Statistical data mining algorithms for optimising analysis of spectroscopic data from on-line NIR mill systems

Thesis submitted by

Justin David Sexton

BSc Mathematics

MSc Mathematics (Research)

In fulfilment of the requirements for the degree of
PhD Mathematics (Research) in the College of Science
and Engineering, James Cook University, Townsville

Thesis submitted:

March 2020

Acknowledgements

I would like to acknowledge and thank my thesis supervisors Dr. Yvette Everingham, Mr. Steve Staunton, Dr. David Donald and Dr. Ronald White, for their support, encouragement and contribution to this thesis. Without their expertise, this work could not have been accomplished.

I would also like to express my appreciation and thanks to Henrique Boriolo Dias, Vinicius Santino Alves and Sosena Mesfin, fellow students whose company and support were invaluable. Finally I would like to thank my parents and family for all their support over the (at times, seemingly endless) duration of the project.

Statement of access

I, the undersigned, author of this work, understand that James Cook University will make this thesis available for use within the University Library and, via the Australian Digital Thesis network, for use elsewhere.

I understand that, as an unpublished work, a thesis has significant protection under the Copyright Act and;

I do not wish to place any further restrictions on access to this work.

Name: Justin David Sexton Signature: Content has been removed for privacy reasons Date: 20/3/2020

Statement of sources declaration

I, declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name: Justin David Sexton Signature: Content has been removed for privacy reasons Date: 20/3/2020

Statement on the contribution of others

Dr. Yvette Everingham, Senior Lecturer, College of Science and Engineering, James Cook University and Dr. David Donald, MicroAg, Cairns supervised this thesis helping to develop the original proposal as well as providing editorial assistance throughout the thesis. Dr. Yvette Everingham provided mathematical and statistical expertise and support while Dr. David Donald provided expertise and support in NIR modelling techniques. As part of the supervisory team, Mr. Steve Staunton, Sugar Research Australia, Gordonvale and Dr. Ronald White, Associate Professor, College of Science and Engineering, James Cook University provided supervision and editorial assistance throughout the thesis. Mr. Steve Staunton acted as a primary supervisor and liaison to Sugar Research Australia during the second half of the PhD project. As such, Mr. Staunton contributed to the further development of the research questions, particularly with development of the final research chapter (Chapter 7).

Contributions to specific chapters were:

- Dr. Everingham, Dr. Donald and Mr. Staunton supplied editorial assistance for the first draft of Chapter 1 and Chapter 2. Editorial comments of the first draft helped to define the overall scope of Chapter 1.
- The research question of Chapter 3 and Chapter 4 was developed with input from Dr. Everingham, Dr. Donald and Mr. Staunton. Editorial assistance was supplied by Dr. Everingham, Dr. Donald, Mr. Staunton and Dr. White.
- Dr. Everingham, Dr. Donald and Mr. Staunton helped develop the research question of Chapter 5. Editorial assistance was supplied by Dr. Everingham, Dr. Donald, Mr. Staunton and Dr. White. Mr. Staunton provided the range of NIR spectral values used in the analysis. As the NIR spectral values was provided by Mr. Staunton on behalf of SRA, the values used were not published. For consistency, with the published version the values are also absent from this manuscript.
- Dr. Everingham, Dr. Donald and Mr. Staunton helped refine the research questions of Chapter 6 and 7. Dr. Everingham suggested the use of Wavelets as a spectral pre-processing step in Chapter 6. Dr. Everingham, Dr. Donald, Mr. Staunton and Dr. White supplied editorial assistance.

The data used throughout the thesis was supplied by Sugar Research Australia with Mr. Staunton acting as liaison, to ensure no breach of etiquette or confidentiality with respect to publications.

Sugar Research Australia Limited (ABN 16 163 670 068) provided a stipend of $32,000 per year as well as an operating budget of $10,000 per year for the two year duration of the thesis. The project was funded as SRA project 2014/019. James Cook University provided $1,500 per year for project support as part of the university's "Minimum Resources Policy" for higher degree research students.

Every reasonable effort has been made to gain permission and acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

Abstract

Despite the economic importance of producing high quality sugar, thousands of tonnes of sugarcane with atypically low quality can pass undocumented through Australian sugarcane mills each season. This 'atypical' cane can represent deteriorated or contaminated sugarcane that affect mill processes and are misrepresented by current rapid assessment techniques such as Near Infra Red (NIR) spectroscopy. Powerful datamining techniques such as support vector machines (SVM) and artificial neural networks (ANN) have often been used to improve NIR rapid assessment tasks due to their ability to model complex relationships. Unfortunately, there has been little research into the use of these techniques to identify atypical cane samples or how estimates of cane quality may be affected by atypical cane samples. The objective of this thesis was to develop and compare statistical data mining methodologies to accurately measure cane attributes for anomalous cases from NIR spectra contained within large NIR databases.

A range of complex and powerful modelling techniques including SVM, ANN and tree-based approaches were compared to simpler techniques that are more commonly used to estimate cane quality parameters such as partial least squares (PLS). Comparisons were used to identify the most effective techniques for estimating cane quality parameters such as Commercial Cane Sugar (CCS) as well as for discriminating between atypical and typical cane samples. A novel methodological framework was then developed to use predicted class probability to apply specific quality estimation models for atypical and typical cane samples. This was achieved by first tuning the probability at which a sample was identified as atypical or typical and then apply an appropriate model to estimate CCS. The ability of this framework to estimate CCS was compared to a baseline PLS model.

There were three important outcomes of this research:

1. The fast and simple PLS performed as well as complex algorithms such as SVM for the estimation of cane quality parameters.
2. Using appropriate data pre-processing and feature selection techniques PLS discriminant analysis was able to classify samples as typical or atypical with approximately 90% accuracy.
3. CCS of atypical samples tended to be overestimated when a single model was used. The methodological framework developed in this thesis was able to remove some of this bias without increasing overall model error.

These results have important implications for the Australian sugarcane industry as well as the broader NIR analysis community.

The classification approach developed here can be used to identify the sources and causes of atypical cane. This will allow for appropriate interventions to be taken and ultimately reduce the occurrences of ‘atypical’ cane consignments. The novel modelling framework developed here can be tuned for a specific task without the need to completely rebuild the classification model. This gives the ability to quickly reflect changes in the risk associated with misclassification. Partial least squares is a lightweight, easy to adjust and interpretable modelling approach. The relatively simple and well-understood nature of PLS means that model maintenance can be performed quickly. Industry familiarity with the technique will also facilitate uptake of the methodologies described in this thesis.

The high skill shown for PLS modelling approaches compared to more complex machine learning techniques is an important contribution as a counterpoint to published research that shows a clear advantage for complex techniques. The results reinforce the need for future researchers to consider a range of modelling approaches and data pre-processing to find the most appropriate modelling framework for the task at hand. The inclusion of the class probability as a tuneable parameter in the methodological framework was a unique example of how classification information can be used in a practical online NIR analysis setup. The outcomes and insights from this thesis can be used to inform future researcher, not only for the case of atypical cane samples but for any application for imbalanced or complex discrimination tasks.

Table of contents

List of tables

List of figures

Publications

Chapter	Details of publication(s)	Status
3	Sexton, J., Everingham, Y. & Donald, D. A comparison of data mining algorithms for improving NIR models of cane quality measures. Proceedings of the Australian Society of Sugar Cane Technologists, 2017 Cairns, Queensland, Australia. 557-567.	Published
4	Sexton, J., Everingham, Y., Donald, D., Staunton, S., & White, R. 2018. A comparison of non-linear regression methods for improved on-line near infrared spectroscopic analysis of a sugarcane quality measure. Journal of Near Infrared Spectroscopy, 26(5), 297–310. https://dx.doi.org/10.1177/0967033518802448	Published
5	SEXTON, J., Everingham, Y. & Donald, D. A feasibility test for detection of atypical cane samples using near infrared spectroscopy Proceedings of the Australian Society of Sugar Cane Technologists, 2018 Mackay, Queensland, Australia.	Published
6	Sexton, J., Everingham, Y., Donald, D., Staunton, S. & White, R. 2020. Investigating the identification of atypical sugarcane using NIR analysis of online mill data. Computers and Electronics in Agriculture*,* 168**,** 105-111. https://dx.doi.org/10.1016/j.compag.2019.105111.	Published

Thesis overview

NIR methods for sugarcane are advanced and work well for most (e.g. 90%) samples. However, calibrations of NIR technologies can fail to estimate the true value of quality measures for atypical or 'outlying' samples accurately. Advances in data science technologies in recent years offer new datamining algorithms and approaches that have not widely been considered before in the Australian sugar industry. The objective of my thesis was to develop and compare statistical data mining methodologies to accurately measure cane attributes for anomalous cases from NIR spectra contained within large NIR databases. Specifically, my research objectives were to:

1. Investigate the use of data mining and machine learning algorithms for improved NIRS estimates of cane quality.
   - Can data mining algorithms improve estimates of cane quality?
2. Investigate the use of NIR spectroscopic analysis for the automatic identification of atypical cane samples.
   - Can NIR analysis be used to identify atypical cane?
3. Investigate the use of NIR classification data to improve estimates of cane quality parameters, for atypical cane samples.
   - Can class predictions be used to improve estimates of cane quality for different classes of cane?

In this thesis I pursued these objectives in a systematic approach, first comparing modelling approaches for estimating cane quality, then extending to classification and finally merging lessons learnt from both to develop class-based quality estimates. This overview outlines how this process is presented within the thesis (Figure 1).

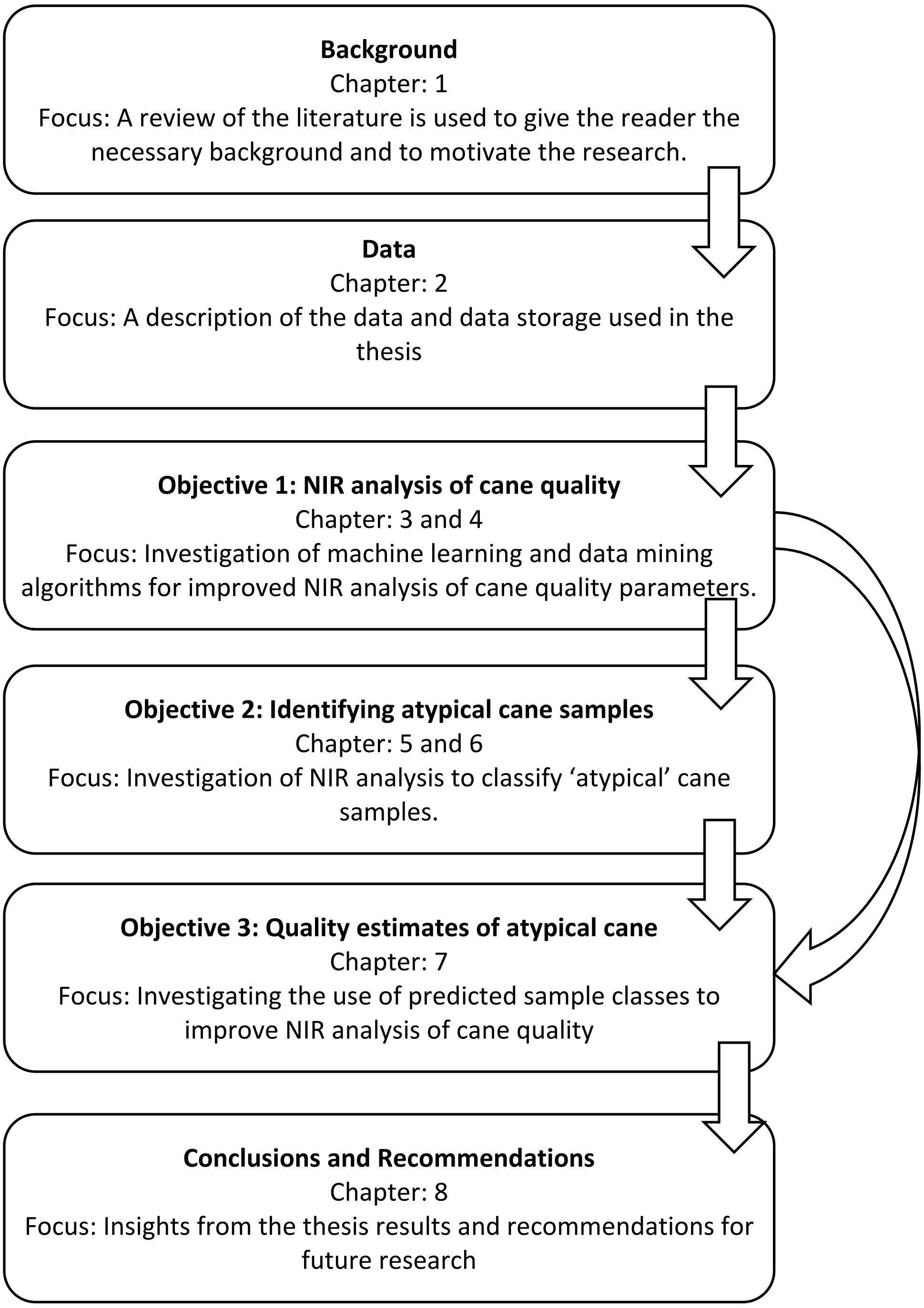
Background
Chapter: 1
Focus: A review of the literature is used to give the reader the necessary background and to motivate the research.
Data
Chapter: 2
Focus: A description of the data and data storage used in the thesis
Objective 1: NIR analysis of cane quality
Chapter: 3 and 4
Focus: Investigation of machine learning and data mining algorithms for improved NIR analysis of cane quality parameters.
Objective 2: Identifying atypical cane samples
Chapter: 5 and 6
Focus: Investigation of NIR analysis to classify 'atypical' cane samples.
Objective 3: Quality estimates of atypical cane
Chapter: 7
Focus: Investigating the use of predicted sample classes to improve NIR analysis of cane quality
Conclusions and Recommendations
Chapter: 8
Focus: Insights from the thesis results and recommendations for future research

Figure 1. Flow diagram of Thesis chapters.

In Chapter 1 I introduce key concepts explored in the thesis such as near infrared spectroscopy, machine learning and sugarcane quality, all in the context of the Australian sugarcane industry. The purpose of Chapter 1 was to help situate the research and build motivation for the thesis objectives. While Chapter 1 gives an overview of key concepts, each chapter is presented as an individual research paper and is therefore readable as a stand-alone document.

The objective of Chapter 2 was to provide an overview of the data sources and data types I have used in the thesis. The data used in my thesis were sourced from a single mill in northern Queensland, Australia. The mill data was collected into a single relational database for simplicity. The relational database made it much simpler to extract and compare data. Each subsequent Chapter uses a selection of data from the database depending on the requirements of the particular experiment. Therefore, details of the data used is provided in each chapter.

In Chapters 3 and 4 I focus on Objective 1: Investigating the use of data mining and machine learning algorithms for improved NIRS estimates of cane quality. Within the Australian sugarcane industry, partial least squares regression (PLSR) has been used to build NIR models of cane quality measures in the lab, on-line and in the field. PLSR relies on the linear relationship between sample constituents and electromagnetic absorption at NIR wavelengths. In practice, this linear relationship can often break down resulting in relationships that are more complex. Recently, machine learning techniques have become popular for their skill with complex data and ability to produce robust calibrations.

The objective of Chapter 3 was to compare PLSR with the machine learning technique support vector regression (SVR). The two techniques were used to estimate three cane quality parameters: brix in juice (Bij), pol in juice (Pij) and apparent purity (Pij/Bij). The results I present in Chapter 3 show that the machine-learning algorithm SVR was comparable to the industry standard approach using PLSR across a range of quality measures. Importantly, the results of Chapter 3 showed that the comparison between PLSR and SVR was similar for each of the quality measures and that many of the same samples were difficult to estimate for both techniques. These results are important because it suggested that there was no advantage to using a different modelling approach for different quality measures. In Chapter 4 I have made use of this fact, concentrating on comparing a wider range of modelling techniques for a single quality measure.

In Chapter 4 I have compared models on their ability to estimate Commercial Cane Sugar (CCS). CCS is the primary quality measure used to calculate cane payments to growers. Therefore it is important to be able to quickly and accurately assess in the mill. PLSR was used as a baseline and was compared to SVR, as well as Artificial Neural Networks (ANN) and gradient boosted regression trees (GBT). The inclusion of ANN and GBT gave a wider range of types of modelling approaches. Similar to SVR, ANN and GBT have shown promise for modelling complex, non-linear relationships. All three techniques approach complexity in different ways. This chapter also placed greater emphasis on variable importance, identifying NIR wavelengths that were influential in each of the models used in the comparison. This type of variable importance investigation has rarely been applied to ANN and SVR models.

The results I present in Chapter 4 confirm that PLSR was as effective as SVR and ANN but that GBT failed to perform as well as other techniques. This was mirrored in the variable importance comparison which showed that PLSR, SVR and ANN placed greater value on similar wavelength regions while GBT placed much higher significance on a small number of wavelengths. This was a valuable contribution to the Australian sugarcane industry and the wider modelling community as it was possible to show why the GBT model underperformed. The variable importance investigation also showed that it was possible to see inside the 'black-box' of ANN and SVR. This is a crucial step in building confidence in using machine learning modelling approaches. The findings and insights of Chapter 3 and Chapter 4 were important for building towards the discrimination between atypical and typical cane samples as there were fewer examples of discrimination or classification tasks within the sugarcane industry. By first focusing on quality estimation I was able to identify that the types of models and comparison approaches used in my thesis were appropriate for the discrimination tasks investigate in Chapters 5 and 6.

In Chapters 5 and 6 I have focused on Objective 2 of the thesis: Investigating the use of NIR spectroscopic analysis for the automatic identification of atypical cane samples. Mill researchers have identified that in any given season, between one and five percent of samples have unusually low laboratory estimates of Pol in juice given their measured Brix in juice. These 'atypical' samples are of particular concern as they can represent deteriorated or contaminated cane samples. Deteriorated or contaminated cane has a number of negative impacts on the cane milling process such as increasing crystallisation times and requiring more frequent cleaning and maintenance periods. Furthermore, lower quality of deteriorated cane means less sugar produced and lower profits for growers and millers. The ability to rapidly identify atypical samples will lead to the ability to track the sources and allow for interventions that can stop atypical cane from arriving at the mill.

The objective of Chapter 5 was to define atypical samples based on laboratory Bij and Pij and test the feasibility of discriminating between atypical and typical samples based on NIR data. I developed a definition for atypical samples was designed based on a linear regression between

Pij and Bij. I then used partial least squares discriminant analysis (PLS-DA) to build discriminate models based on NIR spectral data. In practice only approximately three percent of all samples were defined as atypical. This large imbalance between classes made discrimination a potentially difficult task. The definition of atypical samples that I developed in Chapter 5 was well received when presented at an industry conference as they visually matched atypical samples in a plot of Pij and Bij values. The definition of atypical samples also matched temporal trends in apparent purity. As there was no previous definition for atypical cane it was important that I was able to show my definition was an appropriate and useful measure. Furthermore, The PLS-DA model I developed was able to correctly classify approximately 86% and 92% of atypical and typical samples respectively. This was a crucial result. Not only because discrimination tasks are much rarer in the Australian sugarcane industry, but because atypical samples made up a very small portion of all cane processed by the mill. It was important that I was able to show that it was feasible to discriminate between atypical and typical cane before further investigations could be undertaken because so few examples were available in the literature.

In Chapter 5 I showed that it was feasible to discriminate atypical and typical samples, In Chapter 6 I expanded on this research to compare a range of modelling approaches. Five modelling approaches were considered: PLS-DA, Linear Discriminant Analysis (LDA), random forest (RF), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Furthermore, in Chapter 6 I considered a range of spectral pre-processing techniques including Standard Normal Variate (SNV), Savitzky-Golay first and second derivatives and three wavelet transformations. Finally, given the identification of important wavelengths in Chapter 4, I applied a feature selection process to the best model in order to investigate whether reducing the number of wavelengths improved model performance. It was necessary to consider a range of modelling approaches and spectral pre-treatments in order to identify any important interactions and provide a reference point for future research. The investigation of feature selection was also important not only as a potential method for improving model performance but to show that it was possible to reduce model complexity in a transparent and automatic manner. This was important in the sugarcane context as there was no prior expert knowledge to say what wavelengths may be important for discriminating between atypical and typical cane.

The results I presented in Chapter 6 echoed the results of Chapter 4, showing that the simpler PLS based approach was as or more effective than more complex machine learning approaches. My results also showed that some spectral pre-processing approaches were more effective for certain modelling approaches and that feature selection could improve model performance. The most important result was the ability to discriminate between atypical and typical cane samples using PLS-DA, given that PLS approaches are well understood within industry. This result also meant that it was worthwhile to continue the research to see if this classification data could be used to improve quality estimates in Chapter 7. However, it was also an important contribution to the literature to emphasise the importance of testing data pre-processing and how calibration data is set-up, rather than only testing a range of modelling approaches.

Finally, in Chapter 7 I brought together the methodologies explored in earlier investigations, in order to address Objective 3: Investigating the use of NIR classification data to improve estimates of cane quality parameters for atypical cane samples. In Chapter 7 I developed a process-based approach to estimating cane quality measures for atypical samples (Figure 2). A PLS-DA model was used to predict the probability of a sample being atypical. Three sugarcane quality measures (Pol in juice, Brix in juice and CCS) were then estimated using partial least squares regression. If a sample was identified as atypical an atypical specific PLSR model was used to estimate quality parameters.

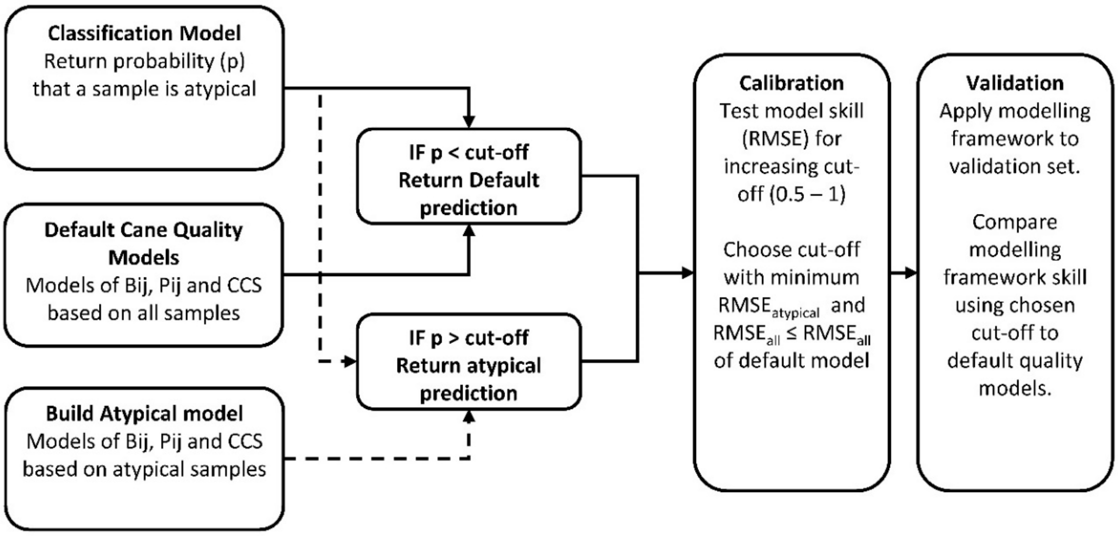
Classification Model
Return probability (p) that a sample is atypical
Default Cane Quality Models
Models of Bij, Pij and CCS based on all samples
Build Atypical model
Models of Bij, Pij and CCS based on atypical samples
IF p < cut-off
Return Default prediction
IF p > cut-off
Return atypical prediction
Calibration
Test model skill (RMSE) for increasing cut-off (0.5 – 1)
Choose cut-off with minimum RMSE$_{atypical}$ and RMSE$_{all}$ ≤ RMSE$_{all}$ of default model
Validation
Apply modelling framework to validation set.
Compare modelling framework skill using chosen cut-off to default quality models.

Figure 2. Methodology overview for modelling cane quality.

The result I present in Chapter 7 showed that Pol-based quality estimates (Pij and CCS) for samples identified as atypical are over-estimated using a baseline PLSR approach. By making use of the probability of a sample being atypical, I was able to reduce this bias without increasing overall model root mean square error. My results show that NIR analysis can be used not only to identify and track atypical samples, but in process control within the mill. By using class probability as a tuneable parameter, it was possible to modify NIR models to achieve a desired outcome. The most novel aspect of the process-based modelling framework developed in Chapter 7 was the use of the class probability as a tuneable parameter. While there is evidence of class based modelling approaches, there was no evidence of this type of flexibility used in the current literature.

In Chapter 8 I present the conclusions of the thesis and discusses the key outcomes and insights from the thesis in terms of the three thesis objectives/research questions. Primarily the outcomes and insights of the thesis are presented in terms of their importance for the Australian sugarcane industry. However, the contribution of the research to the wider research community is also discussed.

Chapter 1

NIR Spectroscopy and sugarcane quality: Current practices and alternatives for the Australian sugarcane industry.

1.1 Introduction

The Australian sugarcane industry strives to remain economically and environmentally sustainable. In order to achieve this, the industry funded Sugar Research Australia (SRA) targets research development and extension programs for the industry (SRA, 2014). The key focus areas of the SRA include variety development, production management, milling efficiency and capability development. In a recent update to their strategic plan, SRA further identified several priority impact areas including plant breeding and maximising productivity along the value chain (SRA, 2015). The key focus area of milling efficiency and technology seeks innovations that improve mill processes and contribute to the long-term sustainability of the milling sector (SRA, 2015). One of the main objectives of this key focus area is to identify solutions for cane quality issues along the value chain. In 2017 alone Australia produced approximately 36 Million tonnes of sugarcane (FAO, 2019). Revenue from sugar exports for 2017 in Australia were valued at 1,500 Million $AUD (http://asmc.com.au/industry-overview/statistics/). Given that sugarcane production in 2017 for the top 5 producing countries ranged from 758 Million tonnes in Brazil to 73 Million tonnes in Pakistan (FAO, 2019, FAO, 2017), the world market for sugarcane can be seen as very competitive, making sugar quality increasingly important. Unsurprisingly then, sugarcane quality are an integral part of the payment system to growers in Australia.

The cane price paid to Australian growers has historically been based primarily on commercial cane sugar (CCS) calculated from measures of Brix in juice, Pol in juice and percent Fibre. Brix in Juice (Bij) can be defined as the concentration of total sugars in grams per 100 gram of solution and can be measured by brix spindle (BSES, 1991) or refractometer (Nawi et al., 2014). Brix is also referred to as Total Soluble Solids (TSS) in other sugar industries (Saxena et al., 2010). Pol in juice (Pij) is a measure of the percent sucrose in juice and is measured by polarimeter. Polarimeters measure the optical rotation of plane polarized light as it passes through a solution. As sucrose is an optically active substance, if it is the only constituent in a solution, the polarimeter reading relates directly to the concentration of sucrose in the solution (McCarthy, 2003). Therefore, time must be taken to clarify sugarcane juice before Pol can be used as a measure of sucrose content (Nawi et al., 2014). These ‘wet’ chemical analyses can be both expensive and time consuming. Fibre content is typically measured as a 3-day rolling average of representative prepared cane sub-samples based on variety groups. The fibre content of a cane sample expressed as a percentage (%Fibre) is used along with Pol and Brix measures in calculating CCS (1-1). CCS is described by Hogarth and Allsopp (2000) as “a measure of pure sucrose that is obtainable from the cane” and is measured as:

$$CCS = PIC - \frac{(BIC - PIC)}{2} \tag{1-1}$$

where,

$$PIC = \frac{Pij(100 - (\%Fibre + 5))}{100} \tag{1-2}$$

and

$$BIC = \frac{Bij(100 - (\%Fibre + 3))}{100} \tag{1-3}$$

Although more recently molasses and Fibre quality have been included in cane payments (Pollock et al., 2007), Brix, Pol and CCS are still important quality measures in Australia.

Deterioration or contamination of sugarcane either pre- or post-harvest can have adverse effects on the measurement of Pol and Brix and is likely to produce atypical samples in sugarcane NIR analysis systems. Deterioration can be caused by bacterial infections. During deterioration, sucrose is metabolised into less economic products such as organic acids, complex polysaccharides (e.g. dextran) and gums (Solomon, 2009). Deterioration due to delays between harvesting and crushing can lead to increased dextran levels and higher Brix readings (Saxena et al., 2010). The presence of complex sugars and gums can cause higher viscosity and longer crystalization times (Solomon, 2009) and hence can result in greater need for mill maintainence. Lionnet (1986), designed mathematical models of cane deterioration indices as delay increased and found that as deterioration increased, Pol became an unreliable measure of sucrose leading to underestimates of sucrose content. Contamination of cane can be considered high levels of leaf matter or soil. While deterioration can affect Pol, contamination can inflate laboratory Brix values calculated by hydrometer. As deterioration and contamination can directly affect measures of quality parameters, they will affect grower payment calculations. Unfortunately, current methods of accounting for products such as dextran are either long and complicated, not specific or expensive and cannot be used in cane payment systems (Van Heerden et al., 2014).

A major innovation of the milling sector was the adoption of efficient near infrared (NIR) spectroscopy technologies for the rapid assessment of cane quality measures on-line (during milling) and the use of these measurements in cane payment calculations. Over the past 2 decades near infrared (NIR) spectroscopic data has been collected and analysed at mills in the Australian Sugar industry by on-line systems. On-line NIR Cane Analysis System used in the sugar industry are capable of assessing cane quality parameters such as Brix, Pol, Fibre, ash content and sugar content. This data is used in grower cane payment calculations as a cost effective and rapid alternative to laboratory analysis and has led to a significant decrease in the costs associated with assessing cane quality. However, calibrations of NIR technologies can fail to estimate the true value of quality measures for anomalous or 'atypical' samples such as deteriorated or contaminated cane. While the effect of cane deterioration on laboratory analysis has been investigated, there is no clear research on the effects of such atypical samples on NIR analysis.

Advances in Data Science technologies in recent years offer new datamining algorithms and approaches that have not widely been considered before in the Australian sugar industry. These technologies should be explored to determine if they can deliver a solution to:

1. Identify atypical samples from NIR spectra and
2. Analyse quality parameters for these samples.

To ascertain the benefits of these newer technologies, it will be important to benchmark existing techniques currently adopted in the Australian sugar industry. Near infrared spectroscopic methods have been used to estimate quality parameters, there is no evidence of their use in the detection or calibration of spectroscopic models for deteriorated cane or atypical samples in general.

1.2 NIR Spectroscopy

Near infrared spectroscopy is a fast, efficient, non-destructive method for analysing the constituents of biological and chemical samples. The underlying physical principle of NIR spectroscopy is that materials absorb energy from electromagnetic radiation resulting in the vibration, rotation and stretching of molecular bonds. The energy absorbed is related to specific wavelengths of the electromagnetic spectrum. Absorption in the near infrared range (700 nm – 2500 nm) and mid infrared range (MIR; 2500 nm – 5x10^4 nm) are related to the vibration of organic and water molecular bonds such as C-H, N-H, O-H and C=O bond (Agelet and Hurburgh, 2010). An important development for NIR analysis was the determination of moisture content in whole grains (Massie and Norris, 1965). Later research allowed the assignment of wavelengths in NIR spectra of agricultural products to food constituents (Osborne et al., 1993b). For example, sucrose can be related to absorption at wavelengths of 1440 nm and 2080 nm. The use of the NIR region of the electromagnetic spectrum has the advantage of enabling transmission through samples intact due to the longer path lengths compared to the MIR yet has the drawback of being sensitive to particle size and sample inhomogeneity.

1.2.1 NIR spectroscopic analysis

The interpretation of NIR absorption spectra would not be possible without appropriate multivariate mathematical techniques. In order to build a model of constituent concentration in a material (e.g. moisture in grains or sucrose in sugarcane) a series of data collection and pre-processing is required before the model can be developed. **Figure 1.1** outlines the model building process. In order to build a model, raw spectral data and reference data need to be collected for each sample. Reference data are the desired predictor variables (e.g. Pol, Brix and Fibre for sugarcane) calculated from standard laboratory methods. Spectral data can be collected using a range of instrumentation. Recently Fourier Transform NIR (FT-NIR) interferometers have increased in popularity and differ from more traditional instruments due to their ability to achieve high signal to noise ratios (Agelet and Hurburgh, 2010).

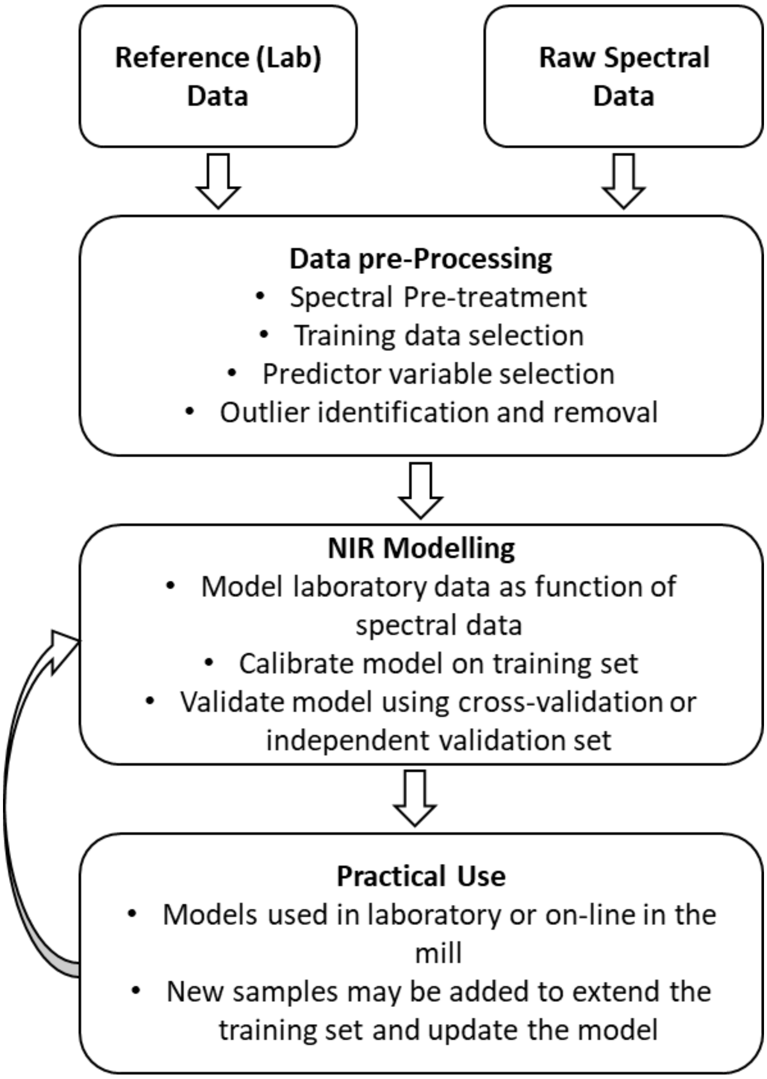
Reference (Lab) Data
Raw Spectral Data
Data pre-Processing
• Spectral Pre-treatment
• Training data selection
• Predictor variable selection
• Outlier identification and removal
NIR Modelling
• Model laboratory data as function of spectral data
• Calibrate model on training set
• Validate model using cross-validation or independent validation set
Practical Use
• Models used in laboratory or on-line in the mill
• New samples may be added to extend the training set and update the model

Figure 1.1. Process of NIR model development. Data collected and pre-processed before a model of the reference values is calibrated and validated. The model can be updated to include new samples identified during operational use**.**

1.2.2 Data pre-processing

1.2.2.1 Spectral pre-treatment

The most common spectral pre-treatments are mean multiplicative scattering correction (MSC) (Geladi et al., 1985), standard normal variate (SNV) transformation (Barnes et al., 1989), and 1st or 2nd derivative spectra (Agelet and Hurburgh, 2010, Osborne et al., 1993b, Rinnan et al., 2009). More recently, wavelet transforms have been used as improvements to NIR pre-treatment (Donald et al., 2006, Mallet et al., 1998, Cen et al., 2006). MSC and SNV both aim to reduce spectral distortion due to the scattering of light off of the sample while spectral derivatives aim to remove the effect of overlapping peaks in the spectra and remove spectral baseline offset (shift) and slope (linear additive variation) across the wavelengths (Agelet and Hurburgh, 2010).

Often some form of MSC or SVN is used in conjunction with a first or second derivative of the spectra.

A key consideration in the use of MSC or SNV is that while SNV works on individual spectra, MSC requires a baseline or reference spectrum and is built up on the whole spectra set (Agelet and Hurburgh, 2010, Sabatier et al., 2014). This requires the storage and maintenance of MSC equations so that the same transformation can be made when used for prediction (Rinnan et al., 2009), which may make it impractical for on-line or large datasets. Rinnan et al. (2009), provide the mathematical basis for MSC and SNV and the reader is referred to their work for a more detailed overview of common pre-processing techniques.

A first order derivative will remove constant (horizontal) baseline shifts while a second order derivative will remove linear sloping shifts that many biological NIR spectra contain. The two most common derivative methods used in NIR spectroscopy are the Norris or Norris-Williams derivative (Norris, 1983, Norris and Williams, 1984) and the Savitzky-Golay filter (Savitzky and Golay, 1964) (Rinnan et al., 2009). While the Norris derivative mimics a finite difference the Savitzky-Golay derivatives are determined by least squares fitting of a polynomial (Rinnan et al., 2009). The two techniques generally will not produce the same derivative spectrum however, modelling accuracy can be similar using either approach (Rinnan et al., 2009).

Spectral pre-treatment in the Australian sugar industry has been largely dependent on proprietary software used in the collection of spectral data such as WinISI^{TM} and Unscrambler^{TM}. Combinations of spectral derivatives and SNV transformations seem to be the most widely used in this industry to date (Berding and Brotherton, 1996, Brotherton and Berding, 1998, O'Shea et al., 2011) although much of the published work does not detail specifics of the treatments used (Berding et al., 1989, Berding and Marston, 2010, Brotherton and Berding, 1995, Staunton et al., 1999, Staunton et al., 2004). Although not explicitly stated in publications, these proprietary software packages tend to use a Norris-Williams approach to calculating spectral derivatives (Guthrie, 2005).

The Savitzky-Golay derivative has often been reported in similar areas of research such as spectral classification of soils, varietal discrimination of wine grapes and best and disease resistance analysis in sugarcane (Araújo et al., 2014, Sabatier et al., 2014, Gutiérrez et al., 2016). How appropriate and effective a given pre-treatment method is will depend on the individual analysis. Subsequently, the lack of detail in current literature on the method used, due in large part to its’ proprietary nature, is a concern as it is difficult to assess if there is a preferential method for analysis of sugarcane quality parameters.

1.2.2.2 Training data selection

In building a calibration model it is important to identify sources of variation in the data and collect samples for use in calibration that are likely to cover future variability. While increasing the number of samples used in a calibration can help cover the likely range of variability, it can also increase noise and cause computational problems as years of data build up. Methods for sampling the training set that better represent the structure of the data such as uniform random sampling, the Kennard-Stone algorithm and the D-optimal method can help improve NIR models (Cao, 2013). Using data selection methods to reduce a large soybean database, Cao (2013) was able to improve partial least squares based NIR models. Cao (2013) recommended D-optimal (de Aguiar et al., 1995) or uniform random selection as efficient and effective alternatives to using all available data. The principle of the D-optimal method is to maximize the determinant of the variance-covariance matrix of the training dataset, while uniform random sample seeks to select samples that cover the whole range of the training dataset.

The uniform random selection process has close parallels to the rectangular distribution approach taken by the sugar industry in attempting to cover the distribution of the reference data (Staunton et al., 1999). As the amount of data collected by the Australian sugar industry has grown, the need to update and remove redundancy in spectral libraries has been addressed. Berding and Marston (2010), describe “combing” the spectral library of a stand-alone NIR analysis system in order to reduce the number of samples while maintaining overall coverage. Statistical methods such as the D-optimal or a formal uniform sampling procedure could be used to improve the stability of NIR models used in the Australian sugarcane industry but have not been explored to date.

1.2.2.3 Variable selection

Variable selection seeks to reduce the number of predictor variables by removing regions of the spectra that are uninformative or lead to better model performance. Predictor variable selection can form part of the model calibration step itself. For example, interval partial least squares (iPLS) (Nørgaard et al., 2000), genetic algorithms (Goicoechea and Olivieri, 2003) and iterative predictors and objects weighting partial least squares (IPOW-PLS) (Forina et al., 2003) have been used as variable selection tools in NIR spectroscopy. The iPLS method seeks to remove uninformative regions by building multivariate models based on spectral regions within a moving window of fixed width. The IPOW-PLS and its' progeny, the modified IPOW-PLS (m-IPOW-PLS) (Chen et al., 2005) were developed to build PLS models from spectral data while simultaneously removing outlier samples and redundant spectral wavelengths. This was achieved by iteratively building PLS models and weighting samples and wavelengths for importance. Genetic algorithms are designed to mimic natural selection and can select well defined spectral regions rather than single points throughout the spectrum (Goicoechea and Olivieri, 2003).

There is little evidence in the literature of what spectral variable selection processes are used within the Australian sugar industry. Interval Partial Least Squares and genetic algorithms have been used in NIR models for sugarcane quality measures (Sorol et al., 2010, Valderrama et al., 2007a) but are not currently used in the Australia sugarcane industry. Sorol et al. (2010), compared several variable selection techniques in building calibrated NIR models for Brix in sugarcane juice and concluded that the genetic algorithm approach outperformed the iPLS approach used in their study. Chen et al. (2005), employed IPOWP-PLS and m-IPOW-PLS to build models of sugars in aqueous solutions, but these techniques also have not been used in the sugarcane industry. Chen et al. (2005), reported improvements compared to standard PLS, their final m-IPOW-PLS model also made use of a wavelet pre-processing step that was not used in the comparative PLS models. Therefore, appropriate pre-processing may have provided similar performance boosts. The added complexity of such modelling processes can be counterproductive for large and complex systems.

1.2.2.4 Spectral outliers

The most commonly used types of outlier detection methods for spectral data are distance based methods such as the Mahalanobis Distance (MD). Distance based outlier detection methods, produce a measure of the distance of a sample from the multivariate mean of the all samples. Outliers are then samples that are relatively far from the multivariate mean. Distance based outlier detection can struggle with high dimensionality and non-linearity. The MD approach to outlier identification can fail for data sets with multiple outliers due to masking and swamping (Egan and Morgan, 1998). Multiple outliers skew the measures of central tendency such that true outliers are not detected (masking) and can make normal observations appear as outliers (swamping). In these cases modified techniques are required to ensure outlier interpretability and the scalability of the technique (Han et al., 2011).

Resampling by Half Means (RHM) and Smallest Half-Volume (SHV) are two approaches that have been suggested as alternatives to traditional MD and leverage approaches. Egan and Morgan (1998) showed that these methods outperform MD and leverage based analyses for a range of data sets while Liu et al. (2005) recommended the use of RHM and SHV be used in place of MD and leverage for the detection of outliers based on their NIR analysis of milk.

Another alternative is to consider datamining techniques more often used for cluster or classification model building. The detection of outliers is similar to clustering and classification in a datamining context (Han et al., 2011). Specifically, outlier detection can be thought of as a clustering or classification problem where we are looking for a very small cluster. This has led to the adoption of datamining algorithms for outlier detection (Han et al., 2011, Campos et al., 2016). Campos et al. (2016) considered the development of baseline datasets for testing outlier detection algorithms. The authors focused on the family of K Nearest Neighbours algorithms, indirectly providing an overview of their use in the detection of outliers.

The detection and removal of spectral outliers from training data sets In the Australian sugar industry is based on Mahalanobis distance (Berding and Marston, 2010). This provides a method of updating the calibrations year to year by adding identified outliers that represent novel samples to the calibration data set. A datamining approach to outlier identification has not been greatly explored in NIR spectroscopy or the Australian sugar industry. However, in various industries datamining algorithms have been used for clustering and classification problems and are therefore familiar in the world of spectroscopic analysis.

1.2.3 NIR Modelling

From the available literature it is apparent that PLS regression is the primary method used for developing NIR models in sugarcane industries worldwide (Sorol et al., 2010, Purcell et al., 2012, O'Shea et al., 2011, Ostatek-Boczynski et al., 2013, Oxely et al., 2012). The principle of partial least squares regression is the assumption that the orginal predictor variables can be replaced by a subset of latent variables expressed as linear combinations of the predictor variables. This is well suited to NIR spectral data where wavelengths can be highly correlated which can lead to unstable regressions without unique solutions (Agelet and Hurburgh, 2010). As with principal components in Principal Component Analysis (PCA), the latent variables are defined as orthogonal to each other. Unlike components in PCA, the latent variable in PCA are chosen to have both high variance and high correlation with the reference variable (the outcome being predicted) (Hastie et al., 2013f).

Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are datamining algorithms that are better suited to large and non-linear data sets but have not found widespread use in the sugar industry. ANN were developed to philosophically mimic neurons in the brain by forming a network of connected input and output ‘neurons’ called nodes (Han et al., 2011). Nodes connections are weighted during the training process and weighted inputs are summed and transformed using a transform function to produce an output node (Alam et al., 2008). Nodes can be grouped into layers with the output nodes of one layer forming the input nodes of another. ANN are highly parallelizable and have a high tolerance for noisy data making them ideal for large datamining jobs (Han et al., 2011) but are not widely used due to their complexity.

SVM is a more recent method for non-linear calibration and have been regarded as an effective alternative to Neural Networks (Agelet and Hurburgh, 2010, Balabin and Lomakina, 2011, Kovalenko et al., 2006) and PLS (Thissen et al., 2004a). SVM has been used in predicting banana quality indices (Sanaeifar et al., 2016), discrimination of adulterated milk (Zhang et al., 2014) and determination of amino acid composition of soybeans (Kovalenko et al., 2006) among many other applications. In their study on predicting soil properties from NIR spectra Araújo et al. (2014) compared the performance of multiple PLS models on subsets of spectra to global PLS as well as boosted regression trees and SVM. The SVM model outperformed the global PLS and performed almost as well as the multiple PLS models. Being able to use a single model can be more desirable as often the performance increase of multiple models is overshadowed by the extra work required to maintain the model calibrations.

It is important to recognise that the effectiveness of a modelling approach may differ between types of problems. For example in a gasoline classification problem, Balabin Balabin et al. (2010) described three classes of classification models:

1. Low performance: e.g. Linear Discriminant Analysis (LDA)
2. Medium performance: e.g. PLS and ANN and
3. High performance: e.g. SVM.

Yet, while SVM outperformed Random Forest and LDA classification models for sugarcane varieties based on hyperspectral data (Everingham et al., 2007), Random Forest outperformed SVM for classifying bruised apples (Che et al., 2018). Wang et al. (2004) showed that ANN performed better at identifying types of fungal contamination while PLS could outperform ANN at discriminating between fungal contaminated and uncontaminated soybeans. In developing a NIR spectroscopic model for either a regression or classification problem it is important to test a range of modelling approaches.

1.3 NIR spectroscopy in the Australia sugar industry

Over the past two decades near infrared (NIR) spectroscopic data has been collected and analyzed by the Australian sugar industry. Berding et al. (1989), introduced NIR spectroscopy for the evaluation of cane quality in clonal trials in a laboratory setting, which was further investigated through the early 90's (Berding et al., 1991, Brotherton and Berding, 1995). By the mid 1990's research had extended to at-line (performed at the mill) analysis (Berding and Brotherton, 1996, Brotherton and Berding, 1998). The late 1990's and early 2000's saw on-line (analysis as part of the mill process itself) NIR Cane Analysis Systems (CAS's) used in the sugar industry (Staunton et al., 1999, Staunton et al., 2004). These systems are capable of assessing cane quality parameters.

NIR estimates of quality parameters are used in cane payment calculations as a cost effective and rapid alternative to laboratory analysis (Pollock et al., 2007, Staunton et al., 2004). This has led to a significant decrease in the costs associated with assessing cane quality measures. For example, Berding and Marston (2010) describe a reduction in analytical operation costs to 14% of standard procedures by implementing an on-line NIR based cane analysis system. However, NIR models are challenged to accurately estimate the true value of quality measures for unusual samples or where spectral signatures are anomalous. Failure to identify or accurately estimate these atypical samples can cause a decline in growers' confidence with NIR technologies. Given the economic advatages of NIR analytical techniques it is important to be able to identify these atypical samples .

While quality measures such as Brix, Pol, Fibre and Commercial Cane Sugar (CCS) have been the primary focus of NIR analysis in the Australian sugar industry, there have also been a range of process control applications explored (Simpson et al., 2011). Simpson et al. (2011) identified maceration rate control (Lloyd et al., 2010); clarifier phosphate addition (Markley et al., 2009) and the first naturally low GI (glycemic index) sugar (Kannar et al., 2009) as process control activities that have used NIR analysis. Mapping of productivity data and nutrient levels and mill maintence scheduling are some of the potential uses for NIR analysis in the Australian sugar industry that still need to be explored. On the global scale, sugarcane industries have used NIR analysis for estimating reducing sugar levels (Valderrama et al., 2007b), trace elements such as nitrogen and silicon in mill by products (Purcell et al., 2012), pest and disease resistance (Sabatier et al., 2014) and cellulose and lignin in sugarcane bagasse (Rodríguez-Zúñiga et al., 2014).

Partial Least Squares is the most often used NIR modelling technique within the Australian sugar industry (Ostatek-Boczynski et al., 2013, Nawi et al., 2013, Berding and Marston, 2010, Fiedler et al., 2001, O'Shea et al., 2011, Oxely et al., 2012, Sorol et al., 2010, Staunton et al., 2004) as well as globally (Rodríguez-Zúñiga et al., 2014, Sabatier et al., 2014, Valderrama et al., 2007b, Valderrama et al., 2007a). Artificial Neural Networks have been used within the Australian sugar industry to classify sugar content of sugarcane from NIR spectra collected by scanning sugarcane rind (Nawi et al., 2013). Nawi et al. (2013) developed ANN classification models for five Brix categories with an average accuracy of 83.1% correct classification rate. Support Vector Machines have also been used in classification problems in the Australian sugar industry, using hyperspectral rather than NIR spectral data. Everingham et al. (2007), were able to correctly classify sugarcane variety and crop class using an SVM model based on Hyperspectral satellite imagery. Unfortunately, there are few other cases of techniques such as ANN or SVM within the Australian sugar industry. Potentially this is because fewer classification tasks have been explored within the Australian sugar industry.

The potential for datamining techniques such as SVM and ANN used to build classification models can in the sugarcane industry can be seen in the example of the classification of adulterated milk. Zhang et al. (2014), were able to discriminate been adulterated and unadulterated cow milk samples from NIR spectra as a form of quality control. NIR analytics have also been used to develop models to classify the phenotype or variety. For example, recently Gutiérrez et al. (2016) developed a model to classify grape phenotypes from NIR spectra. Araújo et al. (2014) were able to improve predictive PLSR models of organic matter and clay percentages in soil samples by first clustering soil NIR spectra using a k-means datamining algorithm. Similar methods can be employed by the sugarcane industry to discriminate between 'normal' and deteriorated or other atypical cane samples. Developing models for different types of samples could then potentially be used to improve NIR predictions of quality measures in a similar manner to that of Araújo et al. (2014).

1.4 Conclusion

NIR spectroscopy is a well-established non-destructive analysis tool with a proven track record in the sugarcane industry. Current data pre-processing, outlier treatment and model calibration techniques used in the Australian sugar industry have been adequate for the determination of cane quality measures. Atypical samples such as deteriorated or contaminated cane can have an adverse effect on milling processes and measurements of cane quality parameters in the laboratory. This can lead to lost productivity and inaccurate cane payment determinations. NIR analysis has been used as fast and accurate tool for determining cane quality measures as well as in mill process control. Despite this there is still a lack of research into the identification and treatment of atypical samples or the effect they may have on NIR analytics. There are a number of data mining techniques that could fill this current lack. Classification or clustering techniques can provide a method for identifying atypical samplse for which to build improved models. Alternatively, datamining regression algorithms could be used to capture non-linear relationships rather than developing multiple models. Future research in the sugarcane industry should consider the advances made in other industries in order to compare current practices with innovative new approaches. In particular more research is needed to identify:

1. If new, more complex modelling approaches can improve NIR analysis methods
2. If NIR analysis can be used to identify potentially atypical cane samples and
3. If identification of atypical cane samples can help improve NIR analysis methods

1.5 Chapter 1 Summary

Maintaining a high level of sugarcane quality is vital for the Australian sugarcane industry to remain competative on a global scale. Deterioration and contamination of sugarcane can lead to lower quality sugar production, increase maintenance costs and can adversly affect cane payment calculations. While many of the causes of deteroriation and contamination are understood there is currently no agreed apon identification measure. Near Infrared spectroscopic analysis has been used widely in the Australian sugarcane industry as a fast and reliable method of estimating cane quality measures as well as for process control automation within sugarcane mills. NIR spectroscopic analysis could potentially be used to identify atypical samples such as deteriorated cane and subsequently manage how these samples are treated in the mill. Chapter 1 gives an overview of the importance of quality in sugarcane and how NIR analysis is used within the sugarcane industry. Chapter 1 aimed to provide an overarching background and context for the three objectives explored in the thesis:

1. Investigate the use of data mining and machine learning algorithms for improved NIRS estimates of cane quality (Chapters 3 and 4).
2. Investigate the use of NIR spectroscopic analysis for the automatic identification of atypical cane samples (Chapters 5 and 6).
3. Investigate the use of NIR classification of cane samples to improve estimates of cane quality parameters (Chapter 7).

Chapter 2

Overview of data collection and storage

2.1 Data Acquisition

Data for this project were sourced from the Sugar Research Australia, experiment station at Meringa. Dr. David Donald and Stephen Staunton of SRA. The data were acquired during a visit to the Meringa SRA experiment station in March 2016 provided the data. The data were supplied on an external drive as a copy of industry backup files. All sensitive industry information such as NIR model calibrations were removed from the data before it was acquired. The data represent NIR on-line research collected over the period 1999 to 2015 from 24 Mills associated with Sugar Research Australia and consisted of data from Cane, Sugar or Bagasse Analysis Systems (CAS, SAS and BAS).

Data collected can be considered as one of three main types of data.

1. **Laboratory Data**: This data was collected primarily as spreadsheet data for each Mill and season. This linked samples with laboratory data for quality measures such as Brix in juice (Bij), Pol in juice (Pij), commercial cane sugar (CCS), ash, fibre and dry matter. Samples in these data were primarily identified with unique sample ID numbers referring to a consignment to the mill. This data is required as reference data on which to build NIR models.
2. **Consignment/Productivity Data**: This data was collected primarily as spreadsheet data for each Mill and season. This linked samples with productivity data such tonnage, quality measures such as Bij, Pij and CCS usually estimated by NIR analysis and metadata such as the Farm, Block and Sub-block the sample originated from as well as the variety, crop class and whether the consignment was burnt or green harvested. Samples in these data were primarily identified with unique sample ID number to protect the confidentiality of growers. Consignment data allows researchers to investigate potential sources of variability in the Laboratory data and possibly in model performance.
3. **NIR spectroscopic data**: This data was collected primarily as binary data files generated by CASs. These files contain NIR spectra as well as metadata about when the NIR scan was taken and the instrument used to produce the spectra. As the NIR systems scan automatically as the cane is fed into the mill, a single sample

(consignment) had a number of NIR scans. Individual scans are identified by sample ID as well as a sequence ID. This data can be used to build models of laboratory based quality measures.

In total approximately one terabyte of data were sourced from the Meringa SRA experiment station. This data has been securely backed up to a JCU computing infrastructure to ensure the data is not lost and remains confidential. In order to link the three main types of data together a relational database framework was designed and implemented as a SQLite database.

2.2 Development of database framework

In conjunction with the supervisory team, a framework for a relational database was developed (Figure 2.1). The database framework was designed to express the relationships between productivity data, laboratory data, NIR analysis results and other consignment data, with the spectroscopic data collected from the on-line analysis systems. As can be seen in Figure 2.1 the relationship between different types of data were connected through the unique sample number. Data tables in the database were designed to have a tall and thin design to reduce the sparsity of tables and improve scalability of the database.

As an example of how the database framework operates, we can consider laboratory data collection. Typical laboratory results may include quality measures such as brix in juice (Bij), but measures such as colour are less likely to be measured. The table *LabValue* in Figure 2.1 shows that each laboratory value (*Lab_Value* column) is matched to a measurement type (*Lab_Type* column) rather than having a separate column for each type of measurement. With this setup, if colour was not measured for a particular sample no data would be added, rather than having an empty cell in the table (reduced sparsity). Similarly, if a new type of measure needs to be added the value can be added to the *LabValue* table and an extra entry (row) added to the table of laboratory data types (*Lab_Type* Table; Figure 2.1). This means that an entirely new column does not have to be created every time a new type of measurement is added to the database (reduced database sparsity and improved scalability).

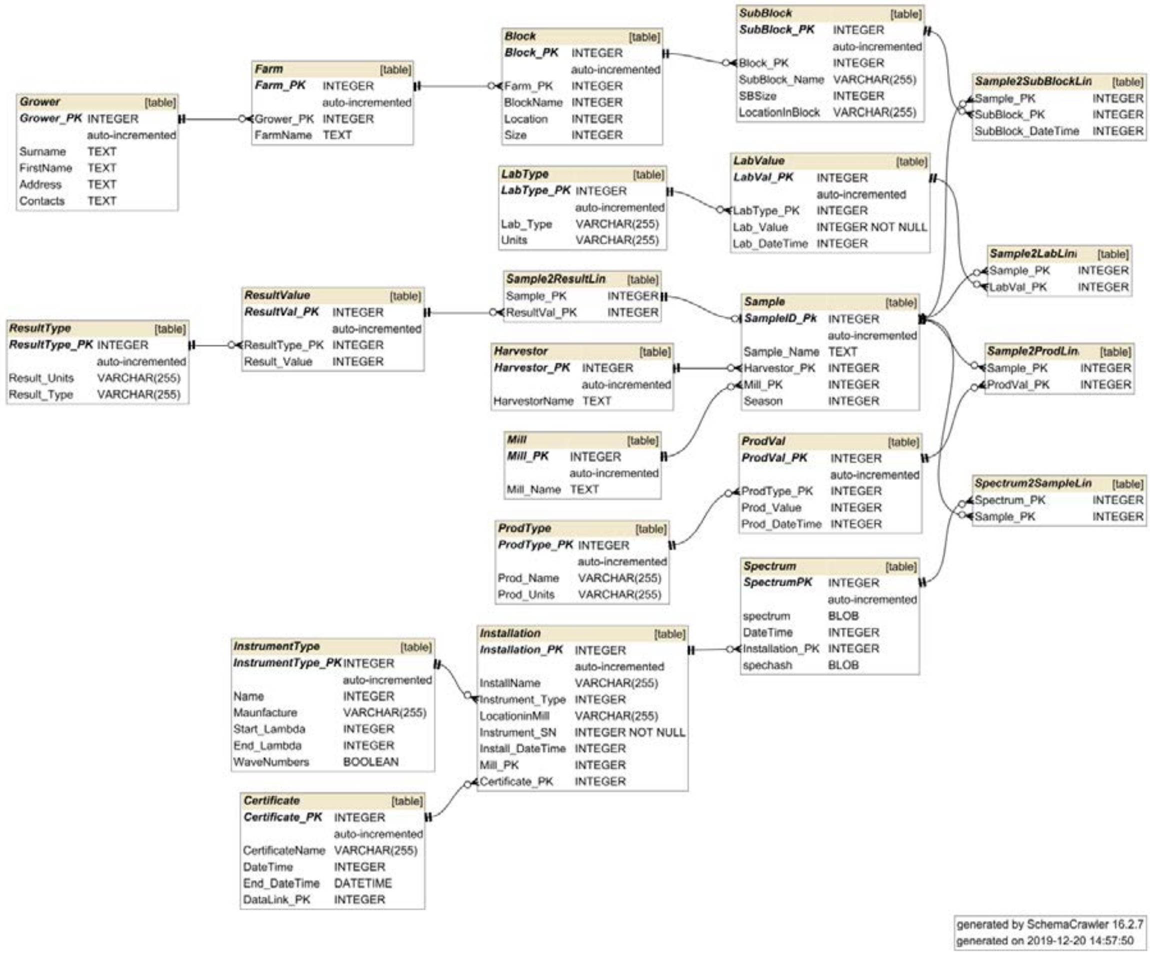

Figure 2.1. Framework template of a relational database designed to store Spectral, productivity and laboratory data for the Australian Sugar Industry. Boxes represent individual tables in the database; labels in black represent columns within each table. Black lines represent relationships between tables and are linked together through "key" columns. E.G., Sample ID's are stored in the Sample table and are linked to a particular Mill through the Mill primary key (Mill_PK).

The database framework has been implemented in the freely available Structured Query Language (SQL) database system SQLite (SQLite 3; www.sqlite.org/copyright.html). A database was created to store the data collected from a northern mill for the period 2004 to 2015. The programming language Python (python 2.7.1; www.python.org) was used to create small programs to collect the data from the numerous source files and store them in the database with some level of automation. Some pre-processing of the source files was required in order to identify relevant data and to check sample ID numbers were consistent between data formats. Some summary statistical analysis were performed in the statistical program R (R Core Team, 2017) by reading data directly from the database.

Data from a northern mill for the period 2004 to 2015 was collected and stored in a single SQLite database. Data collected into the database included spectral data from the on-line NIR Cane Analysis System (CAS), consignment data including farm, block, crop class, variety and productivity data, NIR analysis results and laboratory results for a range of quality measures. The database currently contains >20 GB of data. As the data is now centrally located it is possible to query the database to explore summary statistics for the northern mill. Table 2.1 contains an example of the primary laboratory based quality measure data collected into the database. Measures such as Bij and Pij were analysed for more samples than measures such as fibre. Notably, the 2005 and 2015 laboratory data was not available. The average (mean) of quality measures did not seem to vary greatly between seasons.

Table 2.1. A sample of laboratory analysis data for a northern mill from 2004 to 2015. Quality measures reported here are Brix in juice (Bij), Pol in juice (Pij), Commercial Cane Sugar (CCS) and Fibre. The total number of samples available (N), Mean and Variance (Var) are reported for each season where available.

Season	Bij (%)			Pij (%)			CCS (%)			Fibre (%)		
	N	Mean	Var	N	Mean	Var	N	Mean	Var	N	Mean	Var
2004										25	14.79	0.98
2006	3903	20.74	3.24	3903	18.16	3.75	3903	13.24	2.33	622	15.26	2.08
2007	3392	21.06	3.33	3392	18.39	4.21	3392	13.38	2.71	457	15.15	3.55
2008	3069	22.29	2.29	3069	19.77	2.87	3069	14.44	1.78	492	15.61	3.41
2009	2766	22.02	2.00	2766	19.45	2.34	2766	14.14	1.53	371	15.95	2.71
2010	3193	21.31	2.41	3193	18.66	2.86	3193	13.50	1.85			
2011	3223	22.17	2.42	3223	19.73	3.19	3223	14.40	2.22	455	16.10	3.39
2012	4010	21.83	2.71	4010	19.36	2.67	4010	14.24	1.52	460	14.88	2.60
2013	5631	21.59	1.92	5631	18.96	2.60	5631	13.99	1.99	505	14.65	3.09
2015	4050	21.47	2.58	4050	19.11	3.19	4050	14.30	2.14	434	13.95	2.12

As the database linked laboratory data to consignment and productivity data, it was also possible to consider the distribution laboratory measured quality parameters between factors such as farms (Figure 2.2) or varieties (Figure 2.3) across the 2006 to 2010 period. This period is shown due to the consistent availability of both farm and laboratory data. From Figure 2.2 it is possible to see that although there is often much variability within a particular farm, some farms do stand out. For example, farm "110" had a relatively small variability and a median CCS higher than most farms in the region. By querying the database, it was possible to identify this farm as a research station that had only a limited number of samples. This may explain the low variability and higher performance. From Figure 2.3 it was possible to see that some varieties such as Q220 had a lower median CCS and Purity across the period 2006 to 2010.

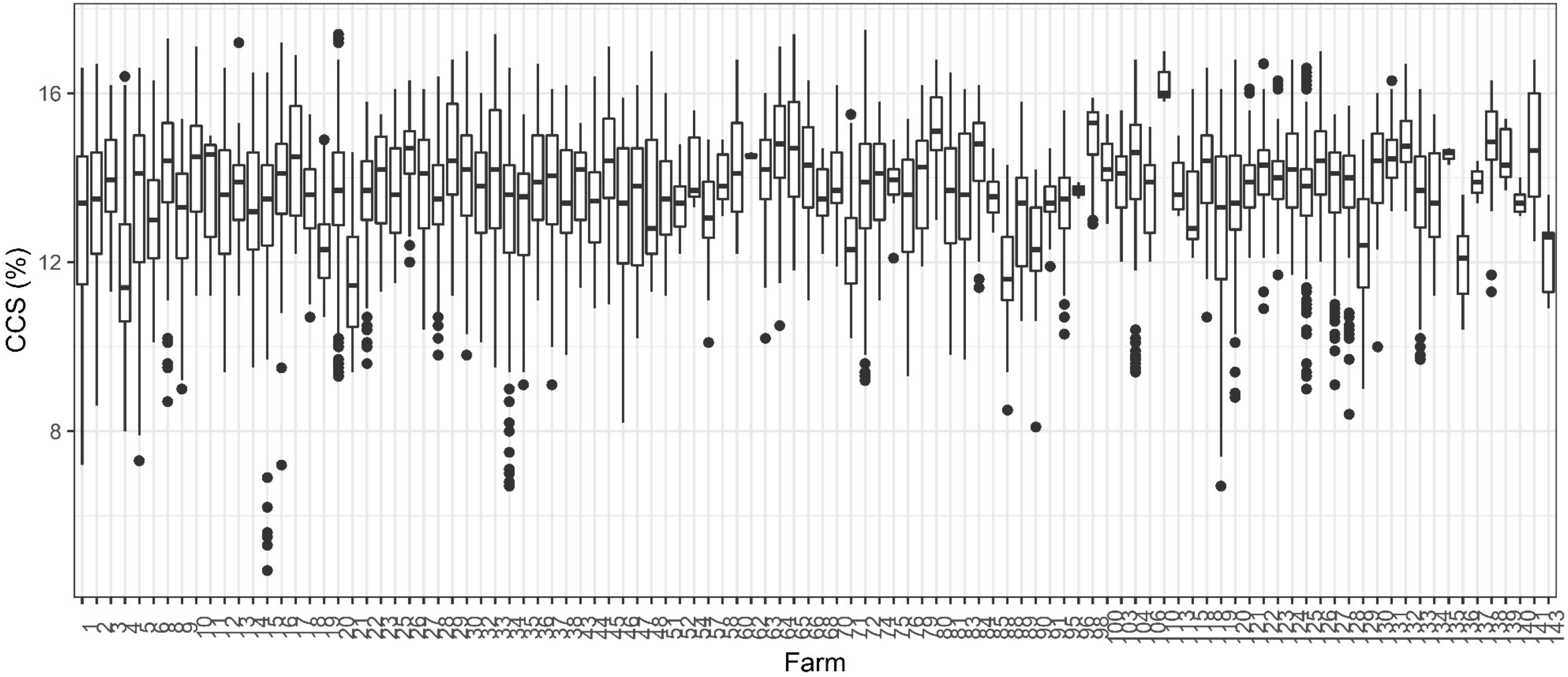

Figure 2.2. Boxplots of laboratory Commercial Cane Sugar (CCS) by farm from a northern mill (2006-2010). Farms are identified by a database specific id number. Boxes represent the 25th to 75th percentiles while solid black lines represent median values. Boxplot “whiskers” cover samples no more than 1.5 times the interquartile range, while points represent ‘outliers’

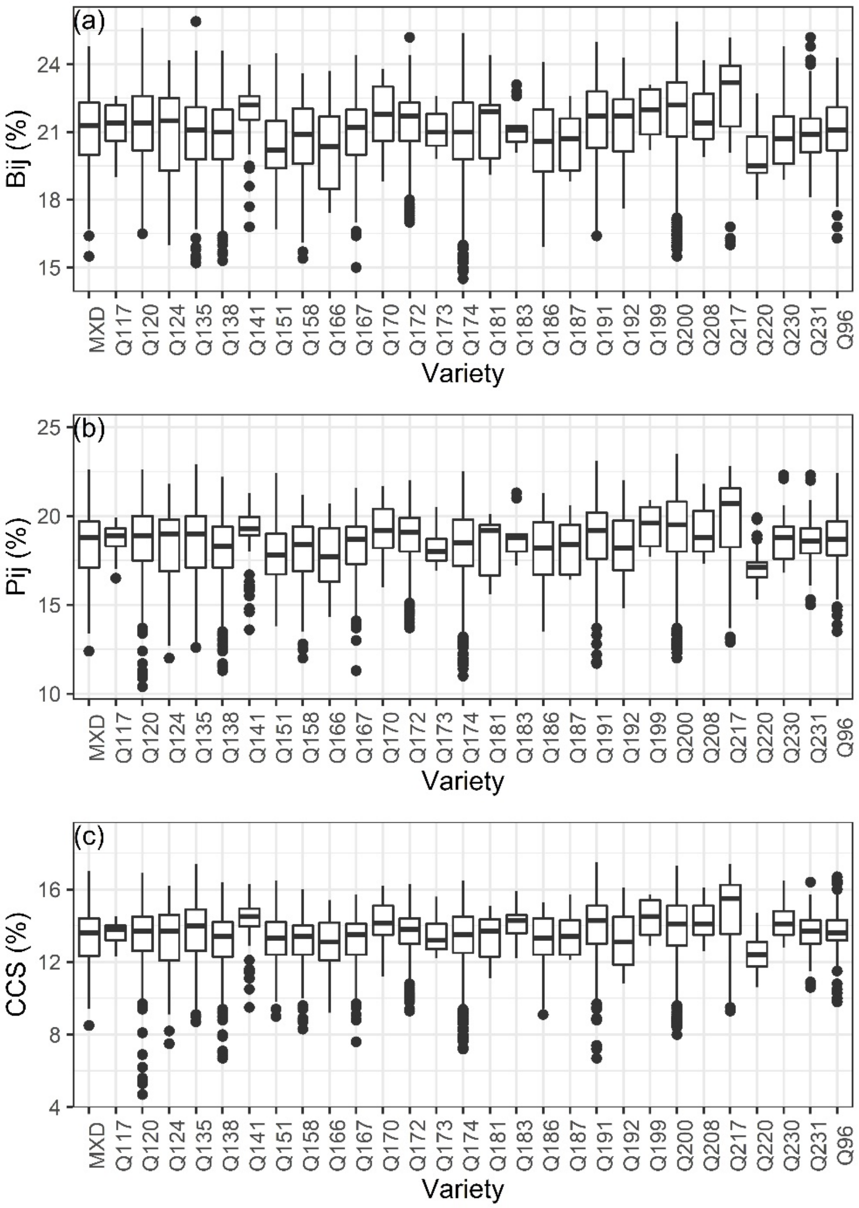

Figure 2.3. Boxplots of laboratory analysis of sugarcane samples by variety from a northern mill (2006 – 2010). Boxplots represent (a) Brix in juice (Bij), (b) Pol in juice (Pij) and (c) Commercial Cane Sugar (CCS). Boxes represent the 25th to 75th percentiles while solid black lines represent median values. Boxplot “whiskers” cover samples no more than 1.5 times the interquartile range, while points represent ‘outliers’.

The ability to link laboratory and consignment or productivity data with the spectral data will enable future research to analyse NIR model performance easily across seasons, farms, varieties and other possible sources of variation. Moving forward, the project will need to consider expanding the database to include more seasons and potentially data from more regions. Considering the size of the database, it may also be necessary to consider alternatives to SQLite as SQLite databases may not be ideally suited to the growing needs of the Australian sugar industry.

2.3 Chapter 2 Summary

For this thesis, productivity, laboratory and spectral data were sourced from the Australian sugarcane industry through Sugar Research Australia. While data were available from several mills, the data used in this thesis was sourced from a single mill in northern Queensland, Australia. The mill data was collected into a single relational database for simplicity. The relational database made it much simpler to extract and compare data. In future, Sugar Research Australia should construct a similar single repository for the NIR spectral analysis. This would also be beneficial in facilitating automation of data storage. The objective of Chapter 2 was to provide an overview of the data sources and data types used in the thesis. Each subsequent Chapter uses a selection of data from the database depending on the requirements of the particular experiment. Therefore, details of the data used is provided in each chapter.

Chapter 3

A comparison of data mining algorithms for improving NIR models of cane quality measures

Relevant publication	Sexton, J., Everingham, Y. & Donald, D. A comparison of data mining algorithms for improving NIR models of cane quality measures. Proceedings of the Australian Society of Sugar Cane Technologists, 2017 Cairns, Queensland, Australia. 557-567.
Statement of intellectual input from co-authors	The research question / objective of Chapter 3 was developed by the candidate with input from Dr. Yvette Everingham, Dr. David Donald and Mr. Steve Staunton. Data for the thesis was provided by SRA through Mr. Staunton. Editorial assistance for Chapter 3 was supplied by Dr. Everingham, Dr. Donald and Mr. Staunton. The candidate developed the methodological framework and ran all simulations. The candidate was also responsible for the write-up of the chapter and produced all tables and figures.
Publication status	Published

3.1 Introduction

Sugarcane quality is regularly measured by near infrared (NIR) analysis on-line in sugarcane mills in Australia (Simpson et al., 2011). On-line NIR Cane Analysis System used in the sugar industry are capable of assessing cane quality parameters such as Brix, Pol, and Commercial Cane Sugar (CCS), which are used in grower cane payment calculations. NIR analysis is used as a cost effective, non-destructive and rapid alternative to standard 'wet chemical' laboratory analysis. Use of NIR technologies has led to a significant decrease in the costs associated with assessing cane quality (Berding and Marston, 2010).

NIR analysis uses chemometric techniques to model the relationship between the absorbance of NIR light by a sample and its chemical composition. The use of the NIR region of the electromagnetic spectrum has the advantage of measuring intact samples. This means samples often require less preparation compared to analysis using other regions of the spectrum such as the mid infrared. NIR analysis was first introduced for the laboratory analysis of cane quality parameters in the late 1980s/early 1990s (Berding et al., 1991, Berding et al., 1989, Brotherton and Berding, 1995). Then, in the late 1990s and early 2000s on-line analysis was introduced, allowing cane quality analysis to become part of the mill process (Staunton et al., 1999, Staunton et al., 2004).

Partial least squares regression (PLSR) is often considered one of the most common, if not the most common method used for developing NIR models (Agelet and Hurburgh, 2010). Within the Australian sugarcane industry PLSR has been used to build NIR models of cane quality measures in the laboratory (Gateway Laboratories; (O'Shea et al., 2011)), on-line (Staunton et al., 1999) and in the field (Nawi et al., 2013). PLSR has also been used to develop NIR-based models for nutrient elements such as carbon and nitrogen (Purcell et al., 2012) and biomass measures including lignin (Oxely et al., 2012). PLSR for NIR analysis relies on the Beer-Lambert law assumption that the quality parameter being estimated and the predictor variables (NIR absorbance at particular wavelengths) is approximately linear (Tange et al., 2015).

In practice, the linear relationship between quality measures and absorption at NIR wavelengths can often break down (Hageman et al., 2005). Non-linearity can be introduced in two main forms: 1) changes to the measuring instrument; and 2) changes in the sample itself (Hageman et al., 2005). Machine wear, repair or replacement generally requires re-calibration of the model (Fearn, 2001). Differences in particle size in the sample, sample deterioration over time or high concentrations of the component being measured can also lead to non-linear effects (Bertran et al., 1999).

Recently there has been increased interest in the use of machine learning algorithms, such as artificial neural networks and support vector regression (SVR) as alternatives to partial least squares, due to their ability to deal with complex data (Tange et al., 2015). However, machine learning algorithms have not been widely considered in sugarcane industries in Australia or internationally. Artificial neural networks have been used within sugarcane industries globally to predict brix and pol from juice samples (Wang et al., 2010). In Australia, artificial neural networks have been used to predict sugar content from cane rind (Nawi et al., 2013).

Previous studies have shown that SVR can outperform artificial neural networks (Thissen et al., 2004a, Balabin and Smirnov, 2012). Recent research as also shown that SVR can produce comparable results to PLSR for sugarcane quality parameters in Japan (Tange et al., 2015, Ramírez-Morales et al., 2016). Current chemometric techniques for cane quality analysis are effective for the vast majority of samples that come through the mill. However, atypical samples such as deteriorated or dirty samples are difficult to model and can result in inaccurate estimates of quality measures. The need to analyse these samples further using standard laboratory techniques can reduce confidence in NIR technologies.

As NIR estimates of cane quality directly influence grower payment calculations, it is vital that we ensure the most robust chemometric techniques available. Given the recent success of machine learning approaches for NIR models, it is timely to consider the use of machine learning techniques such as SVR in Australia. Therefore, the objective of this chapter was to compare SVR with the well-established PLSR for estimating three sugarcane quality parameters (Brix, Pol and Apparent Purity) within the Australian sugar industry.

3.2 Materials and methods

To compare the performance of PLSR and SVR for predicting cane quality measures from NIR spectra, data were collected from a single sugarcane mill located in northern Queensland, Australia. Models for three quality parameters were calibrated and validated using both PLSR and SVR and validation performance was compared. Figure 3.1 outlines the analysis methodology. All data pre-processing, model calibration and model validation were accomplished using the R statistical environment (R Core Team, 2016).

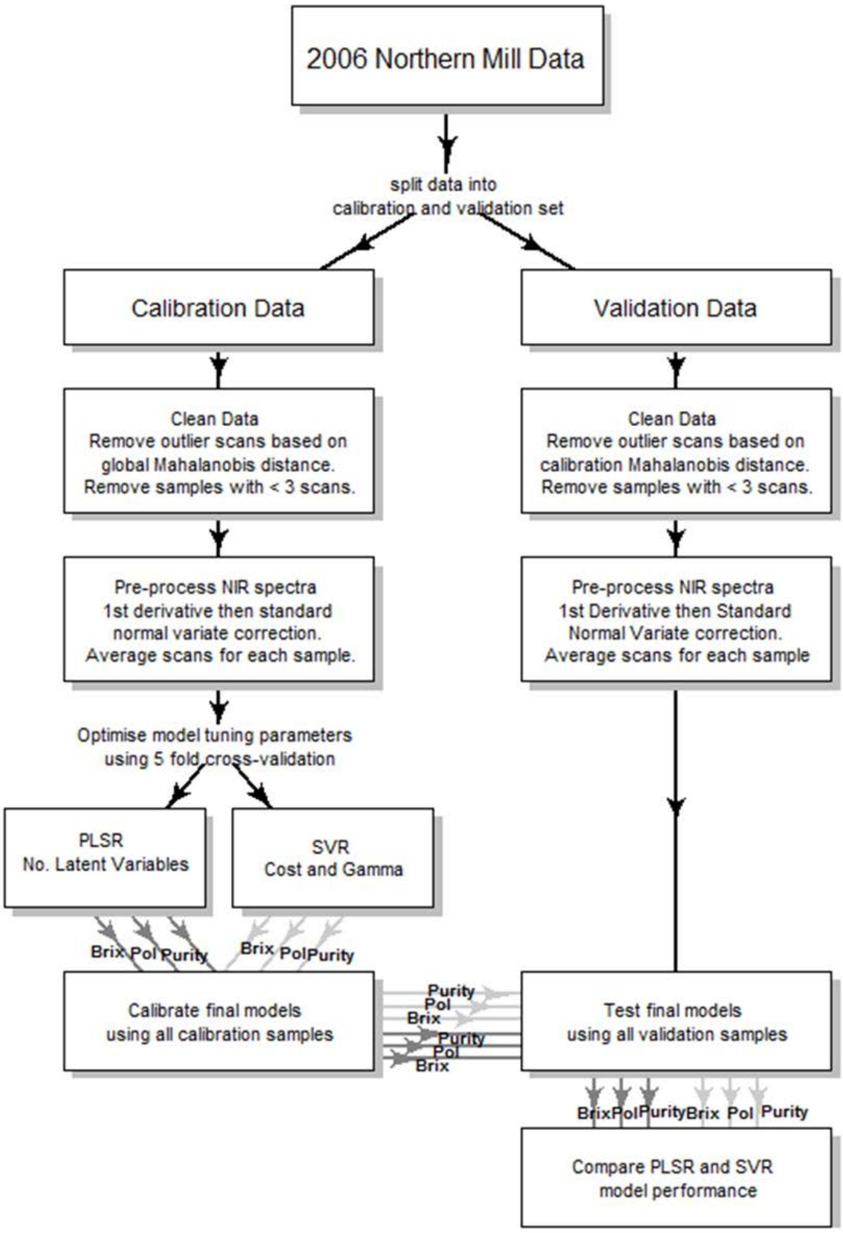
2006 Northern Mill Data
split data into calibration and validation set
Calibration Data
Validation Data
Clean Data Remove outlier scans based on global Mahalanobis distance. Remove samples with < 3 scans.
Clean Data Remove outlier scans based on calibration Mahalanobis distance. Remove samples with < 3 scans.
Pre-process NIR spectra 1st derivative then standard normal variate correction. Average scans for each sample.
Pre-process NIR spectra 1st Derivative then Standard Normal Variate correction. Average scans for each sample
Optimise model tuning parameters using 5 fold cross-validation
PLSR No. Latent Variables
SVR Cost and Gamma
Brix Pol Purity
Brix Pol Purity
Calibrate final models using all calibration samples
Purity Pol Brix
Purity Pol Brix
Test final models using all validation samples
Brix Pol Purity Brix Pol Purity
Compare PLSR and SVR model performance

Figure 3.1. Overview of methodology used in this chapter

3.2.1 Data

Laboratory reference data and NIR spectral data used in this study were collected by the on-line cane analysis system of a single northern mill. Three quality measures were used in the analysis, percent brix in juice (here after referred to as Brix), percent pol in juice (here after referred to as Pol) and apparent purity, calculated as the ratio of Pol to Brix (hereafter referred to as Purity). Data represent on-site laboratory validated samples of the 2006 harvest season. In total 3,794 consignments (samples) were available that had laboratory validated references values of cane quality measures as well as linked NIR data and consignment productivity data including the farm of origin.

For purposes of analysis, 1,899 samples were randomly selected for calibration and the remaining 1,895 samples were used for validation purposes. Samples were split such that no samples from the same farm appeared in both the calibration and validation sets. This was done so that the validation set was as independent as possible from the calibration set.

NIR data linked to each sample were collected using a FOSS ONLINE 5000 system. Spectral wavelengths ranged from 1,100 nm to 2,498 nm at 2 nm intervals. All available wavelengths (700) were included in the analysis. Each sample had multiple scans that were averaged in the data pre-processing phase.

3.2.2 Data pre-processing

There were two key stages to data pre-processing. 1) Data cleaning and 2) data transformation. In cleaning the data, outlier scans were identified in the calibration and validation data using a global Mahalanobis distance. Due to the size of each consignment (sample) delivered to a mill, a single sample may be scanned multiple times resulting in multiple (at times more than 20) NIR spectra (scans) being recorded. Following Staunton et al. (2004), scans with a global Mahalanobis distance greater than 3 were considered outliers and removed from the calibration data set. Scans were removed from the validation data set if they would have been considered outliers in the calibration data set. Following Fiedler et al. (2001), samples were removed from the analysis if they had fewer than three 'clean' scans. Table 3.1 records the final number of samples and the distribution of quality measures used in the calibration and validation data sets. Mean values of quality measures were similar in both the calibration and validation data sets. However, Purity values reached lower levels in the validation data.

Table 3.1. Descriptive statistics of final calibration and validation data sets used in the analysis.

	Calibration (N = 1,857)			**Validation (N = 1,879)**		
	Mean	**SD**	**Range**	**Mean**	**SD**	**Range**
Bij	20.67	1.92	15.2-25.9	20.89	1.57	15.5-24.9
Pij	18.09	2.07	11.7-22.8	18.34	1.68	12.4-22.6
CCS	0.8732	0.0287	0.7368-0.9495	0.8771	0.0251	0.6901-0.9444

NIR spectral scans were transformed using a Savitzky-Golay (Savitzky and Golay, 1964) first derivative with a window width of 17 and scatter corrected using a standard normal variate transformation (Barnes et al., 1989). This combination of pre-treatment was found to result in good models for both PLSR and SVR in preliminary cross-validation tests. The final NIR spectrum for each sample was the average spectra of all clean scans after the transformations were applied.

3.2.3 Partial least squares regression

PLSR linearly transforms the predictor variables into a smaller subset of independent components called latent variables. This allows the model to account for co-linearity in NIR spectral data, which can otherwise lead to unstable regressions without unique solutions (Agelet and Hurburgh, 2010). These latent variables are linear combinations of the original predictor variables (i.e. wavelengths). The latent variables are chosen to have both high variance and high correlation with the dependent variable (i.e. quality measure) (Hastie et al., 2013c). During calibration, the number of latent variables must be chosen. A small number of latent variables that still provides accurate predictions is desirable. PLSR is available in many of the most common chemometric software packages including Unscrambler (Nawi et al., 2013) and WinISI (O'Shea et al., 2011). The key advantages of PLSR are interpretability, ease of use and availability in standard software packages.

3.2.4 Support vector regression

Support vector regression (SVR; (Smola and Vapnik, 1997)) is an extension of the machine learning technique, support vector machines originally designed for classification problems (Cortes and Vapnik, 1995). Similar to PLSR, support vector regression (SVR) seeks a linear relationship between the predictor variables and the dependent variable. Rather than minimising the least squares error, SVR seeks to minimise a ‘cost function’ consisting of a weighted error term with specific constraints (Thissen et al., 2004a). Specifically, this cost function seeks to minimise prediction error (improved accuracy) as well as minimising coefficient size (improved generalisation). The cost weight (cost) and error parameter (epsilon) must be chosen and are usually optimised through cross-validation. The final SVR model uses only observations with an error greater than epsilon. These samples are called the support vectors (Thissen et al., 2004a). This ‘data sparsity’ helps improve generalisation (Tange et al., 2015) resulting in more robust models.

Another key advantage of SVR is the ability to model complex non-linear problems. This is achieved by transforming the original data using a kernel function such as the radial basis function (Tange et al., 2015). Non-linear relationships that exist between the original predictor variables and the dependent variable may become linear after an appropriate transformation, making the relationship easier to model. The parameters of the kernel function must also be optimised from the data. For example, the radial basis function has a single parameter (gamma) that must be tuned.

The major disadvantage of SVR is that the resulting models are difficult to interpret in terms of the original predictor variables. SVR is also largely absent from current commercial NIR analysis programs making it difficult to integrate into established NIR analysis systems. The authors recommend Tange et al. (2015) for a fuller description of SVR theory and its application to sugarcane mill products.

3.2.5 Model tuning and calibration

Both PLSR and SVR model parameters were tuned using a five-fold cross-validation of the calibration data set. Cross-validation estimates the predictive ability of the calibrated model by dividing the data into five ‘folds’. Each fold (~20% of calibration samples) was successively removed from the analysis and the models were built using the remaining four folds (~80% of calibration samples). Cross-validated root mean square error (RMSECV) was then calculated on the data that was left out across all folds.

Individual models were tuned for each of the three cane quality measures. The final model for each variable was selected as the combination of factors that minimised the RMSECV. Final models were then recalibrated using the whole calibration data set. PLSR models were built using the pls package in R (Mevik et al., 2015).The in-built scaling option was used to scale the spectral and reference data to avoid biasing towards wavelengths with higher absorbance. The number of latent variables was the only tuning parameter selected via cross-validation. For each model, up to 40 latent variables were tested during cross-validation. A lower number of latent variables is desired as it results in a less complex model.

SVR models were built using the e1071 package in R (Meyer et al., 2015). As with the PLSR models, an inbuilt scaling function was used to scale the data. Two parameters were tuned during cross-validation, the cost parameter and the gamma parameter for the radial basis function. The epsilon parameter was set to the default value of 0.1. A two stage tuning was performed. The first stage considered parameter values on a coarse exponential grid. Costs used were 2^{-3} to 2^{15} while gamma values used were 2^{-15} to 2^{3}.

3.2.6 Model evaluation

The final calibrated PLSR and SVR models for each quality parameter were applied to the full calibration data set and to the independent validation data set. Model performance statistics for the calibration data set were recorded as model root mean square error of calibration (RMSEC) and calibration coefficient of determination (R^2_c). Calibration results were used to show how well the model fit the data used to generate the model.

Model predictive performance was assessed based on the validation data set. Root mean square error of prediction (RMSEP) and prediction coefficient of determination (R^2_p) were recorded. Root mean square error is a measure of the standard deviation of the model errors and should be close to zero. Root mean square errors were calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{3-1}$$

where N is the number of samples, while y_i and $\hat{y}_i$ are the observed and predicted values of a particular quality measure for sample *i* respectively.

R^2 is a measure of the variance explained by the model such that a value close to 1 is desired. R^2 values were calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{3-2}$$

Here, N is the number of samples, y_i and $\hat{y}_i$ are the observed and predicted values for sample *i* respectively and $\bar{y}$ is the observed mean. Results from the independent validation data set were used to show how well the models perform on a new range of data. The validation set bias, Residual Prediction Deviation (RPD) and the slope of the regression line between predicted and observed data were recorded for completeness. The RPD is a ratio of the observed variance and model error variance (Agelet and Hurburgh, 2010). RPD values for sugarcane quality estimates are rarely provided in the literature. As such RPD was not been used as a primary measure of model performance.

3.3 Results and discussion

3.3.1 Model tuning and calibration

The final model tuning parameters are recorded in Table 3.2. Using SNV-First derivative treated spectra, PLSR models for Brix, Pol and Purity required 20, 22 and 22 latent variables, respectively. This resulted in many fewer variables compared with the original 700 wavelengths of the spectra. The tuning parameters for the SVR models were similar for Brix and Pol while parameters differed somewhat for Purity. The higher cost parameter for the Purity model suggests a more complex model. Simpler models are generally preferred as they are easier to interpret. In the context of on-line analysis, models are often used as a 'black-box' and interpretation of the model may be less important.

Table 3.2. Final model parameters for PLSR and SVR models of Bij, Pij and Apparent Purity. All models used standard normal variate- first derivative pre-treated spectra.

	PLSR	SVR	
	No. Latent Variables	**Cost**	**Gamma**
Bij	20	100	0.00012
Pij	22	200	0.00012
Purity	20	300	0.00020

3.3.2 Model evaluation

For each quality measure, the SVR models better represented the observed values in the calibration data set than the PLSR models (Table 3.3). RMSEC values were lower and R^2_c values higher for SVR models of all quality parameters. This can be expected in the calibration data, as the SVR models are more complex than the PLSR models. The largest difference was for models of purity, where SVR RMSEC was 33% lower than the PLSR RMSEC and 11% more of the observed variation was explained. Both PLSR and SVR models of Purity tended to overestimate low and underestimate high values of Purity (Figure 3.2). Figure 3.2 shows that the PLSR model (Figure 3.2(c)) tended to overestimate lower values of Purity more than the SVR model (Figure 3.2(f)) in the calibration set.

Table 3.3. Model calibration and validation statistics for PLSR and SVR models of Bij, Pij and Apparent Purity. Root Mean Square Error (RMSE) is a measure of the standard deviation of the model errors and should be close to zero. R^2 is a measure of variance explained by the model values close to 1 are desired. Validation Slope, Bias and Residual Prediction Deviation (RPD) were recorded for completeness.

Model		RMSEC	R^2_c	RMSEP	R^2_p	$Slope_p$[a]	$Bias_p$[b]	RPD_p[c]
PLSR	Bij (%)	0.28	0.98	0.30	0.96	0.97	-0.04	5.26
	Pij (%)	0.34	0.97	0.34	0.96	0.96	-0.06	4.98
	Apparent Purity	0.0132	0.79	0.0146	0.66	0.70	-0.0012	1.72
SVR	Bij (%)	0.25	0.98	0.29	0.96	0.96	-0.03	5.33
	Pij (%)	0.28	0.98	0.33	0.96	0.98	-0.06	5.15
	Apparent Purity	0.0088	0.90	0.0148	0.65	0.74	-0.0012	1.70

[a] Slope was calculated as the $\boldsymbol{\beta}$ coefficient of the linear least squares fit of $\hat{\boldsymbol{y}} = \beta \boldsymbol{y} + c$

[b] Bias was calculated as mean difference between predictions and observations $bias = \frac{\sum_{i=1}^{N}(\hat{y_i} - y_i)}{N}$

[c] RPD was calculated as the ratio of the standard deviation of the observations in the validation set and the RMSEP $RPD = \frac{sd(obs_p)}{RMSEP}$ $where\ sd(obs_p) = \sqrt{\frac{\sum_{i=1}^{Np}(\hat{y_i} - \bar{y})^2}{N_p - 1}}$

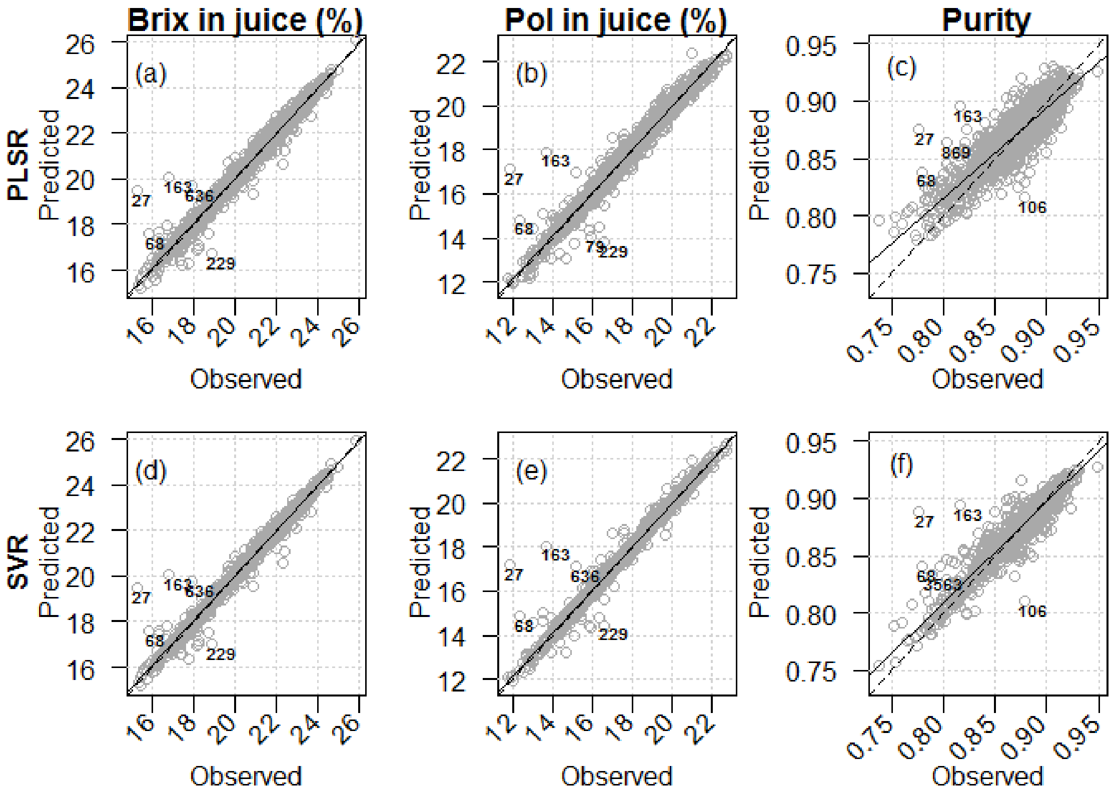

Figure 3.2. Predicted versus Observed values of Brix, Pol and Purity for calibration set data. Points represent individual samples for PLSR models (a), (b), (c) and SVR models (d), (e), (f). Solid line represents the relationship between predicted and observed data. Dashed line represents the 1:1 ratio. Numbers identify the five samples with the largest errors. Points above the dashed line were overestimated while points below the line were underestimated.

Validation performance statistics for PLSR (RMSEP and R^2_p) were comparable with previous results reported for the Australian sugar industry. Staunton et al. (2004) reported R^2_p of 0.91, 0.93 for Brix and Pol, respectively, using bias and temperature corrected PLSR models. The associated standard errors of prediction were 0.34 and 0.32 respectively. This is comparable with the RMSEP of 0.30 (Bij) and 0.34 (Pij) for PLSR achieved in our study (Table 3.3).

Although Staunton et al. (2004) used data from a different Australian mill (Maryborough, Queensland, Australia); these results are encouraging as both studies used a large number of samples from similar on-line cane analysis systems. In particular, the Staunton et al. (2004) study used a FOSS DL 5000 spectrophotometer with a spectral range from 1,100 to 2,500 nm at 2 nm intervals. The consistency of these results suggest that the PLSR method represents a good baseline against which to test new modelling methods. Berding et al. (1991) reported a standard error of prediction (SE_p) for Purity (%) of 1.66% with a correlation coefficient of 0.60 for a laboratory based NIR analysis. The results of Berding et al. (1991) are similar to the RMSEP and R^2_p obtained in our study. However, recent research has achieved more promising results (Berding and Marston, 2010).

SVR models of Brix and Pol slightly outperformed PLSR models when applied to the independent validation data set (Table 3.3). Models of Brix and Pol achieved the same R^2_p as the PLSR models however validation root mean square errors (RMSEP) were 3.3 % (Brix) and 2.9% (Pol) lower for the SVR models. The PLSR model for Purity slightly outperformed the SVR model with an RMSEP 1.37% lower than the SVR model. This was surprising given that the SVR Purity model outperformed PLSR when applied to the calibration data set (Table 3.3). The low R^2_p and relatively high RMSEP of both PLSR and SVR models of Purity suggest that neither is suitable for operational use at this point. Validation values for RPD, slope and bias were similar between PLSR and SVR models supporting the conclusions based on RMSEP and R^2_p.

As with the calibration data, both PLSR and SVMR models of Purity tended to overestimate lower values and underestimate higher values, particularly for Purity samples less than 0.8 (Figure 3.3 (c) and (f)). The overestimation of low Purity values is likely due to low values being under-represented during calibration and the poor ability of NIR models to extrapolate beyond the calibration range. When the calibration set is normally distributed (more samples in the centre of the data), the predictive performance of an NIR model will decrease for new samples further from the centre (Naes and Isaksson, 1989). Furthermore, both PLSR and SVR can perform poorly when extrapolating beyond the calibration range of values (Balabin and Smirnov, 2012).

In our study, both PLSR and SVR were unable to estimate values of Purity lower than the minimum value in the calibration data set (0.7368; Table 3.3) leading to overestimation of samples with lower values such as samples 3284 and 3285 (Figure 3.3(f)). Reducing the calibration data set to a uniform or rectangular distribution in the laboratory data may have improved overall model performance (Cao, 2013). This technique aims to select a uniform number of samples across the range of the quality measure so that low and high values are not under represented.

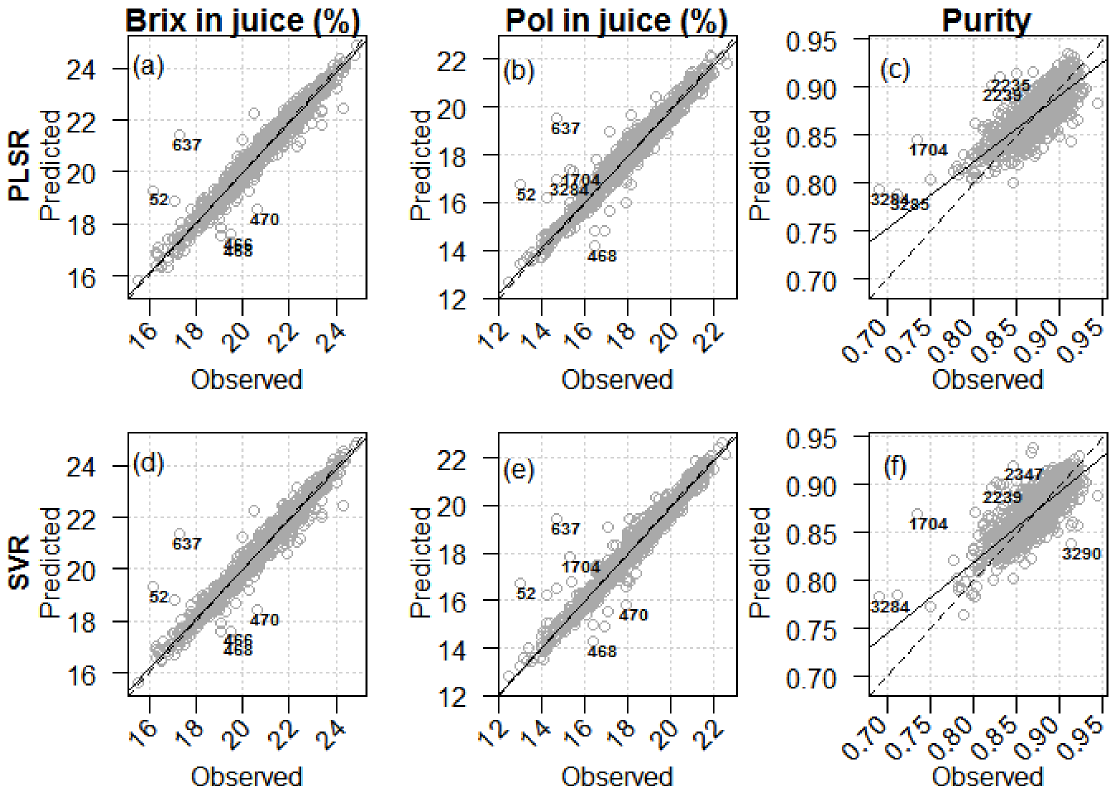

Figure 3.3. Predicted vs Observed values of Brix, Pol and Purity for validation set data. Points represent individual samples for PLSR models (a, b, c) and SVR models (d, e, f). Solid line represents the relationship between predicted and observed data. Dashed line represents the 1:1 ratio. Numbers identify the five samples with the largest errors. Points above the dashed line were overestimated while points below the line were underestimated.

By considering the five samples with the largest errors for each model, it can be seen that the same samples tended to be poorly estimated using either SVR or PLSR models (Figure 3.3). For example, samples 1704 and 2239 in Purity as well as 637 and 52 in Brix and Pol. Some samples may have been poorly estimated because they lay beyond the calibration range such as sample

1704 for Purity. Other samples were poorly estimated despite laying close to the centre of the data. For example, Purity was poorly estimated for sample 2239 while Brix and Pol were poorly estimated for sample 470. As these samples were not identified as outliers in the pre-processing stages, future research should consider trying to identify why these samples were difficult to predict.

The results of this study show the SVR modelling technique offers slight advantages over the traditional PLSR. However, results for Purity suggest that there is still room for improvement. Future research should consider comparisons of other modelling techniques on a more heterogeneous data set. For example, analysing data collected across multiple seasons, multiple NIR systems and multiple locations would allow researchers to better assess the ability of these modelling methods to cope with a greater modelling complexity.

The advantages of PLSR are the relative simplicity of the calculation, the availability in standard analysis software packages and a higher interpretability of the parameters used in the model. In comparison, SVR is difficult to interpret and is not currently available in most software packages used by the Australia sugarcane industry. Although SVR did provide modest improvement for models of Brix and Pol, the advantage in skill was not sufficient to recommend SVR in this study.

3.4 Conclusions

This study compared PLSR and SVR models for three cane quality measures. Results from the PLSR models were consistent with previous industry studies and justified the use of PLSR as a baseline modelling technique to which approaches that are more sophisticated can be compared. SVR models for percent brix and percent pol in juice slightly improved on PLSR models; however, PLSR and SVR models for purity were both considered unsuitable for use operationally. In this study, the slight improvement in model skill using SVR was not considered sufficient to recommend SVR over PLSR, given the relative ease of use and interpretability of PLSR. An important result of this study was that samples that were difficult to estimate with the PLSR models were also difficult to estimate using the SVR models. This suggests that in order to improve our ability to utilize NIR modelling techniques, we require a better understanding of why certain samples are difficult to model and how to identify them.

3.5 Chapter 3 Summary

Near infrared (NIR) analysis systems are used to estimate cane quality measures such as brix and pol in juice and apparent purity. Within the Australian sugarcane industry, partial least squares regression (PLSR) has been used to build NIR models of cane quality measures in the lab, on-line and in the field. PLSR relies on the linear relationship between sample constituents and electromagnetic absorption at NIR wavelengths. In practice, this linear relationship can often break down resulting in relationships that are more complex. Recently, machine learning techniques have become popular for their skill with complex data and ability to produce robust calibrations. The objective of this paper was to compare PLSR with the machine learning technique, support vector regression (SVR). The two techniques were used to estimate three cane quality parameters: brix in juice, pol in juice and apparent purity (Pij/Bij). Results from the PLSR models were consistent with previous industry studies and justified the use of PLSR as a baseline against which to compare approaches that are more sophisticated. The SVR models slightly reduced prediction error compared with PLSR models for brix and pol in juice, but slightly increased prediction error for apparent purity. The marginal improvement in model skill using SVR was not considered sufficient to recommend SVR over PLSR, given the relative ease of use and interpretability of PLSR. However, this study showed that certain samples were difficult to model with either approach.

The focus of Chapter 3 was Objective 1 of the thesis: Investigating the use of data mining and machine learning algorithms for improved NIRS estimates of cane quality. The outcomes of Chapter 3 contributed to the thesis objective in two important ways. Firstly, it was necessary to establish that it was possible to produce models of cane quality with similar skill to those presented in the literature. Comparisons to literature also established that the PLSR was an appropriate method to compare to more complex modelling approaches such as SVR. Secondly, the results of Chapter 3 showed that the comparison between PLSR and SVR was similar for each of the quality measures and that many of the same samples were difficult to estimate for both techniques. These results were important because they contributed to the scope of the Chapter 4 which concentrated on comparing a wider range of modelling techniques for a single quality measure.

Chapter 4

A comparison of non-linear regression methods for improved on-line near infrared spectroscopic analysis of a sugarcane quality measure

Relevant publication	Sexton, J., Everingham, Y., Donald, D., Staunton, S., & White, R. 2018. A comparison of non-linear regression methods for improved on-line near infrared spectroscopic analysis of a sugarcane quality measure. Journal of Near Infrared Spectroscopy, 26(5), 297–310. https://dx.doi.org/10.1177/0967033518802448
Statement of intellectual input	The research question of Chapter 4 was developed by the candidate with input from Dr. Everingham, Dr. Donald and Mr. Staunton. Data for the thesis was provided by SRA through Mr. Staunton. Editorial assistance was supplied by Dr. Everingham, Dr. Donald, Mr. Staunton and Dr. Ronald White. The candidate developed the methodological framework and ran all simulations. The candidate was also responsible for the write-up of the chapter and produced all tables and figures.
Publication status	Published

4.1 Introduction

The world sugarcane market is incredibly competitive. Sugarcane production for the top 5 producing countries ranged from 736 Million tonnes (Brazil) to 62 Million tonnes (Pakistan) in 2014 (FAO, 2017). Unsurprisingly then, cane quality is of paramount importance to sugarcane industries worldwide. In the Australian sugar industry, Commercial Cane Sugar (CCS) is the primary measure of cane quality and is used directly to calculate the payment made to growers. Since the first implementation of online Near Infra-Red Spectroscopy (NIRS) in the Australian sugarcane industry in 1996, millions of CCS measurements have been made using NIRS analysis. Given the importance of cane quality measures, the NIRS models must be both accurate and robust. Within the Australian sugarcane industry, Partial Least Squares Regression (PLSR) has been the primary chemometric algorithm used to build these NIRS models. The growing amount of NIRS data available to the sugarcane industry presents an opportunity to investigate recent advances in machine learning and non-linear data mining algorithms, to maximize the benefits of NIRS.

Within Australia, NIRS analysis was first introduced for the laboratory analysis of cane quality parameters in the late 1980s to early 1990s (Berding et al., 1991, Berding et al., 1989, Brotherton and Berding, 1995). In the mid 1990's at-line analysis was trialled (Berding and Brotherton, 1996, Brotherton and Berding, 1998) and by the late 1990s and early 2000s on-line analysis (Staunton et al., 1999, Staunton et al., 2004) had been introduced. On-line analysis allowed cane quality analysis to become part of the mill process. PLSR is often considered one of the most common, if not the most common method used for developing NIRS models (Agelet and Hurburgh, 2010). Within the Australian sugarcane industry PLSR has been used to build NIRS models of cane quality measures in the laboratory (O'Shea et al., 2011), on-line (Staunton et al., 1999) and in the field (Nawi et al., 2013).

PLSR for NIRS analysis relies on the assumption that the quality parameter being estimated and the predictor variables (NIR absorbance at particular wavelengths) is approximately linear (Miller, 1993, Tange et al., 2015). In practice, the linear relationship between quality measures and absorption at NIR wavelengths can often break down. Non-linearity can be introduced in two main forms: 1) changes to the measuring instrument; and 2) changes in the sample itself (Hageman et al., 2005). Machine wear, repair or replacement generally requires re-calibration of the model (Fearn, 2001). Differences in particle size in the sample, sample deterioration over time or high concentrations of the component being measured can also lead to non-linear effects (Bertran et al., 1999). Machine learning and non-linear algorithms such as support vector regression (SVR), artificial neural networks (ANN) or Gradient Boosted Trees (GBT) may be better suited to these types of complex situations (Balabin and Lomakina, 2011, Hageman et al., 2005).

SVR has been regarded as an effective alternative to ANN (Agelet and Hurburgh, 2010, Balabin and Lomakina, 2011, Kovalenko et al., 2006) and PLSR (Thissen et al., 2004a). The results from Chapter 3 showed that SVR was as effective as PLSR for Brix and Pol based quality measures in Australia. Similarly positive results have been shown in the Japanese sugarcane industry (Tange et al., 2015). Tange et al. (2015) found that SVR had an advantage over PLSR for the determination of sugar quality parameters such as Brix and Pol. Tange et al. (2015) found that a single global SVR calibration, using data from various stages of the milling process, produced a 36% reduction in RMSEP for estimates of Pol compared to a PLSR model built specifically for Molasses NIRS data.

ANN have shown promise at classifying crop quality rather than directly estimating quality parameter values. Nawi et al. (2013) found that hand-held NIRS machines calibrated using ANN could correctly classify sugar content of sugarcane from NIR spectra collected by scanning sugarcane skin with 75% accuracy. More recently, Jam and Chia (2017) showed that an ANN calibrated model was able to correctly classify total soluble solids in pineapples with 73% accuracy when whole pineapples were scanned.

One of the most often stated drawbacks of machine learning techniques is their ‘black box’ nature. However, recent research has shown it is possible to identify influential spectral wavelength regions within SVR models (Feilhauer et al., 2015, Üstün et al., 2007). Similarly, several methods have been explored to better understand and interpret ANN models used in ecological studies (Olden and Jackson, 2002, Olden et al., 2004, Özesmi and Özesmi, 1999). Unfortunately, these methods are not widely evident in NIRS analysis literature. For example, although Üstün et al. (2007) introduced a variable importance measure for SVR in 2007, several comparison studies published since, have not considered variable importance as part of model comparison (Balabin et al., 2010, Cui and Fearn, 2017, Ni et al., 2014, Pierna et al., 2011). If variable importance were explored, it may be possible to link models back to physical properties removing some of the ‘black box’ nature of these methods and help improve adoption by industry.

Data mining algorithms based on regression trees may provide an alternative, interpretable non-linear modelling approach. Tree based methods are structurally interpretable but often do not have the same predictive performance as other datamining techniques (Hastie et al., 2013e). Ensemble tree methods such as Random Forests (Breiman, 2001) and Gradient Boosted Trees (Friedman, 2001) (GBT) have been proposed as methods to improve tree based predictive performance. While we could find no examples of GBT used to assess agronomic quality parameters, GBT has successfully been used for remote sensing and mobile soil testing applications (Loggenberg et al., 2018, Nawar and Mouazen, 2017, Viscarra Rossel and Behrens, 2010).

Recently, Nawar and Mouazen (2017) showed that RF (RMSEP = 0.14) and GBT (RMSEP = 0.20) calibrated Vis/NIR models of soil total carbon, could perform as well or better than an ANN calibration (RMSEP = 0.20) for a mobile soil testing rig at a specific site (Hagg field), using local and regional data. However, results were site and data specific with GBT (RMSEP = 0.20) outperforming ANN (RMSEP = 0.27) when only local data were used. In agricultural industries, RF and extreme gradient boosted trees (XGBT) have recently shown promising results at detecting water stressed vineyards from remotely sensed hyperspectral data (Loggenberg et al., 2018). Loggenberg et al. (2018) showed that RF and XGBT were able to classify water stressed vineyards with up to 83.3% and 80.0% accuracy respectively where spectral data from 473 nm – 708 nm were used.

Given the importance of robust, accurate NIRS analysis of cane quality for the Australian sugarcane industry, we investigate if the additional complexity of machine learning and non-linear data mining algorithms deliver a substantial advantage over a simpler, traditional modelling approach. As many approaches to modelling non-linearity exist, it is also important to test a variety of modelling techniques. Therefore, the objective of this chapter was to compare chemometric models of CCS using four calibration techniques that approach non-linearity in different ways: partial least squares, support vector regression, artificial neural networks and gradient boosted trees. Furthermore, we endeavour to explore how spectral information is used within these different modelling techniques.

4.2 Theory

4.2.1 Partial least squares regression

The approach of partial least squares (PLS) was first developed around 1975 by Herman Wold as a form of two block regression (Wold et al., 2001) and has become one of the most widely used regression techniques in chemometrics. Many variants have since been developed including orthogonalized PLS (Trygg and Wold, 2002) and interval PLS (i-PLS) (Nørgaard et al., 2000) among many others.

Here we describe the basics of the PLS regression (PLSR) algorithm for a single response variable (***y***). The standard form of a multivariate regression in terms of NIR spectroscopic analysis can be expressed as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{f}, \tag{4-1}$$

where ***y*** is a vector (***y*** = [y_1, …, y_N]) of N reference values to be predicted, ***X*** is *N*-by-*P* matrix (***X*** = [$\boldsymbol{x}_1$ …, $\boldsymbol{x}_p$]) of NIR spectroscopic data, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\boldsymbol{f}$ is a vector of model errors. PLSR linearly transforms the predictor variables into a smaller set of independent components called latent variables (***LV*** = [$\boldsymbol{lv}_1$ …, $\boldsymbol{lv}_A$]). This allows the model to account for co-linearity which can lead to unstable regressions without unique solutions (Agelet and Hurburgh, 2010). The latent variables are defined as orthogonal to each other and are related to the original data (***X***) by the *N*-by-*A* weighting matrix ***W*** and loading matrix ***P*** as

$$\boldsymbol{LV} = \boldsymbol{XW}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W})^{-\boldsymbol{1}}. \tag{4-2}$$

The latent variable in PCA are developed iteratively so that the first LV has both high variance and high correlation with the reference values (***y***) (Hastie et al., 2013f). The final form of the PLSR equation can be expressed as

$$\boldsymbol{y} = \boldsymbol{LVq} + \boldsymbol{f}, \tag{4-3}$$

where ***q*** is a vector of regression coefficients in the latent space.

4.2.2 Support vector regression

The support vector machine was first developed as a non-linear approach to binary classification problems (Cortes and Vapnik, 1995) and was later extended to regression problems through two separate approaches; SVR (Smola and Vapnik, 1997) and least squares support vector machines (Cui and Fearn, 2017, Suykens et al., 2002, Thissen et al., 2004b). Similar to PLSR, both support vector regression approaches seek a linear relationship between the predictor variables (***X***) and the dependent variable (***y***). Rather than minimizing the mean square error, SVR seeks to minimize the ε-insensitive loss function (Smola and Vapnik, 1997). This cost function attempts to minimize coefficient size and prediction error(Thissen et al., 2004a). Furthermore, prediction errors are penalized linearly except for absolute errors less than a specified cut-off tolerance of ε and are weighted through a cost parameter (*C*), which can be tuned as a trade-off between model accuracy and simplicity.

The final form of the SVR function can be described as

$$\hat{\boldsymbol{y}} = \sum_{i,j}^{N}(\alpha_i - \alpha_i^*)\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) + b. \tag{4-4}$$

Here $\hat{\boldsymbol{y}}$ represents the vector of model predictions; *N* is the total number of samples and $(\alpha_i - \alpha_i^*)$ represent Lagrangian multipliers that result from the numeric optimization of the cost function. Only samples with errors beyond ±ε (the support vectors) will have a nonzero α value. Samples with an error beyond +ε will have a nonzero α_i while samples with an error beyond -ε will have a nonzero α_i^*. As α differ in size, some samples can be considered more important than others (Thissen et al., 2004a). $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a kernel function used to map the dependent variable matrix to a higher dimensional feature space (Üstün et al., 2007) and *b* is an offset term in the model. A suitable kernel function can allow the SVR to model non-linearities in the data and several may need to be tested for any given scenario. In this study, we considered only the Gaussian radial basis function (Vert et al., 2004)

$$\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\gamma||\boldsymbol{x}_i - \boldsymbol{x}||^2), \tag{4-5}$$

which is often used in SVM literature (Tange et al., 2015, Thissen et al., 2004a, Vert et al., 2004). The radial basis function γ parameter, as well as the cost parameter *C* and ε parameters are usually tuned through cross-validation during model calibration.

4.2.3 Artificial neural networks

In this study, we consider the feed forward single hidden layer network, which forms a two-stage regression model. The network consists of an input layer representing the original dependent variables (***X***), a hidden layer of *M* nodes representing a transformed set of predictor variables $\boldsymbol{h}_1$ to $\boldsymbol{h}_m$ and an output layer representing the predictions ***ŷ*** (Figure 4.1). Each hidden node is a transformation of a linear combination of the weighted original input variables as

$$\boldsymbol{h}_m = f\left(a_{0m} + \boldsymbol{a}_{\mathbf{m}}^{\mathbf{T}}\boldsymbol{X}\right);\ m = 1, \dots, M; \boldsymbol{a}_{\mathbf{m}} = [a_{1m}, \dots, a_{Pm}]. \quad (4\text{-}6)$$

Here, $\boldsymbol{a}_m$ is the vector of weights representing the connections of all *P* inputs to node *m* and a_{0m} represents the bias adjustment of node *m*. The transformation $f(v)$ is called the activation function and is often a sigmoidal function of the form $f(v) = 1/(1e^{-v})$ (Hastie et al., 2013d). The output is then computed as a transformed linear combination of the hidden nodes using

$$\mathbf{o} = g\left(b_0 + \boldsymbol{b}^{\mathbf{T}}\boldsymbol{H}\right);\ \boldsymbol{b} = [b_1, \dots, b_M];\ \boldsymbol{H} = [\boldsymbol{h}_1, \dots, \boldsymbol{h}_M]. \quad (4\text{-}7)$$

Here, ***b*** is a vector of weights for the connections between the *M* nodes of the hidden layer and the output layer while b_0 is a bias term.

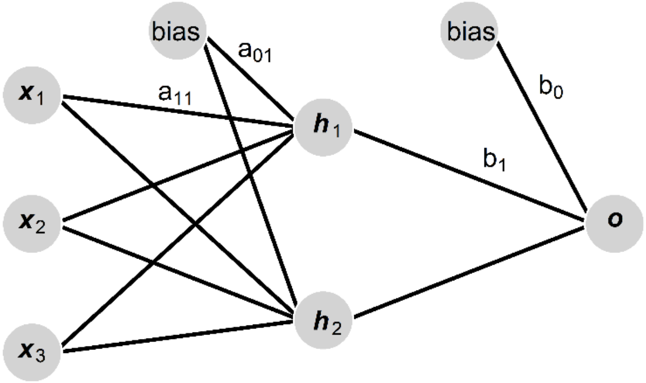

Figure 4.1. A simple neural network with weights. *X* nodes represent the input variables (e.g. NIR data), H nodes represent nodes in the single hidden layer and the O nodes represent the output (ŷ). Bias nodes are identity vectors such that bias weights are simply constants added to the transformations.

In this study the final transformation function $g(v)$ was linear so that $g(b_0 + \mathbf{b}^{\mathbf{T}}\mathbf{H}) = b_0 + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{H}$. However, the softmax transformation is also regularly used (Hastie et al., 2013d). All connection weights (*a*'s and *b*'s) and all bias terms (***a***₀'s and ***b***₀'s) are learned from the data during calibration by minimizing least squares error through gradient decent. The large number of weights mean that the ANN models can be extremely flexible but also run the risk of overfitting. This can be mitigated with a decay parameter (Hastie et al., 2013d).

Networks can be made more complex by adding more nodes or more hidden layers. However, the user must then define the number of layers, the number of nodes in each layer and the decay rate. ANN are highly parallelizable and have a high tolerance for noisy data making them ideal for large data mining jobs (Han et al., 2011).

4.2.4 Gradient boosted trees

Boosting encompasses a range of modelling procedures that attempt to combine the output of several 'weak' models into a single powerful model. In general, a weak model (base learner) is sequentially fitted to slightly modified versions of the data. For boosted tree models, the base learner is a (usually small) regression tree model. In gradient boosted models, each tree is fitted to the residuals of the previous model by regression. In stochastic gradient boosting, the tree model fitted at each iteration is developed on a subset of the training data. Similarly, to the ANN approach a regularization weight (shrinkage rate) is applied to the learning process to help prevent overfitting. This shrinkage rate along with the number of iterations (number of trees built) and the number of terminal nodes of the tree (depth) must be provided. A slower shrinkage rate will require a larger number of iterations (Hastie et al., 2013a).

Here we consider stochastic gradient boosted tree models described by (Friedman, 2001, Friedman, 2002) as implemented by Ridgeway (2015) in the following steps (Hastie et al., 2013a, Ridgeway, 2015):

1. Select a loss function $L(\boldsymbol{y}, f(\boldsymbol{x}))$ (e.g. squared error loss for least-squares regression), the number of trees *J*, the depth of each tree *K* the shrinkage rate λ and subsampling rate *p*.
2. Initialize the predicted values as a constant, $f_0(\boldsymbol{x}) = \arg min_\rho \sum_{i=1}^{N} L(y_i, \rho)$
3. For iteration *j* in 1, …, *J*:
    a. For each observation i in 1, …, N, compute the negative gradient as the temporary response

$$r_{ij} = -\frac{\partial}{\partial f(x_i)} L(y_i, f(x_i))\Big|_{f(x_i)=f_{j-1}(x_i)}. \quad (4\text{-}8)$$

b. Randomly select $p \times N$ cases from the dataset without replacement.

c. Fit a regression tree with *K* terminal nodes to the r_{ij} of the selected cases.

d. Compute the terminal node predictions ρ_k for *k = 1 …,* K_j *as:*

$$\rho_{kj} = \arg\min_{\rho} \sum_{x_i \in S_{kj}} L(y_i, f_{m-1}(x_i) + \rho) \quad (4\text{-}9)$$

where S_{kj} is the subset of $\boldsymbol{x}$ that define terminal node *k*.

e. Update predicted values as

$$f_j(\boldsymbol{x}) = f_{j-1}(\boldsymbol{x}) + \boldsymbol{\lambda} \cdot \sum_{k=1}^{Kj} \rho_{kj}\ (x \in S_{kj}). \quad (4\text{-}10)$$

4. Output final prediction: $\hat{\boldsymbol{y}} = f_J(x)$.

4.2.5 Variable importance measures

PLSR models are considered relatively easy to interpret. A number of measures can identify how input variables contribute to the model. Mehmood et al. (2012) provide an overview of these measures. One simple measure of variable importance within PLSR models are the PLS estimates of the regression coefficients ($\hat{\boldsymbol{\beta}}$) that describe the linear regression $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. By substituting equation (4-2) into equation (4-3) and equating equation (4-1) to equation (4-3), it can be seen that the $\hat{\boldsymbol{\beta}}$ coefficients can be estimated as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{W}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W})\boldsymbol{q}. \quad (4\text{-}11)$$

As linear regression coefficients, $\hat{\boldsymbol{\beta}}$ values give an indication of magnitude and direction of the influence each variable has within the model.

Compared to PLSR, there are few methods of interpreting variable importance or influence within SVR models (Ben Ishak, 2016, Üstün et al., 2007). Üstün et al. (2007) proposed a variable importance measure based on the inner product of the values and the original $\boldsymbol{X}$ matrix. The α values are comparable to the PLS regression coefficients (Jam and Chia, 2017). Only samples with an α value greater than zero are included in the model (the support vectors). The 'P profile' of variable importance can be calculated as:

$$\mathbf{p_r} = \boldsymbol{X}_{\mathbf{sv}}{}' \cdot \boldsymbol{\alpha}_{\mathbf{sv}}, \quad (4\text{-}12)$$

where $\boldsymbol{X}_{sv}$ is an *M*-by-*V* matrix of *V* support vectors for *M* original variables and $\boldsymbol{\alpha}_{sv}$ is a vector of *V* α values for the support vectors. The $\mathbf{p_r}$ values give a magnitude and direction similar to regression coefficients.

Assessing variable importance within neural networks has been explored within ecological studies. Olden et al. (2004) compare nine methods used to assess neural network variable importance in ecological studies. Olden et al. (2004) found that the Olden method of connection weights (Olden and Jackson, 2002) consistently outperformed other methods at identifying true variable importance in Monte Carlo simulations where true variable importance was known. Olden and Jackson (2002) proposed that variable importance within an ANN model could be assessed as the sum of the product of raw connection weights between input nodes, all hidden layer nodes and output node. For example, in the simple network described in Figure 4.2, the Olden index for input variable one would be calculated as $Olden_1 = (a_{11} \times b_1) + (a_{12} \times b_2)$.

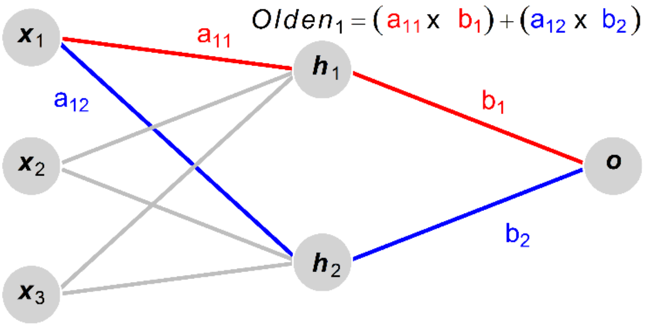

Figure 4.2. A simple neural network with weights. *X* nodes represent the input variables (e.g. NIR data), H nodes represent nodes in the single hidden layer and the O nodes represent the output ($\hat{y}$). The olden connection weight index value for I1 would be calculated as $\mathbf{Olden_1} = (\mathbf{a_{11}} \times \mathbf{b_1}) + (\mathbf{a_{12}} \times \mathbf{b_2})$.

As with the PLSR and SVR indices, the Olden connection weight values give both a magnitude and direction. This was considered an advantage over the similar 'Garson' index (Garson, 1991) which used absolute weights, which could lead to misrepresentation of variable importance when the sign of weights changes between connections (Olden and Jackson, 2002).

Variable importance within a single tree based regression model can be assessed using the approximate relative influence measured as the empirical improvement in squared error ($\hat{I}_i^2\ (T)$) over all splits occurring on that variable (Friedman, 2001). Friedman (2001) extended this to boosted trees, suggesting that for a collection of trees $[T_j]_1^J$ the approximate relative importance of a variable can be calculated as the average across all trees

$$\hat{I}_i^2 = \frac{1}{J}\sum_{j=1}^{J} \hat{I}_i^2\ (T_j). \qquad (4\text{-}13)$$

Unlike the variable importance measures reported for PLSR, SVR and ANN, this variable influence is an actual indication of model improvement and therefore does not give a direction of influence.

4.3 Materials and methods

Four regression models (PLSR, SVR, ANN and GBT) of CCS were built from NIR spectroscopic data collected from the on-line cane analysis system for the 2006 harvest season. Data were randomly divided into a calibration (~50%) and validation (~50%) set such that no samples from the same farm appeared in both sets. All models were calibrated using 5-fold cross-validation with the aim of minimizing root mean square error (RMSECV) while producing simple models. Models were then applied to an independent data set and performance was compared based on predictive root mean square error (RMSEP) and R^2 (R_p^2).

RMSE values were calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y_i}-y_i)^2}{N}}, \qquad (4\text{-}14)$$

while R^2 values were calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y_i}-y_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}. \qquad (4\text{-}15)$$

Here, $\hat{y_i}$ are predicted values and y_i are laboratory observed values and $\bar{y}$ is the mean observed value of all *N* observations.

Model errors were then investigated graphically to assess model performance throughout the range of CCS values. Finally, wavelength importance within each model was investigated. Figure 4.3 outlines the methodological process used in this analysis.

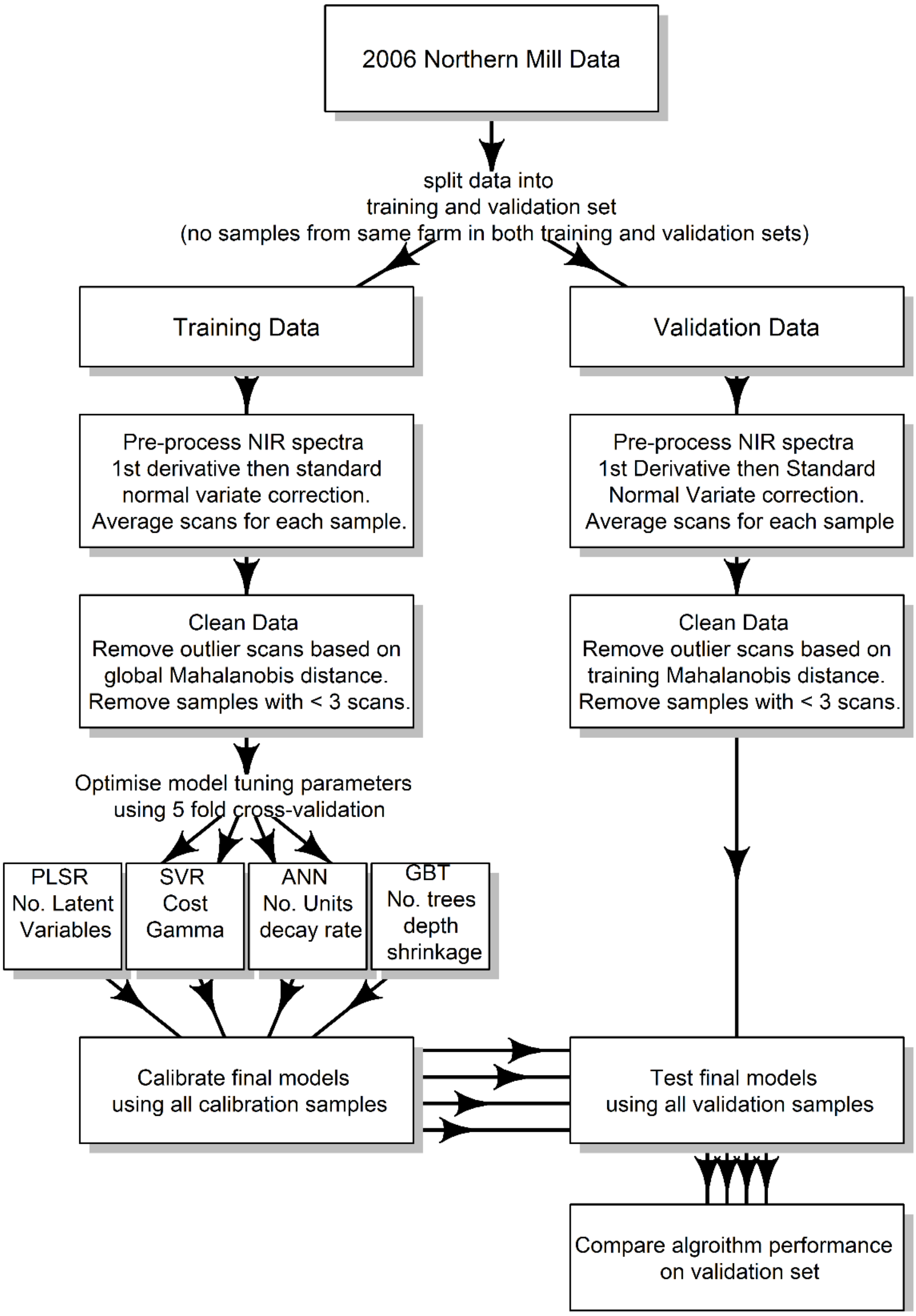
2006 Northern Mill Data
split data into
training and validation set
(no samples from same farm in both training and validation sets)
Training Data
Validation Data
Pre-process NIR spectra
1st derivative then standard
normal variate correction.
Average scans for each sample.
Pre-process NIR spectra
1st Derivative then Standard
Normal Variate correction.
Average scans for each sample
Clean Data
Remove outlier scans based on
global Mahalanobis distance.
Remove samples with < 3 scans.
Clean Data
Remove outlier scans based on
training Mahalanobis distance.
Remove samples with < 3 scans.
Optimise model tuning parameters
using 5 fold cross-validation
PLSR
No. Latent
Variables
SVR
Cost
Gamma
ANN
No. Units
decay rate
GBT
No. trees
depth
shrinkage
Calibrate final models
using all calibration samples
Test final models
using all validation samples
Compare algroithm performance
on validation set

Figure 4.3. A flow diagram of methodology.

All analysis were performed within the R statistical language and computing environment (R Core Team, 2017). The pls package (Mevik et al., 2015) was used to generate the PLSR model. The 'e1071' (Meyer et al., 2015) and nnet (Venables and Ripley, 2002) packages were used to build SVR and ANN models respectively, while the gbm package (Ridgeway, 2015) was used to build the GBT model. Spectral pre-processing algorithms were applied using the prospectr package (Stevens and Ramirez-Lopez, 2013). Variable importance measures for PLSR and ANN were calculated using the plsVarSel (Mehmood et al., 2012) and NeuralNetTools (Beck, 2016) packages respectively.

4.3.1 Data and pre-processing

Sugarcane samples and associated NIR spectra were sourced from an on-line cane analysis system from a sugarcane mill located in Northern Queensland. Cane samples represent consignments sent to the mill for processing during the 2006 crush season. To maintain grower and farm confidentiality, samples and farms were de-identified as unique ID numbers rather than grower personal details.

In total for the 2006 season, 3,794 consignments (samples) were collected that had consignment data, laboratory measured CCS values and NIRS data. The average consignment size for the 2006 season was 22.8 tonnes of cane. A FOSS 5000 on-line NIRS system, collected spectral data on shredded cane as absorbance (log(1/reflectance)). Absorbance was recorded from 1,100 nm to 2,498 nm at 2 nm intervals. Given the size of each consignment, as many as 30 scans were collected as a consignment was processed.

Samples were divided into a training set (1,899 samples) and a test set (1,895 samples) for the analysis. Samples were randomly divided so that no samples from the same farm appeared in both the training and test set. This gave some independence to the test set.

Table 4.1. Description of CCS reference data. One sample was removed from the test set during removal of spectral outlier (1894 samples rather than 1895).

Data set	Number of samples	Mean	Median	Std. Dev.	Range
Training	1,899	13.1%	13.4%	1.68%	7.6% – 16.9%
Test	1,894	13.4%	13.5%	1.33%	8.5% – 17%
Total	3,793	13.2%	13.5%	1.52%	7.6% – 17%

CCS is a measure of the pure sucrose that is obtainable from the cane and is based on the effect impurities in cane have on the mill process (Mackintosh, 2000). CCS is calculated from pol in juice (Pij), brix in juice (Bij) and fibre measures as

$$CCS\ (\%) = \frac{3 \times Pij(95-\%Fibre) - Bij(97-\%Fibre)}{200}\ . \qquad (4\text{-}16)$$

CCS values used in this study were the recorded laboratory values using industry defined methods. Laboratory Pij and Bij values used in the CCS calculation were derived from samples of first expressed juice. Fibre values used in CCS calculation are derived from a fibre class calculation. The range of CCS values represented by the training set included CCS values below that observed in the test set (Table 4.1). This gave the training set a slight skew towards lower values, but ensured that the range of CCS values in the test set were represented in the training set.

In order to match the multiple spectra recorded to the single CCS measure of a sample, a sequence of spectral transformation, outlier removal and spectral averaging was applied to the raw spectral data. All recorded spectra were first transformed using a Savitzky-Golay first derivative (Savitzky and Golay, 1964) with a 17 point window, using a second degree polynomial. This reduced the spectral range from 1,100 nm – 2,498 nm to 1,116 nm – 2,484 nm as the leading and tailing eight wavelengths were removed rather than extrapolated. The standard normal variate of the first derivative spectra was then taken. The pre-processing method used here was chosen as it was previously found to work well for a range of cane quality measures (Chapter 3).

Spectral outliers were then identified using a global Mahalanobis distance based on the transformed training data set. Following the methodology of Chapter 3, scans with a global Mahalanobis distance greater than three were removed from the analysis. For the test data set, global Mahalanobis distances were calculated with respect to the training data set. The final spectra used in the analysis was the average of all remaining scans for each sample. Following the methodology of Chapter 3, samples with less than three scans were removed from the analysis. This resulted in the removal of a single sample from the analysis. Figure 4.4 shows raw and pre-processed mean and standard deviation of all spectra used in the analysis.

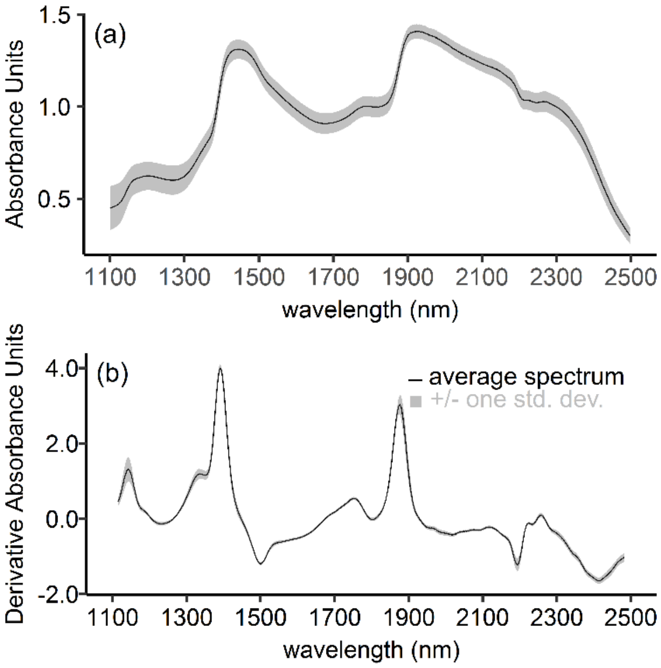

Figure 4.4. Typical Sugarcane spectrum used in the analysis as (a) Raw spectra and (b) First Derivative-SNV spectra. Solid lines represent average spectrum while grey shaded area represents +/- one standard deviation from the mean.

4.3.2 Model calibration

Each calibration technique required the tuning of a number of parameters. Model parameters were tuned using a five-fold cross-validation on the training data set (Hastie et al., 2013b). Table 4.2 records the parameters tuned for each model and the range of values tested. Preliminary testing of parameter values was used to select adequate ranges for each parameter. PLSR has only one tuneable parameter, the number of latent variables so this was the only parameter tuned. The kernel pls method within the pls package was used to develop the PLSR models. Data were auto-scaled using the inbuilt scaling function within the pls package. The SVR models built in this study used Vapnik's ε-insensitive loss function (Cui and Fearn, 2017, Suykens et al., 2002, Thissen et al., 2004b) and a radial basis function. For these models cost (C) and the radial basis function γ parameters were tuned. Following the methodology of Chapter 3, the ε parameter was set to 0.1 for all SVR models. This was considered adequate given the accuracy of the laboratory measurements. As with the PLSR models, data were auto-scaled using the inbuilt scaling function within the e1071 package.

Table 4.2. Parameters tuned through cross-validation for each model and the values tested.

Model	Parameter	Values
PLSR	No. Latent Variables	2, 3, 4, …, 39, 40
SVR	Cost (*C*)	2^{-5}, 2^{-3},2^{-1}, 2, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11},2^{13}, 2^{15}
	Gamma (γ)	2^{-15}, 2^{-13},2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2, 2^{3}, 2^{5}
ANN	No. Hidden Nodes	2, 3, 4, 5, 10, 15, 20
	Decay Rate	0.001, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1
GBT	No. trees	100, 200, 300, …, 9900, 10000
	Tree depth	4, 5, 6, 7, 8, 9
	Shrinkage	0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.90

ANN models built were single hidden layer feed forward neural networks. The number of hidden nodes within the hidden layer and the decay rate were the only parameters tuned. Initial parameter weights were randomly set. Data were scaled to within the range 0-1 to improve model performance in the ANN model. For GBT models, number of trees, tree depth, shrinkage rate and minimum number of observations were tuned. Data were assumed to follow a Gaussian distribution within the gbm package. Regression trees are insensitive to parameter size and no scaling was applied to data for building GBT models.

In each case, the RMSECV was recorded as the mean of the RMSE values for each of the five folds of the training dataset. The final combination of parameters was selected as the simplest model with a RMSECV within one standard error of the absolute minimum RMSECV (Hastie et al., 2013b, Kuhn, 2017, James et al., 2013). The standard error of the RMSECV was calculated from the five RMSE values calculated for each parameter combination. PLSR models were simplified by minimizing the number of latent variables. SVR models were considered simpler if the *C* and γ values were smaller, while ANN models with a smaller number of hidden nodes and a smaller shrinkage rate were considered simpler. For GBT models, a smaller number of trees and a smaller tree depth was considered a simpler model.

4.3.3 Model validation

The final models used the parameter combinations identified by the cross-validation process, re-calibrated on the entire training set. The calibration root mean square error (RMSEC) and R^2_c were recorded. The final models were then applied to the independent test set and predictive statistics were recorded (RMSEP and R^2_p). The test set bias, Residual Prediction Deviation (RPD) and the slope of the regression line between predicted and observed data were recorded as part of the model comparison (section 4.4.1 Model comparison). The relationship between the predicted and laboratory observed CCS values was then explored graphically and the samples with the largest errors were noted as part of a deeper error investigation (section 4.4.2 Error investigation).

4.3.4 Variable importance

Variable (wavelength) importance measures for each final model were calculated as described in the theory section. PLSR coefficient values were used to indicate variable importance within the model. Following Feilhauer et al. (2015), the SVR coefficient based index described by Üstün et al. (2007) was calculated as the dot product of the coefficients and support vectors recorded for the final model by the svm package. The Olden index as calculated by the NeuralNetTools package was used to describe variable importance within the final ANN model, while the relative influence returned by the gbm package in R was used to describe variable importance within the final GBT model. As we were only interested in the relative importance within the model, the absolute values for PLSR, ANN and SVR model indices were used. The regions of the NIR spectrum identified as important within each model were then compared (section 4.4.3 Variable Importance).

4.4 Results and discussion

4.4.1 Model comparison

Validation performance statistics for PLSR (Table 4.3) were comparable with previous results reported for the Australian sugar industry. Staunton et al. (2004) reported $\mathrm{R}^2{}_{\mathrm{p}}$ of 0.87 using a bias and temperature corrected model. The associated Standard Error of Prediction (SEP) was 0.38. The SEP reported by Staunton et al. (2004) can be compared to the RMSEP of 0.37% obtained in this study, given the large number of samples (>1,000) used in validating the models in each study. Staunton et al. (1999) also reported similar SEP and $R^2{}_p$ values for CCS. Using samples from across five mills and three seasons, CCS SEP was 0.33% and $\mathrm{R}^2{}_p$ was reported as 0.957 (Staunton et al., 1999). Although Staunton et al. (1999) used data from a number of Australian sugar mills, the similarity to results presented here is encouraging as both studies used a large number of samples from similar on-line cane analysis systems. The similarity between earlier studies and the results presented here suggest that the PLSR method represents a good baseline against which to test new modelling methods.

Table 4.3. Calibration and predictive statistics of the four models. Validation Slope, Bias and Residual Prediction Deviation (RPD) were recorded for completeness.

Method	Parameter values	RMSEC	$R^2{}_c$	RMSEP	$R^2{}_p$	$Slope_p$[a]	$Bias_p$[b]	RPD_p[c]
PLSR	LV = 18	0.36%	0.96	0.37%	0.92	0.93	-0.04%	3.57
SVR	*C* = 32 $\gamma = 2^{-11}$	0.25%	0.98	0.37%	0.92	0.95	-0.07%	3.60
ANN	Nodes = 2 Decay = 0.02	0.31%	0.97	0.36%	0.93	0.93	-0.05%	3.70
GBT	Trees = 1100 Depth = 8 Shrinkage = 0.05	0.06%	0.99	0.51%	0.85	0.86	-0.05%	2.60

[a] Slope was calculated as the $\boldsymbol{\beta}$ coefficient of the linear least squares fit of $\widehat{\boldsymbol{y}} = \beta \boldsymbol{y} + c$

[b] Bias was calculated as mean difference between predictions and observations $bias = \frac{\sum_{i=1}^{N}(\widehat{y_i} - y_i)}{N}$

[c] RPD was calculated as the ratio of the standard deviation of the observations in the validation set and the RMSEP $RPD = \frac{sd(obs_p)}{RMSEP}$ $where\ sd(obs_p) = \sqrt{\frac{\sum_{i=1}^{N_p}(\widehat{y_i} - \bar{y})\text{^}2}{N_p - 1}}$

The SVR model performed as well as the PLSR model while the ANN model had a slightly lower RMSEP (0.36%) and slightly higher $R^2{}_p$ (0.93). However, the validation bias in SVR was slightly larger in magnitude than for the PLSR model. It is also important to note that the difference between RMSEC and RMSEP was large for the final SVR and GBT models compared to the PLSR and ANN models (Table 4.3). This can indicate that the model was over-fitted to the training set. The final SVR model required 879 support vectors (46% of the training set) while the GBT used a large number of iterations (1,100 in the final model). Further tuning could be applied to reduce the dependence of these models on the training set.

A similar trend was evident in for Bij, Pij and apparent purity in Chapter 3. However, as few comparison studies report calibration statistics, it is difficult to determine from the available literature, whether this is a common feature of SVR spectroscopic models in general. The differences in RMSEP were not likely to be significantly different. This suggests that SVR or ANN could replace PLSR without loss of performance but may not provide a strong improvement in performance.

The final GBT model had a noticeably higher RMSEP and lower R^2_p than any other model (Table 3). This suggests that the GBT model may not be sufficient for the estimation of CCS. The GBT model also had the lowest regression slope (0.86) between the predicted and laboratory observed CCS values and may struggle to estimate extreme low or high CCS values (Williams et al., 2017).

Comparison of RPD, slope and bias between models agreed with results based on RMSEP and R^2_p values. Generally there was little difference between ANN, SVR and PLSR models, with a notably lower RPD score for the GBT model. Following Araújo et al. (2014), RPD is used only in comparison rather than as a primary score of model skill.

Araújo et al. (2014) reported similar RMSEP values for GBT and SVR models of soil organic matter based on Vis-NIR spectroscopic data. GBT and SVR models of organic matter and clay outperformed PLSR models in that study. In a similar study, GBT models for organic carbon, clay and pH failed to improve on PLSR models (Viscarra Rossel and Behrens, 2010). Although the results from Araújo et al. (2014) and Viscarra Rossel and Behrens (2010) are not directly comparable to the results presented here, they do indicate that the use of GBT in regression problems is application specific. Based on comparison with PLSR, SVR and ANN model performance in our study, it would appear GBT is not a viable option for estimating sugarcane quality parameters.

The performance improvement for SVR and ANN models often reported in literature was not observed in this study. For example, recent research in the Japanese sugarcane industry concluded that SVR models did provide an improvement in RMSEP for estimates of two cane quality measures (Tange et al., 2015). Tange et al. (2015) used NIR data of several mill products to produce global calibrations for Brix and sucrose in sugarcane. The greater variability inherent in such as global model may have contributed to the greater difference between SVR and PLSR models.

Balabin and Lomakina (2011), compared SVR, LS-SVM, ANN and PLSR for NIR spectroscopic analysis of quality measures for diesel fuels and concluded that ANN and SVM based techniques outperformed PLS based techniques. More importantly, Balabin and Lomakina (2011) concluded that the advantage of SVM based techniques was more apparent when there was a greater non-linearity within the data. Hageman et al. (2005) reported a similar effect, when comparing calibration techniques for robustness to temperature effects. In our study the relationship between CCS and NIRS data seems to be relatively linear, resulting in little difference in performance between PLSR, SVR and ANN.

The current study was limited to data from one season, machine and geographic region. The ability of SVR and ANN to PLSR in this study, and the reported advantages of machine learning techniques in more complex situations, provides strong evidence that future research within the Australian sugar industry should consider a comparison of models for more global calibrations which include season and/or region variability. Including more geographical and temporal information into the calibration data would improve model robustness for applications to different regions and seasons.

4.4.2 Error investigation

The relationship between predicted and laboratory CCS values was plotted for each model in order to investigate errors visually (Figure 4.5). The spread in the plotted data about the linear regression line (solid black), reflects the overall model performance. A much higher spread is visible for the GBT predictions. Despite all models having a low negative bias overall (Table 4.3), there is some evidence that the models tend to overestimate lower CCS values and underestimate higher CCS values. This can be seen as the linear regression (solid black line) sits above the 1:1 line (dashed line) and is most visible for the GBT model (Figure 4.5(d)) and least visible for the SVR model (Figure 4.5(b)). This suggests that the SVR model was better able to a represent accurately the extreme CCS samples. Chapter 3 showed a similar result for apparent purity. In that study, an SVR model overestimated (underestimated) apparent purity for low (high) values less than a PLSR model. Similar results have also been seen in a comparison of PLSR and SVR applied to the NIR spectroscopic analysis of agricultural seeds for Nitrogen content (Cui and Fearn, 2017).

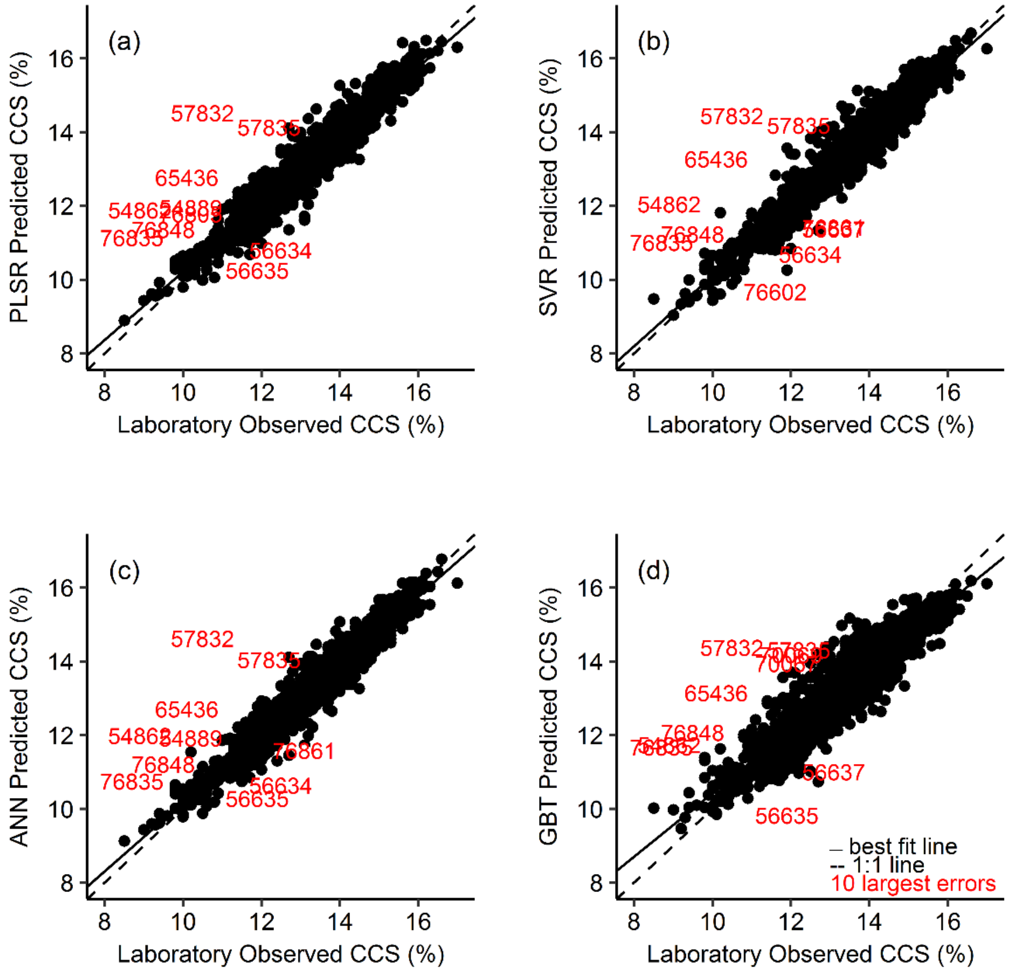

Figure 4.5. Predicted VS Observed CCS values for validation data set using PLSR (a), SVR (b), ANN (c) and GBT (d). Dashed lines represent the 1:1 ratio while solid lines represent the line of best fit. Numbers (red) represent the samples with the largest errors for each model.

Identifying the 10 samples with the largest error for each method showed that some samples were always difficult to estimate regardless of the algorithm used to build the model (Figure 4.5). These samples are of particular interest, as they were not considered spectral outliers during the data cleaning stage. For each model, the majority of the 10 extreme outliers were over predictions (above the 1:1 line) at low-to-mid CCS levels. This agrees with the results of Chapter 3 that showed that the same samples were over-predicted by PLSR and SVR models of brix in juice, pol in juice and apparent purity.

The tendency for all models to misrepresent certain samples may be a result of the calibration set insufficiently representing the variation of CCS and spectral data in the validation set (Agelet and Hurburgh, 2010, Isaksson and Naes, 1990, Naes and Isaksson, 1989). Samples that were difficult to estimate may also represent data entry errors in the reference data, or genuinely atypical samples. For example, it is possible that these samples represent deteriorated or contaminated cane samples. It has been shown that as cane deteriorates the standard refractometer approach to calculating Pij becomes unsuitable (Lionnet, 1986). This may show up as a mismatch in NIR spectroscopic analysis as the nature of the reference measure the NIRS model is trying to predict has changed. This would have flow-on effects to measures such as CCS, which are based on Pij.

4.4.3 Variable importance

Wavelength importance within each model was visualized alongside the correlation between the transformed absorbance spectra and laboratory measured CCS values (Figure 4.6). The PLSR coefficient based importance measure (Figure 4.6 (a)) suggested that the PLSR model generally suppressed the influence of wavelengths above 1,900 nm. This higher wavelength region was noisier than other regions, with few identifiable peaks. In contrast, there were several well-defined regions below 1,900 nm that had high influence within the PLSR model. The 1,600 nm – 1,800 nm and 1,150 nm – 1,250 nm regions were important within the PLSR model. These two regions feature CH first overtones and CH second overtones respectively (Shenk et al., 2008). This suggests that the PLSR model focused on regions of the spectrum with similar information. The ability to identify regions with similar information build confidence in the models ability to extract genuine information from the spectral data.

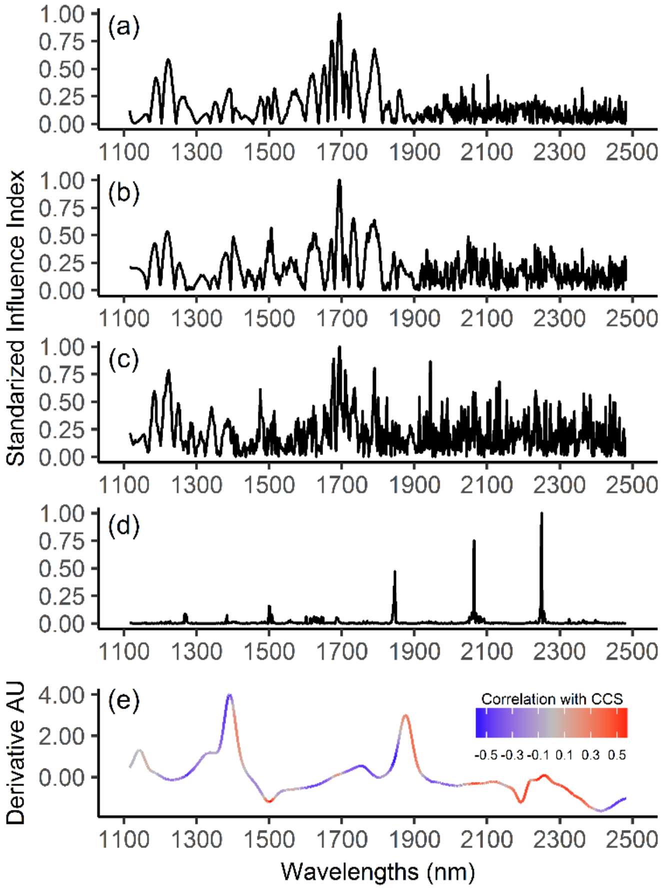

Figure 4.6. Standardized influence of each wavelength for the PLSR (a), SVR (b), ANN (c) and GBT (d) models. Indices were standardized relative to the maximum value such that the most influential wavelength always had a value of one. Plot (e) shows the derivative spectrum used in the analysis and the correlation to CCS at each wavelength.

Wavelengths between 1,600 nm and 1,900 nm that were important within the PLSR model had higher correlations between the transformed spectra and laboratory measured CCS (Figure 4.6(a) and Figure 4.6(e)). However, wavelengths in the 1,150 – 1,250 nm range tended to have lower correlations (Figure 4.6(e)) despite being important in the PLSR model. Wavelengths above 1900 nm were often highly correlated with CCS but were not relatively important in the PLSR model. This may be a result of the PLSR model attempting to suppress wavelengths with a low signal to noise ratio. Within the Australian sugar industry, wavelengths above 1,900 nm are generally removed from analysis due to the low signal to noise ratio. This means that including these higher wavelengths can often result in lower reproducibility using the model, despite the relatively strong correlations.

The SVR (Figure 4.6(b)) and ANN (Figure 4.6(c)) models had similar wavelength importance signatures to the PLSR model. In both machine learning algorithms the 1,150 nm – 1,250 nm and 1,600 nm – 1,900 nm regions had some of the most important wavelengths, generally with well-defined peaks. The GBT model was characterized by a few narrow wavelength regions with relatively high importance. The wavelengths identified as important also had some of the highest correlations with CCS (Figure 4.6(d) and Figure 4.6(e)). This is likely a consequence of the tree based nature of the GBT model. Few, highly important variables are a feature of simple trees and is often considered an advantage in variable selection methods (Feilhauer et al., 2015).

In contrast to the three other models investigated, the GBT model identified wavelengths in the spectral region above 1,900 nm as having high relative importance. The high importance the GBT model placed on wavelengths above 1,900 nm may have contributed to the relatively low predictive performance of the model when compared to PLSR, SVR and ANN. The selection of only a few very important wavelengths may also have contributed to the lower performance. CCS is not derived from a single organic compound, but is instead an estimate of the relationship between several measures including sucrose and fibre. In this respect in may be difficult to summarize CCS using only a small number of wavelengths.

Simpler and more robust models may have been possible if the low signal to noise spectral region above 1,900 nm was removed from the analysis. Although previous research has identified higher wavelengths as problematic within the Australian sugarcane industry, it was worthwhile comparing models using the full spectral range. By using the full spectral range, it was possible to identify that the two machine learning algorithms made use of the same spectral regions as the PLSR algorithm, which is more commonly used in industry. As far as we are aware, this is the first time these indices have been reported as part of the investigation of SVR and ANN models within agricultural industries.

4.5 Conclusions

Three calibrations techniques were compared to partial least squares regression as methods for estimating CCS from a large NIRS data set of sugarcane samples. On a single independent data set, SVR and ANN performed similarly to PLSR and could feasibly replace the more typical approach without loss of overall skill. However, given the similar performance and relative simplicity of the PLSR model, there is no strong evidence to recommend a switch within the Australian sugar industry.

A deeper investigation of sample errors showed that all four of the calibrated models poorly estimated many of the same samples. Difficult to estimate samples, may represent deteriorated cane, or samples that are under-represented in the calibration set. This suggested that understanding why samples are difficult to predict, identifying these samples and adapting models accordingly would be more beneficial than simply applying a new calibration technique. Adaptations could involve something as complex as developing separate models are as relatively simple as developing a calibration set that better represents these samples.

Analysis of wavelength influence in each model showed the techniques emphasized similar spectral regions and in generally suppressed the contribution of wavelengths above 1,800 – 1,900 nm. Wavelength importance also offered some insight into possible reasons the GBT model did not perform as well as other models tested. Future studies comparing algorithms should strongly consider using a similar approach to help better understand how new methods use spectral data compared to methods better understood within the industry, such as PLSR. An understanding of the wavelengths that are influential within the model can also inform the design and development of new cane analysis systems.

By comparing PLSR, SVR, ANN and GBT this research has provided an overview of various machine learning and non-linear calibration techniques that are relatively novel within the Australian sugarcane industry. This overview has highlighted several key recommendations for the Australian sugar industry:

1. SVR and ANN models provided no strong improvement over PLSR, suggesting industry does not need to modify current approaches. However, future research may consider SVR or ANN for situations where non-linear effects may be an issue.
2. Methods for determining the importance of spectral bands within SVR and ANN models were explored. Future research should make use of similar methods to help better understand how/why a novel model does/does not perform well.
3. The use of data from a single season and mill was a limitation of this study. Future research should consider a more global calibration approach by including more temporal or spatial variability in the dataset. This would result in models that are more robust and may be a better comparison of the potential benefits of novel techniques.

4.6 Chapter 4 Summary

On-line near infrared spectroscopic (NIRS) analysis systems, play an important role in assessing the quality of sugarcane in Australia. As quality measures are used to calculate the payment made to growers, it is imperative that NIRS models are both accurate and robust. Machine learning and non-linear modelling approaches have been explored as methods for developing improved NIRS models in a variety of industrial settings, yet there has been little research into their application to cane quality measures. This chapter compared chemometric models of Commercial Cane Sugar (CCS) based on four calibration techniques. CCS was estimated using Partial Least Squares Regression (PLSR), Support Vector Regression (SVR), Artificial Neural Networks (ANN) and Gradient Boosted Trees (GBT). SVR (RMSEP = 0.37%) and ANN (RMSEP = 0.36%) performed similarly to PLSR (RMSEP = 0.37%) on the validation data set, while GBT exhibited a much lower skill (RMSEP = 0.51%). Analysis of important wavelengths in each model showed that PLSR, SVR and ANN techniques emphasized the importance of similar spectral regions. This comparison of variable importance has rarely been provided in previous studies.

The focus of Chapter 4 was Objective 1 of the thesis: Investigating the use of data mining and machine learning algorithms for improved NIRS estimates of cane quality. Results from Chapter 4 confirmed that PLSR was as effective as SVR and ANN but that GBT failed to perform as well as other techniques. The similar performance of PLSR, SVR and ANN models was an important result as PLSR is a straight forward approach that is easy to understand and is already well established within industry applications. This comparison also provides a counterpoint to many studies presented in the literature where ANN and SVR outperform PLSR. As such these results emphasise the importance of comparing modelling approaches.

The results of the variable importance comparison showed that PLSR, SVR and ANN placed importance on similar wavelength regions while GBT placed much higher importance on a small number of wavelengths. This was a valuable contribution to the Australian sugarcane industry and the wider modelling community as it was possible to show why the GBT model underperformed. The variable importance investigation also showed that it was possible to see inside the ‘black-box’ of ANN and SVR. This is a crucial step in building confidence in using machine learning modelling approaches. The lessons learnt from the research of Chapters 3 and 4 were used in developing similar methodology for identifying atypical samples in Chapters 5 and 6.

Chapter 5

A feasibility test for detection of atypical cane samples using near infrared spectroscopy

Relevant publication	Sexton, J., Everingham, Y. Donald, D., Staunton, S. & White, R. A feasibility test for detection of atypical cane samples using near infrared spectroscopy Proceedings of the Australian Society of Sugar Cane Technologists, 2018 Mackay, Queensland, Australia. 382 – 390.
Statement of intellectual input	The research question / objective of Chapter 5 was developed by the candidate with input from Dr. Everingham, Dr. Donald and Mr. Staunton. Data for the thesis was provided by SRA through Mr. Staunton. Dr. Everingham, Dr. Donald, Mr. Staunton and Dr. White supplied editorial assistance. Mr. Staunton provided the range of NIR spectral values used in the analysis. Mr. Staunton on behalf of SRA provided the NIR spectral regions used in the analysis. As such, they were not published. For consistency, with the published version the values are also absent from this chapter. The candidate developed the methodological framework and ran all simulations. The candidate was also responsible for the write-up of the chapter and produced all tables and figures.
Publication status	Published

5.1 Introduction

In any given harvest season, tens of thousands of cane consignments are processed by a sugarcane mill. Of these, thousands of consignments (if not all) are tested under laboratory conditions for quality measures such as Brix in juice (Bij) and Pol in juice (Pij). Mill researchers have noted that in any given season, 1–5% of samples often have unusually low laboratory estimates of Pij given the recorded Bij value (Figure 5.1). These 'atypical' samples are of particular concern as they may represent deteriorated or contaminated cane samples.

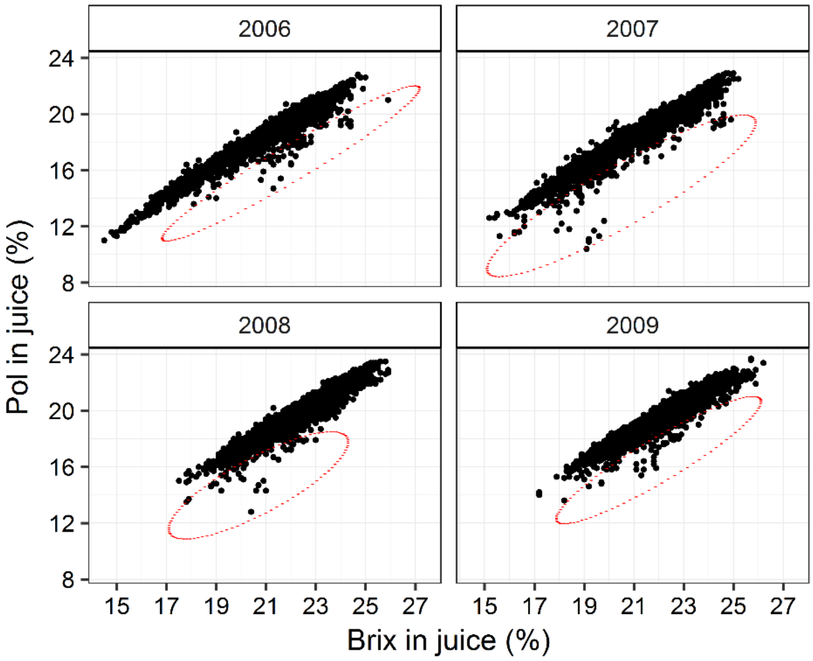

Figure 5.1. Bij and Pij laboratory measured values for data from a northern Queensland sugarcane mill from 2006 to 2009. In each season, a small number of samples appear to have particularly low Pij for their reported Bij (red-circled points). These samples will have low apparent purity compared to cane with similar levels of Bij and may represent deteriorated or contaminated samples.

Deteriorated or contaminated cane has a number of negative impacts on the cane milling process. Deterioration in particular can lead to higher viscosity, longer crystallisation times, elongated crystals, and distorted Pol readings (Solomon, 2009).During deterioration, sucrose is metabolised into less economic products such as organic acids, complex polysaccharides (e.g. dextran) and gums (Solomon, 2009). Deterioration due to delays between harvesting and crushing can lead to increased dextran levels and higher Brix readings (Saxena et al., 2010). Contamination by impurities such as soil can inflate Brix readings, which may lead to reduced quality indices such as Apparent Purity (AP) and Commercial Cane Sugar (CCS).

Given the potential impacts to mill operations, it would be beneficial to be able to identify these atypical samples in real time as they enter the mill. Although many indicators of cane deterioration exist, most are considered impractical for use during the milling process (Van Heerden et al., 2014). As such, observed 'atypical' samples are not recorded as deteriorated or contaminated. Clearly, a rapid, inexpensive indicator is still required within the Australian sugar industry.

Near Infra Red Spectroscopy (NIRS) has been used widely within the sugarcane industry as a rapid and cost effective method for analysing cane properties. In many mills, cane quality measures such as Bij and Pij are calculated using NIRS methods during the milling process. NIRS can accurately estimate Bij, Pij and CCS in an online setting, as shown by Staunton et al. (2004) and in Chapter 4. Unfortunately, there is no evidence in the literature for NIRS classification of deteriorated or otherwise atypical cane samples. Therefore, the purpose of this manuscript was two-fold. Firstly, to develop a method for defining the observed 'atypical' cane samples based on a large collection of laboratory measured Bij and Pij data; and secondly to test the feasibility of detecting these samples using an NIRS model. The ability to identify atypical samples in a rapid and non-invasive manner will be useful in quality control measures within the mill and could lead to improved NIRS models specific to these particular samples.

5.2 Materials and methods

This analysis was performed in two stages. Firstly, two approaches for defining atypical samples based on laboratory data were compared and the most appropriate definition was chosen. Secondly, an NIRS model was built using PLS-DA in order to determine the feasibility of identifying atypical cane samples from NIR data collected during the milling process

5.2.1 Data

Data were collected from a single sugarcane mill in North Queensland, Australia. Data spanned the 2006 to 2009 harvest seasons. Laboratory measures of Bij, Pij and AP were collected with paired NIR spectroscopy data. In total, there were 13,014 samples included in the analysis (Table 5.1).

Table 5.1. Summary of sample laboratory data by harvest season (2006–2009).

Season	Count	Bij (%)		Pij (%)		AP (%)	
		Median	Range	Median	Range	Median	Range
2006	3,794	21.0	14.5-25.9	18.4	11.0-22.8	87.9	69.0-95.0
2007	3,389	21.2	15.2-25.2	18.6	10.4-22.9	87.8	54.5-95.6
2008	3,067	22.3	17.5-25.9	19.8	12.8-23.5	89.1	62.7-94.8
2009	2,764	22.1	17.2-26.2	19.6	13.6-23.7	88.7	72.3-95.5
Total	13,014	21.6	14.5-26.2	19.1	10.4-23.7	88.3	54.5-95.6

NIRS data linked to each sample were collected using a FOSS ONLINE 5000 system. Spectral wavelengths ranged from 1,100 nm to 2,498 nm at two nm intervals. Wavelengths used in the analysis were selected based on industry recommendations.

5.2.2 Defining atypical cane samples

The atypical samples highlighted in Figure 5.1 are difficult to define as cane samples are not regularly identified as deteriorated, contaminated or otherwise ‘atypical’. Therefore, we sought to define these atypical samples based on laboratory records of Bij and Pij. In particular, we were interested in identifying samples that have a particularly low Pij compared to samples with similar recorded Bij values.

These samples were likely to have relatively low AP. Therefore, two approaches to defining atypical samples were explored (i) samples with low AP compared to all observed values and (ii) samples with low residuals from a linear regression of Pij on Bij, compared to all observed values. For each approach, atypical samples were identified based on all available samples and the difference in cane quality measures (Bij, Pij and AP) were explored. A single approach was then chosen as the definition for atypical samples, for use in the second stage of the study.

5.2.2.1 Atypical samples based on apparent purity

Deteriorated samples generally have lower purity than healthy cane. As an initial approach atypical sample were defined as samples with a low AP. Apparent purity was calculated as the ratio of Pij to Bij, expressed as a percentage:

$$AP = \frac{Pij}{Bij} \times 100. \qquad (5\text{-}1)$$

Apparent purity was assumed to have a normal distribution and atypical samples were identified as being below the expected first percentile (the lowest 1% of samples) of the appropriate normal distribution. The appropriate cut-off point for the normal distribution was calculated using the qnorm function in the R statistical programming language (R Core Team, 2017). In practice, samples with an AP of less than 80.96% were considered atypical ‘low AP’ samples (Figure 5.2(a)).

5.2.2.2 Linear regression residuals

A strong linear relationship exists between Bij and Pij (Figure 5.1 and Figure 5.2). As Pij is a subset of all dissolved solids in a solute (Bij), a linear model was built as:

$$\widehat{Pij} = M \times Bij + C. \qquad (5\text{-}2)$$

Where M and C are the slope and intercept of the line of best fit between Bij and Pij. $\widehat{Pij}$ is the model estimate of Pij.

The studentized residuals $(\widehat{Pij} - Pij)$ of this equation were used to identify atypical samples. A negative residual implies the sample has a Pij lower than expected for its measured Bij, while a positive residual implies that the samples has a Pij higher than expected. Samples with a studentized residual lower than the expected cut-off for the first percentile (the lowest 1% of samples) were considered atypical. In practice, samples with a residual lower than -2.33% were considered atypical (Figure 5.2(b)). The linear regression was fitted using the lm function in the R statistical programming language (R Core Team, 2017). The appropriate cut-off point for the t-distribution was calculated using the qt function.

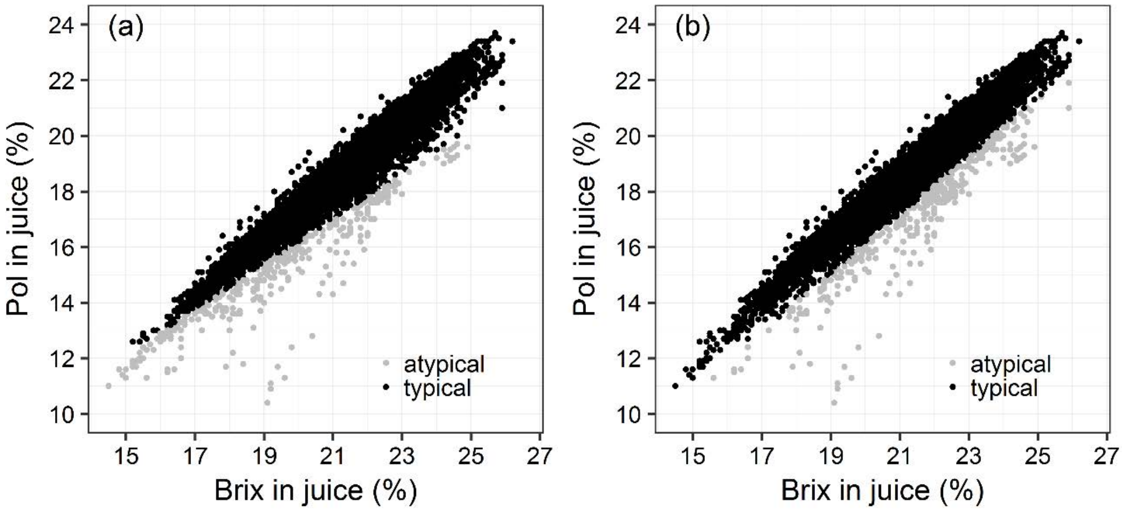

Figure 5.2. Distribution of typical (black) and atypical (grey) samples for (a) AP and (b) residual definitions of atypical. Using the AP definition (a) samples were considered atypical if they had an AP of less than 80.96%. Using the residual definition (b) samples were considered atypical if they had a Pij residual of less than -2.33%. I.E. samples had a Pij 2.33% lower than would be expected for the recorded Bij. For both definitions, 2.8% of all samples were considered atypical.

5.2.3 Detecting atypical cane samples using NIRS

The method of Partial Least Squares discriminant analysis (PLS-DA) was used to build an NIRS classification model to identify atypical and typical samples, based on the definition chosen in stage 1. In order to build PLS models the following steps were taken.

1. Pre-process NIRS data and split into training and test set
2. Tune PLS-DA model on a training set
3. Apply calibrated PLS-DA models to test data and evaluate

5.2.3.1 Partial Least Squares Discriminant Analysis

Although not originally designed for classification problems, partial least squares discriminant analysis (PLS-DA) has been identified as one of the most used classification techniques within chemometrics (Barker and Rayens, 2003). In PLS-DA, a dummy matrix is used to represent the categorical response variable (the known classes) (Song et al., 2016). PLS-DA has recently been used in NIRS analysis to discriminate between organic and Non-organic apples (Song et al., 2016) and has previously been found to be a moderately good classifier for gasoline analysis (Balabin et al., 2010) and an effective method of identifying the geographical origin of a Brazilian sugarcane spirit (Cirino de Carvalho et al., 2016).

5.2.3.2 Data pre-processing

Data were randomly split 50:50 into a training and test set using a stratified approach so that the ratio of atypical to typical samples remain the same in each set. Following Chapter 3, NIR spectral scans were transformed using a Savitzky-Golay (Savitzky and Golay, 1964) first derivative with a window width of 17 and scatter corrected using a standard normal variate transformation (Barnes et al., 1989). As each sample had multiple associated NIRS scans, the final NIR spectrum for each sample was the average spectra after the transformations were applied. All samples in the analysis had at least three associated NIRS scans.

5.2.3.3 Model tuning

The PLS-DA model was tuned using a combination of up-sampling and five-fold cross-validation of the training dataset. The number of latent variables was the only parameter tuned. Models with up to 40 latent variables were tested.

Models were tuned in order to maximise the Receiver Operator Characteristic, Area Under Curve score (AUC). The AUC can be a better measure of model performance for unbalanced classes than the overall accuracy (correct classification rate). Cross-validation estimates the predictive ability of the calibrated model by dividing the data into five “folds”. Each fold (~20% of calibration samples) was successively removed from the analysis and the models were built using the remaining four folds (~80% of calibration samples). Up sampling was used to correct for the unbalanced classes (many more typical than atypical samples). For each fold, atypical samples in the calibration portion were resampled with replacement until there were the same number of typical and atypical samples. Cross-validated ROC AUC was then calculated on the data that were left out across all folds.

To capture uncertainty in the cross-validation results caused by the random nature of the up-sampling process, the 5-fold cross-validation was performed 5 times. The final combination of parameters was selected as the simplest model with a cross-validated AUC within one standard error of the absolute maximum AUC. The standard error of the cross-validated AUC was calculated from the 25 (five times five cross-validation runs) AUC values calculated for each number of latent variables.

The final PLS-DA model using the chosen number of latent variables was re-calibrated on the full training dataset. The PLS-DA model was calibrated and built using the caret package in R (Kuhn, 2017) and used the softmax method to calculate class probabilities. The predicted class was taken as the class with the highest probability.

5.2.2.4 Model evaluation

The final model was applied to the test set and the predictive AUC was recorded. Overall model Accuracy (correct classification rate), Sensitivity (correct classification rate of atypical samples) and Specificity (correct classification rate of typical samples) were also recorded for both the training and test data set. Model skill was also explored by month of harvest, to assess performance throughout the harvest season.

5.3 Results and discussion

5.3.1 Defining atypical cane samples

The AP definition (Figure 5.2(a)) of atypical tended to identify more samples with both low Bij (< 17%) and Pij (< 13%) as atypical than the residual definition (Figure 5.2(b)). In contrast, the residual definition tended to identify more samples with high Bij (>25%) as atypical. Consequently, atypical samples tended to have lower average Pij, Bij and AP using the AP definition (Table 5.2).

The AP definition also had a larger difference in average quality measures between typical and atypical samples. This suggests that defining atypical samples based on low AP, better identifies samples with lower quality measures. However, by comparing Figure 5.1 and Figure 5.2 it is evident that the residual definition better matches the graphically atypical samples. The difference appears to be that the AP definition does not account for changes in AP throughout the harvest period (Figure 5.3).

Table 5.2. Summary of typical and atypical samples mean for each definition of typical and atypical samples. Numbers in brackets represent the standard deviation (SD) of the mean

Definition approach	Sample Type	Percentage of Samples (%)	Bij (%) Mean [SD]	Pij (%) Mean[SD]	AP (%) Mean[SD]
Low Purity	Typical	97.2	21.5[1.73]	19.0[1.84]	88.1[2.36]
	Atypical	2.8	19.3[2.33]	15.0[2.02]	77.9[4.17]
Residual	Typical	97.2	21.5[1.78]	18.9[1.92]	88.1[2.47]
	Atypical	2.8	21.5[1.98]	17.1[2.32]	79.2[4.93]

AP tended to increase until approximately August before plateauing (Figure 5.3). This aligns with sucrose accumulation in the stalk, which increases as the cane matures before beginning to plateau at around 300 days after planting (Muchow et al., 1996). Using the AP definition, 32.45% of May harvested samples were considered atypical (Figure 5.3(a)). As AP is lower in immature cane many of these samples were likely healthy cane that should not be considered atypical.

In comparison, using the residual definition only 12.08% of May harvested samples were considered atypical (Figure 5.3(b)) and the AP of atypical samples tended to increase throughout the harvest season. This suggests that by considering the linear relationship between Bij and Pij, the residual definition of atypical may be more representative of atypical samples that were deteriorated or contaminated rather than cane with low quality measures. For this reason, we consider the residual definition of atypical samples for identification in the NIRS analysis.

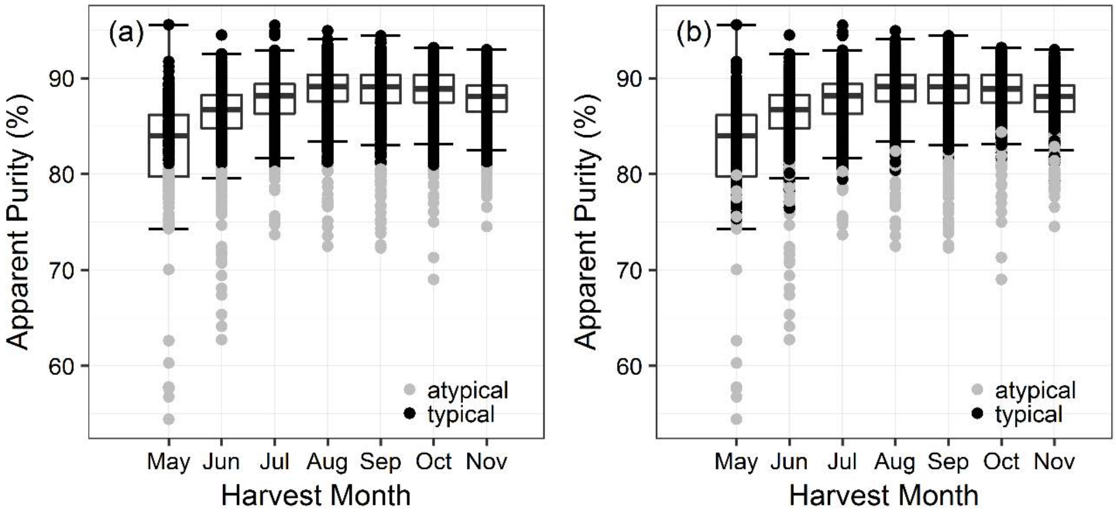

Figure 5.3. Box and whisker plots of the monthly distribution of AP across all years. Boxes represent 50% of all samples per month with thick horizontal lines identifying the median AP. Points represent typical (black) and atypical (grey) samples based on (a) AP and (b) residual definitions of atypical.

5.3.2 Detecting atypical cane samples using NIRS

The final model had a ROC AUC of 0.935 on the test set. This suggests that the model could distinguish between typical and atypical samples based on the residual definition. Likewise, the overall model accuracy, recorded as the correct classification rate, was high (Accuracy = 91.6%). However, given the highly unbalanced nature of the data set (97.2% typical 2.8% atypical) the overall accuracy must be viewed with caution. For example, if all samples were identified as typical the overall CCR would be 97.2%.

To understand model performance the correct classification rates of atypical (Sensitivity) and typical (Specificity) samples were recorded (Table 5.3). On the test set the PLS-DA analysis correctly classified 86.6% of atypical samples and 91.8% of typical samples. These results are encouraging as the model could correctly identify most atypical samples and suggest that it is indeed feasible to identify atypical cane samples online using NIRS analysis.

Table 5.3. PLS-DA model skill for calibration and validation datasets. A high ROC AUC (close to 1) indicates a good model. Model Accuracy was recorded as the overall correct classification rate (%). Sensitivity gives the correct classification rate of atypical samples while specificity gives the correct classification rate of typical samples.

	ROC AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)
Training Set	0.968	92.1	87.2	92.2
Test Set	0.935	91.6	86.6	91.8

Further analysis showed that model skill varied throughout the harvest period (Figure 5.4). For samples harvested early (May) and late (November), the model tended to be over sensitive, correctly classifying all atypical samples (high sensitivity) but also identifying samples that are more typical as atypical (lower specificity).

This drop in model skill may be due to the model being over fitted to samples harvested mid-season. Early and late harvested samples represent the lowest and highest Bij and Pij values. Furthermore, cane is rarely harvested as early as May or as late as November. This suggests that these samples were underrepresented in the modelling process. Future research should consider reducing the training dataset so that low and high Bij and Pij (or early and late harvested) samples are not underrepresented.

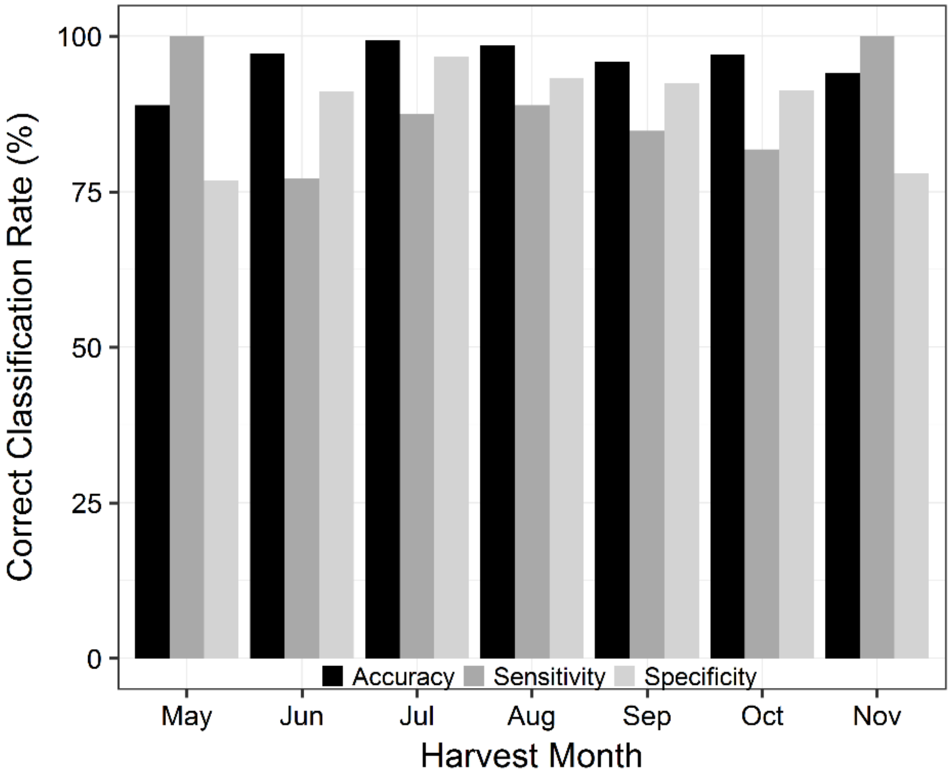

Figure 5.4. Model skill by harvest month. Accuracy represents the overall correct classification rate. Sensitivity represents the correct classification rate of atypical samples and Specificity represents the correct classification rate of typical samples.

The relatively high sensitivity and specificity mask the comparatively large number of typical samples misclassified. In total 519 typical samples were misclassified while only 155 atypical samples were correctly classified. This suggests that the current model may not be suitable for certain tasks in a practical setting. Future research will need to consider ways to improve on the current modelling techniques to ensure that the operational models are fit for purpose.

One such method may be to consider class probabilities. The PLS-DA model can return a pseudo-probability of the class of each sample (Ballabio and Consonni, 2013). Figure 5.5 shows that the misclassified samples tended to sit close to the boundary between typical and atypical samples. This suggests that the number of typical samples misclassified could be reduced by cut-off point at which the model identifies a sample as atypical. By default, the predicted class is the class with a probability of more than 50%. By increasing this value, it would be possible to fine-tune the model to reduce the misclassification of typical samples.

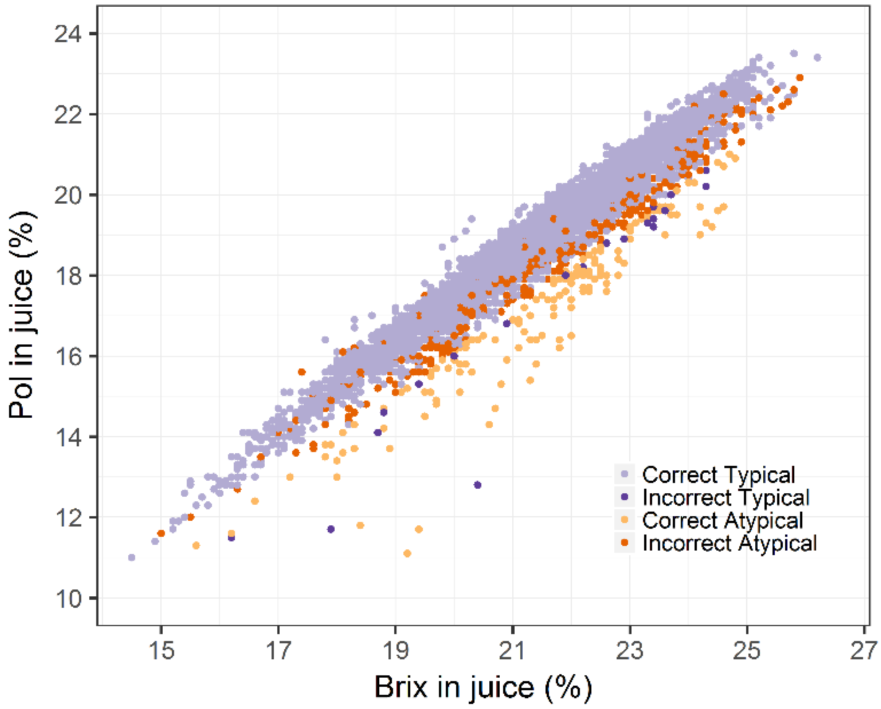

Figure 5.5. Distribution of correctly and incorrectly classified samples within the Pij/Bij space. Correct typical (purple) and correct atypical (light orange) represent samples that were correctly predicted to be typical and atypical respectively. Incorrect typical (dark purple) represent atypical samples that were predicted to be typical, while incorrect atypical (dark orange) represent typical samples that were predicted to be atypical.

The results shown in this analysis are promising. The ability to identify these atypical samples in real time using NIRS analysis will enable researchers to understand when and where these samples are occurring. As NIRS models can struggle to estimate extreme values accurately (Chapter 4), models could be built specifically for atypical samples and applied in real time if a sample is identified as atypical.

5.4 Conclusion

The objective of this research was to define the atypical samples observed in laboratory mill data, and to test if it was feasible to use NIRS instrumentation to identify atypical cane samples as they are processed in the mill. We defined atypical samples using a Pij 'residual' approach. By comparing this 'residual' approach with an approach based on AP, we were able to show that the observed atypical samples were not simply samples with a low apparent purity. The main advantage of the residual approach was that early harvested cane was less likely to be consider atypical then if all samples with a low AP were considered atypical. Once defined, a PLS-DA model could correctly discriminate between typical and atypical samples with high enough accuracy to show that it is indeed feasible to identify these samples during the milling process.

Further refinement is needed before this methodology is ready for industrial deployment. Industry will need to consider carefully, the appropriateness of the definition of 'atypical' used in this research and the effect these atypical samples have on NIRS estimates of cane quality measures. The methodology outlined in this research provides a basis for future research that could easily consider new definitions of atypical samples.

5.5 Chapter 5 Summary

Mill researches have noted that in any given season, 1–5% of samples often have unusually low laboratory estimates of Pol in juice (Pij) given the recorded Brix in juice (Bij) value. These 'atypical' samples are of particular concern as they may represent deteriorated or contaminated cane samples. Deteriorated or contaminated cane has a number of negative impacts on the cane milling process. Deterioration in particular can lead to higher viscosity, longer crystallisation times and overall lower cane purity. Many indicators for cane deterioration have been proposed but most are considered expensive, time consuming or unreliable, making them impractical for use during the milling process. Near Infra Red Spectroscopic (NIRS), analysis has been implemented in many Australian sugarcane mills to replace or supplement laboratory analysis of cane quality. However, there is little evidence in the literature that NIRS has been used to classify atypical samples. The purpose of this Chapter was to test the feasibility of predicting possible atypical cane samples using NIRS analysis. Data were collected from a single Australian sugarcane mill from 2006 to 2009. In total, 13,014 samples were collected with Bij, Pij, apparent purity (AP) and NIR spectroscopic data. Atypical samples were defined based on laboratory Bij and Pij values as cane deterioration/contamination data are not routinely measured. A partial least squares discriminant analysis (PLS-DA) was then used to build an NIRS model to identify the defined atypical cane samples. On a test set, the PLS-DA analysis had a correct classification rate of 91.6% of all samples with 86.6% of atypical samples correctly classified and 91.8% of 'typical' samples correctly classified.

The focus of Chapter 5 was Objective 2 of the thesis: Investigating the use of NIR spectroscopic analysis for the automatic identification of atypical cane samples. In order to do this it was first necessary to define 'atypical' cane. Chapter 5 showed that the definition of atypical samples that was developed matched the samples that appeared atypical graphically. The definition of atypical samples also matched temporal trends in apparent purity. As there was no definition for atypical cane it was important to show that the developed definition was appropriate and useful measure. In practice only approximately three percent of all samples were defined as atypical. This large imbalance between classes made discrimination a potentially difficult task. This made the high accuracy of the PLS-DA modelling approach used an important outcome. The feasibility test presented in Chapter 5 was a necessary to show that the difficult task of correctly identifying atypical samples was possible. This paved the way for a more extensive comparison of modelling approaches.

Chapter 6

Model comparison for online identification of atypical cane samples in a sugarcane mill using NIR analysis

Relevant publication	Sexton, J., Everingham, Y., Donald, D., Staunton, S. & White, R. 2020. Investigating the identification of atypical sugarcane using NIR analysis of online mill data. *Computers and Electronics in Agriculture,* 168**,** 105-111. doi: 10.1016/j.compag.2019.105111.
Statement of intellectual input	The research question / objective of Chapter 6 was developed by the candidate with input from Dr. Everingham, Dr. Donald and Mr. Staunton. Data for the thesis was provided by SRA through Mr. Staunton. Data for the thesis was provided by SRA through Mr. Staunton. Dr. Everingham suggested the use of Wavelets as a spectral pre-processing step. Dr. Everingham, Dr. Donald, Mr. Staunton and Dr. White supplied editorial assistance. The candidate developed the methodological framework and ran all simulations. The candidate was also responsible for the write-up of the chapter and produced all tables and figures.
Publication status	Published

6.1 Introduction

Within the Australian sugarcane industry, sugarcane is routinely analyzed for quality measures such as Pol in juice (Pij) and Brix in juice (Bij), both in the laboratory and using Near Infrared (NIR) spectroscopy. Bij is representative of the concentration of total dissolved sugars and is measured by brix spindle in the laboratory (BSES, 1991). Pij is representative of the concentration of sucrose in juice and is measured by polarimeter in the laboratory. In any given crush season, anywhere from 1% to 5% of laboratory-analysed cane can be observed to have unusually low Pij relative to their Bij. These 'atypical samples' are of particular concern as they may represent deteriorated or contaminated cane that can negatively affect the milling processes and throw off grower payment calculations. In many mills only a fraction of cane consignments are analysed in the laboratory, with fast online NIR analysis performed on the vast majority of samples. This means that the majority of occurrences go undocumented and appropriate interventions at the mill and farm level cannot be applied. On-line NIR analysis systems offer a potential solution but significant challenges must be overcome in order to build an appropriate model calibration.

As cane deteriorates, sucrose is converted to sugars that are more complex. These sugars can deflate laboratory Pol readings and cause crystal elongation leading to longer drying times. Contamination of juice with high levels of dirt or leaf matter can inflate laboratory Bij measures lowering cane payment measures. Unfortunately, there is little published work on spectroscopic analysis of deterioration, contamination or indeed discrimination problems within the sugarcane industry. One possible reason for this is the difficulty of strictly defining the point at which cane should be considered 'deteriorated'. Many measures, such as ethanol levels, ash content, or quality measures such as apparent purity (ratio of Pij/Bij) are either non-discriminatory, difficult to measure or overly inflated by the deterioration process and are not generally used in process control (Van Heerden et al., 2014).

Discrimination of atypical samples in an online system face two further challenges in the scale of variability and the imbalance of available samples. Online systems could include contamination from a range of sources (e.g. soil and leaves) as well as different types of deteriorated cane (e.g. sour or stale cane) (Van Heerden et al., 2014). In a laboratory setup, Tulip and Wilkins (2004) used spectral data in the visual (400 nm – 1,100 nm) and NIR (1,100 nm – 2,500 nm) range to estimate the dirt concentration in cane samples with good accuracy. Their results suggest that predicted soil type could be used to help estimate dirt concentration. However, there was no evidence of how effective this would be in an online system. A similar approach - where contaminated or deteriorated samples are created for a laboratory experiment - has been used in other industries such as decay in oranges (Li et al., 2016), identifying adulterated milk (Zhang et al., 2014), detecting fungal contamination in barley (Senthilkumar et al., 2016) and bacterial spoilage of kiwi juice (Niu et al., 2018). A devised experiment gives ideal conditions and by necessity have a limited size. This means they are unlikely to cover the full range of variability experienced in an online environment.

Atypical samples such as deteriorated or contaminated samples are unlikely to occur as frequently as typical samples. In a laboratory experiment, the number of typical and atypical samples may not accurately reflect the occurrence in a real-world scenario. For example, Fiedler et al. (2001) suggested that it was difficult to classify specific cane varieties using on-line NIR analysis as the majority of consignments were of the same variety. However, Everingham et al. (2007) were able to correctly assign a variety class at the paddock level based on Hyperion satellite image data using Vis/NIR wavelengths (400 nm – 2,400 nm) and machine learning algorithms.

Many different modelling approaches and pre-processing techniques have been explored for discrimination problems using spectral data. In a comparison using gasoline data, Balabin et al. (2010) described three classes of classification models:

1. Low performance including Linear Discriminant Analysis (LDA) and soft independent modelling of class analogy
2. Medium performance including Partial Least Squares (PLS) and artificial neural networks (ANN) and
3. High performance models including K-Nearest Neighbours (KNN) and Support Vector Machine (SVM).

The effectiveness of modelling approaches can differ between problems. For example, SVM outperformed random forest (RF) and LDA classification models for sugarcane varieties based on hyperspectral data (Everingham et al., 2007) while RF outperformed SVM in the identification of bruises in apples (Che et al., 2018). Wang et al. (2004) concluded that PLS could outperform ANN at discriminating between soybeans with and without fungal contamination, yet ANN performed better at discriminating between types of fungi contamination.

More complex models such as SVM and ANN are often better suited to non-linear or complex systems. ANN approach non-linearity by forming a network of linear relationships between predictor variables and the output and using a nonlinear transfer function between layers (Hastie et al., 2013d). SVM tackle non-linearity through use of the 'kernel trick' (Agelet and Hurburgh, 2010), transforming the original predictor variables into a higher dimensional space. These approaches are often difficult to maintain, require a large amount of training data and can be difficult to explain the importance of specific wavelengths. Less complex models such as LDA and PLS are easy to build and maintain, offer clear variable importance and are capable of dealing with small non-linear effects (Bertran et al., 1999). RF models offer an alternative to variable transformation for dealing with non-linearity, by building an ensemble of simple decision trees (Hastie et al., 2013g). RF models have shown to be useful in identifying important wavelengths in NIR analysis (Feilhauer et al., 2015).

Spectral pre-treatments such as spectral derivatives and wavelet transforms are regularly used to enhance spectroscopic model performance. Spectral derivatives such as those proposed by Savitzky and Golay (1964) fit a polynomial to the spectral signature and calculate the derivative at each spectral wavelength. These derivatives are used to enhance chemical signals in the spectra and remove spectral base line shift across wavelengths (Agelet and Hurburgh, 2010). As shown in Chapter 3 and Chapter 4 of this thesis, spectral derivatives are effective in the estimation of cane quality parameters. Alternative methods such as wavelets can be used to compress and extract information within spectral data. Wavelets were developed within the signal processing community, with the aim of compressing as much information from a signal into a compact form. Wavelet transformations have been shown to reduce the number of variables in NIR models without reducing model performance (Trygg and Wold, 1998). Cen et al. (2006) found that wavelet transformation was more effective than derivative spectra when discriminating infant formula while (Donald et al., 2006) showed that adaptive discrete wavelet transformation could improve discrimination between sugarcane samples from different experimental design factors.

Feature selection has also been used to improve performance of NIR analysis, with a wide range of techniques available (Xiaobo et al., 2010). In particular, genetic algorithms have been explored in a range of spectroscopic problems and have been found to be a competitive approach to feature selection (Cirino de Carvalho et al., 2016, Niu et al., 2018, Balabin and Smirnov, 2011, Yang et al., 2017). Within the sugarcane industry, the use of automatic feature selection techniques is not well documented. In general all available spectral data is used or specific wavelength ranges are defined based on industry knowledge of the problem being addressed.

As the majority of sugarcane mills in Australia assess cane using NIR spectroscopic analysis, a simple and fast NIR model to identify these atypical samples is needed. This would enable industry to trace the origin of samples with unusually low Pij or to develop a process control logic to change how atypical samples are managed by the mill. In order to develop such a model it is necessary to overcome the challenges of a large variable data set and a large imbalance in class sizes. It is important that the interactions and possible advantages of different approaches be explored. Unfortunately, there is little evidence of this problem being explored in the agricultural literature. Therefore, the objective of this chapter was to compare a range of modelling and pre-processing techniques for the classification of 'atypical' sugarcane samples using an on-line NIR cane analysis dataset.

6.2 Materials and methods

Five modelling approaches were used to discriminate between atypical and typical samples. The five approaches used were Linear Discriminant Analysis, Partial Least Squares Discriminant Analysis, Random Forest, Support Vector Machine and Artificial Neural Network. For each modelling approach, the use of five spectral pre-treatments and raw spectral data were considered. These included first and second Savitzky-Golay derivatives (Savitzky and Golay, 1964) and three wavelet transformations (Daubechies-8, Least Asymmetric-8, Coiflet-6). The most appropriate pre-treatment for each model was used in model comparisons.

Figure 6.1 shows an overview of the model calibration and validation process. Data with recorded laboratory and spectral data were split evenly into a training set and validation set, using a random stratified approach such that the ratio of atypical to typical samples was maintained in each set. This was important so that the validation set consisted of realistic data (Kuhn and Johnson, 2013b). Each combination of modelling and pre-processing approaches was tuned and calibrated using the training set while model comparisons were based solely on the validation set. Following Cui and Fearn (2017), training and validation splitting was performed multiple times. By repeating the training / validation split it is possible to capture variability in model performance due to the specific data used to build the model. The data was divided into a training and validation set 10 times. Down-sampling of the majority class was used during model calibration in order to correct for the imbalanced nature of the data. Down-sampling or under-sampling selects a subset of the majority class of equal size to the minority class (Kuhn and Johnson, 2013b). In order to capture variability in the model calibrations, each model was built 10 times. This resulted in 100 models built for each combination of pre-processing and modelling techniques.

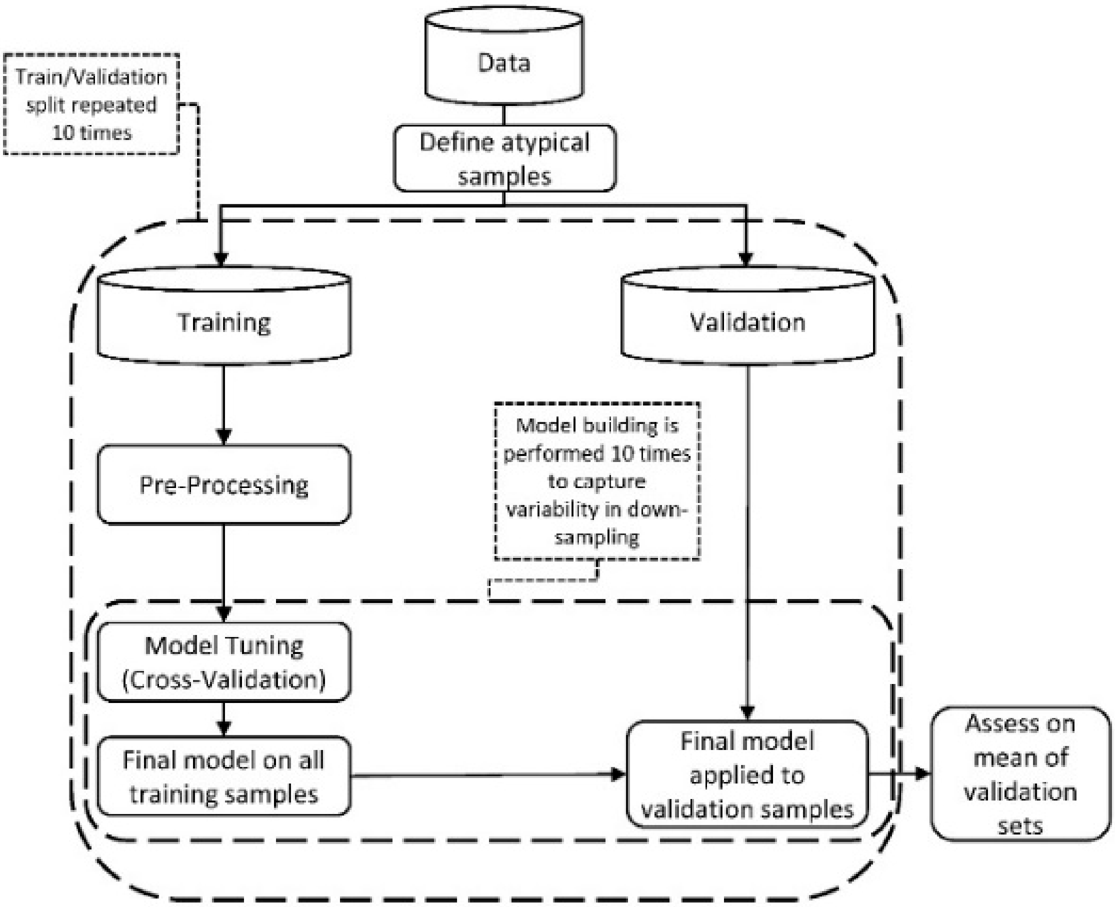
Data
Define atypical samples
Train/Validation split repeated 10 times
Training
Validation
Pre-Processing
Model building is performed 10 times to capture variability in down-sampling
Model Tuning (Cross-Validation)
Final model on all training samples
Final model applied to validation samples
Assess on mean of validation sets

Figure 6.1 Overview of methodology used in this chapter. Data were split evenly into training and validation sets 10 times. For each split, models were rebuilt 10 times to capture variability in random down sampling. Models were tuned using five-fold cross-validation and final models were built using the entire training set. Model performance investigation was performed on validation set results.

Model hyper-parameters were tuned through five-fold cross-validation within each training / validation split. The set of hyper-parameters that maximized the Receiver Operating Characteristic (ROC), Area under curve statistic (AUC) was considered the best model set and was used in model validation. AUC can be a better measure of classifier performance for machine learning classifiers (Bradley, 1997, Jin and Ling, 2005). A perfect model will have an AUC of one while a non-informative model will have a value of 0.5, while a completely incorrect classification would have a value of zero. Models were rebuilt on the entire training data set and applied to the validation set. AUC, correct classification rates for atypical ($CCR_{atypical}$) and typical ($CCR_{typical}$) samples as well as overall Accuracy were calculated as performance measures. $CCR_{atypical}$ was equivalent to model Sensitivity and was calculated as:

$$CCR_{atypical} = \frac{number\ of\ correctly\ classified\ atypical\ samples}{number\ of\ atypical\ samples} \times 100 = \frac{TP}{TP+FN} \times 100, \qquad (6\text{-}1)$$

where TP and FN are the true positive rate and false negative rate respectively. Similarly, CCR$_{\text{typical}}$ was equivalent to model Specificity and was calculated as:

$$CCR_{typical} = \frac{number\ of\ correctly\ classified\ typical\ samples}{number\ of\ typical\ samples} \times 100 = \frac{TN}{TN+FP} \times 100, \qquad (6\text{-}2)$$

Here TN and FP are the true negative and false positive rate respectively. Accuracy was calculated as:

$$Accuracy = \frac{number\ of\ correctly\ classified\ samples}{number\ of\ samples} \times 100 = \frac{TP+TN}{TP+FN+TN+FP} \times 100. \qquad (6\text{-}3)$$

Training and validation set skill were recorded as the average of 100 models. All data analysis was performed using the R statistical language and environment (R Core Team, 2017). Model calibration and validation was performed using the caret package in R (Kuhn, 2017).

6.2.1 Data

Laboratory and spectral data were collected from a single mill in Northern Queensland, Australia for the 2006 to 2009 harvest seasons. Laboratory measures collected were juice Brix (Bij) and Pol (Pij). Bij is a measure of the total dissolved sugars in a solution while Pij is used as a measure of sucrose in juice. Both Bij and Pij were measured on first expressed juice following standard industry practices and reported as a percentage. In total, there were 13,129 samples with measured Bij and Pij values used in defining atypical samples. Following the process outlined in Chapter 5, atypical samples were defined based on the linear relationship between Bij and Pij. Atypical samples made up 2.67% of all samples from 2006 to 2009 and varied from year to year, ranging from 1.92% in 2006 to 3.57% in 2007.

A FOSS 5000 on-line NIRS system, collected spectral data on shredded cane as absorbance (log(1/reflectance)). Absorbance was recorded from 1,100 nm to 2,498 nm at 2 nm intervals. Following, spectral outliers were removed based on a Global Mahalanobis distance (GH). This was done to attempt to remove scans that had a low amount of cane and may be affected by reflection from the chute where the scans take place. Following the methodology of Chapter 4, individual scans with a GH greater than three were removed from the analysis and samples with less than three 'clean' scans were removed from the. The final spectra for each sample was recorded as the average of all scans for that sample. In total, there were 12,798 samples with available laboratory and spectral data.

Removal of outlier spectra did not greatly affect the ratio of atypical samples (2.53% compared to 2.67%), showing that atypical samples were not spectral outliers in particular. Graphical differences between averaged raw spectra for atypical and typical samples were minimal and difficult to identify (Figure 6.2). Absorbance was lower for atypical samples between 1,450 nm and 1,850 nm. Slight difference were also noted around 1,200 nm and between 1,600 nm to 1,800 nm (lower absorbance for atypical samples). The close similarity between spectral signatures of atypical and typical highlight the need for effective data pre-processing and modelling methods to discriminate between sample classes.

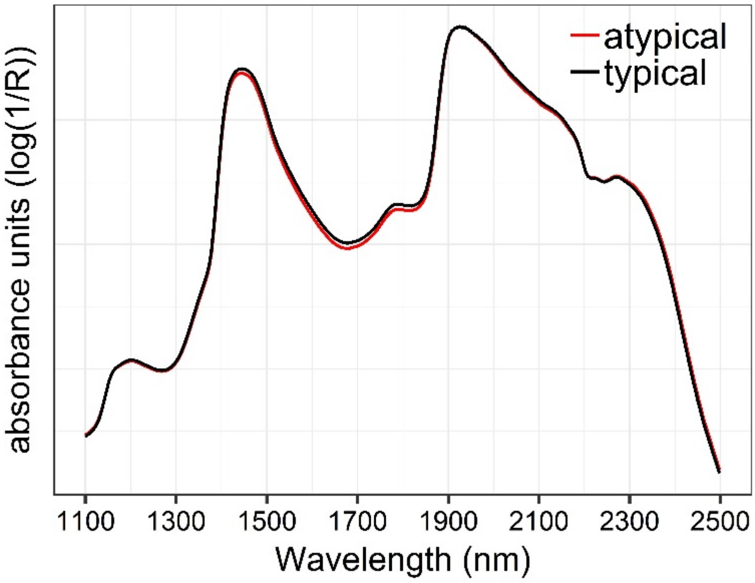

Figure 6.2. Raw spectral data for atypical (red) and typical (black) samples.

6.2.2 Spectral pre-processing

Models were built using raw spectral data and spectral pre-processing techniques:

1. *Raw*: Raw spectral data as described in the Data section. All wavelengths (1,100 nm – 2,498 nm at 2nm) were used as potential predictors.
2. *SNV-First*: A combination of Standard Normal Variate and Savitzky-Golay first derivative transformation.
3. *SNV-Second*: A combination of SNV and Savitzky-Golay Second derivative transformation.
4. *D8*: Daubechies wavelet transformation (Daublet) using eight coefficients.
5. *LA8*: Least asymmetrical wavelet transformation (Symmlet) using eight coefficients.
6. *C6*: Coiflet wavelet transformation using six coefficients.

Savitzky-Golay and wavelet transformations were considered in this study as both have been shown to be effective pre-treatments for the analysis of sugarcane quality measures. For example, Savitzky-Golay was effective for both analysis of sugarcane quality measures as shown in Chapter 4 and in a feasibility test for identifying atypical cane samples, as in Chapter 5. Wavelet transformation has been used in sugarcane spectral classification problems (Donald et al., 2006). However, there is little literature on the effectiveness of different basis functions. Daublet, Symmlet and Coiflet transformations were trialled as some of the most common bases used in the wider literature (Singh et al., 2008). SNV-First and SNV-Second transformations applied SNV before SG derivatives were taken. Both cases used a window length of 13 and a second-degree polynomial to estimate the derivatives. The use of a window length of 13 resulted in the loss of 12 variables, reducing the spectral range to 1,112 nm – 2,486 nm at 2nm intervals. SNV-First and SNV-Second transformations were computed using the prospectr package in R (Stevens and Ramirez-Lopez, 2013).

Prior to wavelet transformations, the spectral range was reduced to 1,112 nm – 2,486 nm. The removal of 12 variables resulted in a length of 688 variables, allowing for a four level decomposition and was similar to the loss of range in the derivative spectra. Wavelet transformations were computed using the wavelets package in R (Aldrich, 2013). All wavelet coefficients and the level four scaling coefficients were used as potential predictor variables. Figures 6.3 to 6.7 show the processed spectra for SNV-First (Figure 6.3), SNV-Second (Figure

6.4), D8 (Figure 6.5), LA8 (Figure 6.6) and C6 (Figure 6.7) pre-processing. As with raw spectral data, graphical differences between atypical and typical samples were minimal and difficult to identify for all pre-processing techniques investigated.

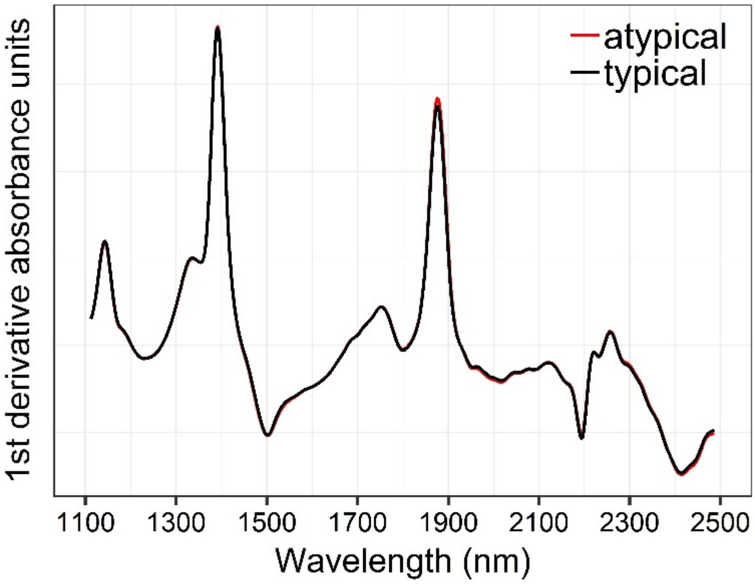

Figure 6.3. Average of SNV-First pre-processed spectra for typical (black) and atypical (red) samples.

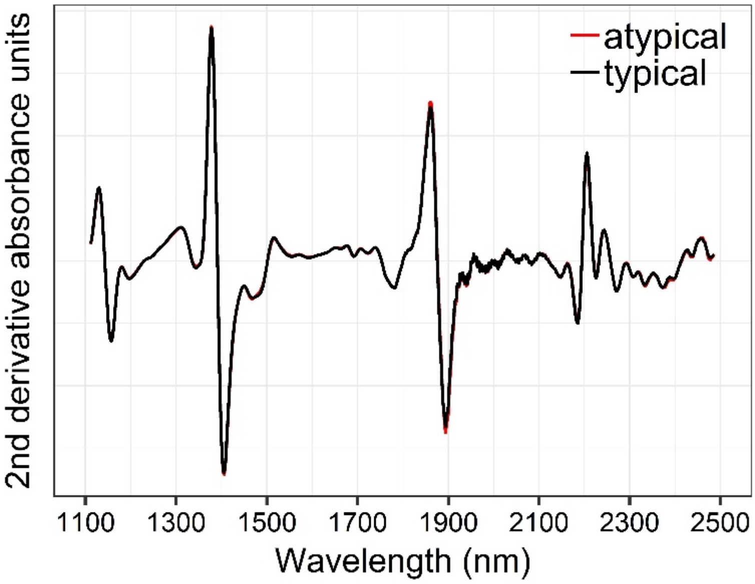

Figure 6.4. Average of SNV-Second pre-processed spectra for typical (black) and atypical (red) samples.

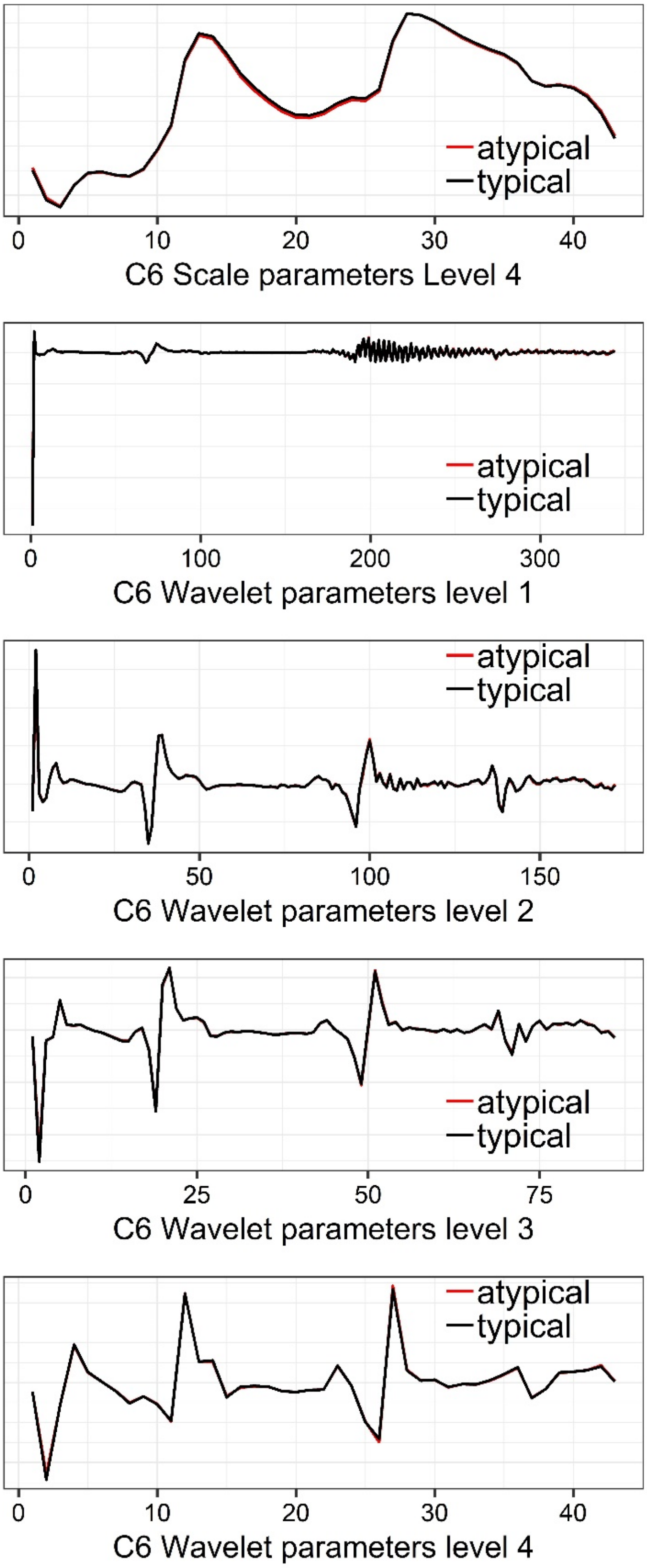

Figure 6.5. Average of Wavelet-C6 pre-processed spectra for typical (black) and atypical (red) samples.

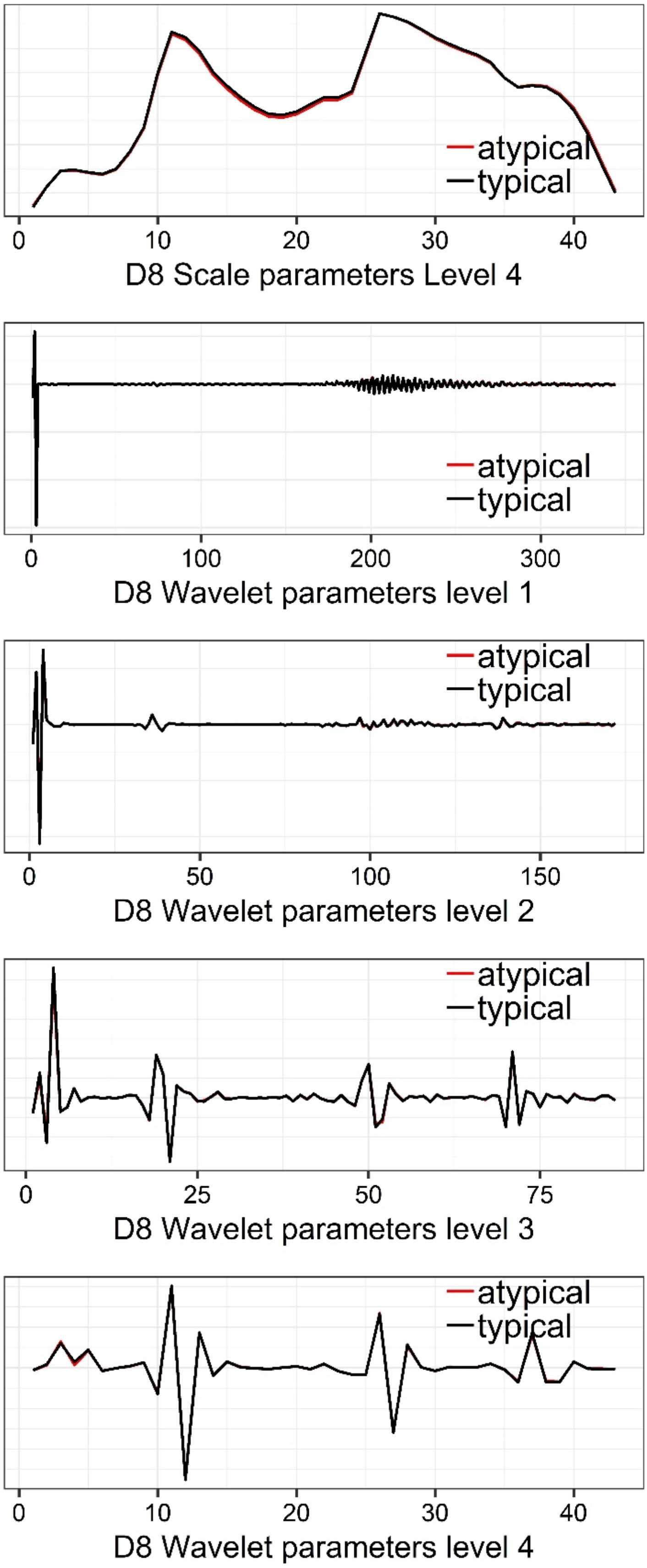

Figure 6.6. Average of Wavelet-D8 pre-processed spectra for typical (black) and atypical (red) samples.

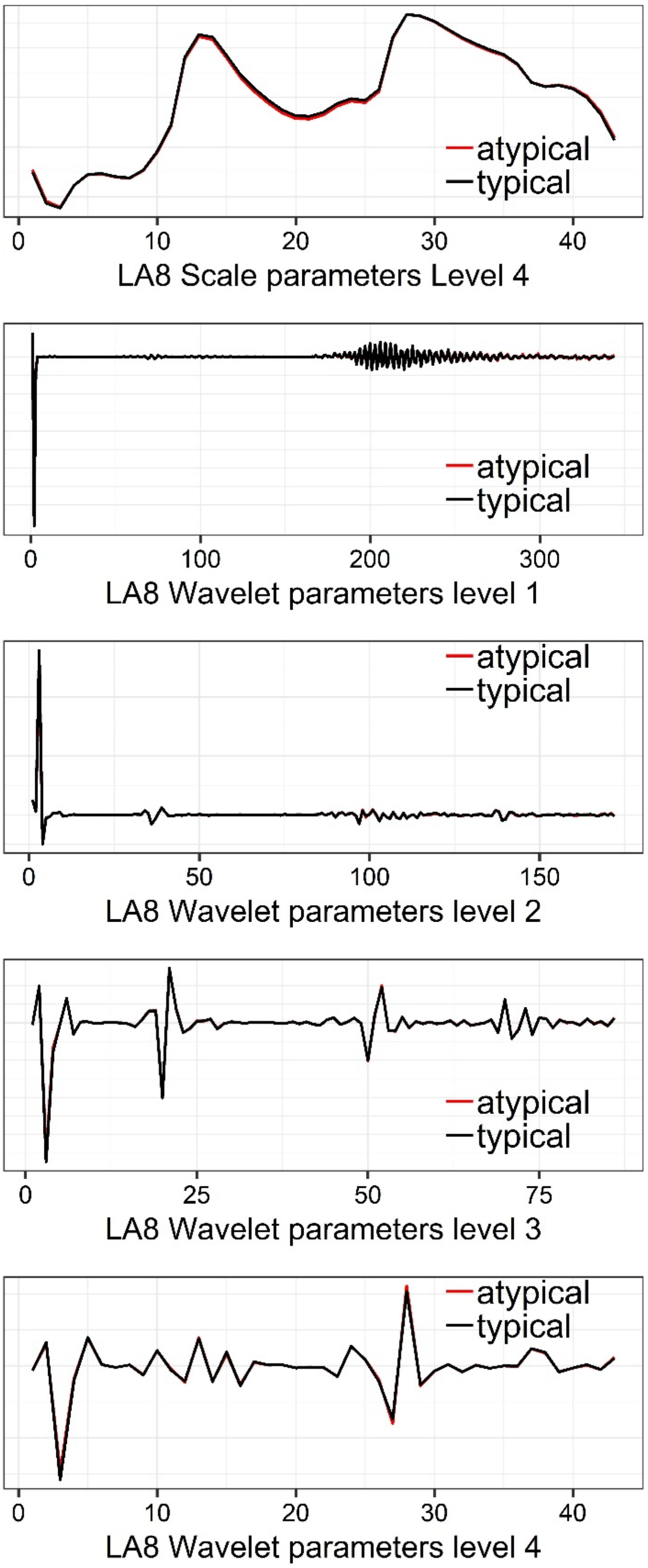

Figure 6.7. Average of Wavelet-LA8 pre-processed spectra for typical (black) and atypical (red) samples.

6.2.3 Model calibration

Five modelling approaches were investigated:

1. LDA: Linear discriminant analysis
2. PLS-DA: Partial least squares for discriminant analysis
3. RF: Random Forests
4. SVM: Support vector machines for classification.
5. ANN: A feed-forward artificial neural network with back-propagation of errors.

These approaches cover a range of modelling philosophies. LDA is simple, linear and is based on Bayesian statistical theory. It is the only method capable of naturally producing a probabilistic prediction. PLS-DA is based on least squares regression and is capable of modelling slight non-linearities, while random forest represents a non-linear, ensemble modelling process. Finally, SVM and ANN are machine-learning algorithms, which approach classification and regression problems from a data driven rather than a statistical perspective. All model building, including cross-validation was managed through the caret package (Kuhn, 2017).

6.2.3.1 Linear discriminant analysis

LDA is a simple technique and by default has no hyper-parameters that require tuning. However, spectral data are often highly correlated. In order to avoid collinearity, pre-processed spectral data were further transformed using Principal Component analysis. The number of PC's retained (*PCs*) was tuned through cross-validation as a hyper-parameter. Models were tuned over 0 – 100 PC's at steps of 4. A 'zero' PC option was included as spectral pre-treatments may effectively remove the need for principal component analysis. LDA modelling was based on the lda function in the MASS package (Venables and Ripley, 2002). Prior probabilities were estimated from the data. In the case of down-sampled data, this prior probability would be 0.50 for both classes. Prior probabilities were considered in the calculation of the predicted posterior probabilities required for AUC. Both class and probability predictions were returned for new samples.

6.2.3.2 Partial least squares discriminant analysis

PLS for discriminant analysis uses the same approach as PLS regression and requires the number of latent variables to be tuned as a hyper-parameter (Barker and Rayens, 2003). The number of latent variables was tuned through cross-validation over a range of 1 – 20. This range was chosen as it was shown to be appropriate in Chapters 3 and 4. PLS-DA models were built using the plsda function in the caret package, which builds on the pls package of Mevik et al. (2015). Within PLS-DA, class labels are replaced by a pair of dummy variables for atypical and typical classes (each sample having a score one in either atypical or typical dummy variable). A multivariate partial least squares regression model is then built to predict the two dependent variables (Martens et al., 1992). Class probabilities were calculated using the softmax function. Class predictions can be considered the class with the highest probability (>0.5).

6.2.3.3 Random Forest

Classification and regression trees (CART) offer a unique way to model complex, non-linear interactions in high dimensional data (Breiman et al., 1984). Random Forest build on this by developing an ensemble of individual trees (Breiman, 2001). Random forests seek to reduce the variance of standard bagging techniques by randomly selecting a subset of variables to test at each split of the tree, making each tree independent of all other trees in the ensemble (Hastie et al., 2013g). Each tree produces a class prediction and the predicted class label was assigned by majority vote across all trees. RF models were built using the randomForest package in R (Liaw and Wiener, 2002). Cross-validation was used to tune the number of variables tried at each split (*mtry*). Twenty values were tested between 2 and all available variables (e.g. for derivative and wavelet data: 2, 38, 74, …, 615, 651, 688). The predicted class probability used in calculating the AUC was estimated as the proportion of votes for each class.

6.2.3.4 Support Vector Machine

Support vector machines are a classification approach from the machine learning community. SVM are data sparse, making use of only a fraction of the available data in the final model and use the ‘kernel trick’ to model non-linearities by projection into a higher dimensional space. Here we use the cost sensitive c-svc approach as implemented within the kernlab package (Karatzoglou et al., 2004). SVM classifiers were built using a radial basis function. The SVM hyper-parameters cost (*cost*) and σ (*sigma*) were tuned through a grid search in the base 2 exponential set. *Cost* values tested were $2^{(-5,-3,-1, 0, 3, 5, 7)}$, while *sigma* values tested were $2^{(-11,-9,-7,-5,-3)}$. Each combination of *cost* and *sigma* were tested through cross-validation and the combination with the highest ROC AUC was chosen. SVM return a prediction between -1 (typical) and +1 (atypical) where sign is used to assign class labels. Pseudo posterior probabilities are estimated using a sigmoidal function (Lin et al., 2007, Platt, 1999). These probabilities were used in calculating the AUC.

6.2.3.5 Artificial Neural Network

ANN models were built as single hidden layer backpropagation networks, with a single output node. ANN models were built using the nnet package (Venables and Ripley, 2002). This package uses a logistic function as the 'activation function' for the hidden nodes and the output node. Internally, entropy (maximum conditional likelihood), rather than squared error was used as the error measure during back-propagation. . The size (number of nodes in the hidden layer) and decay (the decay rate) hyper-parameters were tuned using a grid search. The size parameter was tuned over a range of 2 to 12 at 2 node intervals, while six decay parameter values were tested (0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05). Each combination of size and decay rate were tested through cross-validation and the combination with the highest ROC AUC was chosen. Unlike other models described here, parameters were range-scaled rather than auto-scaled, following the methodology used in Chapter 4. Neural networks produce a probability-like logistic score for the target class (values between 0 and 1), with class labels being assigned as the target class if the score <0.5. The logistic score was used as the probability for the estimation of the AUC.

6.2.4 Feature selection

Genetic algorithm feature selection (GAFS) is a computationally expensive wrapper method for feature selection. As GAFS is based on the idea of genetic evolution, sets of features are retained and mixed from generation to generation, based on some model performance. This results in an iterative improvement in the feature sets. In this study, GAFS was used to improve the PLS-DA model as described in section 2.4.2, with the cross-validated ROC used as the model performance on which selection was based. The GAFS approach was run using the gafs function in caret package (Kuhn and Johnson, 2013a, Scrucca, 2013, Mitchell, 1999). The following default settings were used for the GAFS procedure:

- iters = 100: The maximum number of 'generations',
- popSize = 50: The population size at each iteration,
- pmutation = 0.1: The probability of random mutation,
- pcrossover = 0.8: The probability of crossover,

- elite = 0: The number of the best children to keep into the next generation.

In this setup, the maximum number of parameters (wavelengths or wavelet coefficients) that could be used in any given individual was not restricted. The number of iterations in the GAFS model was tuned through repeated 5-fold cross-validation (5-fold cross-validation is repeated 4 times). Figure 6.8 outlines the GAFS procedure used in this study.

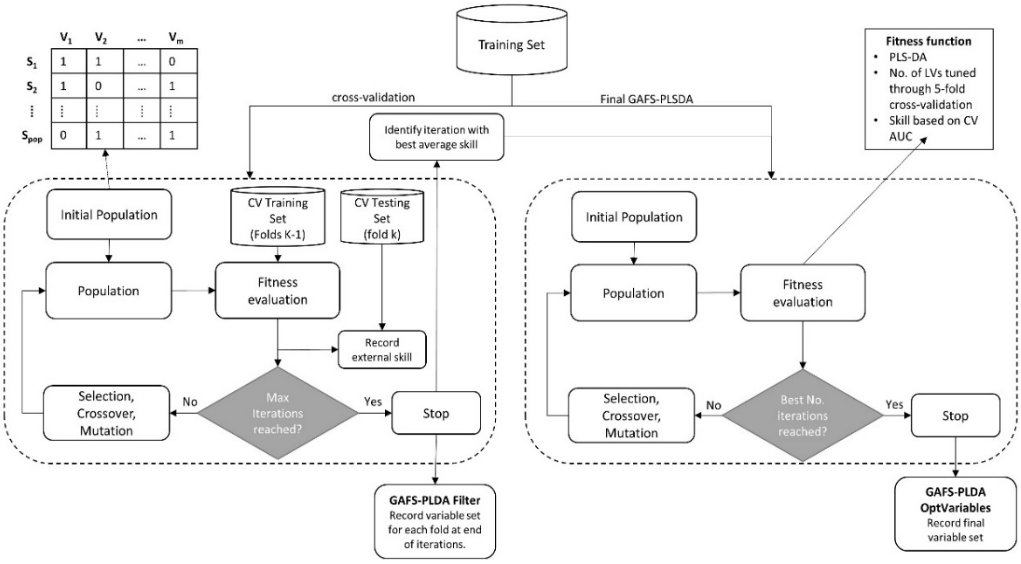

Figure 6.8. Diagram of GAFS-PLSDA feature selection used in this analysis. The GAFS-PLSDA process was performed on the training data set of each of the J = 10 random training/test splits.

Despite tuning the number of iterations, the GAFS procedure can over-fit to the training data set. We investigated the use of the cross-validation results as a filter selection process, in order to reduce the dependence on the specific samples used to train the model. Four selection levels were investigated:

a. OptVariables: A final GAFS model was built using the entire training set and the cross-validated number of iterations, the selected wavelengths were identified for use in building PLS-DA models
b. Imp50: Wavelengths that were selected in > 50% of the 20 cross-validation GAFS models were identified for use in final models,
c. Imp70: Wavelengths that were selected in > 70% of the 20 cross-validation models were identified for use in final models.

d. Imp90: Wavelengths that were selected in > 90% of the 20 cross-validation models were identified for use in final models.

For each ‘selection level’, a PLS-DA model was built following the methodology in section 6.2.3.2, using only the selected wavelengths.

6.3. Results and discussion

6.3.1 Model comparison

All modelling approaches trialled in this study were able to achieve an AUC of greater than 0.9 and an Accuracy of greater than 80% when applied to the validation set using the best pre-processing technique for each model (Table 1). PLS-DA achieved the highest average AUC score for the validation set (AUC = 0.9502). SVM and LDA performed similarly (AUC = 0.9479 and AUC = 0.9465 respectively) while ANN had the lowest validation set AUC (AUC = 0.9154). Similar differences between modelling techniques were also seen in overall model Accuracy, correct classification rates of atypical samples ($CCR_{atypical}$) and correct classification rates of typical samples ($CCR_{typical}$) (Table 6.1). For example, PLS-DA had the highest average Accuracy and CCR scores while ANN had the lowest. The Accuracy shown across all modelling techniques is positive as the majority of samples (at least 83%) could be correctly classified as either typical or atypical. This is consistent with similar research in discriminating deteriorated or agricultural products such as bruised fruits and with spectroscopic classification problems that deal with highly imbalanced data. Importantly, model performance was similar for both atypical and typical cane samples. This suggests that down-sampling was successfully able to promote balanced models that were not biased towards better performance of the majority class.

Table 6.1. Average model skill for each modelling approach using the best combination of pre-processing and feature selection. The average is taken across 10 training/validation splits each with 10 random initializations. The model with the highest overall skill (AUC) is highlighted in bold font. Tuning parameter values recorded as the mode(min, max).

Model	Pre-processing	Tuning	Training				Validation			
			AUC	$CCR_{atypical}$	$CCR_{typical}$	Accuracy	AUC	$CCR_{atypical}$	$CCR_{typical}$	Accuracy
LDA	SNV-First	PCs: 72 (36, 100)	0.9709	94.93%	88.70%	88.86%	0.9465	88.11%	88.34%	88.33%
PLS-DA	**SNV-First**	**LVs: 14 (12, 20)**	**0.9716**	**94.49%**	**89.14%**	**89.27%**	**0.9502**	**88.65%**	**88.75%**	**88.75%**
RF	SNV-Second	mtry: 218 (74-688)	0.9976	100.00%	84.37%	84.76%	0.9260	85.46%	83.99%	84.03%
SVM	SNV-Second	sigma: 2^{-11} (2^{-11}, 2^{-11}) cost: 2^{3} (2^{3}, 2^{7})	0.9881	98.96%	87.99%	88.26%	0.9479	88.96%	87.60%	87.63%
ANN	SNV-First	nodes: 2 (2, 12) decay: $5*10^{-4}$ (10^{-4}, 10^{-3})	0.9636	99.06%	83.99%	84.38%	0.9154	85.27%	83.66%	83.70%

The relative performance of the tested modelling techniques may be somewhat surprising given previous comparison studies. When comparing a range of modelling techniques to classify fuels, Balabin et al. (2010) identified SVM as a highly effective classifier, PLS and ANN as medium effective classifiers and LDA as one of the least effective classifiers. In our study, PLS-DA was the most effective performer while LDA performed similarly to SVM. The relatively good performance of PLS-DA and LDA may suggest that the classification problem was relatively linear, such that the ability to model nonlinearity inherent in SVM did not provide an advantage. Similar comparisons were found for estimating the cane quality parameter CCS (Commercial Cane Sugar) in Chapter 4. PLS-DA and LDA may also have had an advantage over SVM as PLS-DA and LDA both made use of feature extraction techniques (Latent variables in PLS-DA and PCA applied during the LDA process).

Interestingly, although all modelling techniques were improved by the use of some pre-processing technique, the advantage was not always large (Figure 6.9). RF with SNV-Second had the largest increase compared to Raw spectral data, while ANN with SNV-First pre-processing only slightly improved model performance. Figure 6.9 also shows that the use of Wavelet pre-processing could actually lead to decreased model performance compared to Raw spectral data. The interaction between pre-processing and modelling techniques shown in Figure 6.9 highlight the importance of using a pre-processing technique that is appropriate for the modelling technique especially during comparisons of model skill.

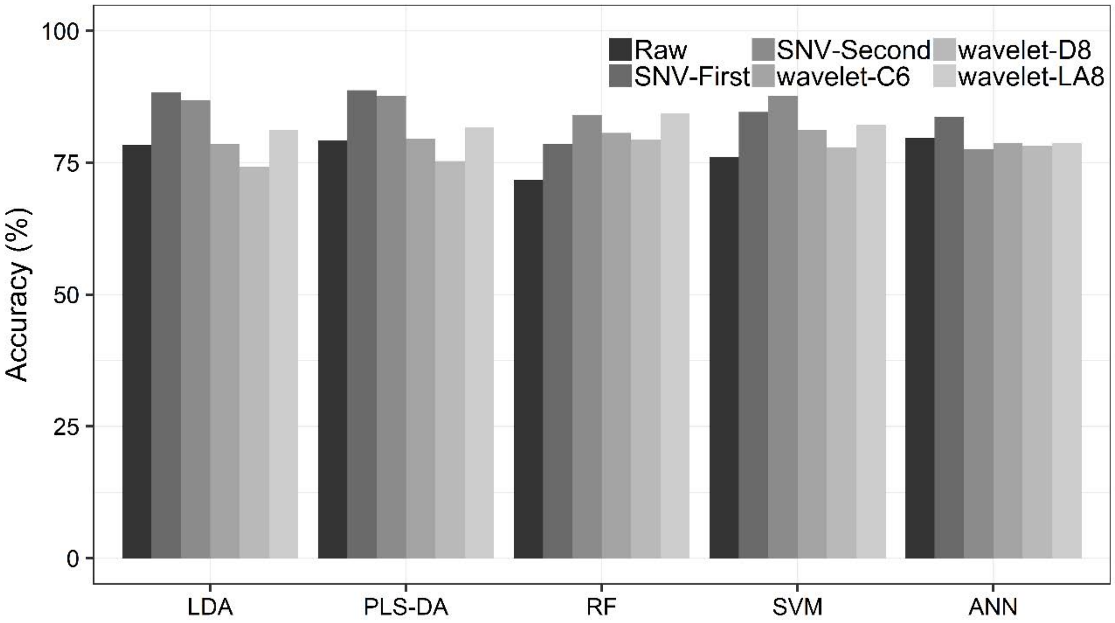

Figure 6.9. Accuracy (correct classification rate of all samples) for each combination of spectral pre-processing and modelling technique. Bars represent the average of all model runs.

Pre-processing techniques such as spectral derivatives and wavelet transformation attempt to highlight information otherwise hidden in the spectral data. Derivation of absorption spectra can help resolve overlapping absorption peaks from different sample constituents (Osborne et al., 1993a). This increased resolution can make it easier to identify the presence of specific compounds and can often lead to simpler more robust models (Agelet and Hurburgh, 2010). For example, PLS-DA required half as many latent variables using SNV-Second compared to Raw spectral data (Table 6.2). Table 6.2 contains the calibration results for all combinations of models and spectral pre-processing, while Table 6.3 contains the validation results. Re-running these analyses using only SNV resulted in model performance similar to Raw spectral data. This suggested that the spectral derivative stage was largely responsible for the improved performance seen using SNV-First and fewer latent variables using SNV-Second (data not shown).

Table 6.2. Average model skill for each modelling approach / pre-processing combination during the calibration phase. The average is taken across 10 train/validation splits each with 10 random initializations. Bracketed skill values are the interquartile range (Q3 – Q1) and represent a robust measure of spread across all individual model runs. Shading represents the best pre-processing for each modelling approach. The model with the highest overall skill (AUC) is highlighted in bold font. Tuning parameter values recorded as the mode (min, max).

Model	Pre-processing	Tuning	AUC	CCR$_{atypical}$	CCR$_{typical}$	Accuracy
LDA	Raw	PCs: 52 (28, 100)	0.9096	88.33%	78.70%	78.94%
			(0.0163)	(4.32%)	(1.78%)	(1.61%)
	SNV-First	PCs: 72 (36, 100)	0.9709	94.93%	88.70%	88.86%
			(0.0068)	(2.47%)	(1.46%)	(1.38%)
	SNV-Second	PCs: 40 (28, 100)	0.9570	92.86%	87.09%	87.24%
			(0.0085)	(3.09%)	(1.53%)	(1.42%)
	wavelet-C6	PCs: 84 (40, 100)	0.9176	89.33%	78.81%	79.08%
			(0.0149)	(3.70%)	(2.10%)	(2.11%)
	wavelet-D8	PCs: 92 (56, 100)	0.8938	87.37%	74.61%	74.93%
			(0.0194)	(3.24%)	(2.25%)	(2.16%)
	wavelet-LA8	PCs: 84 (40, 100)	0.9330	90.29%	81.53%	81.75%
			(0.0149)	(3.70%)	(2.51%)	(2.54%)
PLS-DA	Raw	LVs: 15 (13, 20)	0.9222	91.04%	79.52%	79.81%
			(0.0164)	(3.86%)	(1.80%)	(1.74%)
	SNV-First	**LVs: 14 (12, 20)**	**0.9716**	**94.49%**	**89.14%**	**89.27%**
			(0.0060)	**(2.47%)**	**(1.51%)**	**(1.51%)**
	SNV-Second	LVs: 7 (5, 9)	0.9645	94.17%	88.04%	88.20%
			(0.0081)	(2.01%)	(1.74%)	(1.71%)
	wavelet-C6	LVs: 7 (5, 8)	0.9474	94.78%	80.04%	80.42%
			(0.0242)	(2.47%)	(2.10%)	(2.04%)
	wavelet-D8	LVs: 7 (4, 7)	0.9355	94.85%	75.97%	76.45%
			(0.0313)	(5.56%)	(2.34%)	(2.17%)
	wavelet-LA8	LVs: 5 (5, 7)	0.9569	95.19%	82.04%	82.38%
			(0.0180)	(3.86%)	(2.07%)	(1.89%)
SVM	Raw	sigma: 2^{-11} (2^{-11}, 2^{-9})	0.8662	81.15%	76.61%	76.72%
		cost: 2^{7} (2^{3}, 2^{7})	(0.0271)	(4.48%)	(4.17%)	(3.97%)
	SNV-First	sigma: 2^{-11} (2^{-11}, 2^{-11})	0.9704	98.05%	85.04%	85.37%
		cost: 2^{5} (2^{5}, 2^{7})	(0.0104)	(2.47%)	(1.99%)	(1.91%)
	SNV-Second	sigma: 2^{-11} (2^{-11}, 2^{-11})	0.9881	98.96%	87.99%	88.26%
		cost: 2^{3} (2^{3}, 2^{7})	(0.0054)	(1.85%)	(1.65%)	(1.57%)
	wavelet-C6	sigma; 2^{-11} (2^{-11}, 2^{-9})	0.9913	99.69%	81.64%	82.10%

Model	Pre-processing	Tuning	AUC	CCR$_{atypical}$	CCR$_{typical}$	Accuracy
		cost: 2^3 (2^3, 2^7)	(0.0052)	(0.62%)	(2.43%)	(2.36%)
	wavelet-D8	sigma: 2^{-11} (2^{-11}, 2^{-9})	0.9923	99.91%	78.47%	79.02%
		cost: 2^3 (2^3, 2^7)	(0.0071)	(0.00%)	(2.36%)	(2.30%)
	wavelet-LA8	sigma: 2^{-11} (2^{-11}, 2^{-9})	0.9944	99.88%	82.64%	83.08%
		cost: 2^3 (2^3, 2^7)	(0.0027)	(0.00%)	(2.20%)	(2.13%)
ANN	Raw	nodes: 4 (2, 8)	0.9369	97.09%	80.09%	80.52%
		decay: 5*10^{-4} (10^{-4}, 5*10^{-3})	(0.0177)	(0.15%)	(3.05%)	(2.97%)
	SNV-First	nodes: 2 (2, 12)	0.9636	99.06%	83.99%	84.38%
		decay: 5*10^{-4} (10^{-4}, 10^{-3})	(0.0103)	(1.23%)	(1.70%)	(1.68%)
	SNV-Second	nodes: 6 (2, 12)	0.9510	99.98%	78.20%	78.75%
		decay: 10^{-4} (10^{-4}, 10^{-4})	(0.0094)	(0.00%)	(3.04%)	(2.97%)
	wavelet-C6	nodes: 4 (2, 8)	0.9385	98.44%	79.19%	79.68%
		decay: 5*10^{-4} (10^{-4} , 5*10^{-3})	(0.0138)	(0.62%)	(2.96%)	(2.84%)
	wavelet-D8	nodes: 4 (2 - 12)	0.9228	95.46%	78.67%	79.10%
		decay: 5*10^{-4} (10^{-4}, 10^{-2})	(0.0147)	(0.15%)	(2.83%)	(2.76%)
	wavelet-LA8	nodes: 4 (2 - 10)	0.9330	97.76%	79.15%	79.63%
		decay: 5*10^{-4} (10^{-4}, 5*10^{-3})	(0.0139)	(0.00%)	(2.77%)	(2.62%)
RF	Raw	mtry: 38 (2-700)	0.9664	100.0%	72.89%	73.58%
			(0.0108)	(0.00%)	(3.62%)	(3.53%)
	SNV-First	mtry: 38 (2-688)	0.9896	100.0%	79.01%	79.54%
			(0.0044)	(0.00%)	(2.67%)	(2.60%)
	SNV-Second	mtry: 218 (74-688)	0.9976	100.0%	84.37%	84.76%
			(0.0011)	(0.00%)	(3.94%)	(3.84%)
	wavelet-C6	mtry: 507 (110-688)	0.9961	100.0%	80.96%	81.45%
			(0.0030)	(0.00%)	(2.77%)	(2.70%)
	wavelet-D8	mtry: 507 (38-688)	0.9953	100.0%	79.64%	80.16%
			(0.0032)	(0.00%)	(3.35%)	(3.27%)
	wavelet-LA8	mtry 218 (74-688)	0.9969	100.0%	84.65%	85.04%
			(0.0030)	(0.00%)	(2.90%)	(2.83%)

Table 6.3. Average model skill for each modelling approach / pre-processing combination for the validation stage. The average is taken across 10 train/validation splits each with 10 random initializations. Bracketed skill values are the interquartile range (Q3 – Q1) and represent a robust measure of spread across all individual model runs. Shading represents the best pre-processing for each modelling approach. The model with the highest overall skill (AUC) is highlighted in bold font. Tuning parameter values recorded as the mode (min, max).

Model	**Pre-processing**	**Tuning**	**AUC**	**$\text{CCR}_{\text{atypical}}$**	**$\text{CCR}_{\text{typical}}$**	**Accuracy**
LDA	Raw	PCs: 52 (28, 100)	0.8648	78.77%	78.41%	78.42%
			(0.0140)	(4.63%)	(2.03%)	(1.97%)
	SNV-First	PCs: 72 (36, 100)	0.9465	88.11%	88.34%	88.33%
			(0.0124)	(4.48%)	(1.29%)	(1.31%)
	SNV-Second	PCs: 40 (28, 100)	0.9330	86.73%	86.85%	86.84%
			(0.0136)	(4.32%)	(1.59%)	(1.56%)
	wavelet-C6	PCs: 84 (40, 100)	0.8642	79.28%	78.56%	78.58%
			(0.0124)	(5.09%)	(2.20%)	(2.04%)
	wavelet-D8	PCs: 92 (56, 100)	0.8192	74.95%	74.21%	74.23%
			(0.0212)	(5.71%)	(2.79%)	(2.76%)
	wavelet-LA8	PCs: 84 (40, 100)	0.8860	80.96%	81.23%	81.22%
			(0.0127)	(4.94%)	(2.43%)	(2.27%)
PLS-DA	Raw	LVs: 15 (13, 20)	0.8745	80.12%	79.18%	79.2%
			(0.0149)	(4.32%)	(2.08%)	(2.01%)
	SNV-First	**LVs: 14 (12, 20)**	**0.9502**	**88.65%**	**88.75%**	**88.75%**
			(0.0116)	**(3.70%)**	**(1.69%)**	**(1.63%)**
	SNV-Second	LVs: 7 (5, 9)	0.9386	87.32%	87.65%	87.65%
			(0.0114)	(4.32%)	(1.96%)	(1.95%)
	wavelet-C6	LVs: 7 (5, 8)	0.8754	80.47%	79.6%	79.62%
			(0.0188)	(4.32%)	(2.13%)	(2.01%)
	wavelet-D8	LVs: 7 (4, 7)	0.8344	76.31%	75.31%	75.34%
			(0.0168)	(5.09%)	(2.58%)	(2.45%)
	wavelet-LA8	LVs: 5 (5, 7)	0.8913	81.84%	81.69%	81.69%
			(0.0178)	(4.94%)	(1.89%)	(1.92%)
SVM	Raw	sigma: 2^{-11} (2^{-11}, 2^{-9})	0.8236	72.97%	76.11%	76.03%
		cost: 2^{7} (2^{3}, 2^{7})	(0.0264)	(7.56%)	(4.06%)	(3.91%)
	SNV-First	sigma: 2^{-11} (2^{-11}, 2^{-11})	0.9221	85.94%	84.59%	84.63%
		cost: 2^{5} (2^{5}, 2^{7})	(0.0132)	(4.94%)	(2.10%)	(1.97%)
	SNV-Second	sigma: 2^{-11} (2^{-11}, 2^{-11})	0.9479	88.96%	87.60%	87.63%
		cost: 2^{3} (2^{3}, 2^{7})	(0.0127)	(4.32%)	(1.87%)	(1.87%)
	wavelet-C6	sigma; 2^{-11} (2^{-11}, 2^{-9})	0.8924	81.51%	81.17%	81.18%

Model	Pre-processing	Tuning	AUC	CCR$_{atypical}$	CCR$_{typical}$	Accuracy
		cost: 2^{3} (2^{3}, 2^{7})	(0.0149)	(4.32%)	(2.33%)	(2.27%)
	wavelet-D8	sigma: 2^{-11} (2^{-11}, 2^{-9})	0.8609	78.42%	77.87%	77.89%
		cost: 2^{3} (2^{3}, 2^{7})	(0.0169)	(4.32%)	(2.36%)	(2.15%)
	wavelet-LA8	sigma: 2^{-11} (2^{-11}, 2^{-9})	0.9000	82.26%	82.17%	82.17%
		cost: 2^{3} (2^{3}, 2^{7})	(0.0132)	(4.32%)	(2.03%)	(1.91%)
ANN	Raw	nodes: 4 (2, 8)	0.8725	80.02%	79.69%	79.70%
		decay: $5*10^{-4}$ (10^{-4}, $5*10^{-3}$)	(0.0237)	(4.94%)	(3.35%)	(3.25%)
	SNV-First	nodes: 2 (2, 12)	0.9154	85.27%	83.66%	83.70%
		decay: $5*10^{-4}$ (10^{-4}, 10^{-3})	(0.0126)	(4.32%)	(1.86%)	(1.79%)
	SNV-Second	nodes: 6 (2, 12)	0.8557	78.46%	77.51%	77.54%
		decay: 10^{-4} (10^{-4}, 10^{-4})	(0.0280)	(4.32%)	(2.76%)	(2.72%)
	wavelet-C6	nodes: 4 (2, 8)	0.8769	81.99%	78.67%	78.75%
		decay: $5*10^{-4}$ (10^{-4} , $5*10^{-3}$)	(0.0228)	(5.56%)	(2.55%)	(2.39%)
	wavelet-D8	nodes: 4 (2 - 12)	0.8639	79.87%	78.14%	78.19%
		decay: $5*10^{-4}$ (10^{-4}, 10^{-2})	(0.0287)	(4.94%)	(2.98%)	(2.76%)
	wavelet-LA8	nodes: 4 (2 - 10)	0.8707	80.93%	78.60%	78.66%
		decay: $5*10^{-4}$ (10^{-4}, $5*10^{-3}$)	(0.0233)	(5.09%)	(3.08%)	(2.90%)
RF	Raw	mtry: 38 (2-700)	0.7850	70.41%	71.86%	71.82%
			(0.0250)	(7.56%)	(3.78%)	(3.51%)
	SNV-First	mtry: 38 (2-688)	0.8847	81.24%	78.54%	78.61%
			(0.0203)	(4.32%)	(3.22%)	(2.92%)
	SNV-Second	mtry: 218 (74-688)	0.9260	85.46%	83.99%	84.03%
			(0.0154)	(4.32%)	(3.60%)	(3.41%)
	wavelet-C6	mtry: 507 (110-688)	0.9009	83.43%	80.57%	80.64%
			(0.0243)	(4.94%)	(2.82%)	(2.79%)
	wavelet-D8	mtry: 507 (38-688)	0.8765	80.64%	79.30%	79.33%
			(0.0247)	(4.48%)	(3.19%)	(2.99%)
	wavelet-LA8	mtry: 218 (74-688)	0.9224	84.83%	84.34%	84.35%
			(0.0174)	(4.94%)	(3.36%)	(3.21%)

The higher resolution afforded by SNV-Second pre-processing may explain why SNV-Second produced the best performance for the RF, SVM and ANN models but not the LDA and PLS-DA models. The ability to better separate overlapping peaks may have provided less advantage for LDA and PLS-DA models which both made use of a separate feature extraction technique. These results reinforce the idea that complex modelling techniques may not always offer improvements in model skill. The relatively high performance of PLS-DA is especially encouraging as PLS is widely used within the sugarcane industry and widely available in commercial spectral analysis packages which would reduce the barriers to implementation of a classification system within the Australian sugarcane industry.

6.3.2 Feature selection

As PLS-DA with SNV-First pre-processing was identified as the most effective modelling strategy where all spectral data were used, we looked at feature selection as a possible method of further improving PLS-DA model performance. The genetic algorithm feature selection (GAFS) approach described in section 6.2.4 was applied to the PLS-DA modelling process (GAFS-PLSDA). Figure 6.10 identifies the features selected by the final GAFS-PLSDA model (OptVariables) as well as the features selected using the cross-validation runs (e.g. Imp50 = features selected in >50% of cross-validation runs). Wavelengths above 1,900 nm were selected less often and may be less important in identifying atypical samples. The most commonly selected wavelengths were centred on 1,200 nm, 1,400 nm and 1,700 nm (Figure 6.10). These three regions can all be associated with C-H stretching overtones or combinations. First overtone of O-H stretching, around 1,400 nm may also relate to various sugars such as sucrose (1,440 nm) and glucose (1,480 nm) or moisture (1,450 nm).

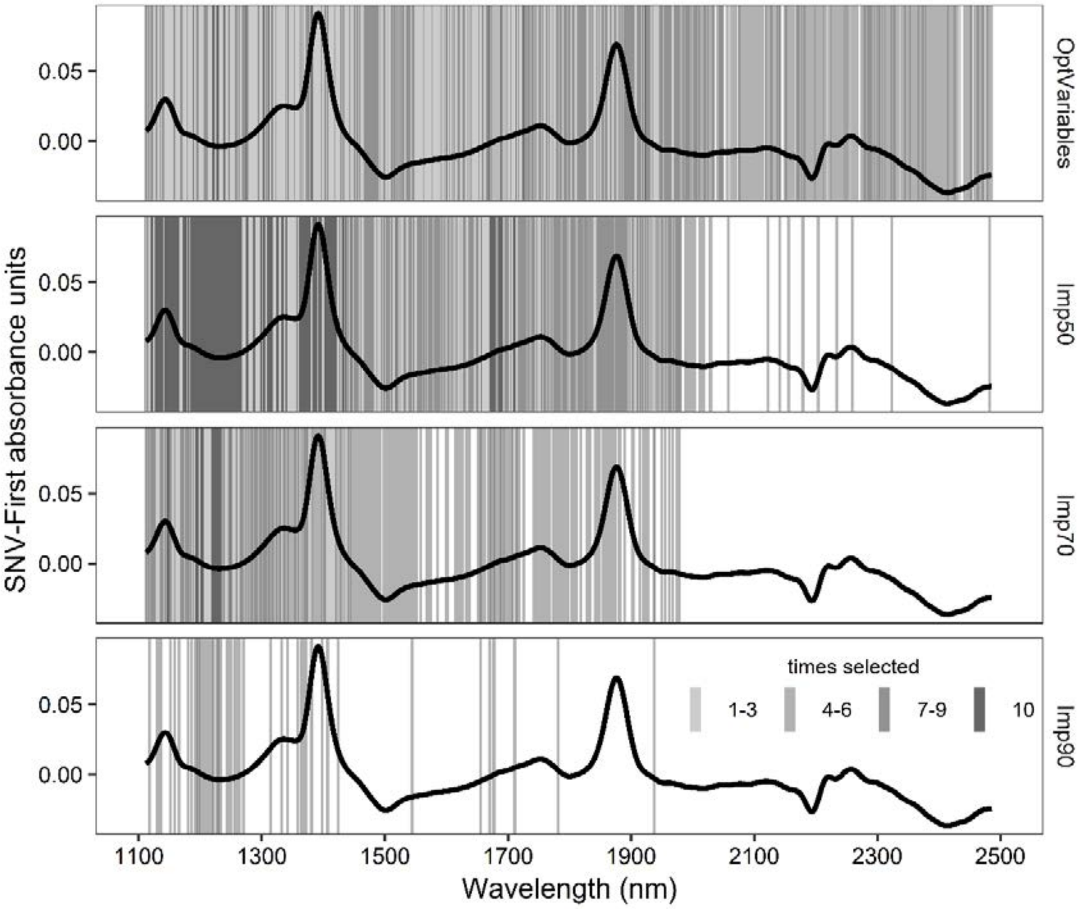

Figure 6.10. Wavelengths selected by GAFS-PLSDA for SNV- First derivative spectra. OptVariables represent features selected by the final GAFS-PLSDA model while Imp50 – Imp90 represent the features that were selected across cross-validation runs. For example Imp50 = features selected in >50% of cross-validation runs. Thick black line represents the average spectral signature. Horizontal bars represent the selected wavelengths. Bars are shaded to represent the number of training sets for which each variable was selected.

The use of GAFS-PLSDA improved model performance for the PLS-DA SNV-First modelling approach (Table 6.4). Using the cross-validation results appears to have had the desired effect of improving test set performance as the largest improvement was obtained using the GAFS-Imp50 approach. By only including features that were selected in more than 50% of cross-validation runs, accuracy improved by 1.82% relative to no feature selection being used, while $CCR_{atypical}$ improved by 2.21% and $CCR_{typical}$ improved by 1.81%. Using the GAFS-Imp50 approach, removed many wavelengths above 1,900 nm, which is known to be a noisier region and may have caused overfitting to the training data set. The more rigorous filter of GAFS-Imp70 and GAFS-Imp90 reduced model performance (Table 6.4). While these methods may have helped identify influential features, the heavy-handed filtering may have broken down relationships between features that were otherwise captured by the genetic algorithm procedure, reducing the effectiveness of the PLS-DA model.

The increase in performance using GAFS for PLS-DA models was relatively modest. The effectiveness of feature selection for PLS-DA may be low due to the inherent feature extraction included in PLS modelling. Latent variables are designed to explain the variance in feature space (spectral data) with respect to the response variable (class labels). This means that non-informative variables should be down-weighted, such that removing them may not improve model performance.

Table 6.4. Average model skill for PLS-DA using SNV-First and GAFS feature selection. The average is taken across 10 train/test splits each with 10 random initializations. The model with the highest overall skill (AUC) is highlighted in bold font. Tuning parameter values recorded as the mode(min - max).

Feature Selection	**Tuning**	**Train**				**Test**			
		AUC	**CCR$_{atypical}$**	**CCR$_{typical}$**	**Accuracy**	**AUC**	**CCR$_{atypical}$**	**CCR$_{typical}$**	**Accuracy**
none	LVs: 14 (12, 20)	0.9716	94.49%	89.14%	89.27%	0.9502	88.65%	88.75%	88.75%
OptVariables	LVs: 14 (12-16)	0.9709	95.04%	90.10%	90.22%	0.9560	89.95%	89.75%	89.76%
Imp50	LVs: **14 (10-16)**	**0.9716**	**94.54%**	**90.63%**	**90.73%**	0.**9619**	**90.65%**	**90.39%**	**90.40%**
Imp70	LVs: 13 (10-18)	0.9976	91.46%	89.58%	89.63%	0.9492	87.56%	89.49%	89.44%
Imp90	LVs: 7 (2-16)	0.9881	69.69%	69.11%	69.12%	0.7243	65.54%	69.10%	69.01%

6.3.3 Limitations and future research

The results of this study showed that NIR spectroscopy could be used to develop models capable of correctly identifying atypical and typical sugarcane samples with equal skill. In particular, the combination of PLS-DA and SNV-First pre-processing resulted in greater than 88% correct classification rate for both atypical and typical samples and greater than 90% when a feature selection stage was used. However, the large imbalance in classes means that even though classification rates were equal, the model still identifies more typical samples as atypical than correctly identified atypical samples (Table 6.5). Table 6.5 shows the confusion matrix for the PLS-DA model using SNV-First pre-processing. On average many more samples are incorrectly classified as atypical (FP) than are correctly classified (TP), despite the high sensitivity (CCR$_{atypical}$).

Table 6.5. Confusion matrix for PLS-DA using SNV-First. Values represent number of samples and are averaged across all model runs. N represents the total number of samples in the validation set. Values in dashed boxes represent sub-totals. TN, FN, FP and TP represent the true negative, false negative, false positive and true positive rates respectively. The same number of typical and atypical samples were present in each model run.

		Predicted		
	N = 6399	Typical	Atypical	
Observed	Typical	TN = 5535.37	FP = 701.63	6237
	Atypical	FN = 18.38	TP =143.62	162
		5553.75	845.25	

In a practical process control endeavour, it may be more beneficial to increase the accuracy of typical samples at the expense of fewer atypical samples being identified. For example, the boundary between 'atypical' and 'typical' is unlikely to be a hard cut-off. This means that some samples currently predicted as 'atypical' may be only slightly deteriorated such that no change in processing is required. The lack of clear definition between atypical and typical was a challenge and potential limitation in this project. To address the issue, future research should consider having models return a probability that a sample is atypical. The probability that a sample is atypical would provide a convenient tuneable cut-off for process control. A next step may also include a more in-depth analysis of the samples that are incorrectly classified this may give a clearer picture of the differences between typical and atypical samples. Unfortunately, this was outside the scope of the current work.

Future research will also need to consider model performance across years and model transfer between NIR systems. Although data from four years was included in this study, data from all years was evenly divided between calibration and validation sets. Year to year variability would affect model performance on data from years not included in the model building process (Guo et al., 2017; Hong et al., 2019; Shetty et al., 2012). One method to overcome this is to update the model to include samples with reference (laboratory) values from the current year (Hong et al., 2019, Huang et al., 2016, Shetty et al., 2012). However, this would require more samples analysed in the laboratory and could prove expensive or inefficient. An alternative may be the use of semi-supervised or 'active learning' to update models using samples without reference data (Gujral et al., 2011, Guo et al., 2017, Nikzad-Langerodi et al., 2018). For example, Guo et al. (2017) were able to improve variety classification of maize seeds based on hyperspectral imaging using a pre-labelling method. Improved classification was achieved by adding selected new samples and their predicted class to the training set and then retraining the model.

The results presented here highlight the importance of considering the influences of pre-processing on comparisons of modelling techniques. Different pre-processing approaches were required to achieve better performance for different modelling techniques. A more subtle effect was the effect of pre-processing on hyper-parameter tuning. Partial least squares is often used for NIR analysis within agricultural industries. In order to build robust models, simple models are often sought. In this case, second derivative or wavelet transformations may have been a more appropriate pre-processing technique. Future model comparisons should consider that the reverse could also be true. The range of valid hyper-parameter values for model tuning may differ depending on the pre-processing approach used.

6.4 Conclusions

Chapter 6 investigated the classification of 'atypical' sugarcane samples in a large online cane analysis dataset. The variability of an online system and the large imbalance in class sizes were particular difficulties faced in this analysis. Despite these challenges, a combination of PLS-DA, SNV-First derivative transform and down-sampling resulted in a well-balanced discriminative model. This methodology can be used to develop a discriminative model that identifies all samples as atypical or typical samples not just those analysed within the laboratory. The methodology can also be used to develop process control logic that allows atypical samples to be treated separately if needed. The relatively high performance of PLS-DA is particularly promising as PLS approaches are already used within the sugarcane industry and are simple and fast to develop and update. Here we have described an initial estimation of 'atypical' and developed a framework for developing discriminatory models. In future, the sugarcane industry will need to investigate the differences between atypical and typical cane. The methodology outlined here could easily be adapted to any new definition of 'atypical' or other highly imbalanced classification problem such as variety discrimination.

6.5 Chapter 6 Summary

In any given season, thousands of tonnes of sugarcane with atypically low quality can pass undocumented through Australian sugarcane mills. This cane can negatively affect mill processes and throw off grower payment calculations. Mill laboratory operators often observe a small subset (1% - 5%) of cane consignments that have an unusually low juice Pol (Pij; a measure of sucrose content) relative to juice brix (Bij; a measure of dissolved sugars), that can indicate deteriorated or contaminated cane. Many mills only test a small subset of cane in the laboratory, with the majority of consignments analysed using fast near infrared (NIR) spectroscopic techniques. This chapter compared five modelling approaches: Linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), random forest (RF), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Model performance was reported as the correct classification rate (CCR) of typical and atypical samples based on independent test sets. The best performance was achieved by PLS-DA ($CCR_{atypical}$ = 88.65% and $CCR_{typical}$ = 88.75%), while ANN had the lowest performance ($CCR_{atypical}$ = 85.27% and $CCR_{typical}$ = 83.66%). PLS-DA accuracy modestly increased if wavelengths were filtered based on genetic algorithm feature selection ($CCR_{atypical}$ = 90.39% and $CCR_{typical}$ = 90.65%).

The focus of Chapter 6 was Objective 2 of the thesis: Investigating the use of NIR spectroscopic analysis for the automatic identification of atypical cane samples. The results of Chapter 6 echoed the results of Chapter 4, showing that the simpler PLS based approach was as or more effective than more complex machine learning approaches such as SVM and ANN. Results also showed that some spectral pre-processing approaches were more effective for certain modelling approaches and that feature selection could improve model performance. The most important result was the ability to discriminate between atypical and typical cane samples using PLS-DA, given that PLS approaches are well understood within industry. The methodology used in this chapter could be used to identify atypical consignments allowing mills to track occurrences to farms and if necessary develop process control operations for atypical cane. Furthermore, the use of a relatively simple modelling technique such as PLS-DA means model updates can be made efficiently and with confidence as PLS is already well established within the industry.

The outcomes of Chapter 6 were also an important contribution to the literature to emphasise the importance of testing data pre-processing and how calibration data is set-up, rather than only testing a range of modelling approaches. The PLS-DA modelling process developed in Chapter 6 was used in the development of a process-based modelling framework for estimating cane quality parameters in Chapter 7.

Chapter 7

Correcting NIR estimates of cane quality for atypical cane samples in an online analysis system

Relevant publication	NA
Statement of intellectual input	The research question / objective of Chapter 7 was developed by the candidate with input from Dr. Everingham and Mr. Staunton. Data for the thesis was provided by SRA through Mr. Staunton. Data for the thesis was provided by SRA through Mr. Staunton. Dr. Everingham, Dr. Donald, Mr. Staunton and Dr. White supplied editorial assistance. The candidate developed the methodological framework and ran all simulations. The candidate was also responsible for the write-up of the chapter and produced all tables and figures.
Publication status	In preparation

7.1 Introduction

For over a decade near infrared (NIR) spectroscopic analysis has played a crucial role in the sugarcane industry as a rapid and inexpensive method for estimating cane quality parameters such as Brix in juice (Bij), Pol in juice (Pij) and Commercial Cane Sugar (CCS). The importance of quality estimation cannot be understated as measures such as CCS are often the primary measure on which grower payments are determined (Pollock et al., 2007). However, the potential of NIR as a process control tool is often overlooked within the Australian sugarcane industry (Simpson et al., 2011).

NIR spectroscopic analysis for quality monitoring has been used online in Australian sugarcane mills for close to two decades. The first online systems were Cane Analysis Systems (CAS) and were designed to estimate quality measures from shredded cane (Staunton et al., 2004, Staunton et al., 1999). These systems were developed to estimate quality measures such as Bij, Pij and CCS(Staunton et al., 2004). But also considered other constituents that can affect millability such as ash, fibre and dry matter percentage in cane, as well as elemental constituents such as nitrogen, potassium, calcium and magnesium (Staunton et al., 1999). Later online systems were developed to assess similar measures for raw sugar (Bevin et al., 2002) and bagasse (Staunton and Wardrop, 2006), with inroads being made towards assesment for mill mud andother mill byproducts (Keeffe, 2013, Ostatek-Boczynski et al., 2013, Purcell et al., 2012). The ability to continuously monitor these properties throughout the milling processes has the potential to be beneficial by allowing any sudden changes in quality parameters to be quickly identified. While there are a range of published results focusing on the accuracy of NIR analysis there are very little available literature on how this data is used by the sugar industry.

In a review of process control within the Australian sugarcane industry Simpson et al. (2011) suggested

> *" …the real power of NIR technology is not solely in the data that is produced,*
> *but in how that data is applied to bring gains to the industry."*

Examples of the use of NIR in process control included maceration rate control (Lloyd et al., 2010); clarifier phosphate addition (Markley et al., 2009) and perhaps most impresively, the development of LoGiCane^{TM}, the first naturally low GI (glycemic index) sugar (Kannar et al., 2009). Traceablility and mapping of productivity data and nutrient levels; mill maintence scheduling and identification of the best use of bagasse were also noted as potential added value from NIR analysis within mills. However, there is no evidence of these approaches being further developed in the current literature.

Maceration is the process of adding wash-water to the milling operation and is crucial for efficent sugar extraction. Lloyd et al. (2010) describe a trial at the Marian Mill in Queensland, Australia. Here, online NIR based estimates of cane fibre were used in conjunction with mill operational data in order to automatically adjust maceration addition rates. The trial was successfully completed and allowed easy control of maceration rates but improvement of standard practices could not be statistically tested. Similarly, Markley et al. (2009) describe small trials at Marina mill in 2010 for phosphate and flocculent clarifier addition. A minimum level of phosphate in juice is required for the proper performance of juice clarifiers that help remove dirt and mud from juice. NIR estimates of juice phosphate levels were used to automatically control the addition of phosphate to the juice. Control logic was also used to adjust flocculant additon based on NIR estimates of incoming ash, which can be indicative of dirt and mud levels. While these reports show that NIR can be used for process control in practice there was little or no evidence of the true benefits or further adoption of these approaches by industry. In comparison, LoGiCane^{TM} provides a very strong case for the possible benefits of using NIR analysis for process control.

Low GI foods have health benefits by reducing the rate of glucose and insulin production in the body compared to high GI foods. LoGiCane^{TM} was developed to have a lower GI than regular white or raw sugars by increasing the presence of polyphenols and minerals (O'Shea et al., 2010). NIR analysis of sugar is used to monitor polyphenols and minerals. Spray applications can then be used to increase their levels as need using molasses extracts (Simpson et al., 2011). For more details interested readers should refer to the original patent application (Kannar et al., 2009). The ability to monitor and control mill processes based on real-time analysis is central to the success of this innovative product. Despite this, there is little evidence in the literature of further developments in process control or other value adding initiatives that make use of online NIR analysis. Some of the most recent applications appear in the South African sugarcane industry where NIR quality monitoring has been shown to be beneficial in mill maintanence by allowing mills to identify that evaporator inversion loss were the cause of undetermine sucrose losses (Dairam et al., 2016) and to produce continuous measures of target purity differences, an indicator of malfunctions in the centrifuge (Gounden and Walthew, 2018).

The apparent lack of development in this area may be due to a lack of trust in NIR analysis and a perceived lack of benefit of process control. As an example, in considering NIR analysis within the South African industry Walford (2019) suggested that two main hurdles to acceptance of NIR analysis were that factorty staff "considered conventional analysis as the absolute truth" and an inherent resistance to change. This means that any errors in NIR analysis would seriously erode trust while any accurate predictions may be met with stoic indifference. Futhermore, the lack of economic or wider benefits in the literature suggest that much of the potential value of online analysis is likely lost on any but those directly involved in milling. Apart from the case of LoGiCane^{TM} and direct quality measure used in cane payment, the uses of online NIR analysis found in the literature are focused on benefits to the mill, largely through automation. This would make it difficult to engage with industry parties in the wider value chain.

One area where online NIR analysis and process control could have a large impact across the value chain is in the identification and treatment of deteriorated, contaminated or otherwise atypical cane. Cane deterioration in particular can be very detrimental to mill processes. Deterioration largely occurs due to bacterial infections. Bacteria enter the cane through any damage and metabolise sucrose into less economic products such as organic acids, complex polysaccharides (e.g. dextran) and gums (Solomon, 2009). Any damage to the cane stalk can be a potential entryway for bacteria. Therefore, insect damage and animal damage pre-harvest can cause adverse deterioration. Mechanical harvesting chops cane stalks into small billet so that any delays in crushing can lead to further deterioration (Saxena et al., 2010). The presence of complex sugars and gums can cause higher viscosity and longer crystalization times (Solomon, 2009) and hence can result in greater need for mill maintainence. Lionnet (1986) showed that cane deterioration can also affect pol readings. In particular as the cane deteriorates Pol as measured by a polimeter becomes an unreliable representation of sucrose. This suggests that Pol is a different measure for deteriorated cane which will have a direct effect on the calculation of cane quality measures used in cane payment schemes in the Australian sugar industry.

Contamination in the form of high levels of leaf trash or dirt can also affect standard analysis of cane quality measures. In particular, contamination can inflate laboratory Brix values calculated by hydrometer. If high levels of contaminates are noticed, laboratory Brix measures can be suitably adjusted. However, the effect of such events on NIR analysis are as yet undocumented. One reason for this is the lack of a consistent industry wide definition for when sugarcane should be considered deteriorated or contaminated. Furthermore, methods for calculating deterioration indicators such as the presence of ethanol and manitol are time consuming and expensive (Van Heerden et al., 2014).

The results of Chapters 5 and 6 have shown that online NIR analysis can be used to identify 'atypical' cane samples. These atypical samples were defined as sugarcane that did not follow the linear relationship between Pij and Bij and had unusally low laboratory Pij compared to their recorded Bij. These samples could represent deteriorated (unusually low Pij) or contaminated (unusually high Bij) samples. Classification rates in Chapter 6 were greater than 80% for both atypical and typical cane samples when the full NIR spectrum was used and greater than 90% when a feature selection process was used. These results suggest that the PLS-DA modelling framework could be used to trace atypical cane samples to identify causes or in process control logic in the mill. However, like many reported NIR analyses, there is as yet no practical examples of how this information could be used to benefit the sugar industry as a whole. Atypical cane samples represent only a small fraction of all samples processed by the mill (approximately 3%; Chapter 5). It is possible then that NIR analysis may struggle to match laboratory estimates of these samples.

The objective of this Chapter was to combine the methodologies for quality estimation and atypical sample detection developed and applied in previous Chapters. The aim was to use the lessons learnt to demonstrate the ability of NIR analysis to estimate three cane quality measures: Brix in juice (Bij), Pol in juice (Pij) and Commercial Cane Sugar (CCS) for atypical samples. Specifically, to demonstrate how a process-based workflow can be used to modify these estimates in a practice. Cane deterioration and contamination can have a large impact on cane quality parameters and is a potentially powerful tool in showing the benefits of using data from online NIR analysis. Furthermore, the importance of cane quality is easily recognised not only within the mill but all along the value chain.

7.2 Materials and methods

The research methodology (Figure 7.1) used the following steps:

1. Partial least squares discriminant analysis (PLS-DA; (Barker and Rayens, 2003)) classification model is built to identify cane samples as typical or atypical following the methodology outlined in Chapter 6. The PLS-DA model using feature selection was used as it provided the greatest accuracy for both typical and atypical samples compared to other classification techniques considered (Chapter 6). Classification models were used to produce a probability that a sample is atypical.
2. Partial least squares regression (PLSR; (Wold et al., 2001)) is then used to build models of Bij, Pij and CCS (Chapter 3 and Chapter 4). For each quality, measure two PLSR models are built: A baseline model using all available samples and a model specifically for atypical samples.
3. A process-based approach was then used to sample quality measures. If a sample was considered atypical, the atypical model was used, otherwise the baseline model was used. A calibration set was used to select the cut-off probability at which a sample was considered atypical based on overall RMSEC.
4. The strengths and weaknesses of the baseline models and the process-based approach were compared based on model performance on an independent validation set.

All models were tuned through cross-validation on a calibration data set and validated on an independent validation set. All models were built using R (R Core Team, 2017).

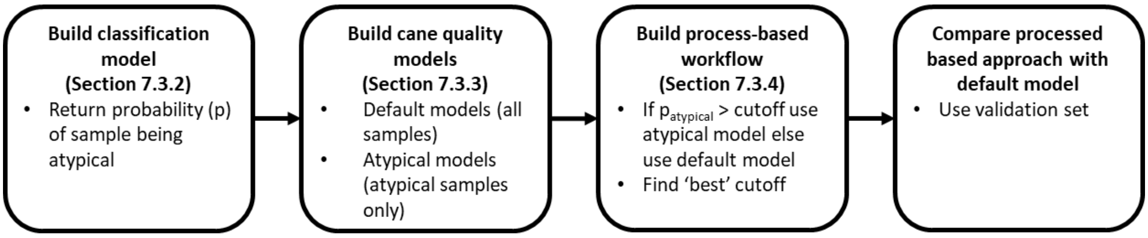
Build classification model (Section 7.3.2)
• Return probability (p) of sample being atypical
Build cane quality models (Section 7.3.3)
• Default models (all samples)
• Atypical models (atypical samples only)
Build process-based workflow (Section 7.3.4)
• If $p_{atypical}$ > cutoff use atypical model else use default model
• Find 'best' cutoff
Compare processed based approach with default model
• Use validation set

Figure 7.1. Overview of methodology

7.2.1 Data

Data were collected from a single sugarcane mill in Northern Queensland, Australia. Data represent consignments from the 2006 - 2009 season that had sufficient NIR spectral data and laboratory analysis of cane quality measures. Brix in juice (Bij), Pol in juice (Pij) and commercial cane sugar (CCS) measures were acquired from the laboratory. Building on the methodologies developed in Chapter 5 and Chapter 6, atypical samples were defined based on the linear relationship between Pij and Bij across all seasons. In total 12,798 samples were included in the analysis. Spectral data were pre-processed using a combination of standard normal variate (Barnes et al., 1989) and Savitzky-Golay First derivative transformation (Savitzky and Golay, 1964). The SG derivative used a second degree polynomial and a window width of 13. Spectral pre-processing was computed using the prospectr package in R (Stevens and Ramirez-Lopez, 2013).

Data were divided evenly into a calibration and validation data set of 6,993 samples each. Data were divided using a stratified random sampling approach such that the proportion of typical and atypical samples was maintained in each set (Table 7.1). The cane quality parameters was relatively evenly distributed between the calibration and validation sets, with similar levels of typical and atypical samples from each season represented in each set. From (Table 7.1) it can be seen that samples defined as atypical tended to have lower Pij and CCS.

Table 7.1. Overview of data used in this Chapter. Pij, Bij, CCS and AP are laboratory measured data represented by the mean and standard deviation (SD).

Set	Season	Type	Number of samples	Pij (SD)	Bij (SD)	CCS (SD)
Calibration	2006	atypical	35	17.7(1.61)	22.01(1.45)	11.99(1.33)
		typical	1844	18.26(1.9)	20.79(1.77)	13.33(1.49)
	2007	atypical	52	16.3(2.58)	20.79(2.16)	11.04(2.32)
		typical	1595	18.5(1.96)	21.1(1.8)	13.51(1.51)
	2008	atypical	39	17.48(1.92)	21.82(1.54)	12.21(1.76)
		typical	1474	19.79(1.68)	22.28(1.54)	14.47(1.29)
	2009	atypical	36	18.02(1.88)	22.34(1.56)	12.12(1.74)
		typical	1324	19.52(1.51)	22.02(1.41)	14.22(1.18)
Validation	2006	atypical	37	17.72(1.46)	22.25(1.38)	11.86(1.15)
		typical	1805	18.22(1.86)	20.75(1.73)	13.31(1.46)
	2007	atypical	53	16.09(2.83)	20.67(2.37)	10.9(2.54)
		typical	1616	18.52(1.93)	21.13(1.78)	13.52(1.49)
	2008	atypical	36	17.49(2.08)	21.85(1.62)	12.11(1.84)
		typical	1472	19.88(1.61)	22.34(1.48)	14.54(1.24)
	2009	atypical	36	18.04(1.74)	22.37(1.47)	12.22(1.64)
		typical	1344	19.48(1.49)	22(1.41)	14.19(1.17)

7.2.2 Discrimination of atypical and typical samples

Following the methodology outlined in Chapter 6, a discrimination model was built using PLS-DA to differentiate between atypical and typical samples (Figure 7.2). The PLS-DA models were built using the caret package in R (Kuhn and Johnson, 2013a, Martens et al., 1992). Class probabilities were calculated using the softmax function. Class predictions were considered the class with the highest probability (>0.5). The number of latent variables was tuned through a five-fold cross-validation over a range of 1 – 20. The cross-validated ROC AUC score was used to select the best number of latent variables as the number of latent variables that maximized the AUC.

A genetic algorithm feature selection process was used in order to identify influential wavelengths (Xiaobo et al., 2010). Based on the results of Chapter 6, the wavelengths used were those that were selected in more than 50% of GAFS runs during cross-validation as this proved to result in higher model accuracy in validation (GAFS-Imp50; Chapter 6). In comparison to Chapter 6, a single realization was used so that misclassifications could be investigated in greater detail.

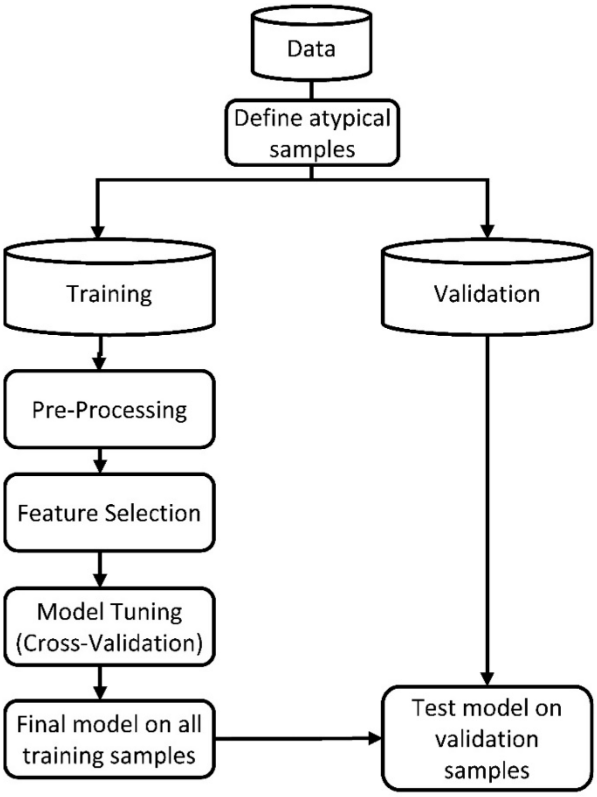
Data
Define atypical samples
Training
Validation
Pre-Processing
Feature Selection
Model Tuning (Cross-Validation)
Final model on all training samples
Test model on validation samples

Figure 7.2. Discrimination model methodology.

Discrimination model performance was recorded as overall Accuracy such that:

$$Accuracy = \frac{number\ of\ correctly\ classified\ samples}{number\ of\ samples} \times 100. \quad (7\text{-}1)$$

Correct classification rates for atypical and typical samples were also recorded, where

$$CCR_{atypical} = \frac{number\ of\ correctly\ classified\ atypical\ samples}{number\ of\ atypical\ samples} \times 100 \quad (7\text{-}2)$$

and

$$CCR_{typical} = \frac{number\ of\ correctly\ classified\ typical\ samples}{number\ of\ typical\ samples} \times 100. \quad (7\text{-}3)$$

Model performance was recorded for both the calibration and validation data sets. When the final model was applied to the calibration and validation data sets, the softmax probabilities were recorded along with class predictions.

7.2.3 Estimating cane quality

Following the methodology outlined in Chapter 4, PLSR models of Bij, Pij and CCS. PLSR models were built using the pls package (Mevik et al., 2015) in R. The number of latent variables was tuned through five-fold cross-validation of the calibration set. The best number of latent variables was selected based on the one standard error approach (Hastie et al., 2013b, Kuhn, 2017), using model RMSECV (root mean square error of cross-validation) as the objective function. The one standard error approach was used to reduce the tendency of the model to overfit to the training set as pre-testing showed that the 'best' result was usually the maximum number of latent variables allowed. Cross-validation sets were split such that each split covered the variability of the target quality measure.

For each quality measure, three PLSR models were built:

1. baseline: all samples from the calibration set were used in developing the PLSR models
2. atypical: Only samples defined as atypical were used in model development

Model performance was investigated based on RMSE, R^2 and Bias. RMSE values were calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y_i}-y_i)^2}{N}}, \qquad (7\text{-}4)$$

while R^2 values were calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y_i}-y_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}. \qquad (7\text{-}5)$$

and Bias was calculated as

$$Bias = \frac{\sum_{i=1}^{N}(\hat{y_i}-y_i)}{N}\,. \qquad (7\text{-}6)$$

Validation set performance statistics RMSEP, R^2_p and $Bias_p$ were recorded for all samples as well as for atypical and typical samples individually. Results from the baseline models were used to investigate the difference in model performance between sample types. Baseline and atypical models were then used in the process-based approach (Figure 7.3). The validation set Residual Prediction Deviation (RPD) and the slope of the regression line between predicted and observed data were also recorded for completeness. The RPD statistic represents a ratio of the observed variance and model error variance and is considered an important statistic in reporting NIRS model analysis (Williams et al., 2017).

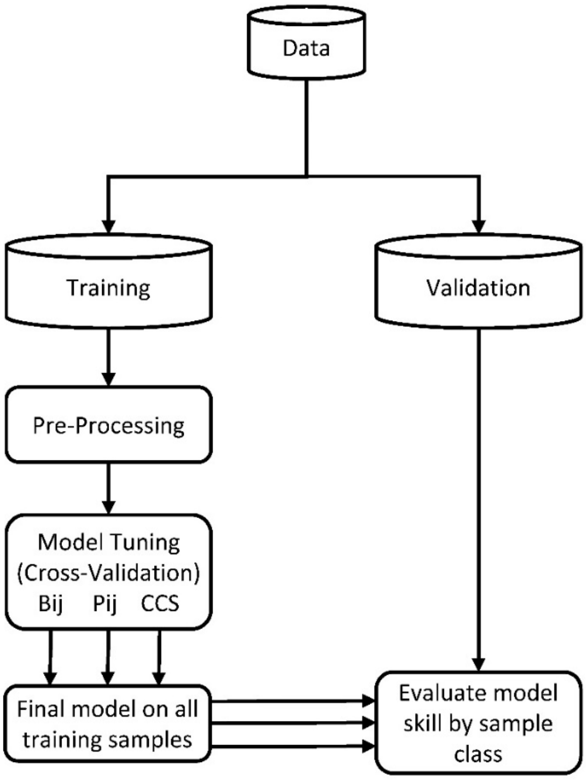

Figure 7.3. Methodology overview for modelling cane quality.

7.2.4 Process-based estimation of cane quality

A process-based approach was used to estimate cane quality of typical and atypical samples independently using predicted class probability (**Figure 7.4**). For each sample, the final PLS-DA model was used to predict the probability that a sample was atypical. If the predicted probability was greater than some cut-off (p) then the final PLSR models for atypical samples were used to predict sample quality parameters. Otherwise, the baseline PLSR models were used. This process-based approach allowed the cut-off value p to be tuned to suite any desired outcome.

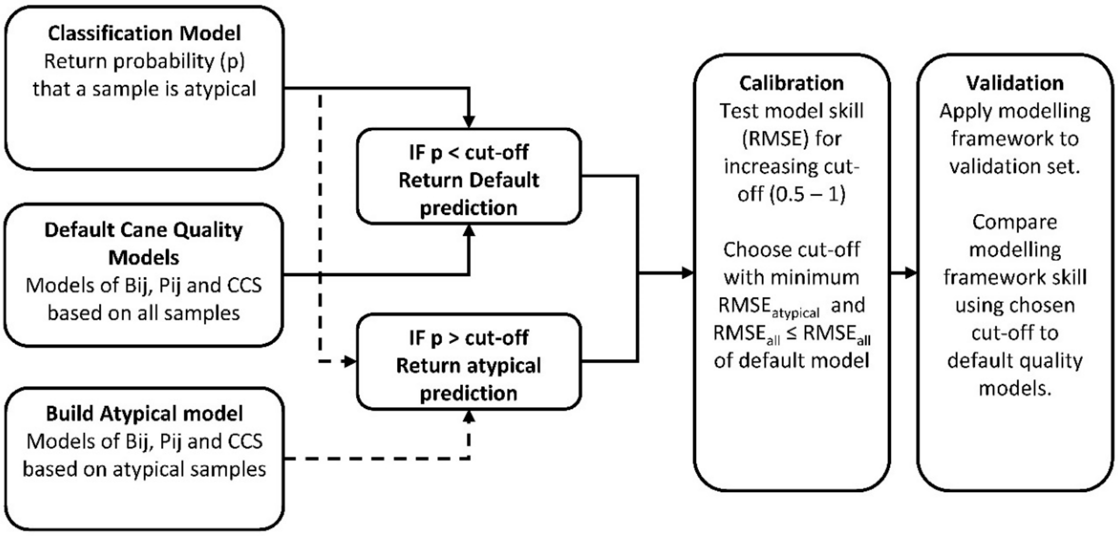
Classification Model
Return probability (p) that a sample is atypical
Default Cane Quality Models
Models of Bij, Pij and CCS based on all samples
Build Atypical model
Models of Bij, Pij and CCS based on atypical samples
IF p < cut-off Return Default prediction
IF p > cut-off Return atypical prediction
Calibration
Test model skill (RMSE) for increasing cut-off (0.5 – 1)
Choose cut-off with minimum RMSE$_{atypical}$ and RMSE$_{all}$ ≤ RMSE$_{all}$ of default model
Validation
Apply modelling framework to validation set.
Compare modelling framework skill using chosen cut-off to default quality models.

Figure 7.4. Methodology overview for process-based estimates of cane quality.

The calibration set was used to tune the cut-of probability (p) for each quality measure. The best p value was defined as the lowest cut-off that resulted in no reduction in overall RMSEP (RMSEP of all samples) compared to the baseline model. The value p was changed from 0 (all samples considered atypical) to 1 (all samples considered typical) at intervals of 0.001. This resulted in the lowest possible RMSEP for atypical samples, without reducing overall model performance. The effect of cut-off value on model performance and the discrimination ROC curve were explored graphically. Finally, model performance on the validation set using the optimum cut-offs were compared to the baseline modelling approach.

7.3 Results and discussion

7.3.1 Discrimination of atypical and typical samples

The final discrimination model used 14 latent variables, based on the 345 wavelengths selected by the genetic algorithm feature selection stage. The discrimination model performed well for both calibration and validation data sets (Table 7.2). The similarity between calibration and validation sets is promising as it suggests that the model was not overfit to the training set. However, performance may be optimistic as the Calibration set was representative of the validation set (Table 7.2). The results presented in Table 7.2 are similar to those reported in Chapter 6 with a slightly higher correct classification rate for atypical samples and slightly lower correct classification rate for typical samples. Model performance was considered appropriate for the further investigation of class probabilities.

Table 7.2. Calibration and validation statistics for classification model. True/False atypical count records the number of samples correctly and incorrectly identified as atypical. Atypical rate records the percentage of samples predicted as atypical. In both calibration and validation sets 2.53% of samples were defined as atypical.

Set	True Atypical Count	False Atypical Count	Atypical Rate	AUC	$CCR_{atypical}$	$CCR_{typical}$	Accuracy
Calibration	149	632	12.21%	0.9708	91.98%	89.87%	89.92%
Validation	152	664	12.75%	0.9629	93.83%	89.35%	89.47%

The results presented in Table 7.2 make use of the default model classification and are equivalent to using a probability cut-off of 0.5. That is, a sample was considered atypical if the predicted probability index was greater than 0.5. Although classification rates are relatively high for both atypical and typical samples, there was more than four times as many false atypical classifications as true atypical classifications (Table 7.2), resulting in a predicted atypical rate of >12%. This is much higher than the defined rate of 2.57%.

The samples incorrectly identified as atypical may be a result of 'arbitrary' cut-off that defines atypical samples. For example, atypical samples were defined as having unusually low laboratory Pij relative to laboratory Bij which may be caused by deterioration or contamination. Samples that have low levels of deterioration or contamination may be identified as atypical even though the lab measured Pij has not been affected.

The use of down-sampling resulted in a model with good skill for both atypical and typical samples. However, the much higher number of false atypical samples may be inappropriate depending on the intended application. Modifying the probability cut-off and exploring the ROC curve in the calibration set showed that using a cut-off of 0.652 resulted in a predicted atypical rate of 2.47%. This was much closer to the defined rate. These results showed that the pseudo-probability cut-off could be used to modify the correct classification rates of typical and atypical samples as needed and therefore could be used as a tuneable parameter for process decisions.

7.3.2 Quality estimation of atypical samples using PLSR

All three quality measures were well estimated by the baseline PLSR model (Table 7.3). RMSEP for Bij (RMSEP = 0.289%), Pij (RMSEP = 0.362%) and CCS (RMSEP = 0.394%) were low and compare well with the results found in Chapter 3 and Chapter 4 as well as previous research (Staunton et al., 1999, Staunton et al., 2004). When all samples were considered bias was also relatively low for all measures. The consistency of these results with previous industry results confirm that the PLSR models used are a good baseline to explore the differences in model skill between atypical and typical samples.

Table 7.3. Validation set skill for baseline PLSR and process based class modelling approaches. Model skill was assessed as Bias, RMSE and R^2 for Bij, Pij and CCS. Model skill was recorded for atypical and typical and across all samples. Validation Slope and Residual Prediction Deviation (RPD) were recorded for completeness.

Model	Measure	Tuning parameters	Type	$Bias_p$	RMSEP	R_p^2	$Slope_p$[a]	RPD_p[b]
Baseline PLSR	Bij	No. LVs: 14	Atypical	-0.089	0.450	0.946	0.961	4.333
			Typical	0.015	0.285	0.973	0.974	6.127
			All	0.012	0.290	0.973	0.973	6.034
	Pij	No. LVs: 15	Atypical	0.717	1.016	0.804	0.876	2.266
			Typical	-0.008	0.331	0.969	0.974	5.663
			All	0.011	0.365	0.963	0.962	5.229
	CCS	No. LVs: 15	Atypical	0.890	1.181	0.651	0.804	1.697
			Typical	-0.015	0.351	0.941	0.954	4.120
			All	0.008	0.394	0.931	0.928	3.810
Process based	Bij	No. LVs: 14	Atypical	-0.093	0.453	0.946	0.963	4.297
		Atypical LV's: 7	Typical	0.015	0.285	0.973	0.974	6.126
		Cut-off: 0.929	All	0.012	0.291	0.972	0.973	6.030
	Pij	No. LVs: 15	Atypical	0.296	0.839	0.867	0.928	2.746
		Atypical LV's: 10	Typical	-0.018	0.347	0.966	0.974	5.401
		Cut-off: 0.601	All	-0.010	0.368	0.963	0.969	5.184
	CCS	LV's: 15	Atypical	0.407	0.937	0.780	0.903	2.139
		Atypical LV's: 11	Typical	-0.028	0.370	0.935	0.959	3.915
		Cut-off: 0.616	All	-0.017	0.394	0.931	0.948	3.812

[a] Slope was calculated as the $\boldsymbol{\beta}$ coefficient of the linear least squares fit of $\hat{\boldsymbol{y}} = \beta \boldsymbol{y} + c$

[b] Bias was calculated as mean difference between predictions and observations $bias = \frac{\sum_{i=1}^{N}(\widehat{y_i} - y_i)}{N}$

[c] RPD was calculated as the ratio of the standard deviation of the observations in the validation set and the RMSEP $RPD = \frac{sd(obs_p)}{RMSEP}$ $where\ sd(obs_p) = \sqrt{\frac{\sum_{i=1}^{N_p}(\widehat{y_i} - \bar{y})^2}{N_p - 1}}$

When RMSEP and $Bias_p$ were considered for atypical and typical samples separately, it was evident that atypical samples had higher RMSE, lower R^2 and larger biases (Table 7.3). In particular, RMSEP of atypical samples was more than twice as large as typical samples for Pij and CCS. Comparison of RPD values between atypical and typical values agreed with the results for RMSEP with higher values for typical samples and lower values for atypical samples suggesting that atypical samples were more difficult to estimate.

The large RMSEP of atypical samples for Pij and CCS would likely not be considered accurate enough for practical application if atypical samples were estimated alone. Bias of atypical samples for Pij (Bias = +0.706) and CCS (Bias = +0.903) were much higher than typical samples. The positive bias represented a tendency to overestimate Pij and CCS for atypical samples. Graphically, the bias of atypical samples was much more evident for Pij (Figure 7.5(c)) and CCS (Figure 7.5(e)) than Bij (Figure 7.5(a)). Interestingly, the tendency to overestimate atypical samples was not limited to samples with low Pij or CCS. While lower values had a larger overestimation atypical samples were overestimated regardless of observed value.

The overestimation and higher RMSE of Pij and CCS for atypical samples is an important result. The baseline PLSR approach has the advantage of being relatively simple to maintain. Furthermore, it is possible to ‘look inside’ the model and identify how spectral data are related to the quality measure estimated. This simple approach has been shown to work as well as more complex machine learning approaches (Chapter 3; Chapter 4). However, the results presented here show that atypical samples represent a definable subset of samples that are consistently miss-represented. This result suggests that identifying atypical samples could be used to directly affect the NIR analysis of cane quality. Specifically a positive bias or overestimated cane samples represent an overpayment for the mill compared to laboratory analysis. Identifying atypical samples and removing this bias would reduce expense for the mill as well as identifying samples that may reduce mill efficiency and allowing for early management interventions.

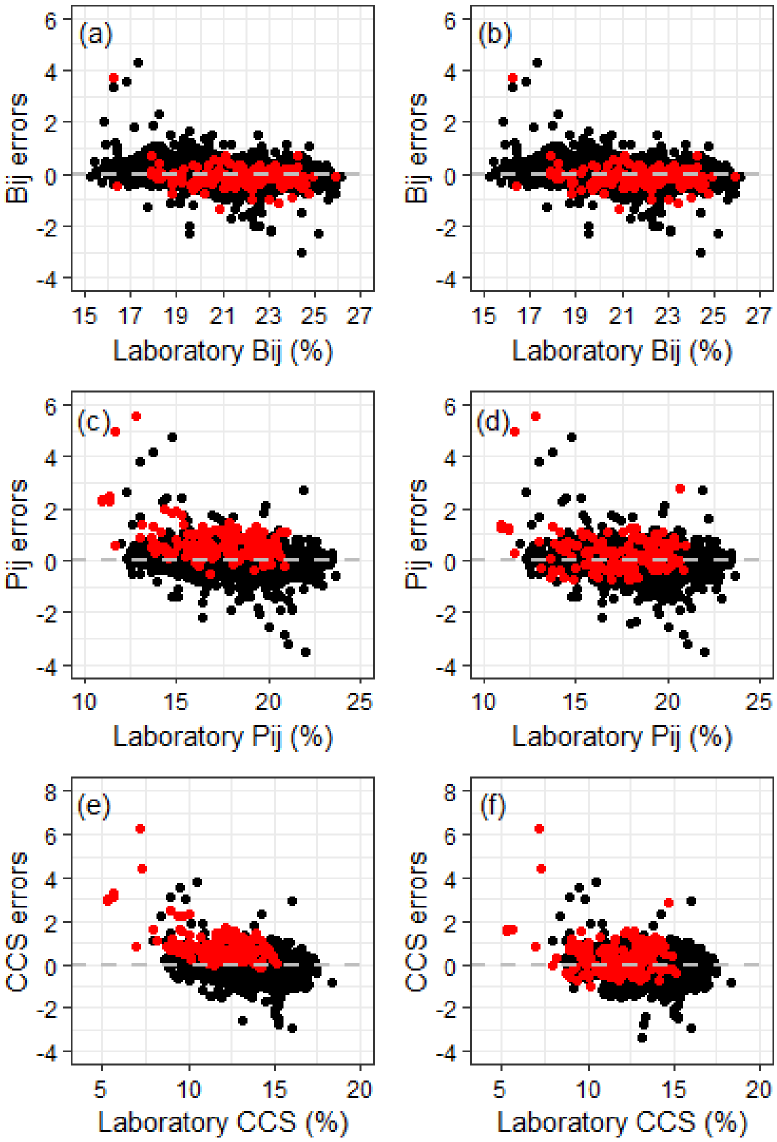

Figure 7.5. Model errors as function of observed value for Bij (a, b), Pij (c, d) and CCS (e, f). Errors represent the baseline PLSR (a, c and e) and process-based (b, d, f) PLSR modelling approaches.

7.3.3 Tuning cut-off for tuned class approach

Modelling atypical samples as a separate population resulted in lower calibration RMSEC compared to the default PLSR models for each quality measure. RMSEC of atypical samples on the calibration set for Bij, Pij and CCS were (0.506%), (0.599%) and (0.585%) respectively. This was a reduction of (2.67%) (36.21%) and (46.33%) relative to the baseline PLSR model. These results suggest that if perfect knowledge of atypical samples was possible, quality estimates for atypical samples could be improved. However, RMSEC of atypical samples was still larger than results for overall model performance.

Probability cut-off points for Bij, Pij and CCS tuned class approach were 0.929, 0.616 and 0.601 respectively (Table 7.3). The cut-off for Pij and CCS reduced the correct classification rate of atypical samples to < 70% but also reduced the predicted atypical rate to < 5% (Figure 7.6(a)). This atypical rate is much closer to the 2.53% observed rate. The selected cut-off reduced the calibration RMSE of Pij (Figure 7.6(c)) and CCS (Figure 7.6(d)) without reducing the overall RMSE. In comparison, the selected cut-off for Bij resulted in all samples being considered typical and therefore did not reduce the calibration RMSE of atypical samples. This suggests that it was not effective to treat atypical samples separately when estimating Bij but was effective for Pij and CCS.

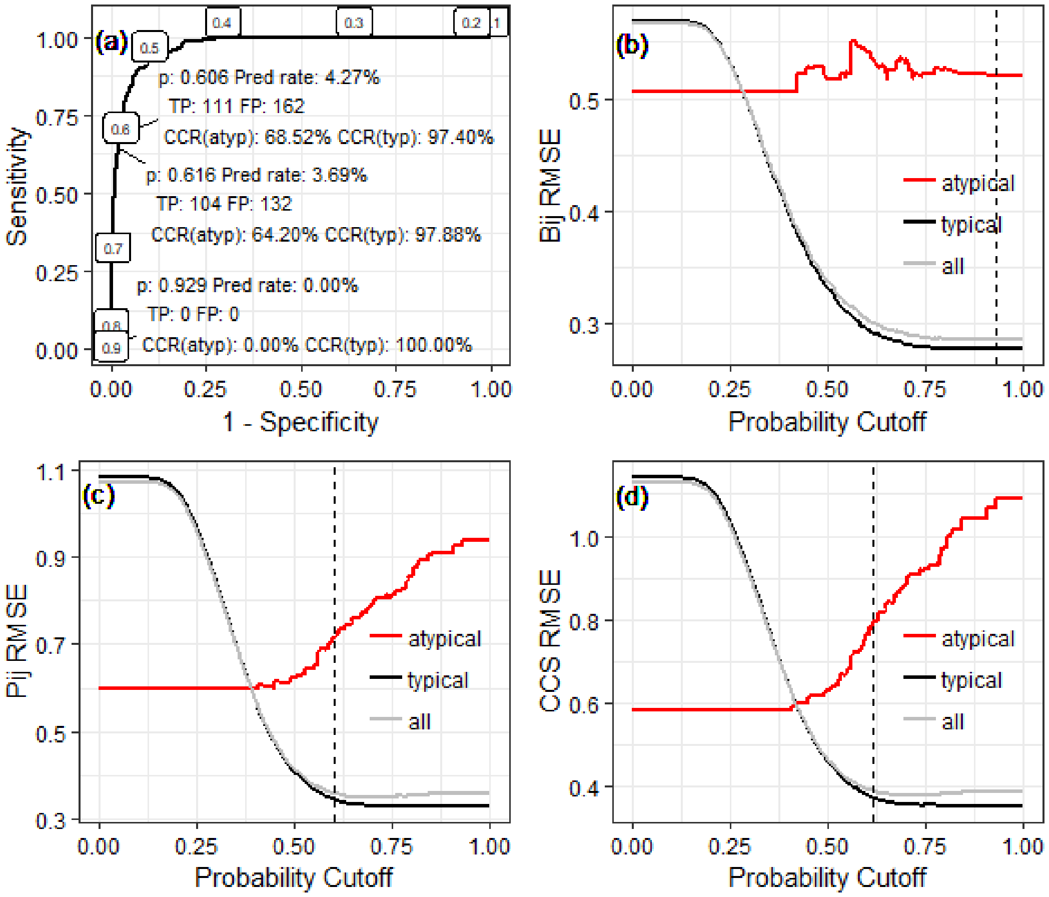

Figure 7.6. Model accuracy measures for calibration data. (a) ROC curve for PLS-DA model. Line shows the trade-off between Sensitivity (CCR$_{\text{atypical}}$) and Specificity (CCR$_{\text{typical}}$) for a given cut-off probability. Figures (b), (c) and (d) show RMSEC at various cut-off probabilities for Bij, Pij and CCS respectively. Lines represent atypical (red), typical (black) and all samples (grey).

7.3.4 Comparison of modelling approaches

When applied to the validation data set, the process-based approach maintained very similar RMSE to the baseline PLSR model for each quality measure (Table 7.3). This shows that tuning on the calibration set was effective. The close match between tuned and baseline PLSR on the validation set may also be attributed to the calibration set being a good representation of the validation set. However, there were evident differences in the RMSEP of atypical samples. In particular, RMSEP of atypical samples for Pij and CCS were approximately 17% and 20% lower than baseline models (Table 7.3). RMSEP of Bij for atypical samples actually increased when the process-based approach was used.

The greatest difference between baseline and process-based approaches was the Bias of atypical samples for Pij and CCS. Bias of Pij and CCS was 0.296% and 0.407% respectively when the process-based approach was used. This was more than 50% lower than the respective baseline models. While RMSEP for quality measures were lower for Pij and CCS the spread of errors may have been higher (Figure 7.5). This spread is likely due to errors in the atypical models themselves as well as errors due to the classification model. For example, the two samples with the highest error in CCS for the baseline model (Figure 7.5(e)) have the same errors using the process-based approach (Figure 7.5(f)). This means that despite being defined as atypical samples, they were not predicted to be atypical by the process-based approach. These may be true outliers that could have been removed from the model building process to improve model accuracy or may be misclassifications due to the change in cut-off value p. These results suggest that both PLS-DA and PLSR models could be improved further.

The advantages of the baseline PLSR modelling approach are the ease of use, relative simplicity and good overall modelling accuracy. The PLSR approach is already widely established within industry and overall accuracy in terms of RMSEP was not effectively improved by using a process-based approach. In comparison the process-based approach requires the maintenance of several algorithms each of which can allow errors to creep into the estimation of cane quality parameters. However, with the added complexity of the process-based approach comes the reduction in bias of atypical samples for Pij and CCS. A drop in CCS bias of 0.483% CCS represents 0.483 tonnes of sugar per 100 tonnes of ‘atypical’ cane produced. At 2.54% atypical rate and a conservative 500,000 tonnes of cane processed by the mill each year, this is equivalent to a saving of approximately 62 tonnes of sugar paid for by the mill when using the process-based approach.

Results from Chapter 6 suggested that identification of atypical samples at the mill could be used to identify when and where atypical samples are occurring allowing for management interventions. The results of this study show that it is also possible to use the information generated by a classification model in process control within a sugarcane mill. In this study we demonstrated that this could have direct effects on quality estimates. However, a similar approach could be used in future to separate atypical cane for use in alternative production lines such as biofuels or to reschedule mill maintenance. By using a probability output rather than a simple binary classification, it is possible to modify the process to suit the desired outcome.

7.4 Conclusion

Sucrose based quality measures such as Pij and CCS of atypical samples tend to be overestimated by NIR analysis. This overestimation makes sense if atypical samples are indeed deteriorated or contaminated. However, to the best of our knowledge, this is the first time such a bias has been reported in the literature. The fact that a definable and identifiable subset of samples is consistently misrepresented will have important implications for the sugarcane industry in terms of cane payment calculations and potentially mill maintenance. By making effective use of NIR analysis that identifies atypical samples it was possible to remove some of this bias. This potential benefit comes at the cost of a more complex modelling approach and the need to maintain multiple models. The use of such a methodology for modifying cane quality estimates will need to be considered carefully, given the importance of quality in cane payment schemes. However, these results highlight the power of NIR analysis for mill processes and the potential benefits of making the most of NIR analysis data. Future research could consider extending this methodology to address a range of early interventions such as mill maintenance scheduling and tracking the source of atypical samples.

7.5. Chapter 7 summary

In any given season, laboratory processes in sugarcane mills identify a certain percentage of atypically low quality cane samples. Recent research has shown that these samples can be detected using NIR analysis systems already used in mills. However, there was still little research on how NIR based quality measures should be estimated for these samples. Chapter 7 explored a process-based approach to estimating cane quality measures for atypical samples using NIR analysis. A PLS-DA model was used to predict the probability of a sample being atypical. Three sugarcane quality measures (Pol in juice, Brix in juice and CCS) were then estimated using partial least squares regression. If a sample was identified as atypical an atypical specific PLSR model was used to estimate quality parameters. Results of this study showed that Pol-based quality estimates (Pij and CCS) for samples identified as atypical are over-estimated using a baseline PLSR approach. By making use of the probability of a sample being atypical, it was possible to reduce this bias without increasing overall model root mean square error.

The focus of Chapter 7 was Objective 3 of the thesis: Investigating the use of NIR classification data to improve estimates of cane quality parameters for atypical cane samples. The results from Chapter 7 showed that Pol-based quality estimates (Pij and CCS) for samples identified as atypical are over-estimated using a baseline PLSR approach. By making use of the probability of a sample being atypical, it was possible to reduce this bias without increasing overall model root mean square error. These results show that NIR analysis can be used not only to identify and track atypical samples, but in process control within the mill. By using class probability as a tuneable parameter, it was possible to modify NIR models to achieve a desired outcome. The most novel aspect of the process-based modelling framework developed in Chapter 7 was the use of the class probability as a tuneable parameter. While there is evidence of class based modelling approaches, there was no evidence of this type of flexibility used in the current literature.

Chapter 8

Thesis conclusions

My thesis investigated statistical methodologies to measure sugarcane attributes for anomalous cases from NIR spectra. Specifically I answered three main research questions:

1. Can data mining algorithms improve estimates of cane quality?
2. Can NIR analysis be used to identify atypical cane?
3. Can class predictions be used to improve estimates of cane quality for atypical samples?

The results I have presented in this thesis answered these questions, and provided a better understanding of the role of data mining algorithms in on-line NIR analysis as well as a framework for improving quality estimates of atypical samples. Specifically, the key outcomes of the thesis showed that:

1. The partial least squares (PLS) modelling framework was easily comparable in performance to the more complex algorithms such as support vector machines and artificial neural networks. This was true for both the regression and classification problems explored.
2. A modelling approach that combined down-sampling and pre-processing was used to develop a balanced PLS discriminant analysis (PLS-DA) model capable of accurately classifying atypical cane samples.
3. A methodological framework that used class probability predictions was developed to remove bias in CCS estimates for atypical samples without reducing overall model error.

My research also provided a number of insights into the detection and treatment of atypical samples that can lead to recommendations for the sugarcane industry in Australia and other primary industries where online NIR analysis is in use.

One major insight drawn from my investigations was the good performance of PLS for both regression and classification tasks compared to the more complex and non-linear approaches tested. This was an important result as the PLS approach to regression is already well established within the Australian sugarcane industry. In a broader modelling context, PLS may be a preferable approach in on-line NIR analysis systems as it is a simple model to maintain and recalibrate, has a sound theoretical background and is easy to interpret. The SVM and ANN approaches were comparable with PLS and could be used in place of PLS models. The major drawback of these approaches would be the difficulty in maintaining the models and the complexity of explaining the modelling process, which may prevent adoption within the Australian sugarcane industry. One advantage of SVR was seen in the estimation of quality parameters. SVR models tended to perform better for samples with values close to the limits of the calibration range. Industry may wish to continue to explore the use of machine learning algorithms especially for tasks where a more global model is required or where more evident non-linear effects are expected. In contrast to SVR and ANN approaches, tree-based approaches tended to perform noticeably worse than PLS models and were not suited to online NIR cane analysis systems for either regression or classification tasks.

Insight 1: PLS models performed as well as more complex models for regression and classification tasks. The sugarcane industry should have confidence in the continued use of these algorithms. Industry should only consider machine learning models for specific tasks where high non-linearity is expected.

Data pre-processing, feature selection and appropriate calibration data selection were used to improve model performance. In exploring the identification of atypical samples, it was shown that a particular spectral pre-processing might be more or less effective for a particular modelling approach. This should be kept in mind in future research, especially if a range of modelling approaches are being explored. Results also showed that for models of quality parameters, PLSR, SVR and ANN placed importance on similar wavelength ranges. This suggests that a faster algorithm such as PLS could be used to identify wavelengths for use with more complex models. This is important as many feature selection routines are computationally expensive and can be impractical to apply to machine learning algorithms that are also computationally expensive.

The importance of an appropriate calibration range for regression is widely accepted. In estimating quality measure, regression model skill was reduced for validation samples that were outside the range of the calibration set. Properly structuring the calibration set was also important for classification tasks. Specifically, down-sampling during model calibration played a large role in ensuring that models performed well for both atypical and typical classes. While down-sampling help improve model performance, it also reduced the variability captured during the calibration process as the majority of typical samples were removed. The relatively low number of atypical samples with observed data is also a concern for calibrating models of quality measures specifically for atypical cane. One potential solution is to collect more data for atypical cane (e.g. deteriorated or contaminated cane samples) however, this may be impractical in commercial applications. An alternative is for future research to explore semi-supervised approaches that could better leverage the NIR data from samples with no available laboratory data. This would extend upon the methodological framework laid out in this thesis.

Insight 2: The importance of appropriate training data samples and training set construction cannot be overstated. Down-sampling and feature selection were important for developing balanced classification models for imbalanced classes. Future research needs consider semi-supervised techniques to make better use of all available data.

My work showed that the CCS of samples with atypically low Pij compared to Bij tended to be overestimated by baseline NIR modelling approaches. In order to reduce this bias, a modelling framework was developed that first classified a sample as atypical then applied an appropriate quality estimation model. This methodological framework made use of the predicted pseudo-probability that a sample would be atypical, in order to tune the point at which as sample should be considered atypical. Such a 'process-based' approach has rarely been used within industry. The novel use of class probability rather than distinct class assignment provides an advantage by allowing the modelling framework to be tuned for a specific task without the need to completely rebuild the classification model. This is important within industry as the modelling framework can be quickly tuned to reflect changes in the risk associated with misclassification. For example, is it more important to catch all atypical samples, or to reduce the number of typical samples that are treated as atypical?

Insight 3: PLS-DA can be used for imbalanced classification tasks in NIR analysis within the Australian sugarcane industry. Using a pseudo-probability allowed for the 'definition' of atypical samples to be tuned to the desired task. Future research need to consider true probabilistic models for discrimination tasks.

In this thesis I have shown that it was possible to identify atypical low quality cane samples using NIR analysis. The ability to classify atypical cane samples could be used to track occurrences and identify causes of atypical samples such as deterioration or soil contamination. Furthermore, a novel use case was presented that used classification information to reduce the bias is estimates of CCS for atypical cane samples. Future research will need to consider using the methodological framework described here in practice alongside standard operations. This would allow industry to validate the framework further and could offer further insight into the types of samples identified as atypical if identified samples can be assessed in the laboratory.

The methodologies I have explored and the methodological framework I have developed in this thesis have strong implications for the Australian sugarcane industry. However, my research also makes important contributions to the broader NIR spectroscopic community. The high skill shown for PLS modelling approaches compared to more complex machine learning techniques is an important contribution as a counterpoint to published research that shows a clear advantage for complex techniques. My results reinforce the need for future researchers to consider a range of modelling approaches and data pre-processing to find the most appropriate modelling framework for the task at hand. Another key contribution was the development of the 'process-based' methodological framework. There are few examples of class based quality or constituent estimation. The inclusion of the class probability as a tuneable parameter was a unique example of how classification information can be used in a practical online NIR analysis setup. The outcomes and insights from this thesis can inform future research, not only for the case of atypical cane samples but for any application for imbalanced or complex discrimination tasks.

References

AGELET, L. E. & HURBURGH, C. R. 2010. A tutorial on near infrared spectroscopy and its calibration. *Critical Reviews in Analytical Chemistry,* 40**,** 246-260. doi: http://dx.doi.org/10.1080/10408347.2010.515468.

ALAM, M. K., STANTON, S. & HEBNER, G. B. 2008. Plastics analysis in two laboratories. *Handbook of near-infrared analysis.* Third Edition ed.: CRC Press/Taylor & Francis Group.

ALDRICH, E. 2013. Wavelets: A package of functions for computing wavelet filters, wavelet transforms and multiresolution analyses. 0.3-0 ed.

ARAÚJO, S. R., WETTERLIND, J., DEMATTÊ, J. A. M. & STENBERG, B. 2014. Improving the prediction performance of a large tropical vis-nir spectroscopic soil library from brazil by clustering into smaller subsets or use of data mining calibration techniques. *European Journal of Soil Science,* 65**,** 718-729. doi: http://dx.doi.org/10.1111/ejss.12165.

BALABIN, R. M. & LOMAKINA, E. I. 2011. Support vector machine regression (svr/ls-svm)-an alternative to neural networks (ann) for analytical chemistry? Comparison of nonlinear methods on near infrared (nir) spectroscopy data. *Analyst,* 136**,** 1703-1712. doi: http://dx.doi.org/10.1039/C0AN00387E.

BALABIN, R. M., SAFIEVA, R. Z. & LOMAKINA, E. I. 2010. Gasoline classification using near infrared (nir) spectroscopy data: Comparison of multivariate techniques. *Analytica Chimica Acta,* 671**,** 27-35. doi: http://dx.doi.org/10.1016/j.aca.2010.05.013.

BALABIN, R. M. & SMIRNOV, S. V. 2011. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. *Analytica Chimica Acta,* 692**,** 63-72. doi: http://dx.doi.org/10.1016/j.aca.2011.03.006.

BALABIN, R. M. & SMIRNOV, S. V. 2012. Interpolation and extrapolation problems of multivariate regression in analytical chemistry: Benchmarking the robustness on near-infrared (nir) spectroscopy data. *Analyst,* 137**,** 1604-1610. doi: http://dx.doi.org/10.1039/C2AN15972D.

BALLABIO, D. & CONSONNI, V. 2013. Classification tools in chemistry. Part 1: Linear models. Pls-da. *Analytical Methods,* 5**,** 3790-3798. doi: http://dx.doi.org/10.1039/C3AY40582F.

BARKER, M. & RAYENS, W. 2003. Partial least squares for discrimination. *Journal of Chemometrics,* 17**,** 166-173. doi: http://dx.doi.org/10.1002/cem.785.

BARNES, R. J., DHANOA, M. S. & LISTER, S. J. 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy,* 43**,** 772-777. doi: http://dx.doi.org/10.1366/0003702894202201.

BECK, M. 2016. Neuralnettools: Visualization and analysis tools for neural networks.

BEN ISHAK, A. 2016. Variable selection using support vector regression and random forests: A comparative study. *Intelligent Data Analysis,* 20**,** 83-104. doi: http://dx.doi.org/10.3233/IDA-150795.

BERDING, N. & BROTHERTON, G. A. 1996. Analysis of samples from sugarcane evaluation trials by near ifra-red spectroscopy using a new at-line, large cassette presentation module. *In:* WILSON, J. R., HOGARTH, D. M., CAMPBELL, J. A. & GARSIDE, A. L. (eds.) *Sugarcane: Research towards efficient and sustainable production.* CSIRO Division of Tropical Crops and Pastures, Brisbane.

BERDING, N., BROTHERTON, G. A., LE BROCQ, D. G. & SKINNER, J. C. 1991. Near infrared reflectance spectroscopy for analysis of sugarcane from clonal evaluation trials: I. Fibrated cane. *Crop Science,* 31**,** 1017-1023. doi: http://dx.doi.org/10.2135/cropsci1991.0011183X003100040035x.

BERDING, N., BROTHERTON, G. A., LEBROCQ, D. G. & SKINNER, J. C. 1989. Application of near infrared reflectance (nir) spctroscopy to the analysis of sugarcane in clonal evaluation trials. *Proceedings of the Australian Society of Sugar Cane Technologists,* 11**,** 8-15.

BERDING, N. & MARSTON, D. H. 2010. Operational validation of the efficacy of spectracane™, a high-speed analytical system for sugarcane quality components. *Proceedings of the Australian Society of Sugar Cane Technologists,* 32**,** 445-459.

BERTRAN, E., BLANCO, M., MASPOCH, S., ORTIZ, M. C., SÁNCHEZ, M. S. & SARABIA, L. A. 1999. Handling intrinsic non-linearity in near-infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems,* 49**,** 215-224. doi: http://dx.doi.org/10.1016/S0169-7439(99)00043-X.

BEVIN, C., STAUNTON, S., STOBIE, R., KINGSTON, J. & LONERGAN, G. 2002. On-line use of near infrared spectroscopy in a sugar analysis system (sas). *Proceedings of the Australian Society of Sugarcane Technologists,* 24.

BRADLEY, A. P. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition,* 30**,** 1145-1159. doi: http://dx.doi.org/10.1016/S0031-3203(96)00142-2.

BREIMAN, L. 2001. Random forests. *Machine Learning,* 45**,** 5-32. doi: http://dx.doi.org/10.1023/a:1010933404324.

BREIMAN, L., FRIEDMAN, J., OHLSEN, R. & STONE, C. 1984. *Classification and regression trees*, Wadsworth International Group.

BROTHERTON, G. A. & BERDING, N. 1995. Near infra-red spectroscopic applications for milling: Prospects and impications. *Proceedings of the Australian Society of Sugarcane Technologists,* 13**,** 21-19.

BROTHERTON, G. A. & BERDING, N. 1998. At-line analysis of well-prepared cane using near infra-red spectroscopy. *Proceedings of the Australian Society of Sugar Cane Technologists,* 20**,** 34-42.

BSES 1991. *The standard laboratory manual for australian sugar mills : Volume 2 analytical methods and tables,* Indooroopili, Queensland, Australia, BSES.

CAMPOS, G. O., ZIMEK, A., SANDER, J., CAMPELLO, R. J. G. B., MICENKOVÁ, B., SCHUBERT, E., ASSENT, I. & HOULE, M. E. 2016. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery***,** 1-37. doi: http://dx.doi.org/10.1007/s10618-015-0444-8.

CAO, N. 2013. *Calibration optimization and efficiency in near infrared spectroscopy.* Ph.D. Dissertation, Iowa State University.

CEN, H., BAO, Y., HUANG, M. & HE, Y. 2006. Comparison of data pre-processing in pattern recognition of milk powder vis/nir spectra. *In:* LI, X., ZAÏANE, O. R. & LI, Z. (eds.) *Advanced data mining and applications: Second international conference, adma 2006, xi'an, china, august 14-16, 2006 proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

CHE, W. K., SUN, L. J., ZHANG, Q., TAN, W. Y., YE, D. D., ZHANG, D. & LIU, Y. Y. 2018. Pixel based bruise region extraction of apple using vis-nir hyperspectral imaging. *Computers and Electronics in Agriculture,* 146**,** 12-21. doi: http://dx.doi.org/10.1016/j.compag.2018.01.013.

CHEN, D., SHAO, X., HU, B. & SU, Q. 2005. Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra. *Analytical Sciences,* 21**,** 161-166. doi: http://dx.doi.org/10.2116/analsci.21.161.

CIRINO DE CARVALHO, L., DE LELIS MEDEIROS DE MORAIS, C., GOMES DE LIMA, K. M., CUNHA JUNIOR, L. C., MARTINS NASCIMENTO, P. A., BOSCO DE FARIA, J. & HENRIQUE DE ALMEIDA TEIXEIRA, G. 2016. Determination of the geographical origin and ethanol content of brazilian sugarcane spirit using near-infrared spectroscopy coupled with discriminant analysis. *Analytical Methods,* 8**,** 5658-5666. doi: http://dx.doi.org/10.1039/C6AY01325B.

CORTES, C. & VAPNIK, V. 1995. Support-vector networks. *Machine Learning,* 20**,** 273-297. doi: http://dx.doi.org/10.1007/BF00994018.

CUI, C. & FEARN, T. 2017. Comparison of partial least squares regression, least squares support vector machines, and gaussian process regression for a near infrared calibration. *Journal of Near Infrared Spectroscopy,* 25**,** 5-14. doi: http://dx.doi.org/10.1177/0967033516678515.

DAIRAM, N., RAMARU, R., NGEMA, S., SUTAR, N. & MADHO, S. 2016. Sucrose losses across the gledhow evaporators determined using nirs predictions. *Proceedings of the Annual Congress - South African Sugar Technologists' Association***,** 391-405.

DE AGUIAR, P. F., BOURGUIGNON, B., KHOTS, M. S., MASSART, D. L. & PHAN-THAN-LUU, R. 1995. D-optimal designs. *Chemometrics and Intelligent Laboratory Systems,* 30**,** 199-210. doi: http://dx.doi.org/10.1016/0169-7439(94)00076-X.

DONALD, D., COOMANS, D., EVERINGHAM, Y., COZZOLINO, D., GISHEN, M. & HANCOCK, T. 2006. Adaptive wavelet modelling of a nested 3 factor experimental design in nir chemometrics. *Chemometrics and Intelligent Laboratory Systems,* 82**,** 122-129. doi: http://dx.doi.org/10.1016/j.chemolab.2005.05.013.

EGAN, W. J. & MORGAN, S. L. 1998. Outlier detection in multivariate analytical chemical data. *Analytical Chemistry,* 70**,** 2372-2379. doi: http://dx.doi.org/10.1021/ac970763d.

EVERINGHAM, Y. L., LOWE, K. H., DONALD, D. A., COOMANS, D. H. & MARKLEY, J. 2007. Advanced satellite imagery to classify sugarcane crop characteristics. *Agronomy for Sustainable Development,* 27**,** 111-117. doi: http://dx.doi.org/10.1051/agro:2006034.

FAO. 2017. *Faostat* [Online]. Food and Agriculture Organisation of the United Nations. Available: http://faostat.fao.org/ [Accessed 18/07/2017].

FAO. 2019. *Faostat* [Online]. Food and Agriculture Organisation of the United Nations. Available: http://faostat.fao.org/ [Accessed 17/12/2019].

FEARN, T. 2001. Review: Standardisation and calibration transfer for near infrared instruments: A review. *Journal of Near Infrared Spectroscopy,* 9**,** 229-244. doi: http://dx.doi.org/10.1255/jnirs.309.

FEILHAUER, H., ASNER, G. P. & MARTIN, R. E. 2015. Multi-method ensemble selection of spectral bands related to leaf biochemistry. *Remote Sensing of Environment,* 164**,** 57-65. doi: http://dx.doi.org/10.1016/j.rse.2015.03.033.

FIEDLER, F. M., EDYE, L. A. & WATSON, L. J. 2001. The application of discriminant analysis to on-line near infrared spectroscopy of prepared sugar cane. *Proceedings of the Australian Society of Sugar Cane Technologists,* 23**,** 317-321.

FORINA, M., CASOLINO, C. & ALMANSA, E. M. 2003. The refinement of pls models by iterative weighting of predictor variables and objects. *Chemometrics and Intelligent Laboratory Systems,* 68**,** 29-40. doi: http://dx.doi.org/10.1016/S0169-7439(03)00085-6.

FRIEDMAN, J. H. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics,* 29**,** 1189-1232.

FRIEDMAN, J. H. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis,* 38**,** 367-378. doi: http://dx.doi.org/10.1016/S0167-9473(01)00065-2.

GARSON, D. G. 1991. Interpreting neural-network connection weights. *AI Expert,* 6**,** 46-51.

GELADI, P., MACDOUGALL, D. & MARTENS, H. 1985. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy,* 39**,** 491-500. doi: http://dx.doi.org/10.1366/0003702854248656.

GOICOECHEA, H. C. & OLIVIERI, A. C. 2003. A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy. *Journal of Chemometrics,* 17**,** 338-345. doi: http://dx.doi.org/10.1002/cem.812.

GOUNDEN, T. & WALTHEW, D. 2018. Nirs as a tool for improved process monitoring. *Proceedings of the Annual Congress - South African Sugar Technologists' Association***,** 350-356.

GUJRAL, P., AMRHEIN, M., ERGON, R., WISE, B. M. & BONVIN, D. 2011. On multivariate calibration with unlabeled data. *Journal of Chemometrics,* 25**,** 456-465. doi: http://dx.doi.org/10.1002/cem.1389.

GUO, D., ZHU, Q., HUANG, M., GUO, Y. & QIN, J. 2017. Model updating for the classification of different varieties of maize seeds from different years by hyperspectral imaging coupled with a pre-labeling method. *Computers and Electronics in Agriculture,* 142**,** 1-8. doi: http://dx.doi.org/10.1016/j.compag.2017.08.015.

GUTHRIE, J. A. 2005. *Robustness of nir calibrations for assessing fruit quality.* PhD, Central Queensland University.

GUTIÉRREZ, S., TARDAGUILA, J., FERNÁNDEZ-NOVALES, J. & DIAGO, M. P. 2016. Data mining and nir spectroscopy in viticulture: Applications for plant phenotyping under field conditions. *Sensors,* 16**,** 236. doi: http://dx.doi.org/10.3390/s16020236.

HAGEMAN, J., WESTERHUIS, J. & SMILDE, A. 2005. Temperature robust multivariate calibration: An overview of methods for dealing with temperature influences on near infrared spectra. *Journal of Near Infrared Spectroscopy,* 13**,** 53-62. doi: http://dx.doi.org/10.1255/jnirs.30910.1255/jnirs.457.

HAN, J., KAMBER, M. & PEI, J. 2011. *Data mining: Concepts and techniques*, Elsevier.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. H. 2013a. Boosting and additive trees. *The elements of statistical learning.* 2 ed.: Springer.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. H. 2013b. Cross-validation. *The elements of statistical learning.* 2 ed.: Springer.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. H. 2013c. *The elements of statistical learning,* New York, USA, Springer.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. H. 2013d. Neural nets. *The elements of statistical learning.* 2 ed.: Springer.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. H. 2013e. "Off-the-shelf" procedures for data mining. *The elements of statistical learning.* 2 ed.: Springer.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. H. 2013f. Partial least squares. *The elements of statistical learning.* 2 ed.: Springer.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. H. 2013g. Random forests. *The elements of statistical learning.* 2 ed.: Springer.

HOGARTH, D. M. & ALLSOPP, P. G. (eds.) 2000. *Manual of cane growing*: Bureau of Sugar Experiment Stations.

HONG, X.-Z., FU, X.-S., WANG, Z.-L., ZHANG, L., YU, X.-P. & YE, Z.-H. 2019. Tracing geographical origins of teas based on ft-nir spectroscopy: Introduction of model updating and imbalanced data handling approaches. *Journal of Analytical Methods in Chemistry,* 2019**,** 8. doi: http://dx.doi.org/10.1155/2019/1537568.

HUANG, M., TANG, J., YANG, B. & ZHU, Q. 2016. Classification of maize seeds of different years based on hyperspectral imaging and model updating. *Computers and Electronics in Agriculture,* 122**,** 139-145. doi: http://dx.doi.org/10.1016/j.compag.2016.01.029.

ISAKSSON, T. & NAES, T. 1990. Selection of samples for calibration in near-infrared spectroscopy. Part ii: Selection based on spectral measurements. *Applied Spectroscopy,* 44**,** 7. doi: http://dx.doi.org/10.1366/0003702904086533.

JAM, M. N. H. & CHIA, K. S. 2017. Investigating the relationship between the reflected near infrared light and the internal quality of pineapples using neural network. *International Journal on Advanced Science Engineering and Information Technology,* 7. doi: http://dx.doi.org/10.18517/ijaseit.7.4.3143.

JAMES, G., WITTEN, D. & HASTIE, T. 2013. Linear model selection and regularization. *In:* CASELLA, G., FIENBERG, S. & OLKIN, I. (eds.) *An introduction to statistical learning (with applications in r).* New York, NY, USA: Springer.

JIN, H. & LING, C. X. 2005. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering,* 17**,** 299-310. doi: http://dx.doi.org/10.1109/TKDE.2005.50.

KANNAR, D., KITCHEN, J. & O'SHEA, M. 2009. *Process for the manufacture of sugar and other food products*. Australia patent application WO/2009/043100.

KARATZOGLOU, A., SMOLA, A., HORNIK, K. & ZEILEIS, A. 2004. Kernlab - an s4 package for kernel methods in r. *Journal of Statistical Software,* 11**,** 1 - 20. doi: http://dx.doi.org/10.18637/jss.v011.i09.

KEEFFE, E. C. 2013. *Rapid nutrient determination of sugarcane milling by-products using near infrared spectroscopy.* Masters by Research.

KOVALENKO, I. V., RIPPKE, G. R. & HURBURGH, C. R. 2006. Determination of amino acid composition of soybeans (glycine max) by near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry,* 54**,** 3485-3491. doi: http://dx.doi.org/10.1021/jf052570u.

KUHN, M. 2017. Caret: Classification and regression training. 6.0-76 ed.

KUHN, M. & JOHNSON, K. 2013a. Applied predictive modeling. Springer.

KUHN, M. & JOHNSON, K. 2013b. Remedies for severe class imbalance. *Applied predictive modeling.* New York, NY: Springer New York.

LI, J. B., HUANG, W. Q., TIAN, X., WANG, C. P., FAN, S. X. & ZHAO, C. J. 2016. Fast detection and visualization of early decay in citrus using vis-nir hyperspectral imaging. *Computers and Electronics in Agriculture,* 127**,** 582-592. doi: http://dx.doi.org/10.1016/j.compag.2016.07.016.

LIAW, A. & WIENER, M. 2002. Classification and regression by randomforest. *R News,* 2**,** 18-22.

LIN, H.-T., LIN, C.-J. & WENG, R. C. 2007. A note on platt's probabilistic outputs for support vector machines. *Machine Learning,* 68**,** 267-276. doi: http://dx.doi.org/10.1007/s10994-007-5018-6.

LIONNET, G. R. E. 1986. Post-harvest deterioration of whole stalk sugarcane. *South African Sugar Technologists Association.* Durban, South Africa: South African Sugar Association.

LIU, R., CHEN, W.-L., XU, K.-X., QIU, Q.-J. & CUI, H.-X. 2005. Fast outlier detection for milk near-infrared spectroscopy analysis. *Guang pu xue yu guang pu fen xi = Guang pu,* 25**,** 207-210.

LLOYD, T., EASTMENT, S. & MITCHELL, P. 2010. Milling train maceration control utilising nir technology. *Proceedings of the Australian Society of Sugar Cane Technologists,* 32**,** 688 - 695.

LOGGENBERG, K., STREVER, A., GREYLING, B. & POONA, N. 2018. Modelling water stress in a shiraz vineyard using hyperspectral imaging and machine learning. *Remote Sensing,* 10**,** 202. doi: http://dx.doi.org/10.3390/rs10020202.

MACKINTOSH, D. 2000. Sugar milling. *In:* HOGARTH, D. M. & ALLSOPP, P. G. (eds.) *Manual of canegrowing.* Brisbane: Bureau of Sugar Experiment Stations.

MALLET, Y., DE VEL, O. & COOMANS, D. H. 1998. Integrated feature extravtion using adaptive wavelets. *In:* LIU, H. & MOTODA, H. (eds.) *Feature extraction, construction and selection: A data mining perspective.* Dordrecht, The Netherlands Kluer Academic.

MARKLEY, J., GRIFFIN, K., STAUNTON, S., THORBURN, P. & CROWLEY, T. 2009. Increasing in-mill nir effectiveness and communicating data to all sectors for improved decision making in the sugarcane value chain. *Project CSR038.* Sugar Research and Development Corporation.

MARTENS, H., NAES, T. & NAES, T. 1992. *Multivariate calibration*, John Wiley & Sons.

MASSIE, D. R. & NORRIS, K. H. 1965. Spectral reflectance and transmittance properties of grain in the visible and near infrared. *Transactions of the ASAE,* 8**,** 598. doi: http://dx.doi.org/10.13031/2013.40596.

MCCARTHY, S. 2003. *The integration of sensory control for sugar cane harvesters.* Doctor of Phyilosophy, University of Southern Queensland.

MEHMOOD, T., LILAND, K. H., SNIPEN, L. & SÆBØ, S. 2012. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems,* 118**,** 62-69. doi: http://dx.doi.org/10.1016/j.chemolab.2012.07.010.

MEVIK, B.-H., WEHRENS, R. & LILAND, K. H. 2015. Pls: Partial least squares and principal component regression. 2.5-0 ed.

MEYER, D., DIMITRIADOU, E., HORNIK, K., WEINGESSEL, A. & LEISCH, F. 2015. E1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien. 1.6-8 ed.

MILLER, C. E. 1993. Sources of non-linearity in near infrared methods. *NIR news,* 4**,** 3-5. doi: http://dx.doi.org/10.1255/nirn.216.

MITCHELL, M. 1999. *An introduction to genetic algorithms,* Cambridge, Massachusetts, The MIT Press.

MUCHOW, R. C., ROBERTSON, M., WOOD, A. & KEATING, B. A. 1996. Effect of nitrogen on the time-course of sucrose accumulation in sugarcane. *Field Crops Research,* 47**,** 143-153. doi: http://dx.doi.org/10.1016/0378-4290(96)00022-6.

NAES, T. & ISAKSSON, T. 1989. Selection of samples for calibration in near-infrared spectroscopy. Part i: General principles illustrated by example. *Applied Spectroscopy,* 43**,** 8. doi: http://dx.doi.org/10.1366/0003702894203129.

NAWAR, S. & MOUAZEN, A. 2017. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line vis-nir spectroscopy measurements of soil total nitrogen and total carbon. *Sensors,* 17**,** 2428. doi: http://dx.doi.org/10.3390/s17102428.

NAWI, N. M., CHEN, G. & JENSEN, T. 2014. In-field measurement and sampling technologies for monitoring quality in the sugarcane industry: A review. *Precision Agriculture,* 15**,** 684-703. doi: http://dx.doi.org/10.1007/s11119-014-9362-9.

NAWI, N. M., CHEN, G., JENSEN, T. & MEHDIZADEH, S. A. 2013. Prediction and classification of sugar content of sugarcane based on skin scanning using visible and shortwave near infrared. *Biosystems Engineering,* 115**,** 154-161. doi: http://dx.doi.org/10.1016/j.biosystemseng.2013.03.005.

NI, W., NØRGAARD, L. & MØRUP, M. 2014. Non-linear calibration models for near infrared spectroscopy. *Analytica Chimica Acta,* 813**,** 1-14. doi: http://dx.doi.org/10.1016/j.aca.2013.12.002.

NIKZAD-LANGERODI, R., LUGHOFER, E., CERNUDA, C., REISCHER, T., KANTNER, W., PAWLICZEK, M. & BRANDSTETTER, M. 2018. Calibration model maintenance in melamine resin production: Integrating drift detection, smart sample selection and model adaptation. *Analytica Chimica Acta,* 1013**,** 1-12. doi: http://dx.doi.org/10.1016/j.aca.2018.02.003.

NIU, C., YUAN, Y., GUO, H., WANG, X., WANG, X. & YUE, T. 2018. Recognition of osmotolerant yeast spoilage in kiwi juices by near-infrared spectroscopy coupled with chemometrics and wavelength selection. *RSC Advances,* 8**,** 222-229. doi: http://dx.doi.org/10.1039/C7RA12266G.

NØRGAARD, L., SAUDLAND, A., WAGNER, J., NIELSEN, J. P., MUNCK, L. & ENGELSEN, S. B. 2000. Interval partial least-squares regression (ipls): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy,* 54**,** 413-419. doi: http://dx.doi.org/10.1366/0003702001949500.

NORRIS, K. Extracting information from spectrophotometric curves. Predicting chemical composition from visible and near-infrared spectra. Food research and data analysis: proceedings from the IUFoST Symposium, September 20-23, 1982, Oslo, Norway/edited by H. Martens and H. Russwurm, Jr, 1983. London: Applied Science Publishers, 1983.

NORRIS, K. & WILLIAMS, P. 1984. Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particle size. *Cereal Chemistry*.

O'SHEA, M. G., STAUNTON, S. & BURLEIGH, M. 2010. Implementation of on-line near infrared (nir) technologies for the analysis of cane, bagasse and raw sugar in sugar factories to improve performance. *Proceedings of the International Society of Sugar Cane Technologists,* 27**,** 15.

O'SHEA, M. G., STAUNTON, S. P., DONALD, D. & SIMPSON, J. 2011. Developing laboratory near infra-red (nir) instruments for the analysis of sugar factory products. *Proceedings of the Australian Society of Sugar Cane Technologists,* 33**,** 1-8.

OLDEN, J. D. & JACKSON, D. A. 2002. Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling,* 154**,** 135-150. doi: http://dx.doi.org/10.1016/S0304-3800(02)00064-9.

OLDEN, J. D., JOY, M. K. & DEATH, R. G. 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling,* 178**,** 389-397. doi: http://dx.doi.org/10.1016/j.ecolmodel.2004.03.013.

OSBORNE, B., FEARN, T. & HINDLE, P. H. 1993a. Physics of the interaction of radiation with matter. *Practical nir spectroscopy with applications in food and beverage analysis.* Harlow, UK: Longman Scientific and Technical.

OSBORNE, B., FEARN, T. & HINDLE, P. H. 1993b. *Practical nir spectroscopy with applications in food and beverage analysis,* Harlow, UK, Longman Scientific and Technical.

OSTATEK-BOCZYNSKI, Z. A., PURCELL, D. E., KEEFFE, E. C., MARTENS, W. N. & O'SHEA, M. G. 2013. Rapid determination of carbon, nitrogen, silicon, phosphorus, and potassium in sugar mill by-products, mill mud, and ash using near infrared spectroscopy. *Communications in Soil Science and Plant Analysis,* 44**,** 1156-1166. doi: http://dx.doi.org/10.1080/00103624.2012.756004.

OXELY, J., FONG CHONG, B., SANT, G. G. & O'SHEA, M. G. 2012. Accelerating the characterisation of sugarcane biomass using near-infrared (nir) spectroscopic techniques. *Proceedings of the Australian Society of Sugar Cane Technologists,* 34**,** 9.

ÖZESMI, S. L. & ÖZESMI, U. 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling,* 116**,** 15-31. doi: http://dx.doi.org/10.1016/S0304-3800(98)00149-5.

PIERNA, J. A. F., LECLER, B., CONZEN, J. P., NIEMOELLER, A., BAETEN, V. & DARDENNE, P. 2011. Comparison of various chemometric approaches for large near infrared spectroscopic data of feed and feed products. *Analytica Chimica Acta,* 705**,** 30-34. doi: http://dx.doi.org/10.1016/j.aca.2011.03.023.

PLATT, J. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers,* 10**,** 61-74. doi: http://dx.doi.org/10.1.1.41.1639.

POLLOCK, J., O'HARA, I. M. & GRIFFIN, K. 2007. Aligning the drivers in the value chain—a new cane payment system for mackay sugar. *Proceedings of the Australian Society of Sugar Cane Technologists,* 29**,** 1-8.

PURCELL, D., OSTATEK-BOCZYNSKI, Z. A., KEEFFE, E. C., MARTENS, W. N. & O'SHEA, M. G. 2012. Development of near infrared (nir) spectroscopic methods to predict carbon, nitrogen, silicon, phosphorus and potassium levels in mill by-products. *Proceedings of the Australian Society of Sugar Cane Technologists,* 34**,** 8.

R CORE TEAM 2016. R: A language and environment for statistical computing. 3.3.1 ed. Vienna, Austria: R Foundation for Statistical Computing.

R CORE TEAM 2017. R: A language and environment for statistical computing. 3.4 ed. Vienna, Austria: R Foundation for Statistical Computing.

RAMÍREZ-MORALES, I., RIVERO, D., FERNÁNDEZ-BLANCO, E. & PAZOS, A. 2016. Optimization of nir calibration models for multiple processes in the sugar industry. *Chemometrics and Intelligent Laboratory Systems,* 159**,** 45-57. doi: http://dx.doi.org/10.1016/j.chemolab.2016.10.003.

RIDGEWAY, G. 2015. Gbm: Generalized boosted regression models. 2.1.1 ed.

RINNAN, Å., BERG, F. V. D. & ENGELSEN, S. B. 2009. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry,* 28**,** 1201-1222. doi: http://dx.doi.org/10.1016/j.trac.2009.07.007.

RODRÍGUEZ-ZÚÑIGA, U. F., FARINAS, C. S., CARNEIRO, R. L., SILVA, G. M., CRUZ, A. J. G., LIMA CAMARGO GIORDANO, R., CAMPOS GIORDANO, R. & ARRUDA RIBEIRO, M. P. 2014. Fast determination of the composition of pretreated sugarcane bagasse using near-infrared spectroscopy. *BioEnergy Research,* 7**,** 1441-1453. doi: http://dx.doi.org/10.1007/s12155-014-9488-7.

SABATIER, D. R., MOON, C. M., MHORA, T. T., RUTHERFORD, R. S. & LAING, M. D. 2014. Near-infrared reflectance (nir) spectroscopy as a high-throughput screening tool for pest and disease resistance in a sugarcane breeding programme. *86th Annual Congress of the South African Sugar Technologists' Association (SASTA 2013), Durban, South Africa, 6-8 August 2013***,** 101-106.

SANAEIFAR, A., BAKHSHIPOUR, A. & DE LA GUARDIA, M. 2016. Prediction of banana quality indices from color features using support vector regression. *Talanta,* 148**,** 54-61. doi: http://dx.doi.org/10.1016/j.talanta.2015.10.073.

SAVITZKY, A. & GOLAY, M. J. E. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry,* 36**,** 1627-1639. doi: http://dx.doi.org/10.1021/ac60214a047.

SAXENA, P., SRIVASTAVA, R. P. & SHARMA, M. L. 2010. Impact of cut to crush delay and bio-chemical changes in sugarcane *Australian Journal of Crop Science,* 4**,** 692-699.

SCRUCCA, L. 2013. Ga: A package for genetic algorithms in r. *2013,* 53**,** 37. doi: http://dx.doi.org/10.18637/jss.v053.i04.

SENTHILKUMAR, T., JAYAS, D. S., WHITE, N. D. G., FIELDS, P. G. & GRÄFENHAN, T. 2016. Detection of fungal infection and ochratoxin a contamination in stored barley using near-infrared hyperspectral imaging. *Biosystems Engineering,* 147**,** 162-173. doi: http://dx.doi.org/10.1016/j.biosystemseng.2016.03.010.

SHENK, J. S., WORKMAN, J. J. & WESTERHAUS, M. O. 2008. Applications of nir spectroscopy to agricultural products. *In:* BURNS, D. A. & CUIRCZAK, E. W. (eds.) *Handbook of near-infrared analysis.* Third ed. 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL, USA: CRC Press/Taylor & Francis Group.

SHETTY, N., GISLUM, R., JENSEN, A. M. D. & BOELT, B. 2012. Development of nir calibration models to assess year-to-year variation in total non-structural carbohydrates in grasses using plsr. *Chemometrics and Intelligent Laboratory Systems,* 111**,** 34-38. doi: http://dx.doi.org/10.1016/j.chemolab.2011.11.004.

SIMPSON, J., STAUNTON, S. P. & O'SHEA, M. G. 2011. A review of nir applications for process control purposes. *Proceedings of the Australian Society of Sugar Cane Technologists,* 33**,** 8.

SINGH, C. B., CHOUDHARY, R., JAYAS, D. S. & PALIWAL, J. 2008. Wavelet analysis of signals in agriculture and food quality inspection. *Food and Bioprocess Technology,* 3**,** 2. doi: http://dx.doi.org/10.1007/s11947-008-0093-7.

SMOLA, A. & VAPNIK, V. 1997. Support vector regression machines. *In:* MOZER, M. C., JORDAN, J. I. & PETSCHE, T. (eds.) *Neural information processing systems.* Cambridge, Massachusetts, USA: MIT Press.

SOLOMON, S. 2009. Post-harvest deterioration of sugarcane. *Sugar Tech,* 11**,** 109-123. doi: http://dx.doi.org/10.1007/s12355-009-0018-4.

SONG, W., WANG, H., MAGUIRE, P. & NIBOUCHE, O. Differentiation of organic and non-organic apples using near infrared reflectance spectroscopy—a pattern recognition approach. SENSORS, 2016 IEEE, 2016. IEEE, 1-3.

SOROL, N., ARANCIBIA, E., BORTOLATO, S. A. & OLIVIERI, A. C. 2010. Visible/near infrared-partial least-squares analysis of brix in sugar cane juice: A test field for variable selection methods. *Chemometrics and Intelligent Laboratory Systems,* 102**,** 100-109. doi: http://dx.doi.org/10.1016/j.chemolab.2010.04.009.

SRA. 2014. Sugar research australia strategic plan 2013/14 - 2017/18. Available: http://www.sugarresearch.com.au/icms_docs/188322_Strategic_Plan_201314-201718.pdf [Accessed 2016-03-02].

SRA. 2015. Sugar research australia strategic plan 2013/14 - 2017/18: 2015 update. Available: http://www.sugarresearch.com.au/icms_docs/220936_Strategic_Plan_2015_Update.pdf [Accessed 2016-03-02].

STAUNTON, S., LETHBRIDGE, P., GRIMLEY, S., STREAMER, R., ROGERS, J. & MACKINTOSH, D. 1999. On-line cane analysis by near infra-red spectroscopy. *Proceedings of the Australian Society of Sugar Cane Technologists,* 21**,** 20-27.

STAUNTON, S., MACKINTOSH, D. & PEATEY, G. 2004. The application of network nir calibration equations at the maryborough sugar factory. *Proceedings of the Australian Society of Sugar Cane Technologists,* 26**,** 1-14.

STAUNTON, S. & WARDROP, K. 2006. Development of an online bagasse analysis system using nir spectroscopy. *Austrailan Society of Sugar Cane Technologists.*

STEVENS, A. & RAMIREZ-LOPEZ, L. 2013. An introduction to the prospectr package. R package. *R package Vignette.* 0.1.3 ed.

SUYKENS, J. A., VAN GESTEL, T. & DE BRABANTER, J. 2002. *Least squares support vector machines*, World Scientific.

TANGE, R. I., RASMUSSEN, M. A., TAIRA, E. & BRO, R. 2015. Application of support vector regression for simultaneous modelling of near infrared spectra from multiple process steps. *Journal of Near Infrared Spectroscopy,* 23**,** 75-84. doi: http://dx.doi.org/10.1255/jnirs.1149.

THISSEN, U., PEPERS, M., ÜSTÜN, B., MELSSEN, W. J. & BUYDENS, L. M. C. 2004a. Comparing support vector machines to pls for spectral regression applications. *Chemometrics and Intelligent Laboratory Systems,* 73**,** 169-179. doi: http://dx.doi.org/10.1016/j.chemolab.2004.01.002.

THISSEN, U., ÜSTÜN, B., MELSSEN, W. J. & BUYDENS, L. M. C. 2004b. Multivariate calibration with least-squares support vector machines. *Analytical Chemistry,* 76**,** 3099-3105. doi: http://dx.doi.org/10.1021/ac035522m.

TRYGG, J. & WOLD, S. 1998. Pls regression on wavelet compressed nir spectra. *Chemometrics and Intelligent Laboratory Systems,* 42**,** 209-220. doi: http://dx.doi.org/10.1016/S0169-7439(98)00013-6.

TRYGG, J. & WOLD, S. 2002. Orthogonal projections to latent structures (o-pls). *Journal of Chemometrics,* 16**,** 119-128. doi: http://dx.doi.org/10.1002/cem.695.

TULIP, J. & WILKINS, K. 2004. Dirt level estimation in prepared cane using vis/vnir spectroscopy. *Proceedings of the Australian Society of Sugar Cane Technologists,* 26**,** 58-58.

ÜSTÜN, B., MELSSEN, W. J. & BUYDENS, L. M. C. 2007. Visualisation and interpretation of support vector regression models. *Analytica Chimica Acta,* 595**,** 299-309. doi: http://dx.doi.org/10.1016/j.aca.2007.03.023.

VALDERRAMA, P., BRAGA, J. W. B. & POPPI, R. J. 2007a. Validation of multivariate calibration models in the determination of sugar cane quality parameters by near infrared spectroscopy. *Journal of the Brazilian Chemical Society,* 18**,** 259-266.

VALDERRAMA, P., BRAGA, J. W. B. & POPPI, R. J. 2007b. Variable selection, outlier detection, and figures of merit estimation in a partial least-squares regression multivariate calibration model. A case study for the determination of quality parameters in the alcohol industry by near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry,* 55**,** 8331-8338. doi: http://dx.doi.org/10.1021/jf071538s.

VAN HEERDEN, P. D. R., EGGLESTON, G. & DONALDSON, R. A. 2014. Ripening and postharvest deterioration. *In:* MOORE, A. D. & BOTHA, F. C. (eds.) *Sugarcane physiology, biochemistry & functional biology.* Wiley Blackwell.

VENABLES, W. N. & RIPLEY, B. D. 2002. *Modern applied statistics with s,* New York, Springer.

VERT, J. P., TSUDA, K. & SCHÖLKOPF, B. 2004. A primer on kernel methods. *In:* SCHÖLKOPF, B., TSUDA, K. & VERT, J. P. (eds.) *Kernel methods in computational biology.* Cambridge, MA, USA: MIT Press.

VISCARRA ROSSEL, R. A. & BEHRENS, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma,* 158**,** 46-54. doi: http://dx.doi.org/10.1016/j.geoderma.2009.12.025.

WALFORD, S. 2019. Near infrared spectroscopy: Rethinking the analysis of sugarcane factory streams.

WANG, D., DOWELL, F. E., RAM, M. S. & SCHAPAUGH, W. T. 2004. Classification of fungal-damaged soybean seeds using near-infrared spectroscopy. *International Journal of Food Properties,* 7**,** 75-82. doi: http://dx.doi.org/10.1081/JFP-120022981.

WANG, X., YE, H. J., LI, Q. T., XIE, J. C., LU, J. J., XIA, A. L. & WANG, J. 2010. Determination of brix and pol in sugar cane juice by using near infrared spectroscopy coupled with bp-ann. *Guang Pu Xue Yu Guang Pu Fen Xi/Spectroscopy and Spectral Analysis,* 30**,** 1759-1762. doi: http://dx.doi.org/10.3964/j.issn.1000-0593(2010)07-1759-04.

WILLIAMS, P., DARDENNE, P. & FLINN, P. 2017. Tutorial: Items to be included in a report on a near infrared spectroscopy project. *Journal of Near Infrared Spectroscopy,* 25**,** 85-90. doi: http://dx.doi.org/10.1177/0967033517702395.

WOLD, S., SJÖSTRÖM, M. & ERIKSSON, L. 2001. Pls-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* 58**,** 109-130. doi: http://dx.doi.org/10.1016/S0169-7439(01)00155-1.

XIAOBO, Z., JIEWEN, Z., POVEY, M. J. W., HOLMES, M. & HANPIN, M. 2010. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta,* 667**,** 14-32. doi: http://dx.doi.org/10.1016/j.aca.2010.03.048.

YANG, M., CHEN, Q., KUTSANEDZIE, F. Y. H., YANG, X., GUO, Z. & OUYANG, Q. 2017. Portable spectroscopy system determination of acid value in peanut oil based on variables selection algorithms. *Measurement,* 103**,** 179-185. doi: http://dx.doi.org/10.1016/j.measurement.2017.02.037.

ZHANG, L. G., ZHANG, X., NI, L. J., XUE, Z. B., GU, X. & HUANG, S. X. 2014. Rapid identification of adulterated cow milk by non-linear pattern recognition methods based on near infrared spectroscopy. *Food Chemistry,* 145**,** 342-348. doi: http://dx.doi.org/10.1016/j.foodchem.2013.08.064.