TITLE Selection of calibration sub-sets to predict ryegrass quality using principle component analysis for near infrared spectroscopy

AUTHORS G.A. Burns, P. O'Kiely, and T.J. Gilliland

SELECTION OF CALIBRATION SUB-SETS TO PREDICT RYEGRASS QUALITY USING PRINCIPLE COMPONENT ANALYSIS FOR NEAR INFRARED SPECTROSCOPY

G.A. BURNS[1,2], P. O'KIELY[1], and T.J. GILLILAND[3]

[1]AGRIC, Teagasc, Grange, Co. Meath, Ireland; [2]Horticulture, Plant Breeding & Land Use Unit, Agri-Food and Biosciences Institute; Co. Armagh, N. Ireland; [3]Agriculure Branch, Agri-Food and Biosciences Institute, N. Ireland.

INTRODUCTION Near infrared reflectance spectroscopy (NIRS) has become the routine method of assessing forage quality on grass evaluation and breeding programmes. NIRS requires predictive calibration models that relate spectral data to reference values developed using a calibration set (Burns *et al.* 2013). The samples that form the calibration set influence the accuracy and reliability of these models and need to be representative of samples that will likely be analysed (Shenk and Westerhaus, 1991). Analysing samples from the calibration set using reference techniques has a significant cost and time associated and needs to be considered in the context of the desired accuracy and robustness of calibration models. Calibration selection techniques can therefore maximise the accuracy and robustness of calibration models whilst reducing the number of samples requiring reference analysis. One such method is principle component analysis (PCA; Shenk and Westerhaus 1991) whereby Shetty *et al.* (2012) reported that the number of samples could be reduced by up to 80% with a minimal loss in accuracy of calibration model. PCA selects representative calibration sub-sets through plotting all the samples in hyper-dimensional space, based on spectral data, and a sample is selected to represent a local neighbourhood cluster of samples for reference analysis. The aim of this research was to assess the accuracy of NIRS calibration models for buffering capacity, *in vitro* dry matter digestibility (DMD) and water soluble carbohydrate (WSC) content developed using calibration sub-sets selected by PCA.

MATERIALS AND METHODS Dried, milled (n = 2076) ryegrass samples were derived from harvested plots from DAFM recommended list trials at Backweston, Co. Kildare. Each sample was analysed for buffering capacity, *in vitro* DMD and WSC content and spectra obtained. Trial management, the reference laboratory methods used and obtaining spectra were carried out as described by Burns *et al.* (2013). A PCA analysis was applied to the spectral data, whereby 18 principle components were selected to describe 99.78 % of variation in the spectra. The 'select' algorithm of Shenk and Westerhaus (1991) was applied, whereby a neighbourhood H (NH) statistic of 0.2, 0.4, 0.6 and 0.8 was applied to select calibrations sub-sets from within the full calibration. Predictive regression models and cross-validations were carried out as per Burns *et al.* (2013).

RESULTS AND DISCUSSION The calibration sub-sets selected using the NH criteria of 0.2, 0.4, 0.6 and 0.8 resulted in a of a 33-34 %, 74 %, 85 % and 90 % reduction in sample numbers from the full calibration set for each quality trait (Table 1). For both buffering capacity and WSC content a decreased calibration sub-set size tended to slightly decrease the R^2 of the calibration model, with the smallest calibration set (NH = 0.8) resulting in a 5 % and a 2 % reduction for buffering capacity and WSC, respectively. In contrast, for *in vitro* DMD there was an increase in R^2 as the calibration set size decreased, however the R^2 of the cross-validation model for *in vitro* DMD decreased as the calibration set size decreased (not reported), indicating a less robust model whilst analysing samples outside the calibration set. Shetty *et al*. (2012) concluded that the methodology of sample selection is more important than the number of samples in the calibration set.

Table 1. NIRS calibration statistics for predicting buffering capacity, *in vitro* DMD and water soluble carbohydrate concentration based on representative calibration subsets selected using four neighbourhood H (NH) values.

Buffering capacity (mEq/kg DM)					*in vitro* DMD (g/kg)				WSC (g/kg DM)			
NH	n	SEC	R^2	SECV	n	SEC	R^2	SECV	n	SEC	R^2	SECV
Full	1985	20.3	0.952	20.7	1978	16.0	0.863	16.3	1945	10.8	0.958	11.1
0.2	1393	22.5	0.935	23.26	1387	16.6	0.869	17.0	1380	12.60	0.951	13.2
0.4	549	24.1	0.924	25.96	547	16.5	0.879	18.2	540	12.83	0.955	14.0
0.6	316	24.4	0.922	27.82	321	18.0	0.862	20.6	316	14.74	0.946	17.4
0.8	209	26.5	0.907	32.72	206	16.8	0.890	20.3	206	15.70	0.940	19.1

SEC – Standard error of calibration; SECV – Standard error of cross validation

CONCLUSION The use of PCA to select representative calibration sub-sets resulted in a small decrease in accuracy and robustness of NIRS calibration models for predicting grass buffering capacity, *in vitro* DMD and WSC content. These decreases however must be considered in the context of associated cost and time savings gained through the reduced requirement of reference laboratory analysis.

ACKNOWLEDGEMENTS Funding provided by the National Development Plan, Research Stimulus Fund administrated by DAFM (RSF – 07 526).

REFERENCES BURNS, G. A., GILLILAND, T. J., GROGAN, D., WATSON S. and O'KIELY, P. (2013). Assessment of herbage yield and quality traits of perennial ryegrasses from a national variety evaluation scheme. *The Journal of Agricultural Science*, **151**, 331-346.

SHENK, J. S. & WESTERHAUS, M. O. (1991). Population definition, sample selection, and calibration procedures for near-infrared reflectance spectroscopy. *Crop Science* **31**, 469-474.

SHETTY N., RINNAN, A. and GISLUM, R. (2012) Selection of representative calibration sample sets for near-infrared reflectance spectroscopy to predict nitrogen concentration in grasses. *Chemometrics and Intelligent Laboratory Systems*, **111**, 59-65.