Banana (Musa L. spp.) is one of the most commonly cultivated crops worldwide and is the top agricultural commodity in many countries (Yeturu et al., 2016; Food and Agriculture Organization of the United Nations, 2017). However, banana production is severely affected by black leaf streak disease (BLSD, also known as black Sigatoka disease), which is caused by the fungal pathogen Pseudocercospora fijiensis. BLSD is considered the most widespread and damaging disease affecting bananas worldwide, causing plant necrosis in six symptomatic stages (Bakache et al., 2019).

BLSD is characterized by a biotrophic phase followed by a necrotrophic phase with visible symptoms. The disease affects the photosynthetic tissues of banana leaves and decreases chlorophyll production (Chaerle et al., 2007), resulting in changes to the structure of the leaves. The first symptoms of the disease are small dark spots on the underside of the leaf that develop to form fine brown lines 2–3 mm long, which are also visible on the adaxial surface of the infected leaves. As the disease progresses, the stripes join together and gradually turn black, showing the first signs of necrosis. The dead zones of the leaves then dry out, causing defoliation and the early maturation of the fruit. The presymptomatic biotrophic phase can last for several weeks, and by the time symptoms are visible the banana plants are irreversibly affected and the disease has already spread (Marin et al., 2003), potentially inducing production losses of up to 85% (Luna-Moreno et al., 2019).
stages of BLSD (i.e., presymptomatic infected leaves and those in stages 1 and 2; see Table 1), the physical changes in the plant are minimal, making the visual identification of leaf damage difficult. Furthermore, asexual and sexual spores develop from stage 2 of the disease onward. Conidial (asexual) spores are waterborne over short distances, whereas ascospores (sexual spores) can be carried over long distances and are responsible for the spread of the disease; therefore, the early detection of BLSD and the timely application of fungicides is crucial for controlling a *P. fijiensis* infestation. Early treatments reduce production costs and improve the health of crops while using shorter treatment times.

Currently, the detection of BLSD relies on the visual observation of symptoms or the use of destructive analyses, such as those based on DNA or immunological assays (Luna-Moreno et al., 2019). To the best of our knowledge, hyperspectral image-based non-destructive methods have not yet been evaluated for the early detection of BLSD.

The optical properties of a leaf can be characterized by (1) the light transmission through the leaf, (2) the light absorbed by chemicals within the leaf (e.g., pigments, water, sugars, lignin, and amino acids), and (3) the light reflected by the leaf surface or internal structures. Light reflectance levels depend on complex biophysical and biochemical interactions within the leaves. Changes in the photosynthetic pigments are evident in the visible region (VIS, 400–750 nm wavelength), changes in leaf structure and the scattering process affect reflectance in the near-infrared region (NIR, 750–1350 nm), and the water content influences the reflectance in the mid-infrared region (MIR, 1350–2500 nm) (Hunt and Rock, 1989). Many reflectance-related changes have been observed in diseased plants (Mahlein et al., 2012). In banana plants, *P. fijiensis* destroys photosynthetic tissue, which induces an increase in fluorescence and heat emission in the leaf blade and modifies the transport pattern of the photoassimilates, affecting the production of chlorophyll (Hidalgo et al., 2006). The resulting necrotic and chlorotic lesions cause variations of reflectance in the VIS and NIR regions of the spectrum.

Structural and chemical changes occurring in leaves during pathogenesis have enabled disease detection using hyperspectral image analyses (Siche et al., 2016). Previous works on sugar beet (*Beta vulgaris* L. subsp. *vulgaris*) (Mahlein, 2011), wheat (*Triticum aestivum* L.) (Ashourloo et al., 2014), tomato (*Solanum lycopersicum* L.), and lettuce (*Lactuca sativa* L.) (Lara et al., 2013) have provided some insights into the relationship between pathogen infections and spectral variations in leaves. In banana plants, hyperspectral image research has mostly focused on fruit measurements (Hu et al., 2015), quality control (Intaravanne et al., 2012), and the differentiation of black Sigatoka from yellow Sigatoka disease (Bendini et al., 2015); however, the use of hyperspectral images for early detection of BLSD has not been evaluated.

Data from hyperspectral leaf images usually present both spatial and spectral dimensions showing (1) high collinearity in the adjacent bands, (2) variability of hyperspectral signatures, and (3) high dimensionality due to the increased sensitivity and resolution of hyperspectral sensors (Lu and Fei, 2014). It is therefore necessary to apply multivariate data processing methods to be able to correlate hyperspectral fingerprints to plant infections.

Partial least squares (PLS) is a dimension reduction technique that is useful when the number of variables is greater than the number of observations. This tool maximizes the covariance between dependent and independent (predictors) variables, extracting from the predictors a set of orthogonal latent factors that are linear combinations of the original variables with the best predictive power (Abdi, 2010). When the response variable is categorical, PLS discriminant analysis (PLS-DA) has been used by other researchers (Breteron and Lloyd, 2014). PLS-DA is a linear classifier and its objective is to find a straight line that separates the regions. In this paper, we use an alternative method based on the algorithm proposed by Bastien et al. (2005), which takes into account the binary nature of the response using logistic regression rather than linear regression. This method is complemented with a hyperspectral biplot (HS biplot) that generalizes the proposal of Oyedele and Lubbe (2015), whose method was based on linear PLS components, whereas our study uses logistic PLS components. Additionally, penalized logistic regressions (PLR) have been introduced into the main algorithm to avoid the separation problems that occur when positive groups are presented separately to the negative groups, which prevents finding the maximum likelihood estimators (Albert and Anderson, 1984). This PLS–PLR model has not previously been used for the early detection of BLSD from hyperspectral images.

The objective of this study was to assess the use of hyperspectral imaging for the early detection of BLSD. The PLS–PLR technique was used to classify banana leaves, and was complemented by an HS biplot representation that facilitates the observation of the influence of the disease-associated changes in the reflectance of artificial light irradiated on a banana leaf.

### METHODS

#### Plant inoculation

Banana plants (*Musa acuminata*, AAA Group, Cavendish sub-group, cultivar ‘Williams’) were obtained from commercial banana propagation facilities (Sociedad Ecuatoriana de Biotecnología C.A. [SEBIOCA], Guayaquil, Ecuador). A total of 100 plants that had been established for 3–4 months in a greenhouse were transported to our greenhouse located at the Centro de Investigaciones Biotecnológicas del Ecuador (Escuela Superior Politécnica del Litoral [ESPOL]) and grown at 28°C in 70% relative humidity and 12 hours of natural light, and were watered every 48 hours. Sixteen plants were randomly selected from this population, of which 10 were inoculated with *P. fijiensis* and six were mock-inoculated for use as a control. The inoculation was carried out as reported by Gbongue et al. (2019). Briefly, isolates of *P. fijiensis* were inoculated onto potato dextrose agar and incubated for two weeks at 30°C. The mycelium was then ground in 10 mL of sterile water and filtered to separate the mycelium from the conidia. The conidial suspension

| Stage | Symptoms |
|-------|----------|
| 1     | Yellowish spots <1 mm in diameter on the abaxial leaf surface |
| 2     | Red or brown streaks from 1 to 5 mm |
| 3     | Similar to stage 2, but streaks are >5 mm |
| 4     | Brown elliptical streaks on the abaxial leaf surface, black streaks on the adaxial leaf surface |
| 5     | The streak is totally black and has spread to the abaxial leaf surface. The streak is surrounded by a yellow halo. |
| 6     | The center of the streak is light gray surrounded by a black ring and a yellow halo. |

TABLE 1. Severity scale of black leaf streak disease (BLSD).
was concentrated by centrifugation at 3000 × g for 10 min at 4°C. Banana leaves were spray-inoculated with the concentrated conidia suspension using an aerograph atomizer (Gerensa, Guayaquil, Ecuador), and disease symptoms were monitored using the severity scale shown in Table 1, as suggested by Fouré (1986). The visual symptoms at each disease stage are shown in Fig. 1. The control plants were mock inoculated with autoclaved distilled water.

**Imaging**

The imaging system used in our experiments (Appendix S1) was exactly as described by Ochoa et al. (2016) and included a spectrometer (ImSpector V10E, Specim, Oulu, Finland) connected to a 12 bits per pixel camera (1500M-GE; Thorlabs Inc., Newton, New Jersey, USA) with high sensitivity in the NIR region. The camera was mounted on a slider in a push-broom configuration and controlled by a computer with a storage capacity of 1 TB, an Intel Core i5 3.1-GHz processor, and 16 GB of RAM. The system operated in a spectral range between 386–1019 nm, with a spectral resolution of 4.55 nm and a spatial resolution of 1040 (rows) by 1392 (columns).

The system allowed the nondestructive scanning of individual leaves from intact plants. Each hyperspectral image was composed of three dimensions, including one spectral and two spatial dimensions, resulting in a hyperspectral cube. The spatial dimensions consisted of the position of each pixel on the image (M × N), whereas the spectral dimension was made of the wavelength of reflected light (J). The system generated a hyperspectral cube for each leaf, with a resolution of 205 rows (M), 198 columns (N), and 520 wavelengths (J). To obtain the width of a hyperspectral cube (N), we set a pixel binning-x camera of 7, resulting in a reduction of 1392 pixels to 198 pixels. To calculate the number of wavelengths in a hyperspectral cube (J), we set a pixel binning-y camera of 2, resulting in a reduction of 1040 pixels to 520 pixels. Finally, the height of a hyperspectral cube (M) was determined according to the acquisition frame rate of the camera to scan the entire leaf holder, resulting in 205 pixels.

The plants were placed in the imaging system and the selected leaves were tagged and scanned. Three leaves were selected from each of the six control plants, but two were damaged due to manipulation, resulting in 16 images of non-infected leaves. Two leaves were selected from each of the 10 inoculated plants. The leaves were scanned every three days for three months. At every scan, the infection symptoms were visually assessed by an expert using the symptoms scale, as detailed in Table 1. During this period, the disease progression in the leaves was unequal. The level 1 symptoms appeared between seven and 31 days after the inoculation and increased irregularly, reaching higher severity levels in different time periods. In some leaves, the disease reached a severity level of 5. Due to manipulation, several leaves were damaged and were therefore discarded during the experiment.

![FIGURE 1](http://www.wileyonlinelibrary.com/journal/AppsPlantSci)  
**FIGURE 1.** Stages of black leaf streak disease (BLSD): (A) non-infected, (B) stage 1, (C) stage 2, (D) stage 3, (E) stage 4, and (F) stage 5. Pen marks highlight the affected areas.
From the images scanned and tagged by experts, we selected those that belonged to infected leaves at the presymptomatic stage and stages 1 and 2 of BLSD. The presymptomatic images (16) were obtained six days before the leaf presented symptoms at severity level 1. The following images were taken in intervals of six days, during the progression of severity level 1 (54) and severity level 2 (18).

The final data set consisted of 104 images (16 non-infected, 16 presymptomatic, 54 severity level 1, and 18 severity level 2). The severity level reported corresponded to the highest disease stage found in the leaf.

Data preprocessing

Standard image calibration methods were applied. First, spectral calibration was applied using a mercury (Hg), argon (Ar), helium (He), and hydrogen (H) light sources to estimate the wavelength corresponding to each charge-coupled device (CCD) line. Second, radiometric calibration was applied to reduce the influence of light intensity variation and the noise of the CCD sensor. For this purpose, white and dark references were imaged before each scanning session and the raw spectral image was normalized. In addition, the spatial calibration of the slider was performed to avoid overlap between leaf regions when the images were acquired.

To retain only the leaf pixels of each hyperspectral cube, we generated a mask using the image at a wavelength of 700 nm, and setting the leaf holder and background pixels to zero. Next, each hyperspectral cube was normalized using the standard normal variate technique (Barnes et al., 1989). This step is necessary to compensate for reflectance variation caused by the relative orientation of the leaf surface and the sensor. Finally, the dimensionality reduction was achieved by calculating the average of the reflectance values measured at each wavelength of the hyperspectral cube, resulting in a matrix of I = 104 rows (one per image) and J = 520 columns (Appendix S2). A preliminary analysis of infected regions showed differences in the spectral patterns of the disease stages. Figure 2 shows the reflectance patterns for severity level 2 and non-infected regions.

Statistical methods

PLS–PLR model and HS biplot—The data were arranged into one vector and one matrix. The Y vector consisted of the binary (infected and non-infected) response variable, whereas the X matrix (X_1, …, X_j) consisted of a set of predictors that correspond to the reflectance intensity at each wavelength.

The general logistic PLS model and its associated biplot were included in the MultiBiplotR package (Vicente-Villardón, 2017). The package implements the algorithm that estimates PLS components and applies a logistic regression on the PLS components and the response vector Y. To prevent the separation problem reported in previous studies (Albert and Anderson, 1984; Santner and Duffy, 1986) and detected in our initial tests, the package applies a ridge penalty (Le Cessie and Van Houwelingen, 1992), which is calculated by the sum of the squares of the coefficients (L2 norm) multiplied by a penalty parameter λ. The parameter λ can have a value between 0 and 1, and can be adjusted by cross-validation. For the purpose of finding the model that best describes the data, we tested values of λ in the [0.1–0.9] range using incremental steps of 0.1 and calculated, for each value of λ, the following goodness-of-fit measures: difference of deviance (DiffDeviance), Cox and Snell’s R^2, Nagelkerke’s R^2, and McFadden’s R^2 (Allison, 2014; Walker and Smith, 2016).

The biplot that represents the labeled leaves, the wavelengths, and the boundary of the prediction regions (i.e., the HS biplot) allows for the visual inspection of the PLS–PLR model. The wavelengths were represented by lines colored according to the spectral band to which they belong. The statistical procedures applied in PLS–PLR and HS biplot are detailed in Appendix 1.

The PLS–PLR model yielded estimated values between 0 and 1. Leaves with model values above or below 0.5 were considered to be predicted-infected or predicted-healthy, respectively. The goodness-of-fit of the logistic regression was calculated using the pseudo R^2 measures described above.

RESULTS

Penalty parameter (λ) selection and model goodness-of-fit

First, the penalty parameter of the PLS–PLR model was selected from different λ coefficients. The goodness-of-fit values obtained after each interaction are shown in Table 2.

We obtained the best goodness-of-fit measures (highest DiffDeviance value of 88.488) using a λ value of 0.1, showing that the fitted model maintained the greatest variance. Furthermore, the P value of 6097 × 10^−20 shows a significant association between the latent variables and the response variable. The pseudo R^2 measures were also estimated; McFadden’s R^2 (0.991) indicated a high explicative power of the model, while the Cox and
Snell’s $R^2$ (0.573) and Nagelkerke’s $R^2$ (0.994) also indicated a high goodness-of-fit. The penalty $\lambda = 0.1$ was applied in each iteration for calculating the coefficients, providing a stable value that maximized the likelihood function while controlling the error. Figure 3 shows the logistic response surface fitted on the space spanned by the first two PLS components. The model is:

$$P_y = \frac{e^{(27.226 + 1.546t_1 + 1.318t_2)}}{1 + e^{(27.226 + 1.546t_1 + 1.318t_2)}}$$

where $P_y$ is the infected leaf probability, $t_1$ is the first PLS–PLR component, and $t_2$ is the second PLS–PLR component.

**HS biplot**

The HS biplot of the training data set is shown in Fig. 4. The first two PLS components contributed 77% of the observed variability. The HS biplot indicates three main groupings: a group of non-infected leaves (blue ellipse) and two other groups of infected leaves (red ellipses). The clustering of non-infected and infected plants was mostly observed in component 1. The wavelengths that contributed the most to the grouping of the non-infected samples ranged from 577 to 651 nm (yellow to red range). The first of the two groups of infected leaves was located near the non-infected group (numbered points inside the red ellipse in Fig. 4) in the HS biplot, mainly influenced by a low density of disease symptoms in the leaves (Table 3). The second group of infected leaves was composed of both presymptomatic and symptomatic leaves. The wavelengths that contributed the most to the grouping of the presymptomatic leaves (turquoise), as well as several leaves at severity levels 1 (blue) and 2 (red), were in the NIR range of the spectrum. However, the grouping of the other leaves at symptomatic levels 1 and 2 was determined by their reflectance in the 577 to 651 nm (yellow to red) range. This was also observed in the external validation data set (Fig. 5). The HS biplot of the test data set showed that the healthy (blue ellipse) and infected (red ellipse) leaves had all been correctly classified, with the exception of two samples, one healthy and one infected.

**Model prediction and validation**

The model was validated using the leave-one-out cross-validation (LOOCV) method, as shown in Appendix S3. During the cross-validation process, 102 leaves were properly classified as non-infected or infected, which represented an overall classification accuracy of 98% (Table 4). Only two non-infected samples were classified as being infected, with a probability of 0.738 and 0.816, respectively. All infected leaves at the different disease stages were correctly classified. The positive predictive value was 98%, while the sensitivity or recall value was 100% (Table 4), indicating that all infected leaves were correctly identified. The estimated global HS biplot goodness-of-fit (the squared correlation coefficient between the adjusted and observed values) was 77.07%.

**External validation**

The PLS–PLR model fitted to the initial training data set was used to predict the presence of the disease in new leaves. A new data set with images of 16 non-infected and 21 infected leaves was used to evaluate the efficacy of the model. The prediction accuracy of the new samples was 95% (35 successful identifications and two errors). Appendix S4 indicates the predicted probability for the external validation test.
Agricultural crops are constantly threatened by pests and plant diseases that reduce production and quality. New agricultural strategies and techniques to identify plant diseases are based on non-destructive methods that detect the disease in early stages, allowing timely management to prevent the spread of the disease and minimize the effect of fungicides in the environment. Remote sensing allows early detection of plant diseases using methods based on reflectance in the VIS and NIR regions. Disease detection involves steps such as image acquisition, image pre-processing, image segmentation, feature extraction, and classification.

In this study, hyperspectral images were used to classify BLSD-infected and healthy banana plants based on the marked differences in the VIS and NIR spectra of both groups. Our results are consistent with previous reports showing that disease symptoms in plants can change spectral reflectance in the VIS (400–700 nm) and NIR (700–1100 nm) ranges (Ayala-Silva and Beyl, 2005). General changes in reflectance occurring during plant–pathogen interactions have been associated with impairments in the leaf structure and chemical composition of the tissue during pathogenesis, which can be observed by the succession from chlorotic to necrotic tissue (Mahlein, 2016).

PLS was originally developed for continuous response variables. In the case of binary responses, a linear regression cannot guarantee that response-fitted values fall at 0 or 1. In this study, we used logistic rather than linear regressions to correlate the response variable with the PLS components. Previous studies on plant disease detection have reported relatively high accuracies and sensitivities similar to PLS models, but the model interpretability is low, and most studies did not include leaves in presymptomatic stages (Rumpf et al., 2010; Mahlein, 2011; Zhu et al., 2017). In our study, the application of PLS–PLR showed a prediction accuracy of 98% in the presymptomatic and early stages of BLSD. The goodness-of-fit measures obtained for the PLS–PLR model were similar to those reported in the literature (Huang et al., 2007; Bock et al., 2010; Cevallos-Cevallos et al., 2018).

As indicated by the HS biplot, which shows leaves as points and wavelengths as lines, the non-infected leaves were characterized by the prominent presence of wavelengths in the VIS spectrum, whereas infected leaves were associated with wavelengths in both the VIS and NIR ranges, depending on the stage of the disease or the proportion of symptomatic to healthy tissue. These results are in agreement with previous studies, as disease
symptoms in plants have been reported to increase spectral reflectance in both the VIS (400–700 nm) and NIR (700–1100 nm) ranges (Ayala-Silva and Beyl, 2005). General changes in reflectance occurring during plant–pathogen interactions have been associated with impairments in the leaf structure and changes in the chemical composition of the tissue during pathogenesis, which can be observed by the succession of chlorotic and necrotic tissue (Mahlein, 2016).

The clustering of non-infected and infected plants was mostly observed within PLS component 1. The wavelengths that contributed the most to PLS component 1 ranged from 577 to 651 nm (yellow to red). Changes in the yellow range of the spectrum suggest that detection of plant chlorosis occurs in the initial stages of BLSD. Chlorosis is caused by insufficient chlorophyll accumulated in the plants, leading to the yellowing of the leaf, which is usually measured with yellowness indexes (Adams et al., 1999). Similarly, changes in the orange–red range of the spectrum suggest the succession of chlorotic to necrotic tissue, as observed in the red or brown streaks that appear in stage 2 of BLSD (Fouré, 1986). Interestingly, presymptomatic plants showing neither chlorosis nor necrosis were clustered apart from the non-infected ones, suggesting changes in the leaf surface of these infected plants despite the lack of visible symptoms. Biotrophic presymptomatic stages do not usually cause observable changes in leaves, but some fungal pathogens can produce structures on the leaf surface that can influence the optical properties of the plant (Mahlein, 2016). Non-infected or infected banana plants were scattered in two groups in the HS biplot, which is in agreement with previous reports showing two groups of non-infected and dwarf banana plants due to the natural biological diversity observed in agricultural conditions (Cevallos-Cevallos et al., 2018). These results confirm the high prediction capacity of the PLS–PLR model and the efficiency of the HS biplot to represent the relationships between the variables and the groups of individuals with non-infected and infected leaves.

Hyperspectral imaging provides a non-destructive method for analyzing plants infected with BLSD and possibly other pathogens. The development of technologies offering larger data storage capacities, faster computers, more sensitive detectors, and different analytical techniques for hyperspectral images, combined with suitable statistical techniques, makes it possible to detect plant diseases even at early stages, and enables the capture and modeling of physiological changes of leaves using close-range hyperspectral data. The early detection of infectious diseases plays a crucial role in both treatment and prevention strategies.

TABLE 4. Confusion matrix for the classification accuracy assessment of the PLS–PLR model.

| True infected leaf | False non-infected leaf | Prediction metrics |
|--------------------|-------------------------|-------------------|
| Test result        |                         | Positive predictive value |
| True positive      | False positive          | 0.98              |
| 88                 | 2                       |
| False negative     | True negative           | Negative predictive value |
| 0                  | 14                      |
| Prediction metrics |                         | Accuracy          |
| Sensitivity        | Specificity             | 0.88              |
| 1                  | 0.98                    |

FIGURE 5. HS biplot of the external validation data set. The diagonal dotted line separates predicted healthy (blue ellipse) and infected leaves (red ellipse). Each wavelength is represented by straight lines colored according to the electromagnetic spectrum colors. PLS component 1 explains 50.68% of variance, while PLS component 2 explains 26.21% of the variance.
of the raw data. The PLS–PLR model provides excellent predictive power, which is complemented by the high level of visual interpretation offered by the HS biplot.

The system presented here combines hyperspectral technology with advanced data analysis and statistical methods to accurately predict plant disease by measuring the reflectance differences resulting from the biophysical and biochemical characteristic changes following infection. It is currently configured as a stationary imaging system and could be used in a laboratory for the recognition of foliar diseases in other plants and for food quality evaluations. In the future, however, the system could be transformed, with few changes, into an airborne imaging system for scanning images of crop fields.

Future research should concentrate on improving the detection of different severity levels of the disease and on the more detailed analysis of the wavelengths that have greater influence, taking into account other factors that can cause spectral changes.

ACKNOWLEDGMENTS

This research was supported by VLIR - UOS grant “VLIR Network Ecuador.”

AUTHOR CONTRIBUTIONS

D.O.D. and J.C.C. formulated the research problem. D.O.D., J.C.C., J.L.V.V., and J.U.F. designed the approaches. D.O.D., R.C.B., and O.B.A. collected the data. J.L.V.V., D.O.D., J.C.C., M.M.Z., R.C.B., O.B.A., and J.U.F. developed the processing workflow. J.L.V.V. and J.U.F. performed the data analysis. All authors contributed to the writing and development of the manuscript. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The two data sets used in this study (i.e., Training data set [data_banana.txt] and Validation data set [data_banana_test.txt]) are available at: https://drive.google.com/drive/folders/1jTHKICPeEdxofiTY_W3YhbV_LBHsBldRL?usp=sharing.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. Hyperspectral imaging system used in this study: (A) camera, (B) spectrometer, (C) light source, (D) slider, (E) holder.

APPENDIX S2. Dimensionality reduction of the hyperspectral cubes. The mean of the reflectance matrix at each wavelength is calculated.

APPENDIX S3. Predicted probability using leave-one-out cross-validation (LOOCV) to assess the PLS–PLR model.

APPENDIX S4. Predicted probability for the external validation of the PLS–PLR model.

LITERATURE CITED

Abdi, H. 2010. Partial least squares regression and projection on latent structure regression (PLS regression). Wiley Interdisciplinary Reviews: Computational Statistics 2(1): 97–106.

Adams, M. L., W. D. Phlipot, and W. A. Norvell. 1999. Yellowness index: An application of spectral second derivatives to estimate chlorosis of leaves in stressed vegetation. International Journal of Remote Sensing 20(18): 3663–3675.

Albert, A., and J. A. Anderson. 1984. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71(1): 1–10.

Allison, P. D. 2014. Measures of fit for logistic regression. In Proceedings of the SAS Global Forum, 1–13. Washington, D.C., USA.

Ashourloo, D., M. Mobasheri, and A. Huete. 2014. Evaluating the effect of different wheat rust disease symptoms on vegetation indices using hyperspectral measurements. Remote Sensing 6(6): 5107–5123.

Ayala-Silva, T., and C. A. Beyl. 2005. Changes in spectral reflectance of wheat leaves in response to specific macronutrient deficiency. Advances in Space Research 35(2): 305–317.

Bakache, A., J.-P. Douzals, B. Bonicelli, E. Cotteux, L. de Lapeyre de Bellaire, and C. Sinforte. 2019. Development of a rapid methodology for biological efficacy assessment in banana plantations: Application to reduced dosages of contact fungicide for black leaf streak disease (BLSD) control. Pest Management Science 75(4): 1081–1090.

Barnes, R. J., M. S. Dhanoa, and S. J. Lister. 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy 43(5): 772–777.

Bastien, P., V. E. Vinzi, and M. Tenenhaus. 2005. PLS generalised linear regression. Computational Statistics and Data Analysis 48: 17–46.

Bendini, H., A. Jacon, A. C. Moreira Pessôa, J. A. Pompeu Pavanelli, W. da Silva Moraes, F. J. Ponzonii, and L. Fonseca. 2015. Spectral characterization of banana leaves (Musa spp.) for detection and differentiation of black sigatoka and yellow sigatoka. XVII Simpósio Brasileiro de Sensoriamento Remoto, João Pessoa, Brazil.

Bock, C., G. Poole, P. Parker, and T. Gottwald. 2010. Plant disease severity estimated visually; by digital photography and image analysis, and by hyperspectral imaging. Critical Reviews in Plant Sciences 29(2): 59–107.

Breton, R. G., and G. R. Lloyd. 2014. Partial least squares discriminant analysis: Taking the magic away. Journal of Chemometrics 28: 213–225.

Cevallos-Cevallos, J. M., C. Jines, M. G. Maridueña-Zavala, M. J. Molina-Miranda, D. E. Ochoa, and J. A. Flores-Cedeno. 2018. GC-MS metabolite profiling for specific detection of dwarf somaclonal variation in banana plants. Applications in Plant Sciences 6(11): e01194.

Chaerle, L., L. Leinonen, H. G. Jones, and D. Van Der Straeten. 2007. Monitoring and screening plant populations with combined thermal and chlorophyll fluorescence imaging. Journal of Experimental Botany 58(4): 773–784.

Demey, J., J. L. Vicente-Villardón, M. Galindo-Villardón, and A. Zambrano. 2008. Identifying molecular markers associated with classification of genotypes by external logistic biplots. Bioinformatics 24(24): 2832–2838.

Food and Agriculture Organization of the United Nations. 2017. FAOSTAT statistical database. Website http://www.fao.org/faostat/en/Home [accessed 15 June 2018].

Fouré, E. 1986. Varietal reactions of bananas and plantains to black leaf streak disease. Banana and Plantain Breeding Strategies 21: 110–113.

Fu, J. W. 1998. Penalized regressions: The bridge versus the lasso. Journal of Computational and Graphical Statistics 7(3): 397–416.

Gbowongue, L.-R., I. Lalaymia, A. Zeze, B. Delvaux, and S. Declerck. 2019. Increased silicon acquisition in bananas colonized by Rhizophagus irregularis MUCL 41833 reduces the incidence of Pseudocercospora fijiensis. Frontiers in Plant Science 9: 1977.

Hidalgo, M., A. Tapia, W. Rodriguez, and E. Serrano. 2006. Efecto de la Sigatoka negra (Mycosphaerella fijiensis) sobre la fotosíntesis y transpiración foliar del banano (Musa sp. AAA, cv. Valery). Agronomía Costarricense 30(1): 35–41.

Hosmer, D. W., T. Hosmer, S. Le Cessie, and S. Lemeshow. 1997. A comparison of goodness-of-fit tests for the logistic regression model. Statistics in Medicine 16(9): 965–980.

Hu, M.-H., Q.-L. Dong, P. K. Malakar, B.-L. Liu, and G. K. Jaganathan. 2015. Determining banana size based on computer vision. International Journal of Food Properties 18(3): 508–520.
Huang, W., D. W. Lamb, Z. Niu, Y. Zhang, L. Liu, and J. Wang. 2007. Identification of yellow rust in wheat using in-situ spectral reflectance measurements and airborne hyperspectral imaging. Precision Agriculture 8(11): 187–197.

Hunt Jr., E. R., and B. N. Rock. 1989. Detection of changes in leaf water content using near- and middle-infrared reflectances. Remote Sensing of Environment 30(1): 43–54.

Intaravanne, Y., S. Sumriddetchkajorn, and J. Nukeaw. 2012. Ripeness level indication of bananas with visible and fluorescent spectral images. In 9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 1–4. IEEE, New York, New York, USA.

Lara, M., B. Diezma Iglesias, L. Lleó García, J.-M. Roger, Y. Garrido, M. Gil, and M. Ruiz-Alisent. 2013. Aplicación de imagen hiperspectral para observar el efecto de la salinidad en hojas de lechuga. In VII Congreso Ibérico de Agroingeniería y Ciencias Horticolas, Madrid, Spain.

Le Cessie, S., and J. C. Van Houwelingen. 1992. Ridge estimators in logistic regression. Journal of the Royal Statistical Society: Series C (Applied Statistics) 41(1): 191–201.

Lu, G., and B. Fei. 2014. Medical hyperspectral imaging: A review. Journal of Biomedical Optics 19(1): 010901.

Luna-Moreno, D., A. Sánchez-Alvarez, I. Islas-Flores, B. Canto-Canche, M. Carrillo-Pech, J. F. Villarreal-Chiu, and M. Rodríguez-Delgado. 2019. Early detection of the fungal banana black Sigatoka pathogen Pseudocercospora fijiensis by an SPR immunosensor method. Sensors 19(3): 465.

Mahlein, A.-K. 2011. Detection, identification, and quantification of fungal diseases of sugar beet leaves using imaging and non-imaging hyperspectral techniques. Doctoral thesis, Rheinischen Friedrich-Wilhelms-Universität Bonn, Bonn, Germany.

Mahlein, A.-K. 2016. Plant disease detection by imaging sensors–parallels and specific demands for precision agriculture and plant phenotyping. Plant Disease 100(2): 241–251.

Mahlein, A., E. Oerke, U. Steiner, and H.-W. Dehne. 2012. Recent advances in sensing plant diseases for precision crop protection. European Journal of Plant Pathology 133: 197–209.

Marin, D. H., R. A. Romero, M. Guzmán, and T. B. Sutton. 2003. Black Sigatoka: An increasing threat to banana cultivation. Plant Disease 87(3): 208–222.

Ochoa, D., J. Cevallos, G. Vargas, R. Cricollo, D. Romero, R. Castro, and O. Bayona. 2016. Hyperspectral imaging system for disease scanning on banana plants. In Sensing for Agriculture and Food Quality and Safety VIII, vol. 9864, 98640M. International Society for Optics and Photonics, Bellingham, Washington, USA.

Oyedele, O. F., and S. Lubbe. 2015. The construction of a partial least-squares biplot. Journal of Applied Statistics 42(11): 2449–2460.

Rumpf, T., A.-K. Mahlein, U. Steiner, E.-C. Oerke, H.-W. Dehne, and L. Plumer. 2010. Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. Computers and Electronics in Agriculture 72(11): 91–99.

Sansntr, T., J., and D. E. Duffy. 1986. A note on A. Albert and J. A. Anderson’s conditions for the existence of maximum likelihood estimates in logistic regression models. Biometrika 73(3): 755–758.

Siche, R., R. Vejarano, Y. Aredo, L. Velasquez, E. Saldaña, and R. Quevedo. 2016. Evaluation of food quality and safety with hyperspectral imaging (HSI). Food Engineering Reviews 8(3): 306–322.

Vicente-Villardón, J. L. 2017. MultiBiplotR: Multivariate analysis using biplots. R package version 0.1.0. Website http://bitplot.usa/multibiplot/multibiplot -in-R/ [accessed 20 November 2019].

Vicente-Villardón, J. L., M. P. Galindo-Villardón, and A. Blázquez-Zabalbos. 2006. Logistic correspondence analysis. In M. Greenacre, and J. Blasius [eds.], Multiple correspondence analysis and related methods. Chapman and Hall/CRC Press, Boca Raton, Florida, USA.

Walker, D. A., and T. J. Smith. 2016. Nine pseudo R² indices for binary logistic regression models. Journal of Modern Applied Statistical Methods 15(1): 848–854.

Yeturu, S., K. Méndez, P. Garrido, S. Serrano, and A. Garrido. 2016. Serological and molecular identification of cucumber mosaic virus (CMV) infecting banana crops in Ecuador. Ecuador Es Calidad: Revista Científica Ecuatoriana 3: 17–22.

Zhu, H., B. Chu, C. Zhang, F. Liu, L. Jiang, and Y. He. 2017. Hyperspectral imaging for presymptomatic detection of tobacco disease with successive projections algorithm and machine-learning classifiers. Scientific Reports 7: 4125.

APPENDIX 1. Partial least-squares penalized-logistic-regression (PLS–PLR) method applied to classify BLSD-infected and healthy banana leaves.

Let Y be the binary response variable and \((X_1, \ldots, X_p)\) a set of predictors. For a sample of size \(n\), the data can be organized in a response vector \(y = (y_1, \ldots, y_n)^T\) and a matrix of predictors \(X = (x_{ij}) (i = 1, \ldots, n; j = 1, \ldots, p)\), where \(y_i\) is either 0 or 1 for the presence or absence of the main characteristic and \(x_{ij}\) is the value of the \(i\)th individual on the \(j\)th predictor. The columns of \(X\) are supposed to be centered and possibly standardized.

We are searching for components that are linear combinations of the predictors and that best explain the response (in a logistic regression manner). Let \(t_h\) the vector containing the scores of each individual on one of those combines components, then \(\tilde{t}_h = \sum_{p=1}^{p} w_{hp} x_j = X w_h\) with \(w_h = (w_{h1}, \ldots, w_{hp})^T\) being the vector of coefficients. Normally we will use \(w\) of those components that are mutually orthogonal.

The logistic PLS regression model is written as

\[
E(y) = \tilde{y} = \frac{1}{1 + e^{-(c_1 + \sum_{h=1}^{m} c_h t_h)}},
\]

or

\[
\logit(\tilde{y}) = \log \left( \frac{\tilde{y}}{1 - \tilde{y}} \right) = c_0 + \sum_{h=1}^{m} c_h t_h.
\]

Or in matrix form \(\logit(\tilde{Y}) = c_01 + Tc\), where \(\tilde{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T\) is the vector of fitted probabilities of the presence of the disease in each individual and \(c = (c_1, \ldots, c_m)^T\) are the coefficients of the regression on the components. The model is a standard logistic regression on the PLS components. The constant \(c_0\) must be kept because the binary variable cannot be centered. In terms of the original variables,

\[
\logit(Y) = c_01 + Tb = c_01 + XWc = c_01 + Xb
\]

where \(W = (w_1, \ldots, w_m)\) and \(b = (b_1, \ldots, b_p)^T\) are the coefficients on the observed variables. The problem now is to estimate \(T, W, c, c_0, \) and \(b\).

**Estimation algorithm**

Here, we use the algorithm developed by Bastien et al. (2005), with slight modifications.

1. Calculation of \(t_1\), the first PLS component.
   a. For each predictor \((j = 1, \ldots, p)\), compute the regression coefficient \(w_{1j}\) of \(x_{1j}\) in the logistic regression of \(y\) on \(x_{1j}\) to obtain \(w_{1j} = (w_{1j1}, \ldots, w_{1jp})^T\).
   b. Normalize the vector \(w_{1j} = w_{1j} / \|w_{1j}\|\)
   c. Compute the component scores \(t_{1j} = X w_{1j} / \|w_{1j}\|^2\)

2. Calculation of \(t_h\), the \(h\)th PLS component. The components \(t_1, \ldots, t_{h-1}\) have been already obtained.
   a. For each predictor \((j = 1, \ldots, p)\), compute the regression coefficient \(w_{hj}\) of \(x_{1j}\) in the logistic regression of \(y\) on \(t_1, \ldots, t_{h-1}\) and \(x_{1j}\) to obtain \(w_{hj} = (w_{hj1}, \ldots, w_{hp})^T\).
b. Normalize the vector $w_h = w_h / \|w_h\|

c. Compute the residual matrix $X_{h-1}$ of the linear regression of $X$ on $t_1, \ldots, t_{h-1}$.
d. Compute the component scores $t_h = Xw_h / w_h^T w_h$.
3. $X$ is factorized as $X = TP$.
4. Logistic regression of $y$ on the retained PLS components
$$
\text{logit}(\hat{y}) = c_0 + \sum_{h=1}^{m} c_h t_h
$$
5. Expression of the model in terms of the original predictors $b = Wc$.

**Remarks**

1. Although the original algorithm is not clear, we think that all the logistic models in steps 1a, 2a, and 3 must include a constant.
2. When the model is good, i.e., when it is able to discriminate almost perfectly among presences and absences, the maximum likelihood method for logistic regression does not converge and the estimators tend to infinity. This is known as the separation problem (Albert and Anderson, 1984; Santner and Duffy, 1986), and is easily solved using a penalty. Here, we use the ridge (Le Cessie and Van Houwelingen, 1992) because of its simplicity. Additionally, when there are many variables that influence the response, as is the case of analysis of this work, this penalty (ridge) offers better results, whereas the lasso penalty (Fu, 1998) is more efficient when fewer variables influence the result. The ridge penalty is equal to the sum of the squares of the coefficients ($L2$ norm) by the penalty parameter $\lambda$. The parameter $\lambda$ may have a value between 0 and 1 and can be adjusted by cross-validation.

PLS–PLR technique was selected and implemented to resolve the bias, overfitting, data separation, and multicollinearity problems. Dimensionality reduction by PLS eliminates perfect or quasi-perfect correlation between predictors. Furthermore, it makes use of the output variable, which reduces the bias, and therefore avoids underfitting. A ridge penalty in logistic regression limits the growth of the regression coefficients, which reduces variance, avoids overfitting, and controls the effects of data separation.

**Goodness of fit of the model**

The goodness-of-fit measures used are described below.

**Deviance difference**—The deviance difference (DiffDeviance) is interpreted as a measure of the variation of the data explained by the model with predictors and the model without predictors (only the constant). This statistic has a chi-square distribution with degrees of freedom equal to the difference in the numbers of the model parameters. Thus, the null hypothesis will be rejected for the significance level $\alpha$ when $\text{DiffDeviance} > \chi^2$, which is equivalent to the $P$ value of the contrast being less than the fixed level of $\alpha$ (Hosmer et al., 1997).

$$
\text{DiffDeviance} = 2LL_M(\beta) - 2LL_0(\beta)
$$

where $LL_M(\beta)$ is the log likelihood of the model and $LL_0(\beta)$ is the log likelihood of the null model.

**Pseudo $R^2$**—Pseudo $R^2$ tells us how well the model can explain/predict the dependent variable based on the independent variables. Several different values were calculated as detailed below:

1. McFadden’s $R^2$ is defined as one minus the ratio between the logarithms of the likelihood for the model with respect to the log likelihood for the intercept only model (null model), with its theoretical range of values being $0 \leq \text{McFadden’s } R^2 \leq 1$. It is usually considered a good quality of fit when $0.2 \leq \text{McFadden’s } R^2 \leq 0.4$ and higher values show an excellent fit.

$$
R^2_{\text{McFadden}} = 1 - \left( \frac{LL_M}{LL_0} \right)
$$

where $LL_M$ is the log likelihood of the model and $LL_0$ is the log likelihood of the null model.

2. Cox and Snell’s $R^2$ is a goodness-of-fit measure that generalizes the $R^2$ of the linear regression. It is based on the comparison of the likelihood of the model ($L_M$) with the likelihood of the null model ($L_0$). Its range of values is between 0 and $(1 - L_0)^{2/n}$.

$$
R^2_{\text{Cox & Snell}} = 1 - \left( \frac{L_0}{L_M} \right)^{\frac{2}{n}}
$$

where $L_M$ is the likelihood of the model and $L_0$ is the likelihood of the null model.

3. Nagelkerke’s $R^2$ is the value of Cox and Snell’s $R^2$ that is standardized based on the maximum value it could take. The maximum value of this pseudo $R^2$ is therefore 1 (Allison, 2014; Walker and Smith, 2016).

$$
R^2_{\text{Nagelkerke}} = \frac{R^2_{\text{Cox & Snell}}}{1 - (L_0)^{2/n}}
$$

where $L_M$ is the likelihood of the model and $L_0$ is the likelihood of the null model.

**Hyperspectral biplot (HS biplot)**—Technically a biplot is a decomposition of a matrix $X$ in the product of two low-rank (usually two or three) matrices and an error matrix.

$$
X \approx TP + E
$$

in such a way that rows and columns can be jointly represented on a scatter diagram using $T$ and $P$ as markers, respectively. In this case, we use the factorization obtained from the logistic PLS regression as a biplot representation. The biplot shows the directions of the space spanned by the columns of $X$ that better separate the presences and absences for the dependent variable.

So, we have a low-rank approximation $\hat{X} = TP$ of the matrix $X$ that captures the part that better explains the response. The percent of variability captured by the approximation is

$$
\rho^2 = \frac{\text{tr}(\hat{X}^T \hat{X})}{\text{tr}(X^T X)} \times 100
$$

It is possible to identify the variables related to the PLS components calculating the amount of variance of each column captured by the approximation as follows:
\[ \rho_j = \frac{\text{tr} \left( \tilde{x}_{[j]}^T \tilde{x}_{[j]} \right)}{\text{tr} \left( x_{[j]}^T x_{[j]} \right)} \times 100 \quad (j = 1, \ldots, p) \]

where \( \tilde{x}_{[j]} \) and \( x_{[j]} \) are the \( j \)th columns of the fitted and original matrix, respectively. Only the columns with high percentages are related to the response. These quantities are called contributions of the components to the variables or predictiveness.

The row scores are also used to predict the binary response in step 4 of the algorithm, then the binary variable can also be projected on the biplot using an external logistic biplot, proposed by Demey et al. (2008) based on the proposal of Vicente-Villardón et al. (2006). The main difference is that, in the original proposal, the scores for the individuals are obtained from the principal coordinates, whereas here they are obtained from the logistic PLS regression. The vector \( \mathbf{c} \) of logistic regression coefficients defines the direction in the space spanned by the columns of \( T \), which better separates the presences and absences and the expected probabilities of having the presence of the characteristic

\[ \text{logit} \left( \hat{y}_i \right) = \text{logit} \left( \hat{P}_i \right) = c_0 + t_{[i]}^c, \]

where \( t_{[i]} \) is the \( i \)th row of \( T \). The expected probability is obtained by projecting the point \( t_{[i]} \) onto the vector \( \mathbf{c} \). The point on that direction that predicts an expected probability of 0.5 in a two-dimensional biplot has the coordinates

\[ x = \frac{-c_0 c_1}{c_1^2 + c_2^2}; \quad y = \frac{-c_0 c_2}{c_1^2 + c_2^2} \]

If we predict presence when the expected probability is greater than 0.5, the direction of \( \mathbf{c} \) divides the representation into two regions, one predicting the presence and the other predicting absence. The boundary of the two regions is a straight line perpendicular to \( \mathbf{c} \) and passing through the point \((x, y)\). For more details, see Demey et al. (2008) or Vicente-Villardón et al. (2006).

The goodness of fit of the logistic regression is measured using pseudo \( R^2 \) measures or the deviance. Although those measures are not completely adequate with the ridge penalizations, they can still be used as descriptive indicators equivalent to the contributions of the continuous biplot.