final report

Project Code: A.MQT.0052
Prepared by: Katja Rosenvold
Date published: November 2012

PUBLISHED BY
Meat and Livestock Australia Limited
Locked Bag 991
NORTH SYDNEY NSW 2059

Accreditation of VISNIR spectroscopy to support MSA grading of beef

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government and contributions from the Australian Meat Processor Corporation to support the research and development detailed in this publication.

Abstract

NIR-calibration models were developed for prediction of five meat attributes used in the MSA grading system (i.e. meat colour, fat colour, ossification score, marble score and ultimate pH). The data collected over 9 days at two plants was split in two datasets each with approximately 50% of the samples evenly distributed among collection days. The first dataset was used to fit the models and the second to validate them. The best performance, based on predictions of the validation data set, was marble score (R^2 = 0.79) followed by ossification (R^2 = 0.78), pH (R^2 = 0.73), meat colour (R^2 = 0.61) and fat colour (R^2 = 0.60). Previously[1], NIR predictions were used on attributing the final star grade of carcasses, when 96% of carcasses were correctly allocated, and the R^2 values for prediction of meat colour, marble score and pH were equal to 0.69, 0.81 and 0.76, respectively. Overall NIR was able to predict correctly the five attributes although with significant variability in the predictions, probably resulting from sampling variability and increase accuracy. Thus an in-house trial to identify ways to reduce sampling variability prior the decision on whether proceed to the next milestone or otherwise is recommended.

Project objectives

1.1 To develop and validate VISNIR calibration models for predicting MSA meat colour, fat colour, ossification score, marble score and ultimate pH in chilled beef carcasses that have prediction accuracies comparable to current conventional MSA measurement methods.

1.2 To attain AUS-MEAT accreditation for the VISNIR measurement of meat colour, fat colour, ossification score, marble score & ultimate pH to support MSA grading of beef.

The outcome will be scientifically substantiated, this work will generate statistically significant proof that Near Infra-Red (NIR) can objectively and reliably predict the required parameters, thus eliminating plant to plant and grader to grader variability and the development of an NIR based method for objective measurement of marbling, ossification, pHu and colour of beef carcases in the chiller.

[1] Project: No RE-221971.

Success in achieving milestone

The NIR project "Accreditation of VISNIR spectroscopy to support MSA grading of beef" aims to develop calibrations to predict five MSA meat attributes from NIR spectra. This requires the collection of two sets of data: NIR spectra and reference measurements. In this project, the reference measurements were the average score for each carcass as evaluated by 3 highly experienced MSA graders. The data collection (reference and NIR) was carried out at Teys (Beenleigh, AU) and Cassino (Casino, AU); both plants processing a variety of cattle, which allowed a wide spread of variation in MSA scores for the attributes of interest. The collected data set spans a wide spread of variation not only in scores but also in plant environments (e.g. chillers, alleys) allowing the development of more representative calibrations that could be applied at different plants. The models for meat colour, fat colour, ossification score, marble score and ultimate pH were developed in this milestone and are herein reported.

Methods

Selection of spectra and pre-processing methods for the calibration models: Spectra were collected across the carcass in 6 different sites: 1 site (the eye muscle) for the three meat quality attributes (i.e. meat colour, marble score and ultimate pH), 4 ossification sites, and 1 site for fat colour (see report for Milestone 2 for details). In the eye muscle, 4 spectra were collected over the meat surface and models could be developed using only one spectrum or average of 2 up to 4 spectra. For each ossification site, two spectra were collected, one on the cap and the other on bone.

The development of NIR calibrations also involves the use of pre-processing methods to reduce effect of sampling variation and other unwanted artefacts associated with the data collection.

All combined, there are a series of possibilities to fit models that includes number of spectra to be averaged (i.e. meat), position where the spectra were collected (i.e. cap vs. bone in ossification sites) and pre-processing methods to be used. Thus a screening methodology was implemented to identify the most promising approach (e.g. number of spectra to be averaged in the meat, the best position to collect spectra for ossification and pre-processing methods to be used), where over 100 models were evaluated for each attribute.

After a series of combinations was screened, the most promising models were further investigated, by selecting spectral regions of interest, changing combinations of pre-processing methods, applying different modelling methodologies including linear, non-linear, and classification models and eliminating outliers. Outliers, defined as samples with large values of leverage (i.e. a measure of the influence of the sample in the calibration) and Q value (residual error of the projection), which are distinctly different from the bulk of the samples. The pre-processing methods investigated were[2]: standard normal variates (SNV), orthogonal signal correction (OSC), generalized least squares weighting (GLSW), multiplicative scatter correction (MSC), extend scatter correction (EMSC), normalization by area (Norm), auto-scaling, mean centering (MNCN), 1st and 2nd derivatives (1st Der, 2nd Der). These applied to the measured reflectance (Refl) or to absorbance (Abs=-$\log_{10}${Reflectance}).

Averaging of scans in the eye muscle: Using the average of 3 spectra enhances the model performance. The 3 spectra that were most similar were used to avoid which any unusual spectra would affect the average. The three spectra were selected using a leave-one-out cross validation for each sample: *i.* one spectrum is left out the other three are used to fit a principal components analysis with one factor; and *ii.* then the spectra left out is projected into this principal component; *iii.* the Q value (residual error of the projection) is then estimated for this projected spectrum. Thus four values

[2] Wold, s. et al. (1998) Chemom. Intell Lab Sys (44) 175.Rinnan, A, et al (2009) TrAC (28) 1201. http://wiki.eigenvector.com/index.php?title=Advanced_Preprocessing:_Multivariate_Filtering

of Q are estimated for each muscle and the one with the largest value of Q is eliminated. The remaining three were averaged and used in the data analysis.
Modelling: Partial least squares[3] (PLS) and PLS discriminant analysis[3] (PLS-DA) were used as the main multivariate linear models for prediction or classification, respectively. Linear discriminant analysis[4] (LDA) was also used for classification. Neural networks, flexible discriminant analysis and generalized additive models were investigated as non-linear models[4]. Hierarchical modelling (HM) was also evaluated, where two approaches were used: *i.* a classification model was fitted to discriminate the spectra among ranges of a given trait and then for each class (i.e. range) a linear model is fitted to predict that trait. When using HM for ossification, spectra are first classified among those carcasses that are lower than 160, between 160 and 200 and higher than 200. Then for each range a model is fitted and the final ossification score is predicted; *ii.* a classification model was used to classify among ranges, and within each range spectra are re-classified using a range specific classification model. Colour was also predicted from the colour parameters L*, a*, b*, hue and saturation index calculated with the visible part of the spectra[5].

Accuracy of the reference measurements[6]: Three MSA scores were assigned to each carcass, except for pH, where only two pH measurements were available for most of the carcasses. For the carcasses where three scores were available the SEP and R^2 were estimated by: *i.* taking the average of two sets of scores (from MSA graders) and using it as reference measurement; *ii.* using the third set of scores (from plant grader) as independent measurement; and then estimating SEP and R^2. For pH, SEP and R^2 were estimated considering one measurement as reference and the other as independent measurement. In this study it was expected that the error estimate calculated for the conventional MSA reference scoring system was under estimate since: the 3 MSA graders were all highly experienced and as such not representative of the "average" company grader; the 3 MSA graders scored each carcass at the same time and their scores cannot be considered strictly independent. Indeed, it was considered the more important priority that the reference measurements were assessed as accurately as possible since all NIR calibrations are based on these reference measurements. Therefore in those challenging situations a discussion among the three graders would help to provide a more accurate score to the carcass.

Calibration and validation sets: The dataset (NIR spectra and reference measurements) was split in two datasets (calibration and validation) with approximately 50% of the original data (when possible) in each set. These two sets were chosen to describe evenly all the collections days. The validation set was only used when the final model for each attribute had been reached and was not used on the decision making process during the modelling. The distribution of samples between calibration and validation datasets for the five attributes is shown in Tables 1A to 5A, respectively, in the Appendix. For ossification, preliminary tests showed that the measurements on the feather bone (site 4) were needed to improve the model performance. At Teys, data could not be collected on the feather bones, thus only data from Cassino was used on the development of a model for ossification. Still the data from Cassino present a wide variation in ossification that allowed fitting calibration models. The pH measurements obtained in the first two days at Teys showed a high level of variability compared to the other days as shown in Figure 1A in the Appendix. Hence, these data were excluded from the analysis.

[3] Nadler, B. and Coifman, RR (2005) J. Chemometrics (19), 45. Westehuis, JA et al (2008) Metabolomics (4) 81.
[4] Hastie T. et al. (2008) The Elements of Statistical Learning(…) 2nd Ed. Springer.
[5] (1991) Guidelines for Meat Color Evaluation, America Meat Science Association (p.11). Little, A.C. (1975) Off on tangent, J. of Food Sci. V40 (p.2). http://www-cvrl.ucsd.edu/cmfs.htm. http://www.hunterlab.com/pdf/color.pdf.
[6] Standard Practices for Infrared Multivariate Quantitative Analysis (STD. ASTM E1655) 1999.

Summary of results

- Calibration models have been developed for MSA meat colour (R^2 = 0.60), fat colour (R^2 = 0.61), ossification score (R^2 = 0.78), marble score (R^2 = 0.79) and ultimate pH (R^2 = 0.73) in chilled beef carcasses (Table 1).
- Colour: NIR assigned correctly 84% of carcasses within the interval of ± 1 score of the reference score for meat colour and 86% in the same interval for fat colour (Table 2).
- Ossification: NIR assigned correctly 74% of carcass within ± 2 steps of the reference score for ossification in the range 100-200 and 82% within ± 2 steps in the range 230-590 (Table 2).
- Marbling: NIR assigned correctly 84% of carcasses within of ± 100 units of marble score (Table 2).
- pH: NIR classified correctly 79% of carcasses with pH≥5.71 and 87% of those with pH < 5.71 (Table 3).

Table 1 Performance measures of the calibration models calculated from predictions of the validation data set (see Table 6A in the Appendix for full description).

		Meat colour	Fat colour	Ossification	MSAMB	pH
Validation	n	271	362	145	371	297
	SEV	0.86	1.04	31	84	0.12
	R^2	0.60	0.61	0.78	0.79	0.73
Base*	R^2			0.70	0.70	
Experienced*	R^2			0.85	0.85	

SEV: Standard error of validation[7]. n: number of samples; * MSA and/or AUSMEAT grading standards that a grader must achieve to gain/retain accreditation (from the tender document).

Table 2 Distribution of samples from validation dataset predicted in the reference meat colour score or in the consecutive scores.

		NIR predicted	NIR predicted (accumulative)	Current standard for MSA graders*
Meat colour	Reference score	42%	42%	65%
	Reference score ± 1	42%	84%	25%
	Reference score ± 2	14%	98%	10%
	Reference score ≥± 3	2%	100%	
Fat colour	Reference score	45%	45%	
	Reference score ± 1	41%	86%	
	Reference score ± 2	12%	98%	
	Reference score ≥± 3	2%	100%	
Ossification 100-200	Reference score	16%	16%	70%
	Reference score ± 1 step#	38%	54%	30%
	Reference score ± 2 steps	20%	74%	
	Reference score ≥± 3 steps	26%	100%	
Ossification 230-590	Reference score			70%
	Reference score ± 1 step	53%	53%	30%
	Reference score ± 2 steps	29%	82%	
	Reference score ≥± 3 steps	18%	100%	
Marble score	Reference score ± 50	57%	57%	70%
	Reference score ± 100	27%	84%	30%
	Reference score > 100	16%	100%	

One step is defined as the difference between two consecutive MSA ossification scores. * MSA and/or Aus-meat grading standards that a grader must achieve to gain/retain accreditation (from the tender document).

Table 3 Classification rate of NIR predicted pH using MSA cut-off value (pH ≥ 5.71).

			Measured (%)	
			< 5.71	≥5.71
Predicted NIR(%)	Reference	< 5.71	**98**	8
		≥5.71	2	**92**
		N	140	106
	Validation	< 5.71	**87**	21
		5.71	13	**79**
		n	186	111

[7] Standard Practices for Infrared Multivariate Quantitative Analysis (STD. ASTm E1655) 1999

Discussion

Meat colour: Figure 1 shows the distribution of NIR predictions for each meat colour score showing a high correlation ($R^2 = 0.98$) between the median of predictions in each class and the reference values. This indicates that NIR is able to capture the meat colour. Figure 1 also shows a high variability on predictions within each class, identified by the length of the boxes (the shorter the box, the closer the predictions are from the reference value, and more accurate is the model). Table 2 presents the distribution of predictions according the reference score and how close the predictions are from it. NIR was able to predict correctly 84% of the samples within an interval of ± 1 unit of the reference score, where 42% was predicted as the reference score. In the MSA grading system, carcasses scoring 4 and above in meat colour are downgrade[8]. NIR classified 29% of carcasses scoring 4 as being less than 4 (18% to be predicted as 3 plus 11% to be predicted below 3, Table 4). While 8% of carcasses scoring over 4 were predicted as lower than 4, Table 4. A grader to gain/retain accreditation in meat colour is allowed to classify 25% of carcasses as ± 1 of the reference score and 10% as ±2 of the reference score (Table 2). Thus there is 17.5% of chance[9] that a grader misclassifies a carcass scoring 4 as being below 4 compared to 29% when using NIR.

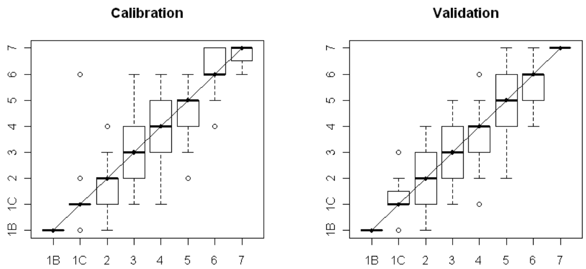


Figure 1 Distribution of meat colour predictions from NIR according the reference value. Each box shows the middle of 50% of the data. The maximum and minimum values (excluding 'outliers') are shown with dashes '-', while dots are 'outlier' samples, whose values are more than 1.5 times the difference between 3rd and 1st quartiles.

Table 4 Rate of classification (%) on meat colour according the colour threshold of 4.

		Graded			
		<3	**3**	**4**	**≥5**
Predicted NIR	**<3**	**80**	30	11	4
	3	11	**25**	18	4
	4	9	44	**55**	21
	5	0	1	16	**71**
	n	46	64	93	68

Fat colour: The best model for prediction of fat colour was a classification model (LDA) applied to estimated values of L*, a*, b*, hue and saturation index. Indeed, the fat colour showed a significant correlation with the parameter b* (yellowness; Figure 2). The same is observed for the NIR predictions (Figure 3). The median in both figures is significantly correlated with fat colour indicating that NIR is sensitive to components giving colour to fat, but also show high variability in the predictions. NIR predicted correctly 86% of the carcasses in the interval ± 1 of the reference score (Table 2), indicating that carcass with a dark fat colour (≥ 7) is unlike to be predicted as light fat colour (≤ 3).

---

8 MSA® Standards Manual for Beef Grading (2012) Meat Standards Australia, ISBN 1 74036 556 3

9 17.5% = 25/2 + 10/2 assuming a symmetric distribution

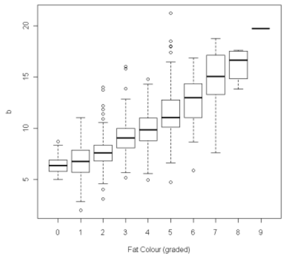


Figure 2 Relationship between colour parameter b* (yellowness) estimated from visible part of the spectra and fat colour (see Figure 1 for description of plot).

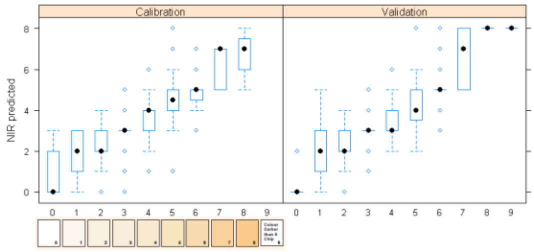


Figure 3 Distribution of fat colour predictions from NIR according the reference value. The colour scale is an illustration and does not reflect the true MSA colour scale (see Figure 1 description of the plot).

Ossification: The pre-screen of models for ossification found that scanning the cap, in general, gave better results than placing the probe over the bone (R^2 increased from 0.6 to 0.8, see Table 5 for example). This indicates that the probe should be positioned on the cap, the region used by graders to evaluate ossification. Two approaches were investigated to develop a calibration for ossification: full-range model and step-wise model. The full range model was obtained by merging the spectra collected from three of the four scanned sites for ossification (sacral, lumbar and feather bones) and fitting a model for the full ossification range. The step-wise model involved a hierarchical modelling (see Methods for description). The full range approach had a R^2 equal to 0.78 (Table 1 and Figure 1A in the Appendix), while the step-wise model achieved a R^2 equal to 0.87 (Table 6A in the Appendix). However, the full range model is recommended as it only uses one model and its performance is comparable to the step-wise model. The accuracy of the full-range model for the ossification ranges between 100-200 and 230-590 is presented in Tables 2, respectively, with worst performance found for the 100-200 range. This difference in performance might be related to sampling variability (see discussion on source of sampling variability).

Table 5 R^2 values for selected models fitted to predict ossification scoring over 200 from NIR spectra collected at the feather bones (site 4).

	Preprocessing	Cap	Bone
Absorbance	MNCN	0.81	0.63
	Norm/MNCN	0.85	0.65
	SNV/MNCN	0.84	0.54
Reflectance	MNCN	0.88	0.67
	Norm/MNCN	0.83	0.60
	SNV/MNCN	0.84	0.55

Marble score: The best predictions for the meat attributes was found for marble score (R^2 = 0.79), where NIR was able to classify correctly 57% of carcasses within ± 50 marbling units and 84% ± 100 marbling units, Table 2. For a grader to gain/retain MSA accreditation, for marbling they must be able to classify correctly 70% of carcasses within ± 50 marbling units from the reference score and 100% between ± 100 units of the reference score (Table 2).

pH: The threshold of 5.7 was used in the MSA grading system to downgrade the carcasses. Thus pH predictions were evaluated according four pH ranges: pH < 5.65, 5.65 to 5.70, 5.71 to 5.74 and pH ≥ 5.75 (Figure 4). NIR correctly classified 82% of carcasses with pH ≥ 5.75, which was comparable to the reference measurements (86%). NIR misclassified 37% of carcasses in the interval 5.71 to 5.75 as being between 6.65-6.70, in comparison, to 19% misclassified in the reference measurements.

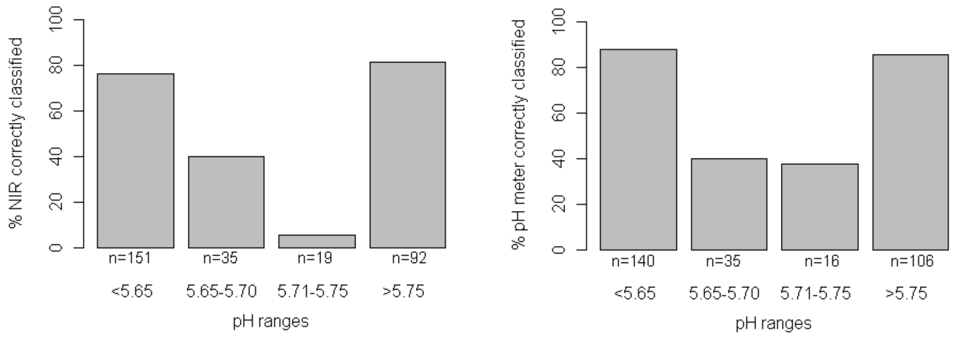


Figure 4 Classification rate according pH ranges when using NIR or when comparing the two pH meters.

Source of variability

NIR was able to predict the five meat attributes, however with significant variability in the predictions. Observations from the trial suggest that the variation in predictions originated from two sources of variability: (1) heterogeneity on the sampling site; (2) muscle variation.

Heterogeneity on the sampling site: The measured NIR signal (i.e. spectrum) is affected by composition and heterogeneity of the sample. If the heterogeneity is too high compared to the signal associated with a given attribute, the sensitivity of NIR to that attribute is reduced.

Several carcasses had a very thin fat layer in the MSA grading site for fat colour, where the NIR probe was positioned. Although the probe was adjusted to capture as much signal from the fat as possible, NIR spectra will inevitable captured information from other tissues. This is also the case for measurement of ossification, especially for spectra collected in the lumbar area, as in this site the probe is placed is surrounded by fat and/or meat. The pre-processing methods applied to the NIR spectra are used to reduce the effect of this heterogeneity, but they cannot increase the sensitivity of NIR. An alternative to increase NIR accuracy is to increase its sensitivity. Each NIR spectrum corresponds to an average of 40 subscans performed with the probe held in the same place. In the pre-trial we observed that this number of scans would allow a good signal to noise ratio and still taking less than 5 seconds to acquire each spectrum. The NIR sensitivity could be enhanced by increasing the number of subscans, however, with an increase in the time required to scan each carcass. In addition, sensitivity could be increased by removing the stand-off used to avoid the meat touching the probe glass window, which could increase the signal captured from the sample. However, if this approach was proven useful, it might require more frequent cleaning of the probe window. In both approaches the calibrations developed to date could still be used, if a correction (calibration transfer function) is applied to spectra collected in the new setup (higher number of subscans and/or without the stand-off).

Muscle variation: When the MSA grader scores meat colour, the score is given to the darkest part of the exposed muscle surface. In contrast, NIR predicts the average colour of the exposed muscle surface. The same is valid for pH which was measured in only one part of the muscle. In this case, NIR is predicting correctly the colour (or pH) but there is a mismatch between MSA assessed colour (or measured pH) and NIR predictions, which can lead to variation in the predictions. This could explain part of the of carcasses predicted ± 1 from the reference value, and also low accuracy in the pH range 5.65-5.75. If the contribution of muscle variation to the variability in NIR predictions were known, it would be possible to estimate how much of this variation in predictions is due to other factors and how much NIR needs to be enhanced to improve its accuracy.

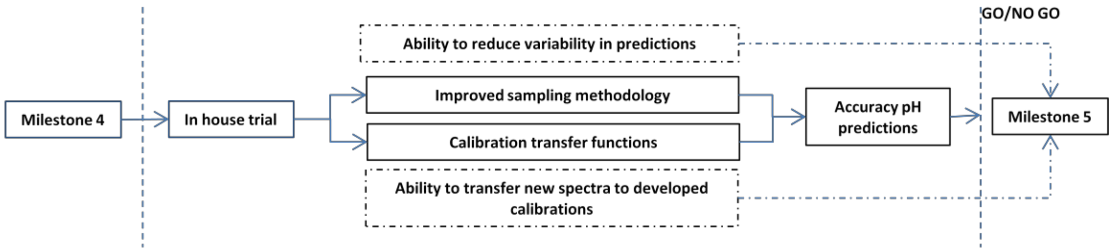
Overall progress of the project

- Calibration models have been developed for MSA meat colour (R^2 = 0.60), fat colour (R^2 = 0.61), ossification score (R^2 = 0.78), marble score (R^2 = 0.79) and ultimate pH (R^2 = 0.73) in chilled beef carcasses.
- Sampling variability prevented more accurate models to be developed, especially for colour.
- The impact of replacing grader scores with the NIR predictions to predict the final MSA grade of carcass estimated by the MSA model is being calculated. This calculation will establish the ability of the predicted NIR scores for five attributes to predict final MSA carcass grade.

Recommendations

1) The performance of NIR needs to be increased to reach the targeted accuracy to proceed towards the next milestone. Thus it is recommended an in-house trial to identify ways to improve sampling methodology, while reducing sampling variability prior the decision on whether proceed to the next milestone or otherwise. This will indicate whether NIR accuracy could reach targeted accuracy. In the in-house trial pH will be used as reference, where both factors leading to sampling variability could be addressed by:
    a. Increasing number of subscans (and/or removing the stand-off) to evaluate whether the NIR sensitivity can be improved or otherwise;
    b. Measuring pH at the same position where each NIR spectrum is collected and collecting spectra and pH across the meat surface, including edges to simulate sampling variability such as observed for fat colour. This will help to quantify the importance of muscle variation and heterogeneity in the variability of predictions and to adjust the sampling methodology;
    c. Developing calibration transfer functions that will allow the use of the developed calibrations. Thus it will be possible to verify whether the adjustments to reduce the variability in predictions were successful using pH as reference.
2) It is suggested that MSA is consulted to advice whether there are other appropriate sites to measure fat colour.

The following the diagram highlights the outcomes from in-house evaluation that will support further development of NIR.



Appendix

Table 1A Summary of the samples distribution in the calibration and validation datasets used to fit a model for meat colour. The number of samples (n), the maximum (Max), minimum (Min) and standard deviation (StdDev) values are shown for each measurement day.

		n		Max		Min	
	Days	**Cal**	**Val**	**Cal**	**Val**	**Cal**	**Val**
Cassino	**Total**	**170**	**125**	**7**	**7**	**1B**	**1B**
	1	35	19	6	6	1B	1B
	2	46	36	7	6	1C	2
	3	47	39	7	7	1B	1C
	4	42	22	7	5	1C	1C
	5		9		6		3
Teys	**Total**	**164**	**146**	**7**	**7**	**1C**	**1C**
	1	36	30	6	5	1C	1C
	2	35	31	6	6	2	2
	3	49	45	6	6	2	1C
	4	44	40	7	7	1C	2

Table 2A Summary of the samples distribution in the calibration and validation datasets used to fit a model for fat colour. The number of samples (n), the maximum (Max), minimum (Min) and standard deviation (StdDev) values are shown for each measurement day.

		n		Max		Min	
	Day	**Cal**	**Val**	**Cal**	**Val**	**Cal**	**Val**
Cassino	**Total**	**198**	**204**	**7**	**9**	**0**	**0**
	1	33	31	7	7	1	1
	2	44	43	7	9	2	2
	3	45	49	5	6	0	0
	4	41	45	7	7	1	1
	5	35	36	4	3	1	1
Teys	**Total**	**165**	**158**	**8**	**8**	**0**	**0**
	1	37	31	8	8	0	0
	2	37	35	8	7	1	2
	3	49	49	7	6	2	2
	4	42	43	6	6	3	3

Table 3A Summary of the samples distribution in the calibration and validation datasets used to fit a model for ossification collected at Cassino. The number of samples (n), the maximum (Max), minimum (Min) and standard deviation (StdDev) values are shown for each measurement day.

	N		Max		Min		StdDev	
Days	**Cal**	**Val**	**Cal**	**Val**	**Cal**	**Val**	**Cal**	**Val**
1	34	34	590	590	120	110	185	194
2	46	43	590	590	120	120	128	145
3	49	52	280	250	120	120	39	40
4	47	50	500	590	110	100	108	123
5	42	38	140	140	100	100	14	12

Table 4A Summary of the samples distribution in the calibration and validation datasets used to fit a model for marble score (MSAMB). The number of samples (n), the maximum (Max), minimum (Min) and standard deviation (StdDev) values are shown for each measurement day.

		n		Max		Min		StdDev	
		Cal	**Val**	**Cal**	**Val**	**Cal**	**Val**	**Cal**	**Val**
Cassino	**Total**	**206**	**204**	**1130**	**1150**	**110**	**110**	**215**	**215**
	Day 1	35	32	530	620	150	160	77	98
	Day 2	46	42	590	520	160	160	87	86
	Day 3	47	50	1130	1150	270	230	257	251
	Day 4	42	46	830	410	150	130	111	79
	Day 5	36	34	370	440	110	110	75	77
Teys	**Total**	**164**	**167**	**1050**	**1050**	**140**	**160**	**133**	**128**
	Day 1	36	33	980	1000	140	170	182	197
	Day 2	35	41	810	670	200	220	121	95
	Day 3	49	50	1050	500	170	160	121	73
	Day 4	44	43	760	1050	170	220	96	126

Table 5A Summary of the samples distribution in the calibration and validation datasets used to fit a model for pH. The number of samples (n), the maximum (Max), minimum (Min) and standard deviation (StdDev) values are shown for each measurement day.

		n		Max		Min		StdDev	
		Cal	Val	Cal	Val	Cal	Val	Cal	Val
	Total	**206**	**204**	**6.74**	**6.77**	**5.41**	**5.38**	**0.22**	**0.23**
Cassino	1	35	32	6.44	6.16	5.48	5.49	0.22	0.17
Cassino	2	46	42	6.05	6.14	5.41	5.41	0.15	0.16
Cassino	3	47	50	6.31	6.77	5.41	5.38	0.13	0.19
Cassino	4	42	46	6.74	6.75	5.49	5.50	0.26	0.26
Cassino	5	36	34	6.37	6.39	5.47	5.52	0.18	0.20
Teys	**Total**	**93**	**93**	**6.86**	**6.78**	**5.45**	**5.48**	**0.22**	**0.24**
Teys	3	49	50	6.22	6.57	5.45	5.53	0.16	0.18
Teys	4	44	43	6.86	6.78	5.53	5.48	0.27	0.29

Table 6A Description of calibrations models and their performance measures.

<table>
<tr><th rowspan="2"></th><th rowspan="2"></th><th rowspan="2">Meat colour</th><th rowspan="2">Fat colour</th><th colspan="5">Ossification</th><th rowspan="2">pH</th><th rowspan="2">MSAMB</th></tr>
<tr><th>100-590</th><th>100-590</th><th><160</th><th>160-200</th><th>>200</th></tr>
<tr><td rowspan="3">Reference</td><td>n</td><td>605</td><td>725</td><td>362</td><td></td><td>209</td><td>83</td><td>70</td><td>297</td><td>741</td></tr>
<tr><td>SEP</td><td>0.47</td><td>0.45</td><td>7</td><td></td><td>3</td><td>6</td><td>12</td><td>0.07</td><td>36</td></tr>
<tr><td>R^2</td><td>0.90</td><td>0.92</td><td>0.99</td><td></td><td>0.97</td><td>0.83</td><td>0.99</td><td>0.93</td><td>0.96</td></tr>
<tr><td rowspan="3">Calibration</td><td>n</td><td>334</td><td>363</td><td>217</td><td colspan="4">218</td><td>299</td><td>370</td></tr>
<tr><td>SEC</td><td>0.81</td><td>1.04</td><td>38</td><td>40</td><td>30</td><td>66</td><td>66</td><td>0.11</td><td>71</td></tr>
<tr><td>R^2</td><td>0.67</td><td>0.6</td><td>0.91</td><td>0.90</td><td>0.32</td><td>0.15</td><td>0.89</td><td>0.75</td><td>0.85</td></tr>
<tr><td rowspan="3">Validation</td><td>n</td><td>271</td><td>362</td><td>145</td><td colspan="4">217</td><td>297</td><td>371</td></tr>
<tr><td>SEV</td><td>0.86</td><td>1.04</td><td>31</td><td>49</td><td>33</td><td>57</td><td>57</td><td>0.12</td><td>84</td></tr>
<tr><td>R^2</td><td>0.60</td><td>0.61</td><td>0.78</td><td>0.87</td><td>0.34</td><td>0.08</td><td>0.81</td><td>0.73</td><td>0.79</td></tr>
<tr><td rowspan="4">Model</td><td>Type</td><td>PLS</td><td>LDA</td><td>PLS</td><td colspan="4">HM: PLS+PLSDA+LDA</td><td>PLS</td><td>PLS</td></tr>
<tr><td>Rgn (nm)</td><td>350-1350</td><td></td><td>350-2500</td><td colspan="4">PLS-S1S1: 450-1930
PLS-S1S4,PLS-DAs: 350-2500</td><td>425-1465</td><td>550-1750</td></tr>
<tr><td>Abs/Refl</td><td>Abs</td><td></td><td>Refl</td><td colspan="4">Refl</td><td>Refl</td><td>Abs</td></tr>
<tr><td>Pre-proc.</td><td>1st Der
GLSW
Auto-scaling</td><td></td><td>1st Der
GLSW
Auto-scaling</td><td colspan="4">PLS-S1S1: GLSW & Auto-scaling
PLS-DA-S1S1: 1st Der & MNCN
PLS-S1S4,PLS-S1S4: MNCN</td><td>EMSC
2nd Der
GLSW
MNCN</td><td>SNV
EMSC
1st Der
GLSW
MNCN</td></tr>
<tr><td colspan="2">Data</td><td>Average 3 scans</td><td>Colour transf.</td><td>S1S1
S1S2S1S4</td><td>S1S1+S1S4</td><td>S1S1</td><td>Average 4 scans</td><td>S1S4</td><td>Average 4 scans</td><td>Average 3 scans</td></tr>
</table>

SEP, SEC and SEV[10]: Standard error of prediction, calibration and validation, respectively. n: number of samples; Abs/Refl: Absorbance and reflectance. S1S1: first scan on the sacral site (site 1 for ossification). S1S2: first scan on the lumbar site(site 2 for ossification). S1S4: first scan on the feather bone site (site 4 for ossification). Rgn: Spectral range.

10 Standard Practices for Infrared Multivariate Quantitative Analysis (STD. ASTm E1655) 1999

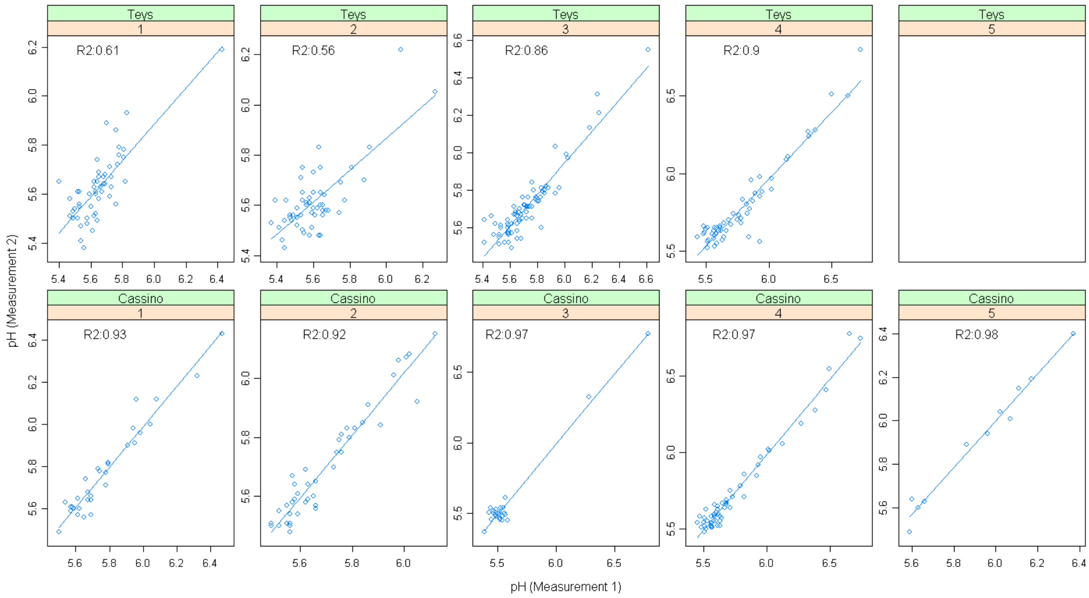


Figure 1A - Relationship between pH measured with two different pH meters for each day/plant.

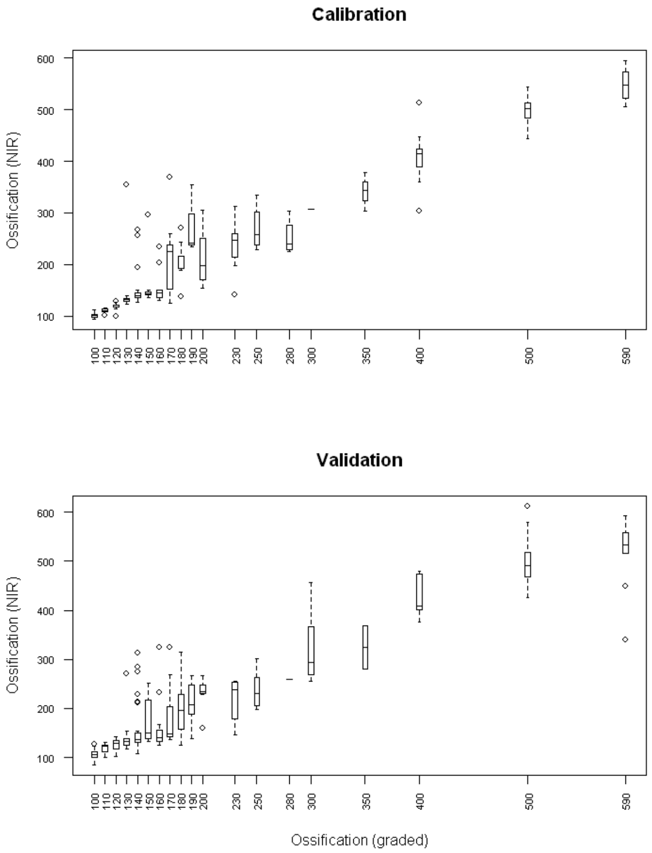


Figure 2A Predictions of the ossification score when using hierarchical model. Median is shown as the horizontal line inside of each box, top and bottom edges of the box corresponds to 3rd and 1st quartiles, the 75th and 25th percentiles, respectively (i.e. each box shows the middle of 50% of the data). Whiskers show the maximum and minimum values (excluding 'outliers'), while dots are 'outlier' samples, whose values are more than 1.5 times the difference between 3rd and 1st quartiles.

Appendix - Following up on Milestone 4

Date:	02/11/2012
To:	Phil Franks, Rajesh Margapuram
Cc:	Stuart Baud
From:	Marlon Reis, Katja Rosenvold
Project:	Accreditation of VISNIR spectroscopy to support MSA grading of beef
Project #:	MLA: A.MQT.0052

Contents

Summary and recommendations

A series of recommendations was suggested in Milestone 4 to improve the accuracy of NIR predictions before proceeding to the next milestone. In order to consider these recommendations, MLA requested that some queries were first addressed:

1) Investigate the effect of using NIR predictions on CMQ4 MSA score used to classify carcasses according to number of stars;
2) evaluate multi-factors models that combine NIR and other carcasses attributes;
3) investigate other factors affecting accuracy of NIR; and
4) review previous work to indentify an alternative site for ossification scanning.

1) A correlation of 0.88 was found between CMQ4 score obtained from NIR predictions and from scores given by a MSA grader for two cuts (silverside[1] and cube roll[2]). In terms of classification on number of stars, the use of NIR showed the following performance:

- Silverside
  - 90% correctly classified as Star4 (n = 41)
  - 97% correctly classified as Star3 (n = 188)
  - 67% correctly classified as ungraded (n = 15)
- Cube roll
  - 83% correctly classified as Star4 (n = 132)
  - 88% correctly classified as Star3 (n = 112)

Marbling was the only attribute where accuracy of NIR predictions significantly affected the rate of classification for the cube roll cut.

2) A series of multi-factors models were investigated by combining NIR with factors originating from grading (i.e. rib fat thickness, eye muscle area, hot standard carcass weight, carcass cold weight, sex, hump height, feed type) and measured attributes (ossification, pH, marbling and meat colour).

A significant improvement was obtained when NIR was combined with measured pH to predict meat colour, where:

- R^2 was increased from 0.60 to 0.66;
- Increase from 71% to 78% (compared to 82% expected for MSA grader) of carcasses correctly classified according to the threshold 4 for meat colour.

This suggests that improving the model for pH could allow the development of a combined model to predict colour with a higher accuracy.

3) During the trials there was a time gap between the time when NIR spectra were collected and the time when the carcass was graded.

- No significant correlations were found between time gap and NIR accuracies (correlations: meat colour= 0.04, pH = -0.08, marbling = 0.02).
- No significant correlation (-0.15) between time gap and oxidation status of the meat was observed.

The temperature of the loin was not correlated with the accuracy of the NIR predictions (correlations: meat colour = 0.15, pH = -0.11, marbling = -0.01).

4) In the previous project[3] the model developed for the ossification range 100-200 using NIR spectra collected at the lumbar area had a R^2 of 0.76. That study used limited number of

---

[1] STA045GRL
[2] CUB045GRL
[3] Project: No RE-221971

animals (n = 30) allowing carcasses to be carefully selected to avoid any type of soft siding the intact cap was available for the measurements). This suggests that the predictions in the ossification range 160-200 from spectra collected in the lumbar area could still be improved. This could lead to a more accurate model for prediction of ossification as the actual model combines spectra collected at the sacral, lumbar and thoracic (i.e. overall accuracy of the model depends on the quality of spectra collected in these three carcass regions).

In summary:

- The accuracy of ossification, pH and meat colour did not affect significantly the CMQ4 score but these attributes are used in plant to down grade carcasses based on cut off values, hence they should be improved;
- Accuracy on predictions of marbling affected CMQ4 score for cube roll;
- Combining NIR with pH improves predictions of meat colour;
- Blooming and loin temperature were not major contributing factors for variability in NIR;

Priorities for R&D in the next step:

1) Refine NIR methodology to improve CMQ4 classification accuracy (i.e. number of stars).
2) Improve NIR accuracy around key MSA cut off values (pH: 5.7; meat colour: 4; ossification: 300).

Recommendations:

1) To conduct an industry consultation involving MLA & AgResearch & Stuart Baud with:
    a. MSA: to present results and to discuss application of NIR within the MSA framework (e.g. plant to plant variability in CMQ4 score, classification of carcasses according to pH, meat colour, ossification and fat colour);
    b. Collaborating meat plants (Teys and Cassino): Update on project results and proposed next steps to enhance NIR supported grading;
    c. PAS/ASD: To address customisation of instrument to be used in the chiller and software to attend MSA needs.

2) To refine and validate protocols for scanning the eye muscle and lumbar region
    a. AgResearch in-house trial to identify effect of muscle variability on NIR predictions and to improve the scanning protocol (pH will be used as reference).
    b. Plant trials:
        i. Validation of protocols for scanning eye muscle for meat colour, pH and marbling. This trial should involve NIR and three experienced graders working independently to compare the effect of muscle variation in NIR and grader scores. This will confirm the improvement in the protocols or otherwise.
        ii. Adjustment of protocols to predict ossification in the range 160-200 (to improve existing model) and to predict ossification over 200 (to address plants where feather bones are removed).
3) To quantify the error variation in the current MSA grading system for pH, meat colour, ossification score and marble score. This will define the comparable error variation for NIR measured traits required for NIR supported grading to attain AUSMEAT accreditation.
    a. Minimum requirement: Captured from MSA training courses and accreditation of plant graders.
    b. Gold standard: Captured in plant trial from 2.bi.

Evaluation of NIR in the MSA model and factors contributing to accuracy of NIR predictions

Effect of replacing grader score by NIR predictions on CMQ4 MSA score

The NIR predictions of four meat attributes included in the MSA model were used to predict MSA CMQ4 scores and compared with the actual MSA CMQ4 scores (i.e. based on MSA grader scores).
The CMQ4 score combines four sensory attributes (tenderness, juiciness, flavour and overall acceptability) into a single parameter, which has a greater discriminatory ability (Watson et al., 2008). In the MSA system cut offs for CMQ4 scores are as follows:

Ungraded	≤ 45 CMQ4
Star3	46 – 63 CMQ4
Star4	64 – 76 CMQ4
Star5	≥ 77 CMQ4

Data from 244 carcasses measured at Cassino were included in the analysis. These were carcasses that had all four attributes predicted by NIR. The MSA model was run by MSA twice:

1. using the MSA grader's scores for meat colour, ossification, marbling and pH;
2. using the NIR predicted scores for meat colour ossification, marbling and pH.

All other attributes in the dataset were identical in both runs.
The analysis was carried out on two cuts: cube roll (CUB045GRL) and silverside (STA045GRL).

This approach is slightly different from the one used previously by Farrell et al. [4] In short, this work compares scores obtained directly from the MSA model, while in Farrell's approach a second model was fitted to related NIR and CMQ4 scores.

A R^2 = 0.88 was observed in both cuts for the relationship between CMQ4 score obtained with NIR predictions and a MSA grader (Figure 1). In terms of classification, when replacing MSA grader's scores with NIR predictions for the silverside, the NIR prediction results in 90% (Star4, n = 41), 97% (Star3, n = 188) and 67% (ungraded, n = 15) correct CMQ4 classification. For the cube roll the results were lower with 83% (Star4, n = 132) and 88% (Star3, n = 112) correct CMQ4 classification, respectively (Table 1).

A data analysis of the NIR variables (i.e. predicted attributes) was carried out to understand which, if any, variable(s) contributed the most to the difference between CMQ4 score resulting from grader assessment and from NIR predictions. This analysis was based in the stratification of CMQ4 score in three sets[5] as summarized in Table 2 and Figure 2. For the cube roll marbling contributed significantly to the deviation in the CMQ4 score when using NIR compared to grader assessed variables (Table 3). This suggests that accuracy in NIR predictions for marbling would have to be improved to reduce this deviation.

[4] Project: No RE-221971

[5] see the note in the last section of this report.

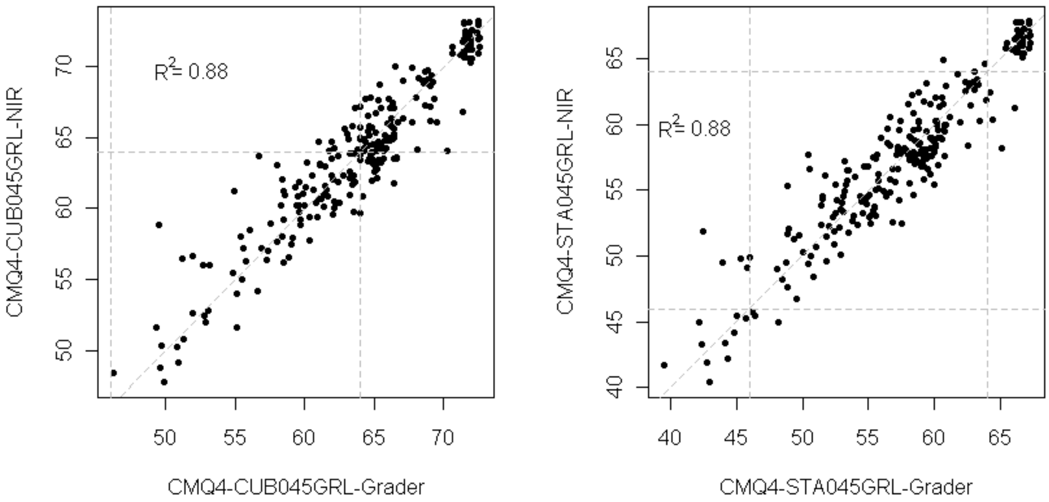


Figure 1 – Relationship between CMQ4 score obtained from the MSA model when using attributes predicted by NIR or scored by a grader. Vertical and horizontal lines refer to cut-off used to qualify carcasses according number of stars.

Table 1 – Rate of correctly or miss classified carcasses according CMQ4 score when attributes are predicted by NIR compared to score obtained from the grader. In this case the CMQ4 resulting from grader is used as reference.

		Grader					
		Cube roll			Silverside		
		Stars4	Stars3	Ungraded	Stars4	Stars3	Ungraded
NIR	Star4	83%	12%		90%	1%	0
	Star3	17%	88%		10%	97%	33%
	Ungraded				0	2	67%
	n	132	112		41	188	15

Table 2 – Percentage of carcasses below 2%, between 2-5% and higher than 5% of absolute deviation on CMQ4 score obtained using attributes from the grader compared to score obtained with attributes predicted by NIR (see also Figure 2).

Cube roll			Silverside		
<2%	2-5%	>5%	<2%	2-5%	>5%
55% (n = 134)	36% (n = 87)	9% (n = 23)	48% (n = 117)	34% (n = 82)	18% (n = 45)

Table 3 – Mean (standard deviation) and P value (for the difference between means corresponding to 2 and 2-5% ranges) for the absolute deviation on attributes scored by a grader compared to score predicted by NIR, according to the absolute deviation (below 2% or between 2-5%) on CMQ4 score obtained with attributes from the grader compared to score obtained with attributes predicted by NIR (see Table 2 and Figure 2).

	Cube roll			Silverside		
	<2%	2-5%	P	<2%	2-5%	P
Meat colour	1.13 (0.98)	1.13 (1.01)	P = 0.95	1.12 (0.96)	1.07 (0.89)	P = 0.72
Ossification	17.31 (20.74)	21.15 (21.75)	P = 0.19	18.38 (21.33)	18.54 (19.51)	P = 0.96
Marbling	44.25 (48.39)	60.92 (39.49)	***P = 0.01***	44.36 (51.52)	49.02 (30.29)	P = 0.42
pH	0.08 (0.09)	0.10 (0.10)	P = 0.44	0.09 (0.09)	0.08 (0.07)	P = 0.34

See the note in the last section of this report document.

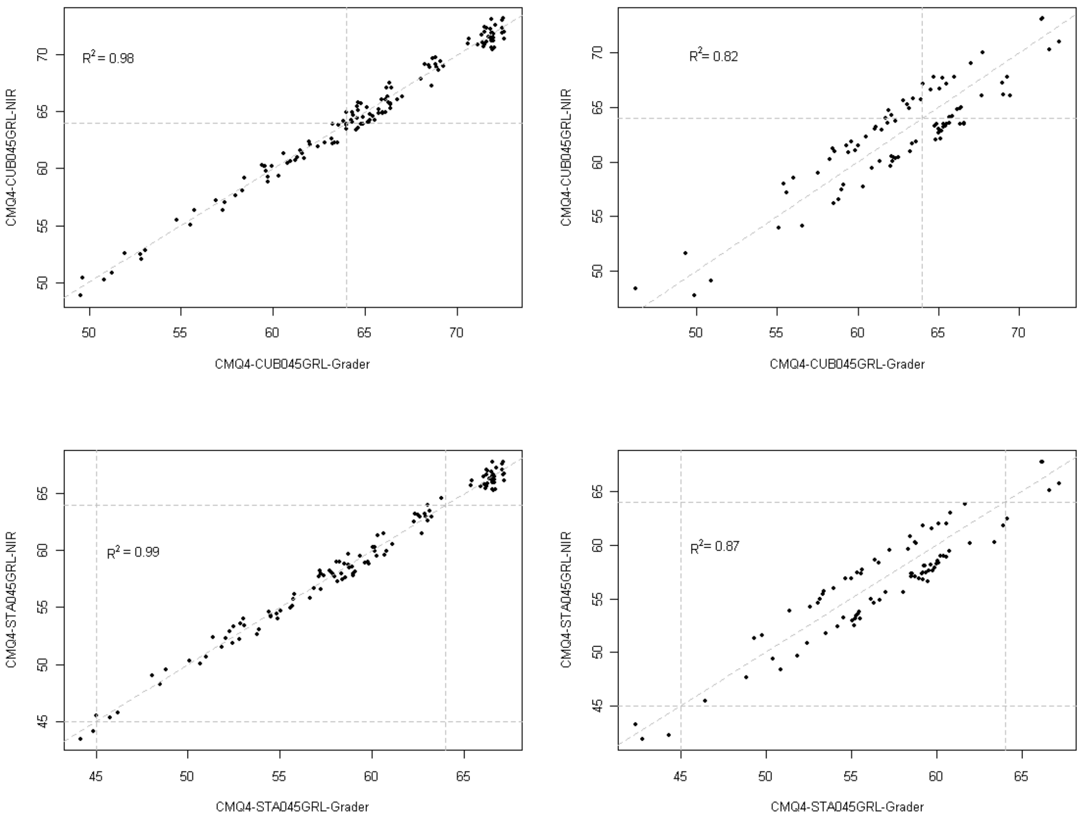


Figure 2 – Relationship between CMQ4 score obtained from the MSA model when using attributes predicted by NIR or scored by a grader. Vertical and horizontal lines refer to cut-offs used to qualify carcasses according number of stars. On the left only carcasses with a deviation between CMQ4 score from grader and NIR below 2% and on the right, carcasses with absolute deviation between 2 and 5%.

Multi-factorial prediction model that includes chemometric and non chemometric factors

The development of multi-factors approach was investigated combining factors obtained during the normal grading process with information derived from NIR. The aim was to identify if the additional information due to other factors other than NIR, would contribute to describe the variation in the meat attributes (meat colour, pH and marbling). The major limitation of this approach is that the range of variation in attributes is limited compared to the original data sets.

Data set: The data used in this analysis was provided by Cassino. During the trial a range of non-MSA carcasses were scanned to ensure that a wide range in values of the attributes was achieved. However the data provided by Cassino only included information about the MSA graded carcasses, where 116 carcasses overlapped with the NIR trial. Thus the data set available for this analysis is limited in size and range of variation among attributes.

Methodology: The evaluation was carried out for meat colour, pH and marbling. In short, the spectra of these 116 carcasses were submitted to the PLS models used to perform the original predictions of these attributes, which generate a set of scores specific for each attribute. Each set of PLS scores was then combined with the factors originating from grading (i.e. rib fat thickness, eye muscle area, hot standard carcass weight, carcass cold weight, sex, hump height, feed type) and measured attributes (ossification, pH, marbling and meat colour, obs.: each of these last three are eliminated when its own model is fitted). In addition, the time gap between the time the carcass was graded during the trial and the time NIR spectra for that specific carcass was collected was included. All these factors were combined and a generalised linear model was fitted to predict each of the 3 attributes.

Results: Overall the use of the additional factors increases the model performance (Table 4).

Table 4 – NIR predictions using multi-factors models.

	NIR	NIR + Factors	NIR + meat colour or pH
pH	0.51	0.67	0.63
Meat colour	0.55	0.69	0.65
Marbling	0.50	0.62	0.51

Meat colour - A significant improvement (R^2 increased from 0.55 to 0.65) is achieved when scores from NIR are combined with measured pH. Additional improvement (Table 4) is achieved when the other factors are added to the model: rib fat thickness, eye muscle area, hot standard carcass weight, carcass cold weight, sex, hump height, feed type, ossification, pH and marbling. It is likely that this additional improvement is due to over-fit of the model.

pH - A significant improvement (R^2 increased from 0.51 to 0.63) is achieved when scores from NIR are combined with measured meat colour. Additional improvement is reached by adding meat colour, feed type, hump height and ossification to the model (Table 4).

Marbling - The model is only improved when all the factors are combined (rib fat thickness, eye muscle area, hot standard carcass weight, carcass cold weight, sex, hump height, feed type, ossification, pH). It is likely that this improvement is due to over-fit of the model.

Meat colour+pH*:* The combined approach ($NIR+pH_{Measured}$) was further investigated using the original data set (including data from Cassino and Teys). When NIR was combined with measured pH the $R^2_{validatio}$ was increased from 0.60 to 0.66. The $NIR+pH_{Measured}$ was able to predict correctly 44% of the reference scores and 90% of the samples within an interval of ± 1 unit of the reference score (Table 5).

Discussion: In the MSA grading system, carcasses scoring 4 and above in meat colour are downgrade[6]. NIR classified 22% of carcasses scoring 4 as being less than 4 (16% were predicted as 3 and 6% were predicted to be below 3, Table 6) 6% of carcasses scoring over 4 were predicted as lower than 4, Table 6. A grader to gain/retain accreditation in meat colour is

[6] MSA® Standards Manual for Beef Grading (2012) Meat Standards Australia, ISBN 1 74036 556 3

allowed to classify 25% of carcasses as ± 1 of the reference score and 10% as ± 2 of the reference score (Table 5). The accepted error for a grader misclassifing a carcass scoring 4 as being below 4 is 18%[7] compared to 22% when using NIR+pH $_{Measured}$.

Table 5 – Rate of classification for NIR and MSA grader.

		NIR+pH$_{Measured}$	NIR+pH$_{Measured}$ (accumulative)	Current standard for MSA graders	
				Rate	Accumulative
Meat colour	Reference score	44%	44%	65%	65%
	Reference score ± 1	46%	90%	25%	90%
	Reference score ± 2	11%	10%	10%	10%

Table 6 - Rate of classification (%) on meat colour according the colour threshold of 4, when using a combined model between NIR and pH.

		Graded			
		<3	**3**	**4**	**≥5**
Predicted NIR	**<3**	77	27	6	2
	3	21	36	16	4
	4	2	34	50	13
	≥5	0	3	28	81
	n	46	64	93	68

Effect of blooming time on prediction of meat colour score

The effect of blooming was investigated comparing the time gap between the time the NIR spectra were collected and the time the carcass was graded during the trial. This was done to understand whether the blooming time had affected the NIR measurements. The time gap (mean = 29 min., standard deviation = 24, n = 605) was then compared with the deviation between NIR predicted values and graded scores for the meat attributes (i.e. meat colour, pH and marbling). No significant correlations were found (i.e. meat colour: 0.04, pH: -0.08, marbling: 0.02). This suggests that blooming was not a contributor to the limited accuracy in NIR predictions. In addition, the time gap was used as a factor in the multi-factor modelling and it did not improve the accuracy on prediction of these attributes.

The ratio between reflectance measured at 525 and 610 nm was also evaluated. According to the guidelines for meat colour evaluation from the American Meat Science Association the ratio between values of K/S[8] for 525 nm and 610 nm is related to percentage of oxymyoglobin. When compared against the time gap between NIR and grading it did not show significant correlation (-0.15).

Altogether these observations suggest that NIR and grading were performed on the bloomed meat and both sets of measurements were taken when the concentration of oxymyoglobin slowly changed to not affect the result and that the models were able to handle this type of variation.

Effect of temperature

The temperature of the loin (mean = 6.3^{o}C, standard deviation = 1.8, n = 741) was compared with the deviation between NIR predicted values and graded scores for the meat attributes (i.e. meat colour, pH and marbling), which also did not show correlation on the accuracy of predictions (correlations: meat colour = 0.15, pH = -0.11, marbling = -0.01).

Alternative scan site for ossification

The prediction of ossification showed low accuracy for the ossification range 160-200. Below 160 there is good relationship between NIR and ossification when using spectra collected at the cap in the sacral range. While for the ossification range 160-200, the NIR spectra collected at the sacral area did not produced good models. Thus the prediction of ossification between 160-200 relies on the quality of measurement collected at the lumbar area. In the previous project that attempted to predict ossification from NIR, one set of spectra was collected in the lumber area, where the predictions for a model developed for the ossification range 100-200 showed a R^2 of 0.76. In that project a limited number of animals was used (n =

[7] 17.5% = 25/2 + 10/2 assuming a symmetric distribution

[8] K/S is the ratio between absorption/scattering coefficients, respectively. It is estimated using $(K/S)=(1-R)^2/2R$, where R is the reflectance. See Guidelines for Meat Color Evaluation, American Meat Science Association, 1991.

30) and carcasses were selected to avoid any type of soft siding. Thus those results suggest that the predictions in the lumbar area (160-200) still could be improved. There are two limitations that need to be addressed:

1) The presence of fat around the cap;
2) The absence of the whole cap due to soft side.

Thus a revised methodology to measure the lumbar area should identify a way to avoid or reduce the effect of fat in the ossification site and scan more than one site in the lumbar area, e.g. operator could place the probe in the site where cap is present. And or identify an additional site in the lumbar region such as rib bone.

Acknowledgments
AgResearch is thankful for the support provided by MSA, in particular Jessira Perovic for performing the predictions of CMQ4 scores and Murray Patrick for providing information regarding MSA required accuracy for graders.

NOTE

The difference in the CMQ4 scores when using NIR predicted attributes compared to scores given by the grader is due to deviation in NIR prediction from score given by the grader, as all the other attributes are the same for both analysis. Thus to investigate the effect of this deviation, the CMQ4 scores were stratified as explained in the following. First the difference between CMQ4 obtained by NIR ($CMQ4_{NIR}$) and CMQ4 obtained from grader ($CMQ4_{Grader}$) was calculated, i.e.:

$$Diff = CMQ4_{NIR} - CMQ4_{Grader}$$

Then the absolute value of this difference ($Diff_{ABS}$) was taken. Lastly the percent deviation was calculated:

$$Diff\% = 100x(Diff_{ABS} / CMQ4_{Grader}).$$

All carcasses showing Diff% below 2% were group in on set called Set<2%. Carcasses with Diff% between 2 and 5% were grouped in the set called Set2-5%. And the remaining carcasses composed the set Set>5%.

In the following the $Diff_{ABS}$ was estimated for each of the four attributes (meat colour ossification, marbling and pH, e.g. $Diff_{ABS\text{-}MeatColour}$). Then, mean and standard deviation of $Diff_{ABS}$ was estimated for the two sets of carcasses: Set<2% and Set2-5%. These two mean values were compared using Student's t-test. The results are reported in Table 3. If the mean value for a given attribute (e.g. MSAMB) is significantly lower for the Set<2% compared to Set2-5%, it indicates that deviation of NIR prediction from score given by the grader is affecting negatively the prediction of the MSA model.