A comparison of classification algorithms within the Classifynder pollen imaging system

Ryan Lagerstrom, Yulia Arzhaeva, Leanne Bischof, Simon Haberle, Felicitas Hopf, and David Lovell

Citation: AIP Conference Proceedings **1559**, 250 (2013); doi: 10.1063/1.4825017
View online: http://dx.doi.org/10.1063/1.4825017
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1559?ver=pdfcov
Published by the AIP Publishing

A Comparison of Classification Algorithms within the Classifynder Pollen Imaging System

Ryan Lagerstrom[a], Yulia Arzhaeva[a], Leanne Bischof[a], Simon Haberle[b], Felicitas Hopf[b] and David Lovell[c]

[a]*CSIRO Computational Informatics. Locked Bag 17, North Ryde, NSW 1670, Australia*
[b]*School of Culture, History and Language, H C Coombs Blg 9, The Australian National University, Canberra, ACT 0200, Australia*
[c]*CSIRO Computational Informatics. GPO Box 664, ACT 2601, Australia*

Abstract. We describe an investigation into how Massey University's Pollen Classifynder can accelerate the understanding of pollen and its role in nature. The Classifynder is an imaging microscopy system that can locate, image and classify slide based pollen samples. Given the laboriousness of purely manual image acquisition and identification it is vital to exploit assistive technologies like the Classifynder to enable acquisition and analysis of pollen samples. It is also vital that we understand the strengths and limitations of automated systems so that they can be used (and improved) to compliment the strengths and weaknesses of human analysts to the greatest extent possible. This article reviews some of our experiences with the Classifynder system and our exploration of alternative classifier models to enhance both accuracy and interpretability. Our experiments in the pollen analysis problem domain have been based on samples from the Australian National University's pollen reference collection (2890 grains, 15 species) and images bundled with the Classifynder system (400 grains, 4 species). These samples have been represented using the Classifynder image feature set. In addition to the Classifynder's native neural network classifier, we have evaluated linear discriminant, support vector machine, decision tree and random forest classifiers on these data with encouraging results. Our hope is that our findings will help enhance the performance of future releases of the Classifynder and other systems for accelerating the acquisition and analysis of pollen samples.

Keywords: Pollen, classification, automation, palynology

INTRODUCTION

Palynologists study samples of particulates such as pollen grains to gain an understanding of the environment under which they are produced. Among other things palynology enables vegetation and climate reconstruction for the assessment of climate change and biodiversity [1,2]. It also underpins the science in areas from allergy research to plant reproductive biology [3,4].

However, the analysis of pollen is a slow and laborious task that involves manually preparing samples, locating and identifying pollen grains under a microscope and finally, quantifying the abundance of various species present in any sample. The palynology community recognizes the need for automation and the role it could play in accelerating the science in these areas. To this end there have been several efforts towards developing systems for automated pollen analysis [5,6,7].

2013 International Symposium on Computational Models for Life Sciences
AIP Conf. Proc. 1559, 250-259 (2013); doi: 10.1063/1.4825017

The Pollen Classifynder [8], developed by Massey University, integrates the hardware and software required to locate, image and classify slide based pollen samples. It combines technologies from microscopy, robotics, pattern recognition, image processing and data science to form an automated pollen analysis system that addresses the needs of palynologists working in labs that deal with pollen counting and classification. Typically a palynologist would build up a library of various pollen species using the Classifynder to image samples and then manually classifying grains. That library can then be used to build a classifier. For example, a palynologist interested in historical biodiversity would build up a library by examining archeological pollen samples and labeling species from the area in question. The library could then be used to train a classifier to assist in larger studies.

The choice of classifier is a very important issue not only in terms of accuracy, but interpretability of results. The Classifynder employs a neural network classifier to perform its classification tasks. Discussion with the Classifynder developers revealed some possible shortcomings in this approach. Because the native neural net strategy does not provide a measure of error for each pollen grain classification, there is no way to streamline a review of classification results. Being able to review (and correct) very obvious misclassifications (such as when a particular species is not in the library or the corruption of a grain observation) would allow for quick improvements in classification results and accelerate the phenotyping process. In palynology Stillman and Fenley [9] recognised a need to investigate classifier choice as early as 1996. Zhang et al. [10] report excellent performance of the neural network classifier based on the features measured on each pollen grain within the Classifynder for five species.

In this paper we assess five classifiers that are typically used in modern data analytics with regard to both accuracy of classification and interpretability of resulting classifications: neural networks [12]; linear discriminant analysis [13]; support vector machines [14]; decision trees [15]; and random forests [16].

DATA

The Classifynder's digital microscopy and software system produces 43 characteristic features for each pollen grain detected in a microscopy slide. The camera scans the slide in low resolution looking for candidate pollen grains. Candidate grains are at this point assessed as to whether they are debris or genuine pollen grains. Once a candidate is deemed to be a genuine pollen grain, a high resolution image is taken at nine different focal depths and a composite image is created. The composite image is converted into hue, lightness and saturation space. Using only the lightness values, the pollen grain is segmented from the background by an edge detector followed by filling the interior. Image feature measurements are then computed from the segmented shape and the lightness values within the shape. The image feature categories, and the number of features are: Geometry (3) Histogram (2) Moments (7) Grey Level Co-occurrence Matrix (5) Grey Gradient Co-occurrence Matrix (12) Gabor (8) and Wavelets (6). More details of these image features are available in Zhang et al. [10].

TABLE 1. A summary of data used in this paper. The first column is the species name. The second column is the data source, ARC – Australian National University Reference Collection and CTS – Classifynder Test Set. Column three is an abbreviation for species. The final column is the number of image samples.

Species Name	Source	Abbreviation	#Images
Acacia Ramoissima	ARC	AR	77
Atriplex Paludosa	ARC	AP	341
Asteraceae	CTS	AS	100
Casuarina Littoralis	ARC	CL	172
Disphyma	CTS	DI	100
Dracophyllum	CTS	DR	100
Euphorbia Hirta	ARC	EH	172
Eucalyptus Fasciculosa	ARC	EF	192
Isoetes Pusilla	ARC	IP	715
Myrsine	CTS	MY	100
Nothofagus Cunninghamii	ARC	NC	113
Nothofagus Discoidea EV	ARC	NE	172
Nothofagus Discoidea PV	ARC	NP	504
Olearia Algida	ARC	OA	121
Phyllocladus Aspleniifolius	ARC	PA	122

Two data sets were available to us to gain a better understanding of the classification capabilities of the Classifynder. The first was provided by Massey University and contains 400 pollen grain images from 4 different species. This data set comes bundled with the Classifynder system to help users gain an understanding of how the analysis part of the system operates. The second was provided by the Australian National University from their pollen reference collection and contains 2980 pollen images from 11 species. The 11 species were selected as common to the Canberra region in Australia. Table 1 summarises the data set while Figure 1 shows example images from all species.

Initial exploratory analysis was carried out to assess the correlation structure between the image feature measurements and to see if there were outlying observations amongst the data. To investigate the correlation structure of our feature set, all observations with labelled species were considered. Looking at the correlation structure and ignoring the species labels could disguise potential discriminability, so correlation between features was examined within species and the minimum over species considered. Using this conditional correlation type approach, it was found that within species correlation differed from overall correlation for one species only, Myrsine. The conditional correlation between seven Grey Gradient Co-occurrence Matrix (GGCM) texture measures was 0.98 or above, while the two first level wavelet features was also 0.98. With the data sets we have, removing five of the GGCM features and one of the first level wavelet features may lead to more simple classification models with improved parameterisations. We assessed classifier performance with and without removing correlated features.

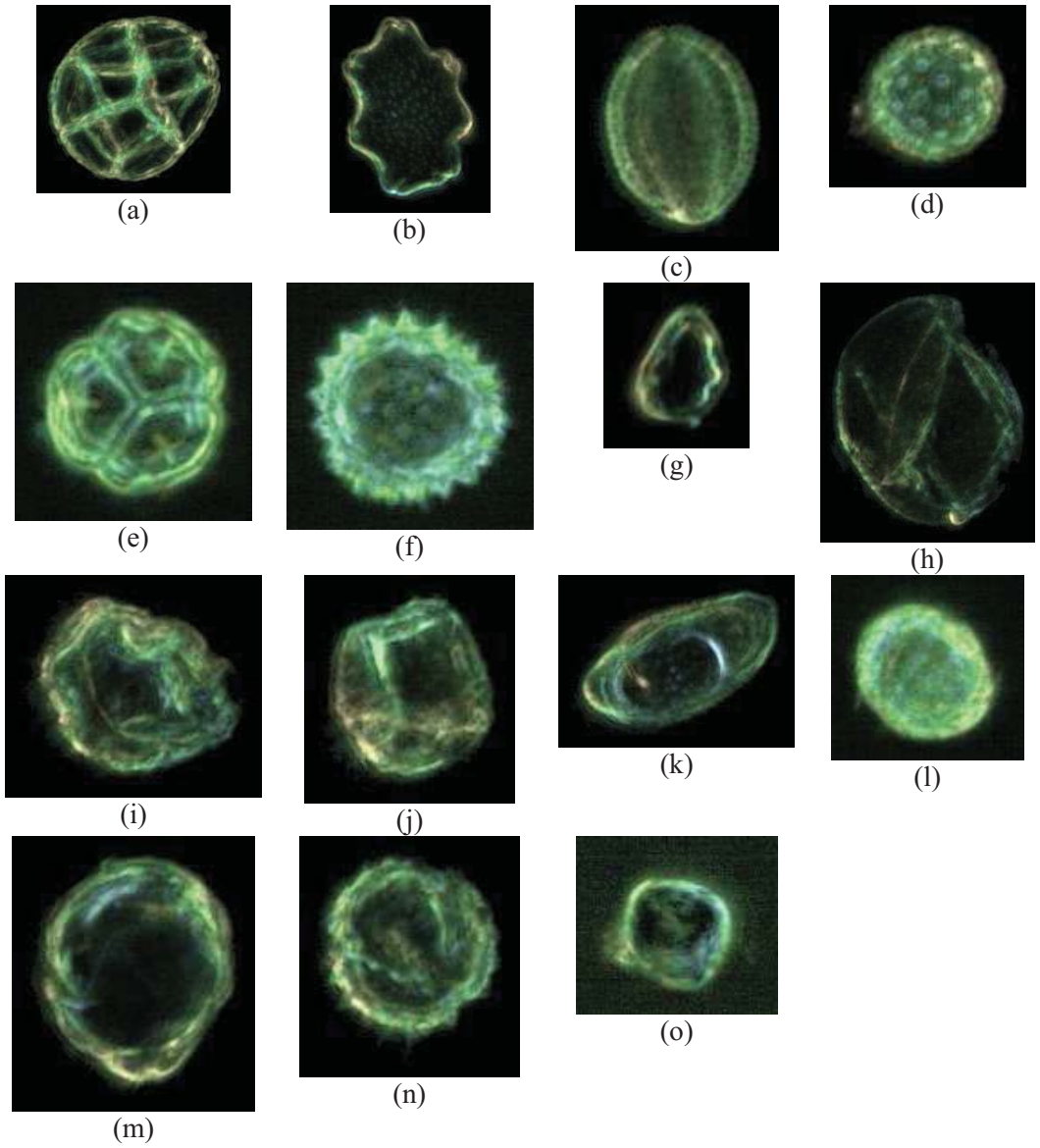

FIGURE 1. Sample images from various species. (a) Acacia Ramoissima, (b) Nothofagus Discoidea EV, (c) Euphorbia Hirta, (d) Atriplex Paludosa, (e) Drocophyllum, (f) Asteraceae, (g) Eucalyptus Fasciculosa (h) Nothofagus Cunninghamii, (i) Isoetes Pusilla, (j) Phyllocladus Aspleniifolius, (k) Nothofagus Discoidea PV (l) Disphyma, (m) Casuarina Littoralis, (n) Olearia Algida and (o) Myrsine .

The influence of outliers on classification depends on the classifier. In an attempt to identify potential outliers in our training sets, we first scaled each variable to have zero mean and unit standard deviation. We found 20 observations with absolute value greater than 10 standard deviations from zero. A selection of five outlying images is shown in Figure 2. They are based on geometry (NC), histogram (IP), moments (EF), GLCM (NC) and Gabor (MY). Because we found just 20 observations from 3290 that

could be characterised as outliers we take the approach of not discarding outliers as the impact on classifier accuracy would be minor.

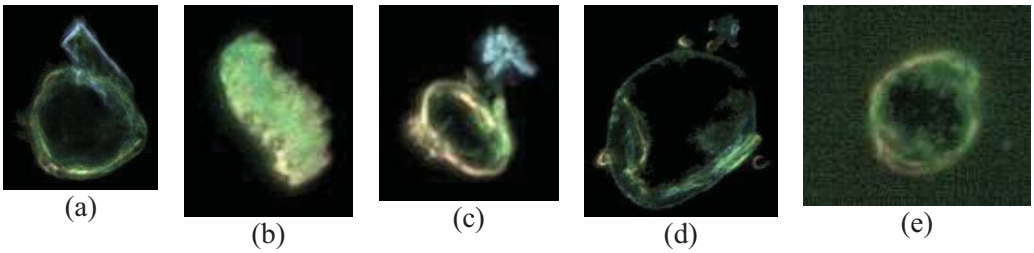

FIGURE 2. Five sample outlying images from different image features. (a) Geometry, (b) Histogram, (c) Moments, (d) Grey Level Co-occurrence Matrix and (e) Gabor.

CLASSIFICATION MODELS

Here we assess five classification models on the labelled species images: neural networks (NN), linear discriminant analysis (LDA), support vector machines (SVM), decision trees (DT) and random forests (RF). This selection of models was chosen to span linear, non-linear and tree based classifiers and represents a typical set of tools a data analyst might use to investigate classification problems in, for example, the data mining area. The analyses were performed in the statistical programming language R [17] where the classifiers are available in the nnet, MASS, kernlab, rpart and randomForest packages.

Our motive for comparing the classifiers is to ascertain whether there is a particular classifier that is especially well suited to pollen data compared to others. The developers of the system have indicated that one of the shortcomings of the neural net is the inability to measure the strength of individual classifications for the purpose of assisted reviewing. With this in mind, a question of particular relevance is how a simple linear classifier (where per observation diagnostics are available) compares to a 'black box' classifier like a neural net.

Our strategy for assessing classifier performance begins by choosing one of the three data sets: the Classifynder test set (CTS), the ANU reference collection (ARC) or the combination of both (COMB). From there we consider a data set where all feature measurements were included (FF) and also when correlated feature measurements were removed (LF). Once the data set was determined, all feature measurements were scaled to have zero mean and unit variance. The data set was then randomly split into equally sized training and test sets. Then each of the five classifier models was built using the training data. Confusion matrices were formed based on the test set and an error measure computed. A performance measure was defined as the number of correct classifications divided by the total number of image observations. The test data was then used for training and the training data for testing in a 2-fold cross validation. This process was repeated ten times and the average performance measure and confusion matrix calculated.

The neural network used was the feed-forward with single hidden layer network. The number of units in the hidden layer was set to 3, initial random weights set to 0.1

with decay 0.0005 and the maximum number of iterations equal to 600. The linear discriminant analysis model used all of the input features (i.e. we did not attempt dimension reduction via principal components or other means). The support vector machine used the C classification model. A Gaussian radial basis kernel was employed with a sigma equal to 0.1 while the cost of constraints violation parameter, C, was set to 10. For the decision tree model, no surrogates were used in the splitting process. For the random forest model, the number of trees parameter was set to 500 while the number of variables randomly sampled as candidates at each split was set to 3.

MODEL PERFORMANCE

Model performance for the five classifier models on the six data sets is summarized in Table 2. The most obvious issue at first glance is the poor relative performance of the DT model on the ARC and COMB data sets. The DT model's performance is comparable to the other models on the CTS data. This may indicate the DT model is not suitable for classification when the number of species is larger. Of the other models, the LDA, SVM and NN models have the best performance over the six data sets. The SVM model outperforms the others on the both forms of the ARC data while the LDA approach outperforms the others on the both forms of the COMB data and the reduced feature form of the CTS data. However, in terms of the performance measure with this data, the difference between models is slight, apart from the DT model.

TABLE 2. This table summarizes the performance of the 5 classification models over the 6 data sets. The performance measure is the sum of the diagonal elements of the corresponding confusion matrix divided by the number of observations. The underlined elements correspond to the best performance for each data set.

Data		NN	LDA	SVM	DT	RF
ARC	**FF**	0.80	0.82	<u>0.83</u>	0.74	0.81
	LF	0.79	0.83	<u>0.84</u>	0.74	0.82
CTS	**FF**	<u>0.97</u>	0.96	0.93	0.94	0.95
	LF	0.97	<u>0.98</u>	0.93	0.93	0.96
COMB	**FF**	0.80	<u>0.84</u>	0.82	0.75	0.83
	LF	0.80	<u>0.83</u>	0.82	0.74	0.82

A full confusion matrix for the NN model on the COMB data with all features is shown in Table 3. The values in the table are percentages with rows corresponding to the truth and columns to classification results, so the sum for a particular row should be 100. The NN model is used natively in the Classifynder system. Firstly, looking at the diagonal elements, the CL, OA and PA species are poorly classified with success rates under 60% and PA in particular at 4%. These 3 species are most frequently confused with the IP species which has the highest number of observations, 715, in the data. The number of observations for CL, OA and PA are 172, 121 and 122 respectively. The morphology and texture of the images from these species are the most similar amongst the species in the data set. Similar observations can be made from the confusion matrices for the SVM and RF models and so their confusion matrices are not displayed here. Table 4 shows the full confusion matrix for the LDA

model on the COMB data. The two species with the lowest classification accuracy are PA with 69% and IP with 72%. The high number of IP observations accounts for most of the model's overall classification error. Another feature of the confusion matrix is that all species other than AR are confused with the PA species. Despite this, the error is more balanced between species which would appear to be a desirable result.

TABLE 3. Full confusion matrix for the NN model on the COMB FF data. The rows correspond to ground truth while the columns represent the classifications results.

	AP	AR	AS	CL	DI	DR	EF	EH	IP	MY	NC	NE	NP	OA	PA
AP	87	0	0	2	0	0	3	4	0	0	0	0	0	3	1
AR	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
AS	1	0	79	0	2	12	0	2	0	0	0	4	0	0	0
CL	0	0	0	49	0	0	0	0	34	1	0	0	15	2	0
DI	0	0	0	0	98	0	0	0	0	2	0	0	0	0	0
DR	0	0	4	0	3	93	0	0	0	0	0	0	0	0	0
EF	3	0	0	0	0	0	80	0	11	0	0	2	0	1	3
EH	0	0	0	1	0	0	0	93	0	4	0	2	0	1	0
IP	1	0	0	2	0	0	2	1	88	0	1	2	1	2	1
MY	4	0	2	0	11	0	0	1	0	80	0	0	2	0	0
NC	0	0	0	0	0	0	0	2	4	0	86	2	4	3	0
NE	1	0	0	6	0	0	0	2	8	0	0	74	9	0	0
NP	0	0	0	2	0	0	0	0	2	0	0	2	93	0	0
OA	3	0	0	0	0	0	3	0	36	0	0	0	0	57	3
PA	7	0	0	5	0	0	7	3	63	0	0	1	0	10	4

TABLE 4. Full confusion matrix for the LDA model on the COMB FF data. The rows correspond to ground truth while the columns represent the classifications results.

	AP	AR	AS	CL	DI	DR	EF	EH	IP	MY	NC	NE	NP	OA	PA
AP	84	0	0	0	0	0	4	3	1	0	0	1	0	2	5
AR	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
AS	0	0	92	0	0	5	0	0	0	0	0	0	0	0	3
CL	0	0	0	78	1	0	3	0	10	0	0	0	2	0	5
DI	0	0	0	0	98	0	0	0	0	0	0	0	0	0	2
DR	0	0	9	0	0	88	0	0	0	0	0	0	0	0	3
EF	1	0	0	3	0	0	83	0	4	0	0	0	0	5	5
EH	0	0	0	0	0	0	0	94	1	0	0	0	0	2	3
IP	1	0	0	7	0	0	3	0	72	2	0	4	0	3	8
MY	0	0	0	0	0	0	0	2	0	93	0	0	0	0	6
NC	0	0	0	4	0	0	0	1	0	1	81	2	12	0	1
NE	0	0	0	10	0	0	0	0	4	0	0	77	7	0	2
NP	0	0	0	3	0	0	0	0	2	0	0	2	92	0	0
OA	4	0	0	0	0	0	0	0	1	0	0	0	0	89	6
PA	2	0	0	4	3	0	0	0	11	0	0	3	2	7	69

SEMI AUTOMATED PERFORMANCE ENHANCEMENT

In practice, the classification results need not be the end point of an investigation. Typically a palynolgist would review and adjust the classification results. The LDA model presents a simple means for assisting the review stage. LDA works by transforming the data into an optimal space for discrimination. For each species, a

mean value for each discriminant is then calculated. The model then measures the Mahalanobis distance between a sample and the set of mean discriminants for each species, with the lowest distance informing the choice of species for classification. This distance itself provides a measure of how far from the training data a particular sample is. Taking the ratio of lowest over the second lowest distance score gives a measure of how "borderline" a classification decision is. So, for example, a ratio close to 0 would indicate a strong decision while a value close to 1 would indicate possible confusion. Table 5 shows the top 20 classification results for the LDA model on the COMB FF data ranked on decreasing values of this ratio. It shows that out of the 20 results only 4 are correctly classified. So in the context of reviewing the data, if a user were to sort their observations based on the ratio, they could easily and efficiently adjust decisions for the most borderline cases. The ratio is similar in spirit to the posterior probability which can be calculated for the LDA models and the RF models. For each classification, a posterior probability is assigned for each class. It is then possible to use this to rank the data in a similar fashion to the ranking, noting that a posterior probability close to 1 corresponds to a strong decision. This would allow the user to use the RF model to perform a similar type of assisted review.

TABLE 5. Worst 20 classification results for the LDA model on the COMB FF data based on distance ratio.

Predicted Species	Species	Distance	Ratio
NP	CL	3.86	0.99
DI	CL	13.84	0.99
PA	IP	3.72	0.99
NE	MY	7.43	0.99
AS	DR	6.79	0.99
OA	IP	2.60	0.99
OA	EF	4.08	0.99
NP	NE	4.99	0.99
CL	IP	3.85	0.99
EH	PA	3.44	0.99
IP	IP	5.21	0.99
IP	IP	4.21	0.99
CL	IP	2.72	0.99
PA	PA	2.48	0.99
PA	IP	2.16	0.99
EH	AP	3.85	0.99
NP	NP	3.28	0.99
IP	NP	3.88	0.99
IP	EF	3.45	0.99
PA	OA	6.58	0.99

Another, more automated approach is to simply exclude a proportion of the classification results based on the ratio. For example, after ranking the results on decreasing values of the ratio, one can exclude the worst N percent of the results. For the LDA model on the COMB FF data, if we exclude 20 percent of the results based on this strategy, the overall performance of the classifier increases to 0.94. This compared to the performance 0.84, when all data is used, is a significant increase. If we excluded 50 percent of the data the performance increases to 0.99. However, when one examines the confusion matrix corresponding to only 50% of the data, the relative

proportions of the species are modified in a reasonably substantial way. For example, the IP, NP and AP species account for 18, 15 and 10 percent of the species present in the COMB FF data, respectively. When we exclude the worst 20% of the data, the relative percentages are 16, 16 and 11 which is not too dissimilar. However, when we exclude 50% of the data, the relative percentages are 5, 23 and 13 which is very different to the known abundances. So if the goal of a palynologist is to study relative abundance of species in a sample, one would need to find an appropriate percentage for exclusion which would preserve relative abundance.

CONCLUSION

We investigated the classification possibilities of data generated by the Classifynder, an automated imaging system for analysing pollen which can locate, image and classify slide based pollen samples. Given pollen's importance, abundance and diversity in nature, it is vital that automated systems for pollen analysis are developed and used in order to overcome the burdens of a historically manually intensive process. We looked at linear models, non linear classification and tree based classifiers from a performance and interpretability point of view. Our findings suggest that in terms of performance, the various models achieved reasonably similar results. However, we also discussed how a conceptually simple classifier like linear discriminant analysis can be exploited to review classification results in a semi-automated or automated manner. By ordering the classification results based on a metric describing how borderline a classification result is, users can efficiently delete or adjust results where classification is questionable. We also outlined an approach to automating this process by sub-setting the results based on this ordering. The benefits of taking this approach not only allow palynologists to increase their accuracy and confidence in their findings, but also accelerate the pollen phenomics process.

REFERENCES

1. C. A. Woodward and J. Shulmeister, "New Zealand chironomids as proxies for human-induced and natural environmental change: transfer functions for temperature and lake production (chlorophyll a)" in *Journal of Paleolimnology*, 36(2006,) pp. 406-429.
2. B.V. Alloway, D.J. Lowe, D.J.A. Barrell, R.M. Newnham, P.C. Almond, P.C. Augustinus, N.A. Bertler, L. Carter, N.J. Litchfield, M.S. McGlone, J. Shulmeister, M.J. Vandergoes, P.W. Williams, NZ-INTIMATE members, "Towards a climate event stratigraphy for New Zealand over the past 30,000 years (NZ-INTIMATE project)", in *Journal of Quaternary Science*, 22 (2007), pp. 9–35.
3. M.P. De Sa-Otero, A.P. Gonzalez, M. Rodriguez-Damian, E. Cernadas, "Computer-aided identification of allergenic species of Urticaceae pollen" in *Grana,* 43 (2004), pp. 224–230.
4. C.M. Costa, S. Yang, "Counting pollen grains using readily available, free image processing and analysis software" in *Annals of Botany,* 104 (2009), pp. 1005–1010.
5. I. France, A.W.G. Duller, G.A.T. Duller, H.F. Lamb, "A new approach to automated pollen analysis", in *Quaternary Science Reviews*, 19 (2000), pp. 537–546
6. A. Boucher, P.J. Hidalgo, M. Thonnat, J. Belmonte, C. Galan, P. Bonton, R. Tomczak, " Development of a semi-automatic system for pollen recognition", in *Aerobiologia,* 18 (2002), pp. 195–201.
7. O. Ronneberger, E. Schultz, H. Burkhardt, "Automated pollen recognition using 3D volume images from fluorescence microscopy", in *Aerobiologia*, 18 (2002), pp. 107–115.

8. K. Holt, G. Allen, R. Hodgson, S. Marsland, J. Flenley, "Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory", in *Review of Palaeobotany and Palynology*, 167,(2011), pp. 175–183.
9. E.C. Stillman, J.R. Flenley, "The needs and prospects for automation in palynology", in *Quaternary Science Reviews*, 15 (1996), pp. 1–5.
10. Y. Zhang, D.W. Fountain, R.M. Hodgson, J.R. Flenley, S. Gunetileke, "Towards automation of palynology 3: pollen pattern recognition using Gabor transforms and digital moments", in *Journal of Quaternary Science*, 19 (2004), pp. 763–768.
11. G.H. Joblove and D. Greenberg, "Color spaces for computer graphics", in *Computer Graphics (SIGGRAPH '78 Proceedings)*, (1978) 12(3):20–25.
12. J.A. Hertz, A. Krogh and R.G. Palmer, "An Introduction to the Theory of Neural Computing". (1991) Addison-Wesley Publishing Company.
13. R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems", in *Annals of Eugenics*, (1936) 7(2), pp179-188.
14. C. Cortes and V. Vapnik, "Support-vector networks", in *Machine Learning*, (1995) 20(3), pp273-297
15. L. Breiman, J.H. Friedman, R.A. Olschen and C.J. Stone, "Classification and Regression Trees", (1999) Chapman Hall
16. L.Breiman, "Random Forests", in *Machine Learning*, (2001) 45(1), pp 5-32.
17. R Development Core Team, "R: A language and environment for statistical computing", *R Foundation for Statistical Computing"*, (2010) Vienna, Austria.