Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2025, 19(9), 50–64
https://doi.org/10.12913/22998624/205690
ISSN 2299-8624, License CC-BY 4.0

Received: 2025.03.04
Accepted: 2025.07.15
Published: 2025.08.01

Hyperspectral imaging and predictive modelling for automated control of a prototype sorting device for kiwiberry (*Actinidia arguta*)

Monika Janaszek-Mańkowska[1], Arkadiusz Ratajski[1*]

[1] Institute of Mechanical Engineering, Warsaw University of Life Sciences, Nowoursynowska 164, 02-787 Warsaw, Poland
* Corresponding autors's e-mail: arkadiusz_ratajski@sggw.edu.pl

ABSTRACT

Efficient post-harvest sorting of kiwiberry (*Actinidia arguta*) is essential for maintaining fruit quality and prolonging the viability of storage and transport. Because kiwiberry is climacteric, mixing ripe and unripe fruits may accelerate over-ripening and shorten shelf life, creating challenges for commercial distribution. This study investigates the integration of hyperspectral imaging and predictive modelling to automate the sorting of ripeness in a prototype device designed for handling kiwiberry. Hyperspectral data from 1,770 fruits were processed using transformation techniques, including Standard Normal Variate, Multiplicative Scatter Correction, Savitzky-Golay filtering, and spectral derivatives. Three regression models – Multivariate Adaptive Regression Splines, Partial Least Squares Regression, and Principal Component Regression – were evaluated to predict soluble solids content as an indicator of ripeness. Based on raw spectra, the Partial Least Squares Regression model obtained the highest accuracy, achieving a determination coefficient of 0.95, root mean squared error for prediction of 0.6778, and residual prediction deviation of 4.18 on the test set. The developed system provides a foundation for real-time, non-invasive sorting, enhancing post-harvest management, reducing waste, and advancing automated fruit sorting technologies as a practical solution for optimizing supply chain logistics in kiwiberry production.

Keywords: *Actinidia arguta*, Automated sorting, hyperspectral imaging, kiwiberry-dedicated prototype sorting device, ripeness prediction

INTRODUCTION

Actinidia arguta (Siebold et Zucc.) Planch. ex Miq. known as mini-kiwi, kiwiberry or hardy kiwi is a vine of the Actinidia genus that produces grape-like fruits with edible green, brownish or purple skin [1]. Native to Northern China, Far Eastern Russia, Japan, and Korea, kiwiberry has become an alternative to brown-skin kiwifruit and has increased its popularity worldwide. In Poland, over 25 years of research have been focused on selecting varieties suitable for commercial cultivation, with 'Weiki,' 'Geneva,' and the Polish-bred 'Bingo' [2] being the most important. Kiwiberry cultivation is concentrated mainly in the Grójecko-Warecki region and Greater Poland. Recognized in Poland as a superfruit, kiwiberry is valued for its health-promoting compounds, including high levels of vitamin C, B vitamins, vitamins A and E, carotenoids (beta-carotene, lutein), polyphenols, and chlorophylls [3, 4]. It also contains minerals (potassium, magnesium, calcium, and more), dietary fiber, and amino acids [1, 5]. These nutrients benefit the immune, metabolism, and nervous systems, while phenols and carotenoids provide anti-allergic, anti-cancer, and anti-inflammatory effects [6, 7]. In Poland and the Central European region, *A. arguta* is highly seasonal and faces commercial challenges due to its climacteric nature and uneven ripening, which limit its shelf life [8,9]. In our region, fruits require optimal post-harvest handling to extend market availability and reduce waste. Since fruits

are collected at the harvest maturity stage when they are still hard, sour, and unsuitable for consumption, technologies for proper storage and ripeness-based sorting are extremely needed but still underdeveloped. Kiwiberry ripeness, primarily determined by the balance of sweetness and acidity, is measured by soluble solids content (SSC) that in fruits consists mainly of sucrose, glucose, and fructose [10, 11]. Growers commonly use handheld refractometers to determine the optimal harvest term, but this technique requires destroying a sample. Thus, it is not suitable for large-scale sorting. Hyperspectral imaging offers a non-invasive solution, combining imaging and spectroscopy to provide spatial and spectral data. Though less precise than traditional spectroscopy, it enables real-time quality control in sorting lines and production facilities. It is a practical tool for optimizing kiwiberry processing and extending its market supply.

This study aimed to demonstrate the potential of hyperspectral imaging combined with three regression techniques to predict the SSC of kiwiberry. The goal was to develop a non-invasive, automated method for real-time ripeness assessment during the post-harvest processing of kiwiberry, enabling precise control of a prototype sorting device.

MATERIALS AND METHODS

Collection of samples

A. arguta fruits (*Weiki* variety) were randomly collected from a commercial orchard in Bodzew, Poland (51°47'50''N, 20°48'43''E, USDA Zone 6B) on 11 September 2021 at a ripeness stage corresponding to SSC of 6.5–7% (°Brix scale) [9, 12]. Stored at 1 °C and 90% relative humidity (RH), fruits were analyzed over two weeks to capture variations in ripeness, starting on the harvest day. A total of 1,770 samples were examined.

Hyperspectral imaging system

The hyperspectral imaging system (HIS) used the push-broom line scanning technique. It was composed of two hyperspectral cameras, FX10 (CMOS detector) and FX17 (cooled InGaAs detector) from SPECIM Ltd. (Oulu, Finland), as well as a 250 W halogen lamp and an external PC. The FX10 operated in the 400–1000 nm range (VNIR) with 448 bands, while the FX17 covered 900–1700 nm (NIR) with 224 bands. FX10 and FX17 cameras featured standard lenses: 15 mm and 17.5 mm, respectively, with a 38° field of view and an F-number of F/2.1. To minimize external light interference, HIS components (excluding the PC) were enclosed in a vision chamber above a conveyor, moving randomly placed fruits forward at 40 mm/s. The chamber was equipped with a forced air exhaust system (FAES) and four transducers (F&F, model MB-AHT-1) connected to a PLC (Siemens, S7-1214C) to monitor and prevent temperature rise caused by the halogen lamp. The vision chamber was the key part of the prototype sorting device dedicated to kiwiberry. The FX10's optical axis was perpendicular to the conveyor belt, while the FX17 was angled at ~12° to capture the same sample area. The halogen lamp was mounted at a 45° angle. Cameras captured 12-bit images with a mean spectral resolution of 1.32 nm (VNIR) and 3.57 nm (NIR). Additional parameters included lens-to-sample distances of 300 mm (FX10) and 315 mm (FX17), a light-to-sample distance of 250 mm, and an integration time of 2.5 ms.

Pre-processing of hyperspectral images

Both hyperspectral image acquisition and pre-processing were handled by a custom application utilizing OpenCV and Silicon Software SDK libraries. Spectral ranges outside 490–921 nm (VNIR) and 970–1617 nm (NIR) were excluded due to high noise levels and low informativeness. The spectrum range below 490 nm probably reflected chlorophyll and beta-carotene absorption peaks (~430 nm for chlorophyll *a*, ~450 nm for chlorophyll *b* and beta-carotene). Overall, 317 VNIR and 185 NIR wavebands were analyzed. Radiometric calibration, correcting for dark current, spectral light variations, and sensor sensitivity, was performed using dark (0% reflectance) and white (99% reflectance) reference images according to the formula [13]:

$$I_C = \frac{I_O - I_D}{I_W - I_D} \; [-] \tag{1}$$

where: I_C and I_O denote calibrated and original images, respectively, I_D corresponds to the dark reference image obtained for a covered camera lens with the light source turned off, whereas I_W corresponds to the image of white reference polytetrafluoroethylene tile captured before each

experimental cycle under the same lighting conditions as original images. Calibrated images were processed to collect spectral data by slicing each hypercube into 16-bit 2D images, representing reflectance in individual spectral bands. The mean spectral values were calculated for each sample in each band, resulting in a dataset of 1.770 cases and 502 reflectance variables. An automatic segmentation was applied to each hyperspectral image series to define a region of interest (ROI) corresponding only to the fruit area. The segmentation procedure was preceded by automatically selecting one image with the highest tonal range within the spectral image series to maximize contrast between sample and background reflectance. The IsoData algorithm was applied to this image, using the mean pixel intensity as the initial threshold. The resulting binary mask was applied to all spectral bands representing the sample to extract the ROI corresponding only to the fruit area and eliminate the background from further analysis. Such a procedure was applied separately for VNIR and NIR images. For each waveband λ, the mean spectral value within the ROI was calculated according to the formula (2):

$$\bar{R}_\lambda = \frac{\sum_{p=1}^{n} R_\mathrm{p}}{n} \; [-] \qquad (2)$$

where: $\bar{R}_\lambda$ is the mean reflectance within ROI in an image of an individual waveband λ, R_p corresponds to the nth pixel reflectance, and n is the number of pixels.

Spectral data processing

Even calibrated spectral data often exhibit variations from baseline drift, nonlinearity, or light scattering caused by the sample, necessitating corrections to remove irrelevant variability. In this study, we used spectral transformations, filter-based, and derivative-based methods to eliminate irrelevant information from spectral data and potentially improve the model's ability for the prediction of kiwiberry ripeness. Calibrated spectra were transformed (separately within VNIR and NIR range) by multiplicative scatter correction (MSC), standard normal variate (SNV) algorithm, Savitzky-Golay (SG) filtering, as well as the first (FD) and the second (SD) derivative. The SG filter frame size was 15, corresponding to bandwidths of 18.56 nm (VNIR) and 48.52 nm (NIR). First and second derivatives were applied after SG smoothing using first- and second-order polynomials to correct baseline shifts and slope and enhance spectral resolution by distinguishing overlapping peaks.

The MSC algorithm corrected offsets and scaled spectra to match a reference spectrum and reduce light scattering [14]. The SNV correction normalized spectra independently by calculating the difference from the mean and dividing by the standard deviation [15], making it unaffected by dataset size or variability. The Savitzky-Golay filter smoothed spectra by fitting low-order polynomial curves using convolution [16]. Derivatives localized baseline shifts (FD) and correct shifts and slopes (SD) but required smoothing to mitigate noise amplification [17, 18].

Spectra transformations, filtering, and derivatives were obtained using the 'mdatools' R package [19].

SSC for ripeness determination

SSC was adopted as the reference ripeness index of kiwiberry. After capturing hyperspectral images, it was measured by squeezing fruit juice onto a refractometer (ATAGO, PAL-1, Tokyo, Japan) with ± 0.1% accuracy in the °Brix scale (0–53%). SSC of fruits ranged from 4.4% to 16.6% (mean 8.91 ± 3.13%), covering a broad ripeness spectrum, including values beyond commercial recommendations.

Data analysis

SSC was the dependent variable for this experiment, with 502 wavelength bands as predictors. The dataset (1,770 cases) was randomly split into training (1,509 cases) and test sets (261 cases) in a ~6:1 ratio. We used a training set for model calibration (C) and evaluated a prediction performance (P) on the test set. Both sets had similar ranges, means, and standard deviations, presented in Table 1.

Data splitting was done using the '*sample'* function from the 'R core package' [20]. We applied three regression methods to predict kiwiberry ripeness: multivariate adaptive regression splines (MARS) with interaction degrees of 1 (M1) and 2 (M2), partial least squares regression (PLSR) and principal component regression (PCR). These

Table 1. SSC statistics in training and test datasets

Dataset	Cardinality	SSC – summary statistics				
		Min [%]	Max [%]	Median [%]	Mean [%]	StDev [%]
Training	1509	4.400	16.600	7.600	8.960	3.143
Test	261	4.400	15.800	7.338	8.626	3.038

Note: Min, Max – minimum and maximum value, Median – second quartile, Mean – arithmetic mean, StDev – standard deviation.

methods have the enormous advantage of reducing data dimensionality by automatic variable selection (MARS) or projection (PLSR, PCR). Models were calibrated using random 10-fold cross-validation (CV) to prevent overfitting and developed for both uncorrected and corrected spectra to compare the effects of different transformations on the accuracy of ripeness prediction. MARS models were calibrated and tested using the '*earth'* R package [21], whereas the '*pls*' R package [22] was used to calibrate and test the PLSR and PCR models. We specified the number of components to fit the PLSR and PCR models based on the standard error of cross-validation residuals, aka One SE rule [23]. Models were evaluated by the determination coefficient (R^2), root mean square error (RMSE), and residual prediction deviation (RPD). The first two statistics were determined at the calibration (R^2_C, RMSEC) and prediction (R^2_P, RMSEP) stage, whereas RPD was calculated only for prediction data. In addition, adjusted R^2 was calculated for both calibration and prediction datasets ($R^2_{adj,C}$, $R^2_{adj,P}$) as a comprehensive measure of model complexity and capability of prediction. The formulae for calculating model evaluation measures were as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \; [-] \quad (3)$$

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1} \; [-] \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - f_i)^2}{n}} \; [-] \quad (5)$$

$$RPD = \frac{StDev}{RMSEP} \; [-] \quad (6)$$

where: y_i represents the *i-th* observed value of dependent variable *Y*, f_i represents the *i-th* value of dependent variable *Y* predicted by the model at calibration or prediction stage, $\overline{y}$ stands for the sample mean of dependent variable *Y*, *n* represents the number of samples used at model calibration or prediction stage, *k* represents the number of model terms, *StDev* stands for standard deviation of dependent variable *Y* observed within the test set, and *RMSEP* represents root mean square error of *Y* prediction. The best model should be characterised by good generalisation properties, high R^2 and R^2_{adj}, low RMSEC and RMSEP values with minimal differences. Regarding RPD, values below 2 indicate poor prediction, 2.0-2.5 suggest coarse predictions, and above 2.5 indicate good or excellent prediction [24, 25].

Feature importance was determined by calculating the variable influence on projection measure (VIP) according to [26]:

$$VIP = \sqrt{k\left(\frac{\sum_{a=1}^{A}\left(W_a^2 \cdot SSY_{comp,a}\right)}{SSY_{cum}}\right)} \; [-] \quad (7)$$

where: *k* is the number of terms in the PLS model, *a* = (1, 2, …, A) is the PLS model dimension, W_a^2 stands for squared loading weight of the dimension *a*, $SSY_{comp,a}$ is the sum of squares of dependent variable explained by the PLS model dimension *a*, whilst SSY_{cum} corresponds to the total sum of squares explained by the PLS model. VIP describes how strong the relationship between each variable included in the model and the model response is, while accounting for all other predictors. According to [27], a variable should obtain a VIP index higher than 1 to be considered valid.

RESULTS AND DISCUSSION

Spectra pre-processing and its relevance for ripeness detection

Kiwiberry fruits exhibited a similar trend in spectral changes within the VNIR and NIR ranges. The overall shape of the raw kiwiberry spectra

mirrored that of kiwifruit reported by [28] and was characteristic for plants consisting of green pigments. Changes in tissue structure and chemical composition during ripening affect reflectance, as evidenced by the significant baseline variability observed among samples at different maturity stages (see Figure 1a, Figure 2a).

Analyzing the raw VNIR spectral curves, the highest variability in reflectance occurred in the visible range of 530–630 nm, which, together with the characteristic valley between 630 and 680 nm (with a global minimum at 680 nm), corresponds to sequential absorption bands of chlorophyll *a* and *b* [28]. As fruit ripens, chlorophyll degradation and the emergence of pigments such as carotenoids and anthocyanins cause reflectance to increase across this range, particularly between 530 and 630 nm. The distinct minimum near 660–680 nm becomes shallower, reflecting a reduction in chlorophyll *a* absorption.

The raw spectra between 730–920 nm showed a similar baseline shift. However, this region was also

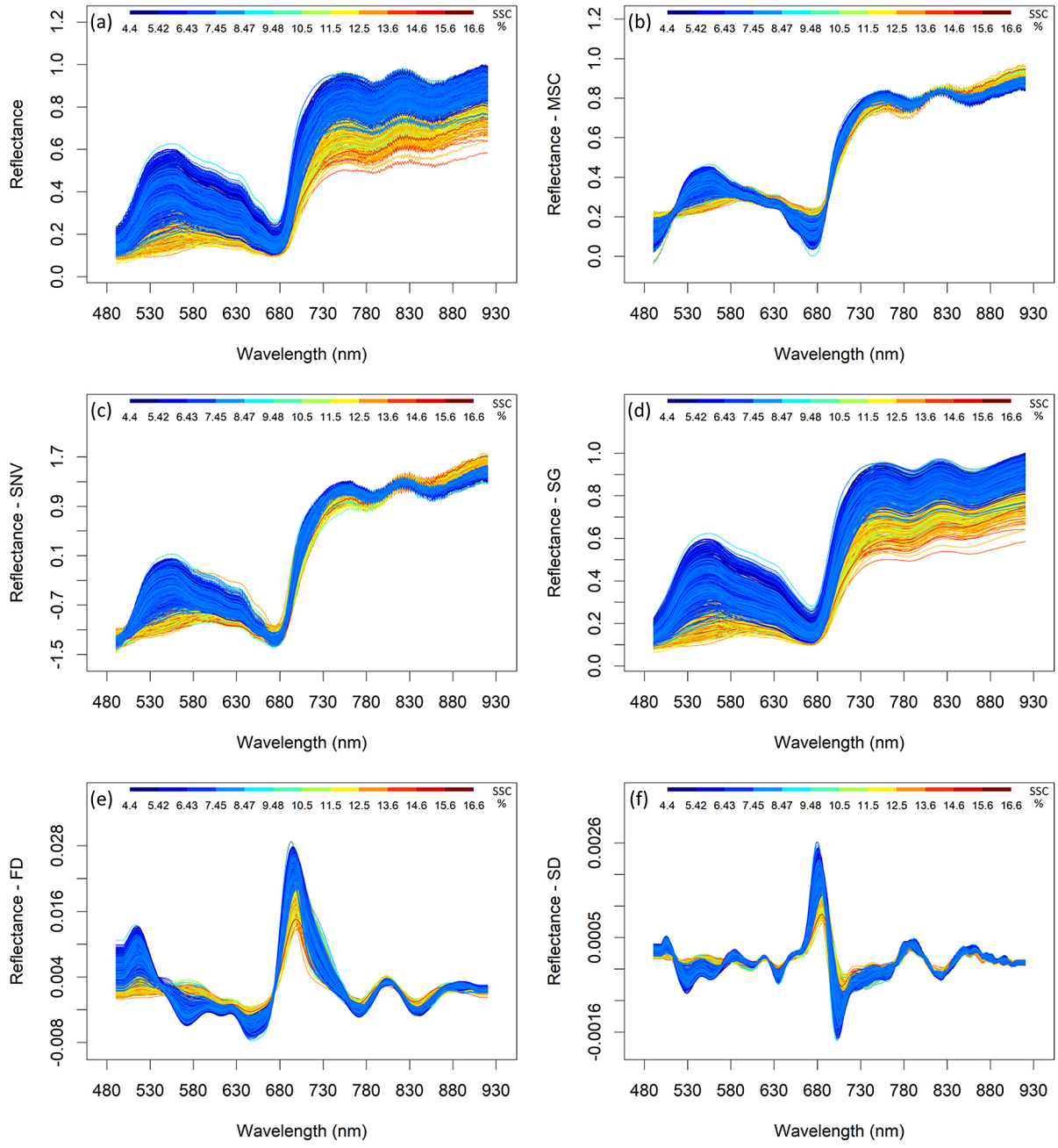

Figure 1. Reflectance spectra of samples within the VNIR range (color scale related to SSC): a) uncorrected, b) effect of MSC, c) effect of SNV, d) effect of SG filtering, e) first derivative, f) second derivative

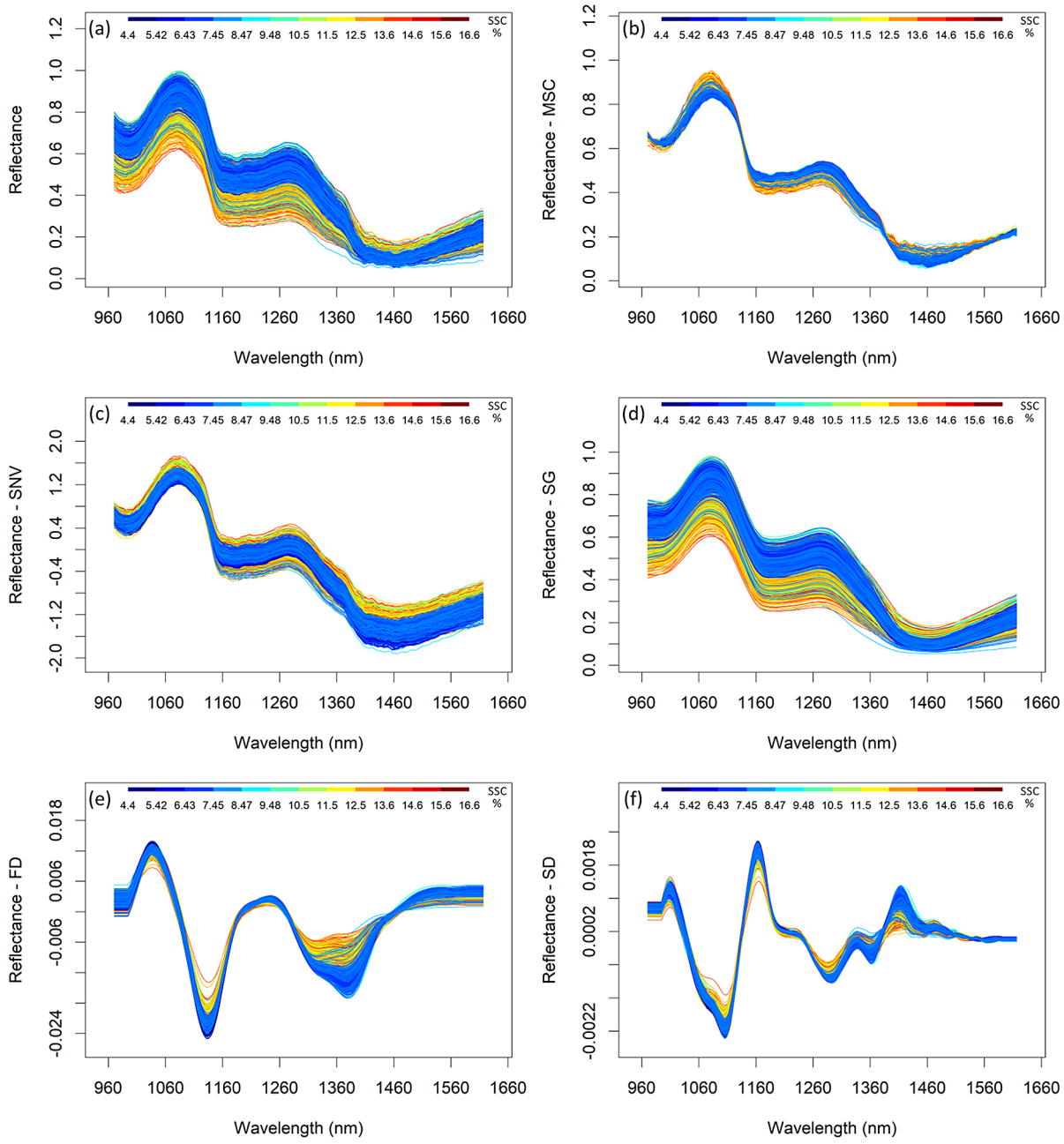

Figure 2. Reflectance spectra of samples within the NIR range (color scale related to SSC): a) uncorrected, b) effect of MSC, c) effect of SNV, d) effect of SG filtering, e) first derivative, f) second derivative

influenced by noise, which was not mitigated by the MSC or SNV algorithms since neither transformation is sensitive to local fluctuations in the adjacent spectral bands (see Figure 1b, Figure 1c). This noise is particularly concerning as it occurs within a spectral range that is informatively linked to the absorption signatures of carbohydrates, water, and sugars in kiwiberry [29, 30]. The probable cause of these minor disturbances is the diffraction or micro-reflection effect caused by the irregular surface of the fruit, which results in signal local fluctuations.

In spectra range 730–920 nm, light is not strongly absorbed by pigments but is intensively scattered in the fruit structure. As the fruit ripens, it softens because its cell walls become less rigid, pectins depolymerize, and intercellular spaces enlarge. This leads to a decrease in the structural integrity of tissues and increased light scattering, which results in higher reflectance. This phenomenon is the basis for assessing fruit firmness in NIR systems.

The NIR spectra showed similar shifts between samples, with three trough absorption peaks

occurring at approximately 1000 nm, 1180 nm, and 1460 nm (Figure 2a). These absorption bands correspond to second overtones of C–H stretching and combination bands of O–H stretching and bending vibrations, which are typical for organic compounds present in fruits. They are directly related to the presence and accumulation of sugars (mainly glucose and fructose), water (affecting the firmness), and acidity of kiwifruit [31, 32].

Spectral pre-processing by transformations MSC and SNV had a notable impact on enhancing features relevant to ripeness detection. The MSC reduced baseline shifts and improved comparability between samples with uneven surface reflection by reducing additive and multiplicative scattering effects. On the other hand, SNV minimized multiplicative effects related to surface structure and improved spectral alignment by correcting for differences in path length and light scattering, which is especially important for improving model robustness in heterogeneous biological samples. Both corrections significantly reduced sample-to-sample variability, improving the comparability of spectra, especially in spectral regions indicative of physicochemical changes during fruit ripening (see Figure 1b-c, 2b-c).

The smoothing effect was visible in spectra corrected by the SG filter, which improved signal quality and interpretability by effectively reducing high-frequency noise without distorting the spectral shape (Figure 1d, Figure 2d).

This smoothing was especially beneficial for enhancing pigment-related features in the visible region and broad absorption bands in the NIR, which are associated with SSC and moisture. By improving the signal-to-noise ratio, SG filtering supported more robust detection of ripening indicators.

Finally, the first and second derivatives highlighted distinctive extremes and changes in the spectra slopes (Figure 1e-f, Figure 2e-f). The FD transformation further improved the visibility of transitions in the spectral slopes. These transitions often reflect changes in skin pigmentation, such as chlorophyll breakdown and anthocyanin synthesis or internal composition. In turn, the SD enhanced the resolution of overlapping absorption bands, particularly in the NIR range, enabling more precise identification of weak yet informative features associated with sugars, water, and organic acids.

Together, these spectral corrections and transformations increased the interpretability of both external (pigmentation, gloss) and internal (SSC, water content, pectins) markers of ripeness.

Minimizing artifacts and enhancing chemical features enabled more accurate monitoring of ripening progression based on hyperspectral data.

Modelling results for uncorrected data

The fit statistics for the models developed on raw data are summarized in Table 2, whereas scatter plots comparing measured and predicted SSC for both calibration and test data are presented in Figure 3.

Models based on latent variables (PLSR and PCR) were the most suitable for SSC prediction, with R^2_P of 0.95 and 0.9497, respectively. The prediction errors (RMSEP) were nearly identical in these models, with a value of approximately 0.68. Although the PCR model had a more complex structure than the PLSR model, it was not accompanied by higher predictive ability. This was evidenced by the $R^2_{adj,P}$ value, which was lower by 0.04 for the PCR model at the prediction stage compared to the PLSR model. In contrast, M2 model produced prediction results comparable to the PCR model, with $R^2_{adj,P} = 0.9076$ utilizing only half the variables. Nevertheless, the RMSEP of 0.921 for the M2 model was significantly higher than that of the PLSR and PCR models. The M2 model outperformed the PCR model in terms of RPD value, which was 0.06 higher than the PCR model. Although the M1 model was the simplest in structure, it yielded the poorest performance at both the calibration and prediction stages.

Table 2. Prediction results of different models based on uncorrected spectra

Model	No. of variables	Calibration			Prediction			
		R^2_C	$R^2_{adj,C}$	RMSEC [%]	R^2_P	$R^2_{adj,P}$	RMSEP [%]	RPD
M1	21	0.8926	0.8911	1.0388	0.8836	0.8734	1.0787	2.83
M2	68	0.9571	0.9550	0.6676	0.9318	0.9076	0.9210	3.31
PLSR	**35**	**0.9649**	**0.9640**	**0.5903**	**0.9500**	**0.9420**	**0.6778**	**4.18**
PCR	123	0.9630	0.9597	0.6055	0.9497	0.9039	0.6796	3.25

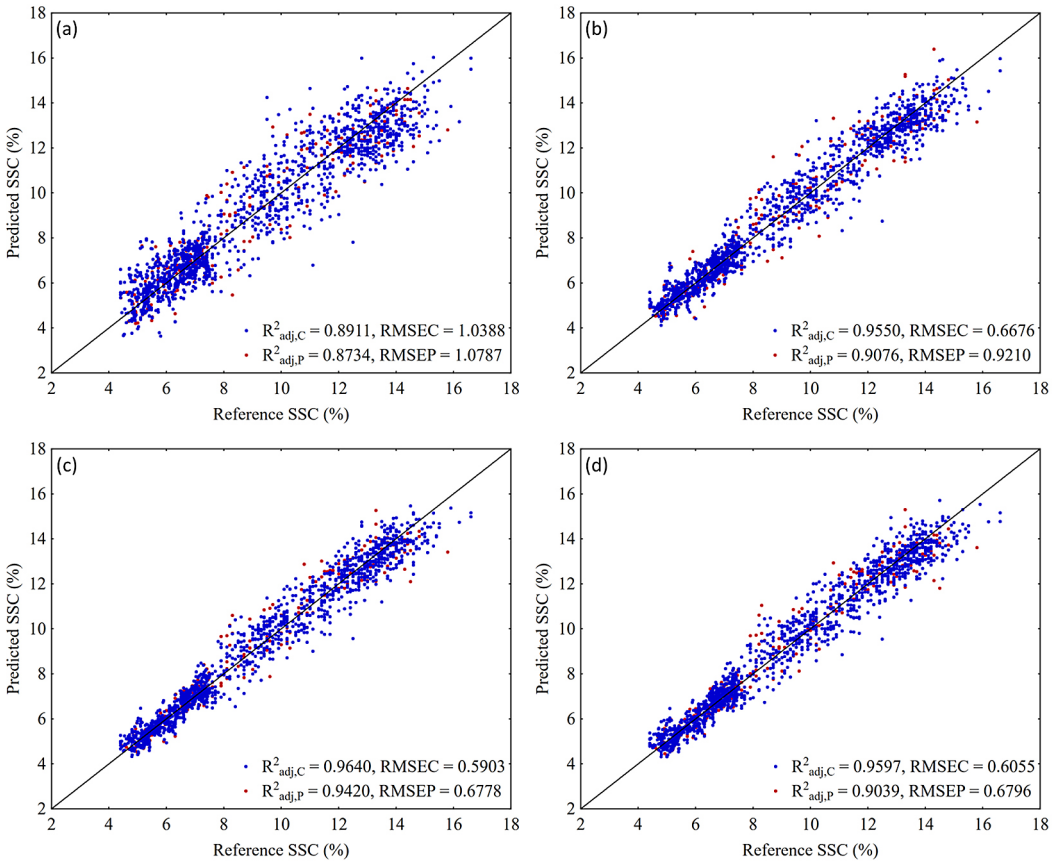

Figure 3. Prediction plots for SSC obtained using different regression methods on uncorrected calibration and test data: a) result for additive model – M1, b) result for the model with degree of interaction 2 – M2, c) result for the PLS model, d) result for the PCR model

Its goodness-of-fit measures at the calibration ($R^2_{adj,C}$ = 0.8911) and prediction stages ($R^2_{adj,P}$ = 0.8734) were the lowest among all models, and the errors in both stages exceeded the value of 1. This model also achieved the lowest RPD value of 2.83. Considering factors such as goodness-of-fit, prediction ability, errors, and complexity, the model established using PLSR on raw data was superior, accounting for 6.97% of the extracted latent variables.

Prediction of ripeness by multivariate adaptive regression splines

Fit statistics for models obtained using MARS are presented in Table 3. Among the M1 models, the SNV+M1 was the simplest in structure but demonstrated the least favourable values of $R^2_{adj,C}$ and $R^2_{adj,P}$, at 0.8293 and 0.8122, respectively. It also exhibited the highest RMSEC and RMSEP values, both exceeding one and achieved the lowest RPD value of 2.32, indicating that the prediction precision of this model was relatively coarse. In contrast, the SNV+M2 model was the most complex and achieved the best calibration statistics with $R^2_{adj,C}$ of 0.9627 and RMSEC of 0.6080. Nonetheless, it performed markedly worse during the prediction stage, ranking third with an $R^2_{adj,P}$ of 0.8816 and RMSEP of 1.0421. Its RPD value was 0.61 higher than that of SNV+M1. Similar results were observed in the SG+M1 and FD+M1 combinations, which recorded $R^2_{adj,C}$ values of 0.9446 and 0.9440, while $R^2_{adj,P}$ accounted for 0.8776 and 0.8887, respectively. The fitting errors of these models were also comparable; unfortunately, during the prediction stage, both models had RMSEP values exceeding one. Despite incorporating additional variables, neither SG+M2 nor FD+M2 significantly outperformed their M1 counterparts. The M1 and M2 models established for MSC-corrected data exhibited similar complexity and fit statistics during calibration. Moreover, the MSC+M1 model achieved the lowest RMSEC of 0.6796 and the highest $R^2_{adj,C}$ of 0.9534. This model ranked second in the prediction stage, following M1+SD, with an

Table 3. Prediction results of MARS models based on different spectra corrections

Model	No. of variables	Calibration			Prediction			
		R^2_C	$R^2_{adj,C}$	RMSEC [%]	R^2_P	$R^2_{adj,P}$	RMSEP [%]	RPD
MSC+M1	65	0.9554	0.9534	0.6796	0.9304	0.9071	0.9233	3.30
SNV+M1	19	0.8315	0.8293	1.3008	0.8259	0.8122	1.3136	2.32
SG+M1	62	0.9469	0.9446	0.7408	0.9068	0.8776	1.0599	2.88
FD+M1	59	0.9462	0.9440	0.7450	0.9139	0.8887	1.0110	3.02
SD+M1	**52**	**0.9520**	**0.9503**	**0.7019**	**0.9289**	**0.9112**	**0.9031**	**3.38**
MSC+M2	66	0.9551	0.9530	0.6822	0.8993	0.8650	1.1131	2.74
SNV+M2	99	0.9651	0.9627	0.6080	0.9267	0.8816	1.0421	2.93
SG+M2	69	0.9491	0.9467	0.7270	0.9130	0.8816	1.0427	2.93
FD+M2	60	0.9494	0.9473	0.7229	0.9147	0.8891	1.0091	3.02
SD+M2	**55**	**0.9564**	**0.9547**	**0.6699**	**0.9337**	**0.9159**	**0.8789**	**3.47**

$R^2_{adj,P}$ of 0.9071, RMSEP of 0.9233, and RPD of 3.30. In contrast, the MSC+M2 model achieved an $R^2_{adj,P}$ of 0.8650, RMSEP of 1.1131, and RPD of 2.74, ranking it lowest regarding predictive ability among all M2 models. The final two combinations, SD+M1 and SD+M2, demonstrated the best predictive outcomes, with the highest $R^2_{adj,P}$ of 0.9112 and 0.9159, and the lowest RMSEPs of 0.9031 and 0.8789. At the calibration stage, both models performed well and secured second place, trailing only the models built on MSC-corrected data. The two models also had minimal differences in complexity and predictive performance. Therefore, when considering adjusted coefficients of determination, error metrics, and the differences between calibration and prediction stages, the SD+M2 model emerged as the most favourable, with an RPD of 3.47, which indicates strong predictive capability. The prediction results for the SD+M1 and SD+M2 models during the calibration and prediction stages are illustrated in Figure 4 for comparison.

Prediction of ripeness by partial least squares regression

The modelling results obtained using PLSR on data subjected to various spectral corrections are summarized in Table 4.

All PLSR models generally produced similar prediction outcomes despite variations in the number of variables used (Table 4). Moreover, these models demonstrated superior fit statistics compared to those obtained using the MARS method.

Errors obtained on the calibration set did not exceed 0.63. The highest prediction error was achieved by the model created on SNV-corrected data and amounted to 0.7336. In contrast, the model created on data corrected by the MSC algorithm had a slightly lower value of RMSEP value of 0.7110. Notably, both models utilized the same number of latent variables. Additionally, applying the Savitzky-Golay filter resulted in a modest improvement in model goodness-of-fit, achieving an $R^2_{adj,P}$ of 0.9386 while also reducing the RMSEP to 0.6737, albeit at the cost of an increased number of variables compared to the two earlier models. The model developed using the first derivative spectra emerged as the most complex, comprising 61 variables. At the prediction stage, it achieved a commendable R^2_P value of 0.9485. However, its $R^2_{adj,P} = 0.9323$ remained nearly on par with the SNV-PLSR model, which had twice as many variables while exhibiting only a marginally lower RMSEP at 0.6883. The second derivative spectra generated the most effective model, showcasing the best fit during calibration and prediction with $R^2_{adj,C} = 0.9634$ and $R^2_{adj,P} = 0.9411$. This model ranked among the top three PLSR models for prediction accuracy, securing second place with an RMSEP of 0.6832. It consisted of the smallest number of components (just 35), accounting for only 6.97% of the extracted latent variables. Figure 5 illustrates the SSC prediction results for this model across both calibration and test datasets.

Prediction of ripeness by Principal Component Regression

Results for PCR models obtained for different methods of data processing are shown in Table 5. PCR models showed the highest complexity among the three methods, with the simplest

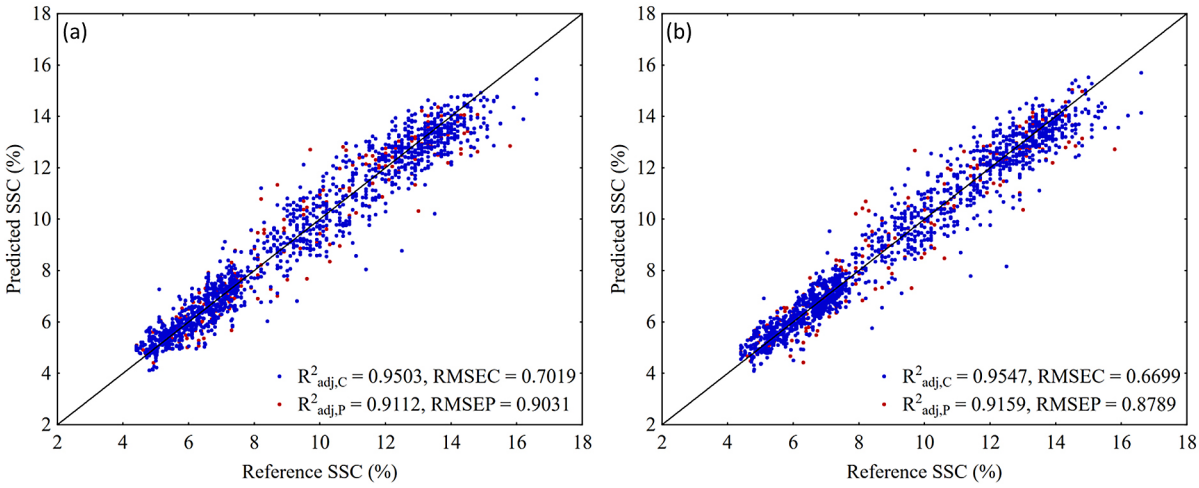

Figure 4. Prediction plots for SSC obtained with MARS on calibration and test data subjected to SD processing: a) results for additive model – M1, b) results for the model with degree of interaction 2 – M2

Table 4. Prediction results of PLSR models based on different spectra corrections

Correction	No. of variables	Calibration			Prediction			
		R^2_C	$R^2_{adj,C}$	RMSEC [%]	R^2_P	$R^2_{adj,P}$	RMSEP [%]	RPD
MSC	32	0.9615	0.9607	0.6176	0.9450	0.9370	0.7110	4.01
SNV	32	0.9609	0.9601	0.6222	0.9414	0.9329	0.7336	3.89
SG	50	0.9635	0.9623	0.6013	0.9506	0.9386	0.6737	4.06
FD	61	0.9614	0.9597	0.6188	0.9485	0.9323	0.6883	3.87
SD	**35**	**0.9643**	**0.9634**	**0.5949**	**0.9492**	**0.9411**	**0.6832**	**4.15**

requiring 114 latent variables. In contrast, the optimal PLSR model achieved better predictions with over three times fewer variables. Despite its complexity (132 components) and a decent R^2_P value of 0.9469, the MSC-PCR model showed the weakest fit, with $R^2_{adj,P}$ and RMSEP values of 0.8337 and 0.6985, respectively.

Models based on SNV and SD processing performed better, achieving $R^2_{adj,P}$ values of 0.8843 and 0.8867, respectively. The former model adopted 132 components, achieving RMSEC = 0.6275 and RMSEP = 0.7209, while the latter improved RMSEC to 0.6023 and RMSEP to 0.6846, albeit with 10 additional components. Models developed on SG and FD processed data delivered comparable calibration results, with SG-PCR model using 127 components, $R^2_{adj,C}$ of 0.9594 and RMSEC of 0.6070, and FD-PCR model requiring just 114 components to gain $R^2_{adj,C}$ of 0.96 and RMSEC of 0.6049. At the prediction stage, SG-PCR and FD-PCR achieved RMSEP values of 0.6709 and 0.6916, with nearly identical $R^2_{adj,P}$ of 0.9035 and 0.9067, respectively. Balancing fit and complexity, the FD-PCR model emerged as the most effective, with an $R^2_{adj,P}$ of 0.9067. Figure 6 highlights its SSC prediction performance on calibration and prediction datasets.

The best prediction models and importance of variables

The results of SSC prediction achieved with various data correction and modelling methods highlighted the superiority of smoothing and derivation techniques, especially SD processing. Among the best models built on corrected data, the SD-PLSR model demonstrated the best performance, balancing prediction accuracy, complexity, and error rate. Notably, the PLSR model built on raw spectral data achieved comparable prediction ability, a slightly higher RPD, and a lower RMSEP, using the same number of variables as the SD-PLSR model. Both models performed well but differed in their reliance on spectral features. Analysis of variable importance in projection (VIP) values revealed that

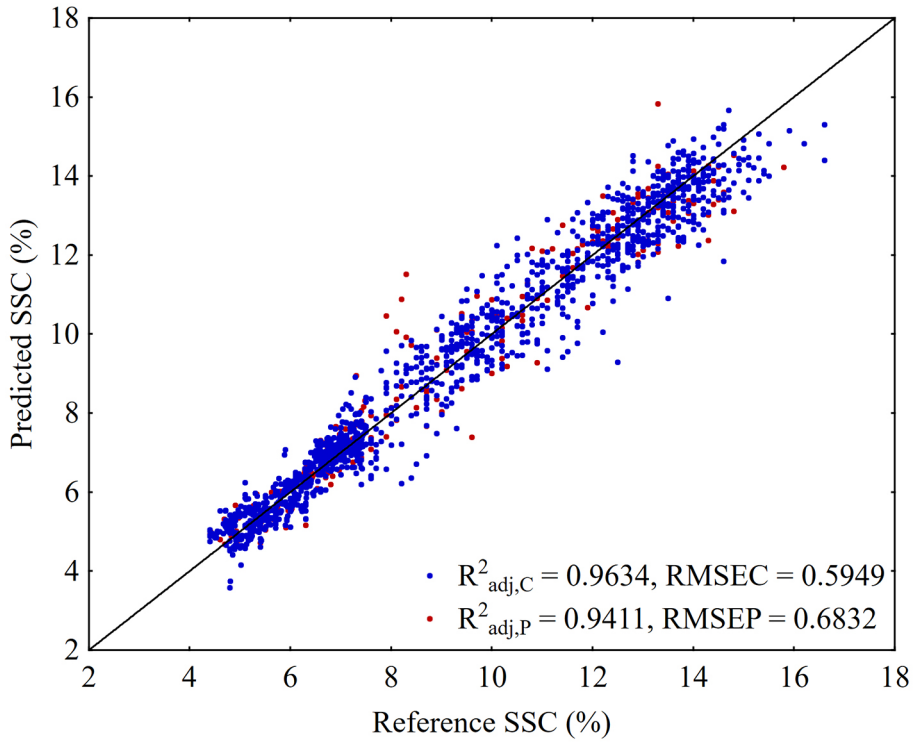

Figure 5. Prediction plot for SSC obtained for PLS regression on calibration and test data subjected to SD processing

Table 5. Prediction results of PCR based on different spectra corrections

Correction	No. of variables	Calibration			Prediction			
		R^2_C	$R^2_{adj,C}$	RMSEC [%]	R^2_P	$R^2_{adj,P}$	RMSEP [%]	RPD
MSC	176	0.9632	0.9583	0.6040	0.9469	0.8337	0.6985	2.48
SNV	132	0.9603	0.9564	0.6275	0.9435	0.8843	0.7209	2.96
SG	127	0.9628	0.9594	0.6070	0.9510	0.9035	0.6709	3.25
FD	**114**	**0.9631**	**0.9600**	**0.6049**	**0.9480**	**0.9067**	**0.6916**	**3.30**
SD	142	0.9634	0.9596	0.6023	0.9490	0.8867	0.6846	3.00

data corrections altered the predictive power of specific wavebands, a phenomenon also noted by [33, 34]. Considering the number of original features to be the same as the number of model components, we identified 35 original features that affected the model's performance the most. Those and all considered valid features are presented in Figure 7.

The raw-data model relied heavily on the NIR range 1396–1517 nm, which is beyond the 729–975 nm region, considered the prime spectra linked to sugar, carbohydrate, and water absorption. VNIR wavebands in 670–686 nm and 704–823 nm ranges had VIP values above one but did not rank among the top 35 features. In contrast, the SD-PLSR model utilized a broader spectrum, including visible and early NIR ranges (626–886 nm and 1102–1446 nm). Only four wavebands overlapped between the two models: 1421.39 nm, 1439.07 nm, 1442.6 nm, and 1446.14 nm, which suggests that spectral corrections influence the between-sample variation, though they may not be essential under consistent acquisition conditions with sufficient calibration samples.

Comparisons with previous studies underscore the robustness of the presented models. Sarkar [30] reported weaker performance with PLSR species-dependent models developed for SSC prediction (range 5–27.2%) in South Korean kiwifruits, achieving an R^2_p of 0.775, RMSEP of 1.8775 and RPD of 2.14. Similarly, Moghimi [35] obtained SSC prediction models with a maximum correlation of 0.93 using combined

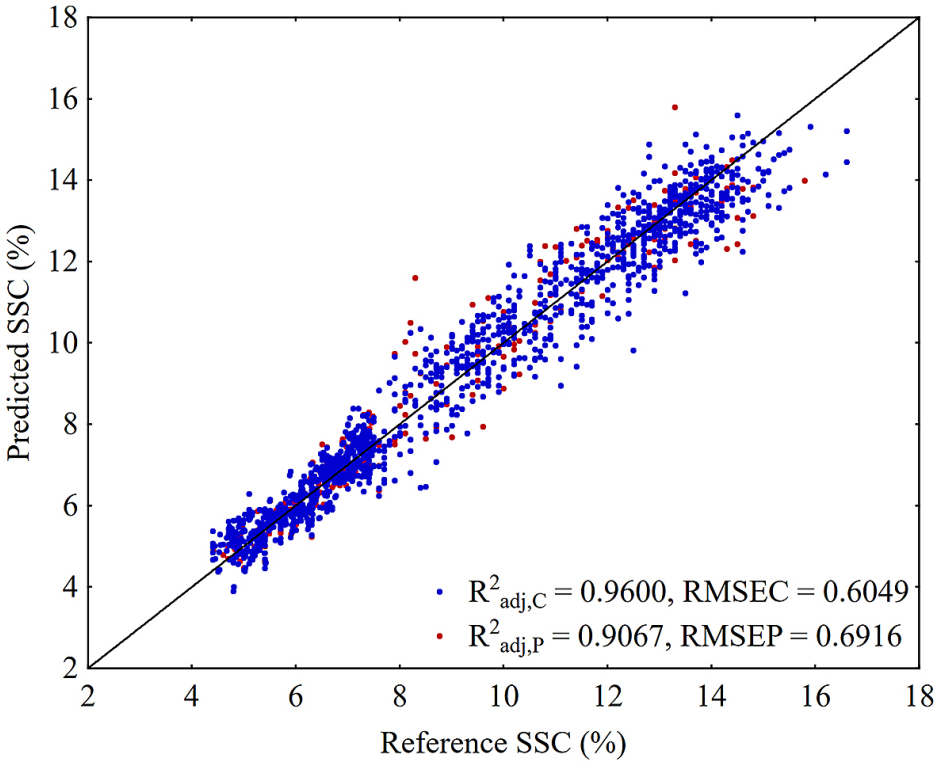

Figure 6. Prediction plots for SSC obtained for PCA regression on calibration and test data subjected to FD processing

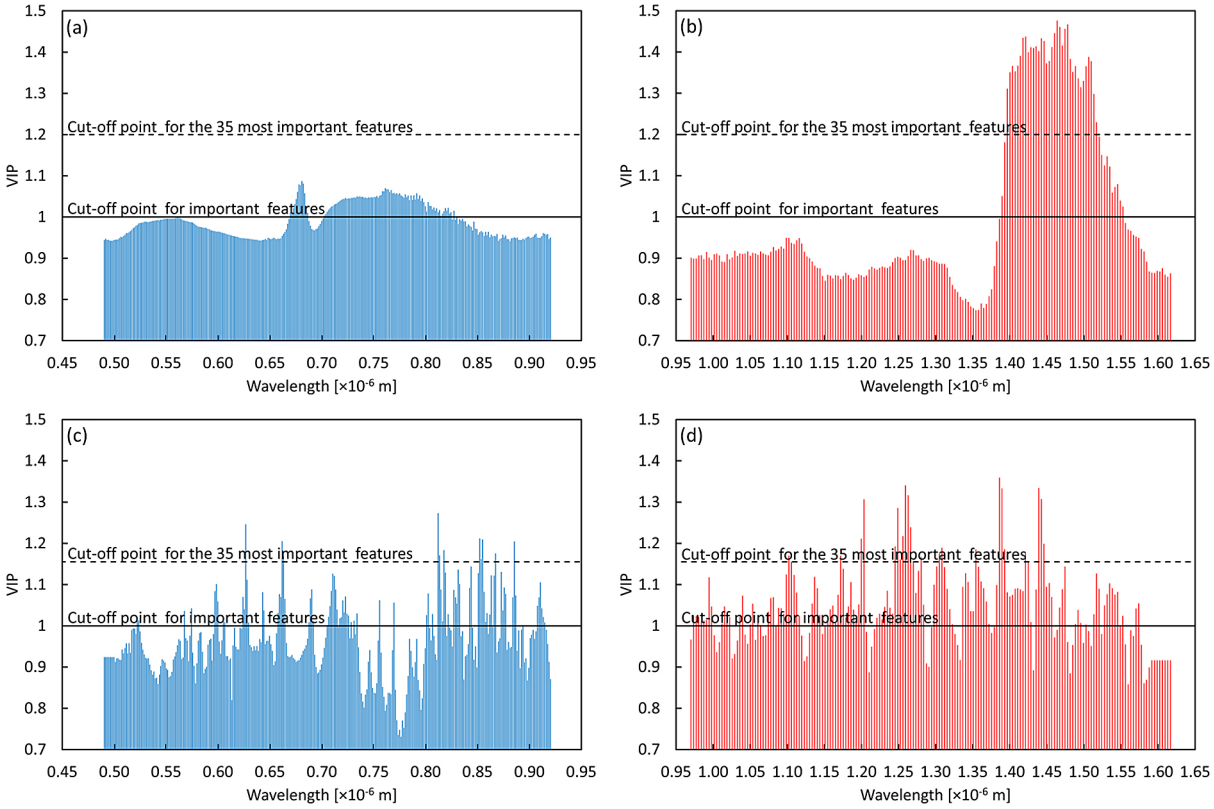

Figure 7. The 35 wavebands most contributing to response predicted by PLSR models developed on: a) uncorrected VNIR spectra, b) uncorrected NIR spectra, c) second derivative VNIR spectra, d) second derivative NIR spectra. Black horizontal lines correspond to cut-off points of considering a feature as valid (solid line) or as one of the 35 most important features (dashed line)

correction methods (the best result obtained for the combination of SNV, median filter and first derivative). These models were restricted to kiwifruits with a narrow SSC range above 11%. Mumford [25] considered a wider range of SSC (from 4.7 to 27.3%) and achieved strong predictions for kiwiberries ('Geneva 3' variety), reporting an R^2 of 0.96 as well as an RPD of 5.08 with a PCR model using only five principal components. However, their results relied solely on cross-validation and a limited spectral range (729–975 nm). Lee [36] developed a model for 'Deliciosa' kiwifruit with an SSC range of 5.30–17.60%, achieving a validation correlation of 0.98 using a broader NIR spectrum (408–2492 nm with spectral sampling of 8 nm). In contrast, Lee [24] applied an orthogonal signal correction to spectra (730–2300 nm with a spectral sampling of 1 nm) for the prediction of SSC in the range of 10.1–25.1% in hardy kiwi, obtaining an R^2 of 0.86, RMSEP of 1.33 and RPD of 2.61. Zhu [37] improved upon this with a PLSR model using successive projections algorithm to predict the SSC of kiwifruit in the range 13.56–18.69%, reporting an R^2_p of 0.9523, RMSEP of 0.4042, and RPD of 3.26, but their dataset was notably small (133 samples). Xu [28] compared hyperspectral and fluorescence spectral imaging to predict kiwifruit SSC (6.5–15.36%). Their best PLSR model, developed with boxing smoothing, achieved an R^2 of 0.76, RMSEP of 0.6876 and RPD of 2.21. Feature selection techniques such as CARS, MASS, and IVISSA combined with the Boss strategy improved the results, yielding an R^2 of 0.87, RMSEP of 0.5661 and RPD of 2.89. Results obtained for other models were similar (ELM model) or slightly worse (PSO-LSSVM combination). Similarly, Kim [29] predicted baby kiwifruit maturity using PLSR on NIR spectra (729–975 nm), obtaining an R^2 of 0.73 and RMSEP of 1.24, but based only on cross-validation procedure. Finally, Benelli [32] combined SNV, first derivative, and mean centering with genetic algorithms to develop models for SSC prediction of kiwifruit. Their best model achieved an R^2 of 0.94 and RMSEP of 0.73, underscoring the importance of optimal preprocessing and variable selection techniques.

Our SD-PLSR and raw-data PLSR models surpass many of these benchmarks, providing promising tools for predicting kiwiberry ripeness under practical conditions.

CONCLUSIONS

This study used hyperspectral imaging as the leading technology to predict the SSC of kiwiberry of the 'Weiki' variety in a non-invasive manner. A total of 1,770 research samples were collected for this experiment. Five different spectra transformation techniques (SNV, MSC, SG, FD, and SD) in combination with three regression approaches (MARS, PLSR, and PCR) were investigated to find the best model for SSC prediction. The findings of this study indicate that hyperspectral imaging, combined with effective acquisition technology, demonstrates significant potential for predicting kiwiberry ripeness based on SSC. The most accurate prediction model was developed using the PLS regression technique and was based on uncorrected data. A second model, nearly identical in its predictive capability, was derived from data subjected to Savitzky-Golay smoothing and second derivative analysis. Both models demonstrate excellent predictive capabilities, evidenced by their RPD values of 4.18 and 4.15, respectively, with goodness-of-fit exceeding 0.94. This is encouraging, as it paves the way for developing devices to sort kiwiberry fruits based on their ripeness. Such advancements are crucial for enhancing the previously developed technology for the commercial cultivation of *A. arguta*, thereby increasing production profitability and facilitating the further development of this plant in Poland. However, the authors recognize that the models calibrated for one specific *A. arguta* variety may yield less accurate predictions for fruits of other varieties. Consequently, ongoing studies are being conducted with various kiwiberry varieties prevalent in Poland. Given the genetic, phenotypic, and inter-sample variances among kiwiberry varieties, it may not be realistic to expect the creation of a universal SSC prediction model for this species without encountering Simpson's paradox (Simpson, 1951). Therefore, calibrating models according to each variety and utilizing multi-seasonal observations appears more pragmatic.

Acknowledgments

The research was supported by the Agency for Restructuring and Modernisation of Agriculture in Poland (ARMA) under the contract 00011.DDD.6509.00015.2019.07.

REFERENCES

1. Latocha P. The nutritional and health benefits of kiwiberry (*Actinidia arguta*) - a Review. Plant Foods for Human Nutrition, 2017; 72(4): 325–334. https://www.doi.org/10.1007/s11130-017-0637-y
2. Latocha P. Some morphological and biological features of 'Bingo' – a new hardy kiwifruit cultivar from Warsaw University of Life Sciences (WULS) in Poland. Yearbook of the Polish Dendrological Society, 2012; 60: 61–67.
3. Latocha P., Łata B., Stasiak A. Phenolics, ascorbate, and the antioxidant potential of kiwiberry vs. common kiwifruit: The effect of cultivar and tissue type. Journal of Functional Foods, 2015; 19: 155–163. https://www.doi.org/10.1016/j.jff.2015.09.024
4. Leontowicz H., Leontowicz M., Latocha P., Jesion I., Park Y.S., Katrich E., Barasch D., Nemiroski A., Gorinstein S. Bioactivity and nutritional properties of hardy kiwi fruit Actinidia arguta in comparison with Actinidia deliciosa 'Hayward' and Actinidia eriantha 'Bidan'. Food Chemistry, 2016; 196: 281–291. https://www.doi.org/10.1016/j.foodchem.2015.08.127
5. Latocha P., Debersaques F., Decorte, J. Varietal differences in the mineral composition of kiwiberry - Actinidia arguta (Siebold et Zucc.) Planch. ex Miq. Acta Horticulturae, 2015; 1096: 479–486. https://www.doi.org/10.17660/ActaHortic.2015.1096.59
6. Pinto D., Delerue-Matos C., Rodrigues F. Bioactivity, phytochemical profile, and pro-healthy properties of *Actinidia arguta*: A review. Food Research International, 2020; 136: 109449. https://www.doi.org/10.1016/j.foodres.2020.109449
7. Wojdyło A., Nowicka P., Oszmiański J., Golis T. Phytochemical compounds and biological effects of Actinidia fruits. Journal of Functional Foods, 2017; 30: 194–202. https://www.doi.org/10.1016/j.jff.2017.01.018
8. Fisk C.L., McDaniel M.R., Strik B.C., Zhao Y. Physicochemical, sensory, and nutritive qualities of hardy kiwifruit (*Actinidia arguta* 'Ananasnaya') as affected by harvest maturity and storage. Journal of Food Science, 2006; 71(3): S204–S210. https://www.doi.org/10.1111/j.1365-2621.2006.tb15642.x
9. Latocha P., Krupa T., Jankowski P., Radzanowska J. Changes in postharvest physicochemical and sensory characteristics of hardy kiwifruit (*Actinidia arguta* and its hybrid) after cold storage under normal versus controlled atmosphere. Postharvest Biology and Technology, 2014; 88: 21–33. https://www.doi.org/10.1016/j.postharvbio.2013.09.005
10. Boyes S., Strübi P., Marsh H. Sugar and organic acid analysis of *Actinidia arguta* and rootstock-scion combinations of Actinidia arguta. Lebensmittel-Wissenschaft und -Technologie, 1996; 30(4): 390–397. https://www.doi.org/10.1006/fstl.1996.0201
11. Nishiyama I., Fukuda T., Shimohashi A., Oota T. Sugar and organic acid composition in the fruit juice of different Actinidia varieties. Food Science and Technology Research 2008; 14(1): 67–73. https://www.doi.org/10.3136/fstr.14.67
12. Stefaniak J., Przybył J.L., Latocha P., Łata B. Bioactive compounds, total antioxidant capacity and yield of kiwiberry fruit under different nitrogen regimes in field conditions. Journal of the Science of Food and Agriculture, 2020; 100: 3832–3840. https://www.doi.org/10.1002/jsfa.10420
13. Xiong Z., Xie A., Sun D.-W., Zeng X.-A., Liu D. Applications of hyperspectral imaging in chicken meat safety and quality detection and evaluation: A review. Critical Reviews in Food Science and Nutrition, 2015; 55: 1287–1301. https://www.doi.org/10.1080/10408398.2013.834875
14. Geladi P., MacDougall D., Martens H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. Applied Spectroscopy, 1985; 39(3): 491–500. https://www.doi.org/10.1366/0003702854248656
15. Barnes R.J., Dhanoa M.S., Lister S.J. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. Applied Spectroscopy, 1989; 43(5): 772–777. https://www.doi.org/10.1366/0003702894202201
16. Savitzky A., Golay M.J.E. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry. 1964, 36(8): 1627–1639. https://www.doi.org/10.1021/ac60214a047
17. Maleki M.R., Mouazen A.M., Ramon H., De Baerdemaeker J. Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. Biosystems Engineering, 2007; 96(3): 427–433. https://www.doi.org/10.1016/j.biosystemseng.2006.11.014
18. Witteveen M., Sterenborg H.J.C.M., van Leeuwen T.G., Aalders M.C.G., Ruers T.J.M., Post A.L. Comparison of preprocessing techniques to reduce nontissue-related variations in hyperspectral reflectance imaging. Journal of Biomedical Optics, 2022; 27(10): 106003. https://www.doi.org/10.1117/1.JBO.27.10.106003
19. Kucheryavskiy S. mdatools – R package for chemometrics. Chemometrics and Intelligent Laboratory Systems, 2020; 198: 103937. https://www.doi.org/10.1016/j.chemolab.2020.103937
20. R Core Team R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2023. https://www.R-project.org
21. Milborrow S. Derived from mda:mars by T. Hastie and R. Tibshirani. earth: Multivariate Adaptive Regression Splines, 2011. R package. http://CRAN.R-project.org/package=earth

22. Liland K, Mevik B, Wehrens R. _pls: Partial Least Squares and Principal Component Regression_. R package version 2.8–3. 2023. https://CRAN.R-project.org/package=pls
23. Hastie T., Tibshirani R., Friedman J. H. The elements of statistical learning : Data mining, inference, and prediction, Second Edition. Springer-Verlag New York Inc. 2009.
24. Lee S., Sarkar S., Park Y., Yang J., Kweon G. Feasibility study for an optical sensing system for hardy kiwi (*Actinidia arguta*) sugar content estimation. Journal of Agriculture & Life Science, 2019; 53(3): 147–157. https://www.doi.org/10.14397/jals.2019.53.3.147
25. Mumford A., Abrahamsson Z., Hale I. Predicting soluble solids concentration of 'Geneva 3' kiwiberries using near infrared spectroscopy. HortTechnology, 2024; 34(2): 172–180. https://www.doi.org/10.21273/HORTTECH05316-23
26. Galindo-Prieto B., Eriksson L., Trygg J. Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). Journal of Chemometrics, 2014; 28(8): 623–632. https://www.doi.org/10.1002/cem.2627
27. Eriksson L., Johansson E., Kettaneh-Wold N., Wold S. Multi- and Megavariate Data Analysis. Principles and Applications. Umetrics Academy, Umeå, Sweden. 2002.
28. Xu L., Chen Y., Wang X., Chen H., Tang Z., Shi X., Chen X., Wang Y., Kang Z., Zou Z., Huang P., He Y., Yang N., Zhao Y. Non-destructive detection of kiwifruit soluble solid content based on hyperspectral and fluorescence spectral imaging. Frontiers in Plant Science, 2023; 13: 1075929. https://www.doi.org/10.3389/fpls.2022.1075929
29. Kim J.G., Park Y., Shin M.H., Muneer S., Lerud R., Michelson C., Il Kang D., Min J.H., Chamidha Kumarihami H.M.P. Application of NIR-Spectroscopy to predict the harvesting maturity, fruit ripening and storage ability of Ca-chitosan treated baby kiwifruit. Journal of Stored Products and Postharvest Research, 2018; 9(4): 44–53. https://www.doi.org/10.5897/JSPPR2018.0257
30. Sarkar S., Basak J.K., Moon B.E., Kim H.T. A comparative study of PLSR and SVM-R with various preprocessing techniques for the quantitative determination of soluble solids content of hardy kiwi fruit by a portable Vis/NIR spectrometer. Foods, 2020; 9(8): 1078. https://www.doi.org/10.3390/foods9081078
31. Wang H., Peng J., Xie C., Bao Y., He Y. Fruit quality evaluation using spectroscopy technology: A review. Sensors, 2015; 15: 11889–11927. https://www.doi.org/10.3390/s150511889
32. Benelli A., Cevoli C., Fabbri A., Ragni L. Ripeness evaluation of kiwifruit by hyperspectral imaging. Biosystems Engineering, 2022; 223(B): 42–52. https://www.doi.org/10.1016/j.biosystemseng.2021.08.009
33. Sharma S., K.C. Sumesh, Sirisomboon P. Rapid ripening stage classification and dry matter prediction of durian pulp using a pushbroom near infrared hyperspectral imaging system. Measurement, 2022; 189: 110464. https://www.doi.org/10.1016/j.measurement.2021.110464
34. Tian P., Meng Q., Wu Z., Lin J., Huang X., Zhu H., Zhou X., Qiu Z., Huang Y., Li Y. Detection of mango soluble solid content using hyperspectral imaging technology. Infrared Physics & Technology, 2023; 129: 104576. https://www.doi.org/10.1016/j.infrared.2023.104576
35. Moghimi A., Aghkhani M.H., Sazgarnia A., Sarmad M. Vis/NIR spectroscopy and chemometrics for the prediction of soluble solids content and acidity (pH) of kiwifruit. Biosystems Engineering, 2010; 106(3): 295–302. https://www.doi.org/10.1016/j.biosystemseng.2010.04.002
36. Lee J.S., Kim S.-C., Seong K.C., Kim C.-H., Um Y.C., Lee S.-K. Quality prediction of kiwifruit based on near infrared spectroscopy. Korean Journal of Horticultural Science and Technology, 2012; 30(6), 709–717. https://www.doi.org/10.7235/hort.2012.12139
37. Zhu H., Chu B., Fan Y., Tao X., Yin W., He Y. Hyperspectral Imaging for Predicting the Internal Quality of Kiwifruits Based on Variable Selection Algorithms and Chemometric Models. Scientific Reports. 2017, 7: 7845. https://www.doi.org/10.1038/s41598-017-08509-6
38. Simpson E. H. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, Series B, 1951; 13: 238–241.