FORECASTING POLLEN AEROBIOLOGY WITH MODIS EVI, LAND COVER, AND PHENOLOGY USING MACHINE LEARNING TOOLS

Huete, Alfredo[1], *Tran, Nguyen Ngoc*[1], *Nguyen, Ha*[1], *Xie, Qiaoyun*[1], *and Katelaris, Constance*[2]

[1]University of Technology Sydney
[2]Western Sydney University

ABSTRACT

Grass pollens are a major source of aeroallergens globally, inducing allergic asthma and hay fever in up to 500 million people worldwide. Pollen forecasting research and methods are site-dependent and tend to be empirically derived composites of expert knowledge and weather data. In this study we utilize satellite-based information of landscape conditions and phenology to better discern and predict grass pollen evolution. We employed machine learning approaches to formulate and better understand relationships between landscape phenology and seasonal flowering-induced pollen concentrations. We show that machine learning approaches significantly improved pollen prediction capabilities and provided key information to better attribute changes in pollen counts driven by shifting ecological landscapes from climate change drivers.

Index Terms— Machine learning, pollen, phenology, satellite, grass

1. INTRODUCTION

Grass pollen allergen exposure in major urban populations have risen dramatically worldwide causing increasing rates of asthma and rhinitis. Changes in climate, invasive species, and landscape modification have been identified as potential environmental drivers for changes in airborne pollen levels [1]. Pollen exposure, including the frequency of high pollen concentration days and thunderstorm pollen events, is projected to intensify with climate variability raising threats of severe public health problems [2, 3].

Conventional methods for monitoring airborne pollen involve the use of labour-intensive spore traps that collect samples of pollen grains. Pollen aerobiology records provide valuable indicators of trends in allergenic grass pollen production resulting from impacts of climate and landscape changes. However, pollen forecasting methods are generally site-dependent and empirically-based using data from spore traps, expert knowledge and meteorological information that includes temperature, relative humidity and precipitation. These methods are hampered by a sparsity of pollen trap sampling and are not amenable to cross-site generalisations. Ecological information on landscape properties, condition, and presence of allergenic plant species are less well incorporated in forecasting models despite their importance in understanding the drivers of pollen aerobiology and to predict future trends of pollen aerobiology.

Satellite imagery overcome the restrictive coverage of in situ pollen networks and provide unique opportunities to probe how land cover modifications and vegetation responses to climate variability relate to grass pollen aerobiology. By virtue of its synoptic coverage and repeatability of measurements, remote sensing imagery is useful in providing insights into space-time environmental influences and offer opportunities to characterise ecosystem changes and sensitivity to climate variability. Generalized Additive Model (GAM) linear regression methods have shown strong correlations between MODIS satellite data and pollen concentrations over Australian and European sites [4], however cross-site relationships have been less successful.

Machine Learning (ML) adds new capabilities for combining satellite data with in situ ground measures, and has been shown useful in crop yield forecasting, aerosol product retrievals, and vegetation and ocean applications. Nowosad et al [5] used Random Forests to forecast pollen from several tree species. ML approaches can potentially couple time series satellite data of grassland phenology with higher frequency daily pollen counts. In this study we address critical knowledge gaps in relating pollen concentrations with ecologically relevant and dynamic landscape variables, through the use of machine learning approaches. Specifically, we investigated the utility of time series satellite data of evolving landscape conditions to inform on grass pollen aerobiology across a gradient of grass surface types of mixed cool- to warm season grass species found in eastern Australia major urban areas.

2. METHODS

2.1. Site description

978-1-5386-9154-0/19/$31.00 ©2019 IEEE

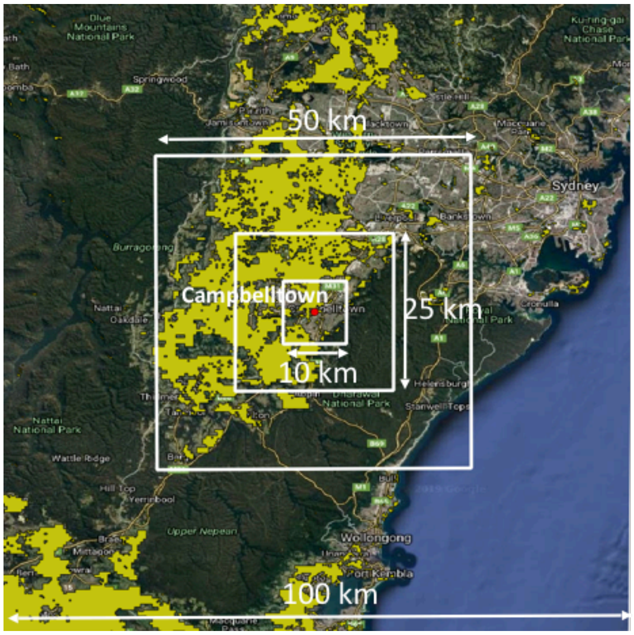

Figure 1. Study site area with Campbelltown pollen spore trap in the center and delineated grass and pasture areas.

An established pollen spore trap, Campbelltown (lat. 34.07S /long.150.796E) near Sydney, Australia and within a 100km x 100km area, was selected for initial model development (Fig. 1). Two additional spore trap locations have recently been deployed across the Sydney basin, which will be used for cross site model development.

The Sydney basin area is in a transition zone of mixed grass functional species that includes both subtropical warm season grasses (C_4) as well as cool season grasses (C_3). This results in a near year-round, bi-modal release of pollen, dominant in the Austral autumn (March-May) and Austral spring (September-November) seasons. Mean annual rainfall is approximately 1200 mm with mean annual maximum and minimum temperatures of 22^0 and 14^0, respectively. Grass and pasture land cover type was determined using the Geoscience Australia, Dynamic Land Cover Data Product.

2.2. Pollen data

Daily atmospheric grass pollen concentration data were obtained for all sites over a 4 year period (2010-2013). Aerobiological monitoring of pollen was conducted with volumetric Burkard spore traps/samplers (Burkard Scientific Ltd, Uxbridge, UK) according to methods described in [6]. Daily pollen concentrations were aggregated into 8-day (weekly) total pollen values to better align with the satellite data.

2.3. Satellite data

MODIS satellite time series data of the Enhanced Vegetation Index (EVI) product from the Terra platform sensor (MOD13A1) was extracted at 500-m and 16-day composites [7]. 16-day EVI time series were interpolated to generated an 8-day annual series of 46 values. EVI values over grass /pasture land cover areas were extracted over a 100 km by 100 km window around the Campbelltown pollen trap location. Four non-overlapping windows of EVI data were analysed as to their distance proximity from the pollen collection trap, to test the importance of grass pollen source distance to pollen counts at the pollen trap. The distance areas included 0-10km from the pollen trap, 10-25km, 25-50km, and 50-100km areas from the pollen trap (Fig. 1). The EVI was used as a spectral surrogate of plant chlorophyll activity and provided intra-annual, phenological, and inter-annual observations of vegetation dynamics on a grass pixel-by-pixel basis. In the initial tests of model development, the EVI grass pixel values within each of the 4 distance classes were averaged.

2.4. Modelling of pollen concentrations

To relate EVI seasonal profiles with pollen concentrations and assess prediction capabilities of grass pollinating periods, we tested Random Forest (RF) machine learning methods to predict pollen concentrations from seasonal and inter-annual MODIS EVI phenology profiles. Two RF models were developed and tested. In the first test, the satellite data and pollen concentration data were randomly split into two groups, 60% as training data and 40% as testing data. The input variables were 8-day EVI for each of the 4 distance classes of grass containing pixels; 8-day sequence EVI with lags of 8-, 16-, 24-, and 32-days; day of year (DOY) in 8-day interval (n=46), and year. The prediction variable was pollen concentrations at 8-day intervals for the entire pollen season and for each of the 4 years.

A second RF machine learning test made in which the RF model was trained with 3 years of 8-day EVI and pollen data, with an entire missing year as the test data. In this case, we wished to assess the potential of satellite EVI data and DOY to predict 8-day pollen for an entire year. In both RF models, we tested the distance variable of importance as well as the 8-, 16-, 24-, and 32 day lag variable importance.

3. RESULTS

As in a previous study, MODIS EVI time series data of grasslands preceded grass pollen seasonality [4]. The peak EVI is an important indicator of the initiation of the grass flowering/ reproductive period. Shortly after peak greenness, the grasses begin their flowering phenophase, followed by the pollen loading and release period (Fig. 2).

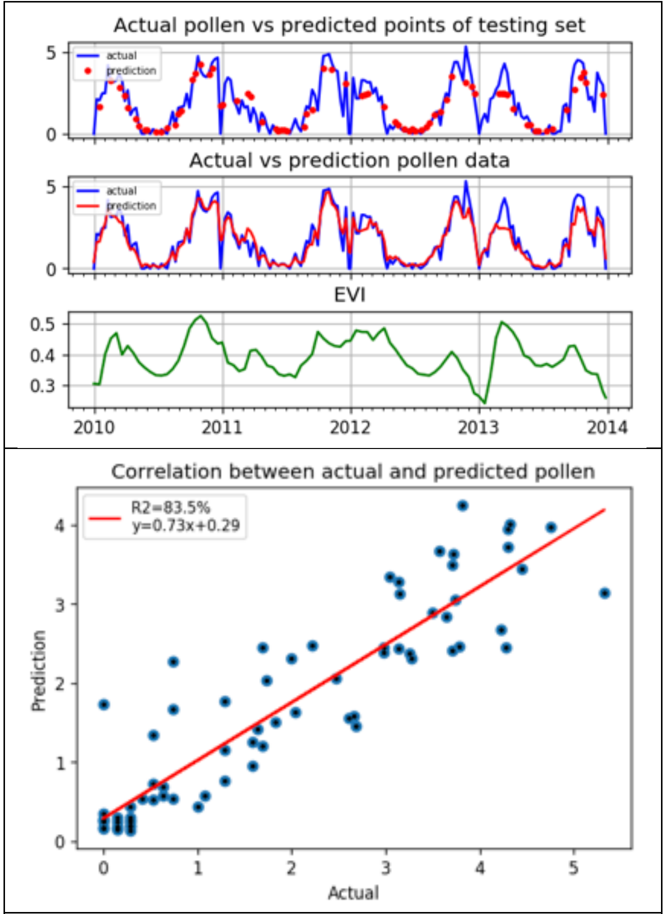

Figure 2. Random Forest pollen concentration model output using 60% randomly selected training data and 40% random test data for a 4 year time series. Top panel shows the predicted test points and their regression with actual pollen data is shown in bottom panel.

The first RF model that used a 60% randomized training data yielded a prediction capability of 83.5% for 8-day pollen concentrations over the remaining test data set (correlation, r of 0.92) (Fig. 2). This model was further found to work best at a zero-lag between EVI and pollen data. The variable importance for these results were DOY, followed by EVI, and year. The distance class was found to be highest at both 11-25 km distance from the pollen trap, as well as 50-100 km from the trap.

The second RF model used 75% training data from 3 separate years and 25% test data, representing an entire missing year with bimodal pollen seasons. The grassland EVI within 11-25 km distance from the pollen trap with zero-lag were used as input variables, based on variable importance. The predictive results were not as strong as in the randomized model run, but still significant, with 63% variance explained and a correlation of 0.8 (Fig. 3). This second RF model was better able to model the Austral autumn pollen peak in the test year of 2013, but not the Austral spring pollen peak in 2013 (Fig. 3), compared with the first RF model, which included a random set of training points over the year in question (Fig. 2).

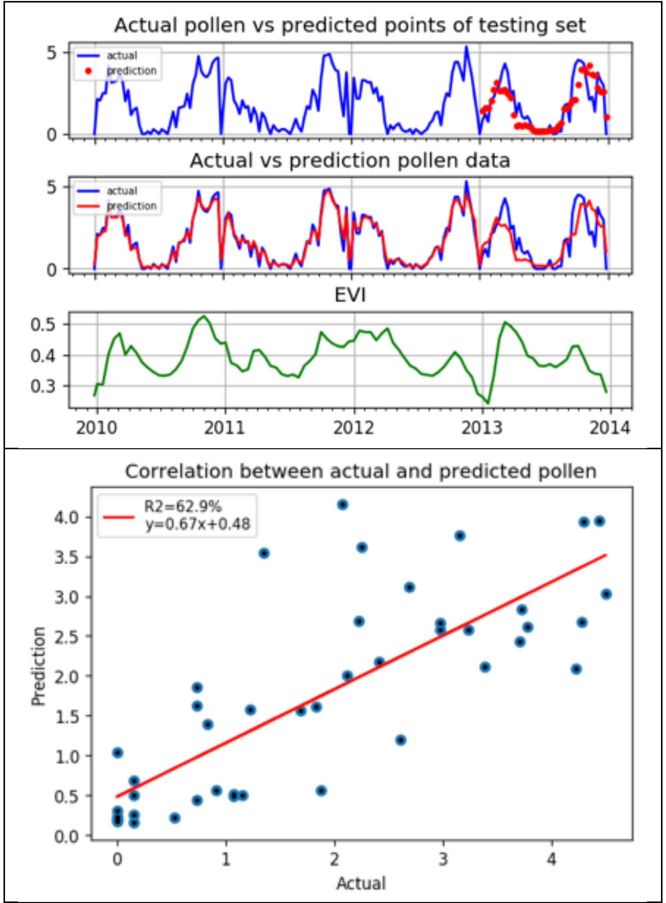

Figure 3. Random Forest pollen concentration model output using 3 years training data and 1 'missing year' test data for the 4 year time series. Top panel shows the predicted test points for the missing 4th year and their regression with actual pollen data is shown in bottom panel.

4. DISCUSSION

Our preliminary RF models show the strong potential of using satellite data alone to predict pollen phenology with fair accuracies. Predicting an entire bi-modal test year was found to be more challenging than the randomized training and testing RF model that included training data in all 4 years. This indicates that for forecasting, or hindcasting applications, there will be some loss in prediction accuracy from the uniqueness of individual years that the RF model has not been trained with.

In Australia there is high year to year climate variability and many years with extreme climate events. To a certain extent, such years may not be accurately predicted through the developed RF models, as it may not recognize an odd or unique year or extreme event, since it only can learn from the data it has been trained with. However, in years with unique and extreme events, the EVI signal will be able to capture the grass response to the unique climate, and thus a certain amount will be reflected in the EVI signal. It remains to be seen to what extent RF models can utilize the EVI response signal to unique years to predict, in turn, the resulting pollen concentrations. In such cases, RF may be able to accommodate some aspects of the unique year, through the EVI signal.

Other aspects needing further attention are the extent to which EVI signals can be used to forecast pollen amounts 1 month, 2 weeks, or 1 week in advance. The lag analysis conducted here yielded mixed and inconclusive results, suggesting separate model runs to ascertain this objective. The distance variable importance also yielded mixed results in that the 100km distance class was as important as the 25km distance class, which was surprising in that grass pollen is not known to be able to travel such large distances to the pollen trap. One possible explanation is that the 100km grass areas, may contain, and thus mimic the same grass species and identical phenology as found in the 11-25 km distance class. The 0-10km distance class was ranked lower in importance due to the large extent of urbanized area.

DOY was the most important variable in both RF models. Basically, the RF models learned the most from DOY by recognizing the annual repetitive patterns of the grass growing season through knowledge of the calendar days. When we ran the same RF model with a missing test year, and without MODIS satellite data, the predictive model explained 50% of the variance (r=0.7). The MODIS data allowed this to be better fine-tuned. The residuals or peaks and troughs in pollen concentrations that were not well modelled are revealing in that it suggests that additional, non-satellite input variables are lacking, most likely meteorological variables, such as wind, temperature, and rainfall. Although much of the climate information may be reflected in the EVI signal, other factors such wind will not be present in the EVI signal yet have significant bearing on the transport of the pollen from source to receptor, pollen trap location.

In the Devadas et al. study [4], generalized additive models (GAMs) were used to predict pollen concentrations from the same Campbelltown pollen trap site. In that study only the Austral spring pollen season was modelled, and approximately 74-79% of the variance in a missing test season could be predicted. In this study our RF models predicted 63 to 85% of the variance but we accommodated two pollen seasons, which to our knowledge is the first attempt at predicting bi-modal pollen years.

In conclusion, our results from the machine-learning analysis conducted here show considerable promise in effectively predicting seasonal pollen activity with key landscape ecological information. The machine-learning approaches were able to effectively consider complex geospatial ecological variables to couple satellite data with output pollen counts., and without the need for meteorology. However, to refine pollen predictions to daily timesteps, the meteorology becomes the key missing information. Future analysis involves integration of satellite ecology with meteorology to further advance pollen forecasting.

The impacts of this research will guide public health strategy and policy on changing pollen exposure conditions, provide better aeroallergen management tools, and improve forecasting of allergenic pollen to potentially reduce the health and socio-economic burden of grass-pollen induced allergies.

5. REFERENCES

[1] L. Ziska, et al., “Recent warming by latitude associated with increased length of ragweed pollen season in central North America”. *Proc. Natl. Acad. Sci. U. S. A.* 108:4248–4251, 2011.

[2] P.J. Beggs, C.H. Katelaris, D. Medek, et al.,”Differences in grass pollen allergen exposure across Australia”. *Aust. N. Z. J. Public Health* 39:51–55, 2015.

[3] J.M. Davies, P.J. Beggs, D.E. Medek, et al. “Trans-disciplinary research in synthesis of grass pollen aerobiology and its importance for respiratory health in Australasia”. *Science of the Total Environment,* 534:85–96, 2015.

[4] R. Devadas, A. Huete, D. Vicendese, et al. “Dynamic ecological observations from satellites inform aerobiology of allergenic grass pollen”, *Science of the Total Environment,* 633, 441–451, 2018.

[5] J. Nowosad, Spatiotemporal models for predicting high pollen concentration level of Corylus, Alnus, and Betula, Int J Biometeorol (2016) 60:843–855, 2016.

[6] S.G. Haberle, D.M.J.S. Bowman, R.M. Newnham, et al. “The macroecology of airborne pollen in Australian and New Zealand urban areas”. *PLoS One*, 9, e97925, 2014.

[7] A. Huete, K. Didan, T. Miura, et al. “Overview of the radiometric and biophysical performance of the MODIS vegetation indices”. *Remote Sens. Environ*. 83:195–213, 2002.