UNIVERSITY OF BIRMINGHAM
University of Birmingham Research at Birmingham

Constructing a pollen proxy from low-cost Optical Particle Counter (OPC) data processed with Neural Networks and Random Forests

Mills, Sophie A.; Bousiotis, Dimitrios; Maya-Manzano, Jose M.; Tummon, Fiona; MacKenzie, A. Rob; Pope, Francis D.

DOI:
10.1016/j.scitotenv.2023.161969

License:
Creative Commons: Attribution (CC BY)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Mills, SA, Bousiotis, D, Maya-Manzano, JM, Tummon, F, MacKenzie, AR & Pope, FD 2023, 'Constructing a pollen proxy from low-cost Optical Particle Counter (OPC) data processed with Neural Networks and Random Forests', *Science of the Total Environment*, vol. 871, 161969. https://doi.org/10.1016/j.scitotenv.2023.161969

Link to publication on Research at Birmingham portal

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

•Users may freely distribute the URL that is used to identify this publication.
•Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
•User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
•Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 15. May. 2024

Science of the Total Environment 871 (2023) 161969

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Constructing a pollen proxy from low-cost Optical Particle Counter (OPC) data processed with Neural Networks and Random Forests

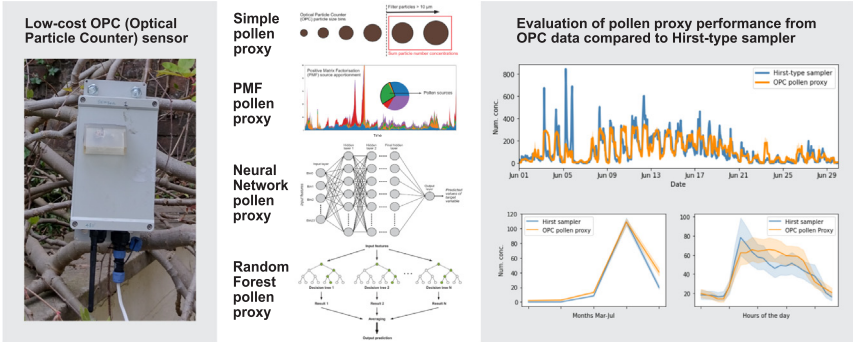
Sophie A. Mills [a,b], Dimitrios Bousiotis [a], José M. Maya-Manzano [c], Fiona Tummon [d], A. Rob MacKenzie [a,b], Francis D. Pope [a,b,*]

[a] School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham B15 2TT, UK
[b] Birmingham Institute of Forest Research, University of Birmingham, Birmingham B15 2TT, UK
[c] Centre of Allergy & Environment (ZAUM), Member of the German Centre for Lung Research (DZL), Technical University and Helmholtz Centre Munich, Munich, Germany
[d] Federal Office of Meteorology and Climatology MeteoSwiss, Payerne, Switzerland

HIGHLIGHTS

- Intercomparison campaign data from automated OPCs and Hirst-type samplers.
- Different approaches for constructing pollen proxies from OPC data evaluated.
- Neural Network and Random Forest models demonstrate promising results.
- Model-constructed proxies can detect temporal trends and high pollen events.
- Attractive low-cost, high time resolution alternative for pollen monitoring.

GRAPHICAL ABSTRACT



ARTICLE INFO

Guest Editor: Pavlos Kassomenos

Keywords:
Aerobiology
Pollen
Automatic monitoring
Optical Particle Counter (OPC)
Low-cost sensors
Machine learning

ABSTRACT

Pollen allergies affect a significant proportion of the global population, and this is expected to worsen in years to come. There is demand for the development of automated pollen monitoring systems. Low-cost Optical Particle Counters (OPCs) measure particulate matter and have attractive advantages of real-time high time resolution data and affordable costs. This study asks whether low-cost OPC sensors can be used for meaningful monitoring of airborne pollen. We employ a variety of methods, including supervised machine learning techniques, to construct pollen proxies from hourly-average OPC data and evaluate their performance, holding out 40 % of observations to test the proxies. The most successful methods are supervised machine learning Neural Network (NN) and Random Forest (RF) methods, trained from pollen concentrations collected from a Hirst-type sampler. These perform significantly better than using a simple particle size-filtered proxy or a Positive Matrix Factorisation (PMF) source apportionment pollen proxy. Twelve NN and RF models were developed to construct a pollen proxy, each varying by model type, input features and target variable. The results show that such models can construct useful information on pollen from OPC data. The best metrics achieved (Spearman correlation coefficient = 0.85, coefficient of determination = 0.67) were for the NN model constructing a *Poaceae* (grass) pollen proxy, based on particle size information, temperature, and relative humidity. Ability to distinguish high pollen events was evaluated using *F*1 Scores, a score reflecting the fraction of true positives with respect to false positives and false negatives, with promising results ($F1 \leq 0.83$). Model-constructed proxies demonstrated the ability to follow monthly and diurnal trends in pollen. We discuss the suitability of OPCs for monitoring pollen and offer advice for future progress. We demonstrate an attractive alternative for automated pollen monitoring that could provide valuable and timely information to the benefit of pollen allergy sufferers.

* Corresponding author at: School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham B15 2TT, UK.
E-mail address: F.Pope@bham.ac.uk (F.D. Pope).

http://dx.doi.org/10.1016/j.scitotenv.2023.161969
Received 21 November 2022; Received in revised form 9 January 2023; Accepted 29 January 2023
Available online 6 February 2023
0048-9697/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Pollen constitutes a significant proportion of atmospheric bioaerosols (primary biological aerosol particles, PBAP) and affects as much as 40 % of the population in industrialised countries with pollen allergies (Fröhlich-Nowoisky et al., 2016) causing health issues such as allergic rhinitis and asthma. Pollen is released from plants as part of their reproductive cycle, carrying the male gamete, and thus are vital for the survival and spread of terrestrial ecosystems. Anemophilous species - including almost all cone-bearing plants (gymnosperms) and many flower-bearing plants (angiosperms) - depend on wind to transport their pollen grains through the atmosphere where the grains can travel long distances at significant concentrations during pollen seasons (Skjøth et al., 2007, Alarcón et al., 2022, Jochner et al., 2015, Grewling et al., 2016, Siljamo et al., 2008, Manninen et al., 2014).

Individual pollen grains can vary in size between 10 and 100 μm (see, e.g., Després et al., 2012, Reponen, 2011, pp. 723, Bradley, 2015, pp. 408–409), though many anemophilous species fall towards the lower end of this range, often below 40 μm. On rupturing in the atmosphere, pollen grains can release smaller subpollen particles which can also contain allergenic proteins relevant to human health (Bacsi et al., 2006), and reside in a more easily respirable size fraction compared to the intact pollen grains (Taylor et al., 2004). Subpollen particle size ranges have been reported between 20 nm and 6.5 μm for giant ragweed (Stone et al., 2021) and <0.25–3.2 μm (aerodynamic diameter) for birch (Burkart et al., 2021). Other work reports subpollen particles containing specific pollen allergens of a similar size range of a few to several micrometres (Miguel et al., 2006). The category of subpollen particles is still loosely defined in the literature and can encompass a wide range of particles, including starch granules and other cytoplasmic debris. As with pollen grains, subpollen particles can be expected to vary in size range across different taxa and species of pollen. However, the wide range of particles included within this term likely makes the distinction between species somewhat difficult.

Pollen grains and subpollen particles have been considered to have potential cloud condensation nucleation (CCN) activity (Pope, 2010; Griffiths et al., 2012; Steiner et al., 2015; Mikhailov et al., 2019), as well as ice nucleation (IN) properties (Diehl et al., 2001, Diehl et al., 2002, Pummer et al., 2012, Tong et al., 2015, Dreischmeier et al., 2017, Gute and Abbatt, 2020, Burkart et al., 2021) which can be affected by atmospheric processing (Gute et al., 2020). This means that these particles could affect the process of water condensation or ice crystal formation in clouds, and thus precipitation and ultimately climate.

Whether in the interest of public health, ecosystem regeneration and recolonisation, or climate change, there is consensus that further focused studies of pollen, and other bioaerosols, are required (Huffman et al., 2019). Yet there are significant limitations to current methods available for studying airborne pollen dynamics. Conventional methods for measuring pollen generally use volumetric Hirst-type samplers which are time consuming and labour intensive, requiring trained staff to identify and count pollen concentrations from collected slides. Data is therefore often only available in daily, or for limited periods up to hourly, time resolution, is not available in real-time, and suffers from relatively large uncertainties, easily up to 30 % (Adamov et al., 2021). An overview of the current state of pollen monitoring stations globally can be found by Buters et al. (2018), which highlights the need for technological advancement.

Attention is now turning towards the use of automated monitoring systems, with an aim to standardise such methods to replace conventional manual techniques (Tummon et al., 2022). Recent automated pollen monitoring instruments have been developed such as PollenSense's PS-300 (Jiang et al., 2022), Hund-Wetzlars BA500 pollen monitor (Oteros et al., 2015, 2020; Plaza et al., 2022), the Rapid-E (Šaulienė et al., 2019; Smith et al., 2022), and the Swisens Poleno (Sauvageat et al., 2020, Sofiev et al., 2022). The first two use an impactor to capture grains on a slide, similarly to conventional manual traps, but the slide imaging and pollen grain classification are automated. The Rapid-E is a cytometer which identifies particles using classification algorithms based on laser scattering signal, fluorescence spectrum and fluorescence lifetime data. Meanwhile the Poleno presents a recent advancement whereby pollen grains are identified in an automated system utilising light scattering, fluorescence, and holographic imaging techniques, trained by machine learning. Fluorescence spectroscopy methods have been utilised to characterise pollen using the Wideband Integrated Bioaerosol Sensor, WIBS (e.g. O'Connor et al., 2014; Savage et al., 2017; Ruske et al., 2018), the WIBS has also been used to identify pollen fragments (Hughes et al., 2020). These and new technologies are in continuous development to meet the present need for more advanced pollen monitoring systems. A comprehensive review, including the use of various machine learning models for pollen modelling, can be found in Maya-Manzano et al. (2020) and in Buters et al. (2022).

These methods do however bring limitations of time resolution in some cases and heavy cost limitations in general. With each instrument costing thousands or tens of thousands of U.S. dollars, such methods become unavailable for the global majority and smaller institutions, let alone for personal use. It would also be infeasible to utilise such sensors in monitoring networks that could offer high spatial resolution data. While high performance sensors with high accuracy have their value, there is also a need for more affordable and portable options, that could be used in networks to provide high spatiotemporal resolution information throughout dynamic pollen seasons.

Low-cost optical particle counter (OPC) sensors have already been explored extensively to monitor particulate matter (PM) – i.e., all generic, internally and externally mixed, solid and liquid particles suspended in the atmosphere - primarily in the context of air quality and anthropogenic pollution (for example: Crilley et al., 2020; Giordano et al., 2021; Dubey et al., 2022; Narayana et al., 2022). These OPC sensors take in a continuous flow of particles from ambient air, direct laser light at the stream of particles and from the scattering pattern of the light intercepted by the particles determines, via Mie theory, counts for particles in different size range bins. Pollen, being an aerosol, should also be detectable by such methods, though it is generally larger in size than most monitored anthropogenic particulate matter (i.e. larger than PM10, which is PM with an aerodynamic diameter of 10 μm or less). There has been very limited previous research and reported success in using a laser optics monitoring system to measure pollen concentrations (Kawashima et al., 2017).

The Alphasense OPC-N3 (Alphasense, Braintree, UK) has capability to measure particles up to 40 μm in size (i.e., optically measured diameter) thereby covering a large proportion of anemophilous pollen species. Costing a few hundred U.S. dollars, it is significantly more affordable than other current pollen monitoring options. Additionally, it comes with advantages of high time resolution (down to 10 s) and flexibility on deployment and functionality. It can also be functionalised to give real-time and remotely accessible data. The challenge, however, is isolating a meaningful pollen signal from the sized particle count information provided by the OPC sensors and this is what this study addresses. Previously, our group has used the OPC-N2, the precursor to the OPC-N3 to monitor for airborne fungal spores in a forest (Baird et al., 2022).

Our aim is to assess whether the data available from low-cost OPC sensors such as these have potential to be used for monitoring pollen, to an extent that provides useful information for public health. We have taken commercially available Alphasense OPC-N3 sensors, functionalised them to log independently outdoors for extended periods, and to deliver their sized particle data online via a mobile phone network in real-time. Through the EUMETNET Autopollen ADOPT - COST Action Intercomparison Campaign in 2021 (Maya-Manzano et al., 2022), we acquired data from our sensors in parallel with baseline pollen data from Hirst-type volumetric samplers. Using these data, we investigated potential methods to determine pollen proxies from OPC data and evaluated for each method its accuracy and potential value for monitoring airborne pollen concentrations.

2. Materials and methods

2.1. Instrumentation

Three commercially manufactured Alphasense optical particle counter (OPC-N3) devices were used in this study. These are small (mass < 105

g), and cost in the range of a few hundred dollars. Their predecessor, the OPC-N2, has been described previously by Sousan et al. (2016) and Crilley et al. (2018) with the main difference being that the OPC-N3 units have an extended measurable size range up to larger particle sizes. While the OPC-N2s measure within a size range of 0.38–17 μm (divided into 16 software bins), the OPC-N3s output raw particle counts segregated into 24 bins of different particle size ranges between 0.35 and 40 μm (see Table S1 in the Supporting Information for individual bin size ranges). This means that, unlike its predecessor, its detection size range overlaps with the typical size range of many anemophilous pollen taxa, including oak (*Quercus*), birch (*Betula*), grass (*Poaceae*), nettle (*Urtica*). Though some taxa are not in this range, such as pine (*Pinus*) pollen, which are typically over 40 μm in diameter (Song et al., 2012).

The OPC device actively draws in a stream of air, at around 5 L min^{-1}, which passes through a 658 nm-wavelength laser beam that is scattered by the incoming particles. Measurements of scattered light intensity are calculated based on Mie scattering theory and used to determine number counts for particles that fall within each bin size range. A refractive index of 1.5 and particle density of 1.65 g mL^{-1} is assumed, and the devices have a maximum particle count of up to 10,000 particles s^{-1} (equivalent to number concentrations of 1.2×10^8 particles m^{-3} at a 5 L min^{-1} flow rate).

A custom-built system was developed with an Arduino MKR GSM1400 microprocessor (Arduino S.r.l., Via Andrea Appiani 25, Monza, 20,900, Italy) to facilitate the independent logging of the OPC-N3 data, both onto internal SD card memory and online via the GSM (Global System for Mobile Communications) network to a webpage where the raw data could be accessed in real-time. The lowest time resolution for the logged data in this case was 1 min, averaged from measurements set to be taken by the OPC every 10 s. Each OPC device and microprocessor-controlled circuit board system were encased inside a plastic weatherproof box and also included BME280 sensors (Bosch Sensortec GmbH, Gerhard-Kindler-Strasse 9, 72770 Reutlingen, Germany) to measure temperature and RH (relative humidity). These were positioned externally to the OPC device but inside the weatherproof box for this study. We found the measurements from the BME280 sensors to be more accurate than the temperature and relative humidity measurements supplied by OPC devices themselves, which are more influences by the heat generated by the OPC-N3 sensors and surrounding electronics.

Fig. S1 in the Supporting Information shows the time series of RH and temperature measured by the BME280 sensors against reference data, obtained from the German National Meteorological Service (DWD) at the München-City station located 7.5 km away from the sampling site. While there is a systematic error for both variables due to the positioning of the BME280 sensors inside the weatherproof box, it is evident they are consistent with the ambient variations over time.

While the OPCs output a number of different variables, including calculated PM mass concentrations, the main data collected that was relevant to this study were the raw particle number counts from each of the 24 size bins, alongside variables that would influence the bin counts, including the laser status, sample period and flow rate.

The benchmark pollen data for this study were obtained from the mean of four collocated Hirst-type samplers (all Burkard Manufacturing Co Lid, Rickmansworth, UK) provided by the coordinators of the EUMETNET AutoPollen – ADOPT COST Action (CA18226) Intercomparison Campaign 2021. Further details on this can be found in the initial overview paper (Maya-Manzano et al., 2022). These data consisted of hourly averages and included pollen number concentrations in grains/m^3 for 16 different pollen taxa: *Alnus* (alder), *Ambrosia* (ragweed), *Artemisia* (mugwort), *Betula* (birch), *Carpinus* (hornbeam), *Fagus* (beech), *Fraxinus* (ash), *Picea* (spruce), *Pinus* (pine), *Plantago* (plantain), *Poaceae* (grass), *Populus* (poplar), *Quercus* (oak), *Taxaceae Cupress* (yew), *Tilia* (lime), *Urtica* (nettle), as well as a total pollen concentration.

2.2. Context

The three OPCs ('OPC1', 'OPC2' and 'OPC3') were placed within a few metres of each other, facing the same direction, on the building roof site used for the AutoPollen Intercomparison Campaign at the Center of Allergy & Environment (ZAUM) in Munich, Germany, alongside all the other sensors participating in the campaign. The campaign ran from early March until July 2021. Hirst pollen concentrations were recorded for each hour of every day between 3rd March and 19th July. The OPCs were deployed in tandem from 9th March until 7th July, however, due to technical issues, not all the datasets were complete for the entire period. OPC1 recorded the most complete dataset, covering the whole period without significant gaps. Meanwhile, OPC2 recorded from 20th May until 7th July and OPC3 from 9th March until 29th June, with only one significant gap of under 48 h between 13th–15th March). Fig. S2 in the Supporting Information displays the data availability of each OPC.

2.3. Data pre-processing

Initial data processing was performed in RStudio using R version number 4.1.2. As standard procedure, timestamps where the laser status variable (which should be around 620 W cm^{-2}) was below 570 or above 670 W cm^{-2} were omitted, as anomalies in laser intensity could make the corresponding particle counts unreliable. This resulted in no data points being omitted for OPCs 1 and 3 and <0.001 % of data points omitted for OPC 2. Subsequently, the raw particle counts were converted to particle number concentrations (PNC) in particles m^{-3} by the following Eq. (1):

$$PNC\left(particles\ m^{-3}\right) = \frac{raw\ counts}{sample\ period(s) \times flow\ rate\left(L\ \ min^{-1}\right) \times \frac{0.001\left(m^3\ L^{-1}\right)}{60\left(s\ \ min^{-1}\right)}} \quad (1)$$

The sample period and flow rate for each timestamp was used in the calculation but the mean sample period for all OPCs during this period was 5.0 s while the mean flow rate was 5.2 L min^{-1}. OPC datasets were averaged over each hour for comparison with the hourly manual pollen data.

2.4. Simple (large particle) pollen proxy

The first method trialled as a pollen proxy, which we refer to here as the 'Simple Pollen Proxy', used the sum of all detected particles in size bins above 10 μm. Pollen grains are generally greater in diameter than 10 μm (Reponen, 2011, pp. 723, Bradley, 2015, pp. 408–409), which means they are also larger than most reported PM. It was observed that the OPC bins between 10 and 40 μm generally showed very little or no activity at normal times when deployed away from sources of coarse dust (e.g., quarries, construction sites, ocean spray, etc). We hypothesised that, when the sensors were placed in the vicinity of significant pollen sources during an active pollen season, particle concentrations present in this particular size range would be dominated by pollen. Thus the total number concentrations greater in diameter than 10 μm could be taken as a proxy for pollen.

2.5. PMF pollen proxy

The second method was a Positive Matrix Factorisation (PMF) source apportionment technique used to isolate a signal for pollen from all OPC bins. This is a multivariate data analysis method developed by Paatero and Tapper (1994), which attempts to find patterns in the variables of a dataset. It then assigns these patterns into factors according to their unique features and provides a relative contribution for each one of them for each timestep of the dataset. This technique is generally used to isolate specific pollutant sources from environmental pollution data that is collectively comprised of background and other various component sources (e.g., Sun et al., 2020). We hypothesised that pollen could be isolated in OPC time series via this method, as has been demonstrated previously for typical pollutant sources (Bousiotis et al., 2022).

This PMF method took the particle counts of the 24 bins as input and output a corresponding time series of the relative contributions of five

different factors at each timestamp. Five factors were chosen on the basis of previous experience with atmospheric aerosol samples. In theory, each factor should equate to a particular source of aerosols picked up by the sensors. For each factor, size distribution information is also available which can be used to identify which factor corresponds with which source. We investigated correlations of each of the factors with total pollen and individual taxa concentrations, and also used the particle size distributions to inform our decision on which factors were most likely to be associated with pollen. Two factors were selected for each OPC dataset that demonstrated the highest correlations with various pollen species and the 'PMF Pollen Proxy' was calculated from the sum of these two factors.

2.6. Supervised machine learning methods

Two further methods – Neural Network (NN) and Random Forest (RF) - were employed in parallel, with identical data for training and testing, in an attempt to construct a pollen proxy from the OPC data. These methods were implemented using Python (version 3.9.7, Anaconda distribution) in Jupyter Notebook.

Both NN and RF are supervised learning methods where a target variable from reference data is used to train and evaluate the model. This target variable was chosen from the Hirst data. Due to the limitations of the OPCs, measuring up to just 40 μm, it was considered that the data may not effectively capture all taxa included within the Hirst 'total pollen' variable. For the purposes of this study, two different target variables were selected to train and evaluate the models: total pollen and *Poaceae* (grass) pollen. The total pollen target variable was the sum of all pollen taxa concentrations collected by the Hirst. The *Poaceae* taxon was chosen here as the Hirst data showed a substantial active season that the OPCs could have a good chance of detecting, and provide sufficient information for the models to learn from. It is also generally ubiquitous and affects many allergy sufferers.

Both the NN and RF require an appropriate selection of input variables or features for the model to extract information from for the learning process. Here we prepared three sets of input variables for the models: number concentrations (particles m^{-3}) from all 24 OPC bins, all 24 OPC bin number concentrations plus meteorological variables relative humidity (RH) (%) and temperature (°C) from the BME280 sensor, and finally, just meteorological variables RH and temperature. As mentioned previously, these values from the BME280 sensor do have a systematic error from ambient conditions, however the variation over time is consistent with ambient conditions (see Fig. S1 in the Supporting Information), and so should serve the purpose sufficiently for these models.

The data for the input features were collated from all three OPC-N3s present at the campaign and split into 'train' and 'test' datasets simultaneously with the target variable using the train_test_split function from the Scikit-learn library (Pedregosa et al., 2011). Data was shuffled before being split and 40 % of all the data (6220 data points) was taken for the test dataset. (Other split percentages were trialled but this was selected as optimal for this dataset based on convergence of NN training and validation learning curves.) The same train and test datasets were used for all models of both methods. Before implementing the Neural Network method, the input feature train and test datasets were normalised using Scikit-learn MinMaxScaler function, with the scaler fit to the train dataset so the test dataset was kept 'unseen'. This was not considered necessary for the Random Forest method.

2.6.1. Neural Network method

The Neural Network (NN) model was a simple multilayer feedforward Neural Network (Svozil et al., 1997; Sazli, 2006) constructed using a Sequential model from the TensorFlow library (Abadi et al., 2015). The Sequential model structure facilitates the construction of linear stacks of layers, each with a chosen number of nodes and an activation function through which the data passes in the training process. The models used here all had five Dense (fully connected) hidden layers, each with 5 times as many nodes as input features (i.e. 130 or 120, with or without meteorological variables) and a ReLU (Rectified Linear Unit) activation function. After each hidden layer a Dropout layer was implemented with a frequency rate of 0.2 (i.e. 20 % of nodes randomly dropped in each layer) to reduce overfitting and improve generalisation. Each hidden layer had a kernel constraint, often used in combination with Dropout layers to manage overfitting, constraining the max norm of incident weights to less than or equal to 1. A Dense single node output layer, also with a ReLU activation function, followed the last hidden layer to produce the output. A generalised model, simplified without visible dropout layers, can be found in Fig. S4 in the Supporting Information.

Models were compiled with the Adam optimiser and mean squared error (MSE) loss metric and fit on the training data subset, using the test data subset for validation, with a batch size of 128. Early stopping was implemented so the training process stopped automatically when the loss in MSE had not improved any further within a given patience parameter of 100 epochs. A model checkpoint saved subsequent models of best MSE so the last to be saved was that which reached the lowest MSE, rather than the final one wherever training stopped. Training generally stopped between 400 and 800 epochs (runs through the whole training dataset).

For each model, training and validation MSE loss curves were plotted to assess the model learning process. Root mean squared error (RMSE), mean absolute error (MAE), explained variance (R^2) and Spearman's rank correlation coefficient (ρ) values were calculated between the predicted target values from the test dataset and the corresponding real target values as metrics to evaluate model performance. Relative root mean squared (RRMSE) error - RMSE divided by the mean of the reference target variable - was also calculated.

2.6.2. Random Forest method

The Random Forest model was implemented using the RandomForestRegressor model from the Scikit-learn library. A generalised diagram of the random forest framework used for regression can be found in Fig. S5 in the Supporting Information. An initial grid search cross validation was used to trial different values for the number of estimators (number of decision trees) and max depth (the number of splits allowed in each decision tree). For the final models, the number of estimators used was 600, the max depth was 10 and the maximum number of features (to be used for each tree) was the square root of the total number of input features. Similarly, RMSE, RRMSE, MAE, R^2 and ρ were calculated between predicted test target values and the real target values to evaluate model performance on unseen data.

The RandomForestRegressor model from the Scikit-learn library enables calculation of feature importance score for each of the variables input into the model. These scores are measured by Gini Importance, or, Mean Decrease in Gini Impurity (MDG) (Breiman, 1984; Menze et al., 2009). The Gini Impurity function determines how useful a node in a decision tree, with a given split rule based on a particular feature, was at separating the observations being passed through. A higher decrease in Gini impurity, which ranges between 0 and 0.5, means that node was more useful. Decrease in Gini impurity values are averaged across all nodes using the same feature, producing MDG, which measures the contribution of each feature across the whole Random Forest in separating the observations according to the labels or target variable. The higher the MDG importance score, the greater effect or predictive power this feature has over the Random Forest model.

This technique was applied to all Random Forest models after training to ascertain the relative influences of each of the OPC size bins, also RH and temperature where applicable, on the predictions output from the model. Due to their more complex nature, this is process is not as easily applied Neural Network models and so such scores were not calculated here. To accompany this, we also calculated Spearman correlation coefficients between each of the input features and target variables. This does not provide a true measure of feature importance on each model since the machine learning methods can learn from more complex variable relationships. However, we still considered it useful to investigate since it can provide

information the feature importances do not, i.e. whether the correlation between input feature and target variable is positive or negative.

2.6.3. Evaluating the spatial footprint of pollen vs pollen proxy

As a further test of the models' ability to accurately identify pollen, wind direction (°) and wind speed (m/s) measurements, also from the DWD München-City meteorology station, were used to assess the spatial footprint of the pollen detected in this campaign. Using the R Openair package (Carslaw and Ropkins, 2012), bivariate polar plots showing pollen concentration by wind direction and speed were constructed for both Hirst and OPC model concentrations. From these plots we can ascertain the direction and approximate distance of the particle sources relative to the sampling site. By comparing Hirst and model concentration plots, we can gauge how accurately the models are identifying particles from the same source of pollen.

2.6.4. Evaluating model ability to distinguish high pollen events

Rather than producing highly accurate pollen concentrations, as is the aim with instruments specifically designed to measure bioaerosols, we hope to be able to use the low-cost OPCs to produce a sufficiently accurate and precise estimate for pollen concentrations making it possible to determine when high pollen events are occurring. Simple information such as this could be useful for allergy sufferers, if suitable thresholds for 'high' concentrations are set.

We evaluated the models using further metrics which test their ability to distinguish high pollen concentration events. The threshold for high *Poaceae* pollen events was set at 50 grains m^{-3}, following standards stated by the UK Met Office (https://www.metoffice.gov.uk/weather/warnings-and-advice/seasonal-advice/health-wellbeing/pollen/what-is-the-pollen-count, Jan 2023), and the threshold was scaled up for total pollen by the ratio of the 99th percentiles. The manual and model-constructed pollen proxy test datasets were split into Positive or Negative categories - equal to/greater than or less than the threshold value respectively. For each data point it was determined whether the constructed pollen proxy was a True Positive (TP), False Positive (FP), True Negative (TN) or False Negative (FN) compared to the manual baseline (see Figs. S5 and S6 in the Supporting Information for more detail). The precision, recall and F1 score (the harmonic mean of precision and recall) were calculated from Eqs. 2, 3 and 4. The closer the F1 score to 1, the better the ability of the model to distinguish successfully high pollen events from the OPC data.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1 score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{4}$$

3. Results and discussion

3.1. Simple pollen proxy

Fig. 1 shows a comparison of the Simple Pollen Proxy for all sensors, engineered from OPC bins >10 μm in particle diameter, in orange against the manual total pollen concentration in blue. To facilitate visual comparison, the Simple Pollen Proxy values have been scaled up by the ratio of the 99th percentiles of the two variables (7.1). During the summer months of June and July the Simple Pollen Proxy suggests a greater frequency of high particle concentration events, coinciding approximately with the *Poaceae* (grass) and *Urtica* (nettle) pollen seasons. However, the resemblance is not close enough to draw significant conclusions. The Spearman correlation coefficients between the Simple Pollen Proxy and each of the total pollen and *Poaceae* pollen target variables were negligible (between −0.1 and 0.1).

While most of the time the Simple Pollen Proxy concentrations are relatively small or equal to zero, there are peaks of very high concentrations which do not appear to coincide with the manual pollen counts. It is possible that some of these peaks are not indicative of pollen grains but rather of other non-biological particles or meteorological conditions at the time such as precipitation, as this method has no way to distinguish from other particles in this size range. Some peaks, particularly in late June and in July, coincide with rainfall events (see Fig. S3 in the Supporting Information), suggesting that rain droplets may have been detected in this size range by the OPCs. Maya-Manzano et al. (2022) reported that some construction work was taking place near the measurement site during the campaign period, which may have affected sensors such as the OPCs. The concomitant occurrence of anthropogenic air pollutants, mineral dust and fungal spores with pollen has been studied by Grewling et al. (2019) and demonstrates the possibility of such particles interfering. Unfortunately, however, we lacked the means to investigate further here.

The significant lack of non-zero concentrations in this size range suggests that the OPC sensors may in fact struggle to capture larger sized particles from ambient air where the sources are varied and potentially some distance away. Larger particles typically are not transported over as long distances as smaller particles, and do not follow air streamlines around sensor inlets (e.g., Hinds, 1999), which can cause sample losses ahead of the detection laser.

3.2. PMF pollen proxy

Fig. 2 shows in orange the PMF Pollen Proxy constructed from all OPC bins of all sensors against the blue Hirst total pollen observations. Again,

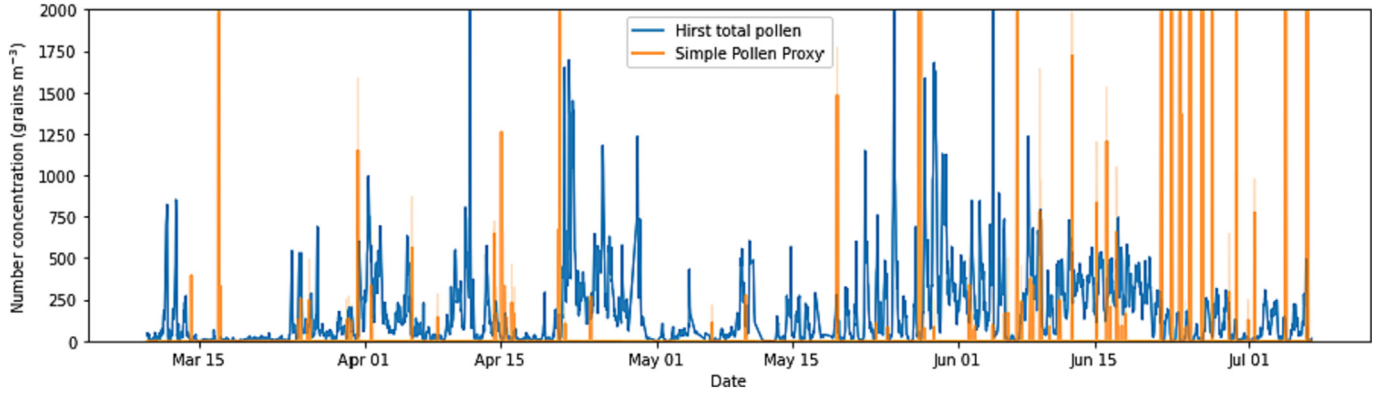


Fig. 1. Time series for Simple Pollen Proxy over the whole campaign duration. Simple Pollen Proxy concentrations from OPC data in orange and total pollen concentrations from Hirst data in blue. The Simple Pollen Proxy has been scaled up by the ratio of the 99th percentile of the two variables to facilitate comparison.

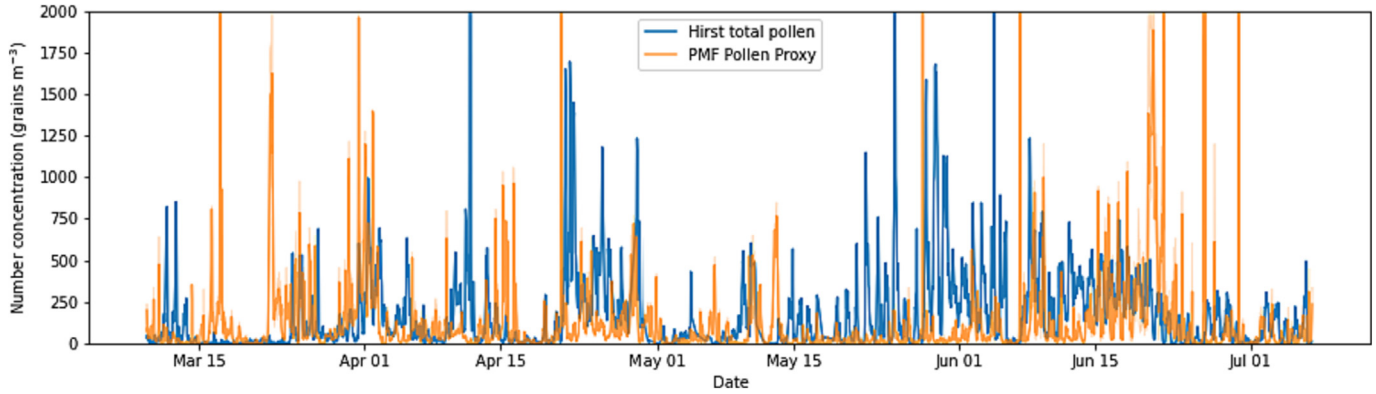


Fig. 2. *Time series for PMF Pollen Proxy over the whole campaign duration.* PMF Pollen Proxy constructed from OPC bin data in blue and Hirst data total pollen in orange. PMF Pollen Proxy values have been scaled up by the ratio of the 99th percentiles of the two variables to facilitate comparison.

the PMF Pollen Proxy values have been scaled up by the ratio of the 99th percentiles of the two variables (51.8) to facilitate comparison. Visually, the PMF Pollen Proxy appears to demonstrate closer resemblance to the manual total pollen counts than the Simple Pollen Proxy. Some activity also coincides with the *Taxaceae Cupressus* (March-early April), *Fraxinus* (April), *Betula* (April), *Quercus* (May), *Poaceae* (June), *Urtica* (June–July) pollen seasons, suggesting that these taxa may be better detected by the OPCs. The PMF Pollen Proxy achieved Spearman correlation coefficients of 0.18 and 0.10 with the target variables total pollen and *Poaceae* respectively.

While the PMF Pollen Proxy, with its more sophisticated processing of the data and inclusion of all bins, performed better than the Simple Pollen Proxy, it still does not provide a significantly accurate indicator for pollen concentrations when compared with the manual pollen measurements. Correlations were also explored between the PMF Pollen Proxy and individual pollen taxa, with certain taxa exhibiting higher correlations with the proxy than other taxa. This suggests that some taxa were more successfully detected by the OPC instruments, due to their grain size and higher concentrations.

A possibility to improve this method further would be to attempt resolving a greater number of factors from the data, which might better isolate some pollen sources from other particle sources. However, this has not been taken further within the scope of this study.

3.3. Neural Network and Random Forest methods

3.3.1. Input feature correlations

The Spearman correlation coefficients between input and target variables are displayed as a bar chart at the top of Fig. 3. Below this, also in Fig. 3, are the Random Forest model feature importances – those with only OPC bin input features above and those including meteorological variables below. These plots provide an idea of which input features have more influence on the model learning process. It may seem surprising that the larger size bins, equivalent to the size of intact pollen grains, do not show significant correlation with the target variables. However, as discussed in Section 3.1, it seems that the OPCs struggle to detect significant particle counts in this range and this could be the reason for the negligible correlations.

Meanwhile, the most positively correlated OPC bins with the target variables are bins 7 and 8, corresponding to a size range of 3.0–5.2 μm. This is a possible size range for pollen fragments and some subpollen particles judging by the measurements of a few different species by Taylor et al. (2004), Miguel et al. (2006), Stone et al. (2021) and Burkart et al. (2021). Interestingly, we then see negative correlations for the smaller size ranges, especially bins 1 and 2 which correspond to 0.46–1.0 μm. It should be noted that the particle size ranges determined by the OPC optical Mie scattering measurements may not correspond absolutely with accurate geometric diameters, but it should be consistent across OPC-N3 instruments in general. Nevertheless, we conclude from this that the models in fact may not be learning primarily from the signal of intact pollen grains. Instead, the models are likely learning from the particle size information primarily associated with pollen fragments and subpollen particles released from ruptured pollen grains. We assume here that such particles coming directly from pollen grains are present simultaneously with pollen in the atmosphere. Yet such particles would be far more numerous than pollen grains (Stone et al., 2021), while likely suspended for increased time periods and transported greater distances compared to intact pollen grains. They would likely be much more readily detected by our OPC instruments. Other particles such as fungal spores and anthropogenic particles may also be present but, since the machine learning methods are supervised using Hirst target data, pollen-correlated particles should be distinguished from these. Thus, we assume that this information can serve as a useful pollen proxy.

Both correlations and calculated importances for meteorological variables are more significant than those of each of the particle bins. RH shows negative correlation while temperature has a positive correlation with pollen concentrations. This is likely a result of diurnal variation where temperature peaks during the day and RH during the night in general, as pollen concentrations tend to be higher in the daytime. The positive temperature correlation could also be due to the onset of significant pollen seasons, including *Pinus*, *Poaceae* and *Urtica*, towards the later summer months, or a more general correlation between temperature and pollen season start (Van Vliet et al., 2002). Negative correlation with RH could also be a result of precipitation occurring when humidity is high, as this could effectively wash airborne pollen grains out of the atmosphere and reduce concentrations. Pollen grains are hygroscopic (Pope, 2010) and therefore swell at high humidity which could also cause the grains to settle more quickly under gravity even without precipitation.

3.3.2. Neural Network and Random Forest models: overview

In total, 8 models were trained and can be categorised by method (Neural Network/Random Forest), input features (with/without RH and temperature), and target variable (*Poaceae*/total pollen). Fig. 4 displays the time series for each of these models for the duration of the campaign, with the reference target variable in blue and the model-constructed proxy in orange. This figure includes only those models which included meteorological input features, whereas an alternative figure also displaying the models without meteorological input features can be found in Fig. S6 in the Supporting Information. Table 1 lists all the models, their defining categories, and the metrics - RMSE, RRMSE, MAE, ρ, R^2 and F1 score - attained by each for the test datasets.

The time series and metrics tables for further models, trained on just RH and temperature input features, can be found in Figs. S7, Fig. S8, Table S2 and Table S3 in the Supporting Information.

From the time series plots in Fig. 4, it is evident that all NN and RF model-constructed proxies performed significantly better than the Simple

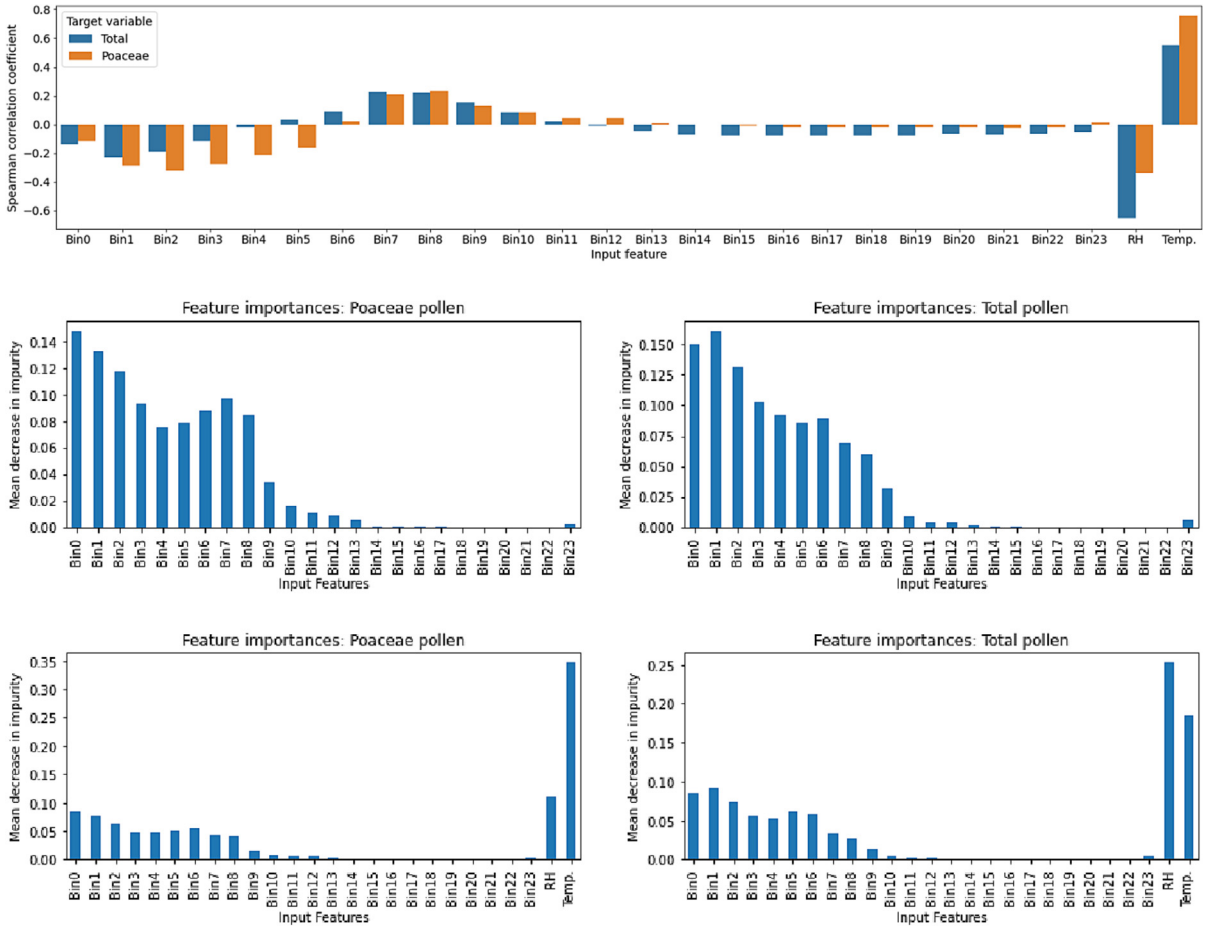


Fig. 3. Bar plots of correlations between OPC input features and Hirst target variables and Random Forest feature importances. Top: Spearman correlation coefficient values of each input feature (24 OPC bins, RH and temperature) with each target variable (Poaceae and total pollen). Middle: Feature importances of Random Forest models with only OPC bins. Bottom: Feature importances of Random Forest models with relative humidity (RH) and temperature (Temp.) input features as well as 24 OPC bins. Corresponding particle size ranges for each OPC bin can be found in Table S1 in the Supporting Information.

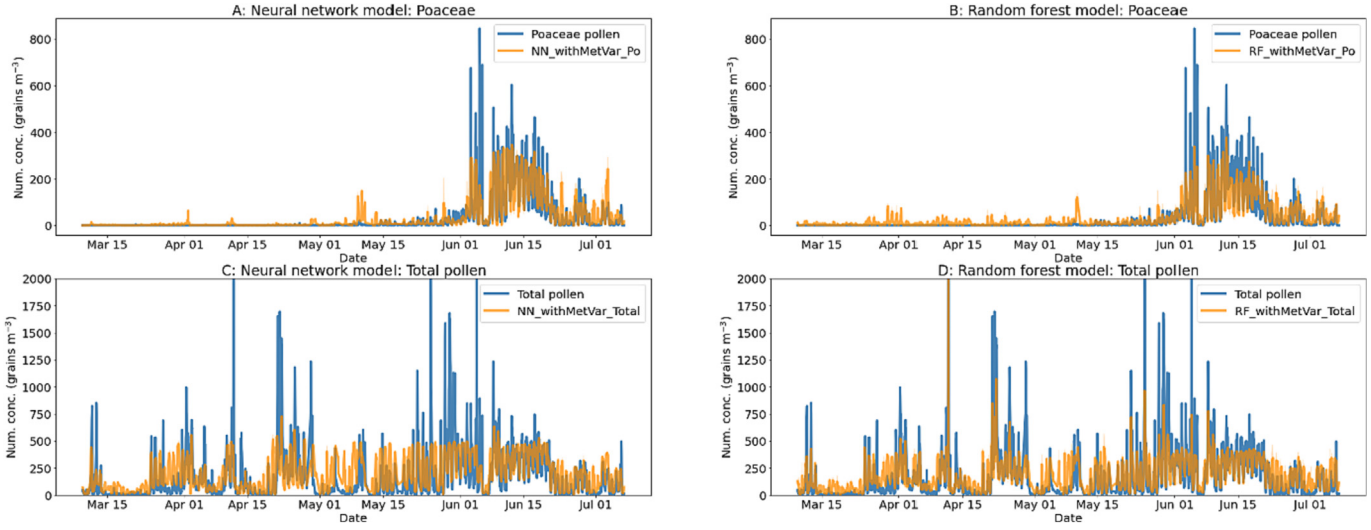


Fig. 4. Time series for NN and RF pollen proxies over the whole campaign duration. Hirst data target variable are in blue and model predictions for models with RH and temperature as input features are in orange. Plots on the left (A and C) are from Neural Network models while those on the right (B and D) are from Random Forest models. A and B are predictions for Poaceae pollen and C and D for total pollen number concentrations in grains m^{-3}.

Table 1
Summary of Neural Network (NN) and Random Forest (RF) models and evaluation metrics. The 8 supervised machine learning models tested to predict pollen concentrations from OPC data, differentiated by model method (NN /RF), input features (with/without meteorological variables RH and temperature) and target variable (Poaceae/total pollen). The best NN and RF results have been emboldened.

Model name	Method	Input features	Target variable	RMSE (grains m^{-3})	RRMSE	MAE (grains m^{-3})	ρ	R^2	F1 Score
NN_Po	Neural Network	24 OPC bins	*Poaceae*	69.5	1.7	34.2	0.62	0.34	0.65
NN_Total	Neural Network	24 OPC bins	Total pollen	240.3	1.3	156.1	0.54	0.15	0.65
NN_withMetVar_Po	**Neural Network**	**24 OPC bins + RH + T**	***Poaceae***	**49.1**	**1.2**	**20.7**	**0.85**	**0.67**	**0.83**
NN_withMetVar_Total	Neural Network	24 OPC bins + RH + T	Total pollen	215.8	1.1	118.0	0.77	0.3	0.78
RF_Po	Random Forest	24 OPC bins	*Poaceae*	71.8	1.8	39.4	0.61	0.29	0.60
RF_Total	Random Forest	24 OPC bins	Total pollen	232.1	1.2	146.5	0.55	0.19	0.67
RF_withMetVar_Po	**Random Forest**	**24 OPC bins + RH + T**	***Poaceae***	**54.3**	**1.4**	**26.5**	**0.79**	**0.59**	**0.80**
RF_withMetVar_Total	Random Forest	24 OPC bins + RH + T	Total	208.5	1.1	113.6	0.78	0.34	0.78

Pollen Proxy and PMF Pollen Proxy, and pollen season variations are to some extent successfully reconstructed. There are false positive signals occurring at times when the manual baseline shows low or zero pollen concentrations, for example *Poaceae* pollen being predicted outside of its season. However, this was observed for most instruments that took part in the campaign (Maya-Manzano et al., 2022). Particularly for the OPCs, non-biological particles - or other biological particles - which are of a similar size to the particles of interest may easily be confused with them causing such false positives. Many different pollen taxa overlap in size ranges, and so may limit the ability of the model to distinguish individual species with the data available. The model-constructed proxies in general appear to underestimate when very high pollen concentrations are observed in the manual counts. This may suggest greater physical limitations for the particles being sampled through the inlet and detected by the OPC than for the Hirst-type sampler.

3.3.3. Difference in model performance due to input variables

The most apparent difference observed among the models is that those with RH and temperature added as input features perform significantly better than those without. This can be seen in Fig. S6 in the Supporting Information. The metrics (see Table 2a) support this where the models without RH and temperature have average RMSE, RRMSE, ρ and R^2 values of 153 grains m^{-3}, 1.5, 0.58 and 0.24 while the models with these features have values of 132 grains m^{-3}, 1.2, 0.80 and 0.48 respectively. The F1 score is also improved from 0.64 to 0.80 demonstrating that RH and temperature significantly add to the accuracy of the models. The likely reasons for the effect of RH and temperature as input features were discussed in Section 3.3.1.

Models trained only on meteorological input features RH and temperature (results shown in Figs. S7 & S8 and Tables S2 & S3 in the Supporting Information) scored higher metrics than models with only particle size information, yet significantly below models with both particle size and meteorological information. This demonstrates, as the input feature correlations in Fig. 3 suggested, that RH and temperature add valuable predictive power to the models. Meanwhile it proves that the combination of both particle size and meteorological information produces the most accurate results.

Table 2
Mean-averaged metrics for models by category. Categorised by input feature in top table A, by model type in middle table B, and by target variable in bottom table C.

A) Input features	RMSE	RRMSE	MAE	ρ	R^2	F1 score
OPC bins	153.4	1.49	94.1	0.58	0.24	0.64
OPC bins + RH + T	131.9	1.20	69.7	0.80	0.48	0.80
B) Model type	**RMSE**	**RRMSE**	**MAE**	**ρ**	**R^2**	**F1 score**
NN	143.7	1.33	82.3	0.70	0.37	0.73
RF	141.7	1.36	81.5	0.68	0.35	0.71
C) Target variable	**RMSE**	**RRMSE**	**MAE**	**ρ**	**R^2**	**F1 score**
Poaceae	61.2	1.52	30.2	0.72	0.47	0.72
Total pollen	224.2	1.17	133.6	0.66	0.25	0.72

3.3.4. Difference in model performance due to model type

The difference between the Neural Network and Random Forest methods is less distinct. From visual inspection, it appears that the Random Forest models adhere a little closer to the Hirst benchmark and may capture more detail of individual peak events. The metrics (see Table 2b) are very close between NN and RF models with values of 144 grains m^{-3}, 1.3, 0.70, 0.37 and 0.73, and 142 grains m^{-3}, 1.4, 0.68, 0.35 and 0.71 for RMSE, RRMSE, ρ, R^2 and F1 Score values respectively.

While this demonstrates presently that both model types are similarly useful for this purpose, we note that the NN methods deployed here have the potential for further development and accuracy improvement in the future. A grid search cross validation method was performed on the RF models first to decide on optimal hyperparameters. Meanwhile the NN method is more complex with many different hyperparameters and possible architecture variations. While different variations were trialled in the process of achieving the final models presented here, this is by no means exhausted or optimised to its full potential.

3.3.5. Difference in model performance due to target variable

For the choice of target variable, the single species *Poaceae* models performed best by all metrics (see Table 2c), except for RRMSE. This is because

total pollen had significantly greater range resulting in inevitably greater error margins. The RRMSE metric however corrects for this and allows for some comparison across different variables. Total pollen models had the lower RRMSE at 1.2 compared to *Poaceae* models at 1.5. Meanwhile *Poaceae* and total pollen models achieved joint best F1 scores at 0.72.

The question of which target variable is preferable to use is not simple, and for future studies no doubt depends on the context, which taxa are present and of interest. While targeting individual taxa like *Poaceae* can result in smaller error, the limitations of the OPC's ability to distinguish between taxa based solely on size information should be kept in mind and false positives are to be expected. Including input features that correlate in some way to season (including temperature) will likely reduce this error yet make the model less generalisable to use in different locations, where seasonal patterns are different or irrelevant. A key point from this study is that using target variables at both the individual taxa and collective pollen levels are worth investigating for further studies. While in many cases the accuracy of precise number concentrations may still be lacking, the results here indicate that reducing the information conveyed to simply the occurrence of high pollen events offers better accuracy across different target variables.

3.3.6. Individual model performances

Of all 12 individual models, the best performing across most metrics was the NN_withMetVar_Po model at RMSE, MAE, ρ, R^2 values and F1 score of 49 grains m^{-3}, 21 grains m^{-3}, 0.85, 0.67 and 0.83 respectively. The RF alternative, RF_withMetVar_Po, performed similarly, also reaching an R^2 value above 0.5 and F1 score of 0.80. While difficult to compare across different target variables, the total pollen models with meteorological variables (both NN and RF) performed best in terms of RRMSE while also demonstrating high F1 scores at 0.78. Meanwhile, the error margins and R^2 scores for these and the other models would be ideal to improve upon.

In the campaign overview (Maya-Manzano et al., 2022), 9 out of 18 of the automated pollen measurement systems (including different classification algorithms for the same instrument) were reported to reach R^2 values above 0.5 for both 3 hourly and daily averaged data. These were high quality instruments, specifically designed to measure pollen with sophisticated algorithms trained directly on pollen samples. In comparison, 2 out of 8 of our models here achieved R^2 values above 0.5, at hourly time resolution. When averaged to daily resolution, all R^2 values improve significantly – 3 further models (including total pollen proxy models) are at or very close to R^2 values of 0.5. The best models – *Poaceae* pollen proxies with meteorological variables – reach R^2 values of 0.75 and 0.86.

It is however interesting that the automated instruments in the intercomparison campaign in general achieved the lowest metrics for *Poaceae* pollen compared to other taxa (with 6 out of 18 systems achieving $R^2 >$ 0.75 for daily resolution data) while we achieved good results for our *Poaceae* models. The comparatively worse performance for *Poaceae* in the campaign overview (Maya-Manzano et al., 2022) was thought to be due to greater variation in size among different species of the Poaceae family (Frenguelli et al., 2010), which could also be expected to affect the ability of the OPC and machine learning algorithm to distinguish this taxon. Nevertheless, we did not study models targeting other taxa here, largely due to considered data limitations for model training, and it is possible they may be able to perform even better given ample training data.

Fig. 5 shows polar plots - using the R Openair package (Carslaw and Ropkins, 2012) - of pollen concentrations plotted according to wind direction and speed, which provides further information on pollen sources. The

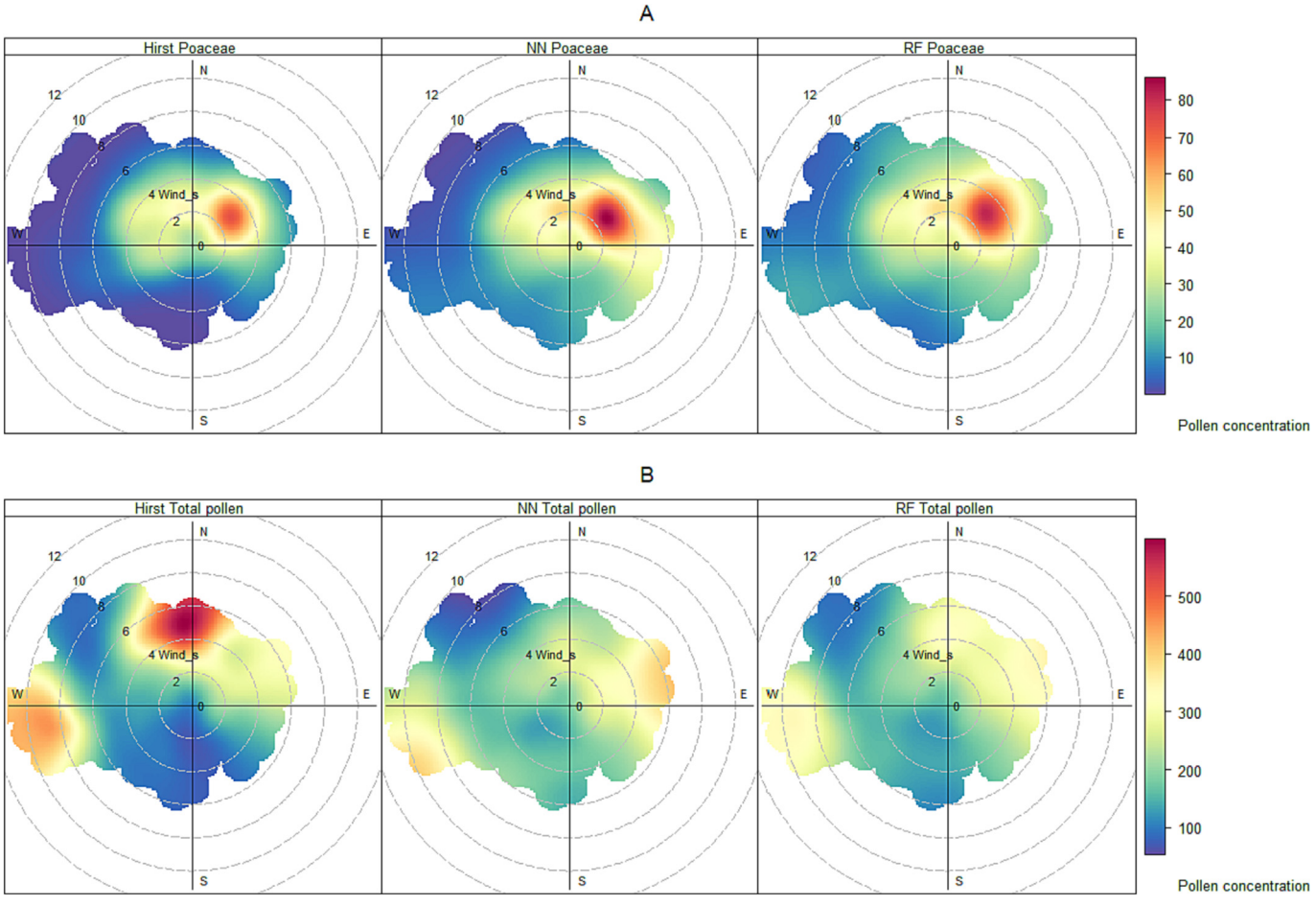


Fig. 5. Polar plots of detected pollen source concentrations according to wind direction and speed. Data for A) Poaceae and B) total pollen for each of the Hirst measurements, NN model pollen proxy and RF model pollen proxy from left to right (models are those which included RH and temperature input variables).

meteorological data here was sourced from a weather station situated 7.5 km away from the intercomparison campaign site and so relates to regional-scale flow rather than local eddies and currents. Hirst, NN model and RF model (only those which included temperature and RH input variables) pollen concentrations are presented (left to right) for each target variable (top to bottom).

The Poaceae models appear to have isolated the same source as the Hirst for this pollen taxa with remarkable success, strengthening the evidence that the models are effective at constructing a pollen proxy. The total pollen models appear less accurate in terms of absolute pollen concentrations, but the source areas are generally consistent with the Hirst. There seems to be a source from the North that the Hirst has collected at high concentrations but the OPCs have been less effective at detecting. This could be due to pollen taxa which are out of the measurable size range of the OPC, likely *Pinus* in the context of this campaign, in the case of total pollen. It may also be that sources brought by the wind from the direction the OPC inlet was facing were detected more efficiently and this may be important to consider for future work.

3.3.7. Monthly and diurnal variation

To answer the question of whether these model-constructed proxies can give meaningful information about real-life pollen trends, we compared monthly and diurnal variations between proxy and manual benchmark. Fig. 6 displays plots of the averaged number concentrations by month on the left and by hour of the day on the right for each target variable – *Poaceae* and total pollen from top to bottom. In each plot, the blue line shows the target variable reference while the other colours show the other models for each model type with and without meteorological input features.

All models adhere approximately to the monthly and diurnal trends (for example increasing in daylight hours), even those without meteorological variable inputs. This demonstrates that the model does not simply rely on RH and temperature to create these trends, but that the particle concentrations also provide useful information. Nevertheless, as proven by the time series and calculated metrics, those models with meteorological variables perform markedly better. These models in fact seem to even overestimate pollen concentrations at the height of daylight hours while the manual concentrations lie in between the model-constructed proxies with and without meteorological variables.

Furthermore, as seen in Fig. 5, there is some evidence that the RF models may be better at detecting some details in temporal variation. For example, the total pollen baseline counts show a small peak in the evening after 8 pm and the RF models follow this noticeably better than the NN models.

3.3.8. Implications for public health

The utility of the methods tested here for public health rely on their fidelity in identifying high pollen events; that is, being able to detect high pollen events with a minimum of false positives and false negatives. Detailed results from the high pollen event thresholding test can be found from Figs. S9 and S10 in the Supporting Information. Fig. S10 shows the percentage of datapoints belonging to each group of the confusion matrix, i.e., true positive, false positive, true negative and false negative. NN and

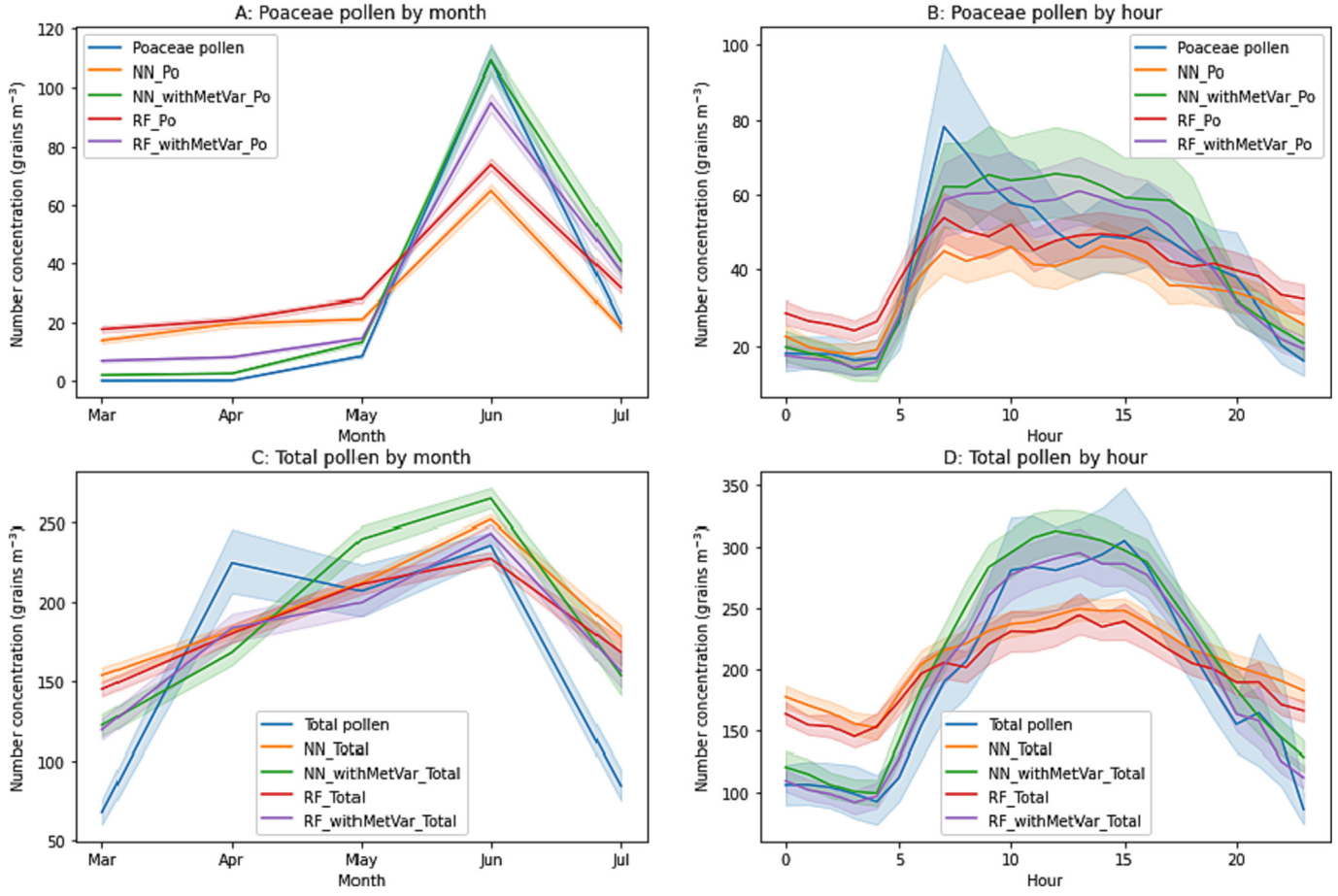


Fig. 6. Monthly and diurnal trends for model-constructed pollen proxies compared to Hirst baseline. Mean-averaged pollen concentrations by month, from March to July, on the left (A and C) and by hour of the day on the right (B and D) for each of the NN and RF models. Top (A and B): Poaceae models; Bottom (C and D): Total pollen models. Blue lines denote the target variable from Hirst data, green and orange lines denote NN models with and without meteorological input variables respectively, purple and red lines denote RF models with and without meteorological input variables.

RF models perform very similarly with respect to true positives; the better performance of the RF model is due mostly to better discrimination between true negatives and false positives (Fig. S10).

The false negative category for all models is below 5 % of all data points (except for the Neural Network *Poaceae* model without meteorological input features which was 6 %). From a public health perspective, this false negative category is the most import to keep minimal. As an instrument that informs allergy sufferers whether to anticipate severe symptoms on going outside, the false positive category would result in an inconvenient false alarm. Meanwhile, the false negative category could result in people at risk setting out when they believe it safe and could potentially result in more severe consequences including hospitalisations. Therefore, with a false negative rate generally under 5 %, we suggest that the capability of the OPC sensors to determine high pollen events above a set threshold is sufficient and can provide valuable information for public health.

4. Conclusions

In this study we have explored four different potential methods to monitor pollen using low-cost OPC instruments, and we have made progress in constructing useful pollen proxies from OPC data. This step towards developing affordable pollen monitoring techniques and forecasting capabilities builds upon that already produced from the EUMETNET AutoPollen ADOPT - COST Action Intercomparison campaign which brought together various automated pollen sensors to facilitate further advancement in this area.

This is, to our knowledge, the first study implementing such techniques to construct a pollen proxy from low-cost OPC sensors conventionally used to measure particulate matter. While inferior in terms of accuracy to other instruments specifically developed to measure pollen, such low-cost sensors can, when coupled with suitable machine learning algorithms, estimate pollen concentrations to a useful degree of accuracy. They would thus supply an attractive alternative that has the potential provide automated, high temporal and spatial resolution data with fleets of sensors deployed in networks. This has not been possible before due to the high cost and labour demands of conventional manual instruments.

We presented here our investigation into finding and demonstrating an appropriate method of processing OPC general particle size data to gain useful information on airborne pollen concentrations. We show that NN and RF methods demonstrate the most success and can predict pollen concentrations to R^2 values reaching above 0.5 when compared to a Hirst-type sampler as a baseline, when meteorological variables RH and temperature are included as training features. The models also demonstrate an ability to distinguish high concentration pollen events above a certain threshold, achieving F1 scores (for reliable event detection) between 60 and 83 %.

We recommend for further investigations that meteorological variables including RH and temperature are used for model training and target variables for both individual and collective pollen taxa should be studied, though the limitations of the OPC should be kept in mind. We find that NN and RF methods achieve similar results, but there is still potential for optimisation of these methods (particularly NN variations) to improve accuracy even further. Future work should include extending the method to other locations and environments and assessing the generalisability across sensors and different environments.

Furthermore, it would be beneficial to investigate application across networks of sensors to see if the information from the models can discern environmental variations on a local scale, as this is where the affordability of the sensors will have their edge. Hybrid networks would likely be required, with high quality instruments integrated across a few of the low-cost network sites, to provide accurate references for the low-cost sensors. High performance automated instruments, such as the Swisens Poleno or Plair Rapid-E, may become increasingly prevalent in populating pollen monitoring networks. Therefore, studies using pollen data from instruments such as these (in particular those that detect particles in-flight similarly to the OPC) as a reference, to provide the target variables in model training, would also be valuable. From here, future possibilities of high spatiotemporal resolution pollen measurement via sensor networks, or even personal pollen monitors, would look promising. This could ultimately provide valuable information to improve pollen forecasting models and help many allergy sufferers.

Author contributions

This study using the low-cost OPCs was conceived and planned by FDP, who gave guidance and contributed to the final manuscript, and SAM who developed the methodology for the machine learning models, performed the analyses and prepared the first draft. ARM also gave guidance and contributed to the final manuscript. DB performed the PMF source apportionment analysis and contributed to the final manuscript. JMM and FT planned and organised the intercomparison campaign in Munich and facilitated the running of the sensors there, providing the context for the data used in this study. JMM and FT also contributed to the final manuscript.

Data availability

Data supporting this publication are openly available from the UBIRA eData repository at doi: https://doi.org/10.25500/edata.bham.00000898.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Andrew Tanner from the University of Birmingham, UK, for his technical work on functionalising the Alphasense OPC sensors for the purposes of this study. We thank the ZAUM (Cordula Ebner von Eschenback, Gudrun Pusch and Christine Weil) and MeteoSwiss (Nina Burgdorfer and Sophie Erb) teams for their assistance facilitating the EUMETNET AutoPollen – COST ADOPT intercomparison campaign, as well as the Helmholtz Zentrum München as the campaign host and their assisting team (Daphne Kolland, George Matuscheck and Benjamin Schnautz). The manual pollen analysts for the campaign Łukasz Kostecki and Agata Szymanska are greatly appreciated for their work facilitating the vital reference data for this campaign. We thank Robert Gebauer, Gisela Nagy and Anton Pointner for their IT support during the campaign.

Funding

The work here involving the OPC sensors and related analysis was funded by the Natural Environment Research Council (NERC) through its Central England NERC Training Alliance (CENTA) doctoral research training consortium, at the University of Birmingham, and the grant “Quantification of Utility of Atmospheric Network Technologies (QUANT)” (NE/T001968/1). The intercomparison campaign where the OPC sensors were deployed and accompanying Hirst reference data was obtained for context of this study was funded by the Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit (LGL) and EUMETNET AutoPollen Programme. Financial support was also received for this from the COST Action CA18226 ADOPT – *New approaches in detection of pathogens and aeroallergens*.

CRediT authorship contribution statement

Sophie Mills: Data curation, Methodology, Original draft preparation, Visualization, Writing- Reviewing and Editing, **Dimitrios Bousiotis**: Methodology, Writing- Reviewing and Editing, **José Maya-Manzano**: Data curation, Methodology, Writing- Reviewing and Editing, **Fiona Tummon**: Data curation, Methodology, Writing- Reviewing and Editing, **Rob MacKenzie**: Supervision, Methodology, Writing- Reviewing and Editing,

Funding acquisition **Francis Pope**: Conceptualization, Supervision, Methodology, Writing- Reviewing and Editing, Funding acquisition.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2023.161969.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. https://doi.org/10.48550/arXiv.1603.04467 arXiv:1603.04467v2 [cs.DC].

Adamov, S., Lemonis, N., Clot, B., Crouzy, B., Gehrig, R., Graber, M.-J., Sallin, C., Tummon, F., 2021. On the measurement uncertainty of Hirst-type volumetric pollen and spore samplers. Aerobiologia https://doi.org/10.1007/s10453-021-09724-5.

Alarcón, M., Periago, C., Pino, D., Mazón, J., Casas-Castillo, M.C., Ho-Zhang, J.J., Linares, C., Rodríguez-Solà, R., Belmonte, J., 2022. Potential contribution of distant sources to airborne Betula pollen levels in northeastern Iberian Peninsula. Sci. Total Environ. 818, 151827. https://doi.org/10.1016/j.scitotenv.2021.151827.

Bacsi, A., Choudhury, B.K., Dharajiya, N., Sur, S., Boldogh, I., 2006. Subpollen particles: carriers of allergenic proteins and oxidases. J. Allergy Clin. Immunol. 118 (4), 844–850. https://doi.org/10.1016/j.jaci.2006.07.006.

Baird, A.B., Bannister, E.J., MacKenzie, A.R., Pope, F.D., 2022. Mass concentration measurements of autumn bioaerosol using low-cost sensors in a mature temperate woodland free-air carbon dioxide enrichment (FACE) experiment: investigating the role of meteorology and carbon dioxide levels. Biogeosciences 19 (10), 2653–2669. https://doi.org/10.5194/bg-19-2653-2022.

Bousiotis, D., Beddows, D.C.S., Singh, A., Haugen, M., Diez, S., Edwards, P.M., Boies, A., Harrison, R.M., Pope, F.D., 2022. A study on the performance of low-cost sensors for source apportionment at an urban background site. Atmos. Meas. Tech. 15, 4047–4061. https://doi.org/10.5194/amt-15-4047-2022.

Bradley, R.S., 2015. Paleoclimatology: Chapter 12 – Pollen, 3rd ed. Elsevier, Massachusetts, USA, pp. 408–409 https://doi.org/10.1016/B978-0-12-386913-5.00012-0.

Breiman, L., 1984. Classification and Regression Trees. 1st ed. Routledge, New York, USA https://doi.org/10.1201/9781315139470.

Burkart, J., Gratzl, J., Seifried, T.M., Bieber, P., Grothe, H., 2021. Isolation of subpollen particles (SPPs) of birch: SPPs are potential carriers of ice nucleating macromolecules. Biogeosciences 18, 5751–5765. https://doi.org/10.5194/bg-18-5751-2021.

Buters, J.T.M., Antunes, C., Galveias, A., Bergmann, K.C., Thibaudon, M., Galán, C., Schmidt-Weber, C., Oteros, J., 2018. Pollen and spore monitoring in the world. Clin. Transl. Allergy 8, 9. https://doi.org/10.1186/s13601-018-0197-8.

Buters, J., Clot, B., Galán, C., Gehrig, R., Gilge, S., Hentges, F., O'Connor, D., Sikoparija, B., Skjoth, C., Tummon, F., Adams-Groom, B., Antunes, C.A., Bruffaerts, N., Çelenk, S., Crouzy, B., Guillaud, G., Hajkova, L., Kofol Seliger, A., Oliver, G., Ribeiro, H., Rodinkova, V., Saarto, A., Sauliene, I., Sozinova, O., Stjepanovic, B., 2022. Automatic detection of airborne pollen: an overview. Aerobiologia https://doi.org/10.1007/s10453-022-09750-x.

Carslaw, D.C., Ropkins, K., 2012. Openair – an R package for air quality data analysis. Environ. Model. Softw. 27–28, 52–61. https://doi.org/10.1016/j.envsoft.2011.09.008.

Crilley, L.R., Shaw, M., Pound, R., Kramer, L.J., Price, R., Young, S., Lewis, A.C., Pope, F.D., 2018. Evaluation of a low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring. Atmos. Meas. Tech. 11, 709–720. https://doi.org/10.5194/amt-11-709-2018.

Crilley, L.R., Singh, A., Kramer, L.J., Shaw, M.D., Alam, M.S., Apte, J.S., Bloss, W.J., Hildebrandt Ruiz, L., Fu, P., Fu, W., Gani, S., 2020. Effect of aerosol composition on the performance of low-cost optical particle counter correction factors. Atmos. Meas. Tech. 13, 1181–1193. https://doi.org/10.5194/amt-13-1181-2020.

Després, V.R., Huffman, J.A., Burrows, S.M., Hoose, C., Safatov, A.S., Buryak, G., Fröhlich-Nowoisky, J., Elbert, W., Andreae, M.O., Pöschl, U., Jaenicke, R., 2012. Primary biological aerosol particles in the atmosphere: a review. Tellus B 64. https://doi.org/10.3402/tellusb.v64i0.15598.

Diehl, K., Quick, C., Matthias-Maser, S., Mitra, S.K., Jaenicke, R., 2001. The ice nucleating ability of pollen: part I: laboratory studies in deposition and condensation freezing modes. Atmos. Res. 58 (2), 75–87. https://doi.org/10.1016/S0169-8095(01)00091-6.

Diehl, K., Matthias-Maser, S., Jaenicke, R., Mitra, S.K., 2002. The ice nucleating ability of pollen: part II: laboratory studies in immersion and contact freezing modes. Atmos. Res. 61 (2), 125–133. https://doi.org/10.1016/S0169-8095(01)00132-6.

Dreischmeier, K., Budke, C., Wiehemeier, L., Kottke, T., Koop, T., 2017. Boreal pollen contain ice-nucleating as well as ice-binding 'antifreeze' polysaccharides. Sci. Rep. 7, 41890. https://doi.org/10.1038/srep41890.

Dubey, R., Patra, A.K., Joshi, J., Blankenberg, D., Sekhara, S., Kolluru, R., Madhu, B., Raval, S., 2022. Evaluation of low-cost particulate matter sensors OPC N2 and PM Nova for aerosol monitoring. Atmos. Pollut. Res. 13 (3), 101335. https://doi.org/10.1016/j.apr.2022.101335.

Frenguelli, G., Passalacqua, G., Bonini, S., Fiocchi, A., Incorvaia, C., Marcucci, F., Tedeschini, E., Canonica, G.W., Frati, F., 2010. Briging allergologic and botanical knowledge in seasonal allergy: a role for phenology. Ann. Allergy Asthma Immunol. 105, 223–227. https://doi.org/10.1016/j.anai.2010.06.016.

Fröhlich-Nowoisky, J., Kampf, C.J., Weber, B., Huffman, J.A., Pöhlker, C., Andreae, M.O., Lang-Yona, N., Burrows, S.M., Gunthe, S.S., Elbert, W., Su, H., Hoor, P., Thines, E., Hoffmann, T., Després, V.R., Pöschl, U., 2016. Bioaerosols in the Earth system: climate, health, and ecosystem interactions. Atmos. Res. 182, 346–376. https://doi.org/10.1016/j.atmosres.2016.07.018.

Giordano, M.R., Malings, C., Pandis, S.N., Presto, A.A., McNeill, V.F., Westervelt, D.M., Beekmann, M., Subramanian, R., 2021. From low-cost sensors to high-quality data: a summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. J. Aerosol Sci. 158, 105833. https://doi.org/10.1016/j.jaerosci.2021.105833.

Grewling, Ł., Bogawski, P., Jenerowicz, D., Czarnecka-Operacz, M., Šikoparija, B., Skjøth, C.A., Smith, M., 2016. Mesoscale atmospheric transport of ragweed pollen allergens from infected to uninfected areas. Int. J. Biometeorol. 60, 1493–1500. https://doi.org/10.1007/s00484-016-1139-6.

Grewling, Ł., Bogawski, P., Kryza, M., Magyar, D., Šikoparija, B., Skjøth, C.A., Udvardy, O., Werner, M., Smith, M., 2019. Concomitant occurrence of anthropogenic air pollutants, mineral dust and fungal spores during long-distance transport of ragweed pollen. Environ. Pollut. 254 (Part A), 112948. https://doi.org/10.1016/j.envpol.2019.07.116.

Griffiths, P.T., Borlace, J.-S., Gallimore, P.J., Kalberer, M., Herzog, M., Pope, F.D., 2012. Hygroscopic growth and cloud activation of pollen: a laboratory and modelling study. Atmos. Sci. Lett. 13 (4), 289–295. https://doi.org/10.1002/asl.397.

Gute, E., Abbatt, J.P.D., 2020. Ice nucleating behaviour of different tree pollen in the immersion mode. Atmos. Environ. 231, 117488. https://doi.org/10.1016/j.atmosenv.2020.117488.

Gute, E., David, R.O., Kanji, Z.A., Abbat, J.D.P., 2020. Ice nucleation ability of tree pollen altered by atmospheric processing. ACS Earth Space Chem. 4 (12), 2312–2319. https://doi.org/10.1021/acsearthspacechem.0c00218.

Hinds, W.C., 1999. Aerosol Technology: Properties, Behaviour, and Measurement of Airborne Particles. 2nd edn. John Wiley & Sons Inc., New York, USA.

Huffman, J.A., Perring, A.E., Savage, N.J., Clot, B., Crouzy, B., Tummon, F., Shoshanim, O., Damit, B., Schneider, J., Sivaprakasam, V., Zawadowicz, M.A., Crawford, I., Gallagher, M., Topping, D., Doughty, D.C., Hill, S.C., Pan, Y., 2019. Real-time sensing of bioaerosols: review and current perspectives. Aerosol Sci. Technol. 54 (5), 465–495. https://doi.org/10.1080/02786826.2019.1664724.

Hughes, D.D., Mampage, C.B.A., Jones, L.M., Liu, Z., Stone, E.A., 2020. Characterisation of atmospheric pollen fragments during springtime thunderstorms. Environ. Sci. Technol. Lett. 7 (6), 409–414. https://doi.org/10.1021/acs.estlett.0c00213.

Jiang, C., Wang, W., Du, L., Huang, G., McConaghy, C., Fineman, S., Liu, Y., 2022. Field evaluation of an automated pollen sensor. Int. J. Environ. Res. Public Health 19 (11), 6444. https://doi.org/10.3390/ijerph19116444.

Jochner, S., Lüpke, M., Laube, J., Weichenmeier, I., Pusch, G., Traidl-Hoffmann, C., Schmidt-Weber, C., Buters, J.T.M., Menzel, A., 2015. Seasonal variation of birch and grass pollen loads and allergen release at two sites in the German Alps. Atmos. Environ. 122, 83–93. https://doi.org/10.1016/j.atmosenv.2015.08.031.

Kawashima, S., Thibaudon, M., Matsuda, S., Fujita, T., Lemonis, N., Clot, B., Oliver, G., 2017. Automated pollen monitoring system using laser optics for observing seasonal changes in the concentration of total airborne pollen. Aerobiologia 33, 351–362. https://doi.org/10.1007/s10453-017-9474-6.

Manninen, H.E., Bäck, J.K., Sihto-Nissilä, S.-L., Huffman, J.A., Pessi, A.-M., Hiltunen, V., Aalto, P.P., Hidalgo, P.J., Hari, P., Saarto, A., Kulmala, M., Petäjä, T., 2014. Patterns in airborne pollen and other primary biological aerosol particles (PBAP), and their contribution to aerosol mass and number in a boreal forest. Boreal Environ. Res. 19, 383–405.. http://hdl.handle.net.10138/165208.

Maya-Manzano, J.M., Smith, M., Markey, E., Clancy, J.H., Sodeau, J., O'Connor, D., 2020. Recent developments in monitoring and modelling airborne pollen, a review. Grana 60 (1), 1–19. https://doi.org/10.1080/00173134.2020.1769176.

Maya-Manzano, J.M., Tummon, F., Abt, R., Allan, N., Bunderson, L., Clot, B., Crouzy, B., Daunys, G., Erb, S., Gonzalez-Alonso, M., Graf, E., Grewling, L., Haus, J., Kadantsev, E., Kawashima, S., Martinez-Bracero, M., Matavulj, P., Mills, S., Niederberger, E., Lieberherr, G., Lucas, R.W., O'Connor, D.J., Oteros, J., Palamarchuk, J., Pope, F.D., Rojo, J., Šaulienė, I., Schäfer, S., Schmidt-Weber, C.B., Schnitzler, M., Šikoparija, B., Skjøth, C.A., Sofiev, M., Stemmler, T., Triviño, M., Zeder, Y., Buters, J., 2022. Towards European automatic bioaerosol monitoring: comparison of 9 automatic pollen observational instruments with classic Hirst-type traps. Sci. Total Environ. 161220. https://doi.org/10.1016/j.scitotenv.2022.161220 (In Press).

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.A., 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 10, 213. https://doi.org/10.1186/1471-2105-10-213.

Miguel, A.G., Taylor, P.E., House, J., Glovsky, M.M., Flagan, R.C., 2006. Meteorological influences on respirable fragment release from Chinese elm pollen. Aerosol Sci. Technol. 40 (9), 690–696. https://doi.org/10.1080/02786820600798869.

Mikhailov, E.F., Ivanova, O.A., Nebosko, E.Y., Vlasenko, S.S., Ryshkevich, T.I., 2019. Subpollen particles as atmospheric cloud condensation nuclei. Izv. Atmos. Ocean. Phys. 55, 357–364. https://doi.org/10.1134/S000143381904008X.

Narayana, M.V., Jalihal, D., Nagendra, S.M.S., 2022. Establishing a sustainable low-cost air quality monitoring setup: a survey of the state-of-the-art. Sensors (Basel) 22 (1), 394. https://doi.org/10.3390/s22010394.

O'Connor, D.J., Healy, D.A., Hellebust, S., Buters, J.T., Sodeau, J.R., 2014. Using the WIBS-4 (waveband integrated bioaerosol sensor) technique for the on-line detection of pollen grains. Aerosol Sci. Technol. 48 (4), 341–349. https://doi.org/10.1080/02786826.2013.872768.

Oteros, J., Pusch, G., Weichenmeier, I., Heimann, U., Möller, R., Röseler, S., Traidl-Hoffmann, C., Schmidt-Weber, C., Buters, J.T.M., 2015. Automatic and online pollen monitoring. Int. Arch. Allergy Immunol. 167, 158–166. https://doi.org/10.1159/000436968.

Oteros, J., Weber, A., Kutzora, S., Rojo, J., Heinze, S., Herr, C., Gebauer, R., Schmidt-Weber, C.B., Buters, J.T.M., 2020. An operational robotic pollen monitoring network based on automatic image recognition. Environ. Res. 191, 110031. https://doi.org/10.1016/j.envres.2020.110031.

Paatero, P., Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5, 111–126. https://doi.org/10.1002/env.3170050203.

Pedregosaet al., n.d.Pedregosa, F., Varoquaux, G., Granfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res., 12 (85), 2825-2830.

Plaza, M.P., Kolek, F., Leier-Wirtz, V., Brunner, J.O., Traidl-Hoffmann, C., Damialis, A., 2022. Detecting airborne pollen using an automatic, real-time monitoring system: evidence from two sites. Int. J. Environ. Res. Public Health 19 (4), 2471. https://doi.org/10.3390/ijerph19042471.

Pope, F.D., 2010. Pollen grains are efficient cloud condensation nuclei. Environ. Res. Lett. 5, 044015. https://doi.org/10.1088/1748-9326/5/4/044015.

Pummer, B.G., Bauer, H., Bernardi, J., Bleicher, S., Grothe, H., 2012. Suspendable macromolecules are responsible for ice nucleation activity of birch and conifer pollen. Atmos. Chem. Phys. 12, 2541–2550. https://doi.org/10.5194/acp-12-2541-2012.

Reponen, T., 2011. Encyclopedia of Environmental Health: Methodologies for Assessing Bioaerosol Exposures. Elsevier, Cincinnati, USA, p. 723 https://doi.org/10.1016/B978-0-12-409548-9.11822-6.

Ruske, S., Topping, D.O., Foot, V.E., Morse, A.P., Gallagher, M.W., 2018. Machine learning for improved data analysis of biological aerosol using the WIBS. Atmos. Meas. Tech. 11 (11), 6203–6230. https://doi.org/10.5194/amt-11-6203-2018.

Šaulienė, I., Šukienė, L., Daunys, G., Valiulis, G., Vaitkevičius, L., Matavulj, P., Brdar, S., Panic, M., Sikoparija, B., Clot, B., Crouzy, B., Sofiev, M., 2019. Automatic pollen recognition with the Rapid-E particle counter: the first-level procedure, experience and next steps. Atmos. Meas. Tech. 12, 3435–3452. https://doi.org/10.5194/amt-12-3435-2019.

Sauvageat, E., Zeder, Y., Auderset, K., Calpini, B., Clot, B., Crouzy, B., Konzelmann, T., Lieberherr, G., Tummon, F., Vasilatou, K., 2020. Real-time pollen monitoring using digital holography. Atmos. Meas. Tech. 13, 1539–1550. https://doi.org/10.5194/amt-13-1539-2020.

Savage, N.J., Krentz, C.E., Könemann, T., Han, T.T., Mainelis, G., Pöhlker, C., Huffman, J.A., 2017. Systematic characterization and fluorescence threshold strategies for the wideband integrated bioaerosol sensor (WIBS) using size-resolved biological and interfering particles. Atmos. Meas. Tech. 10 (11), 4279–4302. https://doi.org/10.5194/amt-11-6203-2018.

Sazli, M.H., 2006. A brief review of feed-forward neural networks. Commun. Fac. Sci. Univ. Ank. Series A2-A3. 50 (1), pp. 11–17. https://doi.org/10.1501/commua1-2_0000000026.

Siljamo, P., Sofiev, M., Severona, E., Ranta, H., Kukkonen, J., Polevova, S., Kubin, E., Minin, A., 2008. Sources, impact and exchange of early-spring birch pollen in the Moscow region and Finland. Aerobiologia 24, 211–230. https://doi.org/10.1007/s10453-008-9100-8.

Skjøth, C.A., Sommer, J., Stach, A., Smith, M., Brandt, J., 2007. The long-range transport of birch (Betula) pollen from Poland and Germany causes significant pre-season concentration in Denmark. Clin. Exp. Allergy 37 (8), 1204–1212. https://doi.org/10.1111/j.1365-2222.2007.02771.x.

Smith, M., Matavulj, P., Mimić, G., Panić, M., Grewling, Ł., Šikoparija, B., 2022. Why should we care about high temporal resolution monitoring of bioaerosols in ambient air? Sci. Total Environ. 826, 154231. https://doi.org/10.1016/j.scitotenv.2022.154231.

Sofiev, M., Sofieva, S., Palamarchuk, J., Šauliene, I., Kadantsev, E., Atanasova, N., Fatahi, Y., Kouznetsov, R., Kuula, J., Noreikaite, A., Peltonen, M., Pihlajamäki, T., Saarto, A., Svirskaite, J., Toiviainen, L., Tyuryakov, S., Šukienė, L., Asmi, E., Bamford, D., Hyvärinen, A.-P., Karppinen, A., 2022. Bioaerosols in the atmosphere at two sites in Northern Europe in spring 2021: outline of an experimental campaign. Environ. Res. 214 (Part 2), 113798. https://doi.org/10.1016/j.envres.2022.113798.

Song, U., Park, J., Song, M., 2012. Pollen morphology of Pinus (Pinaceae) in northeast China. For. Sci. Technol. 8 (4), 179–186. https://doi.org/10.1080/21580103.2012.704973.

Sousan, S., Koehler, K., Hallet, L., Peters, T.M., 2016. Evaluation of the alphasense optical particle counter (OPC-N2) and the grimm portable aerosol spectrometer (PAS-1.108). Aerosol Sci. Technol. 50 (12), 1352–1365. https://doi.org/10.1080/02786826.2016.1232859.

Steiner, A.L., Brooks, S.D., Deng, C., Thornton, C.O., Pendleton, M.W., Bryant, V., 2015. Pollen as atmospheric cloud condensation nuclei. Geophys. Res. Lett. 42 (9), 3596–3602. https://doi.org/10.1002/2015GL064060.

Stone, E.A., Mampage, C.B.A., Hughes, D.D., Jones, L.M., 2021. Airborne sub-pollen particles from rupturing giant ragweed pollen. Aerobiologia 37, 625–632. https://doi.org/10.1007/s10453-021-09702-x.

Sun, X., Wang, H., Guo, Z., Lu, P., Song, F., Liu, L., Liu, J., Rose, N.L., Wang, F., 2020. Positive matrix factorisation on source apportionment for typical pollutants in different environmental media: a review. Environ. Sci.: Processes Impacts 22, 239–255. https://doi.org/10.1039/C9EM00529C.

Svozil, D., Kvasnička, V., Pospíchal, J., 1997. Introduction to multi-layer feed-forward neural networks. Chemom. Intell. Lab. Syst. 39, 43–62. https://doi.org/10.1016/S0169-7439(97)00061-0.

Taylor, P.E., Flagan, R.C., Miguel, A.G., Valenta, R., Glovsky, M.M., 2004. Birch pollen rupture and the release of aerosols of respirable allergens. Clin. Exp. Allergy 34 (10), 1591–1596. https://doi.org/10.1111/j.1365-2222.2004.02078.x.

Tong, H.-J., Ouyang, B., Nikolovski, N., Lienhard, D.M., Pope, F.D., Kalberer, M., 2015. A new electrodynamic balance (EDB) design for low-temperature studies: applicationto immersion freezing of pollen extract bioaerosols. Atmos. Meas. Tech. 8, 1183–1195. https://doi.org/10.5194/amt-8-1183-2015.

Tummon, F., Bruffaerts, N., Celenk, S., Choël, M., Clot, B., Crouzy, B., Galán, C., Gilge, S., Hajkova, L., Mokin, V., O'Connor, D., Rodinkova, V., Sauliene, I., Sikoparija, B., Sofiev, M., Sozinova, O., Tesendic, D., Vasilatou, K., 2022. Towards standardisation of automatic pollen and fungal spore monitoring: best practices and guidelines. Aerobiologia https://doi.org/10.1007/s10453-022-09755-6.

Van Vliet, A.J.H., Overeem, A., De Groot, R.S., Jacobs, A.F.G., Spieksma, F.T.M., 2002. The influence of temperature and climate change on the timing of pollen release in the Netherlands. Int. J. Climatol. 22, 1757–1767. https://doi.org/10.1002/joc.820.