An operational robotic pollen monitoring network based on automatic image recognition

Jose Oteros [a,b], Alisa Weber [c], Suzanne Kutzora [c], Jesús Rojo [a,d], Stefanie Heinze [c], Caroline Herr [c], Robert Gebauer [e], Carsten B. Schmidt-Weber [a], Jeroen T.M. Buters [a,*]

[a] *Center of Allergy & Environment (ZAUM), Member of the German Center for Lung Research (DZL), Technische Universita¨t München/Helmholtz Center, Munich, Germany*

[b] *Department of Botany, Ecology and Plant Physiology, University of C´ordoba, Co´rdoba, Spain*

[c] *Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit (LGL), Munich, Germany*

[d] *Department of Environmental Sciences, University of Castilla La-Mancha, Toledo, Spain*

[e] *IT Consulting Robert Gebauer, Germany*

Keywords:
Allergy
Aerobiology
BAA500
Hirst
ePIN
Quality control

A B S T R A C T

There is high demand for online, real-time and high-quality pollen data. To the moment pollen monitoring has been done manually by highly specialized experts.

Here we evaluate the electronic Pollen Information Network (ePIN) comprising 8 automatic BAA500 pollen monitors in Bavaria, Germany. Automatic BAA500 and manual Hirst-type pollen traps were run simultaneously at the same locations for one pollen season. Classifications by BAA500 were checked by experts in pollen identification, which is traditionally considered to be the "gold standard" for pollen monitoring.

BAA500 had a multiclass accuracy of over 90%. Correct identification of any individual pollen taxa was always >85%, except for *Populus* (73%) and *Alnus* (64%). The BAA500 was more precise than the manual method, with less discrepancies between determinations by pairs of automatic pollen monitors than between pairs of humans. The BAA500 was online for 97% of the time. There was a significant correlation of 0.84 between airborne pollen concentrations from the BAA500 and Hirst-type pollen traps. Due to the lack of calibration samples it is unknown which instrument gives the true concentration.

The automatic BAA500 network delivered pollen data rapidly (3 h delay with real-time), reliably and online. We consider the ability to retrospectively check the accuracy of the reported classification essential for any automatic system.

1. Introduction

We evaluated the performance of the electronic Pollen Information Network (ePIN) in Bavaria, Germany. The ePIN network was built in 2017/2018 to deliver reliable pollen information. The network is based on the automatic BAA500 system that can determine airborne pollen concentrations by using image recognition. The device automatically extracts pollen grains from the atmosphere with a virtual impactor, prepares microscopic slides and analyses and counts the extracted pollen grains with an automated light microscope with an image processing system. The sampling system works with a flow of 1000 l/min, subsampling the collected air on average at 100 l/min. The airflow entering the sampler is from 360◦ in BAA500 while it is only from the dominant wind direction in Hirst (Oteros et al., 2015), ten-fold higher than the traditional Hirst-type pollen trap that is used around the world (Buters et al., 2018).

Hirst-type pollen traps and manual counting are considered the "gold standard" in pollen monitoring (Gala´n et al., 2014; Hirst, 1952). However, this manual system has some limitations:

1. Time delayed information - The manual Hirst method provides pollen data with a delay of at least 1-day, but this is often extended to 7-days due to organisational and budgetary considerations. On the other hand, the BAA500 can send online pollen data with a resolution of currently 3 h (Oteros et al., 2015).

* Corresponding author.
E-mail address: buters@tum.de (J.T.M. Buters).

2. Variability of results - The main source of variability in the manual collection of pollen data is the proficiency of pollen experts (Galán et al., 2014; Smith et al., 2019). However, other sources of variation include differences in the counting method, the subsampling of the environment and technical factors like differences in flow rate between devices and the different mounting media (Oteros et al., 2013, 2017). Many of these can be reduced or eliminated although random human error will always remain (Oteros et al., 2019).

Some of these problems were resolved by the 3rd generation (3G) pollen systems currently available on the market: e.g. KH-3000 (Kawashima et al., 2017), Wibs-4 (O'Connor et al., 2014), BAA500 (Oteros et al., 2015), Plair PA-300/RapidE (Crouzy et al., 2016; Šauliene˙ et al., 2019), PollenSense (pollensense.com), or SwisensePoleno (Chappuis et al., 2019; Huffman et al., 2019). All these systems promise to deliver fast and reproducible pollen information but, to date, no instrument has been evaluated when working within a network (Huffman et al., 2019).

Several of the above-mentioned systems already report an accuracy >80% for numerous pollen types (Huffman et al., 2019), which is the accepted accuracy for the manual system (Galán et al., 2014). Furthermore, it is expected that Artificial Intelligence will outperform humans in many activities in coming years, especially in fields related to visual identification (Grace et al., 2018; He et al., 2015; Silver et al., 2017). For the time being, pollen experts are more accurate than automatic pollen monitors and need to be involved before network pollen data can be reliably reported.

The main aim of this work was to evaluate the performance of the automatic pollen monitor BAA500, used at different locations of the ePIN Network in Bavaria. We evaluated: (A) The accuracy and precision of the automatic classifications delivered by the BAA500 checked by human experts in pollen dentification; (B) The performance of the same instrument at different locations; (C) The results of BAA500 versus manual Hirst-type pollen traps.

2. Material and methods

2.1. Study area

Bavaria is the largest state of the Federal Republic of Germany with a surface of 70,553 km^2. It has a population of nearly 13 million inhabitants. Bavaria has a humid continental climate, classified as Dfb by the Köppen-Geiger system described by Belda et al. (2014) (Belda et al., 2014). This climate is characterized by large seasonal temperature differences, with warm summers and cold winters. Precipitation is distributed throughout the year. By definition, a climate is classified as humid continental when the temperature of the coldest month is below —3 ◦C and when at least four months have mean temperatures above 10 ◦C (Belda et al., 2014). The yearly average temperature in Munich, the capital, is 8.7 ◦C and the annual mean precipitation is 834 mm (1981–2010). More extreme climatic conditions are observed in the southern parts of Bavaria, in the Alps. Thirty-seven percent of the area is covered by forests (bwi.info, accessed 01-08-2019). The vegetation of this climate region includes temperate woodlands, temperate grasslands, temperate deciduous forests, temperate evergreen forests, and coniferous forests.

2.2. Pollen monitoring and data flow

Two pollen monitoring methods were compared in this study: the manual method using Hirst-type pollen traps (Hirst, 1952) and the automatic pollen monitor BAA500 (Oteros et al., 2015). The manual samples where managed and analysed following the VDI norms (VDI4252-4, 2016), fulfilling the EAS minimum requirements as described by Oteros et al. (2019) (Oteros et al., 2019). Pollen experts were selected from an already established European pollen network. The quality control of all pollen experts was carried out during a previous project in Bavaria during 2015 (Smith et al., 2019). Furthermore, a quality control program testing the performance of each pollen expert was implemented during the current experiment. Experts identified pollen on images produced by the BAA500 (eCounters) or slides prepared from samples collected by Hirst-type traps (Hirst counters). Most pollen experts worked on both tasks.

The operation of the BAA500 was described previously (Oteros et al., 2015). All 8 automatic BAA500 traps in the ePIN network were included in the current study. Four Hirst-type pollen traps ran in parallel for one year (Munich, Garmisch-Partenkirchen, Marktheidenfeld and Viechtach), see Fig. 1.

Images of multiple focal planes taken by the BAA500 delivered a stack of 210 pictures of every particle captured in the ePIN network. Out of this stack, a synthetic 2D-picture per particle was built containing all the captured information (the optimized sum of all images taken from a specific particle). All the synthetic images of all captured particles, including pollen, were centrally stored in the Leibnitz-Rechenzentrum (LRZ) supercomputer. A subset (randomly selected) of these images was re-analysed online for evaluating the performance of the BAA500. Software was developed to rapidly access these images, enabling easy manual classification by pollen experts. A randomly selected portion of the particles automatically classified by BAA500 were manually labelled. This validation software can be accessed by stakeholders at https://validation.epin.bayern/using the expert name "Gast" and the password "ePIN-ZAUM".

The BAA500 is currently trained to recognize 40 pollen and spore taxa: *Abies, Acer, Aesculus, Alnus, Ambrosia, Artemisia,* Other *Asteraceae, Betula, Carpinus, Castanea, Chenopodium (Amaranthaceae), Corylus, Cruciferae, Cyperaceae, Erica, Fagus, Fraxinus, Fungus, Galium, Humulus, Impatiens, Juglans, Larix, Picea, Other Pinaceae, Pinus, Plantago, Platanus, Poaceae, Populus, Quercus, Quercus ilex, Rumex, Salix, Sambucus, Secale, Taxus, Tilia, Ulmus, Urticaceae*. Furthermore, a category termed "Varia" is selected when the particle is recognized as a pollen but of an unknown taxa. A category termed "–" ("dash-dash") is reported when the particle is recognized as not being pollen (a non-pollen particle).

The pollen experts working with the digital pictures (eCounters) reported the same pollen categories as the BAA500 including "NoPollen" (when the pollen expert recognizes the particle as not being a pollen grain), and in addition "Unknown Particle" (when the expert cannot

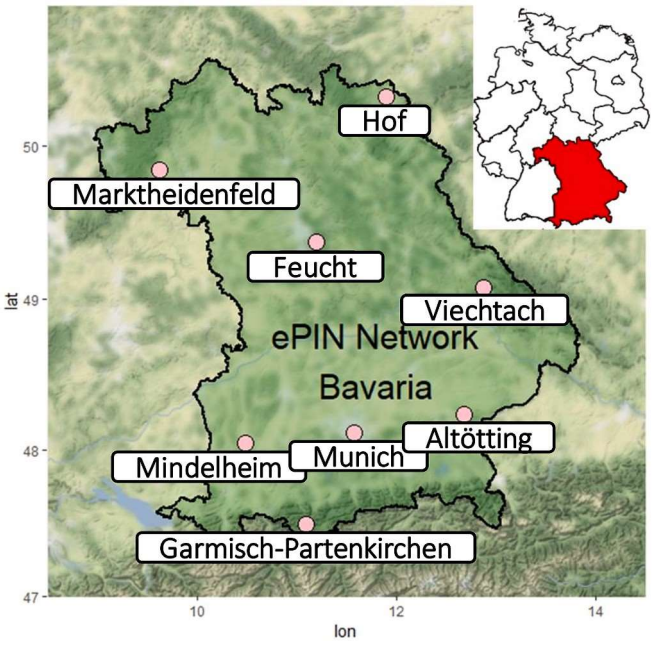

Fig. 1. The 8 pollen monitoring locations of ePIN Network in the state of Bavaria. The insert shows the location of Bavaria within Germany.

classify the particle at all, i.e. it could be a part of a pollen grain), "Unknown Pollen" (when the pollen expert recognizes the particle as a pollen, but she/he does not know which pollen it is) and "Other Pollen" (when the pollen expert recognizes the pollen, but it is not one of the categories trained by BAA500).

In the current study the labelling of each particle by a pollen expert was considered the "true value". The performance of pollen recognition by BAA500 was based on the statistical measures of the performance of a binary classification test. Thus, for each particle four possible categories can be attributed: True Positive (TP), True Negative (TN), False Positive (FP) or False Negative (FN). In the case of a Poaceae pollen, for example, a reanalysed particle can be labelled by the expert as: (1) "True Positive" if it was recognized as Poaceae by both, automatic pollen monitor and human; (2) "False Positive" if it was recognized as Poaceae by the automatic pollen monitor but as a different type by the expert; (3) "True Negative" if it was recognized as a different pollen by the automatic pollen monitor and by the human; (4) "False Negative" if it was not recognized as a Poaceae pollen by the automatic pollen monitor but was by the human (human pollen experts were considered 100% correct).

2.3. Data analysis

Data analyses were performed to ensure: (1) Accuracy; (2) Precision; (3) Reliability.

2.3.1. Accuracy

According to ISO 5725–1 (ISO-5725, 1994), the general term "accuracy" (trueness) is used to describe the closeness of a measurement to the true value. In the current study, the term "accuracy" is taken widely and is approached by calculating two parameters for each particle: Specificity (recall) and positive predictive value, as described by Oteros et al. (2015).

For each pollen type we determined the False Positives (FP), False Negatives (FN), True Positives (TP) and True Negatives (TN). The calculated accuracy was the fraction of the total sample that is correctly identified: (TP + TN)/(TP + TN + FP + FN). Thus, for pollen recognition, calculating accuracy by taking "non-pollen particles" into account did not make much sense because the amount of TN would be almost as large as the number of sampled particles (as the majority of all collected particles were not pollen but other particles e.g. dust particles, pollution …).

Instead of including TN in the calculations we analysed accuracy by two different parameters: sensitivity (recall) and the positive predictive value. Sensitivity measures the proportion of positives that are correctly identified and presented in this study as coinciding manual classifications:

TP/(TP + FN).

The positive predictive value is the fraction of the classifications that are positive and described as correct automatic classifications:

TP/(TP + FP).

Both, sensitivity (humans check BAA500 classifications) and the positive predictive value (BAA500 agrees with humans) are shown as a percentage (%).

Another accuracy parameter calculated was specificity, i.e. the ability of discerning NoPollen from Pollen, which is the proportion of NoPollen properly identified:

TN/FP + TN.

2.3.2. Precision

Precision is the ability of the method to repeat and reproduce the same results under similar conditions. The repeatability (r) of a measuring method is the variation in measurements taken by a single instrument under the same conditions. Reproducibility (R) is the ability of the method to provide the same measurement, but under different conditions (i.e. using a different instrument at a different location). Following the standard ISO 5725, r is calculated by several repetitions of the measure at the same laboratory under the same conditions and R is calculated by an interlaboratory proficiency testing (ISO-5725, 1994). Both, r and R are a measurement of the precision of the method under different conditions: r at the same trap and location; R between different traps and locations. Currently, the real concentration of pollen in not known because there are no calibration samples of concentrations available.

Due to the fact that it is not currently possible to produce known pollen concentrations, and then repeat the same exposure several times, we calculated an approximation of the precision by the discrepancies in accuracy values at each location. In this case, we assumed perfect reproducibility when all the locations showed the same accuracy (positive predictive value), and we quantified precision by the differences in positive predictive value between pairs of stations (Delta error) (ISO-5725, 1994).

2.3.3. Reliability and comparison of BAA500 to hirst-type pollen traps

Reliability of the method was defined as the proportion of the operational time that monitoring was effective. We differentiated between reliability in delivering daily data (pollen/m^3 measured per day) and reliability of diurnal data (pollen/m^3 measured per 3hr periods). Days with at least one data point per day are counted as daily reliable.

We also compared the daily pollen concentrations provided by BAA500 and Hirst-type pollen traps in four locations during one year. Traps were located with the height inlet of a BAA500 at 250 cm and the Hirst-type inlet at 185 cm, both within 5 m of each other. To compare BAA500 and Hirst we calculated the daily ratio between concentrations BAA500/Hirst. We also calculated Spearman's correlations between daily concentrations in both systems.

3. Results

3.1. General performance

The total number of particles registered by BAA500 system and reanalysed by the human pollen analysts was 580,705 from which 484,953 were pollen grains (or the identified spores). Of the particles classified by BAA500 as "–" (or non-pollen particles) humans agreed in 98.7% of the cases (n = 95,752), see Fig. 2A. Thus, the BAA500 performed well in separating particles from pollen.

For further analysis, we focused on the identification by pollen experts of the particles classified by BAA500 as "pollen" (or the identified spores). Fig. 2B shows the distribution of automatically classified pollen in relation to human identification. The BAA500 category "Varia" (similar to the human category "Unknown Pollen", see Fig. 2A) is the most abundant reported pollen category. All the particles in this category were classified by the BAA500 as pollen, but not which pollen types they were. Of these, humans agreed in most cases and classified "Varia" as "Unknown Pollen" (Fig. 2B). Others were classified as "Unknown Particle" or "NoPollen". *Carpinus* is the most abundant pollen hidden in "Varia".

Fig. 3A shows the relative abundance of the most common pollen particles (>1%) registered in the network, with "Varia" being a notable proportion. Figure 3B shows the distribution of the automatic pollen classifications by BAA500 which were also manually labelled by pollen experts, i.e. particles of which both automatic and manual classification were known. Fig. 3B shows only the most abundant pollen types (>50 manual classifications and >50 automatic classifications).

3.2. Accuracy of the network

The positive predictive value (% of correct automatic classifications) and sensitivity (% of coinciding manual classifications) of the BAA500 in the ePIN network are shown in Fig. 4. The graph can be read as follows: on the y-axis 100% of the cases are automatically reported (also the bars plot in the right), and the percentage of the cases to which humans

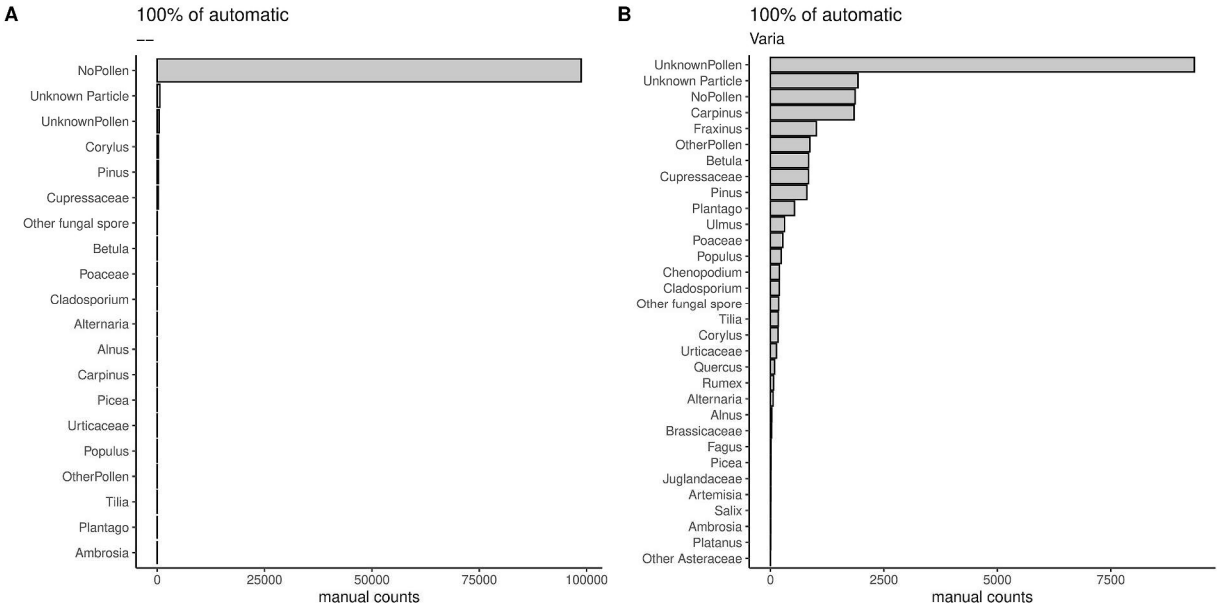

Fig. 2. Interpretation by pollen experts (y-axis) of the automatic classifications by BAA500: (A) '- -' meaning a non-pollen particle and (B) 'Varia - Unknown'.

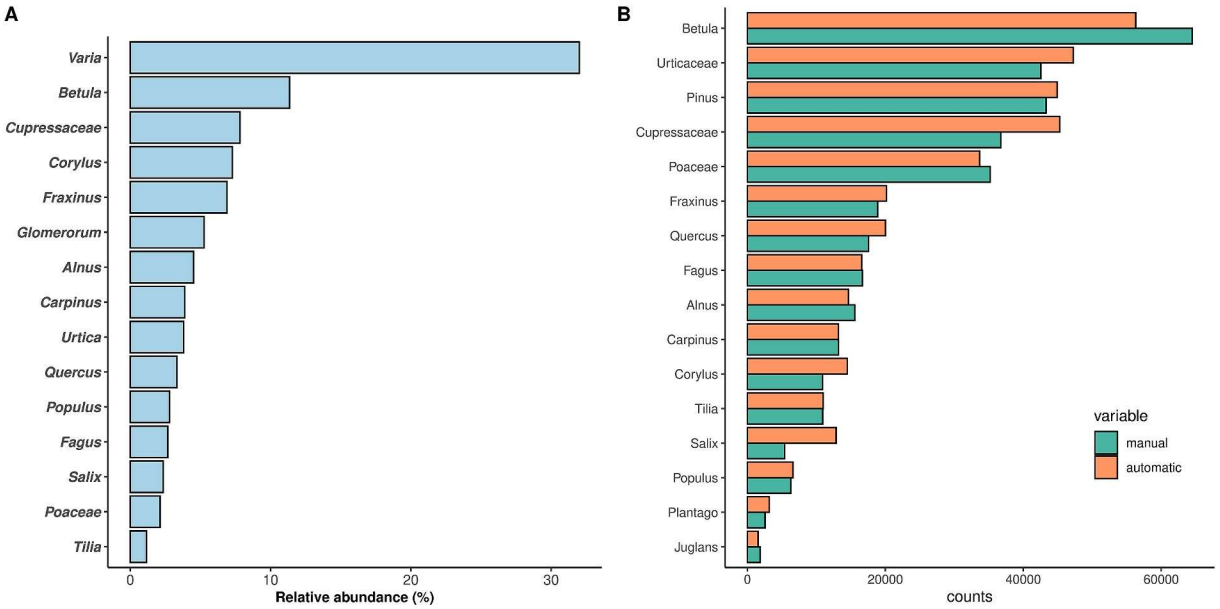

Fig. 3. Fig. 3 (A) Percentage distribution of pollen taxa automatically recognized by BAA500. 'Varia' is used for pollen grains not identified. (B) Comparison of the percentages of automatically and manually recognized pollen taxa. *Glomerorum* is the category used for Fungal spores.

agreed with the automatic pollen monitors are given (i.e. percent of true positives). In the x-axis, and the bars plot at the top, the cases reported by pollen experts were 100%, and the percentage of those which the automatic pollen monitor was also able to recognize is indicated. If a single mistake was >10% it was labelled with its own colour (e.g. in the automatic category *Populus* >10% of the pollen were identified by humans as *Corylus*). Most mistakes are coloured grey, in other words they included all kind of errors but never more than 10% of the same pollen.

The most prominent error was the confusion between *Alnus* and *Corylus*. For *Alnus*, the x-axis value (63%) shows that the automatic BAA500 system has recognized only 63% of the *Alnus* pollen classified by the experts and the bars inform that in 37% of the cases BAA500 was wrongly assigned the pollen to *Corylus* and to other categories each of which under 10%. Y-axis (90%) shows that the experts agreed in 90% of the BAA500 automatic classifications and the column inform that none of the pollen types confused with *Alnus* was >10%.

The most important pollen types regarding allergy in the area are *Betula* (birch) and Poaceae (grasses). It is important to note that the BAA500 classified the majority of these correctly (i.e. True Positives) and less than 10% of the cases were False Positives (TP + FP) (Fig. 4 – y-axis). When examining the manual classifications (Fig. 4 - x-axis), approximately 80% of the *Betula* pollen grains were correctly identified by the BAA500 (True Positives), but about 20% were missed.

In the case of *Salix* (Fig. 4 -y-axis) it can be seen that the BAA500 often wrongly classified *Betula* pollen as *Salix*. This is because *Betula* is the most abundant pollen in the network (Fig. 3B) and an error in classifying *Betula* as *Salix* will give rise to a large error in *Salix*. In environments where *Salix* is plentiful its positive predictive value is likely to increase.

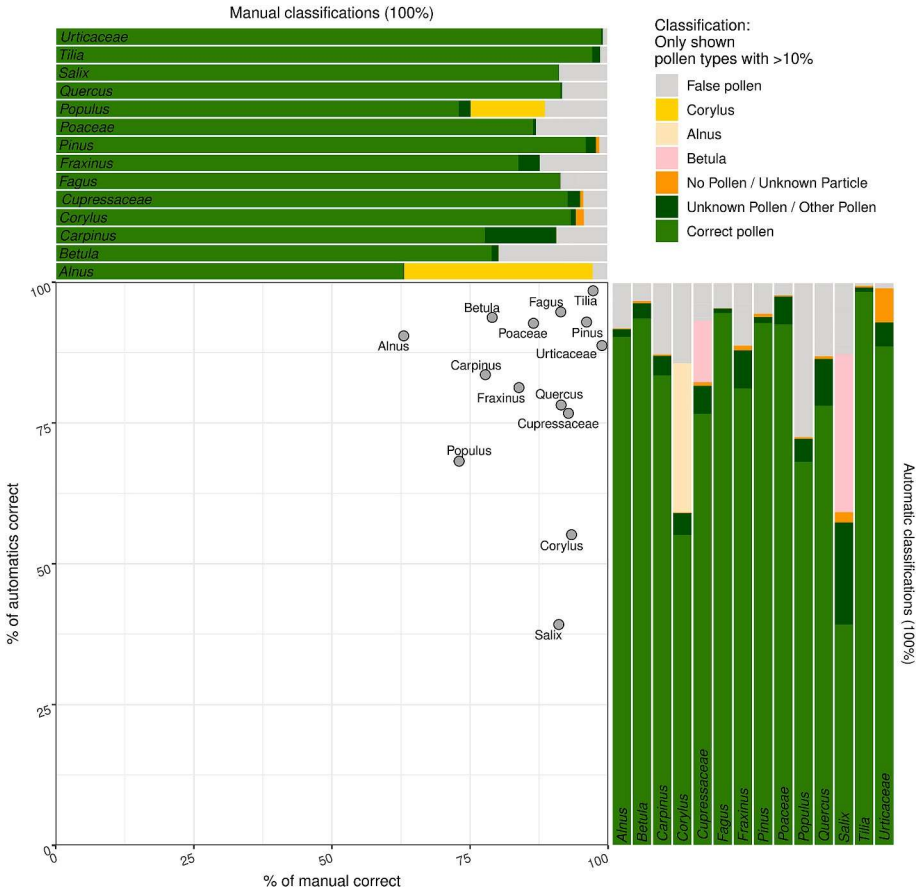

Fig. 4. Combined plot of predictive value (% of correct automatic classifications) and sensitivity (% of correct manual classifications). Vertical (y-axis) human corrected BAA500 classifications (predictive value: % of correct automatic classifications or True Positives) and horizontal (x axis) the BAA500's performance on manually identified pollen (sensitivity: the percent of manual classifications reported correct by the automatic pollen monitor). Only pollen types that were manually classified more than 5000 times are shown. The y-axis contains, for each pollen type, the 100% of the TP + FP. The x-axis contains, for each pollen type, 100% of the TP + FN. In the legend, False pollen (in grey) means other pollen representing <10% of a bar.

Fig. 4 shows the accuracy of the system under real monitoring conditions. However, other automatic pollen monitoring systems, not based on image recognition, do not allow the calculation of True/False-Positives/Negative as the instruments are tested with resuspended pure pollen samples. To compare our results with other published automatic systems, we simplified Fig. 4 to a confusion matrix of the recognition skills limiting the classes to the tested ones, also allowing the calculation of the "Multiclass accuracy", see Table 1. Multiclass accuracy under real-life (not laboratory) circumstances was calculated as the mean percentage of success in the identification of the tested pollen types, which was >90% for BAA500.

3.3. *Precision of the network*

Fig. 5 shows the differences in positive predictive value between pairs of experts (A) and between pairs of automatic monitors (robots) (B), termed "Delta Error". The boxplots in Fig. 5A are made with data points that are the difference in predictive value between each pair of experts for each pollen taxa. The boxplots in Fig. 5B shows these discrepancies in predictive value between robots.

Because all automatic monitors work using the same classification algorithm, a lower variability in discrepancies between robots than between humans was expected. As can be observed in the figure, this is the case for most of the pollen types.

Table 1
Real-life confusion matrix of the BAA500 under environmental monitoring conditions (no feeding of known pollen). Each row is the 100% of actual manual pollen (TP + FN). Multiclass accuracy was >90%. For instance, of 100% manually recognized Betula pollen 86% is classified by BAA500 as such. Aln = *Alnus*; Bet = *Betula*; Car = *Carpinus*; Cor = *Corylus*; Cup = Cupressaceae; Fag = *Fagus*; Fra = *Fraxinus*; Pin = *Pinus*; Poa = Poaceae; Pop = *Populus*; Que = *Quercus*; Til = *Tilia*; Urt = Urticaceae. In bold the main diagonal.

Pollen	*Aln*	*Bet*	*Car*	*Cor*	Cup	*Fag*	*Fra*	*Pin*	Poa	*Pop*	*Que*	*Til*	Urt.
Aln	**64**	0	0	35	1	0	0	0	0	0	0	0	0
Bet	1	**86**	0	4	6	0	2	0	0	1	0	0	0
Car	0	1	**90**	2	0	0	0	2	0	4	1	0	0
Cor	1	0	0	**96**	3	0	0	0	0	0	0	0	0
Cup	0	0	0	2	**96**	0	1	0	0	1	0	0	0
Fag	0	0	3	0	0	**93**	0	1	0	2	1	0	0
Fra	2	3	0	3	2	0	**87**	0	0	1	2	0	0
Pin	0	0	0	0	0	0	0	**100**	0	0	0	0	0
Poa	0	0	0	0	1	0	0	4	***88***	0	7	0	0
Pop	2	1	1	15	4	0	1	0	0	***73***	3	0	0
Que	0	0	0	0	3	0	1	0	0	4	***92***	0	0
Til	0	0	0	0	0	0	0	0	0	0	1	***99***	0
Urt	0	0	0	0	0	0	0	0	0	0	0	0	***100***

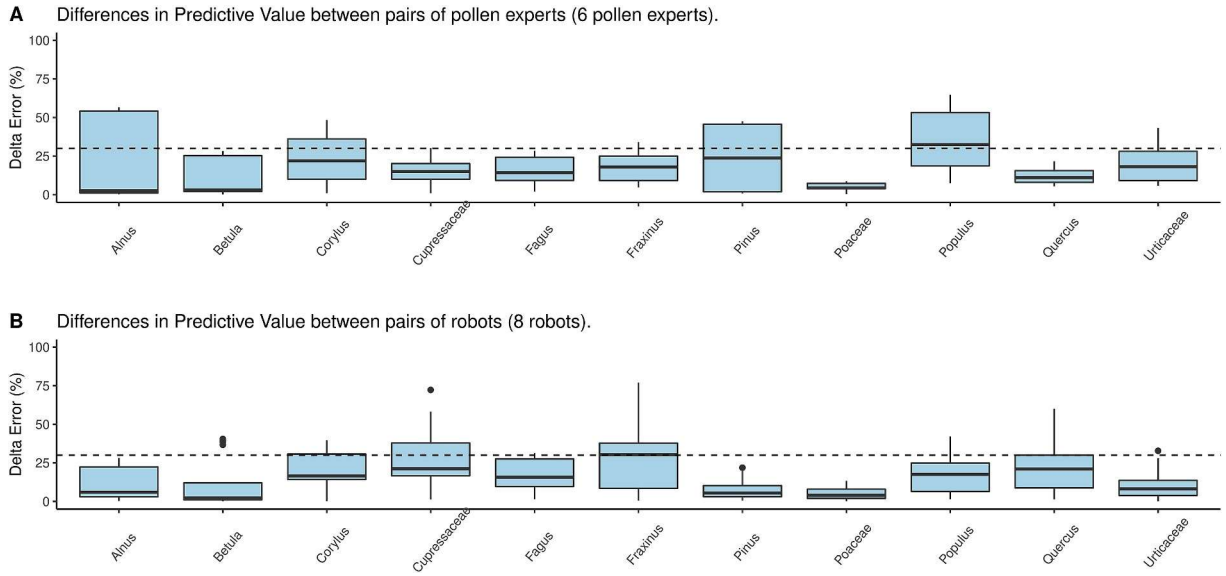

Fig. 5. Discrepancies (Delta Error) between identifications: (A) Differences in the reported accuracy in identifying each pollen type by each pair of human experts (n = 6); (B) Differences in the reported accuracy in identifying each pollen type by each pair of robots (n = 8). Only pollen analysed >500 times for each expert or robot are included.

3.4. Reliability of the network

When automatic pollen monitoring first commenced using the BAA500, reliability (time online) was problematic (Oteros et al., 2015). We therefore checked the reliability in the current network of BAA500. As can be observed in Table 2, the daily reliability of the network was 97% (days with at least 3 h of data). From all the possible 3-h data points, 94% where recorded. Remarkably, the oldest station (DEBIED: Munich) with the most experience had the least amount of down time.

3.5. BAA500 vs. hirst-type pollen trap

We compared the pollen data obtained by BAA500 with pollen data concomitantly measured by Hirst-type pollen traps at the same location. Fig. 6A shows the comparison of the pollen concentrations obtained by both methods at four parallel stations. Here we observed a bias in the pollen amount: Some pollen types like *Alnus* are more abundant in the BAA500, whilst some others like Urticaceae are more abundant in Hirst-type pollen traps. Fig. 6B shows the comparison between BAA500 and Hirst-type pollen traps per pollen type and location. The rate is constant among locations but differs by pollen type.

Table 3 shows the results of statistical analysis between the BAA500 and Hirst-type traps for each pollen type. Pollen types showing Spearman correlation coefficients close to 1 and with low standard deviations suggest that the data are comparable. This is the case, for example, for *Betula* or *Plantago*. This shows the weakness of correlation coefficients for comparing pollen time series with low concentrations, reflecting the inaccuracy of pollen monitoring when concentrations are low. In the case of *Plantago*, the correlation is almost 0 and not significant. However, the values given by both BAA500 and Hirst-type traps are almost the same (Fig. 3B), and the daily ratios "dance" around 1, changing every day depending on the pollen concentrations, resulting in a correlation of 0.

Pollen types showing notable differences in the quantity of both methods are *Carpinus* (with a ratio close to 2, meaning concentrations in the BAA500 were double that recorded in the Hirst-type pollen trap), *Picea* (with a ratio of 0.2, meaning 80% more pollen was recorded in the Hirst-type trap than in the BAA500), Poaceae (ratio of 0.8, 20% more pollen in the Hirst-type trap compared to the BAA500) and Urticaceae (ratio of 0.6, 40% more pollen in the Hirst-type trap). *Artemisia* and *Ambrosia* are not included in the table or in the plots because both have low concentrations in Bavaria.

Table 2
Daily and hourly reliability of BAA500 (%).

location	Daily Reliability	3-Hourly Reliability
DEALTO	93	89
DEBIED	100	99
DEGARM	95	91
DEHOF	95	89
DEMARK	100	98
DEMIND	100	100
DEVIEC	93	90

4. Discussion

We have validated a network of automatic pollen monitors. The BAA500 automatic monitors basically work like manual Hirst-type pollen traps as they use an impactor to catch pollen on an adhesive surface and microscopic image recognition for classification of the pollen type. The main difference between Hirst-type and these automatic pollen monitors is the higher sampling rate, the adhesive mountant, automation and online availability.

The results on the accuracy of a whole network of automatic pollen monitors are similar to the results shown for a single monitor (Oteros et al., 2015). This means that the system is stable temporally and spatially, i.e. the production process of the automatic pollen monitors is reproducible over time and space. Sensitivity and Positive Predictive Value were previously tested for the BAA500 system by Oteros et al. (2015), and agree with the current results showing an average sensitivity of around 75% and a positive predictive value of around 85% for 13 pollen types. The manufacturer has also tested both parameters for 12 pollen types and showed a sensitivity of 95% and a positive predictive value of 93% (Wetzlar, 2009). The difference in results is explained by the manufacturer testing with pure pollen samples (feeding experiments) whereas we used real-life samples (Oteros et al., 2015). This influences at least two factors:

1. Under laboratory conditions, a fixed collection of pollen types is used for building the confusion matrix. As usually this is done with lower

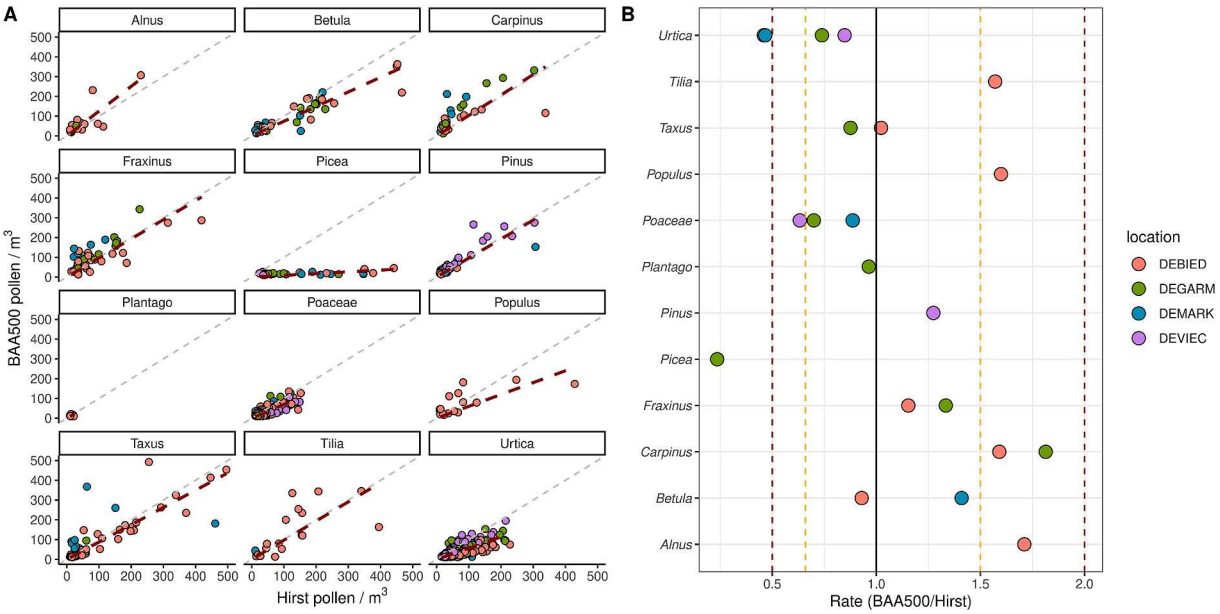

Fig. 6. Comparison of pollen concentrations measured by Hirst-type pollen traps vs BAA500 automatic monitors. (A) Daily pollen concentrations measured by BAA500 versus concentrations measured by Hirst-type pollen trap at the same location per pollen type. The dashed line indicates y = x. Only stations with a complete pollen season of particular pollen with both Hirst-type traps and BAA500 monitors were included. (B) Ratios of Hirst-type versus BAA50 monitors sorted by stations (DEBIED: Munich; DEGARM: Garmisch; DEMARK: Marktheidenfeld; DEVI.

Table 3
BAA500 versus Hirst concentrations with standard deviation, Spearman's correlations between Hirst-type pollen trap and BAA500. Although both instruments correlate well for most pollen, the exact number vary with a fixed coefficient, probably due to the different impaction principles of the instruments.

Pollen type	Ratio (mean)	Ratio (sd)	Correlation	p-value correlation
Alnus	1.680	0.841	0.814	0.000
Betula	1.074	0.667	0.858	0.000
Carpinus	1.929	1.179	0.944	0.000
Fraxinus	1.500	1.217	0.984	0.000
Picea	0.224	0.187	0.318	0.071
Pinus	1.303	0.568	0.857	0.000
Plantago	0.953	0.293	−0.138	0.501
Poaceae	0.769	0.359	0.794	0.000
Populus	1.629	1.399	0.722	0.001
Taxus	1.365	1.310	0.793	0.000
Tilia	1.598	1.541	0.528	0.012
Urticaceae	0.622	0.312	0.743	0.000

diversity of samples than in real-life conditions, meaning that fewer categories can be labelled from a lower diversity of samples. This dramatically decreases the probability of having False Negatives. This means that the more pollen types are allowed in the confusion matrix, the lower the accuracy will be. An extreme example would be a confusion matrix built with only two very easy to differentiate pollen types, there the accuracy and the sensitivity/recall would be 100%.

2. Under real-life conditions, the proportion of Positives/Negatives of each particle is reduced due to the greater diversity of any particles in the environment, which dramatically increases the probability of having False Positives. The more biodiverse the environment, the more chances to producing FPs by the automatic pollen monitor. This phenomenon allowed us to observe an interesting statistical artefact in our evaluation. The more prevalent a pollen type is for one day, the higher the automatic classifications are True Positives (during the peak days, the automatic pollen monitor tends to work perfectly). The most extreme example of this phenomenon would be if we evaluate a system only under lab conditions, by "feeding" monospecific pollen doses. Under these conditions, the probability of getting False Positives is 0 and the Positive Predictive Value should be 100%.

If we compare our results with a previous publication on the same monitor, we also observed an increase in the reliability of the BAA500 (Oteros et al., 2015). The main problems in the past were software and hardware related problems with the camera, which were subsequently solved. The main reason for automatic monitors not reporting data in the current evaluation were human errors while changing the sample holders. The holders contain about 180 sampling substrates, i.e. 23 days of continuous sampling. Changing the holders, in this case by inexperience personnel, lead to initial problems. In Munich (DEBIED) personnel have been changing holders for several years and the reliability was close to 100%.

Another problem we faced was the measurement of the repeatability in the network. In the case of Hirst-type pollen trap, repeatability was tested under laboratory conditions in the first paper describing the trap (Hirst, 1952). Unfortunately, at that time, the technology was not available for measuring the true value of pollen concentration in the air. This is probably a reason for the dramatic effect of wind speed on capture efficiency. The trap was located in a wind tunnel and a known concentration of spores was dispersed and trapped (concentration measured with two different methods). The repeatability of the Hirst-type pollen trap was very high (maximum 20% of difference between measurements taken under the same wind conditions). Later experiments showed an even better capture efficiency for pollen (Mullins and Emberlin, 1997), stating that the trapping efficiency (efficiency for catching the pollen) is increased by increasing the trapping surface by emulating the stigma of an anemophilous flower. The authors therefore suggested that using a cylinder in the Hirst-type trap increases the chance of trapping airborne particles. Due to the fact that we do not know the true pollen concentrations of the environment, we approximated the measurement of precision by calculating the difference in accuracy between each pair of automatic pollen monitors. The precision of pollen experts was assessed in the same way. As expected, the precision of an automatic pollen monitor for image identification is higher

than between pollen experts. However, this could not be calculated for accuracy, as we assumed that pollen experts had an accuracy of 100%. It is known that identification by pollen experts can differ, and differences of about 20% are considered acceptable for most of networks and scientific investigation (Gala´n et al., 2014; Smith et al., 2019).

We compared BAA500 automatic pollen monitors with Hirst-type pollen traps and observed an overall correlation in reported pollen concentrations of 0.84. However, the pollen concentrations differed between the methods depending on pollen type. The reason for this bias could be that the capture efficiency of theBAA500 and Hirst-type pollen traps are dependent on particle size (Mullins and Emberlin, 1997), and the way that particle size impacts upon capture efficiency is different for each system. The lack of a known calibration concentration ("gold standard") currently means that it is not possible to determine which system is closer to the true concentration.

If we want to follow historical time series with a different device the only solution is to calibrate the data of both instruments by running them simultaneously under real-life sampling conditions. Most 3rd generation pollen monitoring systems are designed for 360° monitoring meaning that the data is perhaps not congruent with traditional Hirst-type pollen traps, which capture efficacy is dependent on wind speed and direction (Frenz, 2000).

We consider that having the possibility for pollen experts to retrospectively check each reported pollen event is a key capability of our pollen monitoring network. Long-range transport or Climate Change may result in unexpected pollen events outside the projected pollen calendars. Current automatic pollen monitors are not yet 100% accurate, and so unexpected pollen events could easily be deleted from supervised quality assured pollen reports. Only re-analysis of the raw information can prevent incorrect data auditing. Also, for quality control we found it extremely re-assuring to be able to visually inspect each captured pollen, as all pollen monitoring experts currently do. This enabled us to evaluate the instrument under real-life conditions. In ambient air, most particles (>90%) are NOT pollen. Having unknown particles present while evaluating correct pollen classifications greatly influences the results.

5. Conclusions

- The multiclass accuracy (correct identification of pollen taxa) of the BAA500 was >90%. Human classification of the same pollen was taken as 100%. For the main allergenic pollen from *Betula* and Poaceae, in the area, the accuracy for *Betula* pollen was 86% and for Poaceae 88%. The largest error of the automatic pollen monitors was not wrong classification but "missing" some classifications, i.e. classifying as unknown. The recognition of *Alnus* pollen could be improved, as they are sometimes recognized as *Corylus* pollen.
- In the case of *Salix*, we observed a high accuracy of BAA500 identifying this pollen type (>90%). However, BAA500 misclassified 1% of *Betula* as *Salix.* Given the low amount of *Salix*, the percent of automatics correct decreased below 50%.
- We believe that no pollen monitoring system currently has an accuracy of 100%. Thus, an essential feature of each system should be the possibility of checking whether a classification is correct (reanalysis of past samples).
- When comparing the BAA500 with Hirst-type pollen traps, we found that the precision of the automatic pollen monitors is higher. The discrepancies in identification between automatic pollen monitors are smaller than the discrepancies between humans (i.e. the differences on predictive value between BAA500 were smaller than the differences in predictive value between pollen experts).
- The reliability of the monitor (time on-line) was 97%.
- The ePIN network is currently trained to identify 40 pollen types. Evaluation was only done with pollen taxa that were sufficiently sampled.
- Pollen concentrations obtained from BAA500 and Hirst-type pollen traps show a significant correlation of 0.84.
- The correlations between daily concentrations recorded by the BAA500 and Hirst varies depending on pollen type.

Funding

This study was funded by the Bavarian State Ministry of Health and the Bavarian State Ministry of the Environment and Consumer Protection as project AP-2460-ePIN-V11-U1-D44520/2018 and project TKP01KPB-73815.

Disclosure of potential conflict of interest

Authors declare no conflict of interest. The authors declare they have no actual or potential competing financial interests. We do not have any financial relationship with the company producing the automatic pollen monitor BAA500, Helmut Hund GmbH.

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 the ZAUM team who managed the manual network: Elke Bartusel, Katherina Bartusel, Cordula Ebner v. Eschenbach, Christine Weil, Renate Ostner, Norbert Ertl and Gudrun Pusch. We also thank the assistants at each monitoring location: Margita Herauf, Stefan Kukula, Allison Kolar, Reinhard Kolb, Norbert Oestel, Tobias Haugg, Manuel Maurer, Dr. Franz Schuhbeck, Manuela Ho¨rath, Willi Go¨tz, Josef Dietrich, Daniel Ko¨ppl, Johannes Go¨tz, Paul Beyer. Our special acknowledgement goes to the manual pollen analysts: Mirjana S˘pehar, Anna Schubert, Orsolya Udvardy, Natasa Camprag-Sabo, Weronica Haratym, Gergely Ma´nyoki, Malgorzata Nowak, Łukasz Kostecki & Agata Szymanska. For the IT support we thank Stephan Hachinger (Leibnitz Rechenzentrum) & Robert Gebauer (external IT specialist). We thank to Dr. Matt Smith for all the suggestions reviewing this work.

References

Belda, M., et al., 2014. Climate classification revisited: from Ko¨ppen to Trewartha. Clim. Res. 59, 1–13.

Buters, J.T.M., et al., 2018. Pollen and spore monitoring in the world. Clin. Transl. Allergy 8, 9.

Chappuis, C., et al., 2019. Automatic pollen monitoring: first insights from hourly data. Aerobiologia 1–12.

Crouzy, B., et al., 2016. All-optical automatic pollen identification: towards an operational system. Atmos. Environ. 140, 202–212.

Frenz, D.A., 2000. The effect of windspeed on pollen and spore counts collected with the Rotorod Sampler and Burkard spore trap. Ann. Allergy Asthma Immunol. 85, 392–394.

Gal´an, C., et al., 2014. Pollen monitoring: minimum requirements and reproducibility of analysis. Aerobiologia 30, 385–395.

Grace, K., et al., 2018. When will AI exceed human performance? Evidence from AI experts. J. Artif. Intell. Res. 62, 729–754.

He, K., et al., 2015. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision 1026–1034.

Hirst, J.M., 1952. An automatic volumetric spore trap. Ann. Appl. Biol. 39, 257–265.

Huffman, J.A., et al., 2019. Real-time sensing of bioaerosols: review and current perspectives. Aerosol. Sci. Technol. 1–31.

I, SO-5725, 1994. ISO 5725: 1994: Accuracy (Trueness and Precision) of Measurement Methods and Results. International Organization for Standardization.

Kawashima, S., et al., 2017. Automated pollen monitoring system using laser optics for observing seasonal changes in the concentration of total airborne pollen. Aerobiologia 33, 351–362.

Mullins, J., Emberlin, J., 1997. Sampling pollens. J. Aerosol Sci. 28, 365–370.

O'Connor, D.J., et al., 2014. Using the WIBS-4 (waveband integrated bioaerosol sensor) technique for the on-line detection of pollen grains. Aerosol. Sci. Technol. 48, 341–349.

Oteros, J., et al., 2017. Errors in determining the flow rate of Hirst-type pollen traps. Aerobiologia 33, 201–210.
Oteros, J., et al., 2013. Quality control in bio-monitoring networks, Spanish Aerobiology Network. Sci. Total Environ. 443, 559–565.
Oteros, J., et al., 2015. Automatic and online pollen monitoring. Int. Arch. Allergy Immunol. 167, 158–166.
Oteros, J., et al., 2019. Building an automatic Pollen Monitoring Network (ePIN): selection of optimal sites by clustering pollen stations. Sci. Total Environ. 688, 1263–1274.
Sˇauliene˙, I., et al., 2019. Automatic pollen recognition with the Rapid-E particle counter: the first-level procedure, experience and next steps. Atmos. Meas. Tech. Discuss. 1–33, 2019.
Silver, D., et al., 2017. Mastering the game of Go without human knowledge. Nature 550, 354.
Smith, M., et al., 2019. An abbreviated method for the quality control of pollen counters. Grana 58, 185–190.
V, DI4252-4, 2016. Bioaerosole und biologische Agenzien – Ermittlung von Pollen und Sporen in der Außenluft unter Verwendung einer volumetrischen Methode für ein Messnetz zu allergologischen Zwecken. VDI-Richtlinie 4252 Blatt 4, Entwurf. VDI/DIN-Handbuch Reinhaltung der Luft, Band 1a. VDI/DIN, Beuth, Berlin.
Wetzlar, H., 2009. Pollen Monitor BAA500.