with random forests: a survey and results of new tests. *Pattern Recognition*, 44(2), 330–349.
- Vivekanandan, J., Ellis, S., Oye, R., Zrnica, D., Ryzhkov, A., Straka, J. (1999). Cloud microphysics retrieval using S-band dual-polarization radar measurements. *Bulletin of the American Meteorological Society*, 80(3), 381–388.
- Voukantsis, D., Niska, H., Karatzas, K., Riga, M., Damialis, A., Vokou, D. (2010). Forecasting daily pollen concentrations using data-driven modeling methods in Thessaloniki, Greece. *Atmospheric Environment*, 44(39), 5101–5111.
- Wayne, P., Foster, S., Connolly, J., Bazzaz, F., Epstein, P. (2002). Production of allergenic pollen by ragweed (*Ambrosia artemisiifolia* L.) is increased in CO₂-enriched atmospheres. *Annals of Allergy, Asthma & Immunology*, 88(3), 279–282.
- Wilson, J.W., Weckwerth, T.M., Vivekanandan, J., Wakimoto, R.M., Russell, R.W. (1994). Boundary layer clear-air radar echoes: origin of echoes and accuracy of derived winds. *Journal of Atmospheric and Oceanic Technology*, 11(5), 1184–1206.
- Witten, I.H., & Frank, E. (2005). *Data Mining: practical machine learning tools and techniques*. Morgan Kaufmann.
- Yi, J., & Prybutok, V.R. (1996). A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution*, 92(3), 349–357.
- Zewdie, G.K., Lary, D.J., Liu, X., Wu, D., Levetin, E. (2019). Estimating the daily pollen concentration in the atmosphere using machine learning and nexrad weather radar data. *Environmental Monitoring and Assessment*.
- Zhao, F., Elkelish, A., Durner, J., Lindermayr, C., Winkler, J.B., Ruff, F., Behrendt, H., Traidl-Hoffmann, C., Holzinger, A., Kofler, W., et al. (2016). Common ragweed (*Ambrosia artemisiifolia* L.): allergenicity and molecular characterization of pollen after plant exposure to elevated NO₂. *Plant, Cell & Environment*, 39(1), 147–164.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.