This article is downloaded from

http://researchoutput.csu.edu.au

It is the paper published as:

Author: L. M. Schmidtke, J. P. Smith, M. C. Muller and B. P. Holzapfel
Title: Rapid monitoring of grapevine reserves using ATR-\textsuperscript{\textit{FT}}-IR and chemometrics
Journal: Analytica Chimica Acta ISSN: 0003-2670
Year: 2012 Volume: 732 Pages: 16-25
Abstract: Predictions of grapevine yield and the management of sugar accumulation and secondary metabolite production during berry ripening may be improved by monitoring nitrogen and starch reserves in the perennial parts of the vine. The standard method for determining nitrogen concentration in plant tissue is by combustion analysis, while enzymatic hydrolysis followed by glucose quantification is commonly used for starch. Attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FT-IR) combined with chemometric modelling offers a rapid means for the determination of a range of analytes in powdered or ground samples. ATR-FT-IR offers significant advantages over combustion or enzymatic analysis of samples due to the simplicity of instrument operation, reproducibility and speed of data collection. In the present investigation, 1880 root and wood samples were collected from Shiraz, Semillon and Riesling vineyards in Australia and Germany. Nitrogen and starch concentrations were determined using standard analytical methods, and ATR-FT-IR spectra collected for each sample using a Bruker Alpha instrument. Samples were randomly assigned to either calibration or test data sets representing two thirds and one third of the samples respectively. Signal preprocessing included extended multiplicative scatter correction for water and carbon dioxide vapour, standard normal variate scaling with second derivative and variable selection prior to regression. Excellent predictive models for percent dry weight (DW) of nitrogen (range 0.10–2.65\%DW, median 0.45\%DW) and starch (range 0.25–42.82\%DW, median 7.77\%DW) using partial least squares (PLS) or support vector machine (SVM) analysis for linear and nonlinear regression respectively, were constructed and cross validated with low root mean square errors of prediction (RMSEP). Calibrations employing SVM-regression provided the optimum predictive models for nitrogen (R\textsuperscript{2} = 0.98 and RMSEP = 0.07\%DW) compared to PLS regression (R\textsuperscript{2} = 0.97 and RMSEP = 0.08\%DW). The best predictive models for starch was obtained using PLS regression (R\textsuperscript{2} = 0.95 and RSMEP = 1.43\%DW) compared to SVR (R\textsuperscript{2} = 0.95; RMSEP = 1.56\%DW). The RMSEP for both nitrogen and starch is below the reported seasonal flux for these analytes in Vitis vinifera. Nitrogen and starch concentrations in grapevine tissues can thus be accurately determined using ATR-FT-IR, providing a rapid method for monitoring vine reserve status under commercial grape production.

DOI: http://dx.doi.org/10.1016/j.aca.2011.10.055
URL: http://researchoutput.csu.edu.au/R/-?func=dbin-jump-full&object_id=29728&local_base=GEN01-CSU01
Author Address: lschmidtke@csu.edu.au
CRO Number: 29728
Rapid Monitoring of Grapevine Reserves using ATR-FT-IR and Chemometrics

Leigh M. Schmidtke\textsuperscript{1,2}, Jason P. Smith\textsuperscript{2}, Markus C. Müller\textsuperscript{1,2}, Bruno P. Holzapfel\textsuperscript{2}

\textsuperscript{1}School of Agriculture and Wine Sciences, Charles Sturt University, Wagga Wagga, New South Wales 2678, Australia
\textsuperscript{2}National Wine and Grape Industry Centre, Charles Sturt University Wagga Wagga, New South Wales 2678, Australia

Corresponding Author
Mr Leigh M. Schmidtke
School of Agricultural and Wine Sciences
Charles Sturt University
Locked Bag 588
Wagga Wagga, NSW, 2678
Australia
Tel   +61 (0) 2 69334025
Fax   +61 (0) 2 69332107
Email lschmidtke@csu.edu.au

Keywords:
Fourier transform infrared spectroscopy, vine reserves, nitrogen, starch, \textit{Vitis vinifera}
Abstract

Predictions of grapevine yield and the management of sugar accumulation and secondary metabolite production during berry ripening may be improved by monitoring nitrogen and starch reserves in the perennial parts of the vine. The standard method for determining nitrogen concentration in plant tissue is by combustion analysis, while enzymatic hydrolysis followed by glucose quantification is commonly used for starch. Attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FT-IR) combined with chemometric modelling offers a rapid means for the determination of a range of analytes in powdered or ground samples. ATR-FT-IR offers significant advantages over combustion or enzymatic analysis of samples due to the simplicity of instrument operation, reproducibility and speed of data collection. In the present investigation, 1880 root and wood samples were collected from Shiraz, Semillon and Riesling vineyards in Australia and Germany. Nitrogen and starch concentrations were determined using standard analytical methods, and ATR-FT-IR spectra collected for each sample using a Bruker Alpha instrument. Samples were randomly assigned to either calibration or test data sets representing two thirds and one third of the samples respectively. Signal preprocessing included extended multiplicative scatter correction for water and carbon dioxide vapour, standard normal variate scaling with second derivative and variable selection prior to regression. Excellent predictive models for percent dry weight (DW) of nitrogen (range 0.10-2.65 %DW, median 0.45 %DW) and starch (range 0.25-42.82 %DW, median 7.77 %DW) using partial least squares (PLS) or support vector machine (SVM) analysis for linear and nonlinear regression respectively, were constructed and cross validated with low root mean square errors of prediction (RMSEP). Calibrations employing SVM-regression provided the optimum predictive models for nitrogen ($R^2=0.98$ and RMSEP = 0.07 %DW) compared to PLS regression ($R^2=0.97$ and RMSEP = 0.08 %DW). The best predictive models for starch was obtained using PLS regression ($R^2=0.95$...
and RSMEP = 1.43 %DW) compared to SVR ($R^2=0.95$; RMSEP = 1.56 %DW). The RMSEP for both nitrogen and starch is below the reported seasonal flux for these analytes in *Vitis vinifera*. Nitrogen and starch concentrations in grapevine tissues can thus be accurately determined using ATR-FT-IR, providing a rapid method for monitoring vine reserve status under commercial grape production.

1.0 Introduction

Seasonal patterns of carbohydrates reserve storage and mobilisation in grapevines is strongly influenced by the developmental stage, and the photosynthetic balance between vine growth and maintenance requirements. Developing shoot growth in early spring massively increases carbohydrate demands and mobilisation of storage reserves. As the growing season progresses, reserve dynamics are strongly influenced by factors such as crop load, canopy size, irrigation and climatic conditions that influence photosynthesis and capacity of the vine for carbon acquisition [1]. Consequently, the over-wintering carbohydrate reserve status of vines may vary from season to season according to both climatic and vineyard management factors. Nitrogen reserves are also stored in the perennial tissues of grapevines and play an important role in supporting early season canopy growth [2]. The nitrogen content of perennial tissues is responsive to nitrogen fertilizer applications, and like carbohydrates, concentrations at dormancy and the changes in concentrations between development stages may also vary between seasons [3]. Studies linking carbohydrate and nitrogen reserves in grapevines with vegetative growth and fruiting responses suggest that monitoring stored reserves at dormancy could provide an advanced indication of yield and growth potential in the following season [4-6]. Recently we have also observed that seasonal differences in the composition of Shiraz grapes were associated with a contrasting pattern of carbohydrate reserve storage and mobilization between flowering and harvest [1]. This suggests that monitoring changes in reserve concentrations between key development stages may also
assist with understanding the impact of management practice and climatic factors on berry development and fruit quality.

Regular monitoring of the carbohydrate and nitrogen reserve status of grapevines will be of commercial interest; however some obstacles must first be overcome. Characterization of existing variation in reserve concentrations such that context is provided for the interpretation of analysis undertaken is the first challenge. Current methods of analysis based upon single analyte determination of nitrogen is by combustion, and for starch requires enzymatic hydrolysis with the resulting measurement of glucose, and these methods may be too slow for basing management decisions, or too laborious and expensive for widespread commercial adoption. Hence there is a requirement for a more rapid and lower cost method for measuring carbohydrate and nitrogen concentrations in grapevine tissues.

Attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FT-IR) combined with chemometric modelling offers a rapid means for the determination of a range of analytes in powdered or ground samples. This process involves collection of the infrared spectra of samples and correlating the absorbance at specific wavelengths to analyte concentration with predictive models constructed that can be used to determine the concentration of analytes in other samples. Typically wavelengths in the near infrared (750-2500 nm or 13 333-4000 cm\(^{-1}\)) or mid infrared (2500-25 000 nm or 4000-400 cm\(^{-1}\)) spectrum are used [7]. ATR-FT-IR offers significant advantages over traditional analysis methods due to the simplicity of instrument operation, reproducibility and speed of data collection. ATR-FT-IR has been applied to the detection of adulterants in a variety of edible foods [8], quality control in food manufacture [9], monitoring oil degradation arising from thermal exposure [10], on-line monitoring of nutrient concentration and biomass during antibiotic production [11], the rapid determination of organic acids and carbohydrates in fruits [12] and prediction of olive oil sensory qualities [13]. Within the wine industry FT-IR is used extensively for
determination of ethanol, pH, volatile acidity, glucose and fructose [14], and ATR sampling methods are more commonly being used, with recent investigations employing this technique for discrimination of wine variety [15] and for real time juice and wine analysis [16, 17].

Predictive models for the estimation of analyte concentrations in samples are constructed from the spectra using regression analysis and typically partial least squares (PLS) algorithms are used for this purpose. Data are first pre-processed in an attempt to remove interference and to linearise signal response to analyte concentration. Mathematical models that relate signal intensity at multiple wavelengths with a measured, or known, analyte concentration are then constructed so that the concentration of that analyte can be accurately predicted in ‘unknown’ samples. Advantages of multivariate calibrations using PLS regression are efficiency in data collection, sample handling and expedition of results, hence the popularity of applying mathematical models to chemical problems. The underlying assumption of a linear dependency of signal response to analyte concentration does not always hold, particularly in complex samples containing interfering substances and thus non-linear predictive models may offer solutions in which PLS regressions model underperform [18]. Support vector regression (SVR), with its origins in machine learning and statistical learning theory [19], are becoming increasingly popular chemical models to handle non-linear data. The basic concept of support vector regression is to map the calibration data to a high dimension space, termed hyperplanes, using a kernel function from which a linear model can be constructed. Deviations of predicted quantity (or class) from actual are allowed within an error margin (denoted ε) whilst maintaining a constraint that minimises individual sample weighting but maximises sample distance for each hyperplane. Such constraints are not always possible to achieve and a penalty error or cost function (denoted C) is fitted to the model to explicitly deal with samples which cannot be thus fitted. The deviation above the error margin is termed the slack variable in soft margin SVR, and the cost function essentially
determines the compromise between data fit and the size of the slack variables [20], and is referred to as the $\varepsilon$-insensitive loss function [19]. An analogy is a tube with a diameter equal to the acceptable error margin fitted to the data and samples that lie on or beyond the boundary are support vectors. Thus for small values of $\varepsilon$ larger numbers of support vectors will be used for modelling purposes with lower prediction errors in the model but this increases the risk of over fitting the data [18]. Conversely a lower number of support vectors increases the predictive error of the model. Large values of $C$ effectively force samples with slack margins back within the error margin of the fitted model and thus decreases the number of support vectors.

The ability of support vector regression to accommodate non-linear data arises from the kernel based manner in which the data is initially mapped into a high dimension subspace. Radial basis functions (RBF) are commonly employed for mapping data to the high dimension space [18], although numerous alternatives such as linear, polynomial, sigmoidal, splines, multi-layer perceptrons, wavelets and tensor product functions have been described [19, 21-23]. In the application of RBF transformations the diameter of the radius must also be carefully chosen. If the radial diameter, (denoted $\gamma$) is too small an overly complex solution is determined with greater numbers of support vectors that will lead to data overfit, whereas a flatter radius will lead to larger predictive error. Often the RBF tuning parameter is described in terms of the standard deviation of the overall data (denoted $\sigma$) rather than $\gamma$ with an inverse squared relationship existing between them; hence low $\sigma$ will infer a larger $\gamma$ and vice versa [18]. Again, a trade-off between complexity and optimum fit is required and in practise the best approach for determining the optimum RBF parameter ($\gamma$ or $\sigma$), acceptable error margin ($\varepsilon$) and the cost penalty ($C$) is a grid search with cross validation to find minima of predictive errors [22]. One must be aware however, that a workable solution may not exist
for all data sets in spite of computational programs providing values for each of these parameters [18].

The purpose of the present investigation was to investigate the feasibility of ATR-FT-IR spectroscopy for the rapid measurement of important vine reserve components. The sample set was designed to be representative of the variability within the range of analytical attributes typically found in vine tissue samples across a number of vineyard sites and *Vitis vinifera* varieties.

## 2 Experimental

### 2.1 Materials

All chemicals were of analytical grade and purchased from BDH Australia. Deionized water (18 MΩ cm$^{-1}$) was prepared using a MilliQ filtration system. Thermostable α-amylase, amyloglucosidase and the glucose oxidase/peroxidase based assay were purchased from Megazyme International.

### 2.2 Instrumentation

Infrared spectra of dried and powdered vine samples were collected using two Alpha Fourier transform spectrophotometers (Bruker) configured for attenuated total reflectance (ATR) at ambient temperature. Each sample was measured once where the ATR spectra were averaged from 64 scans over the region 374-7496 cm$^{-1}$ at 1.4 cm$^{-1}$ resolution, with a background measurement against air conducted every 10 samples. Spectra were acquired using OPUS software version 6.5 provided by Bruker. This approach enabled approximately 30 samples to be measured every hour. Data was converted to ascii format and imported into Matlab version 7.4.0.287 R2007a (The Mathworks, Natick). For the purposes of exploratory data and regression analysis ATR data between 374 to 3958 cm$^{-1}$ was extracted from each
sample measurement and converted to absorbance to correct for differences in sample penetration of the beam at different wavelengths. Five independent spectra of deionised water and powdered starch were also collected and averaged as reference for use in extended multiplicative scatter correction.

2.3 Procedure

2.3.1 Vine samples

Wood and root tissue samples for modelling purposes were selected from an archived collection of several thousand potential candidate samples collected from viticultural research trials or vineyard surveys conducted between 2006 and 2009 (Table 1). The purposes of the trial sites was to investigate the implications of nitrogen and water supply on the reserve dynamic of the perennial structure and on the composition of the grapes, while the survey sites were aimed to gain a better understanding of the intra-seasonal changes of winter vine reserves in relation to yield variation and prediction. A detailed description of these studies is beyond the scope of the present paper, and background information has been presented in an earlier review [1]. Samples were chosen to represent a range of analytical values, growing years, conditions and grape varieties experienced during the conduct of these projects. A total of 1880 grapevine (Vitis Vinifera L.) tissue samples were collected from thirty-five vineyards over a period of four years consisting of three field trials and thirty-two survey sites (Tables 1 and 2), these were located within the Riverina and south eastern slopes of New South Wales (Australia) and in the Rheingau wine region in Hessen (Germany).

Thirty-two Shiraz and Chardonnay blocks in commercially operated vineyards were selected across the Riverina and Gundagai, Tumbarumba and Hilltops regions. Within each survey block, five panels of four vines were selected after leaf fall in 2005. The position of each replicated panel was not chosen based on any pre-existing spatial information, but where
possible, one panel of vines was located near the centre of the block, and the other four
approximately within each quarter of the block. Prior to pruning in 2005, 2006 and 2007
root, wood and cane samples were collected from each panel of four vines. Root samples
(diameter range ~3-7 mm) were manually dug from near the base of each vine. Wood
samples were collected with a 4.8 mm drill bit to a depth visually estimated as the centre of
the cordon and trunk. Two 4-node canes were collected from each vine, of which the basal
two nodes (subsequently referred to as spurs) of four canes were used for carbohydrate and
nitrogen analysis.

The Semillon and Riesling samples were collected from two commercially operated
vineyards in the Riverina at five key phenological stages (budbreak, flowering, veraison,
harvest and dormancy) commencing in 2007 and ending in July 2009. Each vineyard trial
consisted of 24 plots with five vines being sampled at each stage. Root samples (diameter
range ~3-7 mm) were manually dug from near the base of each vine, while wood samples
were collected with a 4.8 mm drill bit to a depth visually estimated as the centre of the cordon
and trunk. From the third experimental vineyard located in the Rheingau only wood samples
were collected at four key phenological stages (budbreak, flowering, veraison and harvest) in
the growing season 2008 and 2009. This vineyard was subdivided in 12 plots with four vines
being sampled per date, the samples were collected by randomly cutting off one spur
including the cane of the cordon. The collected spur was considered as two year old wood
and the cane sample as one year old wood and was separated in spur and cane sections prior
to further sample preparation.

2.3.2 Nitrogen and starch analysis

Roots were washed in phosphate free detergent and rinsed with deionized water, both
wood and root tissues were either oven dried at 70°C (survey sites, Riesling Rhinegau) or
freeze dried (Riverina trials). Following the drying processes the samples were ground to
0.12 mm with a heavy duty cutting mill (Retsch ZM2000, Haan, Germany) and an ultra-
centrifugal mill (Retsch ZM200); the samples from the Rhinegau trial were ground to 0.2 mm
with a cutting mill (Retsch SM200). Non-structural carbohydrate concentrations were
determined on a 20 mg sub-sample of each tissue as outlined below. For total nitrogen, an
equal proportion of tissue from the five replicate locations in each vineyard was bulked, and a
nitrogen content of a 50 mg sub-sample determined with a VarioMAX combustion analyser
(Elementar, Hanau, Germany).

For starch determination, free soluble carbohydrates first removed from a 20 mg
sample with three 1 ml aliquots of 80% v v\(^{-1}\) ethanol. The first two aliquots were incubated in
a water bath at 80°C for 10 minutes, and the third at room temperature for 10 minutes. Starch
from the remaining wood sample was then solubilized in 200 μL of dimethylsulfoxide at
98°C for 10 minutes. Thermostable α-amylase (29 U mL\(^{-1}\)) in 300 μL of MOPS buffer (pH
7) was then added; the samples incubated for 15 minutes, and then allowed to cool.
Amyloglucosidase (32 U mL\(^{-1}\)) in 400 μL of sodium acetate buffer (pH 4.5) was added, the
samples incubated at 50°C for 60 minutes and then centrifuged at 10,000 rpm for 2 minutes.
The supernatant was diluted 1:6 (wood) or 1:11 (roots), and the glucose concentration
determined using a glucose oxidase/peroxidase based assay. Samples were tested with either
a Konelab Arena 20 XT or in microplate format using a Biotek μQuant microplate reader
after incubation at 40°C for 30 minutes. Glucose standard curves in the range of 0-5000 mg
L\(^{-1}\) were prepared for each run of samples and starch concentration in the original 20 mg
sample was then calculated as a percentage of dry weight.

2.3.3 Exploratory data analysis
Spectra from all samples, and the 95% confidence interval for absorbance, was plotted (Fig. 1) along with spectra obtained for water, carbon dioxide and starch (Fig. 1S) to identify regions with interesting variation. A principal component analysis (PCA) of the standard normal variate (SNV) scaled and mean centred (MC) spectral data between the regions 875-1800 cm\(^{-1}\) was conducted using the singular value decomposition (SVD) algorithm with principal components (PC) 1-2 and associated loadings plotted (Fig. 2) to identify specific groups of samples.

### 2.3.4 Selection of samples for calibration and test sets

Two approaches to constructing calibration and test data sets were conducted; firstly random assignment of samples to a calibration or test set based so that approximately two thirds of samples were designated calibration and one third as an independent test set. The second approach involved interrogation of the MIR spectra to identify outlier samples. PCA was conducted as described above and the chosen number of PC was determined from the relative decrease in eigenvalues and the variance captured for each PC. Sample outlier and influence plots were constructed (Fig. 3) to identify samples with Q residuals or Hotelling \(T^2\) values exceeding the 95% confidence interval. Samples with high Q residuals are poorly modelled with the chosen number of principal components [24], whereas samples with high Hotelling \(T^2\) have specific attributes that deviate further from mean observations and may exert a disproportionate leverage upon the models [25, 26]. For the purposes of further modelling, samples with high Hotelling \(T^2\) and/or Q residuals were excluded. Calibration samples were chosen from the remaining sample matrix based upon the D-optimal criteria [27] such that two thirds of samples were assigned to the calibration set. D-optimal computation was conducted on the SNV treated spectra using PLS toolbox (Eigenvector, Wenatchee). Samples not selected for calibration were assigned to the test data set.
Calibration and test set sample details are shown in Tables 1 and 2 for random and D-optimal criteria allocation respectively, and data set attributes are shown in Table 3.

2.3.5 Data pre-processing and regression analysis

Calibration and test data were pre-processed using SNV scaling followed by calculating the second derivative using a Savitzky–Golay [28] filter prior to variable selection. Alternatively, multiplicative scatter correction (MSC) or extended multiplicative scatter correction (EMSC) using reference spectra for water and carbon dioxide vapour were applied to the calibration data set using a second order polynomial filter in accordance with described procedures [29-31] prior to variable selection. Spectra for carbon dioxide were obtained from an online database [32] at a resolution of 0.1 cm$^{-1}$, were de-resolved and re-sampled to match the reflectance spectra of samples. For the purposes of applying MSC and EMSC, the mean spectra of the calibration samples were used as the reference spectra.

Variable selection within the calibration block was conducted using the forward interval PLS (iPLS) regression for each analyte and model combination [33].

2.3.6 Regression modelling of data

Predictive models of the spectra for nitrogen and starch were constructed using either PLS regression or support vector regression analysis. For PLS regression the pre-treated data were modelled using the SIMPLS algorithm with cross validation using random sample subsets with ten data splits of the calibration block. The number of latent variables chosen for each model was determined from eigenvalues and minima of the root mean square errors of calibration, cross validation and prediction. Support vector regression using a radial basis kernel was conducted using the LIBSVM algorithm with a grid search for optimising cost function (C), RBF parameter ($\gamma$) and error margin ($\varepsilon$) with RMSECV determined with a five-fold data split for cross validation. Feature selection, Hotelling $T^2$, Q residuals, PLS and
support vector regressions were conducted in PLS Toolbox (version 6.2, Eigenvector Research Inc, Wenatchee, Washington). Predictive model characteristics for nitrogen and starch are presented in Tables 4 - 5 and figs. 4 - 5 respectively.

3.0 Results and Discussion

3.1 Exploratory Data Analysis

The ATR-FT-IR spectra of the dried and powdered vine tissue samples (Fig. 1) is characterised by dominant peaks between 375-800, 875-1800 and a broad peak between 3000-3500 cm\(^{-1}\). A broad absorbance band between 3000-3500 cm\(^{-1}\) characteristic of O-H bond stretching dominates the spectra and could be related to the presence of carbohydrates or residual moisture [34] (Fig. 1-supp). The spectral region with the most interesting variation occurs between 875-1800 cm\(^{-1}\) (Fig. 1) and PCA of the SNV and MC treated spectra from this region clearly separates samples according to grapevine tissue type (root and wood; Fig. 2) with the first two PC modelling approximately 84% of the data variability.

Distinction of samples based upon grapevine tissue is expected to occur as the tissue vary significantly in their reserve composition [1, 35]. Sample groupings according to location are also evident with samples from the Rheingau grouped tightly together in the centre of the plot of PC1 and PC2. Samples from other locations appear well dispersed, as do growing years and grape variety. The spectral bands characterising carbohydrates are typically observed in the 3600-2800 cm\(^{-1}\), 1500-1200 cm\(^{-1}\), 1200-950 cm\(^{-1}\) and 950-700 cm\(^{-1}\) regions corresponding to O-H and C-H bond stretching; bending of symmetric HCH bonds; C-O and C-C bond stretching and COH, CCH and OCH bond bending [34]. Loading plots for PC1 and PC2 show the spectral variations responsible for this discrimination, with regions between 1650-1550 cm\(^{-1}\) highly positively loaded onto PC1, with a smaller contribution between 1500 to
1300 cm\(^{-1}\). Regions between 1150 to 1050 cm\(^{-1}\) are negatively correlated with PC2 and the region between 1000-875 cm\(^{-1}\), are positively correlated.

Examination of the score plots also reveals the spread of samples allocated to the calibration and test data sets by random or the D-optimal criteria (Fig. 2). Test and calibration samples appear to be well distributed along the first 2 PCs for both approaches to calibration and test sample set allocation. For the purposes of regression modelling samples from the entire data set were retained rather than separated into their distinct groups (i.e. *Vitis* species, locations or tissue type) with calibration and test sets allocated as described above.

An important consideration when using SVR is to ensure that data over fitting does not arise as there are no specific rules for fitting the data unlike PLS [18]. Comparison of the number of support vectors used with the total number of samples in the calibration models show that between 43 and 45% of samples have been classified as support vectors for the nitrogen models and 54% of samples used for starch models. This compared favourably with the previous reports in which around 85% of samples have been used in least squares SVR [36]. Various spectral processing was conducted for each model (data not shown) to determine the optimum pre-processing steps for constructing predictive models. Spectral correction using SNV attempts to normalise signal intensity between samples without least squares fitting of a reference spectrum but can be prone to signal noise [37]. MSC removes non-linearities in spectral data caused by light scatter by particles in the sample, thus correcting for variations in particle size distribution between samples [29] and EMSC expands MSC so that wavelength correction for known spectra of interferents or desired components are effectively filtered with a baseline correction [37]. EMSC thus enables spectral correction of water vapour and atmospheric interference from carbon dioxide [31]. In the present investigation signal pre-processing using MSC produced better predictive
models (lowest RMSEP) for nitrogen and starch when samples were randomly allocated to
calibration and test data sets (Tables 4-5). SNV treated spectra produced the optimum
predictive models for both nitrogen and starch using SVR once spectral outliers were
removed and samples allocated to the calibration and test sets using the D-optimal criteria.
EMSC proved to be most effective for the prediction of nitrogen using PLS regression of the
D-optimal allocated data sets. These results illustrate the requirement to consider carefully
data pre-processing to optimise predictive model performance.

3.2 Prediction of nitrogen

The total nitrogen content of the samples ranged from 0.10 to 2.65 % dry weight and
high correlations for predicted and measured percent dry weight nitrogen concentration of the
vine samples was obtained for both PLS and SV regression models (Fig. 4, Table 4). PLS
and SV regression models exhibit similar co-efficient of determination ($R^2 = 0.97$) for
randomly allocated samples and RMSEP = 0.09 and 0.08 respectively. SV regression was
superior in predictive performance for nitrogen when samples were allocated according to the
D-optimal criteria with $R^2 = 0.98$ and RMSEP = 0.07 (Table 4) and was marginally better
then PLS regression of the D-optimally allocated samples ($R^2 = 0.97$; RMSEP =0.08). The
RMSEP for the models compares favourably with the analytical uncertainty of the reference
method which has been determined as 0.05% DW (unpublished data). Plots of the residuals
for nitrogen (Fig. 2S) are normally distributed for the SVR models demonstrating that
nitrogen percent dry weight of the dried vine tissue samples could be accurately predicted for
all samples using ATR-FT-IR, and no deterioration in predictive performance of the models
occurs at varying nitrogen concentrations. The RMSEP obtained for the determination of
total nitrogen content in this investigation compares very favourably with reported predictive
errors for the estimation of nitrogen in grasses using similar sample preparation and
analytical approaches [38]. The RMSEP obtained for regression models in this project is also
well below the seasonal change in nitrogen concentration measured for vine samples [35]
thus making these models suitable for the determination of inter and intra season flux of total
nitrogen.

3.3 Prediction of starch

Excellent predictive models for starch were also obtained using both PLS and SV
regression (Fig. 5, Table 5). Regression analysis with randomly allocated samples produced
models with co-efficient of determination of 0.94 and 0.96 for PLS and SV regression
respectively and RMSEP of 1.73 and 1.60 respectively. Interestingly the lowest predictive
error was obtained using PLS regression (RMSEP = 1.43) compared to SV regression
(RMSEP = 1.56) in the D-optimal allocated samples, with equal co-efficient of determination
of 0.95. These values compare most favourably with the analytical uncertainty associated
with the calibrant was has been determined to be 1.62% DW (unpublished data). An
interesting feature of the SV regressions is the lack of negative predicted values at low
measured starch concentrations, whereas the PLS regression analysis does not predict low
starch quantities as well. Plots of starch residuals (Fig. 3S) do not display a normal
distribution with negative skew apparent at higher starch concentrations. Residual skew is
particularly apparent in the PLS models and the reduced data fit at higher starch
concentrations may have arisen from inefficiencies with enzymatic digestion of ground vine
tissue for starch extraction and subsequent measurement [35]. Improved data fit at low starch
concentrations and decreased residual skew at higher concentrations suggests SV regression
models may be a better predictive tool for the estimation of starch in the samples rather than
the PLS regressions. The observed magnitude of starch flux within growing seasons is
reported to be between approximately 2 and 4% dry weight thus RMSEP below this quantity
indicates that SV and PLS models provide a suitable means for the rapid determination and
monitoring of vine starch reserves.
It is recognised that the RMSEP of predictive models for analytes is generally an overestimation of the actual error associated with the predictand as the calibrants have associated systematic analytical error [39-42]. Random errors within the calibrants are well modelled using PLS regression with a tendency to cancel out and thus can be considered to be inconsequential to the prediction error [43]. In the absence of a quantified analytical error for the calibrants, a sample specific error of prediction could be determined from the sample leverage in the predictive model which has been shown to be lower than the RMSEP [44]. Such error estimates may thus become important for samples with low intra-seasonal analyte variation. Overall the predictive models from this investigation are useful for monitoring the seasonal flux of nitrogen and starch reserves in grapevine tissue.

4.0 Conclusion

The results from this investigation show that starch and nitrogen reserves in grapevine tissue can be measured with good precision and accuracy using ATR-FT-IR and PLS or SV regression. SV regression models were marginally better than PLS regression models and predictive errors were below the reported seasonal flux for total nitrogen and starch for vine samples. In the present investigation, the RMSEP is well below the reported flux of storage reserve nitrogen and starch in *Vitis vinifera*. The predictive models could be applied to vine samples from different regions and varieties thus providing a useful analytical tool for monitoring vine carbohydrate and nitrogen reserve status. Total analytical time for ATR-FT-IR is around 60 seconds per sample once dried and ground, compared to several days to extract and quantify by enzymatic analysis. Sample preparation is relatively simple and this technique will make it possible for vineyard managers and personnel to rapidly monitor winter and intra-seasonal changes in vine reserve starch and nitrogen and provide information about potential vine growth and development in the following season.
5.0 Acknowledgements

The loan of ATR-FT-IR instruments from Bruker Optics, Australia is gratefully acknowledged. Vine reserve trials were funded by the Australian wine and grape industry through the Grape and Wine Industry Development Corporation, Australia, while the Riesling trial in Germany received funding from the Geisenheim Research Centre. Considered discussion of support vector regression by Donal O’Sullivan is gratefully acknowledged.

6.0 References


Figure 1. ATR-FTIR mean spectra and 95 percent confidence interval of all samples prior to outlier detection.
Figure 2. Identification of sample groupings by PCA of ATR-FT-IR between 875-1850 cm\(^{-1}\) and associated loadings of the first 2 principal components.
Figure 3. Identification of sample outliers and influence using Hotelling $T^2$ and Q residuals.

Figure 4. Regression plots for prediction of nitrogen by ATR-FT-IR.
Figure 5. Regression plots for prediction of starch by ATR-FT-IR.

Supplementary Figure Captions
Figure 1-supp. ATR-FTIR spectra of filters for extended multiplicative scatter correction.
Figure 2-supp  Plot of residuals for predictive models for Nitrogen.
Figure 3-supp Plot of residuals for predictive models for Starch.
Table 1. Sample details (year, location and vine tissue) in calibration and test sets with random allocation.

<table>
<thead>
<tr>
<th>Year</th>
<th>Calibration Set (n=1239)</th>
<th>Test Set (n=641)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Riesling</td>
<td>Semillon</td>
</tr>
<tr>
<td>2006</td>
<td>148</td>
<td>148</td>
</tr>
<tr>
<td>2007</td>
<td>47</td>
<td>98</td>
</tr>
<tr>
<td>2008</td>
<td>203</td>
<td>158</td>
</tr>
<tr>
<td>2009</td>
<td>163</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>Gundagai</td>
<td>74</td>
<td>74</td>
</tr>
<tr>
<td>Hilltops</td>
<td>302</td>
<td>345</td>
</tr>
<tr>
<td>Riverina</td>
<td>39</td>
<td>39</td>
</tr>
<tr>
<td>Tumbarumba</td>
<td>111</td>
<td>111</td>
</tr>
<tr>
<td>Rheingau</td>
<td>270</td>
<td>178</td>
</tr>
<tr>
<td>Wood</td>
<td>143</td>
<td>167</td>
</tr>
<tr>
<td>Root</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Sample details (year, location and vine tissue) in calibration and test sets following outlier removal and D-optimal Criteria selection

<table>
<thead>
<tr>
<th></th>
<th>Calibration Set (n=1114)</th>
<th></th>
<th>Test Set (n=575)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Riesling</td>
<td>Semillon</td>
<td>Shiraz</td>
<td>Total</td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>147</td>
<td>147</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>43</td>
<td>83</td>
<td>239</td>
<td>365</td>
</tr>
<tr>
<td>2008</td>
<td>190</td>
<td>111</td>
<td>121</td>
<td>422</td>
</tr>
<tr>
<td>2009</td>
<td>100</td>
<td>80</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gundagai</td>
<td>59</td>
<td>59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hilltops</td>
<td>81</td>
<td>81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Riverina</td>
<td>254</td>
<td>274</td>
<td>331</td>
<td>859</td>
</tr>
<tr>
<td>Tumbarumba</td>
<td>36</td>
<td>36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rheingau</td>
<td>79</td>
<td>79</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Vine Tissue</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wood</td>
<td>180</td>
<td>120</td>
<td>328</td>
<td>628</td>
</tr>
<tr>
<td>Root</td>
<td>153</td>
<td>154</td>
<td>179</td>
<td>486</td>
</tr>
</tbody>
</table>
Table 3. Calibration and test set characteristics for regression analysis

<table>
<thead>
<tr>
<th></th>
<th>Calibration Set</th>
<th>Test Set</th>
<th>Calibration Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nitrogen (%DW)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.65</td>
<td>0.68</td>
<td>0.70</td>
<td>0.51</td>
</tr>
<tr>
<td>Median</td>
<td>0.46</td>
<td>0.57</td>
<td>0.53</td>
<td>0.29</td>
</tr>
<tr>
<td>Range</td>
<td>0.11-2.65</td>
<td>0.10-2.31</td>
<td>0.10-2.39</td>
<td>0.11-1.81</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.50</td>
<td>0.50</td>
<td>0.52</td>
<td>0.41</td>
</tr>
<tr>
<td><strong>Starch (%DW)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>10.24</td>
<td>11.08</td>
<td>11.05</td>
<td>8.38</td>
</tr>
<tr>
<td>Median</td>
<td>7.49</td>
<td>7.77</td>
<td>8.01</td>
<td>7.01</td>
</tr>
<tr>
<td>Range</td>
<td>0.25-47.85</td>
<td>0.30-42.82</td>
<td>0.34-47.85</td>
<td>1.26-28.78</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>7.55</td>
<td>8.04</td>
<td>8.06</td>
<td>4.18</td>
</tr>
<tr>
<td>Number of samples</td>
<td>1239</td>
<td>641</td>
<td>1114</td>
<td>575</td>
</tr>
</tbody>
</table>
### Table 4  Regression Performance of Predictive Models for Nitrogen

<table>
<thead>
<tr>
<th>Calibration and Test Set Sample Allocation</th>
<th>Regression Performance</th>
<th>PLS</th>
<th>SVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Sample Allocation</td>
<td>Pre-processing(^\d)</td>
<td>SNV/(^2)Der/MC/iPLS</td>
<td>MSC/(^2)Der/MC/iPLS</td>
</tr>
<tr>
<td></td>
<td>(R^2)</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>RMSEP</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Bias</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>No. LV/SV(^\d)</td>
<td>6</td>
<td>6/562</td>
</tr>
<tr>
<td></td>
<td>Hyperparameter values</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cost function (C)</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Error margin ((\varepsilon))</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RBF parameter ((\gamma))</td>
<td>0.032</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outlier Removal D-Optimal Allocation</th>
<th>Pre-processing</th>
<th>EMSC, (^2)Der, MC, iPLS</th>
<th>SNV/(^2)Der/MC/iPLS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(R^2)</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>RMSEP</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Bias</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>No. LV/SV</td>
<td>6</td>
<td>6/472</td>
</tr>
<tr>
<td></td>
<td>Hyperparameter values</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cost function (C)</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Error margin ((\varepsilon))</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RBF parameter ((\gamma))</td>
<td>0.032</td>
<td></td>
</tr>
</tbody>
</table>

\(^\d\)SNV: standard normal variate scaling  
MSC: multiplicative scatter correction  
EMSC: extended multiplicative scatter correction  
MC: mean centre  
iPLS: interval partial least squares variable selection  
\(^2\)nd Der: Savitzky-Golay \(^2\)nd derivative  
\(^\d\)Calculated as the mean \(R^2\) for calibration, cross validation and independent test sets  
\(^\d\)Number of latent variables used for data compression/number of support vectors used for modelling
Table 5  Regression Performance of Predictive Models for Starch

<table>
<thead>
<tr>
<th>Calibration and Test Set Sample Allocation</th>
<th>Regression Performance</th>
<th>PLS</th>
<th>SVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Sample Allocation</td>
<td>Pre-processing †</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R²</td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>RMSEP</td>
<td>1.73</td>
<td>1.60</td>
</tr>
<tr>
<td></td>
<td>Bias</td>
<td>-0.14</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td>No. LV/SV§</td>
<td>8</td>
<td>6/668</td>
</tr>
<tr>
<td></td>
<td>Hyperparameter values</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cost function (C)</td>
<td>31.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Error margin (ε)</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RBF parameter (γ)</td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>Outlier Removal D-Optimal Allocation</td>
<td>Pre-processing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SNV/2nd Der/MC/iPLS</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>RMSEP</td>
<td>1.43</td>
<td>1.56</td>
</tr>
<tr>
<td></td>
<td>Bias</td>
<td>0.20</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>No. LV/SV</td>
<td>6</td>
<td>6/595</td>
</tr>
<tr>
<td></td>
<td>Hyperparameter values</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cost function (C)</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Error margin (ε)</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RBF parameter (γ)</td>
<td>0.032</td>
<td></td>
</tr>
</tbody>
</table>

† SNV: standard normal variate scaling
MSC: multiplicative scatter correction
EMSC: extended multiplicative scatter correction
MC: mean centre
iPLS: interval partial least squares variable selection
2nd Der: Savitzky-Golay 2nd derivative

§ Calculated as the mean R² for calibration, cross validation and independent test sets

§ Number of latent variables used for data compression/number of support vectors used for modelling