Title	Combined analysis of near-infrared spectra, colour, and physicochemical information of brown rice to develop accurate calibration models for determining amylose content
Author(s)	Diaz, Edenio Olivares; Kawamura, Shuso; Matsuo, Miki; Kato, Mizuki; Koseki, Shigenobu
Citation	Food Chemistry, 286, 297-306 https://doi.org/10.1016/j.foodchem.2019.02.005
Issue Date	2019-07-15
Doc URL	http://hdl.handle.net/2115/78880
Rights	© 2019. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/
Rights(URL)	https://creativecommons.org/licenses/by-nc-nd/4.0/
Type	article (author version)
Additional Information	There are other files related to this item in HUSCAP. Check the above URL.
File Information	Revised manuscript-HUSCAP.pdf

[Instructions for use](#)

**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

29 **1. Introduction**

30

31 Rice (*Oryza sativa* L.) is an important source of nutrition and energy for mankind; its main
32 constituents are proteins, moisture, and starch (Kondo & Kawamura, 2013). Moisture content
33 affects rice storage quality and milling characteristics, whereas protein content and starch contribute
34 to the texture and eating quality of cooked rice (Siebenmorgen, Grigg & Lanning, 2013). Starch,
35 protein content, and moisture content are therefore the most important determinants of rice quality.
36 Starch makes up 90% of milled rice endosperm dry weight (Patindol, Siebenmorgen & Wang,
37 2015) and comprises two types of glucose polymers: branched amylopectin and the mostly linear
38 amylose (Biselli et al., 2014). Amylopectin, which accounts for most of the starch fraction, has
39 recently been linked to rice eating quality (Li, Prakash, Nicholson, Fitzgerald & Gilbert, 2016);
40 however, (apparent) amylose content remains the most important factor defining the palatability of
41 rice (Li, Prakash, Nicholson, Fitzgerald & Gilbert, 2016).

42 Granule-bound starch synthase 1 (GBSS-1), encoded by the Waxy gene, synthesizes amylose in the
43 endosperm during the grain-filling period (Biselli et al., 2014). Ambient air temperature strongly
44 influences amylose content. Rice varieties are classified according to amylose content as low (<
45 20%), medium (21–25%), and high (26–33%) (Biselli et al., 2014).

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

50 Iodine-binding, also known as iodine colourimetry or amylose-iodine, is the only validated and
51 most commonly used method for determining amylose content (Biselli et al., 2014). It is based on
52 measuring the absorbance of the amylose-iodine complex at 620 or 720 nm using a
53 spectrophotometer, and calculating the amylose content from a calibration curve (Fitzgerald et al.,
54 2009). Fitzgerald et al. (2009) highlighted the main adjustments made to the amylose-iodine method

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

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

66 Bagchi, Sharma and Chattopadhyay (2016) highlighted the most important developments in NIR
67 determination of amylose content using brown and milled rice flour, a single milled kernel, and
68 bulk brown and milled rice. However, Sampaio, Soares, Castanho, Almeida, Oliveira and Brites
69 (2018) reported that valuable rice amylose reference data and model performance in these studies
70 were incomplete. The authors also showed that the main challenge to determine amylose content
71 was to select the most accurate wavelength or spectral region. Accordingly, they proposed a
72 variable selection chemometric technique to define the most suitable spectral region for accurate
73 amylose content determination.

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

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

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

91 When farmers deliver the harvested rough rice at grain elevators in Japan, an automatic system
92 comprising a NIR spectrometer and a visible light (VIS) grain segregator grades its quality based on
93 highly precise and accurate measurement of moisture, protein content, and the percentage of whole
94 kernels of brown rice. The NIR spectrometer determines moisture and protein content, whereas the
95 VIS grain segregator determines the percentage of whole kernels (Kondo & Kawamura, 2013).

96 Inclusion of amylose content in the set of properties is expected to add greater precision and
97 accuracy when predicting rough rice quality.

98 At present, there are no accurate NIR calibration models for determining rice amylose content at
99 grain elevators based on brown rice information. In this study, we assessed the accuracy of three
100 models developed by combining NIR spectra, colour, and physicochemical information of brown
101 rice. We also identified the most accurate and robust model by comparing their validation statistics.

102 A suitable model could guarantee an accurate, precise, and efficient grain quality screening at
103 Japanese grain elevators. This would support the production of high-quality rice and help address
104 the demand for improved technology during rice postharvest processing.

105

106 **2. Materials and methods**

107

108 2.1 Rice samples

109

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

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

121

122 2.2 Sample preparation

123

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

129

130 2.3. Methods of measurement and devices

131

132 2.3.1 Reference analysis of amylose content of milled rice

133

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

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

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

148

149 2.3.2 Near-infrared transmittance spectra of brown rice

150

151 NIR spectra of brown rice were obtained from a bulk of whole kernels using an Omega Analyzer G
152 BR-5000 NIR transmittance spectrometer (Bruins Instruments, Fukuroi, Shizuoka, Japan) with an
153 effective spectral range of 850–1048 nm, a 2-nm increment, and a path length of 30 mm. This NIR
154 spectrometer was similar to a previously described automatic system (Kondo & Kawamura, 2013).
155 The NIR spectrum of each sample was determined for an average of three repetitions, with each
156 repetition representing an average of seven successive scans taken while scanning approximately
157 400 g of whole brown rice kernels.

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

159

160 2.3.3 *Physicochemical properties of brown rice*

161

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

166

167 2.3.3.1 *Physical properties and colour information of brown rice*

168

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

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

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

179

180 2.3.3.2 *Predicted protein content of brown rice*

181

182 Protein content is predicted based on a calibration model, which is inserted in the NIR spectrometer,
183 when rice is delivered by farmers at grain elevators. Here, the model was developed using predicted
184 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 2nd order Savitzky-Golay derivative, with 2nd 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 2nd 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 2nd 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”.

211 In the second model, the AC_{pred} result was first determined using the procedure explained in the
212 “Only NIR” model and was then linearly combined with R1G1B1, R2G2B2, and R3G3B3 colour
213 information (CI) of brown rice obtained from the VIS grain segregator. This procedure identified
214 which AC_{pred} result values correlated most closely to AC_{ref} and minimized the prediction error,
215 based on multiple linear regression (MLR) using the test set validation method. The ensuing AC_{ref}
216 result and AC_{pred} result values were used for validation statistics analysis (Fig. 1B). This model was
217 named “NIR+CI”. As two chemometric techniques (PLS and MLR) were used, the model was
218 classified as a dual-step calibration model. This method had proved highly accurate in previous
219 studies when applied to milled rice data (Kato, Kawamura, Jo, Olivares Diaz, Yokoe & Koseki,
220 2016; Kawamura, Kato, Olivares Díaz, Yokoe & Koseki, 2017).

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

230

231 2.4.3 Selection of calibration and validation sample sets

232

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

237 (2017) and two latest (2016–2017) production year samples. This strategy sought to span the range
238 of amylose content between the calibration and validation sets as uniformly as possible (Westad &
239 Marini, 2015). First, the models were calibrated by samples produced between 2009 and 2016, then
240 they were validated by samples produced in 2017. Second, the models were calibrated by samples
241 produced between 2009 and 2015, and then validated by samples produced between 2016 and 2017
242 (Fig. 2).

243

244 *2.4.4 Development and validation of the accuracy of each model*

245

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

249 Briefly, all collected samples were divided into “Low amylose varieties” and “Ordinary amylose
250 varieties” groups (explained in section 2.1). Given the large number of samples (Table 1), the
251 calibration models of the two groups were developed individually following the selection of
252 calibration and validation sets (explained in section 2.4.3). Next, a combined model was created by
253 merging the validation results of the “Low amylose varieties” and the “Ordinary amylose varieties”
254 models. The accuracy and robustness of the combined model were analysed by the coefficient of
255 determination (r^2), systematic error (Bias), SEP, and RPD (Williams, Dardenne & Flinn, 2017) (Fig.
256 2).

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

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

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

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).

275

3. Results and Discussion

277

3.1 Reference values of amylose content in milled rice

279

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%.

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

319

320 *3.2 Near-infrared transmittance spectra of brown rice*

321

322 Two absorption bands, corresponding to those identified in the raw NIR spectra (Fig. S2), were
323 detected also in the transformed NIR spectra of brown rice obtained from the Savitzky-Golay 2nd
324 derivative (Fig. 3). This indicated that functional groups in the sample molecules absorbed light
325 when irradiated in the NIR wavelength range. We assigned the strongest absorption band, at
326 approximately 982 nm, to the second overtone stretching (O-H bond) of the bound -OH alcohol
327 group. The other absorption band, at approximately 916 nm, was assigned to the third overtone of
328 symmetric stretching (C-H bond) of the methylene combination -CH₂ group (Ozaki, 2015) (Fig. 3).
329 Amylose, moisture, and proteins, which are the major constituents of rice, cause O-H and C-H bond
330 vibrations (Sampaio, Soares, Castanho, Almeida, Oliveira & Brites, 2018). However, because
331 absorption bands are complex and comprise broad and numerous overlapping bands, the
332 identification and correlation of an absorption band with a specific constituent is difficult and
333 inaccurate (Porep, Kammerer & Carle, 2015). To overcome this issue and identify the information
334 in the NIR spectra that best correlated to AC_{ref} , we applied PLS regression and maximized the
335 covariance between NIR spectra and AC_{ref} values.

336

337 *3.3 Physicochemical properties and colour of brown rice*

338

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

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

348

349 *3.4 Models for determining rice amylose content at grain elevators*

350

351 *3.4.1 Model validated using the 2017 production year samples*

352

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

356 While developing the “Only NIR” model, we identified 9 as the optimal PLS factor for both the
357 “Low amylose varieties” and “Ordinary amylose varieties” models based on the high explained
358 variance for AC_{ref} . Also, we removed two outliers from both models based on the large leverage
359 and high residual variance for AC_{ref} reported by the PLS regression results (CAMO Software As,
360 2014b).

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

characteristic of to the starch group (Font, Vélez, Del Río-Celestino, De Haro-Bailón, & Montoro, 2005; Sampaio, Soares, Castanho, Almeida, Oliveira & Brites, 2018).

While developing both the “NIR+PC+PP” and “NIR+CI” models, we identified and removed two outliers from both the “Low amylose varieties” and the “Ordinary amylose varieties” models based on the high residual variance for AC_{ref} reported by the MLR result (CAMO Software As, 2014a).

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 (Williams, Dardenne & Flinn, 2017).

We plotted the AC_{ref} values against AC_{pred} values obtained 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. 4A), the “NIR+CI” (Fig. 4B), and the “NIR+PC+PP” (Fig. 4C) models.

The SEP values of the “Only NIR” model (0.70%) (Fig. 4A), the “NIR+CI” model (0.65%) (Fig. 4B), and the “NIR+PC+PP” model (0.59%) (Fig. 4C) were below 1%, which is deemed suitable for industrial application (Kato, Kawamura, Jo, Olivares Diaz, Yokoe & Koseki, 2016). In addition, the RPD values of the “Only NIR” model (2.51) (Fig. 4A), the “NIR+CI” model (2.67) (Fig. 4B), and the “NIR+PC+PP” model (2.95) (Fig. 4C) were within 2.5–2.9, a range suitable for industrial screening (Prieto, Pawluczyk, Dugan & Aalhus, 2017). These results show that the three developed models were highly accurate for the non-destructive determination of rice amylose content.

However, because of differences in the range of AC_{ref} values between samples in the calibration and validation sets (see section 3.1), we focused our analysis on a comparison among models.

The r^2 , SEP, and RPD values of the “NIR+CI” model (Fig. 4B) were better than those of the “Only NIR” model (Fig. 4A). Kato, Kawamura, Jo, Olivares Diaz, Yokoe and Koseki (2016) and Kawamura, Kato, Olivares Diaz, Yokoe and Koseki (2017) reported a similar result on milled rice. Including the information about colour of brown rice in the model barely increased the r^2 (0.02), barely decreased the SEP (0.05%), and increased the RPD (0.16) in comparison with the “Only NIR”

model. The reduction in SEP derived from a lower standard deviation of AC_{pred} values and, consequently, increased the RPD.

The “NIR+PC+PP” model (Fig. 4C) was the most accurate, reflecting the result by Olivares Díaz, Kawamura, Jo, Kato, Matsuo and Koseki (2017) on milled rice. Inclusion of the physicochemical properties of brown rice in the model barely increased the r^2 (0.04 compared to “Only NIR” and 0.02 compared to “NIR+CI”), slightly decreased the SEP (0.11% compared to “Only NIR” and 0.06% compared to “NIR+CI”), but considerably increased the RPD (0.44 compared to “Only NIR” and 0.28 compared to “NIR+CI”) (Fig. 4A, B, C). The reduction in SEP was due to a lower standard deviation of AC_{pred} values and led to an increased RPD.

Given that the “NIR+PC+PP” model provided the best result, we defined the model equations obtained for the “Low amylose varieties” (AC_{Low}) (Eq. 4) and “Ordinary amylose varieties” ($AC_{Ordinary}$) models (Eq. 5) as follows:

$$AC_{Low} = 3.7362 + 0.9981 \cdot AC_{pred} + 0.0729 \cdot PC_{pred} + 0.0004 \cdot MK - 0.0160 \cdot IK - 0.8749 \cdot L + 0.2275 \cdot W \quad \text{Eq. 4}$$

$$AC_{Ordinary} = -6.1558 + 0.9631 \cdot AC_{pred} - 0.0069 \cdot PC_{pred} - 0.0061 \cdot MK - 0.0185 \cdot IK + 0.5944 \cdot L + 1.5979 \cdot W \quad \text{Eq. 5}$$

Considering that MLR regression in Unscrambler is based on analysis of variance (ANOVA) (CAMO Software As, 2014a), we defined the “Low amylose varieties” and “Ordinary amylose varieties” models, which comprised the “NIR+PC+PP” model, as highly significant based on p-values from ANOVA ($p < 0.001$).

Accordingly, the dual-step “NIR+PC+PP” model appears to be the most accurate and robust among the developed models.

3.4.2 Model validated using the 2016–2017 production year samples

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

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

422 While developing the “Only NIR” model, we identified 10 as the optimal PLS factor and removed
423 two outliers from the “Low amylose varieties” model and four from the “Ordinary amylose varieties”
424 model. The regression coefficients for $PLS = 10$ indicated that the highest spectral variation was at
425 916 nm, which correlated best with AC_{ref} for both the “Low amylose varieties” (Fig. S3C) and the
426 “Ordinary amylose varieties” model (Fig. S3D). As explained in section 3.4.1, this peak defined
427 absorption by the C-H bond in the 2nd derivative plot (Fig. 3) and likely corresponded to the starch
428 group.

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

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

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

438 The AC_{ref} values range was higher upon validation with the 2016–2017 production year samples
439 (14%–24%) (Fig. S1B) than with the 2017 production year samples (18%–24%) (Fig. S1A). We
440 increased the variability in AC_{ref} values and thus the standard deviation of the validation set by
441 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, Pawluczuk, 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 ($p < 0.001$).

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

Eq. 6

$$AC_{Ordinary} = -4.0877 + 0.9770 \cdot AC_{pred} - 0.0117 \cdot PC_{pred} - 0.0013 \cdot MK + 0.0020 \cdot IK + 0.1148 \cdot L + 1.3561 \cdot W \quad \text{Eq. 7}$$

Based on these results, the dual-step “NIR+PC+PP” model displayed the highest accuracy and robustness among the developed models.

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

3.5.1 Comparison among models validated using the 2017 production year samples

To further compare accuracy and robustness among models, we calibrated the models using the 2009–2013, 2009–2014, 2009–2015, and 2009–2016 production year samples and validated each calibration set with the 2017 production year samples.

While developing the models within each production year, we identified 9 as the optimal PLS factor. We also identified and removed two outliers from both the “Low amylose varieties” and the “Ordinary amylose varieties” models from each of the three models.

For all three models, accuracy increased as the production year increased. This augmented also the variability of the AC_{ref} value by increasing the number of samples in the calibration set (Esteve Agelet & Hurburgh, 2014) (Table 2A).

The “NIR+CI” model exhibited better accuracy within each production year and among them than the “Only NIR” model (Table 2A). Inclusion of colour information in the model slightly increased the r^2 , slightly decreased the SEP, and increased the RPD compared to the “Only NIR” model. As explained in section 3.4.1, this occurred as a result of reducing the standard deviation of AC_{pred} values.

The “NIR+PC+PP” model displayed the best accuracy within each production year and among them (Table 2A). Inclusion of physicochemical properties in the model decreased the SEP while 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

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

549

550 **Acknowledgements**

551

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

556

557 **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.

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

584 Font, R., Vélez, D., Del Río-Celestino, M., De Haro-Bailón, A., & Montoro, R. (2005).

585 Screening inorganic arsenic in rice by visible and near-infrared spectroscopy.

586 *Microchimica Acta*, 151(3–4), 231–239. <https://doi.org/10.1007/s00604-005-0404-x>.

587 Inatsu, O. (1988). *Studies on Improving the Eating Quality of Hokkaido Rice. Report No. 66*.

588 Sapporo, Japan. ISSN: 0367-6048.

589 Kato, M., Kawamura, S., Jo, A., Olivares Díaz, E., Yokoe, M., & Koseki, S. (2016,

590 November). Determination of Rice Amylose Content by Combined Use of a Near-

591 infrared Spectrometer and a Visible Light Segregator. Oral presentation at the meeting of

592 the 5th Asian Near-Infrared Symposium and the 32nd Japanese NIR Forum, Kagoshima,

593 Japan.

594 Kawamura, S., Kato, M., Olivares Díaz, E., Yokoe, M., & Koseki, S. (2017, June). Non-

595 destructive determination of rice amylose content: Improvement of eating quality of

596 Hokkaido grown rice by sorting amylose and protein contents. Oral presentation at the

597 meeting of the International Food Machinery & Technology Exhibition FOOMA

598 JAPAN 2017 Academic Plaza, Tokyo, Japan.

599 Kinoshita, N., Kato, M., Koyasaki, K., Kawashima, T., Nishimura, T., Hirayama, Y.,

600 Takamure, I., Sato, T., & Kato, K. (2017). Identification of quantitative trait loci for rice

601 grain quality and yield-related traits in two closely related *Oryza sativa* L. subsp.

602 japonica cultivars grown near the northernmost limit for rice paddy cultivation. *Breeding*

603 *Science*, 67(3), 191–206. <https://doi.org/10.1270/jsbbs.16155>.

604 Kondo, N., & Kawamura, S. (2013). Postharvest automation: fundamentals and practices. In

605 Q. Zhang & F. J. Pierce (Eds.), *Agricultural Automation* (pp. 367–383). Boca Raton,

606 Florida, USA: CRC Press Taylor & Francis Group.

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

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

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

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

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

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

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

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

633 <https://doi.org/10.1002/star.201400174>.

634 Porep, J. U., Kammerer, D. R., & Carle, R. (2015). On-line application of near infrared (NIR)

635 spectroscopy in food production. *Trends in Food Science & Technology*, 46(2), 211–230.

636 <https://doi.org/10.1016/j.tifs.2015.10.002>.

637 Prieto, N., Pawluczyk, O., Dugan, M. E. R., & Aalhus, J. L. (2017). A Review of the

638 Principles and Applications of Near-Infrared Spectroscopy to Characterize Meat, Fat,

639 and Meat Products. *Applied Spectroscopy*, 71(7), 1403–1426.

640 <https://doi.org/10.1177/0003702817709299>.

641 Sampaio, P. S., Soares, A., Castanho, A., Almeida, A. S., Oliveira, J., & Brites, C. (2018).

642 Optimization of rice amylose determination by NIR-spectroscopy using PLS

643 chemometrics algorithms. *Food Chemistry*, 242, 196–204.

644 <https://doi.org/10.1016/j.foodchem.2017.09.058>.

645 Siebenmorgen, T. J., Grigg, B. C., & Lanning, S. B. (2013). Impacts of Preharvest Factors

646 During Kernel Development on Rice Quality and Functionality. *Annual Review of Food*

647 *Science and Technology*, 4, 101–115. [https://doi.org/10.1146/annurev-food-030212-](https://doi.org/10.1146/annurev-food-030212-182644)

648 182644.

649 Westad, F., & Marini, F. (2015). Validation of chemometric models - A tutorial. *Analytica*

650 *Chimica Acta*, 893, 14–24. <https://doi.org/10.1016/j.aca.2015.06.056>.

651 Williams, P., Dardenne, P., & Flinn, P. (2017). Tutorial: Items to be included in a report on a

652 near infrared spectroscopy project. *Journal of Near Infrared Spectroscopy*, 25(2), 85–90.

653 <https://doi.org/10.1177/0967033517702395>.

654 Williams, V. R., Wu, W.-T., Tsai, H. Y., & Bates, H. G. (1958). Rice Starch, Varietal

655 Differences in Amylose Content of Rice Starch. *Journal of Agricultural and Food*

656 *Chemistry*, 6(1), 47–48. <https://doi.org/10.1021/jf60083a009>.

657

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 -CH₂ 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

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 calibration set
model validation set

2009-16
2017

2009-15
2016-17

Ordinary calibration set
AC model validation set

2009-16
2017

2009-15
2016-17

Low AC
validation result

Ordinary AC
validation result

r^2
Bias
SEP
RPD

Accurate models to assess rice amylose non-destructively at grain elevators

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			
		<i>Daichinohoshi</i>	<i>Fukkurinko</i>	<i>Hoshimaru</i>	<i>Hoshinoyume</i>	<i>Kitakurin</i>	<i>Kirara-397</i>	<i>Nanatsuboshi</i>	<i>Sorayuki</i>			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;

r², coefficient of determination; Bias, difference between reference and predicted values;

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