Combined analysis of near-infrared spectra, colour, and physicochemical information of brown rice to develop accurate calibration models for determining amylose content

Edenio OLIVARES DÍAZ*, Shuso KAWAMURA, Miki MATSUO, Mizuki KATO, and Shigenobu KOSEKI

Graduate School of Agricultural Science, Hokkaido University,
Kita-9 Nishi-9 Kita-Ku, Sapporo 060-8589, Japan

*Corresponding author: Tel. & Fax: +81-011-706-2558
E-mail: edenio@bpe.agr.hokudai.ac.jp, edeniood@gmail.com

Abstract

Amylose content is an important determinant of rice quality. Accurate non-destructive determination of amylose content remains a primary challenge for the rice industry. Here, we analysed the accuracy of three models for the non-destructive determination of amylose content. The models were developed by combining near-infrared spectra, colour, and physicochemical information relative to 832 brown rice samples from ten varieties produced between 2009 and 2017 in various regions of Hokkaido, Japan. Models describing low and ordinary amylose varieties were developed individually, merged, and validated using production year samples (2016–2017) different from the calibration set (2009–2015). The resulting accuracy was suitable for industrial application. With standard error of prediction = 0.70% and ratio of performance deviation = 3.56, the combination of near-infrared spectra and physicochemical information produced the most robust model, enabling more precise rice quality screening at grain elevators.

Keywords: rice quality; amylose content; near-infrared spectroscopy; calibration model accuracy; chemometric techniques

1. Introduction

Rice (*Oryza sativa* L.) is an important source of nutrition and energy for mankind; its main constituents are proteins, moisture, and starch (Kondo & Kawamura, 2013). Moisture content affects rice storage quality and milling characteristics, whereas protein content and starch contribute to the texture and eating quality of cooked rice (Siebenmorgen, Grigg & Lanning, 2013). Starch, protein content, and moisture content are therefore the most important determinants of rice quality. Starch makes up 90% of milled rice endosperm dry weight (Patindol, Siebenmorgen & Wang, 2015) and comprises two types of glucose polymers: branched amylopectin and the mostly linear amylose (Biselli et al., 2014). Amylopectin, which accounts for most of the starch fraction, has recently been linked to rice eating quality (Li, Prakash, Nicholson, Fitzgerald & Gilbert, 2016); however, (apparent) amylose content remains the most important factor defining the palatability of rice (Li, Prakash, Nicholson, Fitzgerald & Gilbert, 2016).

Granule-bound starch synthase 1 (GBSS-1), encoded by the Waxy gene, synthesizes amylose in the endosperm during the grain-filling period (Biselli et al., 2014). Ambient air temperature strongly influences amylose content. Rice varieties are classified according to amylose content as low (< 20%), medium (21–25%), and high (26–33%) (Biselli et al., 2014).

Low-amylose content varieties, which have a soft and sticky texture when cooked, are very palatable to Japanese and Northeast Asian consumers; whereas high-amylose content varieties, which cook into firm and separate grains, are more appealing to European and Latin-American consumers (Biselli et al., 2014).

Iodine-binding, also known as iodine colourimetry or amylose-iodine, is the only validated and most commonly used method for determining amylose content (Biselli et al., 2014). It is based on measuring the absorbance of the amylose-iodine complex at 620 or 720 nm using a spectrophotometer, and calculating the amylose content from a calibration curve (Fitzgerald et al., 2009). Fitzgerald et al. (2009) highlighted the main adjustments made to the amylose-iodine method

since its introduction. Because conventional methods are labour-intensive, time-consuming, chemical-dependent, and vulnerable to random error (Manley, 2014) they are unsuitable for laboratory and/or industrial uses where large volumes of samples need to be processed.

Near-infrared (NIR) spectroscopy in combination with chemometric techniques represents an alternative method for determining rice amylose content (Sampaio, Soares, Castanho, Almeida, Oliveira & Brites, 2018). Once the calibration model has been developed (Manley, 2014), it is rapid, chemical-free, easy to use, and non-destructive. The accuracy of the model, however, depends on factors such as physical conditions, temperature at scanning, cleanliness, and colour of rice samples, as well as the number of samples used to develop the model, the accuracy of reference amylose content values (Esteve Agelet & Hurburgh, 2014), and the chemometric analysis used (Sampaio, Soares, Castanho, Almeida, Oliveira & Brites, 2018).

Bagchi, Sharma and Chattopadhyay (2016) highlighted the most important developments in NIR determination of amylose content using brown and milled rice flour, a single milled kernel, and bulk brown and milled rice. However, Sampaio, Soares, Castanho, Almeida, Oliveira and Brites (2018) reported that valuable rice amylose reference data and model performance in these studies were incomplete. The authors also showed that the main challenge to determine amylose content was to select the most accurate wavelength or spectral region. Accordingly, they proposed a variable selection chemometric technique to define the most suitable spectral region for accurate amylose content determination.

In general, the accuracy of the developed models has been insufficient for screening rice amylose content in an industrial setting. Nevertheless, the influence of amylose content on rice whiteness and the degree of translucency through the grain has meant that combined analysis of NIR spectra and colour of milled rice has enabled the approximate determination of amylose content at grain elevators (Kato, Kawamura, Jo, Olivares Diaz, Yokoe & Koseki, 2016; Kawamura, Kato, Olivares Díaz, Yokoe & Koseki, 2017). The inclusion of colour in the model has improved validation statistics by decreasing the standard error of prediction (SEP), while increasing both the coefficient

of determination (r^2) and the ratio of performance deviation (RPD). Still, developing more accurate calibration models remains the desired outcome.

Based on the relationship between amylose content and physicochemical properties (Olivares Diaz, Kawamura, Jo & Koseki, 2016), Olivares Díaz, Kawamura, Jo, Kato, Matsuo & Koseki (2017) combined NIR spectra with physicochemical properties of milled rice to improve the model's validation statistics by decreasing the SEP and increasing both the r^2 and the RPD. Recently, the incorporation of rice cultivar genetics has been found to further improve the validation statistics for determining rice amylose content at grain elevators (Matsuo, Kawamura, Kato, Olivares Diaz & Koseki, 2018). Compared to a model developed using all the cultivars, the combined validation of four cultivar calibration models decreased the SEP and increased both the r^2 and RPD.

When farmers deliver the harvested rough rice at grain elevators in Japan, an automatic system comprising a NIR spectrometer and a visible light (VIS) grain segregator grades its quality based on highly precise and accurate measurement of moisture, protein content, and the percentage of whole kernels of brown rice. The NIR spectrometer determines moisture and protein content, whereas the VIS grain segregator determines the percentage of whole kernels (Kondo & Kawamura, 2013). Inclusion of amylose content in the set of properties is expected to add greater precision and accuracy when predicting rough rice quality.

At present, there are no accurate NIR calibration models for determining rice amylose content at grain elevators based on brown rice information. In this study, we assessed the accuracy of three models developed by combining NIR spectra, colour, and physicochemical information of brown rice. We also identified the most accurate and robust model by comparing their validation statistics. A suitable model could guarantee an accurate, precise, and efficient grain quality screening at Japanese grain elevators. This would support the production of high-quality rice and help address the demand for improved technology during rice postharvest processing.

2. Materials and methods

2.1 Rice samples

A total of 832 rough rice samples were collected from different rice producing regions of Hokkaido, Japan, between 2009 and 2017; they included Hidaka, Hiyama, Iburi, Ishikari, Kamikawa, Oshima, Rumoi, Shiribeshi, and Sorachi regions. The sample set comprised ten non-waxy Japonica rice varieties, namely *Daichinohoshi*, *Fukkurinko*, *Hoshimaru*, *Hoshinoyume*, *Kitakurin*, *Kirara-397*, *Nanatsuboshi*, *Sorayuki*, *Oborozuki*, and *Yumepirika*. The number of samples collected per variety and per year of production is reported in Table S1.

Within the collected samples, two groups were identified based on amylose content level. One group comprised only *Oborozuki* and *Yumepirika* varieties, which are Hokkaido low-amylose rice varieties, and was named "Low amylose varieties". The other group included the ordinary-amylose content Hokkaido rice varieties, excluding *Oborozuki* and *Yumepirika*, and was named "Ordinary amylose varieties" (Table 1).

2.2 Sample preparation

The rough rice samples were dried to approximately 15% of moisture content using a laboratory grain test dryer (Shizuoka Seiki Co., Ltd, Fukuroi, Shizuoka, Japan). Next, the dried samples were hulled using an FC2K impeller huller (Otake, Aichi, Japan) to obtain brown rice. The impeller huller was set to 3,500 rpm to obtain a high hulling rate and to avoid scratches on the rice kernel surface, which is essential for obtaining high-quality NIR spectra.

2.3. Methods of measurement and devices

2.3.1 Reference analysis of amylose content of milled rice

Reference amylose content (AC_{ref}) was determined based on iodine colourimetry using a Solid Prep III auto-analyser (Bran-Luebbe, Norderstedt, Germany) following the protocol of Williams, Wu, Tsai, and Bates (1958) with modifications by Inatsu (1988). Absorption of the amylose-iodine complex was measured at 620 nm with a spectrophotometer comprising an auto-analyser. The amylose content was quantified against a calibration curve. The *Hoshinoyume* variety (moisture content 13.1%, amylose content 21.1%) and *Hakuchoumochi* glutinous rice (amylose content 0%) grown in Hokkaido were used as standards.

Amylose content was calculated for an average of three repetitions per sample and expressed as a percentage of milled rice using an auto-analyser (%, MR, Auto-analyser). The AC_{ref} standard error of the laboratory (SEL) was estimated from the standard deviation of the three repetitions of each sample.

To develop an accurate and precise NIR calibration model (Esteve Agelet & Hurburgh, 2014), AC_{ref} values were determined using the same auto-analyser device in the same institution throughout the experimental period (2009–2017).

2.3.2 Near-infrared transmittance spectra of brown rice

NIR spectra of brown rice were obtained from a bulk of whole kernels using an Omega Analyzer G BR-5000 NIR transmittance spectrometer (Bruins Instruments, Fukuroi, Shizuoka, Japan) with an effective spectral range of 850–1048 nm, a 2-nm increment, and a path length of 30 mm. This NIR spectrometer was similar to a previously described automatic system (Kondo & Kawamura, 2013). The NIR spectrum of each sample was determined for an average of three repetitions, with each repetition representing an average of seven successive scans taken while scanning approximately 400 g of whole brown rice kernels.

To ensure accuracy, sample temperature was maintained at approximately 25±1 °C during scanning.

2.3.3 Physicochemical properties of brown rice

To assess the models' suitability for industrial application, the physicochemical properties and colour of brown rice were determined following a similar automatic procedure as described previously (Kondo & Kawamura, 2013). Accordingly, an ES-1000 VIS grain segregator (Shizuoka Seiki Co., Ltd) and an Omega Analyzer G BR-5000 NIR spectrometer were used.

2.3.3.1 Physical properties and colour information of brown rice

Physical properties, such as estimated percentages of mature kernels (MK) and immature kernels (IK), kernel length (L), and kernel width (W) were obtained from approximately 1,000 kernels of brown rice scanned by the VIS grain segregator.

Colour information such as Red, Green, and Blue (RGB) reflectance detected from the top surface of the rice kernel (R1G1B1), from the bottom surface of the kernel (R2G2B2), and transmittance detected through the kernel (R3G3B3), were also obtained from approximately 1,000 kernels of brown rice scanned by the VIS grain segregator.

The physical properties and colour information of each sample were determined for an average of five repetitions, whereas the SEL was estimated from the standard deviation of the five repetitions of each sample.

2.3.3.2 Predicted protein content of brown rice

Protein content is predicted based on a calibration model, which is inserted in the NIR spectrometer, when rice is delivered by farmers at grain elevators. Here, the model was developed using predicted protein content (PC_{pred}) rather than the protein content determined by the Kjeldahl reference method.

185 PC_{pred} was determined by chemometric analysis. Partial least squares regression (PLS) was applied
186 to analyse the reference protein content previously determined by the Kjeldahl method and the NIR
187 spectra of brown rice obtained from the NIR spectrometer.
188 The PC_{pred} standard error was estimated from the SEP.
189
190 *2.4 Data analysis*
191
192 Unscrambler Version 10.3, Upgrade 10.3.0r4 (CAMO software, Oslo, Norway) was used to process
193 the data.
194
195 *2.4.1 Preprocessing of near-infrared spectra*
196
197 A 2^{nd} order Savitzky-Golay derivative, with 2^{nd} polynomial order and two left-side and right-side
198 smoothing points was applied for preprocessing of raw NIR transmittance spectra of brown rice.
199 The 2^{nd} derivative was applied to resolve overlapping bands and to correct for baseline effects, as
200 well as to emphasize small spectral variations in the raw NIR spectra (Esteve Agelet & Hurburgh,
201 2010).
202
203 *2.4.2 Development of calibration models*
204
205 Three models were developed for the non-destructive determination of rice amylose content.
206 The first model was a conventional calibration model. AC_{ref} and NIR spectra of brown rice
207 transformed by the Savitzky-Golay 2^{nd} derivative were analysed by PLS regression using the test set
208 validation method. The resulting predicted amylose content values (AC_{pred} result) that best matched
209 the AC_{ref} result with the lowest error were used for validation statistics analysis (Fig. 1A). This
210 model was named “Only NIR”.

In the second model, the AC_{pred} result was first determined using the procedure explained in the "Only NIR" model and was then linearly combined with R1G1B1, R2G2B2, and R3G3B3 colour information (CI) of brown rice obtained from the VIS grain segregator. This procedure identified which AC_{pred} result values correlated most closely to AC_{ref} and minimized the prediction error, based on multiple linear regression (MLR) using the test set validation method. The ensuing AC_{ref} result and AC_{pred} result values were used for validation statistics analysis (Fig. 1B). This model was named "NIR+CI". As two chemometric techniques (PLS and MLR) were used, the model was classified as a dual-step calibration model. This method had proved highly accurate in previous studies when applied to milled rice data (Kato, Kawamura, Jo, Olivares Diaz, Yokoe & Koseki, 2016; Kawamura, Kato, Olivares Díaz, Yokoe & Koseki, 2017).

In the third model, the AC_{pred} result was again determined using the procedure explained in the "Only NIR" model and was then linearly combined with the PC_{pred} obtained by PLS regression and the physical properties (PP) of brown rice (MK, IK, L, and W) obtained from the VIS grain segregator. To identify which AC_{pred} result values correlated most closely to AC_{ref} and minimized the prediction error, MLR regression using the test set validation method was applied. The ensuing AC_{ref} result and AC_{pred} result values were used for validation statistics analysis (Fig. 1C). This model was named "NIR+PC+PP" and was also classified as a dual-step calibration model as both PLS and MLR were applied. The method had also proven highly accurate in previous studies when applied to milled rice data (Olivares Díaz, Kawamura, Jo, Kato, Matsuo & Koseki, 2017).

2.4.3 Selection of calibration and validation sample sets

Rice quality constituents, including amylose, vary with variety, production year, and location; they are influenced by factors such as soil type and fertility, ambient air temperature, irradiation, day length, relative humidity, and panicle characteristics (Kinoshita et al., 2017). Therefore, the models were validated by sample sets distinct from those used in the calibration model, such as the latest

(2017) and two latest (2016–2017) production year samples. This strategy sought to span the range of amylose content between the calibration and validation sets as uniformly as possible (Westad & Marini, 2015). First, the models were calibrated by samples produced between 2009 and 2016, then they were validated by samples produced in 2017. Second, the models were calibrated by samples produced between 2009 and 2015, and then validated by samples produced between 2016 and 2017 (Fig. 2).

2.4.4 Development and validation of the accuracy of each model

A model was developed by combining the "Low amylose varieties" and "Ordinary amylose varieties" results, which were developed individually (Fig. 2) as suggested by Matsuo, Kawamura, Kato, Olivares Diaz and Koseki (2018).

Briefly, all collected samples were divided into "Low amylose varieties" and "Ordinary amylose varieties" groups (explained in section 2.1). Given the large number of samples (Table 1), the calibration models of the two groups were developed individually following the selection of calibration and validation sets (explained in section 2.4.3). Next, a combined model was created by merging the validation results of the "Low amylose varieties" and the "Ordinary amylose varieties" models. The accuracy and robustness of the combined model were analysed by the coefficient of determination (r^2), systematic error (Bias), SEP, and RPD (Williams, Dardenne & Flinn, 2017) (Fig. 2).

The r^2 reports the goodness-of-fit of the model. It was determined by plotting AC_{ref} against AC_{pred} of the validation set. The r^2 value ranges from 0 to 1 and is unitless. When close to 1, it indicates a perfect linear relationship between AC_{pred} and AC_{ref} (Porep, Kammerer & Carle, 2015).

The Bias was determined from the average difference between the AC_{pred} and AC_{ref} of the respective number of samples in the validation set (n_{val}) (Eq. 1) and was expressed as a % (Porep, Kammerer & Carle, 2015).

$$Bias = \sum_{i=1}^{n_{val}} \frac{(AC_{pred} - AC_{ref})}{n_{val}} \quad \text{Eq. 1}$$

The SEP, which is defined as the standard deviation of the predicted residuals (Eq. 2), was expressed as a % (Porep, Kammerer & Carle, 2015). Lower SEP values indicate a more accurate and robust calibration model.

$$SEP = \sqrt{\frac{\sum_{i=1}^{n_{val}} (AC_{pred} - AC_{ref} - Bias)^2}{n_{val} - 1}} \quad \text{Eq. 2}$$

The RPD defines the ability of the model to predict future data in relation to the initial variability of calibration data. It is defined as the ratio between the standard deviation (SD) of AC_{ref} values and the SEP (Eq. 3). It, too, is unitless (Porep, Kammerer & Carle, 2015).

$$RPD = \frac{SD}{SEP} \quad \text{Eq. 3}$$

A higher RPD value indicates a better predictive ability and is caused by high variability in reference values and/or by a small SEP relative to the standard deviation of the predicted residuals (Prieto, Pawluczyk, Dugan & Aalhus, 2017).

3. Results and Discussion

3.1 Reference values of amylose content in milled rice

The average AC_{ref} value varied among varieties and production years. The difference was due to soil type and fertility, day length, relative humidity, panicle characteristics, and particularly ambient air temperature during the grain filling period. In Hokkaido, warm temperatures during the grain filling period increase the quality and palatability of rice by reducing amylose content (Kinoshita et al., 2017).

We increased the total number of samples by increasing both the number of varieties per production year, as well as the number of production years. This resulted in high variability among AC_{ref}

287 values. However, the low SEL values among varieties and production years (average of 0.17%)
288 indicated that AC_{ref} values were determined with a high accuracy (Table 1). These are crucial
289 factors for developing an accurate NIR calibration model (Esteve Agelet & Hurburgh, 2014;
290 Manley, 2014).
291 The *Oborozuki* and *Yumepirika* varieties, which comprised the "Low amylose varieties" group, had
292 a mean AC_{ref} of 14.2% and 16.2% respectively; whereas the varieties included in the "Ordinary
293 amylose varieties" group, had a mean AC_{ref} of approximately 20.5% (Table 1).
294 The histogram of AC_{ref} of samples used in calibration and validation sets and then validated by the
295 2017 production year (Fig. S1A) shows that AC_{ref} in the calibration set (upper bars) ranged between
296 10% and 25%. Specifically, the "Low amylose varieties" group (grey bars) ranged between 10%
297 and 21%, with the highest frequency being 15%; whereas the "Ordinary amylose varieties" group
298 (black bars) ranged between 17% and 25%, with the highest frequency being 20% (Fig. S1A). The
299 validation set (lower bars) encompassed a narrower range, between 18% and 24%. In particular, the
300 "Low amylose varieties" group (grey bars) ranged between 18% and 21%, with the highest
301 frequency being 19%; whereas the "Ordinary amylose varieties" group (black bars) ranged between
302 20% and 24% with the highest frequency being 23% (Fig. S1A).
303 The above result indicates that the low temperature during the grain filling period increased the
304 AC_{ref} values of the latest production year samples (2017). The AC_{ref} range and average values per
305 variety and production year were higher compared with the previous years (Table 1). Based on the
306 different range of AC_{ref} values between calibration and validation sets, we do not recommend these
307 AC_{ref} values be used for validating an industry-oriented model (Westad & Marini, 2015). To
308 compare the accuracy and robustness among models, we validated the analyses using 2017 and
309 2016–2017 production year samples (Fig. S1B). The histogram of AC_{ref} samples used in the
310 calibration set (upper bars) revealed that the "Low amylose varieties" group (grey bars) ranged
311 between 10% and 21%, with the highest frequency being 15%; whereas the "Ordinary amylose
312 varieties" group (black bars) ranged between 17% and 25%, with the highest frequency being 20%.

The validation set (lower bars) ranged between 14% and 24%; specifically, the "Low amylose varieties" group (grey bars) ranged between 14% and 21%, with the highest frequency being 16%, and the "Ordinary amylose varieties" group (black colour bars) ranged between 19% and 24%, with the highest frequency being 20% (Fig. S1B). The proximity in ranges of AC_{ref} values between calibration and validation sets suggests that the latter was suitable for validating a model aimed for industrial application (Westad & Marini, 2015).

3.2 Near-infrared transmittance spectra of brown rice

Two absorption bands, corresponding to those identified in the raw NIR spectra (Fig. S2), were detected also in the transformed NIR spectra of brown rice obtained from the Savitzky-Golay 2^{nd} derivative (Fig. 3). This indicated that functional groups in the sample molecules absorbed light when irradiated in the NIR wavelength range. We assigned the strongest absorption band, at approximately 982 nm, to the second overtone stretching (O-H bond) of the bound -OH alcohol group. The other absorption band, at approximately 916 nm, was assigned to the third overtone of symmetric stretching (C-H bond) of the methylene combination $-CH_2$ group (Ozaki, 2015) (Fig. 3). Amylose, moisture, and proteins, which are the major constituents of rice, cause O-H and C-H bond vibrations (Sampaio, Soares, Castanho, Almeida, Oliveira & Brites, 2018). However, because absorption bands are complex and comprise broad and numerous overlapping bands, the identification and correlation of an absorption band with a specific constituent is difficult and inaccurate (Porep, Kammerer & Carle, 2015). To overcome this issue and identify the information in the NIR spectra that best correlated to AC_{ref}, we applied PLS regression and maximized the covariance between NIR spectra and AC_{ref} values.

3.3 Physicochemical properties and colour of brown rice

Physicochemical properties (Table S2) and colour information (Table S3) pertaining to brown rice obtained from the VIS grain segregator varied somewhat by variety and production year. This result was due to genetic similarities among Japonica varieties bred for producing rice in Hokkaido (Kinoshita et al., 2017).

We used MLR regression to develop the "NIR+CI" and "NIR+PC+PP" models. The number of variables in the "NIR+CI" model (R1, G1, B1, R2, G2, B2, R3, G3, and B3) and in the "NIR+PC+PP" model (PC_{pred}, MK, IK, L, and W) was smaller than the total number of samples (832). Also, each variable used for developing the models was a linearly independent variable. These are important requirements for obtaining a stable MLR output (CAMO Software As, 2014a).

3.4 Models for determining rice amylose content at grain elevators

3.4.1 Model validated using the 2017 production year samples

To compare the accuracy and robustness among models (see section 3.1), we calibrated the three models using the 2009–2016 production years (738 samples) and validated them by the 2017 production year (94 samples).

While developing the "Only NIR" model, we identified 9 as the optimal PLS factor for both the "Low amylose varieties" and "Ordinary amylose varieties" models based on the high explained variance for AC_{ref}. Also, we removed two outliers from both models based on the large leverage and high residual variance for AC_{ref} reported by the PLS regression results (CAMO Software As, 2014b).

The regression coefficients for PLS = 9 indicated that the wavelength at 916 nm had the highest spectral variation and thus better correlated with AC_{ref} for both the "Low amylose varieties" (Fig. S3A) and the "Ordinary amylose varieties" model (Fig. S3B). This large regression coefficient corresponded to the absorption band of the C-H bond in the 2nd derivative plot (Fig. 3), which is

365 characteristic of to the starch group (Font, Vélez, Del Río-Celestino, De Haro-Bailón, & Montoro,
366 2005; Sampaio, Soares, Castanho, Almeida, Oliveira & Brites, 2018).
367 While developing both the "NIR+PC+PP" and "NIR+CI" models, we identified and removed two
368 outliers from both the "Low amylose varieties" and the "Ordinary amylose varieties" models based
369 on the high residual variance for AC_{ref} reported by the MLR result (CAMO Software As, 2014a).
370 The optimal PLS factor in the "Only NIR" model and the number of outliers removed in each
371 model were within the recommended range for considering the model as highly accurate (Williams,
372 Dardenne & Flinn, 2017).
373 We plotted the AC_{ref} values against AC_{pred} values obtained from the three developed models. We
374 also defined the "Low amylose varieties" and "Ordinary amylose varieties" groups comprising the
375 regression between AC_{ref} values and AC_{pred} values for the "Only NIR" (Fig. 4A), the "NIR+CI"
376 (Fig. 4B), and the "NIR+PC+PP" (Fig. 4C) models.
377 The SEP values of the "Only NIR" model (0.70%) (Fig. 4A), the "NIR+CI" model (0.65%) (Fig.
378 4B), and the "NIR+PC+PP" model (0.59%) (Fig. 4C) were below 1%, which is deemed suitable for
379 industrial application (Kato, Kawamura, Jo, Olivares Diaz, Yokoe & Koseki, 2016). In addition, the
380 RPD values of the "Only NIR" model (2.51) (Fig. 4A) , the "NIR+CI" model (2.67) (Fig. 4B), and
381 the "NIR+PC+PP" model (2.95) (Fig. 4C) were within 2.5–2.9, a range suitable for industrial
382 screening (Prieto, Pawluczyk, Dugan & Aalhus, 2017). These results show that the three developed
383 models were highly accurate for the non-destructive determination of rice amylose content.
384 However, because of differences in the range of AC_{ref} values between samples in the calibration and
385 validation sets (see section 3.1), we focused our analysis on a comparison among models.
386 The r^2, SEP, and RPD values of the "NIR+CI" model (Fig. 4B) were better than those of the "Only
387 NIR" model (Fig. 4A). Kato, Kawamura, Jo, Olivares Diaz, Yokoe and Koseki (2016) and
388 Kawamura, Kato, Olivares Diaz, Yokoe and Koseki (2017) reported a similar result on milled rice.
389 Including the information about colour of brown rice in the model barely increased the r^2 (0.02),
390 barely decreased the SEP (0.05%), and increased the RPD (0.16) in comparison with the "Only NIR"

391 model. The reduction in SEP derived from a lower standard deviation of AC_{pred} values and,
392 consequently, increased the RPD.
393 The "NIR+PC+PP" model (Fig. 4C) was the most accurate, reflecting the result by Olivares Díaz,
394 Kawamura, Jo, Kato, Matsuo and Koseki (2017) on milled rice. Inclusion of the physicochemical
395 properties of brown rice in the model barely increased the r^2 (0.04 compared to "Only NIR" and
396 0.02 compared to "NIR+CI"), slightly decreased the SEP (0.11% compared to "Only NIR" and
397 0.06% compared to "NIR+CI"), but considerably increased the RPD (0.44 compared to "Only NIR"
398 and 0.28 compared to "NIR+CI") (Fig. 4A, B, C). The reduction in SEP was due to a lower
399 standard deviation of AC_{pred} values and led to an increased RPD.
400 Given that the "NIR+PC+PP" model provided the best result, we defined the model equations
401 obtained for the "Low amylose varieties" (AC_{Low}) (Eq. 4) and "Ordinary amylose varieties"
402 ($AC_{Ordinary}$) models (Eq. 5) as follows:
403 $$AC_{Low} = 3.7362 + 0.9981 \cdot AC_{pred} + 0.0729 \cdot PC_{pred} + 0.0004 \cdot MK - 0.0160 \cdot IK - 0.8749 \cdot$$
404 $$L + 0.2275 \cdot W \quad \text{Eq. 4}$$
405 $$AC_{Ordinary} = -6.1558 + 0.9631 \cdot AC_{pred} - 0.0069 \cdot PC_{pred} - 0.0061 \cdot MK - 0.0185 \cdot IK +$$
406 $$0.5944 \cdot L + 1.5979 \cdot W \quad \text{Eq. 5}$$
407 Considering that MLR regression in Unscrambler is based on analysis of variance (ANOVA)
408 (CAMO Software As, 2014a), we defined the "Low amylose varieties" and "Ordinary amylose
409 varieties" models, which comprised the "NIR+PC+PP" model, as highly significant based on p-
410 values from ANOVA ($p < 0.001$).
411 Accordingly, the dual-step "NIR+PC+PP" model appears to be the most accurate and robust among
412 the developed models.
413
414 *3.4.2 Model validated using the 2016–2017 production year samples*
415

In this case, we calibrated the three models using the 2009–2015 production year (637 samples) and validated them with the 2016–2017 production year (195 samples), with the scope of a) determining whether their accuracy was suitable for industrial application and b) comparing accuracy and robustness among models.

We analysed the PLS factor, outliers, regression coefficient, etc. following the same procedure as in section 3.4.1.

While developing the "Only NIR" model, we identified 10 as the optimal PLS factor and removed two outliers from the "Low amylose varieties" model and four from the "Ordinary amylose varieties" model. The regression coefficients for PLS = 10 indicated that the highest spectral variation was at 916 nm, which correlated best with AC_{ref} for both the "Low amylose varieties" (Fig. S3C) and the "Ordinary amylose varieties" model (Fig. S3D). As explained in section 3.4.1, this peak defined absorption by the C-H bond in the 2nd derivative plot (Fig. 3) and likely corresponded to the starch group.

While developing both the "NIR+PC+PP" and "NIR+CI" models, we identified and removed two outliers from the "Low amylose varieties" model and four from the "Ordinary amylose varieties" model.

The optimal PLS factor in the "Only NIR" model and the number of outliers removed in each model were within the recommended range for considering the model as highly accurate.

We plotted the AC_{ref} against AC_{pred} values resulting from the three developed models. We also defined the "Low amylose varieties" and "Ordinary amylose varieties" groups comprising the regression between AC_{ref} values and AC_{pred} values for the "Only NIR" (Fig. 4D), the "NIR+CI" (Fig. 4E), and the "NIR+PC+PP" (Fig. 4F) models.

The AC_{ref} values range was higher upon validation with the 2016–2017 production year samples (14%–24%) (Fig. S1B) than with the 2017 production year samples (18%–24%) (Fig. S1A). We increased the variability in AC_{ref} values and thus the standard deviation of the validation set by including the 2016 production year samples. This led to an increase in SEP and RPD values of the

models (Fig. 4D, E, F) compared to those obtained upon validation with the 2017 production year samples (Fig. 4A, B, C).

The SEP values of the "Only NIR" model (0.73%) (Fig. 4D), the "NIR+CI" model (0.73%) (Fig. 4E), and the "NIR+PC+PP" model (0.70%) (Fig. 4F) were still below 1%. RPD values of the "Only NIR" model (3.41) (Fig. 4D) and the "NIR+CI" model (3.39) (Fig. 4E) were within 3.0–3.4, a range deemed suitable for industrial quality control; whereas the RPD value for the "NIR+PC+PP" model (3.56) (Fig. 4F) was within 3.5–4.0, a range suitable for industrial process control (Prieto, Pawluczyk, Dugan & Aalhus, 2017). These results indicate that the three models were highly accurate for the non-destructive determination of rice amylose content at grain elevators.

The "Only NIR" (Fig. 4D) and "NIR+CI" analyses (Fig. 4E) yielded similar results, which differed from those obtained when the 2017 production year samples were used to validate the model (Fig. 4A, B). Colour information, therefore, did not influence the accuracy of the models when validated with the 2016–2017 production year samples. An increase in the number of samples within the validation set reduced the influence of colour information by augmenting the variability of AC_{ref} values.

The "NIR+PC+PP" model was characterized by $r^2 = 0.93$, SEP = 0.70%, and RPD = 3.56 (Fig. 4F). As with validation using the 2017 production year samples (Fig. 4C), "NIR+PC+PP" was the most accurate model. Inclusion of physicochemical properties in the model increased the RPD value as a result of decreasing the SEP and reducing the standard deviation of AC_{pred} values.

We defined the model equations for the "Low amylose varieties" group (AC_{Low}) (Eq. 6) and the "Ordinary amylose varieties" group ($AC_{Ordinary}$) (Eq. 7). We also defined the "Low amylose varieties" and "Ordinary amylose varieties" models, which comprised the "NIR+PC+PP" model, as highly significant considering the p-values from ANOVA ($p < 0.001$).

$$AC_{Low} = 1.6359 + 0.9945 \cdot AC_{pred} - 0.0352 \cdot PC_{pred} - 0.0100 \cdot MK - 0.0102 \cdot IK + 0.1591 \cdot L - 0.3930 \cdot W \quad \text{Eq. 6}$$

467 $AC_{Ordinary} = -4.0877 + 0.9770 \cdot AC_{pred} - 0.0117 \cdot PC_{pred} - 0.0013 \cdot MK + 0.0020 \cdot IK +$

468 $0.1148 \cdot L + 1.3561 \cdot W$ Eq. 7

469 Based on these results, the dual-step "NIR+PC+PP" model displayed the highest accuracy and

470 robustness among the developed models.

471

472 *3.5 Comparison among models based on year of production of the calibration set*

473

474 *3.5.1 Comparison among models validated using the 2017 production year samples*

475

476 To further compare accuracy and robustness among models, we calibrated the models using the

477 2009–2013, 2009–2014, 2009–2015, and 2009–2016 production year samples and validated each

478 calibration set with the 2017 production year samples.

479 While developing the models within each production year, we identified 9 as the optimal PLS factor.

480 We also identified and removed two outliers from both the "Low amylose varieties" and the

481 "Ordinary amylose varieties" models from each of the three models.

482 For all three models, accuracy increased as the production year increased. This augmented also the

483 variability of the AC_{ref} value by increasing the number of samples in the calibration set (Esteve

484 Agelet & Hurburgh, 2014) (Table 2A).

485 The "NIR+CI" model exhibited better accuracy within each production year and among them than

486 the "Only NIR" model (Table 2A). Inclusion of colour information in the model slightly increased

487 the r^2, slightly decreased the SEP, and increased the RPD compared to the "Only NIR" model. As

488 explained in section 3.4.1, this occurred as a result of reducing the standard deviation of AC_{pred}

489 values.

490 The "NIR+PC+PP" model displayed the best accuracy within each production year and among

491 them (Table 2A). Inclusion of physicochemical properties in the model decreased the SEP while

492 increasing the r^2 and RPD in comparison with both "Only NIR" and "NIR+CI" models. The effect

493 was highest on RPD values as a result of decreased SEP and lower standard deviation of AC_{pred}
494 values (Table 2A).
495 These results suggest that the dual-step "NIR+PC+PP" model presents the highest accuracy and
496 robustness among the developed models within and between production years.
497
498 *3.5.2 Comparison among models validated using the 2016–2017 production year samples*
499
500 To compare the accuracy and robustness among models, we calibrated the models using the 2009–
501 2013, 2009–2014, and 2009–2015 production year samples and validated each calibration set by the
502 2016–2017 production year samples.
503 While developing the models within each production year, we identified 10 as the optimal PLS
504 factor. We also identified and removed two outliers from the "Low amylose varieties" model and
505 four from the "Ordinary amylose varieties" model from each of the three models.
506 The "Only NIR" and "NIR+CI" models shared similar accuracy within each production year and
507 among them (Table 2B). The SEP value of the "Only NIR" and "NIR+CI" models barely decreased,
508 whereas the r^2 and RPD values of both models barely increased as we increased the production year
509 (Table 2B). This result contrasted with the findings obtained when the 2017 production year
510 samples were used to validate the model (Table 2A). The increase in the number of samples of both
511 calibration and validation sets reduced the influence of colour information by increasing the
512 variability of AC_{ref} and AC_{pred}.
513 The "NIR+PC+PP" model yielded the highest accuracy within each production year and among
514 them (Table 2B). As with validation using the 2017 production year samples (Table 2A), inclusion
515 of physicochemical properties had the highest effect on the RPD value (Table 2B). The increment in
516 production year increased variability among AC_{ref} values by increasing the number of samples in
517 the calibration set. Additionally, inclusion of physicochemical properties decreased the SEP by

518 reducing the standard deviation of AC_{pred} values. As a result, the RPD value increased within and
519 among production years.
520 These results suggest that the dual-step "NIR+PC+PP" model was the most accurate and robust
521 among the developed models within each production year and among production years.
522
523 **4. Conclusions**
524
525 The results of this study indicated that the developed models: "Only NIR", "NIR+CI", and
526 "NIR+PC+PP" calibrated by the 2009–2015 and validated by the 2016–2017 production year
527 samples were highly accurate for the non-destructive determination of rice amylose content at grain
528 elevators. Throughout the study period, we increased the variability of the AC_{ref} by yearly increases
529 in the number of samples in the calibration set. Also, owing to the large number of samples, each
530 model was developed by combining the "Low amylose varieties" and "Ordinary amylose varieties"
531 results, yielding models with accuracy suitable for industrial application.
532 The regression coefficients of each wavelength related to the optimal PLS factor indicated that the
533 highest spectral variation and thus highest correlation with AC_{ref} was attained at 916 nm,
534 corresponding to the third overtone of symmetric stretching (C-H bond) of the methylene
535 combination -CH_2 group.
536 The inclusion of physicochemical properties was a relevant factor in improving the accuracy of the
537 model. The "NIR+PC+PP" model revealed the highest accuracy and robustness. Physicochemical
538 properties had the greatest effect on RPD among the various validation statistics, decreasing the
539 SEP by reducing the standard deviation of AC_{pred}. As a result, the RPD value increased within each
540 production year and among production years.
541 Given that a NIR spectrometer and a VIS grain segregator comprise the automatic system for
542 inspecting rough rice quality at grain elevators, it is easy to derive the PC_{pred} and the PP of brown
543 rice to include in the "NIR+PC+PP" model. This allows the addition of amylose content to the set

of properties analysed to predict rough rice quality based on brown rice information obtained when farmers deliver the harvested product. The resulting information could improve the accuracy, precision, and efficiency of grain quality screening at Japanese grain elevators, in turn stimulating the production of high-quality rice and addressing the demand for improved technology during rice postharvest processing.

Acknowledgements

This research was supported by grant from the project of the National Agriculture and Food Research Organization (NARO) in Japan, Bio-oriented Technology Research Advancement Institution (BRAIN) titled the special scheme project on vitalizing management entities of agriculture, forestry and fisheries.

Declarations of interest: none.

References

Bagchi, T. B., Sharma, S., & Chattopadhyay, K. (2016). Development of NIRS models to predict protein and amylose content of brown rice and proximate compositions of rice bran. *Food Chemistry*, *191*, 21–27. https://doi.org/10.1016/j.foodchem.2015.05.038.

Biselli, C., Cavalluzzo, D., Perrini, R., Gianinetti, A., Bagnaresi, P., Urso, S., Orasen, G., Desiderio, F., Lupotto, E., Cattivelli, L., & Valè, G. (2014). Improvement of marker-based predictability of Apparent Amylose Content in japonica rice through GBSSI allele mining. *Rice*, *7*(1), 1. https://doi.org/10.1186/1939-8433-7-1.

CAMO Software As. (2014a). Multiple Linear Regression. In CAMO Software AS (Ed.), *The Unscrambler X v10.3 User Manual* (pp. 583–616). Oslo, Norway: CAMO Software AS.

CAMO Software As. (2014b). Partial Least Squares. In CAMO Software AS (Ed.), *The Unscrambler X v10.3 User Manual* (pp. 675–742). Oslo, Norway: CAMO Software AS.

Esteve Agelet, L., & Hurburgh, C. R. (2010). A Tutorial on Near Infrared Spectroscopy and Its Calibration. *Critical Reviews in Analytical Chemistry*, *40*(4), 246–260. https://doi.org/10.1080/10408347.2010.515468.

Esteve Agelet, L., & Hurburgh, C. R. (2014). Limitations and current applications of Near Infrared Spectroscopy for single seed analysis. *Talanta*, *121*(April 2014), 288–299. https://doi.org/10.1016/j.talanta.2013.12.038.

Fitzgerald, M., Bergman, C., Resurreccion, A., Möller, J., Jimenez, R., Reinke, R. F., Martin, M., Blanco, P., Molina, F., Chen, M., Kuri, V., Romero, M., Habibi, F., Umemoto, T., Jongdee, S., Graterol, E., Reddy, K. R., Bassinello, P. Z., Sivakami, R., Rani, N., Das, S., Wang, Y. J., Indrasari, S. D., Ramli, A., Ahmad, R., Dipti, S. S., Xie, L., Lang, N. T., Singh, P., Castillo Toro, D., Tavasoli, F., & Mestres, C. (2009). Addressing the Dilemmas of Measuring Amylose in Rice. *Cereal Chemistry*, *86*(5), 492–498.

https://doi.org/10.1094/CCHEM-86-5-0492.

Font, R., Vélez, D., Del Río-Celestino, M., De Haro-Bailón, A., & Montoro, R. (2005). Screening inorganic arsenic in rice by visible and near-infrared spectroscopy. *Microchimica Acta*, *151*(3–4), 231–239. https://doi.org/10.1007/s00604-005-0404-x.

Inatsu, O. (1988). *Studies on Improving the Eating Quality of Hokkaido Rice. Report No. 66*. Sapporo, Japan. ISSN: 0367-6048.

Kato, M., Kawamura, S., Jo, A., Olivares Díaz, E., Yokoe, M., & Koseki, S. (2016, November). Determination of Rice Amylose Content by Combined Use of a Near-infrared Spectrometer and a Visible Light Segregator. Oral presentation at the meeting of the 5th Asian Near-Infrared Symposium and the 32nd Japanese NIR Forum, Kagoshima, Japan.

Kawamura, S., Kato, M., Olivares Díaz, E., Yokoe, M., & Koseki, S. (2017, June). Non-destructive determination of rice amylose content: Improvement of eating quality of Hokkaido grown rice by sorting amylose and protein contents. Oral presentation at the meeting of the International Food Machinery & Technology Exhibition FOOMA JAPAN 2017 Academic Plaza, Tokyo, Japan.

Kinoshita, N., Kato, M., Koyasaki, K., Kawashima, T., Nishimura, T., Hirayama, Y., Takamure, I., Sato, T., & Kato, K. (2017). Identification of quantitative trait loci for rice grain quality and yield-related traits in two closely related Oryza sativa L. subsp. japonica cultivars grown near the northernmost limit for rice paddy cultivation. *Breeding Science*, *67*(3), 191–206. https://doi.org/10.1270/jsbbs.16155.

Kondo, N., & Kawamura, S. (2013). Postharvest automation: fundamentals and practices. In Q. Zhang & F. J. Pierce (Eds.), *Agricultural Automation* (pp. 367–383). Boca Raton, Florida, USA: CRC Press Taylor & Francis Group.

Li, H., Prakash, S., Nicholson, T. M., Fitzgerald, M. A., & Gilbert, R. G. (2016). The

importance of amylose and amylopectin fine structure for textural properties of cooked rice grains. *Food Chemistry*, *196*, 702–711. https://doi.org/10.1016/j.foodchem.2015.09.112.

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chem. Soc. Rev.*, *43*(24), 8200–8214. https://doi.org/10.1039/C4CS00062E.

Matsuo, M., Kawamura, S., Kato, M., Olivares Diaz, E., & Koseki, S. (2018, June). Effect of Rice Cultivars on Accuracy for Determining Amylose Content using Near-infrared Spectroscopy. Poster session presentation at the meeting of the 6th Asian NIR Symposium (ANS2018) and The 7th Chinese National NIR Conference, Kunming, China.

Olivares Díaz, E., Kawamura, S., Jo, A., & Koseki, S. (2016, May). Relationship Between Rice Kernel Maturity and Physicochemical Properties. Oral presentation at the meeting of the 8th International Symposium on Machinery and Mechatronics for Agricultural and Biosystems Engineering (ISMAB 2016), Niigata, Japan.

Olivares Diaz, E., Kawamura, S., Jo, A., Kato, M., Matsuo, M., & Koseki, S. (2017, November). Combined Analysis of Near-Infrared Spectra and Physicochemical Properties for Determination of Amylose Content of Milled Rice. Oral presentation at the meeting of the 33rd NIR Forum held under Japan Council for Near Infrared Spectroscopy (JCNIRS), Tsukuba, Ibaraki, Japan.

Ozaki, Y. (2015). Appendix B Table of group frequency. In S. Yabuki (Ed.), *Near-Infrared Spectroscopy (in Japanese)* (pp. 269–272). Tokyo, Japan: The Spectroscopical Society of Japan.

Patindol, J. A., Siebenmorgen, T. J., & Wang, Y. J. (2015). Impact of environmental factors on rice starch structure: A review. *Starch/Staerke*, *67*(1–2), 42–54.

https://doi.org/10.1002/star.201400174.

Porep, J. U., Kammerer, D. R., & Carle, R. (2015). On-line application of near infrared (NIR) spectroscopy in food production. *Trends in Food Science & Technology*, *46*(2), 211–230. https://doi.org/10.1016/j.tifs.2015.10.002.

Prieto, N., Pawluczyk, O., Dugan, M. E. R., & Aalhus, J. L. (2017). A Review of the Principles and Applications of Near-Infrared Spectroscopy to Characterize Meat, Fat, and Meat Products. *Applied Spectroscopy*, *71*(7), 1403–1426. https://doi.org/10.1177/0003702817709299.

Sampaio, P. S., Soares, A., Castanho, A., Almeida, A. S., Oliveira, J., & Brites, C. (2018). Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms. *Food Chemistry*, *242*, 196–204. https://doi.org/10.1016/j.foodchem.2017.09.058.

Siebenmorgen, T. J., Grigg, B. C., & Lanning, S. B. (2013). Impacts of Preharvest Factors During Kernel Development on Rice Quality and Functionality. *Annual Review of Food Science and Technology*, *4*, 101–115. https://doi.org/10.1146/annurev-food-030212-182644.

Westad, F., & Marini, F. (2015). Validation of chemometric models - A tutorial. *Analytica Chimica Acta*, *893*, 14–24. https://doi.org/10.1016/j.aca.2015.06.056.

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*(2), 85–90. https://doi.org/10.1177/0967033517702395.

Williams, V. R., Wu, W.-T., Tsai, H. Y., & Bates, H. G. (1958). Rice Starch, Varietal Differences in Amylose Content of Rice Starch. *Journal of Agricultural and Food Chemistry*, *6*(1), 47–48. https://doi.org/10.1021/jf60083a009.

Figure Captions

Fig. 1 Procedure for developing the "Only NIR" model (A), "NIR+CI" model (B), and "NIR+PC+PP" model (C). AC_{ref}, reference amylose content; AC_{pred}, predicted amylose content; CI, colour information; PC_{pred}, predicted protein content; PP, physical properties; PLS, partial least squares; MLR, multiple linear regression

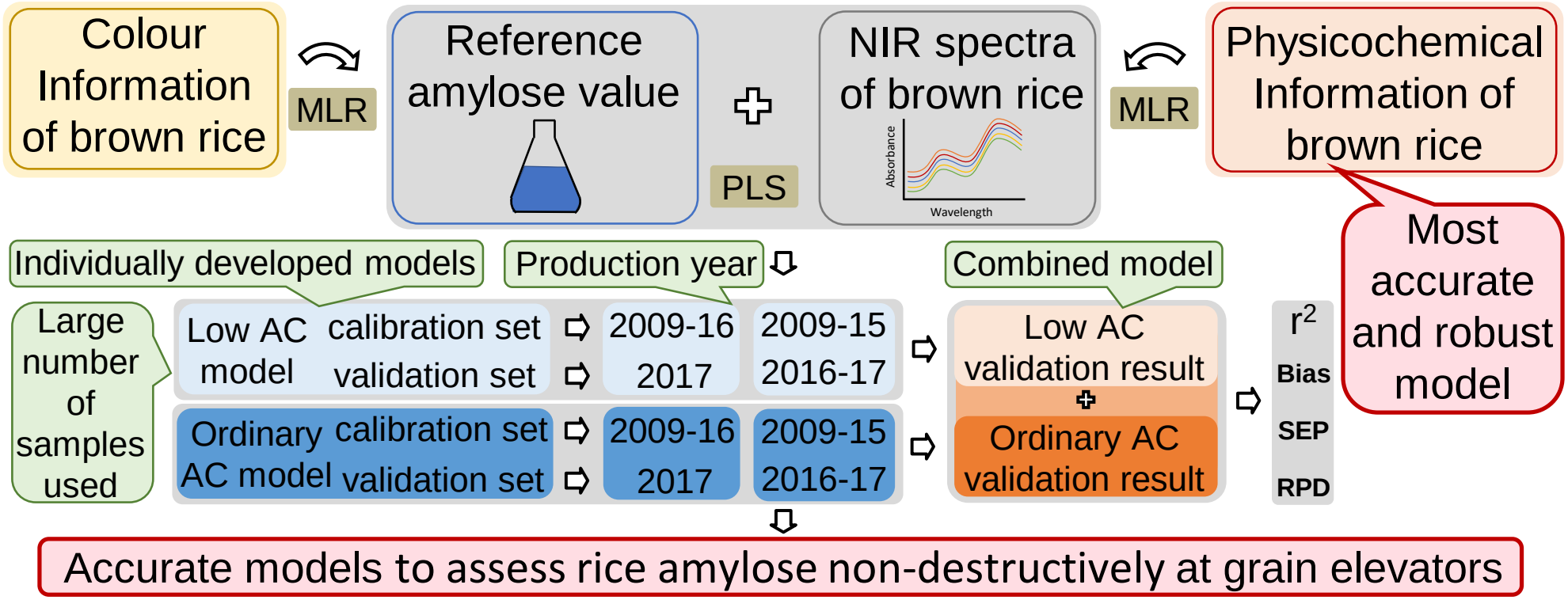
Fig. 2 Procedure for developing and validating the accuracy of each model. r^2, coefficient of determination; Bias, systematic error; SEP, standard error of prediction; RPD, ratio of performance deviation

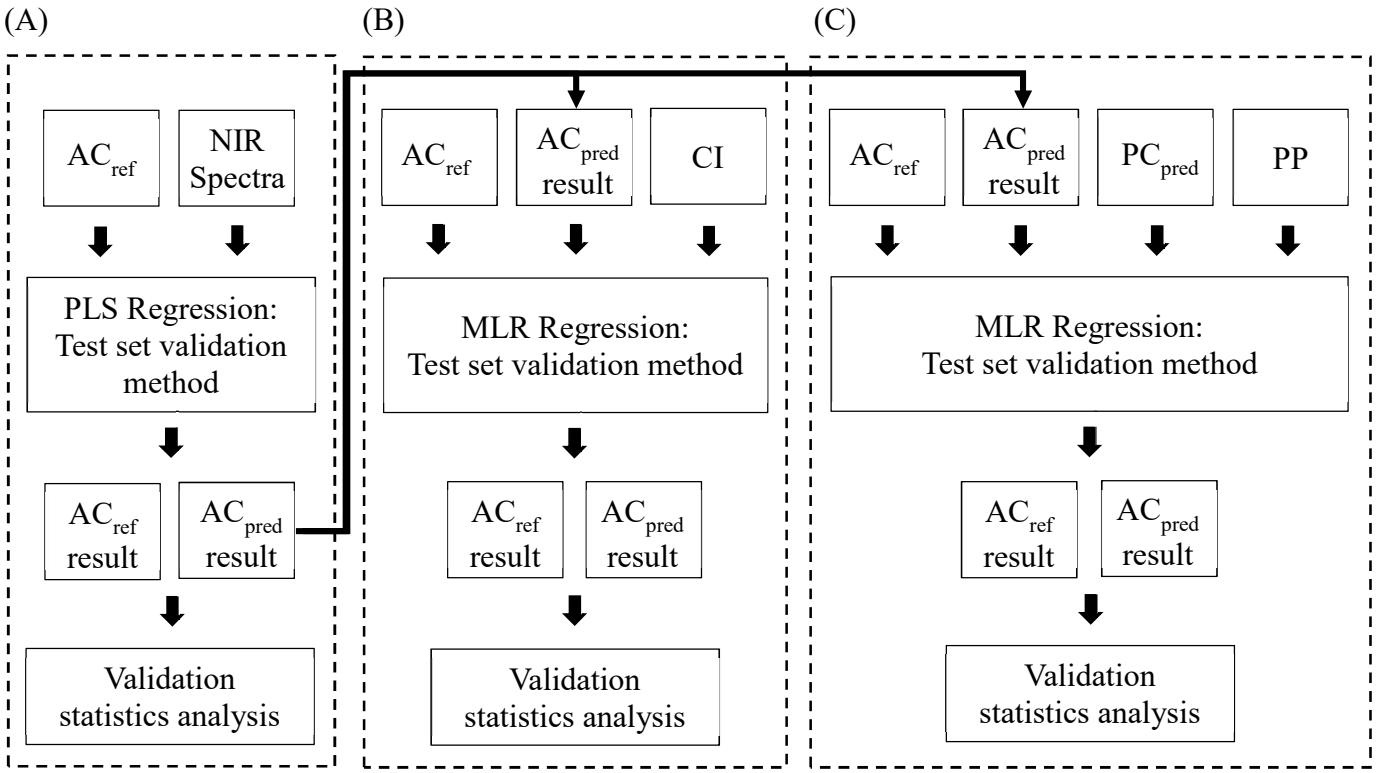
Fig. 3 NIR spectra transformed by the Savitzky-Golay 2nd derivative. 982 nm: strongest absorption band assigned to the second overtone stretching (O-H bond) of the bound -OH alcohol group. 916 nm: second absorption band assigned to the third overtone of symmetric stretching (C-H bond) of the methylene combination -CH_2 group

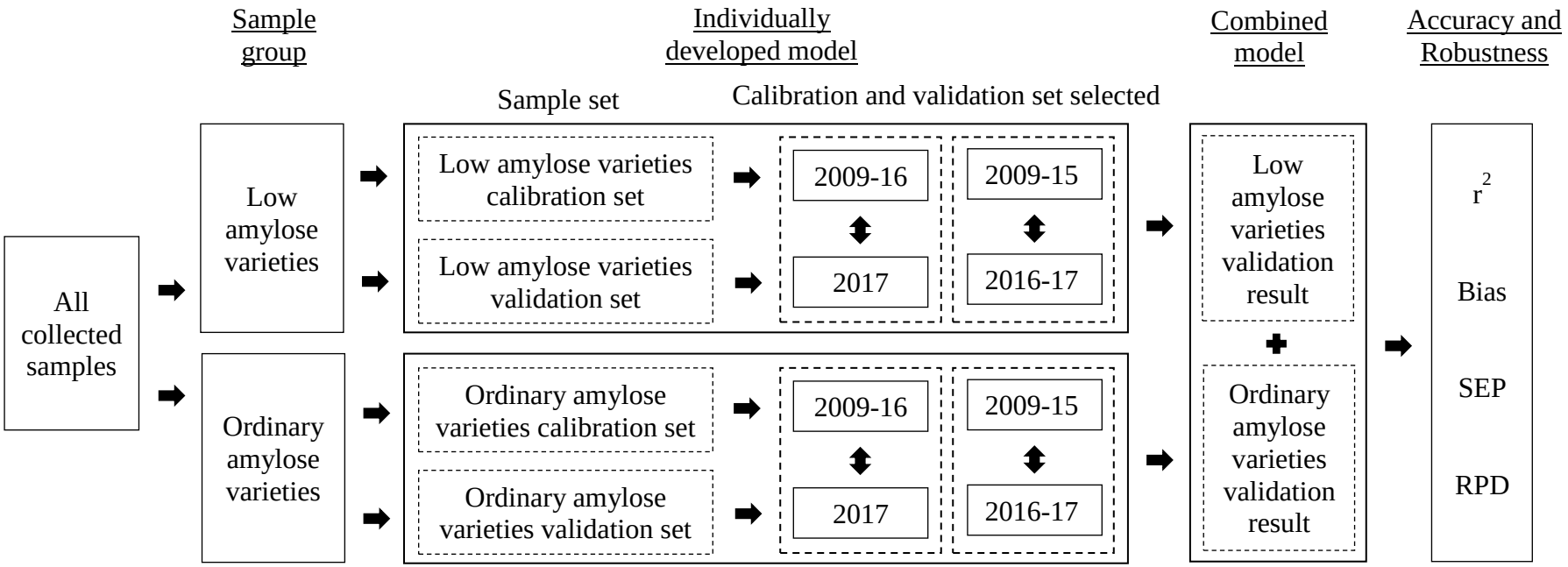
Fig. 4 Relationship between the reference and predicted amylose content values derived from each model validated either with the 2017 production year samples: "Only NIR" model (A), "NIR+CI" model (B), and "NIR+PC+PP" model (C); or with the 2016–2017 production year samples: "Only NIR" model (D), "NIR+CI" model (E), and "NIR+PC+PP" model (F). r^2, coefficient of determination; Bias, systematic error; SEP, standard error of prediction; RPD, ratio of performance deviation; n, number of samples

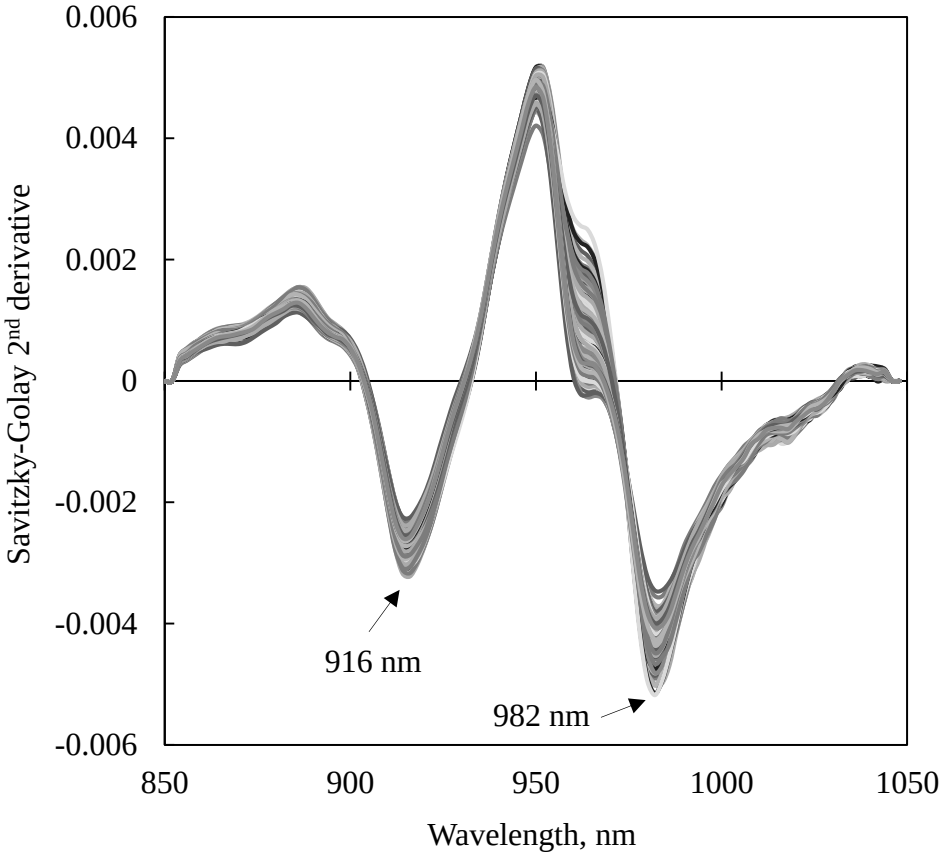
Highlights

- Development of models to assess amylose non-destructively using brown rice data and chemometrics.
- Models developed by merging the low and ordinary amylose levels validation results.
- Accuracy of developed models was suitable for industrial application.
- Combination of NIR spectra and physicochemical data revealed the most robust model.
- Contribution to more precise rice quality screening at grain elevators.


Colour Information of brown rice
MLR
Reference amylose value
PLS
NIR spectra of brown rice
Absorbance
Wavelength
MLR
Physicochemical Information of brown rice
Most accurate and robust model
Individually developed models
Production year
Combined model
Large number of samples used
Low AC model
calibration set
validation set
2009-16
2017
2009-15
2016-17
Ordinary AC model
calibration set
validation set
2009-16
2017
2009-15
2016-17
Low AC validation result
Ordinary AC validation result
r2
Bias
SEP
RPD
Accurate models to assess rice amylose non-destructively at grain elevators


(A)
(B)
(C)
AC ref
NIR Spectra
PLS Regression: Test set validation method
AC ref result
AC pred result
Validation statistics analysis
AC ref
AC pred result
CI
MLR Regression: Test set validation method
AC ref result
AC pred result
Validation statistics analysis
AC ref
AC pred result
PC pred
PP
MLR Regression: Test set validation method
AC ref result
AC pred result
Validation statistics analysis


Sample group
Individually developed model
Combined model
Accuracy and Robustness
Sample set
Calibration and validation set selected
All collected samples
Low amylose varieties
Ordinary amylose varieties
Low amylose varieties calibration set
Low amylose varieties validation set
Ordinary amylose varieties calibration set
Ordinary amylose varieties validation set
2009-16
2017
2009-15
2016-17
Low amylose varieties validation result
Ordinary amylose varieties validation result
r2
Bias
SEP
RPD


0.006
0.004
0.002
0
-0.002
-0.004
-0.006
Savitzky-Golay 2nd derivative
916 nm
982 nm
850
900
950
1000
1050
Wavelength, nm

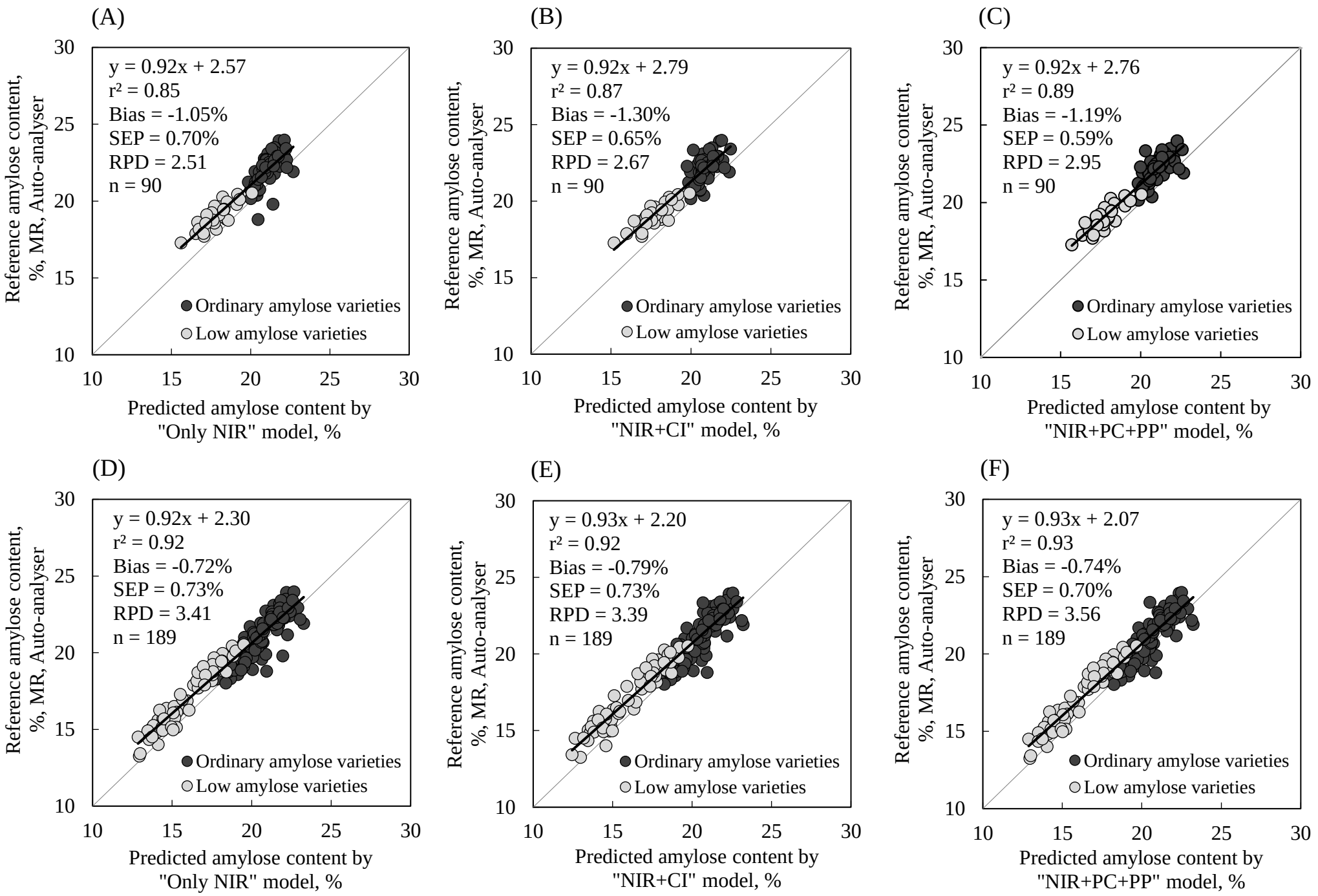

(A)
y = 0.92x + 2.57
r² = 0.85
Bias = -1.05%
SEP = 0.70%
RPD = 2.51
n = 90
Ordinary amylose varieties
Low amylose varieties
Reference amylose content, %, MR, Auto-analyser
Predicted amylose content by "Only NIR" model, %
(B)
y = 0.92x + 2.79
r² = 0.87
Bias = -1.30%
SEP = 0.65%
RPD = 2.67
n = 90
Predicted amylose content by "NIR+CI" model, %
(C)
y = 0.92x + 2.76
r² = 0.89
Bias = -1.19%
SEP = 0.59%
RPD = 2.95
n = 90
Predicted amylose content by "NIR+PC+PP" model, %
(D)
y = 0.92x + 2.30
r² = 0.92
Bias = -0.72%
SEP = 0.73%
RPD = 3.41
n = 189
(E)
y = 0.93x + 2.20
r² = 0.92
Bias = -0.79%
SEP = 0.73%
RPD = 3.39
n = 189
(F)
y = 0.93x + 2.07
r² = 0.93
Bias = -0.74%
SEP = 0.70%
RPD = 3.56
n = 189
10
15
20
25
30

Table 1. Summary of reference amylose content per variety and year of production.

Sample set	Year of production	Ordinary amylose variety								Low amylose variety		Average within production year			
		Daichinohoshi	*Fukkurinko*	*Hoshimaru*	*Hoshinoyume*	*Kitakurin*	*Kirara-397*	*Nanatsuboshi*	*Sorayuki*	*Oborozuki*	*Yumepirika*	n	Range, %	Mean, %	SEL, %
Cal	2009	-	-	-	21.6	-	-	-	-	-	18.8	35	18.8-21.6	20.2	0.15
	2010	18.7	17.9	18.3	19.1	-	18.5	17.6	-	11.9	13.6	136	11.9-19.1	16.9	0.24
	2011	21.0	20.0	-	20.5	-	20.4	19.8	-	12.8	16.6	110	12.8-21.0	18.7	0.17
	2012	-	19.4	-	-	-	-	19.1	-	-	14.7	42	14.7-19.4	17.7	0.16
	2013	19.6	18.8	-	20.3	19.5	19.5	18.6	-	13.8	14.2	99	13.8-20.3	18.0	0.23
	2014	21.7	20.8	-	21.6	20.5	20.8	19.9	-	15.3	16.3	103	15.3-21.7	19.6	0.17
	2015	23.2	20.9	-	22.5	21.2	21.6	20.1	22.9	14.2	16.8	112	14.2-23.2	20.4	0.11
Cal / Val	2016	21.0	19.9	20.0	20.5	19.8	20.4	19.1	19.4	13.9	15.5	101	13.9-21.0	19.0	0.18
Val	2017	23.7	22.4	22.5	-	21.5	23.0	21.6	21.8	17.8	19.1	94	17.8-23.7	21.5	0.14
Average within variety	n	35	80	12	45	21	104	191	18	51	275	832			
	Range, %	18.7-23.7	17.9-22.4	18.3-22.5	19.1-22.5	19.5-21.5	18.5-23.0	17.6-21.6	19.4-22.9	11.9-17.8	13.6-19.1		11.9-23.7		
	Mean, %	21.3	20.0	20.3	20.9	20.5	20.6	19.5	21.4	14.2	16.2			19.3	
	SEL, %	0.22	0.19	0.16	0.20	0.14	0.15	0.18	0.13	0.15	0.16				0.17

Cal, calibration; Val, validation; n, number of samples; SEL, average standard error of the laboratory (reference method)

Table 2. Validation statistics of each model by year of production validated with 2017 production year samples (A) and 2016-2017 production year samples (B).

(A)

Production year of the calibration set	Model	n_{cal}	n_{val}	r^2	Bias, %	SEP, %	RPD, -	Regression line equation
2009-2013	Only NIR	422	90	0.83	-1.18	0.79	2.21	y = 0.82x + 4.70
	NIR+CI	422	90	0.85	-1.53	0.83	2.11	y = 0.78x + 5.89
	NIR+PC+PP	422	90	0.87	-1.34	0.70	2.48	y = 0.85x + 4.35
2009-2014	Only NIR	525	90	0.85	-1.16	0.78	2.26	y = 0.82x + 4.80
	NIR+CI	525	90	0.87	-1.32	0.73	2.42	y = 0.83x + 4.76
	NIR+PC+PP	525	90	0.88	-1.30	0.67	2.63	y = 0.85x + 4.30
2009-2015	Only NIR	637	90	0.84	-1.06	0.77	2.29	y = 0.84x + 4.32
	NIR+CI	637	90	0.86	-1.22	0.70	2.52	y = 0.87x + 3.89
	NIR+PC+PP	637	90	0.88	-1.22	0.68	2.59	y = 0.85x + 4.29
2009-2016	Only NIR	738	90	0.85	-1.05	0.70	2.51	y = 0.92x + 2.57
	NIR+CI	738	90	0.87	-1.30	0.65	2.67	y = 0.92x + 2.79
	NIR+PC+PP	738	90	0.89	-1.19	0.59	2.95	y = 0.92x + 2.76

(B)

Production year of the calibration set	Model	n_{cal}	n_{val}	r^2	Bias, %	SEP, %	RPD, -	Regression line equation
2009-2013	Only NIR	422	189	0.91	-0.54	0.75	3.29	y = 1.02x + 0.23
	NIR+CI	422	189	0.90	-0.72	0.80	3.10	y = 1.00x + 0.74
	NIR+PC+PP	422	189	0.91	-0.54	0.74	3.34	y = 1.05x - 0.32
2009-2014	Only NIR	525	189	0.92	-0.72	0.74	3.34	y = 0.93x + 1.98
	NIR+CI	525	189	0.91	-0.80	0.74	3.34	y = 0.94x + 1.92
	NIR+PC+PP	525	189	0.92	-0.76	0.72	3.48	y = 0.95x + 1.65
2009-2015	Only NIR	637	189	0.92	-0.72	0.73	3.41	y = 0.92x + 2.30
	NIR+CI	637	189	0.92	-0.79	0.73	3.39	y = 0.93x + 2.20
	NIR+PC+PP	637	189	0.93	-0.74	0.70	3.56	y = 0.93x + 2.07

n_{cal}, number of samples of the calibration set; n_{val}, number of samples of the validation set;

r^2, coefficient of determination; Bias, difference between reference and predicted values;

SEP, standard error of prediction; RPD, ratio of performance deviation; [-], non-dimensional