# Using machine learning to estimate atmospheric Ambrosia pollen concentrations in Tulsa, OK

Authors: Liu, Xun, Wu, Daji, Zewdie, Gebreab K, Wijerante, Lakitha, Timms, Christopher I, et al.

Source: Environmental Health Insights, 11(1)

Published By: SAGE Publishing

URL: https://doi.org/10.1177/1178630217699399

# Using machine learning to estimate atmospheric *Ambrosia* pollen concentrations in Tulsa, OK

Environmental Health Insights
Volume 11: 1–10
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1178630217699399

Xun Liu[1], Daji Wu[1], Gebreab K Zewdie[1], Lakitha Wijerante[1], Christopher I Timms[1], Alexander Riley[1], Estelle Levetin[2] and David J Lary[1]

[1]The University of Texas at Dallas, Richardson, TX, USA. [2]The University of Tulsa, Tulsa, OK, USA

**ABSTRACT:** This article describes an example of using machine learning to estimate the abundance of airborne *Ambrosia* pollen for Tulsa, OK. Twenty-seven years of historical pollen observations were used. These pollen observations were combined with machine learning and a very complete meteorological and land surface context of 85 variables to estimate the daily *Ambrosia* abundance. The machine learning algorithms employed were Least Absolute Shrinkage and Selection Operator (LASSO), neural networks, and random forests. The best performance was obtained using random forests. The physical insights provided by the random forest are also discussed.

**KEYWORDS:** Pollen, machine learning

**RECEIVED:** November 1, 2016. **ACCEPTED:** February 13, 2017.

**PEER REVIEW:** Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 1845 words, excluding any confidential comments to the academic editor.

**TYPE:** Original Research

**FUNDING:** The author(s) received no financial support for the research, authorship, and/or publication of this article.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Xun Liu, The University of Texas at Dallas, Richardson, TX 75080, USA. Email: xxl135130@utdallas.edu

## Introduction

*Ambrosia* (ragweed) pollen with concentrations of 5–20 pollen grains per cubic meter is allergenic for many people.[1] The *Ambrosia* genus consists of more than 40 species. Of all the Ambrosia species, *A. artemisiifolia* (common ragweed) has the highest allergenic potency and can produce millions of pollen grains per day. Figure 1 shows the *A. artemisiifolia* life cycle. Figure 1 is plotted based on Solter et al.[2] Ragweed typically blooms and produces large amounts of pollen between August and October.

The latest National Health and Nutrition Examination Survey (NHANES) III estimates that 26.2% of the US population is sensitive to *Ambrosia* pollen.[3] A single plant can release about a billion pollen grains in a season.[4] Typically, the size of a single pollen grain of *Ambrosia* is between 15 and 25μm.[5] Particles of this size do not typically go deep into the human peripheral airways. However, smaller particles with a size of less than 10μm can go deep into the peripheral airways.[6] *Ambrosia* pollen can fragment into smaller particles ranging in size from 0.5 to 4.5μm in size.[7]

Allergic conditions such as asthma and rhinitis can be worsened by pollen. According to the World Health Organization (WHO),[8] 9% of US students younger than 18 experienced seasonal hay fever symptom in 2008; three quarters of these are believed to be caused by *Ambrosia* pollen. Approximately, 50 million Americans have allergic diseases. On average, each day in the USA, 44,000 people have an asthma attack. On average, in the USA, asthma causes 36,000 kids to miss school, 27,000 adults to miss work, and 4,700 people to visit the emergency room (with 1,200 of these emergency room visits leading to a hospital admission) each day. Unfortunately, on average, nine of those admitted with asthma dies.

Early warning of imminent high pollen levels could be valuable for people with conditions such as asthma and chronic obstructive pulmonary disease (COPD). However, giving these accurate early warnings is a challenging task. The traditional approach of measuring the atmospheric pollen abundance with a Burkard trap is labor intensive, involving manual counting of the number of pollen particles under a microscope. Manual counting is also necessary because it has an inbuilt latency, often of approximately a week.

In this paper, we show that the pollen abundance can be estimated using machine learning and a suite of environmental parameters from meteorology and remote sensing. Some previous studies have used neural networks (NNs) to estimate pollen.[9–13] In this article we use machine learning to explore the relative importance of a variety of environmental factors in estimating the airborne abundance of *Ambrosia* pollen over a 27-year period in Tulsa, OK.

## Previous Work

Howard and Levetin[14] measured and analyzed the long-term *Ambrosia* pollen counts observed at the University of Tulsa and developed a multi-linear forecasting model to predict the next day's pollen concentration. In this model, they associated the pollen concentration with the long-term phenology[15] and a set of meteorological factors that included the minimum temperature $T_{min}$, precipitation $P$, and the mean dew point $DP$:

$$\ln(C) = -0.505 - 0.018 \times T_{min} - 0.108 \times P + 0.013 \times DP + 0.970 \times PH \quad (1)$$

where $C$ is the pollen concentration and $PH$ the phenology. The phenology is the mean pollen count for that day of the year for all prior years of *Ambrosia* pollen observations in Tulsa, OK.

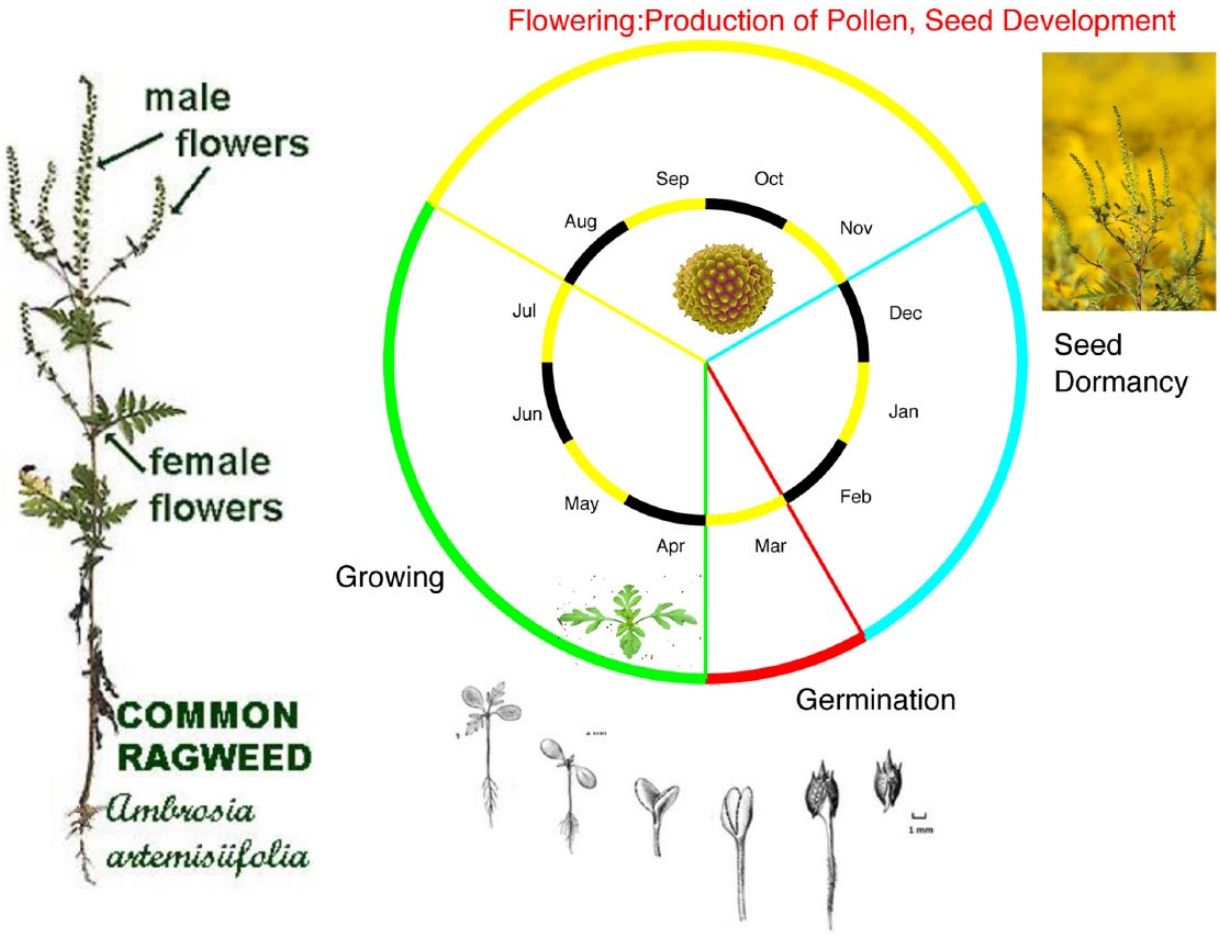


**Figure 1.** A schematic showing the *Ambrosia* life cycle.

The accuracy of this multi-linear model was examined by Howard and Levetin.[14] Figure 2 shows a scatter diagram of this multi-linear model, where the *x*-axis shows the estimated pollen count and the *y*-axis the actual observed pollen count. For a perfect prediction all the points lie on a straight line with a slope of one and an intercept of zero. This figure is used as a benchmark for the comparison of results obtained later, using a variety of machine learning approaches. In Figure 2 the correlation coefficient is 0.59.

A key point to note is that this multi-linear model shown in equation (1) makes use of the phenology (i.e. the observed mean pollen count for that day over the 25 years of observations). In this study we have partnered with Levetin, using the same data presented in Howard and Levetin,[14] except that here we use machine learning instead of multi-linear regression and that *the phenology* was not used as an input variable. Instead our goal was to be able to estimate pollen based only on a comprehensive environmental context.

The goal of this study was to accurately estimate the pollen count in Tulsa, OK, using just the readily available contextual information such as meteorological analysis, weather radar, and satellite data. In the linear model, it can be observed that the phenology item has a much higher weight than the other factors. Obtaining an accurate phenology for a given location is very labor intensive, and is rather expensive as a result.

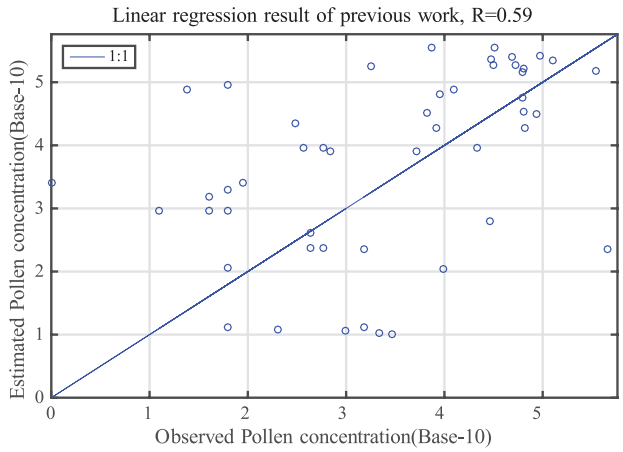


**Figure 2.** Correlation of the model-predicted pollen concentrations with observed validation data for 2013. Plotted based on equation (1), using data from Howard and Levetin [14] and Rienecker et al.[16]

In contrast, the contextual meteorological data are readily available. The goal of the current study was to show that an accurate pollen estimate can be provided from these contextual data alone, thereby allowing the possibility of dispensing with labor-intensive phenology information.

In this endeavor, a set of machine learning approaches was used. As shown below, some perform better than others. Let us now examine these machine learning approaches in turn,

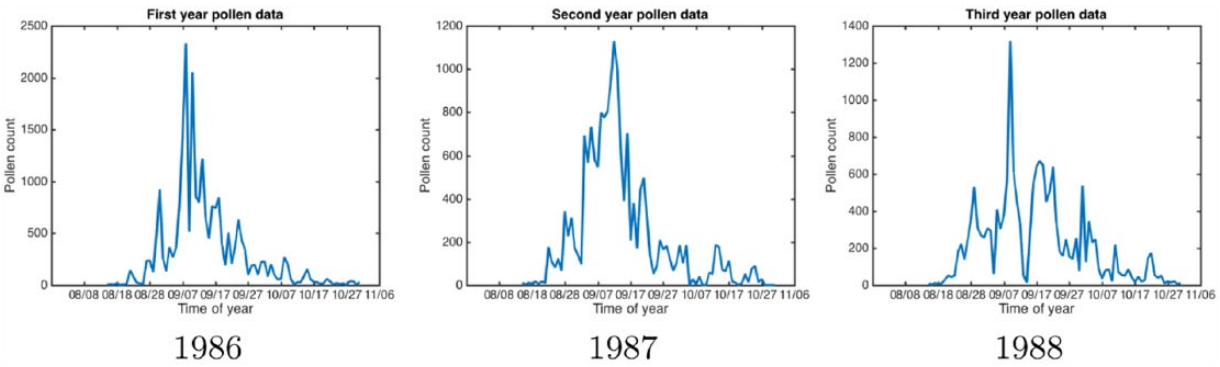


**Figure 3.** Annual pollen data through 1986 to 1988.

starting with the best performing algorithm and finishing with the poorest performing algorithm.

## Data Sets Used

Two types of data were used in this study. First, observational data of the abundance of airborne *Ambrosia* pollen (e.g. Figure 3) which was previously reported by Howard and Levetin.[14] Second, a comprehensive meteorological and land surface context for the pollen observations provided by the NASA MERRA meteorological reanalysis.[16–18]

The daily airborne pollen concentration was obtained at the University of Tulsa in Tulsa, Oklahoma. During the time period of 1986 to 2014, a Burkard Volumetric Spore Trap was deployed on the roof of Oliphant Hall, collecting airborne pollen day and night. Inside the Burkard trap, the pollen is deposited onto a greased strip of Melenex tape that is affixed to a rotating drum. Tapes were collected each week, divided into strips for each day, and then examined at a magnification of 400× for pollen grain identification and counting under a microscope. Once the pollen counts were obtained, they were multiplied by a conversion factor to yield the overall atmospheric pollen concentration.[14]

Figure 3 shows the *Ambrosia* pollen counts at Tulsa, OK, for three consecutive years, 1986–1988. We note that for each year, the duration of the *Ambrosia* pollen season is similar, as is the timing of the peak pollen counts. The average *Ambrosia* pollen counts at Tulsa, OK, over all 27 years of observation is shown in Figure 4. The *Ambrosia* pollen season starts in August, the peak concentrations are reached in September, and then slowly decline through October. Figure 4 shows the average time variation for the 27 years of pollen data.

For every day of the 27 year period from 1987 to 2013, for which pollen data were available at Tulsa, OK, the hourly values of 85 environmental parameters were retrieved from the NASA MERRA meteorological analysis that describe the surface meteorology and soil state.[16] These 85 variables are listed in Table 2 of the appendix and comprehensively characterize both the air close to the land surface and the land surface itself. Since the pollen data are only available as daily values, three summary statistics were also calculated for each of the 85 environmental parameters: the mean, minimum, and

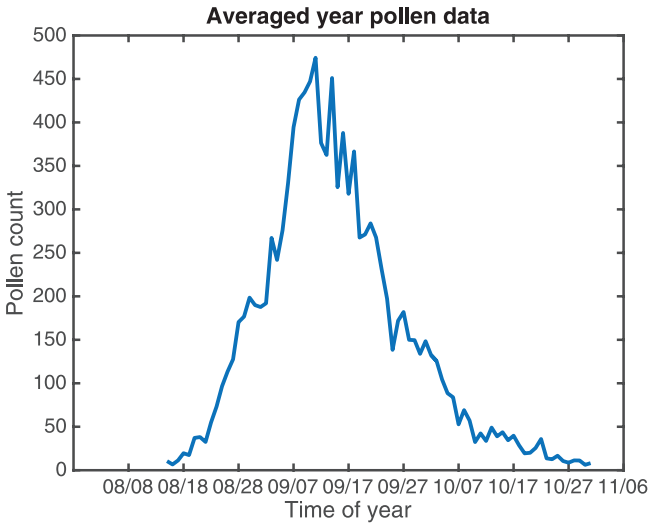


**Figure 4.** Averaged 1986–2014 annual pollen data.

maximum. According to life experience, weather plays a key role in time, concentration, and for how long pollen is released by plants. For example, windy dry weather typically leads to higher levels of pollen that are rapidly dispersed. When it rains, pollen is quickly washed out of the atmosphere. Since the plant's likelihood of releasing pollen on any given day is naturally affected by that plant's recent history, we also time lagged each of the 85 parameters by a delay that varied from 1 to 30 days. This leads to a total of $85 \times 30 \times 3 = 7{,}650$ variables that were used in our machine learning studies. Of these 7,650 variables, some are not important for estimating the pollen count. The machine learning automatically highlighted for us which variables are the most significant (Figure 6(c)).

A comparison of three machine learning regression approaches to show which performs best in estimating atmospheric pollen abundance was done. A brief overview of each approach is provided.

## Machine Learning

Machine learning is an automated implementation of the scientific method,[19] following the same process of generating, testing, and discarding or refining hypotheses. While a scientist

may spend his or her entire career coming up with and testing a few hundred hypotheses, a machine-learning system can do the same in a fraction of a second. Machine learning provides an objective set of tools for automating discovery. It is therefore no surprise that machine learning revolutionizing many areas of science, technology and business.[20]

For each machine learning approach we used, the performance was quantified using a scatter diagram. In the scatter diagram the actual observations were plotted against the current study machine learning estimates. A perfect scatter diagram is a straight line with a slope of one and an intercept of zero. In each case, the data were randomly split into two independent samples; one sample was used for training and the second sample for an independent validation, that is, the validation data were *not* used in the training stage of the algorithms. Table 1 shows the correlation coefficients for the various machine learning approaches used in this study. The best performing approach, namely the random forest, is listed first. Here $R_T$ is the correlation coefficient for the training dataset and $R_V$ is the correlation coefficient for the totally independent validation dataset.

**Table 1.** Correlation coefficients for the various machine learning approaches used in this study, with the best performing approach listed first. Here $R_T$ is the correlation coefficient for the training dataset and $R_V$ is the correlation coefficient for the totally independent validation dataset.

| MACHINE LEARNING APPROACH WITHOUT PHENOLOGY | CORRELATION COEFFICIENT | |
|---|---|---|
| | TRAINING, $R_T$ | VALIDATION, $R_V$ |
| Random forest | 1 | 0.98 |
| NN | 0.91 | 0.61 |
| LASSO | 0.53 | 0.56 |
| Prior multi-linear study with phenology | 0.68 | |

### *Random forest*

A random forest is an ensemble statistical learning approach, consisting of an ensemble of decision trees.[21–23] A schematic representation of a random forest is shown in Figure 5. Random forests have proved to be a very useful multi-variable, non-linear, non-parametric approach for both regression and supervised classification. Ensemble methods are less prone to over-learning the noise of the data and typically provide better generalization. A random forest also provides a useful ranking of the relative importance of the predictors, an example of which is shown in Figure 6(c) for estimating pollen abundance. To decide how many trees we should use in our random forest, we examined how the error decreased as the number of trees is increased (Figure 6(e)).

A random forest can facilitate estimation of the pollen count as a multi-variate, non-parametric function of N input variables, i.e.

$$pollen\ count = f_{Random\ Forest}(x_1,\ldots,x_N) \tag{2}$$

where $x_1,\ldots,x_N$ are the N readily available environmental parameters (listed in the appendix).

Two enhancements were then made for a standard random forest implementation that allowed both improvement of the performance and provided an estimated error for each pollen count that is estimated. The enhancement was inspired by Newton–Raphson iteration.

A series of iterations were executed, for each iteration, a random forest was used to estimate the pollen count as indicated in equation (2). Then, the estimated pollen count was compared with the actual pollen count to calculate an error, that is:

$$\text{error} = \text{observation} - \text{estimation} \tag{3}$$

Next, an additional random forest was used to learn this error. After each iteration, the random forest estimate of the pollen count was then corrected using the error estimated by this additional random forest, that is, by rearranging equation (3) and replacing the observation by our random forest estimate of the pollen count, and by replacing the error with the estimated error provided by the second random forest:

$$\text{improved estimation} = \text{initial estimation} + \text{estimated error} \tag{4}$$

This was then repeated for a set of n iterations (we used $n$ = 10). After each iteration, the estimated pollen count, and estimated pollen count error were added as additional input variables for the next iteration. This considerably improved the reliability of our estimated pollen count as can be seen by comparing verification scatter diagrams in Figure 6(a) and (b). In these scatter diagrams, the $x$-axis shows the observed amount of pollen and the $y$-axis shows the estimated amount of pollen. The error bars show the estimated uncertainty. As shown, these estimates do not require the phenology to be specified, yet show a substantial improvement in a prior study shown in Figure 2. Figure 2(a) shows the scatter diagram for the first iteration and Figure 2(b) shows the much improved scatter diagram after the last iteration. The approach offers very good performance. Interestingly, when the pollen estimations were tested using a completely independent data sample not used in the training (the validation dataset), the correlation coefficient is actually even better than that for the training dataset. *These scatter diagrams show the remarkable ability of the iterative random forest approach to accurately estimate the airborne pollen count*.

Figure 6(c) shows the relative importance of the 20 most important variables for estimating pollen count. The random forest indicated that the five most important parameters in estimating the pollen count are: the vegetation greenness 26 days prior, the current surface roughness length for sensible heat, the displacement height 15 days prior, the energy stored in all land reservoirs 30 days prior, and the current surface humidity.

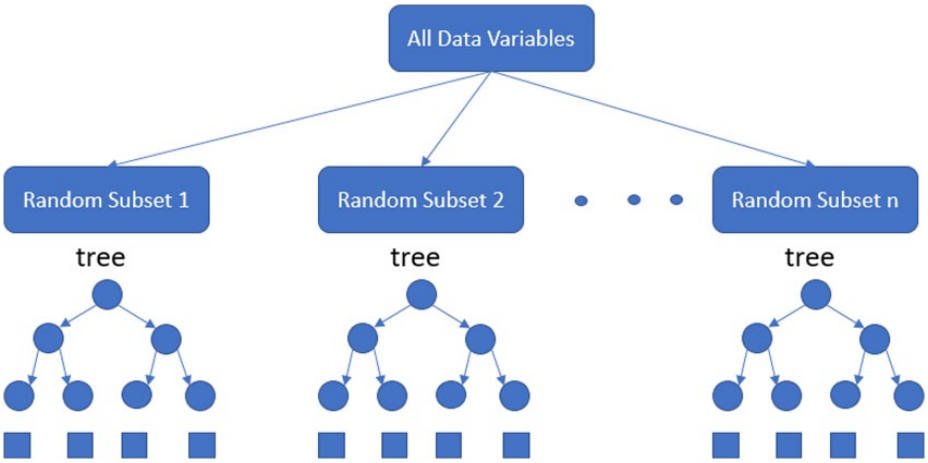


**Figure 5.** Schematic of a random forest. A random forest is an ensemble of decision trees.

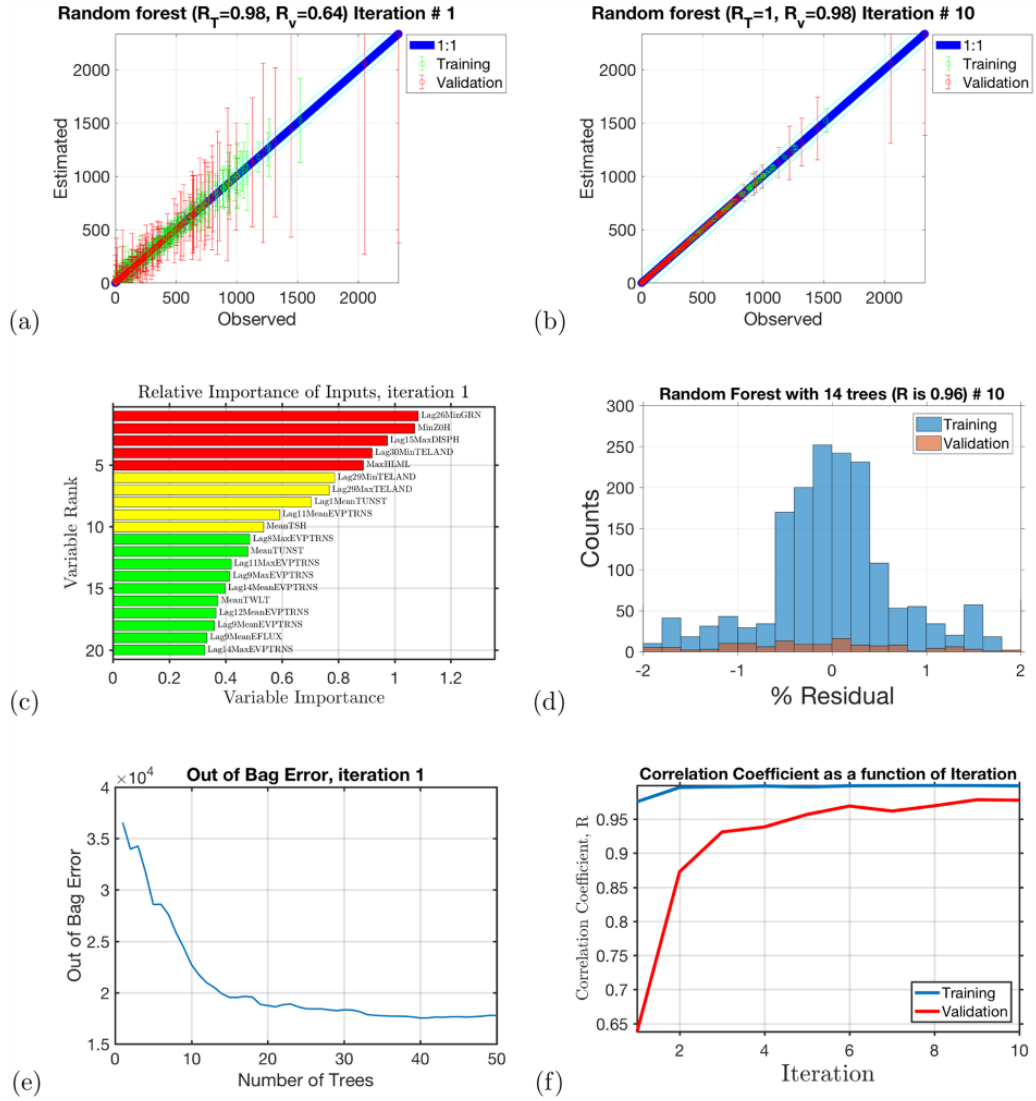


**Figure 6.** Descriptions for the random forest result. (a), (b) Verification scatter diagrams, with the *x*-axis showing the observed amount of pollen and the *y*-axis showing the estimated amount of pollen, while the error bars show the estimated uncertainty. We note that these estimates do *not* require the phenology to be specified. In (a) we show the scatter diagram for the first iteration and in (b) we show the much improved scatter diagram after the last iteration. (c) The relative importance of the 20 most important variables for estimating the pollen count. (d) Histogram of the residuals between the observed and estimated pollen counts. (e) Variation of the error as a function of the number of trees in the random forest. (f) The correlation coefficient for the training and independent validation datasets as a function of iteration.

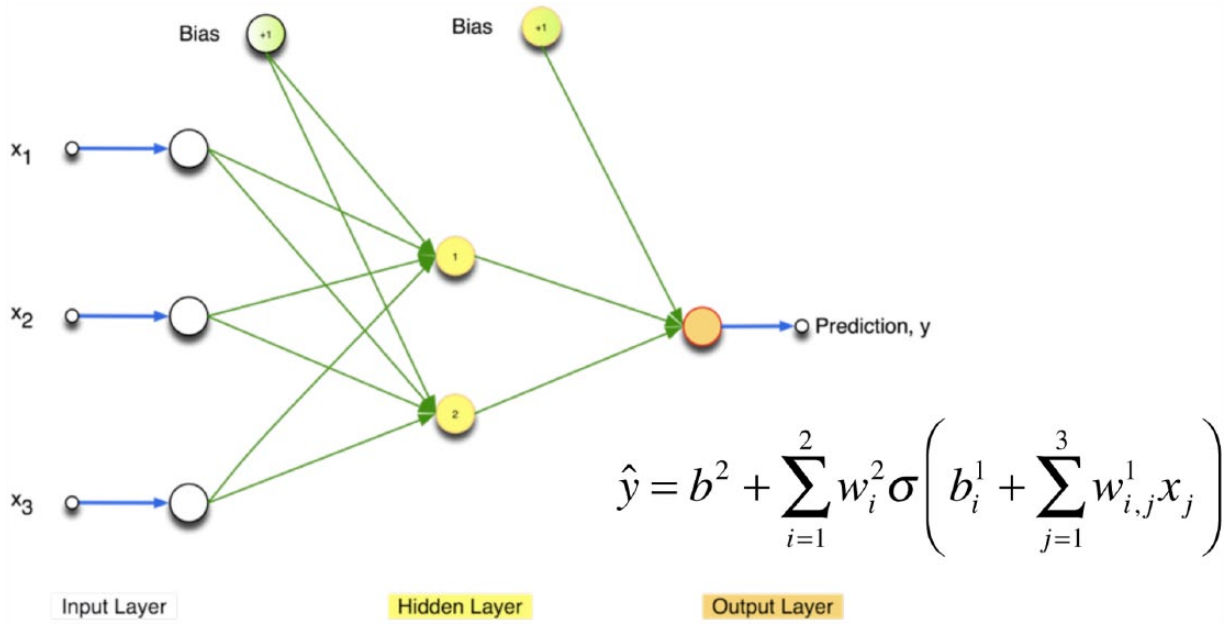


**Figure 7.** Schematic of a single hidden layer, feed-forward NN. Each arrow corresponds to a real-valued parameter, or a weight, of the network. The values of these parameters are tuned in the network training. Here *b* are the biases, *w* are the weights, and $\sigma$ is the activation function.

For air flows over the ground, when the scale of the land surface irregularities is much greater than the viscous scale, then a high surface roughness causes a local equilibrium breakdown by momentum transfer due to local pressure gradients at a height comparable with the vertical dimension of the surface irregularities, thereby affecting the boundary scale roughness length, $z_0$ .[24] The random forest highlighted this phenomenon, indicating that the current surface roughness length for sensible heat (sensible heat is related to changes in temperature with no change in phase) and the displacement height 15 days prior were both significant factors in estimating the pollen count.

Figure 2(d) shows a histogram of the residuals between the observed and estimated pollen counts.

Figure 2(e) shows how the error varies as a function of the number of trees in the random forest. It is obvious that, the error approaches a constant after the number of trees reaching 20. Thus, number of tree estimators should be larger than 20 for good performance of regression. It was set to 50 in this article.

Figure 2(f) shows the correlation coefficient for the training and independent validation datasets as a function of iteration. The training data error approaches a constant after four iterations. So iterations was set to 10 (i.e. more than four) in this article.

### *Neural Network*

NNs are non-linear, non-parametric learning algorithms inspired by biological networks such as those found in the human brain.[25–27] NNs are capable of approximating non-linear functions by the adaptive adjustment of their weights using a training algorithm. Figure 7 shows a schematic of a single hidden layer, feed-forward NN. Each arrow corresponds to a real-valued parameter, or a weight, of the network. The values of these parameters are tuned in the network training (*b* are the biases, *w* are the weights, and σ is the activation function). Associated with each node interconnection is a weight and a bias. These weights start as random numbers and during the process of training, they are iteratively updated.

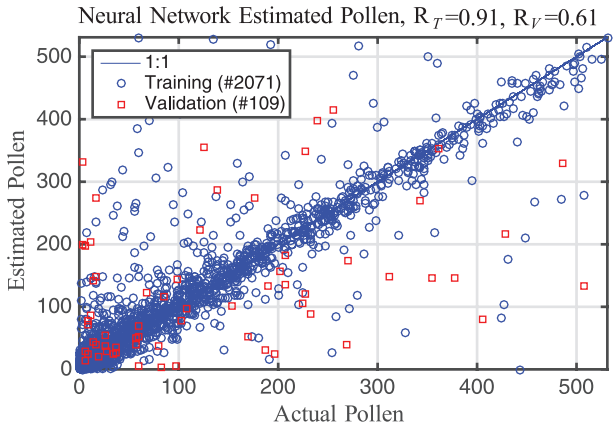


**Figure 8.** Scatter diagram for the airborne pollen estimates made using a NN.

Figure 8 shows the neural network scatter diagram. The validation correlation coefficient, $R_V = 0.61$, is not as good as that for the random forest.

### *LASSO Method*

The Least Absolute Shrinkage and Selection Operator (LASSO) is a linear regression method that involves both variable selection

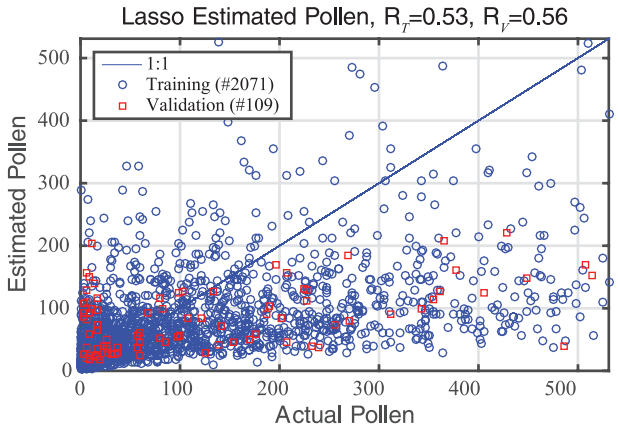


**Figure 9.** Scatter diagram for the airborne pollen estimates made using the LASSO approach.

and regularization.[28] The main benefit of using the LASSO approach is that it highlights the most important subset of parameters that can best describe the pollen concentration. The LASSO approach is similar to Pearson correlation analysis that is often used with classic linear regression models. The LASSO approach uses only a subset of the original predictors.

Figure 9 shows a scatter diagram for the LASSO pollen estimate. The *x*-axis shows the observed pollen amount and the *y*-axis shows the LASSO estimated pollen amount. The blue circles depict the training dataset, which has a correlation coefficient, $R_T = 0.53$. The red squares depict the independent validation dataset, which has a correlation coefficient, $R_V = 0.56$.

## Conclusion

In this article, a new *Ambrosia* pollen estimation model for Tulsa, OK, has been developed. The pollen concentration was described as a non-linear multi-variate function of the input variables, where the multi-variate function is provided by three different machine learning algorithms: LASSO, NNs, and random forests. The input environmental parameters are readily available from the NASA MERRA meteorological and land surface analysis. The random forest performed the best, and also provided insight into the relative importance of the 85 input variables. The most important input variables were *vegetation greenness*, *displacement height*, *roughness length of sensible heat*, *soil evaporation*, and *energy stored in all reservoirs*.

In future studies we will be exploring the additional information that can be provided by LANDSAT and weather radar. LANDSAT provides the surface reflectivity in multiple wavelengths. When ragweed blooms there will be a change in the surface reflectivity over multiple wavelengths. Weather radar detects airborne particles such as precipitation. The radar signal is also reflected by other particles such as smoke, pollen, and even insects.

## REFERENCES


1. Bacsi A, Choudhury BK, Dharajiya N, Sur S, Boldogh I. Subpollen particles: carriers of allergenic proteins and oxidases. *J Allergy Clin Immunol* 2006; 118(4): 844–850.
2. Solter U, Starfinger U, Verschwele A. HALT Ambrosia-complex research on the invasive alien plant ragweed (Ambrosia artemisiifolia L.) in Europe. *Julius-Kühn-Archiv*. 2012; 434: 627.
3. Arbes SJ, Gergen PJ, Elliott L, Zeldin DC. 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* 2005; 116(2): 377–383.
4. Thompson JL, Thompson JE. The urban jungle and allergy. *Immunol Allergy Clin N Amer* 2003; 23(3): 371–387.
5. Bacsi A, Choudhury BK, Dharajiya N, Sur S, Boldogh I. Subpollen particles: carriers of allergenic proteins and oxidases. *J Allergy Clin Immunol* 2006; 118(4): 844–850.
6. Weber RW, Adkinson Jr N, Bochner B, et al. Aerobiology of outdoor allergens. *Middleton's Allergy: Principles and Practice* 2013; 69: 430.
7. Oswalt ML, Marshall GD. Ragweed as an example of worldwide allergen expansion. *Allergy Asthma Clin Immunol* 2008; 4(3): 1.
8. World Health Organization. 7 million premature deaths annually linked to air pollution. http://www.who.int/mediacentre/news/releases/2014/air-pollution/en (Published 25 March 2014; accessed 12 August 2014).
9. Astray G, Fernandez-Gonzalez M, Rodriguez-Rajo F, Lopez D, Mejuto J. Airborne castanea pollen forecasting model for ecological and allergological implementation. *Science Total Environ* 2016; 548: 110–121.
10. Castellano-Mendez M, Aira M, Iglesias I, Jato V, Gonzalez-Manteiga W. Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air. *Int J Biometeorol* 2005; 49(5): 310–316.
11. Iglesias-Otero M, Fernandez-Gonzalez M, Rodriguez-Caride D, Astray G, Mejuto J, Rodriguez-Rajo F. A model to forecast the risk periods of Plantago pollen allergy by using the ANN methodology. *Aerobiologia* 2015; 31(2): 201–211.
12. Navares R, Aznarte JL. What are the most important variables for Poaceae airborne pollen forecasting? *Sci Total Environ* 2016; in press.
13. Sanchez JMB, Lugilde DN, de Linares Fernandez C, de la Guardia CD, Sanchez FA. Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. *Expert Syst Applic* 2007; 32(4): 1218–1225.
14. Howard LE, Levetin E. Ambrosia pollen in Tulsa, Oklahoma: aerobiology, trends, and forecasting model development. *Ann Allergy Asthma Immunol* 2014; 113(6): 641–646.
15. Chapman DS, Haynes T, Beal S, Essl F, Bullock JM. Phenology predicts the native and invasive range limits of common ragweed. *Global Change Biology* 2014; 20(1): 192–202.
16. Rienecker MM, Suarez MJ, Gelaro R, et al. MERRA: NASA's modern-era retrospective analysis for research and applications. *J Climate* 2011; 24(14): 3624–3648.
17. Bosilovich M, et al. MERRA-2: Initial evaluation of the climate. *NASA Technical Report Series on Global Modeling and Data Assimilation* NASA/TM-2015-104606.39:139.
18. Koster RD, McCarty W, Coy L, et al. MERRA-2 Input Observations: Summary and Assessment. 2016. https://ntrs.nasa.gov/search.jsp?R=20160014544
19. Domingos P. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books, 2015.
20. Lary DJ, Alavi AH, Gandomi AH, Walker AL. Machine learning in geosciences and remote sensing. *Geosci Front* 2016; 7(1): 3–10.
21. Breiman L. Random forests. *Machine Learning* 2001; 45(1): 5–32.
22. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Machine Intell* 1998; 20(8): 832–844.
23. Breiman L. Random forests. *Machine Learning* 2001; 45(1): 5–32.
24. Martano P. Estimation of surface roughness length and displacement height from single-level sonic anemometer data. *J Appl Meteorol* 2000; 39(5): 708–715.
25. Werbos P. Beyond regression: New tools for prediction and analysis in the behavioral sciences. 1974.
26. Bishop CM. *Neural networks for pattern recognition*. Oxford: Oxford University Press, 1995.
27. Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning (Springer Series in Statistics*, Vol. 1). Berlin: Springer, 2001.
28. Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Ser B Methodological* 1994; 58: 267–288.
29. Burgan RE, Hartford RA. Monitoring vegetation greenness with satellite data. 1993. https://www.fs.fed.us/rm/pubs_int/int_gtr297.pdf
30. Brutsaert W. The roughness length for water vapor sensible heat, and other scalars. *J Atmospheric Sci* 1975; 32(10): 2028–2031.

## Appendix: Full List of the Environmental Variables Used

**Table 2.** Variable names, abbreviations and units.

| VARIABLE | DESCRIPTION | UNITS |
|---|---|---|
| EFLUX | latent heat flux(positive upward) | $W \cdot m^{-2}$ |
| EVAP | Surface evaporation | $kg \cdot m^{-2} \cdot s^{-1}$ |
| HFLUX | Sensible heat flux (positive upward) | $W \cdot m^{-2}$ |
| TAUX | Eastward Surface wind stress | $N \cdot m^{-2}$ |
| TAUY | Northward Surface wind stress | $N \cdot m^{-2}$ |
| TAUGWX | Eastward gravity wave surface stress | $N \cdot m^{-2}$ |
| TAUGWY | Northward gravity wave surface stress | $N \cdot m^{-2}$ |
| PBLH | Planetary boundary layer height | m |
| DISPH | Displacement height | m |
| BSTAR | Surface buoyancy scale | $m \cdot s^{-1}$ |
| USTAR | Surface velocity scale | $m \cdot s^{-1}$ |
| TSTAR | Surface temperature scale | K |
| QSTAR | Surface humidity scale | kg |
| RI | Surface Richardson number | non-dimensional |
| ZOH | Roughness length, sensible heat | m |
| ZOM | Roughness length, momentum | m |
| HLML | Height of center of lowest model layer | m |
| TLML | Temperature of lowest model layer | m |
| QLML | Specific humidity of lowest model layer | kg |
| ULML | Eastward wind of lowest model layer | $m \cdot s^{-1}$ |
| VLML | Northward wind of lowest model layer | $m \cdot s^{-1}$ |
| RHOA | Surface air density | $kg \cdot m^{-3}$ |
| SPEED | Three-dimensional wind speed for surface fluxes | $m \cdot s^{-1}$ |
| CDH | Surface exchange coefficient for heat | $kg \cdot m^{-2} \cdot s^{-1}$ |
| CDQ | Surface exchange coefficient for moisture | $kg \cdot m^{-2} \cdot s^{-1}$ |
| CDM | Surface exchange coefficient for momentum | $kg \cdot m^{-2} \cdot s^{-1}$ |
| CN | Surface neutral drag coefficient | non-dimensional |
| TSH | Effective turbulence skin temperature | K |
| QSH | Effective turbulence skin humidity | kg |
| FRSEAICE | Fraction of sea-ice | Fraction |
| PRECANV | Surface precipitation flux from anvils | $kg \cdot m^{-2} \cdot s^{-1}$ |
| PRECCON | Surface precipitation flux from convection | $kg \cdot m^{-2} \cdot s^{-1}$ |
| PRECLSC | Surface precipitation flux from large-scale | $kg \cdot m^{-2} \cdot s^{-1}$ |

**Table 2.** (Continued)

| VARIABLE | DESCRIPTION | UNITS |
|---|---|---|
| PRECSNO | Surface snowfall flux | $kg \cdot m^{-2} \cdot s^{-1}$ |
| PRECTOT | Total surface precipitation flux | $kg \cdot m^{-2} \cdot s^{-1}$ |
| PGENTOT | Total generation of precipitation | $kg \cdot m^{-2} \cdot s^{-1}$ |
| PREVTOT | Total re-evaporation of precipitation | $kg \cdot m^{-2} \cdot s^{-1}$ |
| GRN | Vegetation greenness fraction | Fraction |
| LAI | Leaf area index | $m^2$ |
| GWETROOT | Root zone soil wetness | fraction |
| GWETTOP | Top soil layer wetness | fraction |
| TPSNOW | Top snow layer temperature | K |
| TUNST | Surface temperature of unsaturated zone | K |
| TSAT | Surface temperature of saturated zone | K |
| TWLT | Surface temperature of wilted zone | K |
| PRECSNO | Surface snowfall | $kg \cdot m^{-2} \cdot s^{-1}$ |
| PRECTOT | Total surface precipitation | $kg \cdot m^{-2} \cdot s^{-1}$ |
| SNOMAS | Snow mass | $kg \cdot m^{-2}$ |
| SNODP | Snow depth | m |
| EVPSOIL | Bare soil evaporation | $W \cdot m^{-2}$ |
| EVPTRNS | Transpiration | $W \cdot m^{-2}$ |
| EVPINTR | Interception loss | $W \cdot m^{-2}$ |
| EVPSBLN | Sublimation | $W \cdot m^{-2}$ |
| RUNOFF | Overland runoff | $kg \cdot m^{-2} \cdot s^{-1}$ |
| BASEFLOW | Baseflow | $kg \cdot m^{-2} \cdot s^{-1}$ |
| SMLAND | Snowmelt | $kg \cdot m^{-2} \cdot s^{-1}$ |
| FRUNST | Fractional unsaturated area | fraction |
| FRSAT | Fractional saturated area | fraction |
| FRSNO | Fractional snow-covered area | fraction |
| FRWLT | Fractional wilting area | fraction |
| PARDF | Surface downward PAR diffuse flux | $W \cdot m^{-2}$ |
| PARDR | Surface downward PAR beam flux | $W \cdot m^{-2}$ |
| SHLAND | Sensible heat flux from land | $W \cdot m^{-2}$ |
| LHLAND | Latent heat flux from land | $W \cdot m^{-2}$ |
| EVLAND | Evaporation from land | $kg \cdot m^{-2} \cdot s^{-1}$ |
| LWLAND | Net downward longwave flux over land | $W \cdot m^{-2}$ |

*(Continued)*

**Table 2.** (Continued)

| VARIABLE | DESCRIPTION | UNITS |
|---|---|---|
| SWLAND | Net downward shortwave flux over land | $W \cdot m^{-2}$ |
| GHLAND | Downward heat flux at base of top soil layer | $W \cdot m^{-2}$ |
| TWLAND | Total water store in land reservoirs | $kg \cdot m^{-2}$ |
| TELAND | Energy store in all land reservoirs | $J \cdot m^{-2}$ |
| WCHANGE | Total land water change per unit time | $kg \cdot m^{-2} \cdot s^{-1}$ |
| ECHANGE | Total land energy change per unit time | $W \cdot m^{-2}$ |
| SPLAND | Spurious land energy source | $W \cdot m^{-2}$ |
| SPWATR | Spurious land water source | $kg \cdot m^{-2} \cdot s^{-1}$ |
| SPSNOW | Spurious snow source | $kg \cdot m^{-2} \cdot s^{-1}$ |
| PM2.5 | Airborne Particulate | $\mu g \cdot m^{-3}$ |
| Soil | Soil type | non-dimensional |
| Lithology | Lithology | non-dimensional |
| Topography | Topography | m |
| PopulationDensity | Population density | |
| Type | Surface type | non-dimensional |
| AlbedoWSABand1 | Surface reflectivity at 470 nm | non-dimensional |
| AlbedoWSABand2 | Surface reflectivity at 555 nm | non-dimensional |
| AlbedoWSABand3 | Surface reflectivity at 670 nm | non-dimensional |
| AlbedoWSABand4 | Surface reflectivity at 858 nm | non-dimensional |
| AlbedoWSABand5 | Surface reflectivity at 1,240 nm | non-dimensional |
| AlbedoWSABand6 | Surface reflectivity at 1,640 nm | non-dimensional |
| AlbedoWSABand7 | Surface reflectivity at 2,130 nm | non-dimensional |