2- and 3-way analysis of near infrared scans from seed crossings

Torbjörn A. Lestander,[a] Paul Geladi[b] and Per Christer Odén[a]

[a]*SLU Forest Seed Science Centre, Department of Silviculture, Swedish University of Agricultural Sciences, SE – 901 83 Umeå, Sweden*

[b]*Department of Chemistry, Umeå University, SE – 901 87 Umeå, Sweden*

Introduction

Seed quality from a physiological and genetic viewpoint is of major interest in plant breeding and reforestation programmes. Fast and automatic selection of good tree seeds is a vision within forest seed management programmes.[1] One aspect that needs studying within this vision is the possibility of collecting genetic information about the seeds using non-destructive methods such as near infrared (NIR) spectroscopy. In this paper we have used Scots pine (*Pinus sylvestris* L.) full sib seeds measured by NIR reflectance spectroscopy. This study is, to our knowledge, one of the first of its kind and only a few studies have explored the field to collect genetic information directly from NIR spectra,[2] whereas there are many studies that utilise the calibrations between NIR spectra and defined traits [3–6] of interest in plant [7–12] and animal[13] breeding.

In a Scots pine seed, the embryo, which is of both maternal and paternal origin, is embedded in a megagametophyte surrounded by a seed coat. The two latter components are in gymnosperms of only maternal origin. These components have been analysed[14] in Scots pine giving the ratio 25 : 66 : 9 between the mass of seed coat, megagametophyte and embryo, respectively. Thus, seed tissue of maternal origin is expected to have a high influence on NIR spectra, whereas there can be problems in collecting information about tissues of paternal origin.

The aim of the study was to use 2- and 3-way analysis to determine maternal, paternal and interaction effects in NIR spectra from controlled seed crossings.

Material and methods

Genetic material

The crossings emanated from a factorial mating design of Scots pine (*Pinus sylvestris* L.) using six maternal clones and ten paternal clones growing in clonal tree archives in northern Sweden. Three samples (one sample and two re-samples) containing *ca* 30 seeds from each of the 60 controlled crossings (full sib families) were measured by spinning cup NIR reflectance spectroscopy.

Spectra

An NIR spectrophotometer (Foss NIRSystems model 6500) was used to collect reflectance (R) from each seed sample as the mean spectrum of 32 scans at every 2nd wavelength between 400 and 2498 nm, *i.e.* 1050 wavelengths. Absorbance (A) was calculated as $A = \log(1/R)$ for each wavelength.

Near Infrared Spectroscopy: Proceedings of the 10th International Conference
© IM Publications Open LLP 2002

Data treatment

The absorbances at 1050 wavelengths were replaced by seven significant ($p < 0.05$) latent variables. The latent variables were principal components of the 180 × 1050 data set and calculated by the software SIMCA[15] after mean-centring.

The used models for univariate (ANOVA, random effects) and multivariate (MANOVA, fixed effects) analysis of variance were:

$$z_{ijk} = b_{0k} + M_{ik} + F_{jk} + M_{ik} \times F_{jk} + E_{ijk}$$

where at the *k*th principal component z_{ij} is the score value of the *ij*th seed crossing, b_0 is the intercept, M_i is the contribution from the *i*th mother, F_j is the contribution from the *j*th father, $M_i \times F_j$ is the interaction between mother *i* and father *j* and E_{ij} is the residual. The model was calculated in SPSS[16] and was one of the possible interpretations of the genetic background information.

For the parallel factor analysis [17] (PARAFAC) the replicates were replaced by their mean spectra. The Savitzky–Golay (window size 31, order 4, 1st derivative) derivative smoothing and the PARAFAC analysis were carried out in MATLAB[18] using the PLS-Toolbox.[19] Reduction of wavelengths was not used. Another PARAFAC model was also made using mean-centred raw data.

A three-way data array has three ways or modes. The array used in this paper has the modes; mother, father and wavelength. PARAFAC allows the decomposition of the 6 × 10 × 1050 array in factors (called loadings) representing the mothers, the fathers and the wavelengths and gives a residual three-way array. The number of factors used is the three-way pseudorank.

Results

The NIR spectra all look the same by visual inspection. Any interpretation of the raw spectra only relates to common properties of the seeds and not to genetic differences.

The seven PCA components used in the analysis of variance explained 99.6% of the sum of squares of the NIR data. The random effect ANOVA-model on each of the seven principal components showed that the maternal effect was highly ($p > 0.001$) to moderately ($p > 0.01$) significant for the first six principal components. The interaction effects were highly significant ($p > 0.001$) for all components and the paternal effect was only significant ($p > 0.05$) for the second component. For the fixed effect MANOVA model on the seven components together, highly to moderately significant maternal, paternal and interaction effects were found for all components.

The PARAFAC analysis on the first-order smoothing derivative gave a model explaining almost 100% of the total sum of squares of the data. A pseudorank of three was considered meaningful. The first major component explained the average spectral content of all the seed samples. The second and third components were the main ones used for showing genetic differences. The results of the three-way analysis are best seen in loading plots (Figures 1 and 2).

Because of the huge loading for the first PARAFAC component, a new model was made

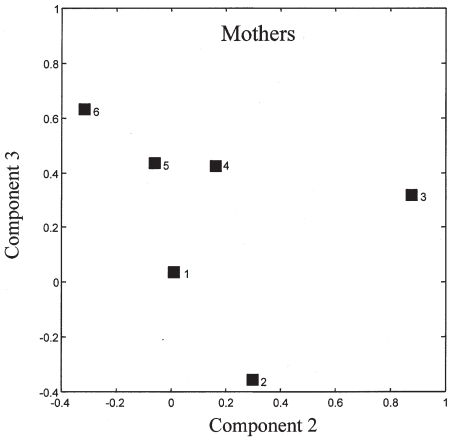

Figure 1. Second and third PARAFAC loading plot for maternal influence (mothers) on NIR spectra from controlled Scots pine seed crossings.

© IM Publications Open LLP 2002

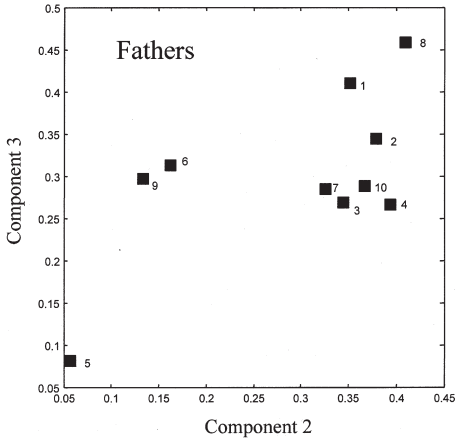

Figure 2. Second and third PARAFAC loading plot for paternal influence (fathers) on NIR spectra from controlled Scots pine seed crossings.

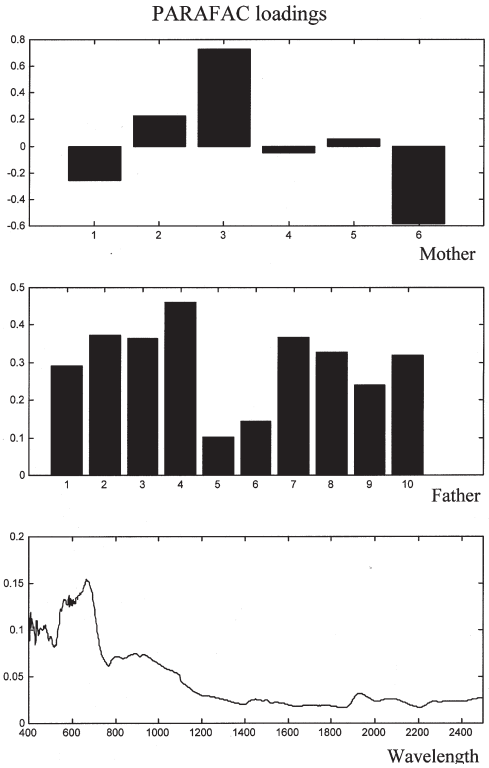

Figure 3. The PARAFAC loadings for the model on mean-centred NIR data from controlled Scots pine seed crossings.

using mean-centred raw data. This model gave one component explaining 70% of the sum of squares (Figure 3).

Discussion

The ANOVA and MANOVA on the seven significant principal components showed the importance of the maternal contribution to the seeds, which was expected, but there were also significant paternal and interaction contributions. This is in accordance with the genetic background knowledge.

Both PARAFAC models showed similar results. There was a huge influence of the mean spectrum and this had nothing to do with genetic influences. The PARAFAC loadings allowed visualisation of mainly maternal, but also of paternal influences. Because of the nature of the model, these remain mixed with the interactions.

Three-way analysis gives an advantage compared to doing a principal component analysis on a 60 × 1050 matrix. The PARAFAC decomposition is more parsimonious and the maternal and paternal effects are separated in the loadings. PARAFAC was used on NIR data earlier.[20,21] It has been shown [21] that raw NIR data are not ideal for PARAFAC analysis and that Savitzky–Golay smoothing derivatives are much better.

The loading plot of wavelengths showed a high contribution from the visual wavelength range, which is mainly an effect by the maternal seed coat. The region where NIR-radiation penetrates deepest into organic matter seems to have a high contribution to the PARAFAC model.

We have shown that it is possible to determine maternal and paternal influence as well as interactions of maternal and paternal origin in NIR spectra on seeds from controlled seed crossings. This finding opens up new possibilities in plant breeding and seed separation. The bulk of the seed has high

© IM Publications Open LLP 2002

maternal influences as expected, but only the embryo inside the seed has the detailed genetic information. Therefore, NIR imaging or confocal microscopy of single seeds is expected to be useful for extracting that information in more detail.

Acknowledgement

The authors acknowledge financial support from the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning and the Unizon-project NIRCE within the EU Interreg Kvarken-MidScandia cooperation. We also thank associate professors Bengt Andersson and Jan-Erik Nilsson for lending us the seed crossings for non-destructive NIR measurements and finally Margareta Söderström and Josefina Nyström for carrying out the NIR measurements.

References

1. T.A. Lestander and P.C. Odén, *Seed Science and Technology* **30,** 383 (2002).
2. B.L. Greaves, L.R. Schimleck, N.M.G. Borralho, A.J. Michell and C.A. Raymond, *Appita Journal* **49(6),** 423 (1996).
3. B.G. Osborne, T. Fearn and P.H. Hindle, *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis*, 2nd Edition. Longman Scientific & Technical, Harlow, UK (1993).
4. S.J. Lister, M.S. Dhanoa, J.L. Stewart and M. Gill, *Animal Feed Science and Technology* **86,** 231 (2000).
5. A. Reverter, D.J. Johnston, H.-U. Graser, M.L. Wolcott and W.H. Upton, *J. Animal Science* **78,** 1786 (2000).
6. L. Velasco and H.C. Becker, *Genetic Resources and Crop Evolution* **47(3),** 231 (2000).
7. B.E. Zher, S.R. Eckhoff, W.E. Nyquist and P.L Keeling, *Crop Science* **36,** 1159 (1996).
8. S.R. Delwiche, R.A. Graybosch and C.J. Peterson, *Cereal Chem.* **76(2),** 255 (1999).
9. K.G. Campbell, C.J. Bergman, D.G. Gualberto, J.A. Anderson, M.J. Giroux, G. Hareland, R.G. Fulcher, M.E. Sorrells and P.L. Finney, *Crop Sci.* **39,** 1184 (1999).
10. M.Q. Lu, L. O'Brien and I.M. Stuart, *Australian J. Agric. Res*. **50,** 1425 (1999).
11. M.Q. Lu, L. O'Brien and I.M. Stuart, *Australian J. Agric. Res*. **51,** 247 (2000).
12. M.Q. Lu, L. O'Brien and I.M. Stuart, *Australian J. Agric. Res*. **52,** 85 (2001).
13. G.D. Snowder, J.W. Walker, K.L. Launchbaugh and L.D. van Vleck, *J. Animal Science* **79,** 486 (2001).
14. P.B. Reich, J. Oleksyn and M.G. Tjoelker, *Canadian Journal of Forest Research* **24,** 306 (1994).
15. Umetrics AB, *Simca-P8.0*, Umetrics AB, Umeå, Sweden.
16. SPSS Inc., *SPSS for Windows 10.0.5*, SPSS Inc. Headquarters, Chicago, Illinois, USA (1999).
17. R. Coppi and S. Bolasco (Eds), *Multiway Data Analysis*. North Holland, Amsterdam, The Netherlands (1989).
18. The MathWorks Inc., *Matlab* 5.3.0, The MathWorks, Inc., Natick, MA, USA (1999).
19. B. Wise and N. Gallagher, *PLS Toolbox 2.0 for use with MATLAB*, Eigenvector Research, Manson, WA, USA (1998).
20. N. Allosio, P. Boivin, D. Bertrand and P. Courcoux, *J. Near Infrared Spectrosc.* **5,** 157 (1997).
21. P. Geladi and P. Åberg, *J. Near Infrared Spectrosc.* **9,** 1 (2001).

© IM Publications Open LLP 2002