Proceeding Paper

Assessment of Tomato Quality through Near-Infrared Spectroscopy—Advantages, Limitations, and Integration with Multivariate Analysis Techniques †

Mima Todorova 1, Petya Veleva 1-*, Stefka Atanassova 1, Tsvetelina Georgieva 2, Miroslav Vasilev 3 and Zlatin Zlatev 3

1 Faculty of Agriculture, Trakia University–Student’s Campus, 6000 Stara Zagora, Bulgaria; mima.todorova@trakia-uni.bg (M.T.); stefka.atanassova@trakia-uni.bg (S.A.)
2 Department of Automatics and Mechatronics, University of Ruse, Studentska 8, 7017 Ruse, Bulgaria; cgeorgieva@uni-ruse.bg
3 Faculty of Technics and Technologies, Trakia University, 38 Graf Ignatiev Str., 8602 Yambol, Bulgaria; miroslav.vasilev@trakia-uni.bg (M.V.); zlatin.zlatev@trakia-uni.bg (Z.Z.)
* Correspondence: petya.veleva@trakia-uni.bg; Tel.: +359-896601537
† Presented at the International Conference on Electronics, Engineering Physics and Earth Science (EEPES'24), Kavala, Greece, 19–21 June 2024.

Abstract: This study conducted a systematic analysis and investigation of near-infrared (NIR) spectroscopy for assessing tomato fruits’ quality. This was combined with the integration of multivariate techniques for analysis to improve existing methods for nondestructive quality assessment of tomato fruits. Variations in tomato characteristics were revealed by selecting spectral indices, leading to changes in reflection spectra at different wavelengths. To achieve more accurate predictions, the use of vectors containing various spectral indices became necessary. This study found that predicting key tomato characteristics relies more on the informativeness and predictive ability of vectors derived from spectral indices than on the method used to reduce data volume. Regression models were developed to automate the prediction of six tomato characteristics based on the spectral data. The developed models for vitamin C content, titratable organic acids, and lycopene showed the highest predictive efficiency (over 90%) among the investigated features. These results highlight the potential of regression models to be integrated into automated systems for assessing the quality of tomato fruits, providing sufficiently accurate information to enhance quality control processes.

Keywords: tomato quality; nondestructive evaluation; multivariate analysis; lycopene; sustainability; spectral imaging

1. Introduction

Tomatoes (Solanum lycopersicum) belong to the Solanaceae family and are the most widely cultivated vegetable crop globally.

Consuming tomatoes offers health benefits due to their low fat and calorie content. They are rich in fiber, natural antioxidants, vitamins, carotenoids, and phenols. Scientific evidence supports the protective role of tomato consumption in preventing chronic degenerative diseases such as cancer, cardiovascular issues, and neurodegenerative pathologies [1].

The taste and appearance of tomatoes significantly influence consumer preferences. These characteristics depend on the content of soluble solids, sugars, and organic acids. Soluble solids are the collective term for soluble compounds in tomato fruits that directly impact taste [2,3]. The sweet taste of the vegetable primarily results from fructose and glucose, while acidity arises from citric and malic acids [4]. Consequently, the sugar and acid content often serve as essential indicators for assessing tomato flavor.

Measuring the quality characteristics of tomatoes is crucial to ensure they meet consumer expectations regarding taste, appearance, and nutritional value. Quality assessment
plays a vital role throughout the tomato production and supply chain. It helps growers optimize cultivation practices, enables distributors and retailers to maintain suitable storage conditions, and ensures compliance with food safety requirements.

Near-infrared (NIR) spectroscopy is one of the methods for nondestructive quality assessment of tomato fruits, due to its noninvasive nature and rapid analysis. By evaluating the interaction of light with the tomato surface, this method provides fast and reliable measurements without damaging the fruits, making it suitable for high-throughput analyses in the food and flavor industry. Additionally, NIR spectroscopy does not require chemicals, making it more environmentally friendly and cost-effective compared to other techniques. Through multivariate analysis techniques such as principal component analysis and partial least squares regression, NIR reflection spectra can be used to assess various quality parameters simultaneously, including ripeness, firmness, sugar content, and acidity, facilitating an overall evaluation of vegetable quality. As a drawback, successfully implementing the method requires instrument calibration specific to tomato varieties and the measured parameters. External conditions such as temperature, lighting, and humidity can also influence measurement accuracy, necessitating careful management of experimental conditions. Furthermore, NIR spectroscopy has a limited penetration depth, which may restrict the ability to measure internal parameters in specific tomato varieties. Despite these limitations, NIR spectroscopy remains a reliable method when used in conjunction with other techniques to provide a comprehensive assessment of tomato quality.

It has been proven that Vis–NIR spectroscopy allows for tracking the biochemical changes occurring in tomatoes during ripening, their optimal harvest time, and assessing their ripeness after harvest [5,6].

In studies by Clément et al. [7], Bunghez et al. [8], and Szuvandzsiev et al. [9], the assessment of tomato quality is based on their lycopene content. By using nondestructive quality measurement through Vis–NIR spectroscopy, information about lycopene content and associated color changes in tomatoes can be obtained. Vis–NIR spectroscopy is an effective method for determining specific quality parameters but it is not applicable for predicting other parameters such as pH and soluble solids in tomatoes due to sample homogeneity. The use of portable spectrometers for assessing the quality of tomato puree provides rapid and real-time evaluation of soluble solids, lycopene, and polyphenols. The method’s accuracy is influenced by the calibration period of the measuring devices. Moreover, UV–Vis spectrophotometry and FTIR spectroscopy are suitable methods for quantitatively determining lycopene in tomato powder samples. However, FTIR spectroscopy requires prior sample preparation.

In other studies by Saad et al. [10], Governici et al. [11], and Dobrin et al. [12], combining Vis–NIR spectroscopy with partial least squares regression (PLSR) is proposed for tomato quality assessment. The aim is to track the physicochemical parameters during tomato storage and accurately predict the soluble solids content, degree of ripeness, and, to a lesser extent, the lycopene content.

Integrating various techniques for tomato quality assessment enhances the efficiency of spectral imaging for comprehensive evaluation, necessitating additional methods for data processing. Despite the sufficient accuracy in measuring tomato quality using Vis–NIR spectroscopy, the possibility of predicting important characteristics for these vegetables has not been thoroughly explored, including the generalization of the obtained regression models.

Najjar et al. [13] and Duckena et al. [14] investigated the potential of Vis–NIR spectroscopy for the analysis of the internal quality characteristics of tomato fruits. They have developed models for predicting taste index, as well as the content of lycopene, flavonoids, β-carotene, total phenols, and dry matter in intact vegetables based on Vis–NIR reflectance spectra. Their research demonstrates high correlation coefficients for determining lycopene content and dry matter. The authors emphasize the need for additional studies to optimize the application of Vis/NIR spectroscopy under various production conditions. Improving the robustness of models for predicting additional parameters is also necessary to ensure reliable tomato quality control.
Radzevičius et al. [15] investigated the potential of NIR spectroscopy in assessing the maturity of tomatoes. A correlation between the data obtained through NIR and physicochemical analyses was determined. The authors commented on how the quality parameters of tomato fruits change during ripening. The discovered correlation for the three studied parameters (dry matter, soluble solids, and fruit firmness) is in the range of 0.82–0.96. The study presented important data regarding the potential of NIR spectroscopy for assessing the maturity characteristics of tomatoes, but this investigation requires a more comprehensive examination of the various physicochemical parameters and possible factors affecting the quality of tomatoes.

In another study, the applicability of NIR spectroscopy was tested in field conditions for predicting dry matter values in tomatoes grown in five regions of Brazil. The regression models obtained through PLS regression showed a significant accuracy of over 90%. It was also highlighted that for the commercialization of fresh fruits, improvement in the models and inclusion of different seasons and varieties, as well as a larger number of samples, are necessary. Moreover, the study was focused on a single quality parameter and may not provide a complete picture of the tomatoes’ suitability and nutritional benefits [16]. A team of researchers, applying NIR spectroscopy, investigated the possibility of assessing the level of maturity and predicting the textural properties of “Momotaro” tomatoes. The PCA model with mean normalization demonstrated high distinguishability, classifying ripe green, pink, and red tomatoes with over 90% accuracy. For the assessment of the soluble solids content in fresh tomatoes, the obtained values of the coefficient of determination were 0.8, while the prediction of alcohol-insoluble solids in tomatoes was determined as unsuccessful by the authors [17].

From the analysis of the available literature, it can be summarized that assessing the quality of tomatoes is important from the perspective of producers, distributors, and consumers. Vis–NIR spectroscopy is a suitable, fast, and nondestructive method for conducting such quality assessments. However, it has limitations related to the surface sensitivity of tomatoes and the need for calibration of measurement devices. NIR spectroscopy, which also requires calibration, offers several advantages over Vis spectroscopy for tomato quality assessment. Firstly, NIR spectroscopy allows analysis across a broader spectrum of light, enabling the assessment of a greater number of quality parameters in tomatoes when seeking a comprehensive evaluation. Additionally, NIR spectra penetrate deeper into the studied structure, allowing more detailed measurements of internal characteristics related to various nutrients. Due to the specific wavelengths in the NIR range, this technique typically provides greater precision in determining water content, sugars, acids, and other components. Furthermore, NIR spectroscopy enables rapid and efficient scanning of large samples, speeding up the analysis process and allowing for the processing of larger volumes of data. Despite the satisfactory accuracy of Vis–NIR spectroscopy, the method’s potential for predicting important tomato characteristics and generalizing obtained regression models has not been fully explored. Additionally, in numerous studies using the Vis–NIR spectral range, it is observed that it covers only a small portion of the entire NIR spectrum, limiting the informative wavelengths and indices relevant to specific tomato varieties. This fact leads to the need for additional research and the development of models that incorporate a broader NIR spectrum to ensure greater informativeness of measurements and better prediction of individual characteristics. Additional efforts to identify suitable indices from predefined informative wavelengths across the entire NIR spectrum can lead to more accurate and generalizable models for predicting tomato quality, which is essential for the agricultural industry. In the conducted review of studies involving the application of NIR spectroscopy for tomato quality assessment, it is evident that most of them focus on one or two evaluation parameters. Integrating techniques for multivariate analysis and encompassing a wider range of parameters could enhance the efficiency of quality assessment. Further research is necessary to overcome limitations associated with using NIR spectra for evaluating the quality of tomato fruits, considering different varieties and cultivation methods.
The purpose of this study is to explore the potential of NIR spectroscopy as a non-destructive method for assessing tomato quality. To achieve this goal, it is necessary to investigate the advantages and limitations of NIR spectroscopy in evaluating various quality parameters such as dry matter, vitamin C, titratable organic acids, total dyes, lycopene, and beta-carotene. Additionally, the integration of multivariate analysis techniques needs to be studied to enhance the efficiency of NIR spectroscopy in the comprehensive assessment of tomato quality.

The results of this study can contribute to the development of more effective and reliable methods for assessing tomato quality, thereby increasing consumer satisfaction and meeting food safety requirements during production and distribution.

2. Material and Methods

Three varieties of tomatoes were used for the examination: Manusa, Mirsini, and Red Bounty, grown under greenhouse conditions. The fruits were grown under natural sunlight, without additional heating during the summer season. They were harvested at peak ripeness, ensuring their optimal taste and color.

The main characteristics of the tomato fruits were determined using the methodologies presented in [18–20] as follows:

- Dry matter (%)—by drying at 70 °C until constant mass.
- Titratable organic acids (%)—by titration of tomato juice or extract with 0.1 N NaOH (sodium hydroxide), using phenolphthalein as an indicator until reaching neutral pH. The consumed milliliters of 0.1 N NaOH during titration determine the acidity, expressed as a percentage of malic acid.
- Vitamin C (mg %) by Tilman’s reaction with 2,6-ichlorophenolindophenol.
- Total pigments (mg %)—by extracting lycopene and beta-carotene from tomatoes with hexane, acetone, and ethanol. Lycopene (mg%) is a red pigment and a powerful antioxidant, found in tomatoes. High-performance liquid chromatography (HPLC) is the gold standard for measuring lycopene. This involves extracting lycopene from the tomato sample and analyzing it with HPLC to determine its concentration.
- Beta-carotene (mg %) is another carotenoid pigment present in tomatoes. Similar to lycopene, this component is quantitatively determined by HPLC after extraction from the tomato sample.

In the paper, the characteristics are denoted by alphanumeric symbols according to the following scheme:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Dry matter, (%)</td>
</tr>
<tr>
<td>C2</td>
<td>Vitamin C (mg %)</td>
</tr>
<tr>
<td>C3</td>
<td>Titratable organic acids (%)</td>
</tr>
<tr>
<td>C4</td>
<td>Total pigments (mg %)</td>
</tr>
<tr>
<td>C5</td>
<td>Lycopene (mg %)</td>
</tr>
<tr>
<td>C6</td>
<td>Beta-carotene (mg %)</td>
</tr>
</tbody>
</table>

The chemical analysis was followed by spectrophotometric measurement of each sample. NIR reflectance spectra of tomato fruits were obtained. Each sample was measured at five independent locations on the object’s surface. The NIRQuest512 spectrophotometer (Ocean Optics Inc., Orlando, FL, USA) was used in the spectral range of 900–1700 nm, with a Hamamatsu G9204-512 InGaAs detector (Hamamatsu Photonics K. K., Japan). Its optical resolution is 3.1 nm.

The RRReliefF method [21] was used to select the informative spectral indices. RRReliefF is an algorithm for selecting informative features used in machine learning. It evaluates the relevance and importance of features by comparing their values on close instances, focusing on distinguishing instances from different classes. Spectral indices defined by Ju et al. [22] and Mendiguren et al. [23] were used in this work. An advantage of these indices is their adaptability, as they are not limited to fixed spectral wavelengths. Six
informative wavelengths ($\lambda$, nm) specific to the product under study need to be selected. Subsequently, the calculated indices can serve as input data for classification, regression, and clustering tasks. These indices are shown below (Equations (1)–(15)):

\[ SI_1 = \frac{R_{\lambda_2} - R_{\lambda_1}}{R_{\lambda_2} + R_{\lambda_1}} \]  
\[ SI_2 = \frac{R_{\lambda_1} - R_{\lambda_2}}{R_{\lambda_1} + R_{\lambda_2}} \]  
\[ SI_3 = \frac{R_{\lambda_1}}{R_{\lambda_2}} \]  
\[ SI_4 = R_{\lambda_1} - R_{\lambda_2} \]  
\[ SI_5 = 2.5 \frac{R_{\lambda_1} - R_{\lambda_2}}{R_{\lambda_2} + 6R_{\lambda_1} - 7.5R_{\lambda_3}} \]  
\[ SI_6 = \frac{R_{\lambda_2} - R_{\lambda_5}}{R_{\lambda_2} + R_{\lambda_5}} \]  
\[ SI_7 = \frac{R_{\lambda_2} - R_{\lambda_6}}{R_{\lambda_2} + R_{\lambda_6}} \]  
\[ SI_8 = \frac{R_{\lambda_2}}{R_{\lambda_5}} \]  
\[ SI_9 = \frac{R_{\lambda_4} - R_{\lambda_1}}{R_{\lambda_4} + R_{\lambda_1} - R_{\lambda_3}} \]  
\[ SI_{10} = R_{\lambda_1} \]  
\[ SI_{11} = R_{\lambda_2} \]  
\[ SI_{12} = R_{\lambda_3} \]  
\[ SI_{13} = R_{\lambda_4} \]  
\[ SI_{14} = R_{\lambda_5} \]  
\[ SI_{15} = R_{\lambda_6} \].  

The preliminary assessment of forecasting feasibility was conducted using PCR and PLSR methods [24]. PCR employs principal component analysis to reduce multicollinearity in regression. PLSR identifies latent variables explaining variations in both predictors and responses in regression. A second-order regression model was employed [25], describing the relationship between independent and dependent variables as follows:

\[ z = b_0 + b_1x + b_2y + b_3x^2 + b_4xy + b_5y^2, \]  

where $z$ is the dependent variable; the independent variables are $x$ and $y$; model coefficients are denoted as $b$.

Model assessment includes metrics such as coefficient of determination ($R^2$), standard error (SE), $p$-value, Fisher’s test (F), and residual analysis. Validation of regression models was performed through the relationship between actual and predicted values of tomato characteristics. Validation was conducted using the coefficient of determination ($R^2$), mean
squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and standard error (SE). These errors are calculated using the following formulas:

\[ \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_m - y_p)^2; \quad \text{(17)} \]

\[ \text{RMSE} = \sqrt{\text{MSE}}; \quad \text{(18)} \]

\[ \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_m - y_p|; \quad \text{(19)} \]

\[ \text{SE} = \sqrt{\frac{\sum_{i=1}^{n} (y_m - y_p)^2}{n - 2}}, \quad \text{(20)} \]

where \( n \) is the number of data; \( y_m \) is the actual measured values; \( y_p \) is the predicted values.

The choice of an error metric depends on the goals and characteristics of the regression problem being solved. Absolute error values are used when the robustness and interpretability of models are crucial.

The validation used tomato data that were not included in the creation of the regression models. The processing of the experimental data was performed in the software Matlab 2017b (The Mathworks Inc., Natick, MA, USA). All data were processed at a significance level of \( \alpha = 0.05 \).

### 3. Results

Figure 1 depicts the spectral characteristics of tomatoes from the three varieties (Manusa, Mirsini, and Red Bounty 27 5a), within the range of 900–1700 nm. Across the entire spectral range, there is an overlap in characteristics for the three tomato varieties.

![Figure 1](image1)

**Figure 1.** Spectral characteristics of tomatoes from three varieties, as follows: (a) Manusa, (b) Mirsini, and (c) Red Bounty 27 5a.

Selection of informative wavelengths of the spectral waves was made. Figure 2 illustrates graphs of the weighting coefficients of the respective wavelengths for each of the technological characteristics of the tomatoes.

![Figure 2](image2)

**Figure 2.** Selection of informative spectral wavelengths.
The six informative wavelengths, a result of the selection by the RReliefF method, are presented in Table 1. The wavelengths are sorted according to their informativeness, determined by the corresponding weighting coefficients.

**Table 1. Informative spectral wavelengths.**

<table>
<thead>
<tr>
<th>Wavelength</th>
<th>( \lambda_1 )</th>
<th>( \lambda_2 )</th>
<th>( \lambda_3 )</th>
<th>( \lambda_4 )</th>
<th>( \lambda_5 )</th>
<th>( \lambda_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value, nm</td>
<td>1430.61</td>
<td>1448.11</td>
<td>1525.83</td>
<td>1654.76</td>
<td>915.53</td>
<td>1270.5</td>
</tr>
</tbody>
</table>

Fifteen spectral indices were computed, using the selected wavelengths of the spectra. In Table 2, data on the informativeness of spectral indices are shown depending on the selected tomato characteristics. As seen from the selection of spectral indices and their informativeness, changes in each tomato characteristic lead to changes in reflection spectra at different wavelengths. Hence, forecasting these tomato characteristics requires the use of different spectral indices. The smallest number of informative indices is calculated for changes in lycopene content (C5), while the highest number of indices is selected for beta-carotene content (C6). Those spectral indices that have a weighting factor greater than 0.6 are selected.

**Table 2. Informative spectral indices.**

<table>
<thead>
<tr>
<th>No.</th>
<th>Characteristic</th>
<th>SI1</th>
<th>SI2</th>
<th>SI3</th>
<th>SI4</th>
<th>SI5</th>
<th>SI6</th>
<th>SI7</th>
<th>SI8</th>
<th>SI9</th>
<th>SI10</th>
<th>SI11</th>
<th>SI12</th>
<th>SI13</th>
<th>SI14</th>
<th>SI15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C1</td>
<td>0.73</td>
<td>0.73</td>
<td>0.16</td>
<td>0.66</td>
<td>1.00</td>
<td>0.15</td>
<td>0.15</td>
<td>0.71</td>
<td>0.04</td>
<td>0.05</td>
<td>0.00</td>
<td>0.23</td>
<td>0.64</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>C2</td>
<td>0.70</td>
<td>0.70</td>
<td>0.01</td>
<td>0.63</td>
<td>0.52</td>
<td>0.65</td>
<td>1.00</td>
<td>0.42</td>
<td>0.00</td>
<td>0.48</td>
<td>0.47</td>
<td>0.53</td>
<td>0.60</td>
<td>0.88</td>
<td>0.61</td>
</tr>
<tr>
<td>3</td>
<td>C3</td>
<td>0.93</td>
<td>0.93</td>
<td>0.69</td>
<td>1.00</td>
<td>0.93</td>
<td>0.63</td>
<td>0.21</td>
<td>0.00</td>
<td>0.10</td>
<td>0.19</td>
<td>0.19</td>
<td>0.20</td>
<td>0.24</td>
<td>0.63</td>
<td>0.21</td>
</tr>
<tr>
<td>4</td>
<td>C4</td>
<td>0.27</td>
<td>0.27</td>
<td>0.78</td>
<td>0.28</td>
<td>0.62</td>
<td>0.00</td>
<td>0.26</td>
<td>0.04</td>
<td>1.00</td>
<td>0.67</td>
<td>0.67</td>
<td>0.63</td>
<td>0.26</td>
<td>0.07</td>
<td>0.20</td>
</tr>
<tr>
<td>5</td>
<td>C5</td>
<td>0.12</td>
<td>0.12</td>
<td>0.64</td>
<td>0.17</td>
<td>0.66</td>
<td>0.00</td>
<td>0.62</td>
<td>0.11</td>
<td>1.00</td>
<td>0.32</td>
<td>0.62</td>
<td>0.32</td>
<td>0.28</td>
<td>0.15</td>
<td>0.21</td>
</tr>
<tr>
<td>6</td>
<td>C6</td>
<td>0.37</td>
<td>0.37</td>
<td>0.32</td>
<td>0.64</td>
<td>0.36</td>
<td>0.00</td>
<td>0.67</td>
<td>0.21</td>
<td>1.00</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
<td>0.63</td>
<td>0.80</td>
<td>0.68</td>
</tr>
</tbody>
</table>

The following vectors of informative spectral indices were selected:

\[
\text{FVC1} = \text{SI1} \text{ SI2} \text{ SI4} \text{ SI5} \text{ SI9} \text{ SI14};
\]
\[
\text{FVC2} = \text{SI1} \text{ SI2} \text{ SI4} \text{ SI6} \text{ SI7} \text{ SI13} \text{ SI14} \text{ SI15};
\]
\[
\text{FVC3} = \text{SI1} \text{ SI2} \text{ SI3} \text{ SI4} \text{ SI5} \text{ SI6} \text{ SI14};
\]
\[
\text{FVC4} = \text{SI3} \text{ SI5} \text{ SI9} \text{ SI10} \text{ SI11} \text{ SI12};
\]
\[
\text{FVC5} = \text{SI3} \text{ SI5} \text{ SI7} \text{ SI9} \text{ SI11};
\]
\[
\text{FVC6} = \text{SI4} \text{ SI7} \text{ SI9} \text{ SI10} \text{ SI11} \text{ SI12} \text{ SI13} \text{ SI14} \text{ SI15}. \]

A preliminary analysis was conducted on the possibility of forecasting key characteristics of tomatoes using the PCR and PLSR methods. The necessary number of principal components and latent variables was determined to describe over 95% of the variance in the data vectors containing spectral indices from each of the six analyzed tomato characteristics. It was found that two principal components and two latent variables are necessary to describe these data in all cases.

The results of this analysis are shown in Table 3. It is evident that the forecasting of the main characteristics of tomatoes does not depend on the method used to reduce the volume of data, but, rather, on the informativeness of the vectors of indices used and their
predictive ability. In this case, there is a preference for reduction via principal components, as the regression method used shows lower errors and a slightly higher coefficient of determination compared to PLSR.

Table 3. Results from PCR and PLSR.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Criteria</th>
<th>PCR</th>
<th>PLSR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SE</td>
<td>$R^2$</td>
<td>RMSE</td>
</tr>
<tr>
<td>C1</td>
<td>0.06</td>
<td>0.87</td>
<td>0.02</td>
</tr>
<tr>
<td>C2</td>
<td>1.74</td>
<td>0.96</td>
<td>0.11</td>
</tr>
<tr>
<td>C3</td>
<td>0.01</td>
<td>0.92</td>
<td>0.01</td>
</tr>
<tr>
<td>C4</td>
<td>1.02</td>
<td>0.94</td>
<td>0.08</td>
</tr>
<tr>
<td>C5</td>
<td>0.29</td>
<td>0.96</td>
<td>0.05</td>
</tr>
<tr>
<td>C6</td>
<td>0.19</td>
<td>0.74</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Nonsignificant coefficients with a $p$-value > 0.05 were removed from the main model. The following regression models were obtained:

\[
C_1 = f(PC_1, PC_2) \quad C_1 = 3.93 + 3.61PC_1 - 13.9PC_1^2 - 7.7PC_2^2 + 25.5PC_1PC_2 \quad (27)
\]

\[
C_2 = f(PC_1, PC_2) \quad C_2 = 19.5 + 34.2PC_1 + 19.7PC_2 - 89.2PC_1^2 + 263PC_1PC_2 \quad (28)
\]

\[
C_3 = f(PC_1, PC_2) \quad C_3 = 0.33 + 0.73PC_2 + 1.23PC_1^2 - 3.13PC_1PC_2 \quad (29)
\]

\[
C_4 = f(PC_1, PC_2) \quad C_4 = 2.5 + 11.7PC_2 - 20.5PC_1^2 - 59.8PC_2^2 + 86.3PC_1PC_2 \quad (30)
\]

\[
C_5 = f(PC_1, PC_2) \quad C_5 = 1.95 + 8.8PC_1 + 4.7PC_2 - 37.2PC_2^2 - 56.4PC_1PC_2 \quad (31)
\]

\[
C_6 = f(PC_1, PC_2) \quad C_6 = 0.15 + 2.6PC_2 + 195.1PC_1^2 + 291.3PC_2^2 - 473PC_1PC_2 \quad (32)
\]

The values of the criteria for the regression models assessment are indicated in Table 4. Each model represents a function of the principal components (PC1 and PC2).

Table 4. Assessment of regression models.

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
<th>$F$</th>
<th>F-Critical</th>
<th>$p$-Value</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1 = f(PC1,PC2)</td>
<td>0.95</td>
<td>F(4, 145) = 671.7</td>
<td>2.43</td>
<td>0.00</td>
<td>0.06</td>
</tr>
<tr>
<td>C2 = f(PC1,PC2)</td>
<td>0.98</td>
<td>F(4, 145) = 1866</td>
<td>2.43</td>
<td>0.00</td>
<td>0.57</td>
</tr>
<tr>
<td>C3 = f(PC1,PC2)</td>
<td>0.95</td>
<td>F(3, 146) = 944.95</td>
<td>2.67</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>C4 = f(PC1,PC2)</td>
<td>0.98</td>
<td>F(4, 145) = 1460</td>
<td>2.43</td>
<td>0.00</td>
<td>0.14</td>
</tr>
<tr>
<td>C5 = f(PC1,PC2)</td>
<td>0.98</td>
<td>F(4, 145) = 1394.2</td>
<td>2.43</td>
<td>0.00</td>
<td>0.15</td>
</tr>
<tr>
<td>C6 = f(PC1,PC2)</td>
<td>0.81</td>
<td>F(4, 145) = 150.24</td>
<td>2.43</td>
<td>0.00</td>
<td>0.17</td>
</tr>
</tbody>
</table>

The coefficient of determination ($R^2$) indicates how well each model explains the variance in the dependent variable. According to Fisher’s criterion, the computed value of $F$ is much greater than the critical F-critical value. Based on this criterion, it can be inferred that the obtained regression models have sufficient accuracy.

In Figure 3, the most informative models for predicting key characteristics of tomatoes are presented in general.
Figure 3. Regression models—general view for (a) \( C_2 = f(PC_1,PC_2) \); (b) \( C_4 = f(PC_1,PC_2) \); (c) \( C_5 = f(PC_1,PC_2) \).

Figure 4. Residual analysis of regression models for (a) \( C_2 = f(PC_1,PC_2) \); (b) \( C_4 = f(PC_1,PC_2) \); (c) \( C_5 = f(PC_1,PC_2) \).

Figure 5. Validation of regression models for (a) \( C_2 = f(PC_1,PC_2) \); (b) \( C_4 = f(PC_1,PC_2) \); (c) \( C_5 = f(PC_1,PC_2) \).

The validation process involves comparing the predictive ability of the regression models under identical measurement conditions. Figure 5 depicts the distribution of measured and predicted values of tomato characteristics relative to the appropriate regression line.

The errors in validating the obtained regression models for the key characteristics of tomatoes are presented in Table 5. The obtained models show low errors according to the used criteria, indicating strong predictive ability. This was evidenced by high values of \( R^2 \), which signify the proportion of variance explained by the model.
Table 5. Errors between real and predicted values of tomato characteristics.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Criterion</th>
<th>MSE</th>
<th>RMSE</th>
<th>MAE</th>
<th>SE</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td></td>
<td>0.86</td>
<td>0.93</td>
<td>2.85</td>
<td>2.99</td>
<td>0.83</td>
</tr>
<tr>
<td>C4</td>
<td></td>
<td>0.88</td>
<td>0.94</td>
<td>2.88</td>
<td>3.04</td>
<td>0.84</td>
</tr>
<tr>
<td>C5</td>
<td></td>
<td>0.88</td>
<td>0.94</td>
<td>2.89</td>
<td>3.07</td>
<td>0.84</td>
</tr>
</tbody>
</table>

The results from validating the regression models for predicting tomato characteristics across the three varieties (C2, C4, and C5) consistently demonstrate stable performance. In all features, the models exhibit low error values, indicating a sufficiently high level of accuracy. The mean squared error (MSE) and the root mean squared error (RMSE) are adequately low, ranging from 0.86 to 0.88 for MSE and 0.93 to 0.94 for RMSE. These metrics suggest that the models have a minimal average discrepancy between actual and predicted values, which is a marker of their effectiveness.

The mean absolute error (MAE), ranging from 2.85 to 2.89, further confirms the models’ strong predictive ability, although they are slightly higher than the error metrics based on squared differences. Additionally, the standard error (SE) values show a tendency toward increase compared to MAE, indicating some variability in prediction errors. However, they remain relatively low, ranging from 2.99 to 3.07.

The high values of the coefficient of determination (R²), ranging from 0.83 to 0.84, imply that a significant portion (83–84%) of the variation in actual tomato characteristics could be explained by the regression models.

4. Discussion

As a result of the measurements, calculations, and analyses conducted, the results of Alsina et al. [26] have been improved. Despite these authors applying spectral indices specifically developed for predicting lycopene and carotenoids, they achieved an accuracy of 73–87% in predicting these tomato characteristics. In the present study, after selecting sufficiently informative spectral indices and creating vectors from them, the accuracy in predicting the mentioned characteristics exceeds 90%.

Bunghez et al. [8] achieved over 90% accuracy in predicting lycopene in tomatoes but using less accessible devices characterized by their high cost. Such devices include FTIR spectrophotometers and HPLC analysis. In the current study, similar accuracy is achieved using data in the NIR spectral range, through advanced methods of data analysis.

Using spectral characteristics in the NIR range, Radzvevičius et al. [15] obtained a predictive model with $R^2 = 0.81$ for dry matter for a single tomato variety, whereas in the present study, it is 0.96.

The recommendations of Najjar et al. [13] and Duckena et al. [14] concerning the necessity of improving the robustness of models for predicting additional parameters to ensure reliability in controlling the quality of tomatoes were implemented.

5. Conclusions

This study complements and enhances the known methods for nondestructive assessment of tomato quality through systematic analysis and investigation of NIR spectroscopy for assessing tomato quality and integrating multivariate analysis techniques.

From the selection of spectral indices, it was found that changes in each tomato characteristic lead to variations in reflection spectra at different wavelengths. Therefore, vectors containing different spectral indices should be used to forecast these vegetable characteristics.

It was demonstrated that predicting the main characteristics of tomatoes depends less on the method used to reduce the volume of data and more on the informativeness of the indices vectors used and their predictive ability.

Regression models for the automated prediction of six tomato characteristics based on their spectral features were developed.
It was found that models for vitamin C, titratable organic acids, and lycopene have the highest predictive capability compared to the other investigated tomato characteristics. The regression models can be used in creating automated systems for assessing tomato quality.

In conclusion, it is noted that the continuous development of noncontact analysis techniques requires careful consideration and taking into account new methods and technologies that have the potential to improve the accuracy and reliability of predictive models for tomato quality improvement. It is also important to recognize that producers, traders, and consumers have different requirements for tomato quality. This requires flexibility and adaptability of prediction models to meet the specific needs and preferences of different stakeholders.

Author Contributions: Conceptualization, Z.Z. and P.V.; methodology, Z.Z. and M.V.; software, Z.Z.; validation, S.A. and P.V.; formal analysis, P.V.; investigation, P.V.; resources, M.T., P.V., and S.A.; data curation, P.V., M.V., and Z.Z.; writing—original draft preparation, Z.Z. and P.V.; writing—review and editing, P.V.; visualization, Z.Z. and P.V.; supervision, T.G. and S.A.; project administration, S.A.; funding acquisition, T.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created.

Acknowledgments: This work was supported by the Bulgarian Ministry of Education and Science under the National Research Program “Smart crop production” approved by the Decision of the Ministry Council No. 866/26.11.2020.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Friedman, M. Anticarcinogenic, cardioprotective, and other health benefits of tomato compounds lycopene, α-tomatine, and tomatidine in pure form and in fresh and processed tomatoes. J. Agric. Food Chem. 2013, 61, 9534–9550. [CrossRef] [PubMed]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.