



Estimating the daily pollen concentration in the atmosphere using machine learning and NEXRAD weather radar data

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

Received: 13 April 2017 / Accepted: 6 November 2017
© Springer Nature Switzerland AG 2019

Abstract Millions of people have an allergic reaction to pollen. The impact of pollen allergies is on the rise due to increased pollen levels caused by global warming and the spread of highly invasive weeds. The production, release, and dispersal of pollen depend on the ambient weather conditions. The temperature, rainfall, humidity, cloud cover, and wind are known to affect the amount of pollen in the atmosphere. In the past, various regression techniques have been applied to estimate and forecast the daily pollen concentration in the atmosphere based on the weather conditions. In this research, machine learning methods were applied to the Next Generation Weather Radar (NEXRAD) data to estimate the daily *Ambrosia* pollen over a 300 km × 300 km region centered on a NEXRAD weather radar. The Neural Network and Random Forest machine learning methods have been employed to develop separate models to estimate *Ambrosia* pollen over the region. A feasible way of estimating the daily pollen concentration using only the NEXRAD radar

data and machine learning methods would lay the foundation to forecast daily pollen at a fine spatial resolution nationally.

Keywords Pollen · NEXRAD · Weather · Machine learning · Neural network · Random forest · Environmental public health

Introduction

Plants producing large amount of pollen such as *Ambrosia artemisiifolia* (common ragweed) are becoming invasive (Laaïdi et al. 2003). Consequently, seasonal allergies due to pollen are widespread. Large numbers of people in the northern temperate regions, such as North America and Europe, are affected by pollen. For example, about 50 million adult Americans, and nearly one-tenth of children below the age of 18, have experienced some form of allergic reaction to pollen (Howard and Levetin 2014). Similarly, in Europe, it is believed that more than 30 million people are allergic to ragweed pollen, and the number of affected people is expected to double by 2041–2060 (Lake et al. 2016). This is a sharp increase compared with the 15 million people affected by pollen in 1991 (D’Amato and Spieksma 1991).

The recently high impact of allergic pollen is associated with global climate change. Climate change affects the growing region, the amount of pollen in the atmosphere, and the onset and end dates of a

G. K. Zewdie (✉) · D. J. Lary · X. Liu · D. Wu
William B. Hanson Center for Space Sciences, The
University of Texas at Dallas, Richardson, TX 75080, USA
e-mail: gkz130030@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

pollen season (Kinney 2008; Voukantsis et al. 2010; D'Amato et al. 2011). Other major air pollutants such as CO₂ and NO₂ are also known to affect the abundance and allergic potency of pollen (Wayne et al. 2002; Zhao et al. 2016).

Weather conditions are important for the production, release, and dispersal of pollen from the source plants. Previous studies have shown that atmospheric parameters such as temperature (mean, minimum, maximum), wind speed and direction, humidity, cloud coverage, and precipitation are among the top environmental predictors of pollen (Howard and Levetin 2014; Stark et al. 1997; Molina et al. 2001; Kasprzyk 2008).

These environmental variables can be used together with advanced mathematical regression and machine learning methods to estimate the daily pollen concentration in the atmosphere. Previous studies used less advanced regression techniques such as linear regression and the polynomial fits to these environmental variables to estimate the airborne pollen abundance (Sánchez-Mesa et al. 2002). It has been found that using the NEXRAD weather measurements with machine learning also provides the capability to estimate pollen over a large spatial scale and open up the opportunity for providing pollen alerts on a much finer spatial scale than previously possible.

The NEXRAD radar network measures the direction and speed of the wind, both directly associated with the release and dispersal of pollen. High radar reflectivity values are associated with precipitation, rainfall, and cloud coverage which are controlling variables for pollen production and release. Low radar reflectivity values are characteristic of clear skies. Spectral width measurements of the weather radar indicate the presence/absence of thunderstorms. Thunderstorms are associated with high asthma outbreaks, especially during the pollen season (D'Amato et al. 2007).

In this paper, a pollen estimation capability is presented using only the NEXRAD radar measurements. The paper is a continuation of the first article Zewdie et al. (in press 2019) which used environmental and NEXRAD radar measurements to estimate pollen. In this study, we use both the Random Forest and Neural Network machine learning methods to develop two different models to estimate pollen using only the NEXRAD radar measurements. The models are developed employing the NEXRAD measurements at the

two lowest elevations plus the wind vectors at a square area of 10 km × 10 km centered at the pollen observation site. The models are then applied over a 300 km by 300 km region at a resolution of 0.5 km × 0.5 km (excluding pixels having no radar measurements).

Materials and methods

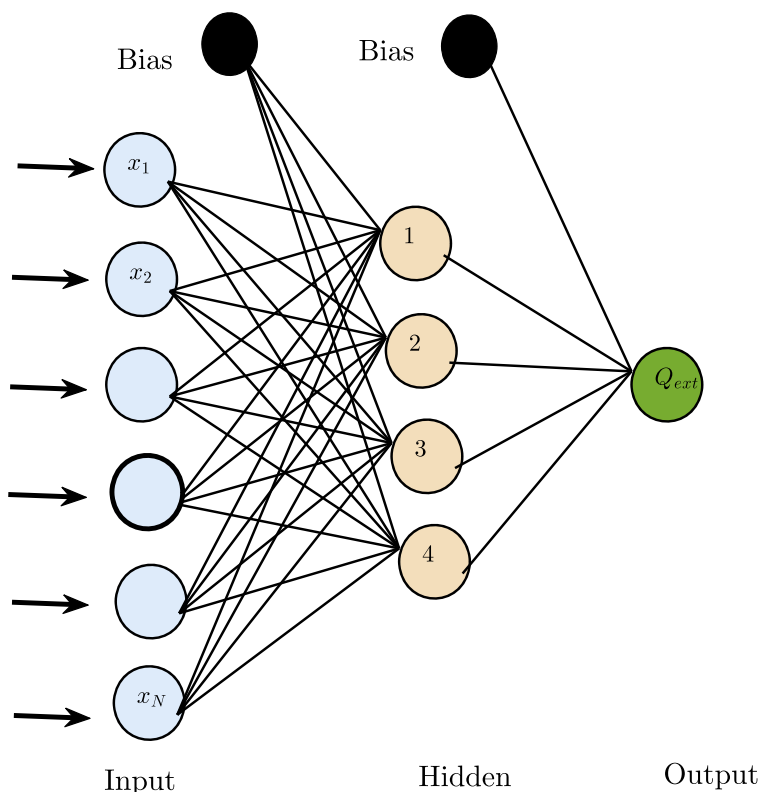
Machine learning

Machine learning employs mathematical and statistical approaches that 'learn' by example from a training data set. Ideally, the training dataset should be a large sample of data containing as many relevant predictors as possible. Machine learning is suitable for problems in which we do not have the exact theoretical functional relation between the output parameter(s) one would like to estimate and the available set of predictors. In this case, we use the machine learning to 'learn' from the data and to develop an empirical model. The training data is randomly partitioned using the k-fold method (Kohavi 1995) into a portion used for training the machine learning models and a portion used as an independent validation dataset to test the performance of our model.

Previously, a few studies have employed advanced machine learning methods to estimate pollen using environmental parameters as predictors (Csépe et al. 2014; Rodríguez-Rajo et al. 2010; Arizmendi et al. 1993; Sánchez-Mesa et al. 2002). Most of them involve ground based meteorological parameters such as mean, minimum, maximum and range temperature, relative humidity, rainfall, wind speed, and air pressure as input predictors. Few studies (Lary et al., 2018; Liu et al., 2017; Nowosad, 2016) used climate model data as predictors. In this study, we used only the NEXRAD radar measurements of reflectivity, spectral width, and Doppler velocity at the lowest elevations as well as the direction and speed of wind as predictors for the machine learning to estimate daily *Ambrosia* pollen.

Here, two popular machine learning methods have been used, Neural Networks and Random Forests. The Neural Network machine learning algorithm is a multivariate nonparametric approach that mimics the interconnection and information flow of the neuron's in the human brain (Lary 2010; Haykin 1994; Haykin 2009, 2001). Information is passed to a receiving

Fig. 1 Example of a Neural Network architecture with N input variables, four hidden neurons and one output layer



neuron as a weighted sum of the inputs. The weights are optimized for the synaptic learning. The receiving neuron computes the activation function, f , as shown in Eq. 2 where b represents the bias and w_{ij} is the weight from the i th to the j th neuron. The training is iterative and both the biases and weights start out as random numbers. Figure 1 shows the Neural Network architecture, with input, hidden, and output layers.

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

Random Forests are an effective machine learning approach for both classification and regression (Breiman 2001). Random Forest works on the basis of random subspaces and bagging (Verikas et al. 2011). An ensemble of trees (Fig. 2) are grown by taking a random sample from the training data (Breiman 2001; Biau et al. 2008). At each node along the tree, a random subset of input parameters are selected to find the best split among those predictors. After the trees are grown to the maximum depth, they vote (for classification) and average (for regression) to obtain prediction for the new sample. Predictions are made by

aggregating predictions of all the random subspaces, i.e., averaging for regression and majority vote for the case of classification (Liaw and Wiener 2002).

The Random Forest can highlight the most important variables by estimating the increase in the out-of-bag (OOB) error when the variable is permuted while fixing the other variables (Liaw and Wiener 2002).

NEXRAD data

The NEXRAD radar network includes about 160 interconnected weather radars sampling the atmosphere in three dimensions. This information makes a significant contribution to modern weather forecasting. The primary products are the reflectivity, providing the amount of scattering from different sources ranging from large aerosols to precipitation and cloud coverage. The spectral width, providing the turbulent state of the scatterers showing their velocity field. The Doppler velocity, providing the line of sight speed of the scatterers, which in most cases is the wind field.

Recently, the radar networks have been upgraded to double (the horizontal and vertical) polarization signals making them suitable to derive three additional

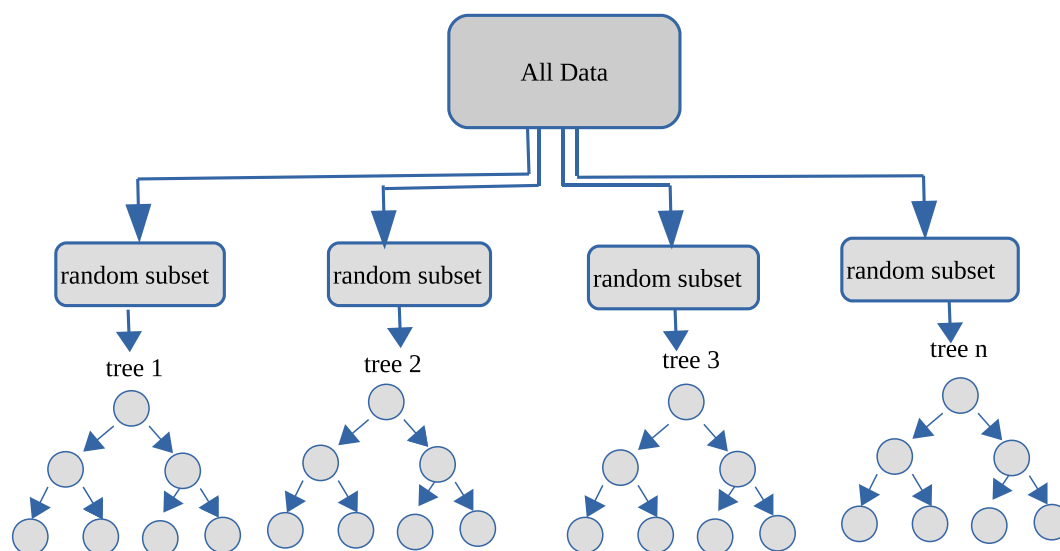


Fig. 2 Showing the Random Forest general structure consisting of n trees

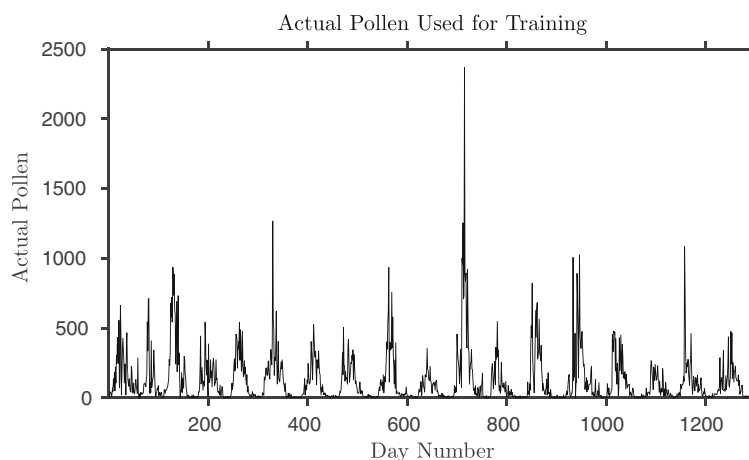
measurements, the polarimetric parameters. The polarimetric measurements are useful for aerosol size and shape identification thereby inferring the type of aerosol. The polarimetric measurements include the differential reflectivity, differential phase, and correlation coefficient. They are derived based on the horizontal and vertical polarization signals. Due to limited availability of the newer polarimetric and large data requirement of the machine learning, use of the polarimetric data is not within the scope of this study.

Pollen observation

Observations of pollen at The University of Tulsa (geographic location: 36.1511° N, 95.9446° W) have

been made on a routine basis since 1987. Detailed description of the pollen data used to supervise the Neural Network and Random Forest machine learning methods is presented by Howard and Levetin (2014). The Burkhard trap apparatus was placed at the top of a building which is 12 m from the ground. Inside the Burkhard trap, a sticky plate is attached to a drum that rotates at 2 mm per h in order to let in air through a small orifice. Pollen entering through the orifice will become affixed to a sticky tape. At the end of the week of sampling, the tape is cut into daily lengths (48 mm) and the sample carefully observed under a microscope at a magnification of $\times 400$ after applying a solution of glycerin jelly and fuchsin stain.

Fig. 3 Showing the actual *Ambrosia* pollen data used to train the Neural Network and Random Forest machine learning methods



For this study, the NEXRAD radar located at 36.1750° N, 95.5644° W which is about 34 km from the University of Tulsa is used to derive the radar measurements. NEXRAD measurements are relatively fully accessible since 1995. Consequently, we have used the pollen observations made from 1995 to 2014 to train our machine learning algorithms. The data consisting of 8 predictors, all of which are NEXRAD measurements of reflectivity, line of sight Doppler velocity, and spectral width at the lower two elevations plus the direction and speed of wind near the surface were combined with the time series of pollen count data and then separated into training set and test set. The ratio of the training and test set is roughly 90 to 10%.

Figure 3 shows the actual pollen data used to supervise the machine learning. A preliminary analysis shows that the Neural Network and Support Vector machine learning methods work better when the response data and the predictors are normalized. Consequently, we used a simple normalization technique

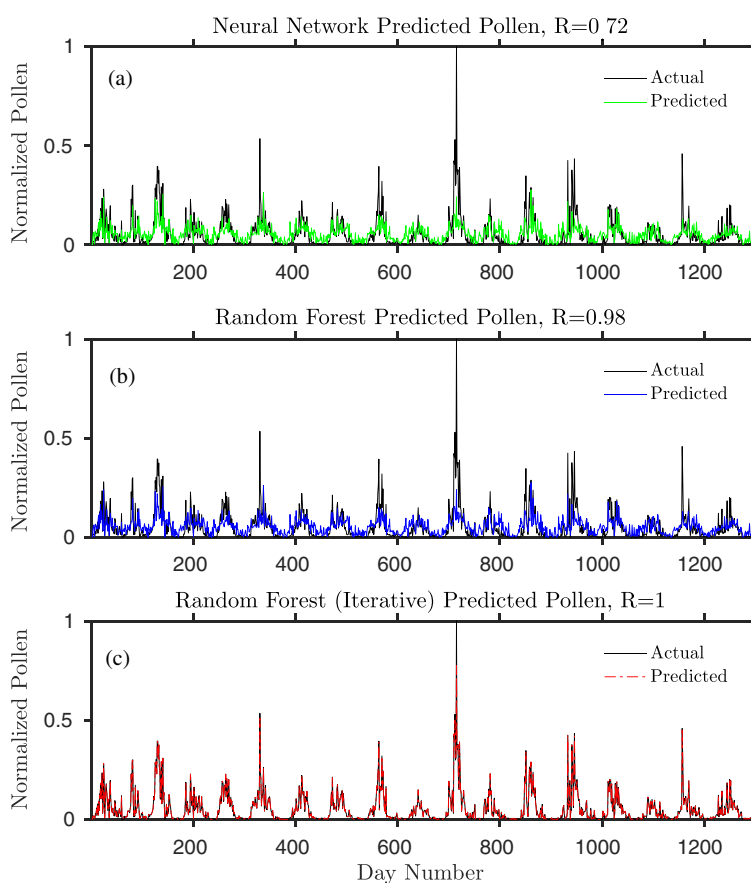
by dividing each parameter by its maximum value. However, we observed that the Random Forest did not require any data normalization.

Results and discussion

After partitioning using the k-fold random sampling, the training data is then used to develop the model for the Neural Network and Random Forest methods. The Random Forest model is developed using an ensemble of 200 decision trees. We observed the out-of-bag error converges to its minimum after around 100 trees. The feed-forward architecture with 20 hidden number of neurons and the Levenberg-Marquardt training function is used for Neural Network machine learning method (Hagan and Menhaj 1994).

Figure 4 shows the predictions made by the Neural Network, Random Forest, and iterative Random Forest methods using the same data used to develop the model. Panel a shows a prediction made by the

Fig. 4 Estimating the daily Ambrosia pollen using the training data to predict the normalized pollen. Panels **a**, **b** and **c** present the predicted pollen using, respectively, the Neural Network, Random Forest and iterative Random Forest machine learning methods. The actual pollen data is normalized before using for the training



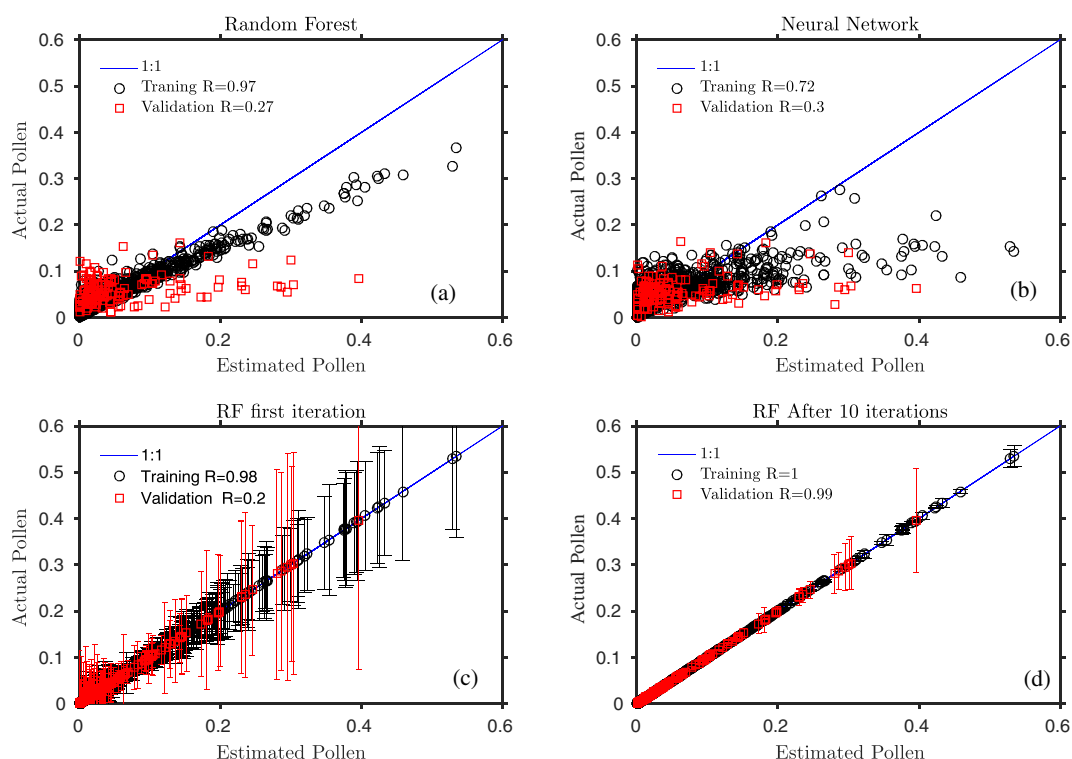


Fig. 5 Scatter diagrams showing the performance of the various machine learning algorithms. Results of the Random Forest (RF) and Neural Network (NN) are shown, respectively, in

panels (a) and (b). The iterative method applied to the Random Forest is given by panel (c)

feed-forward Neural Network, panel b gives predictions using the Random Forest, and panel c presents predictions of the iterative Random Forest after about 10 iterations. In all cases, the black curve shows the actual pollen used to supervise the training and the green, red, and blue curves are predictions using the input parameters used to train the model. The high correlation coefficients between the actual and predicted pollen are expected for the training data set.

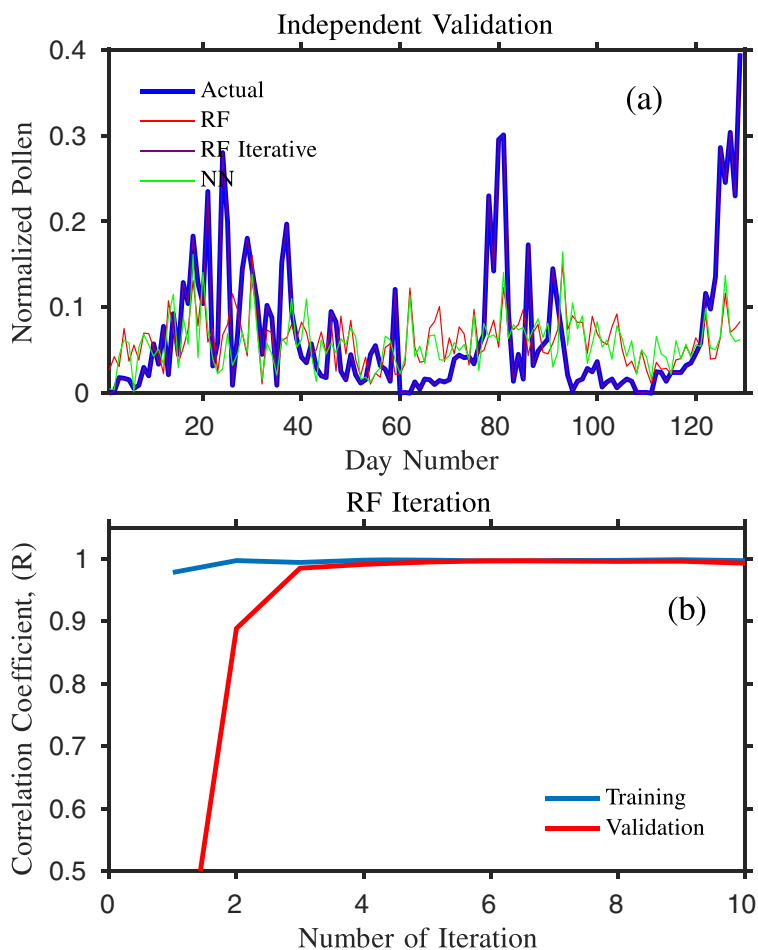
Scatter plots of results of the model estimates using the 10% data withheld for independent validation are shown in Fig. 5. In this figure, the validation data (shown as red squares) is totally independent and was not used to develop the models. Panel a presents results from the Random Forest for the training data (black circles) and independent validation data (red squares). In this case, the Random Forest is used without incorporating the error corrections in the iterative training. Panel b shows results obtained using the feed-forward Neural Network for the training data (black circles) and validation data (red squares). From panel a and panel b in Fig. 5, we observe that both the

Random Forest and Neural Network machine learning methods produced comparable results with the Neural Network producing slightly better results than the Random Forest in estimating the pollen data.

Panels c and d in Fig. 5 show error plots of the Random Forest at the first and 10th iteration respectively for the training and independent validation data. The Random Forest is trained with 200 trees. We observe that the error values are highly reduced and the correlation coefficient converges towards one as the number of iterations increases. After a few iterations, the iterative Random Forest significantly improved the results for the cross validation dataset.

Panel a in Fig. 6 shows the output of the model using the 10% independent validation data. In this case, the estimated pollen is compared with the actual pollen retained for independent cross validation. Both the Neural Network and Random Forest models produced comparable results. However, the application of the Random Forest to estimate the error iteratively has shown superior performance. Plots of the correlation coefficient between actual pollen and estimated pollen

Fig. 6 Panel **a** a time-series plot showing the independent validation for the Random Forest, Neural Network and the iterative improvement R andom Forest. Panel **b** shows the plot of the correlation coefficient as a function of the number of iterations



for a few iterative Random Forest for both the training and cross validation data are shown in Fig. 6 panel b. We see that after a few iterations, the correlation coefficient value converges towards 1 for both the training and validation datasets.

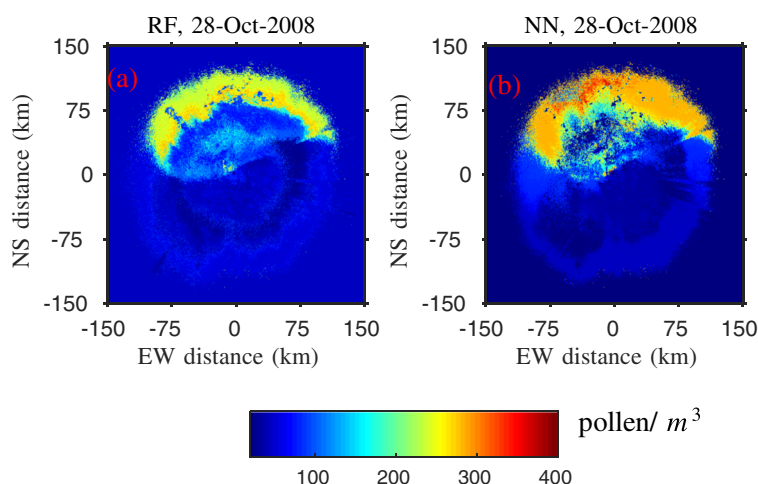
Neural Network and Random Forest machine learning models were developed using the NEXRAD weather radar data at a $10 \text{ km} \times 10 \text{ km}$ pixel centered at the pollen collection site. The eight predictor variables presented in “NEXRAD data” and the pollen measurement are used to train the models. The models are then applied to all available pixels over the $300 \text{ km} \times 300 \text{ km}$ region centered on the NEXRAD weather radar.

Figure 7 shows example results of the application of the Neural Network and iterative Random Forest for all pixels over the $300 \text{ km} \times 300 \text{ km}$ region for 28 October 2008. Panels a and b, respectively, show results of the Random Forest and Neural Network

machine learning methods applied to all pixels. The approaches used by the Random Forest and Neural Network are completely independent. The Random Forest uses the principle of random sampling of the data to grow forest of trees for bagging (Breiman 2001; Verikas et al. 2011) whereas the Neural Network works based on the analogy of the connection and information flow in the human brain. Despite very different approaches used by the Neural Networks and Random Forest, the two machine learning methods highlighted high levels of pollen in approximately the same regions.

In Fig. 7, we observed interesting pollen distribution structures that presumably are associated with the wind direction and speed. Further analysis of the distribution of the ragweed plant over the region in relation to the direction and speed of the wind is needed to describe the pollen structures shown in Fig. 7.

Fig. 7 Map showing the pollen distribution using the Random Forest (panel **a**) and Neural Network (panel **b**) machine learning methods for 28 October 2008.



Conclusion

In this paper, we used the NEXRAD radar measurements of reflectivity, line of sight Doppler velocity, spectral width at the lowest two elevations, direction and speed of wind, and machine learning to estimate the daily *Ambrosia* pollen at the University of Tulsa, Oklahoma. Random Forest and Neural Network machine learning models were developed for a $10 \text{ km} \times 10 \text{ km}$ region. The models are developed using the training data and validated by an independent data obtained using the k-fold random partitioning technique. The models were applied to both the training dataset and the independent validation dataset. The validation results using the training data back into the model show that both the Neural Network and Random Forest models produced results having a high correlation coefficient. The model performance using the independent validation dataset was good. We also applied an iterative approach where we performed error estimation and error correction on our Random Forest estimates.

The model developed over the $10 \text{ km} \times 10 \text{ km}$ region was then applied to estimate pollen using the predictor variables over a $300 \text{ km} \times 300 \text{ km}$ region. The results show that the two machine learning approaches produced pollen distribution maps highlighting relatively similar pollen abundance over similar areas. The potential estimation of pollen from the NEXRAD data without using other environmental measurements will lay the foundation to predict allergic pollen of different species (provided their size is detectable by the NEXRAD radar) over a large spatial scale covered by NEXRAD weather radars.

References

- 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.
- Biau, G., Devroye, L., Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9, 2015–2033.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- 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.
- D'Amato, G., Liccardi, G., Frenguelli, G. (2007). Thunderstorm-asthma and pollen allergy. *Allergy*, 62(1), 11–16.
- D'Amato, G., Rottem, M., Dahl, R., Blaiss, M.S., Ridolo, E., Cecchi, L., Rosario, N., Motala, C., Ansotegui, I., Annesi-Maesano, I. (2011). Climate change, migration, and allergic respiratory diseases: an update for the allergist. *World Allergy Organization Journal*, 4(7), 121.
- Hagan, M.T., & Menhaj, M.B. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE transactions on Neural Networks*, 5(6), 989–993.
- 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. (2009). *Neural networks and learning machines* (Vol. 3). Upper Saddle River: Pearson.
- 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. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (pp. 1137–1143). San Francisco: Morgan Kaufmann.
- Laaïdi, M., Laaïdi, 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.
- Lake, I.R., Jones, N.R., Agnew, M., Goodess, C.M., Giorgi, F., Hamaoui-Laguel, L., Semenov, M.A., Solomon, F., Storker, J., Vautard, R., et al. (2016). Climate change and future pollen allergy in Europe. *Environmental Health Perspectives*, 125(3):385–391.
- 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. (2018). Machine learning applications for earth observation. *Earth Observation Open Science and Innovation*, 165.
- Liaw, A., & Wiener, M. (2002). Classification and regression by RandomForest. *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, 1–10.
- Molina, R.T., Palacios, I.S., Rodríguez, A.F.M., Muñoz, J.T., Corchero, A.M. (2001). Environmental factors affecting airborne pollen concentration in anemophilous species of plantago. *Annals of Botany*, 87(1), 1–8.
- Nowosad, J. (2016). Spatiotemporal models for predicting high pollen concentration level of *Corylus*, *Alnus*, and *Betula*. *International Journal of Biometeorology*, 60(6):843–855.
- Rodríguez-Rajo, F., Astray, G., Ferreiro-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.
- 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.
- 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.
- 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.
- 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.
- Zewdie, G.K., Liu, X., Wu, D., Lary, D.J. (in press 2019). Applying machine learning to forecast daily ambrosia pollen using environmental and NEXRAD radar parameters. *Environmental Monitoring and Assessment*, 191.
- Zhao, F., Elkelish, A., Durner, J., Lindermayr, C., Winkler, J.B., Rüeff, 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.