

RESEARCH ARTICLE

Open Access



# Investigation into the predictive performance of colorimetric sensor strips using RGB, CMYK, HSV, and CIELAB coupled with various data preprocessing methods: a case study on an analysis of water quality parameters

Nutthatida Phuangsaibai<sup>1</sup>, Jaroon Jakmunee<sup>1,2</sup> and Sila Kittiwachana<sup>1,2,3\*</sup> 

## Abstract

The potential use of colorimetric sensors has received significant attention due to its feasibility for use in various applications. After reacting with a sample, the image of the colorimetric sensor can be captured and converted into digital data using several different color models. The analytical data can then be processed with various chemometric methods. This research study investigated the predictive performance of calibration models established using color models commonly used in analytical chemistry including RGB, CMYK, HSV and CIELAB. A total of eight commercially available colorimetric sensors were used to determine the presence of manganese ( $\text{Mn}^{2+}$ ), copper ( $\text{Cu}^{2+}$ ), iron ( $\text{Fe}^{2+}/\text{Fe}^{3+}$ ), nitrate ( $\text{NO}_3^-$ ), phosphate ( $\text{PO}_4^{3-}$ ), sulfate ( $\text{SO}_4^{2-}$ ), as well as total hardness and pH values. As external validation tests, real water samples collected in Chiang Mai, Thailand were used. Based on the resulting data obtained using the synthetic test samples, the color that was most similar to the appearing color of the chemical sensor could offer satisfactory results. However, it was not always the case especially when the strips composed of multiple colorimetric sensors or sensor array were used. When tested with external validation, the predictive performance could be improved using appropriate data preprocessing and, in this research study, a normalization method was recommended to guarantee the accuracy of the calibration models.

**Keywords:** Colorimetric sensor, RGB, CMYK, HSV, CIELAB, Chemometrics

## Introduction

In recent years, colorimetric sensors have received substantial attention because they are relatively cost-effective and allow for chemical detection to be done without the need for complicated analytical instruments.

The detection of colorimetric sensors is based on the assumption that an analyte sample should interact with a chemoresponsive chemical sensor resulting in an intense color change. Colorimetric sensors are composed of either a single or several sensing chemicals that are coated on a supporting material. A plain paper or some hydrophobic material, such as polyethylene terephthalate film (LaGasse et al. 2014) and polyvinylidene fluoride (Kim et al. 2016), are among the common supporting materials where the sensor can be positioned as a narrow strip allowing for

\* Correspondence: [sila.k@cmu.ac.th](mailto:sila.k@cmu.ac.th)

<sup>1</sup>Department of Chemistry, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand

<sup>2</sup>Center of Excellence for Innovation in Chemistry, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

practical and easy use (Pla-Tolós et al. 2016). The color change could be simply detected using unaided eyes for the purposes of screening or semi-quantitative analysis. To improve the detection ability, the images of the colorimetric sensor strips can be captured using a digital camera or an ordinary scanner. After that, the sensing colors can be converted into digital coded values of various color models allowing the recorded images to be digitally analyzed by simple processing tools such as mobile phones, laptop computers and/or miniature computing devices based on the Raspberry Pi platform (Pla-Tolós et al. 2016; Cantrell et al. 2010).

In analytical chemistry, RGB (red, green, blue), CMYK (cyan, magenta, yellow, black), HSV (hue, saturation, value), and CIELAB are the among the most common color models used to display the color spectrum (Cantrell et al. 2010; Sharifzadeh et al. 2014; Peng et al. 2020; Ravindranath et al. 2018; Wang et al. 2020). RGB is regarded as an additive model wherein color is a combination between the three primary colors of red, green, and blue. On the other hand, CMYK is a subtractive model meaning that color can be generated by subtracting the primary colors from white light. For example, white minus red is cyan. Magenta is as a result of white minus green. HSV is another color-coding system that represents color using a conical geometric shape comprised of hue (H), saturation (S), and value (V). The parameters of H are used to indicate colors using angular dimensions. For instance, red, green, and blue are located at 0°, 120°, and 240°, respectively. The horizontal axis of the HSV model utilizes the saturation of color to indicate the grayness or the intensity of white or black, where the S value ranges from 0 (white/black) to 1 (pure color). On the vertical axis, the V parameter or lightness describes the amount of darkness in the color. A lightness value of 0 represents black and transitions toward increased lightness all the way to 1. In CIELAB or CIEL\*a\*b\*, a color is expressed using three parameters; a\*, b\*, and lightness (L). The parameter L indicates the lightness of the color in a similar way to the V parameter used in HSV. Additionally, the a\* and b\* parameters indicate the amount of green-red and blue-yellow color components, respectively (Korifi et al. 2013).

The use of these color models depicts color as a multi-dimensional model resulting in data of a multivariate nature. For example, a pH test strip composed of four pH indicators generates a total of 12 variables comprised of three digitally coded RGB parameters that have been determined from four different chemical-sensing mechanisms ( $3 \times 4$ ). In this circumstance, the analytical decision should be made based on the use of all available data in a multivariate manner. To quantitatively analyze the samples, multivariate linear regression (MLR) and partial least squares (PLS) regression are usually used in

addition to the more traditional univariate linear regression (ULR) (Brereton 2007). In previously published literature, various color models have been differently applied to solve analytical problems. For example, RGB has been used to extract the concentrations of trimethylamine (TMA) in meat-borne substances using PLS (Xiao-wei et al. 2016). Gonzalez-Miret et al. also reported on the use of CIELAB in analyzing the concentrations of mineral contents in honey based on the established MLR equation (González-Miret et al. 2005). On the other hand, HSV could be used to effectively predict the degree of deprotonation of the indicator molecule under different membrane thickness conditions (Cantrell et al. 2010). In addition, the color models can be applied to monitor the changes in fluorescent sensing for detecting alkaline phosphatase activity (Upadhyay et al. 2020a),  $\text{Cr}^{3+}$ ,  $\text{Cu}^{2+}$ , and  $\text{Hg}^{2+}$  (Upadhyay et al. 2018; Anand and Sahoo 2019). Since each of the color parameters contains a fraction of information that could be differently related to the predicted response, selecting an appropriate color model can be a very important matter that can affect the predictive performance of the prediction models.

In addition to the choice of the color models, several data preprocessing methods, such as scaling, standardization, and normalization, could also influence the predictive accuracy of the developed models. The use of different data preprocessing methods can reveal different aspects of the data (Brereton 2003). The appropriate data preprocessing method could ensure that the right trends are being studied, and a further analysis of the methods will prevent the confusion caused by the emergence of nonessential information. Using the standardization method, variations of each studied parameter can be adjusted to appear approximately on the same level so that they can be more easily compared. On the other hand, the data normalization method can be used to balance the overall variations of each sample. For example, a row scaling normalization method mathematically adjusts the summation of the parameter values to be a constant value, e.g., 1. Although the use of the data preprocessing method can dramatically affect the predictive performance of the calibration models, its potential use has not yet been much studied from an analytical perspective of this type of the data.

This research study aimed to investigate the effects of various color models on the prediction performance of calibration models that had been constructed using colorimetric sensors. A set of commercially available colorimetric sensors used to monitor some quality parameters, such as the concentrations of manganese ion ( $\text{Mn}^{2+}$ ), copper ion ( $\text{Cu}^{2+}$ ), iron ion ( $\text{Fe}^{2+}/\text{Fe}^{3+}$ ), nitrate ( $\text{NO}_3^-$ ), phosphate ( $\text{PO}_4^{3-}$ ), sulfate ( $\text{SO}_4^{2-}$ ), pH, and total hardness, were used for the purpose of demonstration. The test samples were made up of synthetic solutions and

natural waters. In addition, the data preprocessing methods that seriously affected the predictive performance of the calibration models were investigated. This research study has emphasized that the choice of color models is important to the quantitative analysis using colorimetric sensors, while the selection of the data preprocessing methods could have significant influence on the predictive performance of the calibration models.

## Experimental

### Colorimetric sensors

In this research study, the detected analytes were the chemical properties used to assess natural water quality including manganese ion ( $\text{Mn}^{2+}$ ), copper ion ( $\text{Cu}^{2+}$ ), iron ion ( $\text{Fe}^{2+}/\text{Fe}^{3+}$ ), nitrate ( $\text{NO}_3^-$ ), phosphate ( $\text{PO}_4^{3-}$ ), sulfate ( $\text{SO}_4^{2-}$ ) contents, along with total hardness and pH values. A total of eight colorimetric strips were used in the investigation. The  $\text{Fe}^{2+}/\text{Fe}^{3+}$  (REF 913 44),  $\text{NO}_3^-$  (REF 913 13),  $\text{PO}_4^{3-}$  (REF 913 20),  $\text{SO}_4^{2-}$  (REF 913 29), and total hardness (REF 912 902) test strips were purchased from Macherey-Nagel (Germany). The  $\text{Mn}^{2+}$  (REF 1.10080.0001),  $\text{Cu}^{2+}$  (REF 1.10003.0001), and pH test strips (REF 1.09535.0001) were obtained from Merck (Germany). The detection of  $\text{Mn}^{2+}$ ,  $\text{Cu}^{2+}$ ,  $\text{Fe}^{2+}/\text{Fe}^{3+}$ ,  $\text{NO}_3^-$ , and  $\text{PO}_4^{3-}$  was based on the color changes of a single chemical sensor. The detection of  $\text{SO}_4^{2-}$ , total hardness, and pH were determined based on 4, 5, and 4 chemoresponsive sensors, respectively. The details of the colorimetric strips are summarized in Table 1. These strips were ideally designed to serve as a screening test for semi-quantitative purposes. For example, the color changes of the  $\text{Mn}^{2+}$  sensor strip were calibrated in accordance with the presence of  $\text{Mn}^{2+}$  in discrete concentration ranges of 0, 2, 5, 20, 50, and 100 ppm.

### Standard reagents

Chemical solutions, such as nitric acid ( $\text{HNO}_3$ ), acetic acid ( $\text{CH}_3\text{COOH}$ ), ascorbic acid ( $\text{C}_6\text{H}_8\text{O}_6$ ), sodium acetate ( $\text{CH}_3\text{COONa}$ ), sodium phosphate ( $\text{Na}_2\text{HPO}_4$ ), and ammonium acetate ( $\text{CH}_3\text{COONH}_4$ ), were prepared using analytical-grade chemicals (Merck, Germany). To construct the calibration models, a series of chemical solutions was prepared, the details of which are summarized in Table 1. The  $\text{Mn}^{2+}$ ,  $\text{Fe}^{3+}$ , and  $\text{Cu}^{2+}$  standard samples were diluted from certified reference materials (CRMs) purchased from Loba Chemie, India. Preparations of  $\text{NO}_3^-$ ,  $\text{SO}_4^{2-}$ , and  $\text{PO}_4^{3-}$  solutions were based on the standard methods that had been previously described (Franson 1992). For each of the sensor strips, the solutions were separated into two sets. These were defined as a training set that was used to establish the calibration model and a test set that was used to validate the predictive performance of the established calibration model.

### Acquisition of color data

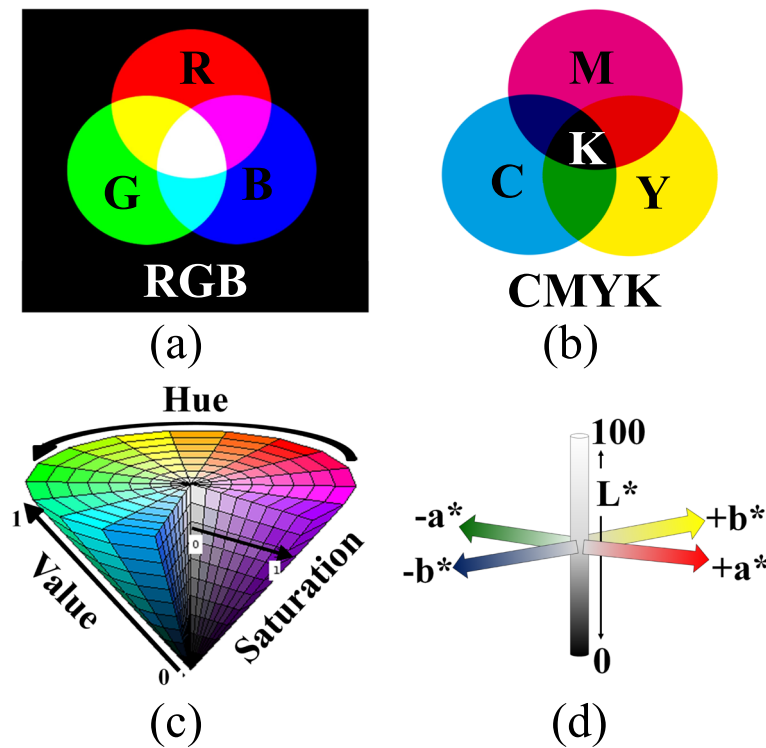
The colorimetric sensor detection method assumes that the changed color of the sensing materials could be related to the analyte concentration. To make a comparison of predictive performance, the color change of the chemical strips was converted into digital data using various color models such as RGB, CMYK, HSV, and CIELAB.

### RGB

RGB (red, green, blue) is regarded as the most common digitally coded color model and employs the three primary colors of red, green, and blue, as illustrated in Fig. 1a. RGB is an additive color model meaning that the presented colors are generated through combinations of the three primary colors. For example, yellow is representative

**Table 1** Detail of colorimetric strips used in this investigation

| Analyte                                      | No. of sensors | Concentration range                               | Concentrations of training samples                                   | Concentrations of validation samples                       |
|--|----------------|---|--|--|
| Manganese ion ( $\text{Mn}^{2+}$ )           | 1              | 0, 2, 5, 20, 50, and 100 (ppm)                    | 0.00, 3.00, 7.00, 15.00, 25.00, 45.00, 60.00, 80.00, and 100.00      | 2.00, 5.00, 10.00, 20.00, 35.00, 50.00, 70.00, and 90.00   |
| Copper ion ( $\text{Cu}^{2+}$ )              | 1              | 0, 10, 30, 100, and 300 (ppm)                     | 0.00, 10.00, 30.00, 80.0, 120, 170.00, 200.00, and 300.00            | 5.00, 20.00, 40.00, 100.00, 150.00, 180.00, and 2500.0     |
| Iron ion ( $\text{Fe}^{2+}/\text{Fe}^{3+}$ ) | 1              | 0, 2, 5, 10, 25, 50, and 100 (ppm)                | 0.00, 3.00, 7.00, 15.00, 25.00, 45.00, 60.00, 80.00, and 100.00      | 2.00, 5.00, 10.00, 20.00, 35.00, 50.00, 700.00, and 90.00  |
| Nitrate ( $\text{NO}_3^-$ )                  | 1              | 0, 10, 25, 50, 100, 250, and 500 (ppm)            | 0.00, 10.00, 20.00, 30.0, 50.0, 100.00, 200.0, 350.00, and 500.00    | 5.00, 15.00, 25.00, 40.0, 75.0, 150.00, 250.00, and 450.00 |
| Phosphate ( $\text{PO}_4^{3-}$ )             | 1              | 0, 3, 10, 25, 50, and 100 (ppm)                   | 0.00, 2.00, 5.00, 10.00, 20.00, 30.00, 50.0, 70.0, 90.00, and 100.00 | 1.00, 3.00, 7.00, 15.00, 25.00, 40.00, 60.00, and 80.00    |
| Sulfate ( $\text{SO}_4^{2-}$ )               | 4              | < 200, > 400, > 800, > 1200, and > 1600 (ppm)     | 0.00, 200.00, 400, ..., 1600   | 100,300, 500, ..., 1500                                    |
| Total hardness ( $\text{CaCO}_3$ )           | 5              | < 54, > 90, > 180, > 270, > 360, and > 445 (ppm)  | 0.00, 59.00, 75.00, 100.00, 137.00, 250.00, and 450.00               | 50.00, 62.00, 87.00, 106.00, 150.00, 200.00, and 375.00    |
| pH   | 4              | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14 | 0.93, 2.86, 4.70, 6.66, 9.10, 10.97, and 12.66                       | 1.83, 3.85, 5.83, 7.20, 9.64, and 11.76                    |



**Fig. 1** Schematic diagram showing the difference among the **a** RGB, **b** CMYK, **c** HSV, and **d** CIELAB color models

of a combination of green and red. Additionally, a full combination of red, green, and blue or [255(R) 255(G) 255(B)] results in white (Upadhyay et al. 2020b; Upadhyay et al. 2020c; Krongchai et al. 2020). In this research study, the acquisition of the images was performed using a scanner (L220, Epson, Philippines). On the scanner, the resolution was set to 300 dpi with 8 bits of color depth. The RGB images were recorded in true image file format (TIFF) without data compression. The images were then processed with a set of scripts and the function was developed using Matlab 2010. In this process, the first step was to screen out the pixels that were not considered part of the sensors and, therefore, should not be included in the calculation (Cantrell et al. 2010). In this step, the R, G, and B values of each pixel were compared with a threshold fraction calculated using the values recorded from the images of each sensor. In this research study, an average intensity of  $\pm 3$  standard deviations (SD) was used to set the thresholds. These values were obtained from all pixels of a selected sensing area of an image. The recorded intensities, which were either smaller or greater, were interpreted as background or outliers and were discarded from the analysis.

#### CMYK

CMYK (cyan, magenta, yellow, black) color model consists of three secondary colors, cyan, magenta, and

yellow, and is presented in Fig. 1b. In contrast to RGB, it is a subtractive color model as the final color is created by subtracting the primary colors from a white surface. Black or the key (K) is included to improve the density range and allow for the available color gamut of the generated color. Like RGB, CMYK is a device-dependent color model, meaning that the characteristics of each data acquisition tool could alter the recorded values. In other words, the recorded color intensities obtained from different scanners or cameras could be different. In this research study, CMYK values were converted from RGB values using the methods described previously (Cantrell et al. 2010). With 8 bits of color depth, color intensities ranged from 0 to 255. For example, absolute white is a combination of [0 0 0 0] in the CMYK color model.

#### HSV

HSV color model describes the color spectrum as a multidimensional model of three different factors including hue (H), saturation (S), and value (V) (Fan et al. 2019). The hue parameter specifies the color as an angular value between  $0^\circ$  and  $360^\circ$ . For example, red, green, and blue are located at  $0^\circ$ ,  $120^\circ$ , and  $240^\circ$ , respectively (Thajee et al. 2018). The saturation or chroma (S), which is also known as purity, indicates the amount of color ranging from 0 to 100% or 0 to 100%. A color with 100%

saturation is the purest color possible; whereas, with less saturation, a grayer color will appear. The value of  $V$  describes the brightness or intensity of the color ranging from 0 to 1, where 0 refers to absolute black and 1 indicates 100% brightness with no black mixed into the color at all. These three components can be visualized using a color geometric cone as illustrated in Fig. 1c, where hue is a given point on the color wheel. Saturation is represented as the radius of the cone, while the value component is represented as the height of the color combination cone.

### CIELAB

CIELAB or CIE  $L^*a^*b^*$  is a color model that imitates human perception of image colors (Sharifzadeh et al. 2014). CIELAB color model characterizes colors using three components. The first component is the lightness of color ( $L^*$ ), and the others are the chromatic components ( $a^*$  and  $b^*$ ) as shown in Fig. 1d. Usually, the  $L^*$  value ranges from 0 to 100, where  $L^* = 0$  and 100 will refer to black and white colors, respectively. The  $a^*$  value defines the distance through the red-green axis, and the positive and negative values indicate red and green, respectively. On the other hand, the  $b^*$  value defines the distance through the yellow-blue axis, where the positive and negative values indicate yellow and blue. The scaling and limits of the  $a^*$  and  $b^*$  axes depend upon the specific implementation, and they run in the range of  $\pm 110$  for the 8-bits of color depth. This color model is similar to the HSV color model where the intensity is decoupled from the chromaticity. Because the tonality changes are linear, the chromaticity differences can be computed using the Euclidean distance. Therefore, the CIELAB is considered a device-independent color model (Sharifzadeh et al. 2014).

### Data preprocessing of the decoded digital values

Data preprocessing involves applying a mathematical modification to the analytical values prior to initiating the data analysis methods. A numerical dataset may be shifted or rescaled in accordance with the variables/samples to place emphasis on the desired information obtained from the analysis. The use of a data preprocessing method could seriously affect not only the predictive ability but also the interpretation of the models (Famili et al. 1997). Common data preprocessing methods can be applied prior to application of the calibration model. These data pre-processing methods include data scaling, standardization, normalization and various distance-based methods. In this research study, the decoded digital values obtained from  $I$  samples with the  $J$  color parameters were organized into a data matrix  $X$ . The value of  $x_{ij}$  was positioned in the  $i$ th row and the  $j$ th

column element of the data matrix  $X$ . The preprocessed data were presented in a data matrix  $Z$ .

### Basic data preprocessing

To analyze the color spectrum data, many researchers have reported on the use of raw data without applying any data preprocessing method for prediction purposes (Pla-Tolós et al. 2016). In many cases, simple mathematical modifications can also be performed prior to making a prediction. Color change can be evaluated using the different intensities obtained from before and after chemical exposure (Zaragoza et al. 2015). The difference between the RGB values obtained from the before and after images was calculated using the absolute value of the color change (Tahir et al. 2016). Some studies have reported on the use of the differences in the values after the colors were converted into a gray scale format (Morsy et al. 2016). In addition, the obtained color intensities were subtracted by the background where no optical dyes were present in order to compensate for the non-uniformity in the background light (Chen et al. 2017). Lastly, the logarithm of the ratio between the color intensity determined from before and after the reaction was calculated to represent the effective intensity (De Almeida et al. 2015).

### Standardization

Standardization is an extension of mean centering and can be carried out by dividing each variable by its standard deviation after mean centering (Brereton 2003).

$$^s z_{ij} = (x_{ij} - \bar{x}_j) / (\sum_{i=1}^I (x_{ij} - \bar{x}_j)^2 / (I-1))^{1/2} \quad (1)$$

where  $\bar{x}$  is the mean value of intensities in the  $j$ th variable. After standardization, all the variables will be adjusted on the same scale. This can be useful when there are some variables that might be abundant in all samples, but when their variation is not very significant. For example, the change in concentration of a minor chemical might be significant in relation to the underlying system. Standardization can be used to ensure that the significance of the variables with low intensity is comparable to those of higher intensities. Without standardization, the predictive results may rely only on the variables having relatively high intensities.

### Normalization

Normalization can be done by dividing each element of a sample by a constant. If a sample is scaled to a constant total, this means that the sum over all variables after normalization is equal to 1. In this research study, each sample was normalized to the summation of the square values of the color intensities as follows:



$${}^{rs}z_{ij} = (x_{ij}^2 / \sum_{j=1}^J x_{ij}^2)^{1/2} \quad (2)$$

Normalization can be useful if it is difficult to precisely control the absolute intensities of the samples. The variations in each experimental run that may be not completely controlled could be minimized after data normalization. After normalization, each parameter contains information about the relative proportion of the compounds in the sample, and a comparison between the samples reveals information about how the ratios between the compounds differ.

#### Dissimilarity-based methods

Instead of using the decoded digital values directly for the calculations, it was possible to evaluate the color change using the dissimilarity between the colors of the tested sensors and the background of the reference images (Hoang et al. 2017). The color dissimilarities can be evaluated using Euclidean and Mahalanobis distances. The Euclidean distance for sample  $i$  (represented by row vector  $\mathbf{x}_i$ ) is calculated to the centroid of  $\mathbf{x}_c$  by

$$Eu_z^2 = (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \quad (3)$$

where  $z^2$  is the squared Euclidean distance between sample  $i$  and the centroid. In applications of CIELAB, this is a linear distance of the 3-dimensional space in the color components. The Mahalanobis distance is regarded as a standardized Euclidean distance and can be calculated by

$$Maz_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})S^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})' \quad (4)$$

where  $S$  is a diagonal matrix corresponding to the standard deviation of each variable.

#### Regression methods

In this research study, quantification was achieved based on the color models using single linear regression (ULR), multivariate linear regression (MLR), and partial least square (PLS) regression.

##### ULR and MLR

ULR is a simple linear regression that describes the relationship between one predictive (independent) and one response (dependent) variable into a first-degree polynomial regression assuming the data distribution has a linear association as follows:

$$y = ax + c \quad (5)$$

where  $y$  is a response variable or, in this research study, a concentration of the sample analyte which is an estimation based on a single-color parameter of  $x$ . The  $a$

and  $c$  parameters are the slope and intercept of the univariate linear regression line, respectively.

MLR is similar to ULR in that the prediction of the response is based on the simple linear regression. However, the influence of more than one independent variable contributes to the estimation of the response variable with the use of simple matrix operations:

$$y = X \cdot b \quad (6)$$

where  $y$  is a vector containing the response values estimated by the predictive data contained in matrix  $X$ , while  $b$  is a vector containing the coefficient of each predictive variable indicating the relationship between the predictive variables in estimation of the  $y$  responses. The calculation of  $b$  using the pseudo-inverse matrix operation has been effectively described previously (Brereton 2009).

##### PLS

PLS is a multivariate calibration model that captures both variations from the predictive ( $X$ ) and response ( $y$ ) data (Albayrak et al. 2019). A prediction is carried on based on the assumption that the covariance between those extracted variations is maximized (Brereton 2007). Therefore, in most cases, PLS is regarded as a powerful calibration method that could provide satisfactory predictive results (Xiaowei et al. 2015; Bordbar et al. 2018). In this research study, the prediction of PLS was achieved following the PLS1 algorithm that has been previously described (Brereton 2003). The number of PLS latent vectors was optimized using leave-one-out cross-validation (LOO-CV) (Alberti et al. 2020).

##### Statistical merits

The predictive performance level achieved in this research study was evaluated using the root mean square errors of calibration (RMSEC), the root mean square errors of prediction (RMSEP), the cross-validated explained variances ( $R^2$  and  $Q^2$ ) and the ratio between the RMSEP and RMSEC values ( $R_x$ ). The RMSEP is calculated as follows:

$$RMSEP = \sqrt{\sum_{i=1}^N (y_i - \hat{y}_i)^2 / I} \quad (7)$$

where  $\hat{y}_i$  is the predicted concentration for the test sample  $x_i$  having the actual value of  $y_i$ . The number of samples is denoted by  $I$ . The RMSEC value can be calculated using the same equation as the RMSEP value but by also using the training samples. The  $Q^2$  and  $R^2$  values for the predictive results of the test samples and training samples, respectively, can be calculated by using the same equation:

$$Q^2 = 1 - \left( \sum_{i=1}^I (\hat{y}_i - y_i)^2 / \sum_{i=1}^I (y - \bar{y})^2 \right) \quad (8)$$

Ideally, the predicted concentration should lie along the diagonal line indicating that the predicted and actual concentration are the same. In this case, the  $R^2$  or  $Q^2$  values will be high and close to 1, and this implies that the greater degree of variation within the data is modeled by the calibration model. In addition, the ratio of RMSEP and RMSEC are calculated as follows:

$$R_x = \text{RMSEP}/\text{RMSEC} \quad (9)$$

This statistical ratio can be used to indicate the robustness of the developed model. It is expected that the ratio will be as close to 1 as possible. A value that is higher than 1 implies that the model could be prone to an overfitting problem. The predictive result could be inconsistent when there is no change in the training samples, or when it is used to estimate the unknown variations of the test samples.

### Real samples

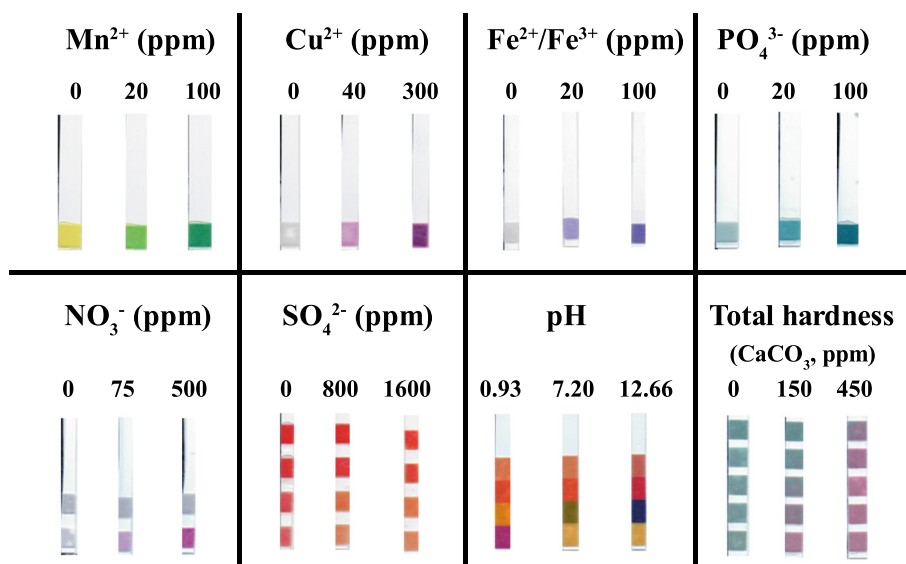
Four water samples were collected in Chiang Mai, Thailand representing the three water types that were investigated in this study. These included sewage canal (W1), wastewater treatment pond (W2), and water reservoirs (W3 and W4). The water samples were collected in December 2018. The concentration of metal ions was determined using the Air–Acetylene Flame method (Rice et al. 2017). The turbidimetric method was used to measure the concentrations of  $\text{SO}_4^{2-}$  (Rice et al. 2017).

$\text{NO}_3^-$  was assessed following the Brucine colorimetric method (George 2012). The pH values were measured using a pH meter (SevenCompact pH/Ion meter S220, Mettler-Toledo, Switzerland). Total hardness was evaluated using the titrimetric method (Rice et al. 2017). Lastly, an ascorbic acid method was used to quantify the concentrations of  $\text{PO}_4^{3-}$  (Rice et al. 2017).

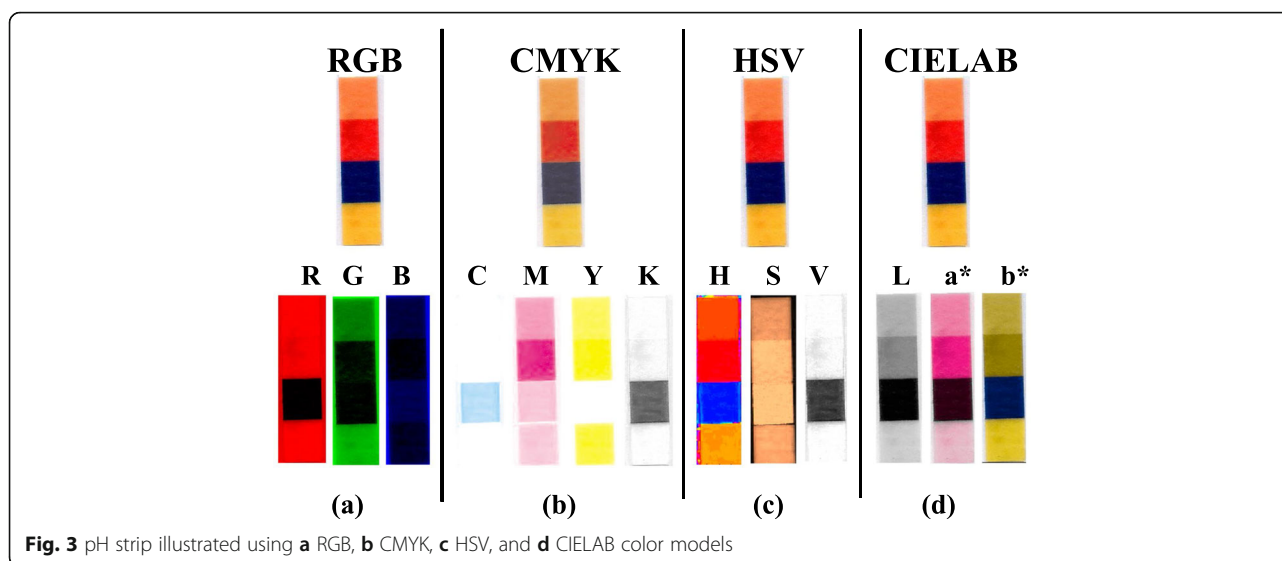
## Results and discussion

### Image visualization using different color models

Figure 2 presents images of the colorimetric sensors and sensor arrays after reacting to the standard solutions of different concentrations. Detection of  $\text{Mn}^{2+}$ ,  $\text{Cu}^{2+}$ ,  $\text{Fe}^{2+}/\text{Fe}^{3+}$ ,  $\text{NO}_3^-$  and  $\text{PO}_4^{3-}$  values was based on the uses of colorimetric strips having a single sensing chemical, whereas the detection of  $\text{SO}_4^{2-}$ , pH, and total hardness was based on colorimetric strip sensor array composed of 4, 4, and 5 sensing elements, respectively. Figure 3 displays the images of a pH strip visualized by the four different color models wherein the differences can hardly be noticed. Color models can be used to visualize the same color spectrum, although different color parameters need to be used. However, differences could be observed when images were decomposed and presented through the use of an individual color parameter. The RGB (Fig. 3a) and CMYK (Fig. 3b) models characterized the image sensors based on the primary and secondary colors, respectively. On the other hand, the colors were characterized by the three different color factors employed in cases involving HSV and CIELAB.



**Fig. 2** Colorimetric sensors and sensor arrays after immersion in standard solutions at different concentrations. It was noted that only the first reactive zone that was used to measure  $\text{NO}_3^-$ . The second zone was used to measure the concentration of nitrite ion ( $\text{NO}_2^-$ ) which was not used in this research study



### Exploratory data analysis of the color data

Principal component analysis (PCA) score plots of the RGB, CMYK, HSV, and CIELAB digitally coded values are presented in Fig. 4. Due to the fact that each of the color models were characterized by different numbers of parameters, the scores of each dataset were adjusted through Procrustes transformation so that the different types of the measurements could be comparable and presented within the same PCA space (Andrade et al. 2004). The linearity trends can be observed in the color data acquired from most colorimetric strips. For example, the concentrations of  $\text{Mn}^{2+}$  increased along with the increases of the PC1 and PC2 scores. Therefore, the samples were distributed along the diagonal line of the PCA space implying the sensible use of the linear calibration methods. The scattering trend in the total hardness detection could be due to the fact that the strip was originally fabricated for the purposes of determining the concentration levels of  $\text{CaCO}_3$  or for scanning purposes (see concentration ranges presented in Table 1). In each PCA score plot, when a sample was presented using different color models, they were located differently on the PCA space. This implied that variations in the different color data could create alternations in the predictive performance when they were used to establish the applicable calibration models.

### Effect of different color models

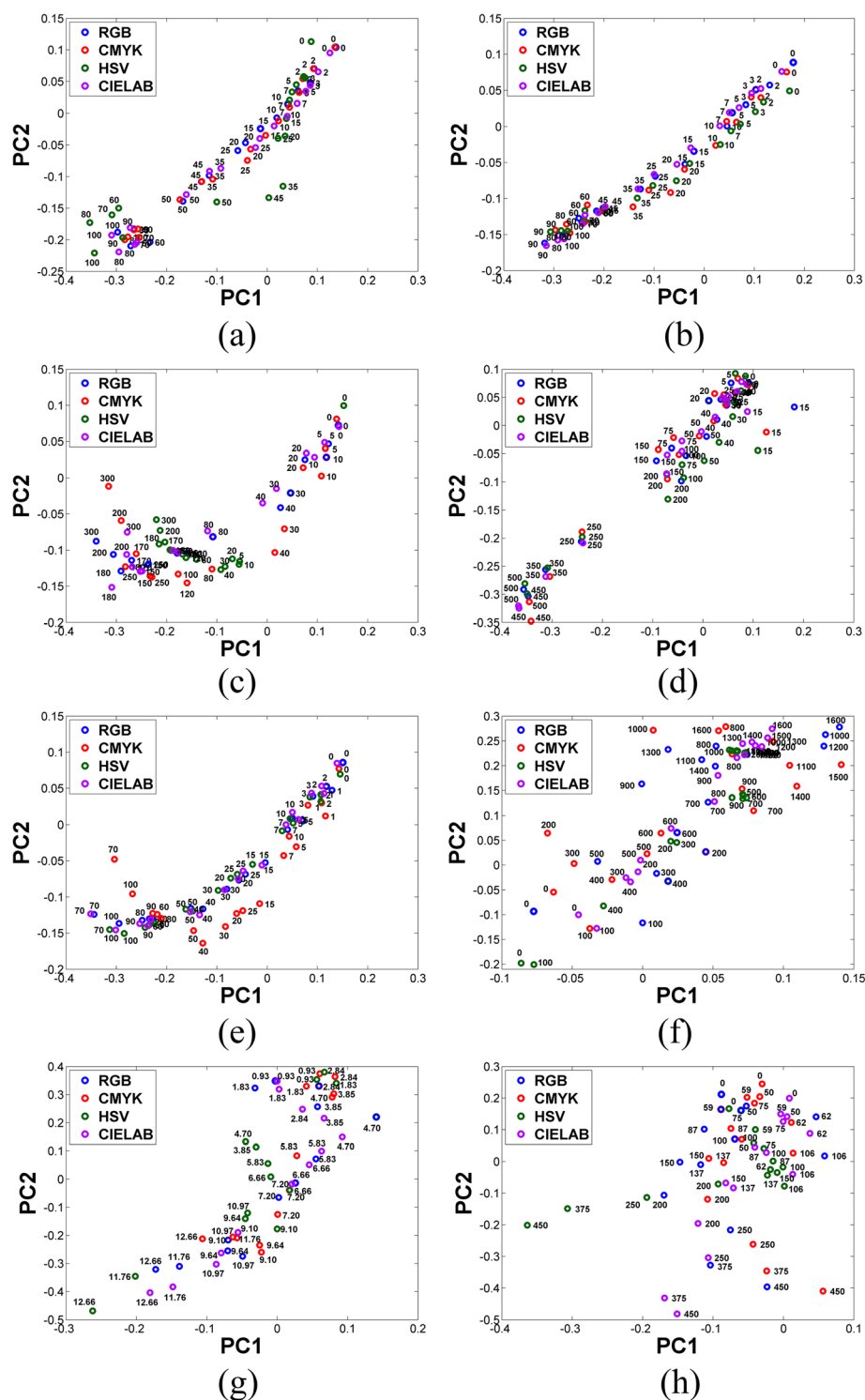
To compare the predictive abilities when the calibration models were calibrated using different color parameters, the ULR models were constructed using the training samples of each colorimetric strip. In this comparison, the digitally coded values were used without any data pretreatment (no data preprocessing). For the sensor

array with multiple chemical detections, the multivariate data was calibrated using the PLS method. The correlation between the actual and predicted concentrations of the test samples for each of the colorimetric strips are illustrated in Fig. 5. The predictive results of the established models are also summarized in Supplementary Materials Table S1.

### Prediction of colorimetric strips using a single chemical sensor

In Fig. 5a–e, variations in the predictive results can be observed in the ULR prediction using different color parameters. For example, in the detection of  $\text{Mn}^{2+}$  (Fig. 5a), the ULR calibration model based on the G parameter or the greenness of the image resulted in the lowest RMSEP value of 4.14 with a  $Q^2$  value of 0.9813. As a result, the samples were placed closer along the diagonal line of the correlation graph. On the contrary, the ULR calibration model using the B parameter resulted in relatively higher predictive errors with RMSEC and RMSEP values of 26.92 and 18.58, respectively. For detection of  $\text{Mn}^{2+}$ ,  $\text{PO}_4^{3-}$ , and  $\text{NO}_3^-$  values, the colors that were most similar to the appearing color of the chemical sensor could offer satisfactory results. For example, the detections of  $\text{PO}_4^{3-}$  and  $\text{Mn}^{2+}$ , where the sensor change can be correlated to an increase in the color green, the lowest RMSEP value was reported with the ULR model using the G parameter. Identically, the ULR model based on the detection of the color magenta (M parameter) resulted in the best detection of  $\text{NO}_3^-$ . In this case, the color of the strip was changed from white to magenta. Though, it is not always the case, the detections of  $\text{Fe}^{2+}/\text{Fe}^{3+}$  and  $\text{Cu}^{2+}$  were achieved by the appearance of royal purple (blue-magenta) and dark magenta, respectively. However, K and G parameters could result in

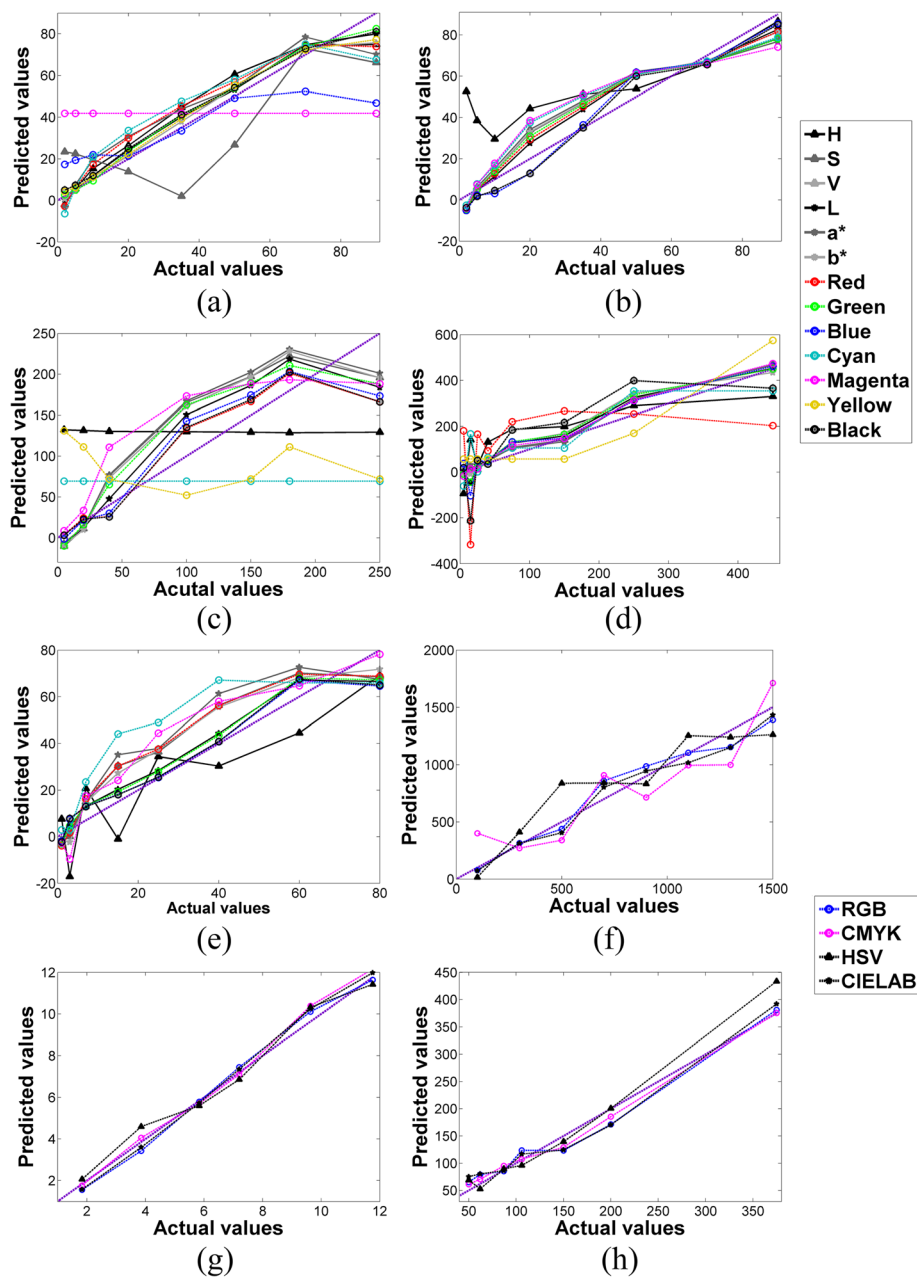




**Fig. 4** Procrustes PCA score plots of **a**  $\text{Mn}^{2+}$ , **b**  $\text{Fe}^{2+}/\text{Fe}^{3+}$ , **c**  $\text{Cu}^{2+}$ , **d**  $\text{NO}_3^-$ , **e**  $\text{PO}_4^{3-}$ , **f**  $\text{SO}_4^{2-}$ , **g** pH, and **h** total hardness test strips

significantly lower RMSEP values in the detection of  $\text{Fe}^{2+}/\text{Fe}^{3+}$ . In the detection of  $\text{Cu}^{2+}$ , the use of the parameters of R, G, and B could result in low predictive errors of RMSEP at 36.15, 39.79, and 35.83, respectively,

whereas the use of the M parameter, which was supposed to be more closely related to changes in  $\text{Cu}^{2+}$  concentrations, resulted in a relatively higher predictive error of 47.89 of RMSEP.



**Fig. 5** Correlation graphs comparing the predicted values and actual values presented in Supplementary Materials Table S1 using ULR prediction **a**  $\text{Mn}^{2+}$ , **b**  $\text{Fe}^{2+}/\text{Fe}^{3+}$ , **c**  $\text{Cu}^{2+}$ , **d**  $\text{NO}_3^-$ , and **e**  $\text{PO}_4^{3-}$  and PLS prediction of **f**  $\text{SO}_4^{2-}$ , **g** pH, and **h** total hardness test strips

With regard to the use of HSV and CIELAB, in this situation, they obtained poorer predictive performance levels resulting in low  $Q^2$  and high RMSEP values as presented in Supplementary Materials Table S1. These color modes decomposed the color spectrum into different light components. For example, in HSV, the H parameter represents the actual color, and S and V represent color saturation and brightness, respectively. Since the H parameter mainly indicates the difference

between the colors, it could not effectively detect changes in color intensity. For instance, in the detection of  $\text{Cu}^{2+}$  (Fig. 5c), where the sensor changed from light magenta to dark magenta, the use of the H parameter resulted in a significantly high prediction error (RMSEP of 88.84) when compared with the use of the B parameter (RMSEP of 35.83). In addition, when the light component parameters (S and V parameters) were separately used for the purposes of prediction, they could be more

affected by the inconsistency of the light source of the scanner. The inconsistency of the light source will be demonstrated and discussed later in the next section.

#### **Prediction of colorimetric strips using multiple chemical sensors**

In addition, with regard to the detections of the  $\text{SO}_4^{2-}$ , pH, and total hardness using chemical strips composed of more than one sensing element, the multivariate regression using PLS could outperform both of ULR and MLR models (Supplementary Materials Tables S7-S9). Variations could still be observed when the predictions were based on different color models. For example, in the detection of  $\text{SO}_4^{2-}$ , the PLS-RGB model produced relatively lower RMSEP values when compared with the PLS-CMYK model yielding RMSEP values of 94.25 and 206.69, respectively. Interestingly, in the case where intensity data were used without any data preprocessing, the prediction errors based on the multivariate use of the HSV or CIELAB color models were not apparently high when compared with those based on individual uses (Supplementary Materials Table S1). These predictive outcomes demonstrate the advantage of the multivariate calibration where all of the recorded parameters were simultaneously used to establish the prediction model in order to achieve the best predictive performance.

#### **Effects of different data preprocessing methods**

To investigate the effects of different data preprocessing methods on the predictive performance of the colorimetric calibration models, all preprocessing methods were applied to each of the digitally coded data prior to establish the model constructions. The predictive results are presented in Supplementary Materials Tables S2-S9. The colorimetric prediction models which offered the top five lowest RMSEP values are listed in Table 2 for comparison.

In addition to the choice of the color model used, a different preprocessing method could dramatically affect the predictive performance of the calibration models. Compared with the established model employing raw data (no data preprocessing), the predictive results could be either improved or diminished when various other data preprocessing methods were applied. For example, in the determination of  $\text{NO}_3^-$ , the RMSEP value was decreased from 44.31 to 28.28 after the RGB values were treated with logarithm scaling prior to the PLS modeling. In contrast, using the HSV values, the PLS model resulted in higher prediction error values with an RMSEP value of 59.58 when data were standardized prior to analysis.

#### **Prediction of colorimetric strips using a single chemical sensor**

The detections of  $\text{Fe}^{2+}/\text{Fe}^{3+}$ ,  $\text{Cu}^{2+}$ , and  $\text{PO}_4^{3-}$  values were performed using the colorimetric strips with a single chemical sensor. In this research study, the calibration models employed without any data preprocessing of the colorimetric strips could result in the lowest predictive errors of the RMSEPs. These prediction models were based on a simple linear regression of the change in color intensity.

In the case of  $\text{Fe}^{2+}/\text{Fe}^{3+}$  determination, although the use of ULR provided the smallest RMSEP value of 5.80, the prediction of the training samples was relatively poor while the RMSEC value was as high as 12.60. Consequently, this model was prone to the underfitting problem with an  $R_x$  ratio of 0.48. In this situation, it was recommended that a more stable model be used with a slight increase in the RMSEP value. For example, the RGB-PLS model should be employed with absolute color changes as a preferred data preprocessing method. However, the errors in the prediction of the test samples were slightly increased to 8.02. In the detection of the  $\text{Fe}^{2+}/\text{Fe}^{3+}$ , MLR and PLS values provided comparably satisfactory predictive results with high  $R^2$  and  $Q^2$  values and low RMSEC and RMSEP values. This was the case because all parameters can be significant for model prediction and because the optimum number of PLS latent variables were close or the same as the number of the parameters used.

In contrast to the detections of  $\text{Cu}^{2+}$ ,  $\text{Fe}^{2+}/\text{Fe}^{3+}$ , and  $\text{PO}_4^{3-}$ , the lowest errors identified in the prediction process of the  $\text{Mn}^{2+}$  and  $\text{NO}_3^-$  concentrations were related to the utilization of the multivariate modelings of PLS and MLR. Notably, the best predictive results of  $\text{Mn}^{2+}$  were obtained from the RGB-PLS model with standardization. However, it may not be recommended that spectroscopy data, such as UV-Vis and NIR, be pretreated by standardization (Brereton 2009). After standardization, the variations of all variables were adjusted to be within the same range. Since the absorbance of not all wavelengths contained useful information in relation to the predicted responses or they presented only pattern of noise, the standardization process might amplify the noise variations, and this undesirable variation could then lead to wrongful predictions. In this case, although the PLS model with standardization could result in the lowest RMSEP value of 3.42, the  $R_x$  value of 2.28 was considered relatively high. This implied that the errors in the auto prediction process and those of the test modes were too different and the model was prone to an overfitting problem (Wongsaipun et al. 2018). Therefore, the second best PLS model, wherein the RGB values were subtracted by the background (R-

**Table 2** Five best predictive performance results of each colorimetric sensor

| Analyte           | No. | Chemometric method | Color model | Data preprocessing | $R^2$  | $Q^2$  | RMSEC  | RMSEP | $R_x$ | RMSEP <sub>eyes</sub> |
|-------------------|-----|--------------------|-------------|--------------------|--------|--------|--------|-------|-------|-----------------------|
| $Mn^{2+}$         | 1   | PLS                | RGB         | S                  | 0.9981 | 0.9872 | 1.50   | 3.42  | 2.28  | 5.35                  |
|                   | 2   | MLR, PLS           | RGB         | F                  | 0.9879 | 0.9870 | 3.74   | 3.45  | 0.92  |                       |
|                   | 3   | PLS                | RGB         | SD                 | 0.9993 | 0.9845 | 0.92   | 3.76  | 4.08  |                       |
|                   | 4   | MLR                | CMYK        | None               | 0.9893 | 0.9829 | 3.53   | 3.95  | 1.12  |                       |
|                   | 5   | PLS                | CMYK        | F                  | 0.9935 | 0.9796 | 2.75   | 4.31  | 1.57  |                       |
| $Fe^{2+}/Fe^{3+}$ | 1   | ULR                | CMYK        | K                  | 0.8729 | 0.9632 | 12.16  | 5.80  | 0.48  | 8.62                  |
|                   | 2   | ULR                | CMYK        | AK                 | 0.8765 | 0.9571 | 11.98  | 6.26  | 0.52  |                       |
|                   | 3   | ULR                | RGB         | B                  | 0.8794 | 0.9555 | 11.84  | 6.37  | 0.54  |                       |
|                   | 4   | ULR                | CIELAB      | A                  | 0.8857 | 0.9450 | 11.53  | 7.09  | 0.62  |                       |
|                   | 5   | ULR                | RGB         | R                  | 0.8794 | 0.9555 | 11.84  | 6.37  | 0.54  |                       |
| $Cu^{2+}$         | 1   | ULR                | RGB         | B                  | 0.9359 | 0.8208 | 24.87  | 35.83 | 1.44  | 48.18                 |
|                   | 2   | ULR                | RGB         | AR                 | 0.9638 | 0.8200 | 18.70  | 35.90 | 1.92  |                       |
|                   | 3   | ULR                | RGB         | AB                 | 0.9317 | 0.8196 | 25.67  | 35.95 | 1.40  |                       |
|                   | 4   | ULR                | RGB         | R                  | 0.9555 | 0.8176 | 20.71  | 36.15 | 1.75  |                       |
|                   | 5   | ULR                | CMYK        | K                  | 0.9499 | 0.8132 | 21.98  | 36.58 | 1.66  |                       |
| $NO_3^-$          | 1   | MLR                | CMYK        | F                  | 0.9688 | 0.9736 | 29.44  | 23.55 | 0.80  | 33.03                 |
|                   | 2   | PLS                | CMYK        | D                  | 0.9596 | 0.9628 | 33.50  | 27.96 | 0.83  |                       |
|                   | 3   | MLR                | CMYK        | D                  | 0.9696 | 0.9624 | 29.04  | 28.12 | 0.97  |                       |
|                   | 4   | PLS                | RGB         | I                  | 0.9774 | 0.9619 | 25.06  | 28.28 | 1.13  |                       |
|                   | 5   | PLS                | CMYK        | A                  | 0.9652 | 0.9614 | 31.09  | 28.46 | 0.92  |                       |
| $PO_4^{3-}$       | 1   | ULR                | RGB         | G value            | 0.8864 | 0.9442 | 12.02  | 6.38  | 0.53  | 9.74                  |
|                   | 2   | PLS                | HSV         | S                  | 0.8790 | 0.9439 | 12.41  | 6.40  | 0.52  |                       |
|                   | 3   | ULR                | CMYK        | AK                 | 0.8870 | 0.9397 | 11.99  | 6.64  | 0.55  |                       |
|                   | 4   | ULR                | CMYK        | K                  | 0.8813 | 0.9381 | 12.29  | 6.72  | 0.55  |                       |
|                   | 5   | ULR                | RGB         | B                  | 0.8777 | 0.9380 | 12.47  | 6.73  | 0.54  |                       |
| $SO_4^{2-}$       | 1   | PLS                | RGB         | I                  | 0.9696 | 0.9705 | 90.10  | 78.75 | 0.87  | 198.89                |
|                   | 2   | PLS                | CIELAB      | None               | 0.9705 | 0.9667 | 88.74  | 83.64 | 0.94  |                       |
|                   | 3   | PLS                | CIELAB      | SO                 | 0.9568 | 0.9598 | 107.35 | 91.86 | 0.86  |                       |
|                   | 4   | PLS                | RGB         | N                  | 0.9408 | 0.9583 | 125.66 | 93.55 | 0.74  |                       |
|                   | 5   | PLS                | RGB         | None               | 0.9439 | 0.9577 | 122.33 | 94.25 | 0.77  |                       |
| pH                | 1   | PLS                | RGB         | N                  | 0.9996 | 0.9955 | 0.08   | 0.22  | 2.91  | 0.74                  |
|                   | 2   | PLS                | RGB         | SD                 | 1.0000 | 0.9946 | 0.03   | 0.25  | 9.47  |                       |
|                   | 3   | PLS                | CMYK        | N                  | 1.0000 | 0.9920 | 0.02   | 0.30  | 14.20 |                       |
|                   | 4   | PLS                | CIELAB      | None               | 0.9984 | 0.9912 | 0.16   | 0.31  | 1.96  |                       |
|                   | 5   | PLS                | RGB         | None               | 0.9999 | 0.9916 | 0.03   | 0.31  | 10.79 |                       |
| Total hardness    | 1   | PLS                | CMYK        | F                  | 1.0000 | 0.9905 | 0.18   | 10.22 | 57.28 | 48.74                 |
|                   | 2   | PLS                | CMYK        | None               | 0.9998 | 0.9882 | 2.21   | 11.37 | 5.15  |                       |
|                   | 3   | PLS                | RGB         | F                  | 0.9999 | 0.9880 | 1.58   | 11.48 | 7.25  |                       |
|                   | 4   | PLS                | RGB         | A                  | 0.9998 | 0.9817 | 2.12   | 14.15 | 6.67  |                       |
|                   | 5   | PLS                | CMYK        | A                  | 0.9997 | 0.9803 | 2.27   | 14.68 | 6.47  |                       |

A = Color change (different intensities before and after the chemical exposure), AR = Change of red parameter, B = Intensity of blue parameter, F = Color intensities were subtracted by background, I = Logarithm, N = Normalization, S = Standardization, SO = Standardization of a\* and b\* parameters, AB = Change of blue parameter, AK = Change of black parameter, D = Absolute value of color change, G = Intensity of green parameter, K = Intensity of black parameter, R = Intensity of red parameter, SD = Standardization of color change, RMSEP<sub>eyes</sub> = Predictive errors using naked eye,

Rwb, G-Gwb, B-Bwb), should be used instead to avoid this problem of inconsistency.

The determination of  $\text{NO}_3^-$  demonstrated that the color data put through an appropriate data preprocessing step could positively affect the predictive performance of the model. In this case, the CMYK with the background subtraction could offer the best predictive performance of the synthetic test samples. Normally, PLS could provide satisfactory predictive results. However, in this case, only the comparable predictive results between PLS and MLR were obtained. This could have occurred because in situations where the number of variables were not high, PLS could result in similar outcomes since the number of latent variables were not quite different or were the same as the number of the parameters in the original data space.

#### Prediction of colorimetric strips using multiple chemical sensors

The determinations of  $\text{SO}_4^{2-}$ , pH, and total hardness were based on the colorimetric strips equipped with multiple chemical sensors. To utilize all of the colorimetric sensor arrays, ULR would not be practical since it can only model one independent variable each time for a prediction of the response. Although the MLR model

could be calculated, the calibration algorithm resulted in high predictive errors in all cases. This was due to the fact that the numbers of the modeling samples were lower than the numbers of the predictive variables; therefore, the pseudo-inverse method was determined to be mathematically inappropriate (Saxena and Prathipati 2003). Therefore, PLS provided the best predictive results with low predictive errors for the colorimetric strips equipped with multiple chemical sensors. The high value of  $Q^2$  with low  $R_x$  implied that the developed model could be practically used for the prediction of unknown or test samples.

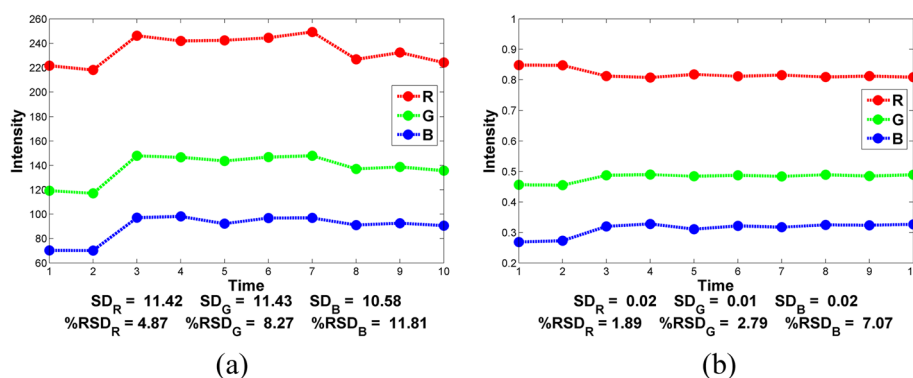
#### Monitoring of real water samples

Table 3 presents the predictive results of the sensor strips used for determining the concentrations of the selected water monitoring parameters using the PLS calibration models. By the standard methods, the concentrations of  $\text{Mn}^{2+}$ ,  $\text{Fe}^{2+/3+}$ ,  $\text{Cu}^{2+}$ ,  $\text{NO}_3^-$ ,  $\text{PO}_4^{3-}$ , and  $\text{SO}_4^{2-}$  were very low or non-detectable. Therefore, to evaluate the accuracy of the developed calibration models, the analytes were spiked into the samples, and the predictive results were reported as the %recovery.

**Table 3** Predictive results of samples using RGB color model after normalization

| <b>Mn<sup>2+</sup> (ppm)</b>                  |                 |                     |           | <b>Fe<sup>2+</sup>/Fe<sup>3+</sup> (ppm)</b> |                     |           |
|---|-----------------|---------------------|-----------|--|---------------------|-----------|
| Sample  | [Spiked sample] | RGB + normalization | %Recovery | [Spiked sample]                              | RGB + normalization | %Recovery |
| W1  | 40              | 39.85 ± 1.91        | 95–104    | 40   | 57.82 ± 3.39        | 136–153   |
| W2  | 40              | 31.74 ± 1.28        | 76–82     | 40   | 65.09 ± 2.78        | 156–170   |
| W3  | 40              | 35.31 ± 7.05        | 71–106    | 40   | 61.30 ± 4.22        | 143–164   |
| W4  | 40              | 40.45 ± 0.56        | 100–103   | 40   | 67.24 ± 3.19        | 160–176   |
| <b>Cu<sup>2+</sup> (ppm)</b>                  |                 |                     |           | <b>NO<sub>3</sub><sup>-</sup> (ppm)</b>      |                     |           |
| Sample  | [Spiked sample] | RGB + normalization | %Recovery | [Spiked sample]                              | RGB + normalization | %Recovery |
| W1  | 150             | 160.27 ± 6.87       | 102–111   | 250  | 292.47 ± 33.37      | 104–130   |
| W2  | 150             | 147.98 ± 4.89       | 95–102    | 250  | 297.02 ± 24.57      | 109–129   |
| W3  | 150             | 150.93 ± 3.42       | 98–103    | 250  | 286.75 ± 16.77      | 108–121   |
| W4  | 150             | 152.57 ± 10.15      | 95–108    | 250  | 260.39 ± 23.95      | 95–114    |
| <b>PO<sub>4</sub><sup>3-</sup> (ppm)</b>      |                 |                     |           | <b>SO<sub>4</sub><sup>2-</sup> (ppm)</b>     |                     |           |
| Sample  | [Spiked sample] | RGB + normalization | %Recovery | [Spiked sample]                              | RGB + normalization | %Recovery |
| W1  | 40              | 41.51 ± 3.06        | 77–89     | 800  | 813.24 ± 180.29     | 79–124    |
| W2  | 40              | 59.79 ± 1.39        | 106–112   | 800  | 860.69 ± 108.44     | 94–121    |
| W3  | 40              | 35.14 ± 1.83        | 67–74     | 800  | 848.61 ± 60.37      | 98–113    |
| W4  | 40              | 36.94 ± 3.07        | 68–80     | 800  | 932.06 ± 13.90      | 115–118   |
| <b>Total hardness (CaCO<sub>3</sub>, ppm)</b> |                 |                     |           | <b>pH</b>                                    |                     |           |
| Sample  | [Sample]        | RGB + normalization | %Error    | pH   | RGB + normalization | %Error    |
| W1  | 93.3            | 62.82 ± 19.96       | 11–54     | 7.43   | 7.06 ± 0.11         | 3–6       |
| W2  | 189.3           | 183.95 ± 18.25      | 7–12      | 7.96   | 6.11 ± 0.13         | 22–25     |
| W3  | 85.3            | 61.03 ± 18.25       | 7–5       | 7.32   | 7.84 ± 0.07         | 8–6       |
| W4  | 24              | 16.90 ± 26.97       | 83–142    | 7.32   | 6.72 ± 0.33         | 4–13      |





**Fig. 6** RGB intensity recorded from the pH test strips using experimental scanner for 10 consecutive hours **a** without using data preprocessing and **b** normalization

With the real sample waters, the strips could successfully identify the sample concentrations when the %recoveries did not deviate much from 100%. In this case, the digitally coded values were preprocessed with normalization prior to the standard protocol of the chemometric calibration models. In this experiment, errors may have occurred which could not be controlled, and these could have caused changes in the overall intensity of the recorded colors. Examples of these errors include those that are related to laboratory conditions, inconsistency in the light source of the scanner, the amount of the liquid samples or the chemical sensors used, and the scanning time. These errors could create fluctuations in the predictive results. It is important to have consistent measurements, or, in other words, the recorded images should be calibrated to remove these systematic biases prior to the chemometric analyses. Using the normalization equation, the color intensities were divided by a constant value that could be directly calculated from each sample. In this work, the sum of the squares of all color parameters of each sample was equal to 1. Consequently, the variations of each sample could be immobilized to be appropriately the same and the data from each measurement could be comparable with each other. This idea is quite similar to the recommendation of a physical addition of some internal standard chemicals in the analysis of HPLC or GC (Harvey 2011). In this case, the overall presence of variations for each test was calibrated to be a constant value. However, for the internal standard, the peak intensities of each sample were related to the peak intensity of the internal standard chemical.

Figure 6 shows the intensity of the RGB colors recorded from the same pH chemical sensor when the time was prolonged from 1 to 10 h. In Fig. 6a, although the same scanner was used, the recorded RGB intensities were randomly deviated. Using these raw RGB intensities, it could lead to variations in the predictive results. After normalization, as shown in Fig. 6b, the recorded

RGB intensities became approximately the same. The %RSD of each color parameter then became gradually smaller. This could guarantee that variations in the recorded color change exclusively corresponded to the concentration of the targeted analyte enhancing the predictive performance of the colorimetric strip detections.

## Conclusion

The selection of the color models was an important factor affecting the predictive performance of the calibration models. In addition, the use of an appropriate data preprocessing method could be important to guarantee the accurate predictive results from the chemometric calibration models. In general, PLS offered a satisfactory level of predictive performance. Ideally, several potential color models could be tested to obtain the best prediction. In this research study, when tested with the real samples, the use of normalization as a data preprocessing method was recommended. This data preprocessing method could minimize the error variation from different experimental conditions because it would be impractical to reestablish a new calibration model each time for the sensor detection of the unknown samples.

## Abbreviations

CIE: International Commission on Illumination; CMYK: Cyan, magenta, yellow, black; CRM: Certified reference materials; HSV: Hue, saturation, value; LOO-CV: Leave-one-out cross-validation; MLR: Multivariate linear regression; PLS: Partial least squares; RGB: Red, green, blue; RMSEC: Root mean square errors of calibration; RMSEP: Root mean square errors of prediction; ULR: Univariate linear regression

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40543-021-00271-9>.

**Additional file 1: Table S1.** Predictive performance of colorimetric sensors using ULR model and colorimetric sensor arrays using PLS model without data preprocessing. **Table S2.** Predictive performance of  $Mn^{2+}$  test strips using ULR, MLR and PLS coupled with various forms of data-preprocessing. **Table S3.** Predictive performance of  $Fe^{2+}/Fe^{3+}$  test strips

using ULR, MLR and PLS coupled with various forms of data-reprocessing. **Table S4.** Predictive performance of  $\text{Cu}^{2+}$  test strips using ULR, MLR and PLS coupled with various forms of data-reprocessing. **Table S5.** Predictive performance of  $\text{NO}_3^-$  test strips using ULR, MLR and PLS coupled with various forms of data-reprocessing. **Table S6.** Predictive performance of  $\text{PO}_4^{3-}$  test strips using ULR, MLR and PLS coupled with various forms of data-reprocessing. **Table S7.** Predictive performance of  $\text{SO}_4^{2-}$  test strips using ULR, MLR and PLS coupled with various forms of data-reprocessing. **Table S8.** Predictive performance of pH test strips using ULR, MLR and PLS coupled with various forms of data-reprocessing. **Table S9.** Predictive performance of total hardness test strips using ULR, MLR and PLS coupled with various forms of data-reprocessing.

## Acknowledgements

This research was supported by Chiang Mai University. The Center of Excellence for Innovation in Chemistry (PERCH-CIC), Office of the Higher Education Commission, Ministry of Education (OHEC) is also gratefully acknowledged. N. Phuangsaibai would like to thank the Science Achievement Scholarship of Thailand (SAST).

## Authors' contributions

Nutthatida Phuangsaibai: Methodology, Validation, Formal analysis, Investigation, Writing—original draft, Visualization. Jaroon jakmunee: Resources, Supervision. Sila Kittiwachana: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing—original draft, Visualization, Supervision, Resources, Writing—review & editing, Funding acquisition. The authors read and approved the final manuscript.

## Funding

Not applicable

## Availability of data and materials

The datasets are available upon request. All analyzed data during this study are included in this published article and its supplementary information file.

## Declarations

## Ethics approval and consent to participate

Not applicable

## Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup>Department of Chemistry, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand. <sup>2</sup>Center of Excellence for Innovation in Chemistry, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand. <sup>3</sup>Environmental Science Research Center (ESRC), Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand.

Received: 17 January 2021 Accepted: 23 March 2021

Published online: 08 April 2021

## References

- Albayrak M, Demirkaya-Miloglu F, Senol O, Polatdemir E. Design, optimization, and validation of chemometrics-assisted spectrophotometric methods for simultaneous determination of etodolac and thiocolchicoside in pharmaceuticals. *J Anal Sci Technol*. 2019;10:16. <https://doi.org/10.1186/s40543-019-0176-2>.
- Alberti G, Zanoni C, Magnaghi LR, Biesuz R. Low-cost, disposable colourimetric sensors for metal ions detection. *J Anal Sci Technol*. 2020;11:30. <https://doi.org/10.1186/s40543-020-00221-x>.
- Anand T, Sahoo SK. Cost-effective approach to detect  $\text{Cu(II)}$  and  $\text{Hg(II)}$  by integrating a smartphone with the colorimetric response from a NBD-benzimidazole based dyad. *Phys Chem Chem Phys*. 2019;21:11839–45. <https://doi.org/10.1039/C9CP00002J>.
- Andrade JM, Gómez-Carracedo MP, Krzanowski W, Kubista M. Procrustes rotation in analytical chemistry, a tutorial. *Chemom Intell Lab Syst*. 2004;72:123–32. <https://doi.org/10.1016/j.chemolab.2004.01.007>.
- Bordbar MM, Tashkhourian J, Hemmateenejad B. Qualitative and quantitative analysis of toxic materials in adulterated fruit pickle samples by a colorimetric sensor array. *Sensors Actuators B Chem*. 2018;257:783–91. <https://doi.org/10.1016/j.snb.2017.11.010>.
- Brereton RG. Chemometrics: data analysis for the laboratory and chemical plant: Wiley; 2003.
- Brereton RG. Applied chemometrics for scientists: Wiley; 2007.
- Brereton RG. Chemometrics for pattern recognition: Wiley; 2009.
- Cantrell K, Erenas MM, de Orbe-Payá I, Capitán-Vallvey LF. Use of the hue parameter of the hue, saturation, value color space as a quantitative analytical parameter for bitonal optical sensors. *Anal Chem*. 2010;82:531–42. <https://doi.org/10.1021/ac901753c>.
- Chen Y, Fu G, Zilberman Y, Ruan W, Ameri SK, Zhang YS, et al. Low cost smart phone diagnostics for food using paper-based colorimetric sensor arrays. *Food Control*. 2017;82:227–32. <https://doi.org/10.1016/j.foodcont.2017.07.003>.
- De Almeida LKS, Chigome S, Torto N, Frost CL, Pletschke BI. A novel colorimetric sensor strip for the detection of glyphosate in water. *Sensors Actuators B Chem*. 2015;206:357–63. <https://doi.org/10.1016/j.snb.2014.09.039>.
- Famili A, Shen W-M, Weber R, Simoudis E. Data preprocessing and intelligent data analysis. *Intell Data Anal*. 1997;1:3–23. [https://doi.org/10.1016/S1088-467X\(98\)00007-9](https://doi.org/10.1016/S1088-467X(98)00007-9).
- Fan YZ, Dong JX, Zhang Y, Li N, Liu SG, Geng S, et al. A smartphone-coalesced nanoprobe for high selective ammonia sensing based on the pH-responsive biomass carbon nanodots and headspace single drop microextraction. *Spectrochim Acta A*. 2019;219:382–90. <https://doi.org/10.1016/j.saa.2019.04.073>.
- Franson MAH. In: Greenberg AE, Clesceri SL, Eaton AD, editors. APHA method 4500- $\text{NO}_3^-$ : standard methods for the examination of water and wastewater. Washington, DC: American Public Health Association; 1992. p. 87–94.
- George WLJ. Official methods of analysis of AOACAOAC official method 97350. Rockville: AOAC International; 2012.
- González-Miret ML, Terrab A, Hernanz D, Fernández-Recamales MÁ, Heredia FJ. Multivariate correlation between color and mineral composition of honeys and by their botanical origin. *J Agric Food Chem*. 2005;53:2574–80. <https://doi.org/10.1021/jf048207p>.
- Harvey D. Analytical chemistry 2.0—an open-access digital textbook. *Anal Bioanal Chem*. 2011;399:149–52. <https://doi.org/10.1007/s00216-010-4316-1>.
- Hoang AT, Cho YB, Kim YS. A strip array of colorimetric sensors for visualizing a concentration level of gaseous analytes with basicity. *Actuators B Chem*. 2017;251:1089–95. <https://doi.org/10.1016/j.snb.2017.06.142>.
- Kim S-Y, Li J, Lim N-R, Kang B-S, Park H-J. Prediction of warmed-over flavour development in cooked chicken by colorimetric sensor array. *Food Chem*. 2016;211:440–7. <https://doi.org/10.1016/j.foodchem.2016.05.084>.
- Korifi R, Le Dréau Y, Jean Francois A, Valls R, Dupuy N. CIEL\*a\*b\* color space predictive models for colorimetry devices - analysis of perfume quality. *Talanta*. 2013;104C:58–66. <https://doi.org/10.1016/j.talanta.2012.11.026>.
- Krongchai C, Jakmunee J, Kittiwachana S. Development of colorimetric sensor array for instant determination of sodium metabisulfite in dried longan. *Food Anal Methods*. 2020;13:1717–25. <https://doi.org/10.1007/s12161-020-01797-1>.
- LaGasse MK, Rankin JM, Askim JR, Suslick KS. Colorimetric sensor arrays: interplay of geometry, substrate and immobilization. *Sensors Actuators B Chem*. 2014; 197:116–22. <https://doi.org/10.1016/j.snb.2014.01.102>.
- Morsy MK, Zór K, Kotesha N, Alström TS, Heiskanen A, El-Tanahi H, et al. Development and validation of a colorimetric sensor array for fish spoilage monitoring. *Food Control*. 2016;60:346–52. <https://doi.org/10.1016/j.foodcont.2015.07.038>.
- Peng B, Xu J, Fan M, Guo Y, Ma Y, Zhou M, et al. Smartphone colorimetric determination of hydrogen peroxide in real samples based on B, N, and S co-doped carbon dots probe. *Anal Bioanal Chem*. 2020;412:861–70. <https://doi.org/10.1007/s00216-019-02284-1>.
- Pla-Tolós J, Moliner-Martínez Y, Verdú-Andrés J, Casanova-Chafer J, Molins-Legua C, Campíns-Falcó P. New optical paper sensor for in situ measurement of hydrogen sulphide in waters and atmospheres. *Talanta*. 2016;156:79–86. <https://doi.org/10.1016/j.talanta.2016.05.013>.
- Ravindranath R, Periasamy AP, Roy P, Chen Y-W, Chang H-T. Smart app-based on-field colorimetric quantification of mercury via analyte-induced enhancement of the photocatalytic activity of  $\text{TiO}_2$ -Au nanospheres. *Anal Bioanal Chem*. 2018;410:4555–64. <https://doi.org/10.1007/s00216-018-1114-7>.
- Rice A, Baird EW, Eaton RB. APHA 2017 standard methods for examination of water and wastewater. Washington: American Public Health Association,

- American Water Works Association, Water environment federation ISBN; 2017.
- Saxena AK, Prathipati P. Comparison of MLR, PLS and GA-MLR in QSAR analysis\*. SAR QSAR Environ Res. 2003;14:433–45. <https://doi.org/10.1080/10629360310001624015>.
- Sharifzadeh S, Clemmensen LH, Borggaard C, Støier S, Ersbøll BK. Supervised feature selection for linear and non-linear regression of L\* a\* b\* color from multispectral images of meat. Eng Appl Artif Intell. 2014;27:211–27. <https://doi.org/10.1016/j.engappai.2013.09.004>.
- Tahir HE, Xiaobo Z, Xiaowei H, Jiyong S, Mariod AA. Discrimination of honeys using colorimetric sensor arrays, sensory analysis and gas chromatography techniques. Food Chem. 2016;206:37–43. <https://doi.org/10.1016/j.foodchem.2016.03.032>.
- Thajee K, Paengnakorn P, Wongwilai W, Grudpan K. Application of a webcam camera as a cost-effective sensor with image processing for dual electrochemical – colorimetric detection system. Talanta. 2018;185:160–5. <https://doi.org/10.1016/j.talanta.2018.03.055>.
- Upadhyay Y, Bothra S, Kumar R, Kumar Sk A, Sahoo SK. Mimicking biological process to detect alkaline phosphatase activity using the vitamin B6 cofactor conjugated bovine serum albumin capped CdS quantum dots. Colloids Surf B: Biointerfaces. 2020a;185:110624. <https://doi.org/10.1016/j.colsurfb.2019.110624>.
- Upadhyay Y, Bothra S, Kumar R, Sahoo SK. Smartphone-assisted colorimetric detection of Cr3+ using vitamin B6 cofactor functionalized gold nanoparticles and its applications in real sample analyses. ChemistrySelect. 2018;3:6892–6. <https://doi.org/10.1002/slct.201801289>.
- Upadhyay Y, Kumar R, Sahoo SK. Vitamin B6 cofactors conjugated ovalbumin-stabilized gold nanoclusters: application in alkaline phosphatase activity detection and generating white-light emission. Microchem J. 2020b;156:104859. <https://doi.org/10.1016/j.microc.2020.104859>.
- Upadhyay Y, Kumar R, Sahoo SK. Developing a cost-effective bioassay to detect alkaline phosphatase activity and generating white light emission from a single nano-assembly by conjugating vitamin B6 cofactors with lysozyme-stabilized fluorescent gold nanoclusters. ACS Sustain Chem Eng. 2020c;8:4107–13. <https://doi.org/10.1021/acssuschemeng.9b06563>.
- Wang K, Yang J, Xu H, Cao B, Qin Q, Liao X, et al. Smartphone-imaged multilayered paper-based analytical device for colorimetric analysis of carcinoembryonic antigen. Anal Bioanal Chem. 2020;412:2517–28. <https://doi.org/10.1007/s00216-020-02475-1>.
- Wongsapun S, Krongchai C, Jakmunee J, Kittiwachana S. Rice grain freshness measurement using rapid visco analyzer and chemometrics. Food Anal Methods. 2018;11:613–23. <https://doi.org/10.1007/s12161-017-1031-y>.
- Xiaowei H, Xiaobo Z, Jiewen Z, Jiyong S, Zhihua L, Tingting S. Monitoring the biogenic amines in Chinese traditional salted pork in jelly (Yao-meat) by colorimetric sensor array based on nine natural pigments. J Food Sci Technol. 2015;50:203–9. <https://doi.org/10.1111/jifs.12620>.
- Xiao-wei H, Zhi-hua L, Xiao-bo Z, Ji-yong S, Han-ping M, Jie-wen Z, et al. Detection of meat-borne trimethylamine based on nanoporous colorimetric sensor arrays. Food Chem. 2016;197:930–6. <https://doi.org/10.1016/j.foodchem.2015.11.041>.
- Zaragoza P, Fuentes A, Ruiz-Rico M, Vivancos J-L, Fernández-Segovia I, Ros-Lis JV, et al. Development of a colorimetric sensor array for squid spoilage assessment. Food Chem. 2015;175:315–21. <https://doi.org/10.1016/j.foodchem.2014.11.156>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)