Plant Methods

RESEARCH

Open Access

Bayesian modelling of phosphorus content in wheat grain using hyperspectral reflectance data

Rosa Angela Pacheco-Gil[1*], Ciro Velasco-Cruz[3], Paulino Pérez-Rodríguez[3], Juan Burgueño[1], Sergio Pérez-Elizalde[3], Francelino Rodrigues[4], Ivan Ortiz-Monasterio[2], David Hebert del Valle-Paniagua[3] and Fernando Toledo[1]

Abstract

Background As a result of the technological progress, the use of sensors for crop survey has substantially increased, generating valuable information for modelling agricultural data. Plant spectroscopy jointly with statistical modeling can potentially help to assess certain chemical components of interest present in plants, which may be laborious and expensive to obtain by direct measurements. In this research, the phosphorus content in wheat grain is modeled using reflectance information measured by a hyperspectral sensor at different wavelengths. A Bayesian procedure for selecting variables was used to identify the set of the most important spectral bands. Additionally, three different models were evaluated: the first model assumes that the observations are independent, the other two models assume that the observations are spatially correlated: one of the proposed models, assumes spatial dependence using a Conditionally Autoregressive Model (CAR), and the other through an exponential correlogram. The goodness of fit of the models was evaluated by means of the Deviance Information Criterion, and the predictive power is evaluated using cross validation.

Results We have found that CAR was the model that best fits and predicts the data. Additionally, the selection variable procedure in the CAR model reveals which wavelengths in the range of 500–690 nm are the most important. Comparing the vegetative indices with the CAR model, it was observed that the average correlation of the CAR model exceeded that of the vegetative indices by 23.26%, − 1.2% and 22.78% for the year 2010, 2011 and 2012 respectively; therefore, the use of the proposed methodology outperformed the vegetative indices in prediction.

Conclusions The proposal to predict the phosphorus content in wheat grain using Bayesian approach, reflect with the results as a good alternative.

Keywords Bayesian statistics, Hyperspectral reflectance, Wheat, Spatial analysis, Wavelength, Phosphorus

Background

Wheat, like all plants, requires nutrients and macro nutrients for its development. A balanced contribution of these, leads to good grain yield and a good quality product. Phosphorus is a nutrient which is as a source of energy necessary for all metabolic processes in the wheat plant to take place. Its deficiency makes it impossible for the plant to complete these metabolic processes normally. The two critical moments in which its presence

*Correspondence:
Rosa Angela Pacheco-Gil
r.a.pacheco@cgiar.org
Full list of author information is available at the end of the article

is fundamental are: (1) in germination because it favors rapid root growth, and (2) in pre-flowering and growing because it provides the necessary energy for both grain synthesis and transport of photosynthesized sugars [1]. The phosphorus content in wheat grain, in addition to impacting the performance behavior, represents a product with a high nutritional content [1]. The determination of phosphorus content by using traditional methods is expensive and time consuming, therefore new methods to estimate this content in more efficient ways are needed.

Different vegetative indices have been used to indirectly measure the phosphorus content in wheat grain, such as the Simple Ratio [4], Normalized difference vegetation index [4], Green normalized difference vegetation index [11], Soil adjusted vegetation index [12] and Optimized soil adjusted vegetation index [19], however, these indices have been developed for monitoring N in plants.

A vegetative index that monitors phosphorus in the wheat plant is P_1808_1460 [13], for which it is necessary to adjust the reflectance data to obtain values of the complete light spectrum including the shortwave infrared region (SWIR), we can't compare our results with this vegetative index because there was no information on this light spectrum.

We can also find studies to estimate phosphorus using reflectance data in other crops, such as [17], where they used neural networks to ascertain the key wavelengths for phosphorus prediction in savanna grass; but nothing specific for predict the phosphorus content in wheat grain, for that reason this research proposes a method to estimate it using hyperspectral reflectance.

We propose to evaluate three models: Model 1 assumes that the observations are independent, the other two models assume that the observations are spatially correlated, following either modeled by an exponential correlogram or using a Conditionally Autoregressive Model (CAR). In all models, to satisfy the assumptions of normality the dependent variable is the natural logarithm of the phosphorus concentration in the wheat grain; the independent variables are the wavelength bands measured in nanometers.

The organization of the article is as follows. In material and methods section, we describe a real dataset used for studied the predictive ability of 3 different models and to be able to compare the results with different vegetation indices. We describe the methodology and criteria for select variables, and how we did the cross-validation and predictions in each model. Next, we describe a simulation experiment to show that the proposed models work. Finally, we present the results and discussion. We include an appendix on the derivation of the full conditional distributions necessary to implement Gibbs sampler and Metropolis-Hastings algorithms and the supplementary material includes R codes that implement the proposed algorithms.

Materials and methods

Field experiment and data acquisition

An experiment was carried out at the International Maize and Wheat Improvement Center (CIMMYT), located in the Yaqui valley near to Obregon, Sonora, in the northwest from Mexico. The experiment aimed to investigate the effect of different levels of phosphorus (P) fertilization on P content in the wheat grain. No fertilization was performed with P in the experimental area during the previous four wheat cycles of the experiment to avoid residual effect. The climate in the Yaqui valley is semi-arid with variable precipitation rates averaging 280 mm per year and an average daily temperature of 24 °C. The soils in this region are clay coarse sandy, mixed with montmorillonite clay.

An experiment in split plots to evaluate the phosphorus in wheat grain was carried out in 3 cycles corresponding to the years 2010, 2011 and 2012. Two levels of phosphorus, 0 kg/ha and 80 kg/ ha, were considered in the main plots and 21 different wheat genotypes in subplots. 3 repetitions were performed. Each of the 126 plots, was 4 beds of 0.8 m with two rows on the top by 5 m long.

Hyperspectral reflectance was measured in each plot using a JAZ spectroradiometer Z31 with a CC-3 cosine corrector attached to the optical fiber with a FOV (field of view) aperture of 25° (Ocean Optics, Dunedin, FL, EE.UU.). The sensor has a spectral range of 339 to 1029 nm (nm) with a bandwidth of 0.38 nm, giving a total of 2048 bands. A dark reading was taken just before measurements to set a lower reflectance point of the device. A diffuse white reflectance target, Spectralon (Labsphere, North Sutton, NH, EE.UU.) was used from time to time for field measurements as reference for the upper reflectance point of the device. The data was downloaded and subsequently calibrated using SpectraSuite software (Ocean Optics, Dunedin, FL, EE.UU.). Measurements were taken at the center of each of the four beds, typically from 11:00 am. to 2:00 p.m., targeting at the canopy at a constant height. This procedure was done 2 weeks after anthesis.

At the start this experiment was carried out for purposes of fertilization studies, but for our purpose only the data was retrieved because there were 21 wheat genotypes in 2010 and 2011, in addition to another 21 in 2021, this would allow us to capture the variability not only between individuals but also the spatial one and be able to compare the results under 3 scenarios in the sense that it can be observed variability in each year (Fig. 1) and it is

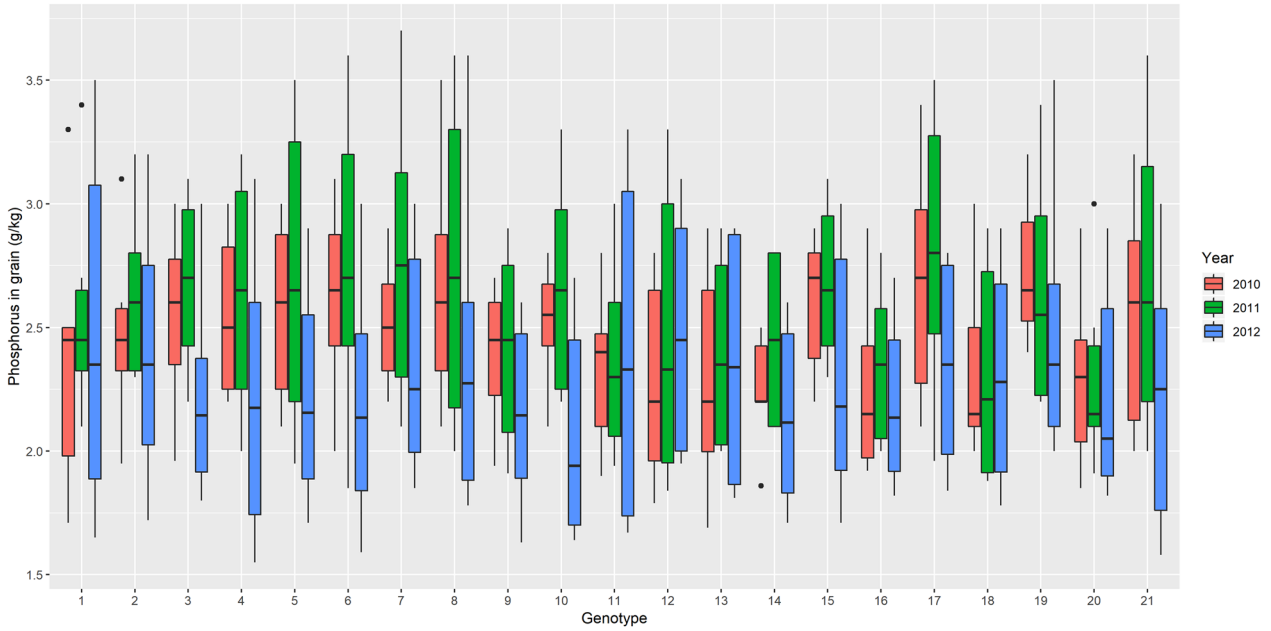

Fig. 1 Boxplot of wheat grain phosphorus in each of the 21 genotypes for each of the 3 years

due to this capability of capture the variability that predictions can improve or worsen.

Figure 1 shows the distribution of phosphorus content in wheat grain in the genotypes of the years 2010, 2011 and 2012. In 2012, 21 different genotypes were included than in the previous years and it is generally observed that these genotypes have lower phosphorus content in the grain.

Pre-processing of hyperspectral reflectance data

Information about 2048 spectral bands was available for use as independent variables in the models. Nevertheless, it is observed that there is up to 65% of data lost in both the first 296 bands and the last 592 bands, so we decided to narrow the range of light spectrum, considering only the range from 450 to 850 nm. Additionally, it is known from several studies that multicollinearity exists between the bands of the spectrum and in our case the information was available with a narrowband of 0.38 nm bandwidth, with which linear combination in parts could be generated to make a resampling using Bsplines [7] and stay with a bandwidth of 4 nm, resulting in 101 total wavelengths used for the data analysis.

Models

Consider the model:

$$y_i = x_i^t \beta + \epsilon_i \tag{1}$$

where y_i is the natural logarithm of phosphorus in wheat grain in case $i = 1, \ldots, n$, $x_i^t = (x_{i1}, \ldots, x_{ip})$ represents the reflectance of light value in each wavelength and $\epsilon_i|\sigma^2 \sim NIID\left(0, \sigma^2\right)$ where "NIID" stands for normal independent and identically distributed random variables, β is the vector of effects for each of the independent variables and σ^2 is the variance component associated to the residuals.

Model (1) can be further extended to include spatial correlation between observations, the so called geo-spatial model which can be written as:

$$y_i = x_i^t \beta + w_i + \epsilon_i \tag{2}$$

where $w = (w_1, \ldots, w_n)\prime$ is the spatial random effects vector with distribution $N\left(0, \sigma^2 \mathrm{H}(\phi)\right)$ and τ^2 is the variance component of y, $\mathrm{H}(\phi) = exp\left(-\phi||s_i - s_j||\right)$, is the isotropic correlation function, where $||s_i - s_j||$ is the Euclidean distance between the site i and j [2].

Another model used frequently in spatial statistics for dealing with aerial data is CAR, the model can be written as:

$$y_i|y_{j:i\neq j} = x_i^t \beta + \sum\nolimits_{j=1}^{n} c_{ij}\left(y_j - x_j^t \beta\right) + \epsilon_i \tag{3}$$

where $\epsilon_i|\tau^2 \sim NIID\left(0, \tau^2\right)$, $\tau^2 > 0$ and $c_{ij} > 0$ are covariance parameters, with $c_{ii} = 0$ for all i. For the set of full conditional distributions to determine a well-defined joint distribution for y we need to consider:

$$y \sim N\left(X\beta, \tau^2 (D_M - \phi M)^{-1}\right)$$

where τ^2 is the variance component of y, M the neighborhood matrix, $D_M = \sum_{j=1}^{n} m_{ij}$ and ϕ is the autocorrelation parameter related to the ordered eigenvalues $\left(\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(n)}\right)$ of ${D_M}^{1/2} M {D_M}^{1/2}$ (for see more details consult [2]).

Selection Criteria

Intrinsic to the adjustment, the method includes the selection of bands (variables), by adapting the method proposed by [10], which induces variable selection. As a result, the posterior probability is obtained, $p\left(\beta_j \neq 0 | data\right)$, where β_j is the regression coefficient corresponding to the j-th band. Or similarly, $p\left(\gamma_j = 1 | data\right)$, where γ_j is an indicator variable corresponding to the j-th band, which is equal to 1 if $\beta_j \neq 0$, and 0 otherwise. With this information the spectral bands can be selected under two criteria:

Those with $p\left(\beta_j \neq 0 | data\right) \geq 0.6$ [3].

The γ vector represents a submodel, its probability was obtained by counting the frequency of each submodel, in this way we can classify the submodels and identify which one is the most likely (avgmod).

Once the previous criteria for band selection have been applied and their respective parameters have been estimated in each model, the Deviance Information Criterion (DIC) [20] is calculated in order to select the best model.

Sampling model and likelihood function

Assuming a random sample from (1), (2) and (3) respectively. Then the conditional distribution for (1) $y_i|\beta, \sigma^2$ is normal with mean $x_i^t \beta$ and variance σ^2. The conditional distribution for (2) $y_i|\beta, \sigma^2, \tau^2, \phi, w$ is normal with mean $x_i^t \beta + w_i$ with variance τ^2. Finally, conditional distribution for (3) $y_i|\beta, \tau^2, \phi$ is normal with mean $x_i^t \beta$ and variance $\tau^2 \left(\sum_{j=1}^{n} m_{ij} - \phi m_{i.}\right)^{-1}$. In general, we the joint conditional distribution of $y|\theta$ is given by $p\left(y|\theta\right) = \prod_{i=1}^{n} p\left(y_i|\theta\right)$, where θ denote the model unknowns for each model.

Prior distributions

In order to complete the specification of the models, we assign prior distribution to the model unknowns. Let $p\left(\beta_j|\gamma_j\right) = \left(1 - \gamma_j\right) p_{spike}\left(\beta_j\right) + \gamma_j p_{slab}\left(\beta_j\right)$, with $p(\gamma_j = 0) = p_j$, $p_{spike}\left(\beta_j\right) = I_{(0)}(\beta_j)$ and $p_{slab}\left(\beta_j\right) = N(\beta_j|\mu_j, v_j)$ for each model. For model (1) $\sigma^2|\alpha_{11}, \eta_{11} \sim IG(\alpha_{11}, \eta_{11})$. In model (2) $\sigma^2|\alpha_{21}, \eta_{21} \sim IG(\alpha_{21}, \eta_{21})$, $\tau^2|\alpha_{22}, \eta_{22} \sim IG(\alpha_{22}, \eta_{22})$ and for $\phi > 0, \phi|min, max \sim U(min, \max)$. And for model (3) $\tau^2|\alpha_{31}, \eta_{31} \sim IG(\alpha_{31}, \eta_{31})$ and $\phi|\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}} \sim U\left(\frac{1}{\lambda_{(n)}}, \frac{1}{\lambda_{(1)}}\right)$.

With $IG(\alpha, \eta)$ we denote an inverse gamma distribution, whose probability density function is $f(x; \alpha, \eta) = \frac{\eta^{\alpha}}{\Gamma(\alpha)} (1/x)^{\alpha+1} exp(-\eta/x)$ where α and η correspond to the shape and rate parameters, respectively. The joint priori distribution $p(\theta|H)$ of each model unknows is given by:

$p\left(\beta, \sigma^2|H_1\right) \propto p(\beta|\gamma) p\left(\sigma^2|\alpha_{11}, \eta_{11}\right)$ for model (1), where $H_1 = \{\alpha_{11}, \eta_{11}\}$ is the set of hyper-parameters.

$p\left(\beta, \sigma^2, \tau^2, \phi, w|H_2\right) \propto p(\beta|\gamma) p\left(\sigma^2|\alpha_{21}, \eta_{21}\right)$ $p\left(\tau^2|\alpha_{22}, \eta_{22}\right) p(\phi|min, max) p\left(w|\sigma^2, \phi\right)$ for model (2), where $H_2 = \{\alpha_{21}, \eta_{21}, \alpha_{22}, \eta_{22}, min, max\}$ is the set of hyper-parameters.

$p\left(\beta, \tau^2, \phi|H_3\right) \propto p(\beta|\gamma) p\left(\tau^2|\alpha_{31}, \eta_{31}\right) p\left(\phi|\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}}\right)$ for model (3), where $H_3 = \left\{\alpha_{31}, \eta_{31}, \frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}}\right\}$ is the set of hyper-parameters.

Posterior distributions

The joint posterior distribution of all quantities can be obtained by applying the Bayes' theorem, so we obtain $p(\theta|data, H) \propto p\left(y|\theta\right) p(\theta|H)$.

The hierarchical structure of this distribution allows us to obtain the conditional distributions necessary to implement the Gibbs sampler [8] and draw samples from the joint posterior distribution, in other form this distribution is analytically un-tractable. Not all full conditional distributions have a closed form, for that reason was necessary to implement the Metropolis–Hastings algorithm [6]; the algorithms are described in the Appendix.

The hyper-parameters for the inverse gamma distributions are set as $\alpha = \eta = 0.01$ because it provides a weakly informative prior [14]; for the uniform distributions are set as min =0, max = 1, and $\lambda_{(1)}, \lambda_{(n)}$ are the eigenvalues mentioned in model (3).

Software

The algorithm to fit models was implemented in a program written in R [18]. The input arguments are the response vector y, the matrices X,w,M, the number of Markov Chain Monte Carlo (MCMC) iterations, a burn-in period and the hyper-parameters. The outputs provide the mean of the predictive distribution obtained through the MCMC algorithm and provides us with the variables selected under the selection criteria previously described, it also computes the DIC [20] and it provides the correlations that are obtained when making the cross validations. Three libraries from R were used, MCMCpack [15], mvtnorm [9] and truncnorm [16].

Simulations

To evaluate the behavior of the variable selection procedure in each model, a simulation experiment

was carried out, this consisted of generating X with 100 random variables from a normal standard distribution, based on linear combinations $x_i + x_j = x_k$ with $i \in \{1, 3, 5, \ldots, 99\}$, $j \in \{2, 4, 6, \ldots, 100\}$ and $k \in \{101, 102, \ldots, 150\}$ were generated 50 more variables, this to simulate the multicollinearity expected in real dataset. The neighborhood matrix was the same as in the real dataset. β with values close to zero was also generated to specify variables with little effect and large values for the most important variables, let $A = \{2, 6, 13, 33, 25, 67, 71, 77, 85, 94, 96, 99\}$ and $B = \{1, 2, 3, \ldots, 150\}$ then $\beta_j = 0.9$ with $j \in A$, $\beta_j = 0.01$ with $j \in B - A$; and values were set for the parameters $\sigma^2 = 0.3, \tau^2 = 0.45, \phi = 0.21$. Using those data and conditional distribution we simulated y for each model. Finally, we used the simulated data as an input in our R script in order to check that in each model the desired parameters are being correctly estimated.

Cross validation and predictions

In order to have a reference regarding the predictive power of the models used, cross validation was performed. The response vector, y, was divided into 2 disjoint sets, randomly, always considering 25 observations for the testing data set and 101 for the training set. In such a way that $y = (y_{training}, y_{testing})'$. Also, the matrix of covariates or bands, was divided in a way that corresponded to the values of y, such as $X = (X_{training}, X_{testing})'$ The models were fitted with the information corresponding to the "training", and the adjusted model was used to predict the response of the "testing" set. This procedure was repeated 5 times.

Results

The results presented below are derived from 100,000 iterations. The first 50,000 were discarded, to ensure convergence (which was validated graphically using the trace plots). As a sample, one in 5 of the last 50,000 were considered, with which the averages of each parameter involved, called a point estimate, were calculated. For example, for the model without spatial correlation, the point estimation of the parameters is denoted as $\left(\widehat{\beta}, \widehat{\sigma}^2\right)$. For the CAR model, the point estimation of the parameters is denoted as $\left(\widehat{\beta}, \widehat{\tau}^2, \widehat{\phi}\right)$, and $\left(\widehat{\beta}, \widehat{\sigma}^2, \widehat{\tau}^2, \widehat{\phi}\right)$ for the geospatial model. The DIC was calculated for each model. $p(\beta_j \neq 0 | data)$ was calculated as a measure of importance of each band and the most probable models, and as a measure of the predictive power of the models the Pearson's correlation between $y_{testing}$ and $\widehat{y}_{testing}$ (the prediction of response variable based on the model adjusted with the information in the "training" set) of cross validation.

In Figs. 2, 3, 4, we show the trace plots and posterior means of parameters of interest for each year when considering all wavelengths for models (1), (2) and (3). Note that the variability of the data is very similar between 2010 and 2011 (see $\widehat{\sigma}^2$ in the case of the model without spatial correlation and $\widehat{\tau}^2$ in the CAR). The 2012 variance is smaller than in the previous years. This difference could be attributed to the fact that

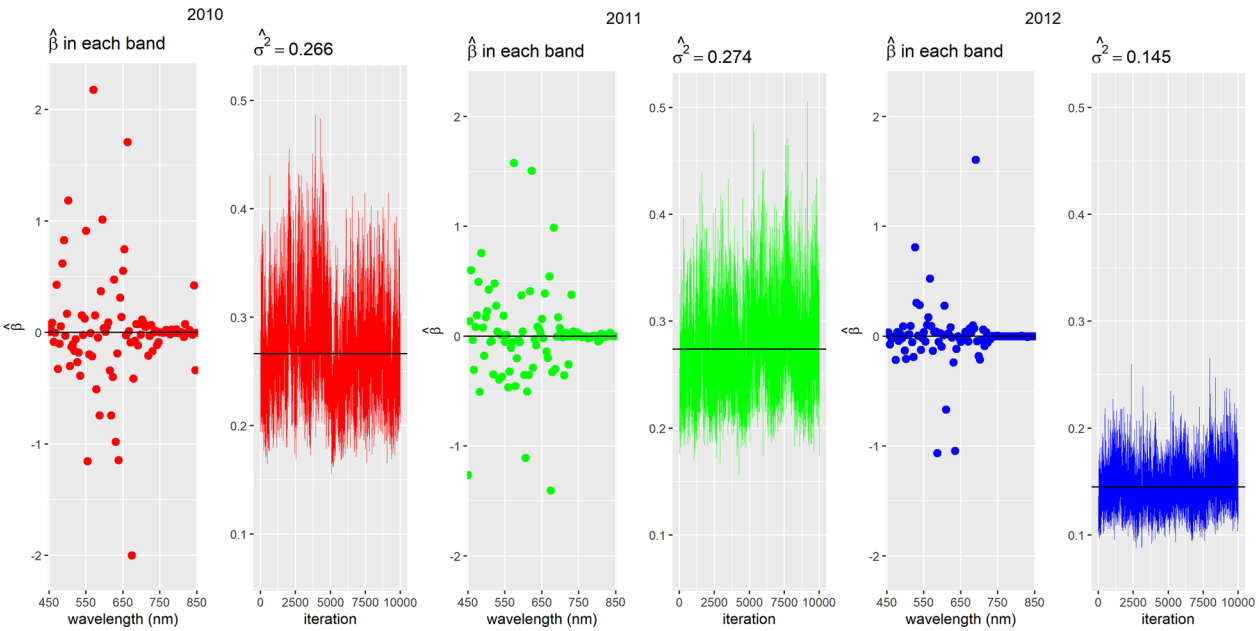

Fig. 2 Trace plots and posterior means of estimated parameters in the model without spatial correlation per year

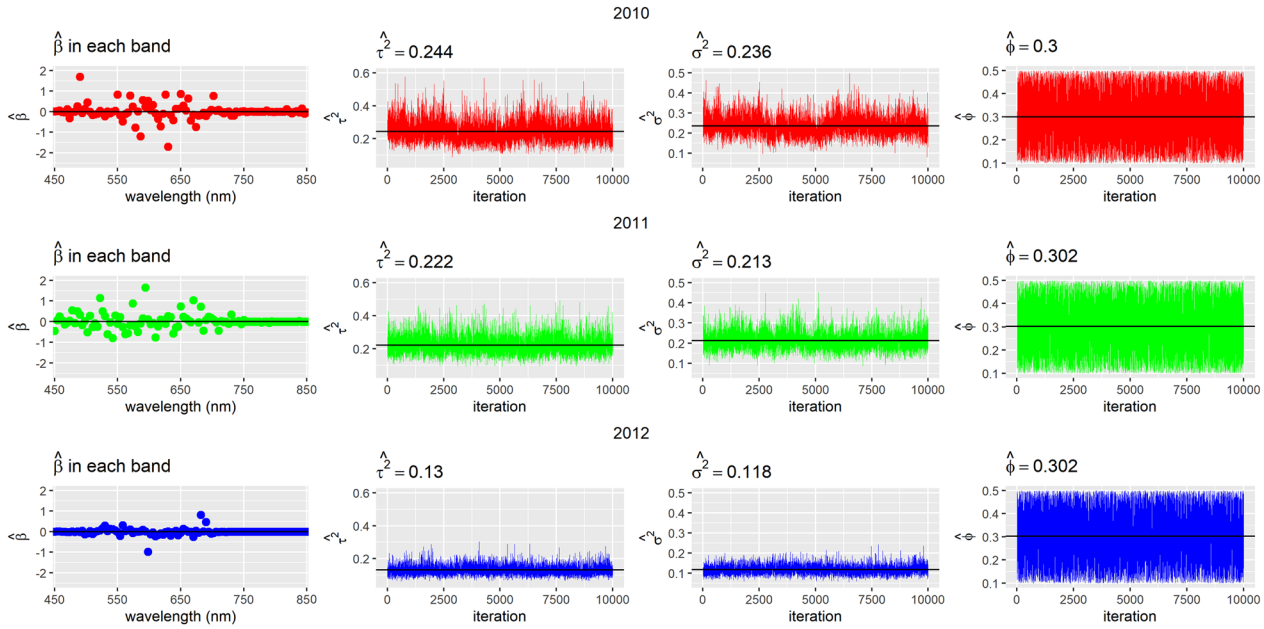

Fig. 3 Trace plots and posterior means for parameters in the geospatial model by year

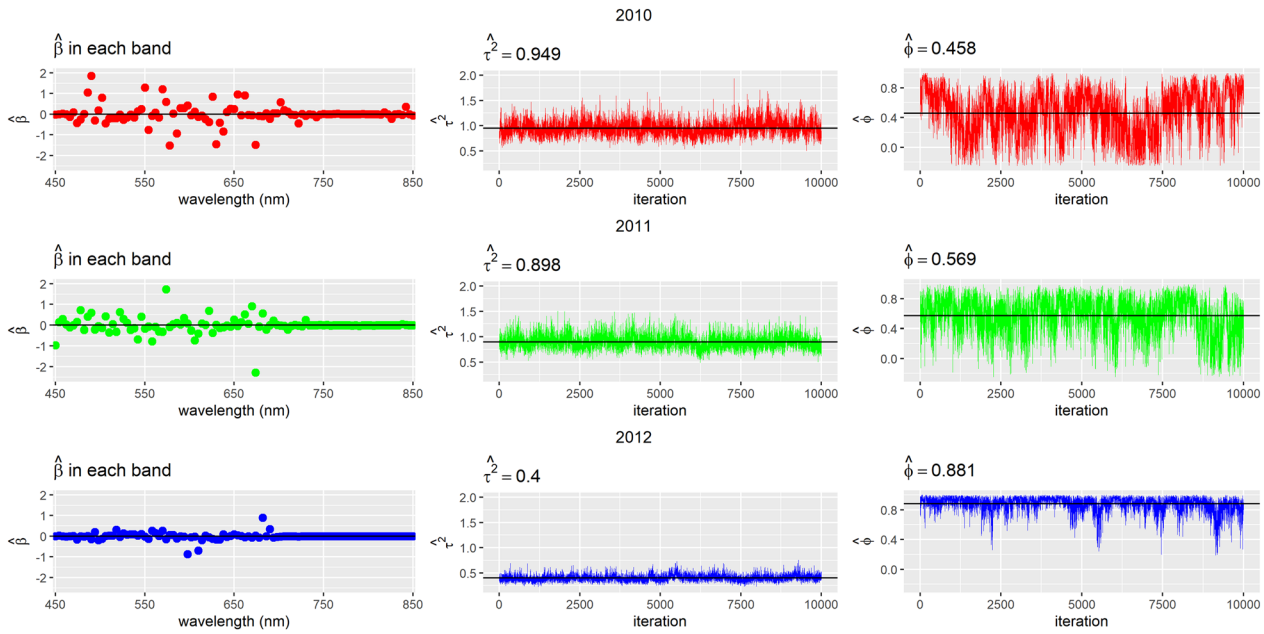

Fig. 4 Trace plots and posterior means of parameters in the CAR model per year

different genotypes were used for the experiment in this last year.

With respect to the parameter of spatial variation $\left(\widehat{\phi}\right)$ in the CAR model, values were found around and above 0.5, depending on the year, which indicates a positive dependence between the plots studied. In the case of the geospatial model it is observed that the maximum distance from which there is spatial dependence is $\widehat{\phi} = 0.3$ and that the maximum variability in the absence of spatial dependence is $\widehat{\sigma}^2 = 0.2$, approximately.

In Figs. 5, 6, 7 the value of the posteriori probability for each spectral band is observed at each point, that is $p\left(\beta_j \neq 0 | data\right)$ graphically, the light spectrum to which they belong is also illuminated. The black dots represent

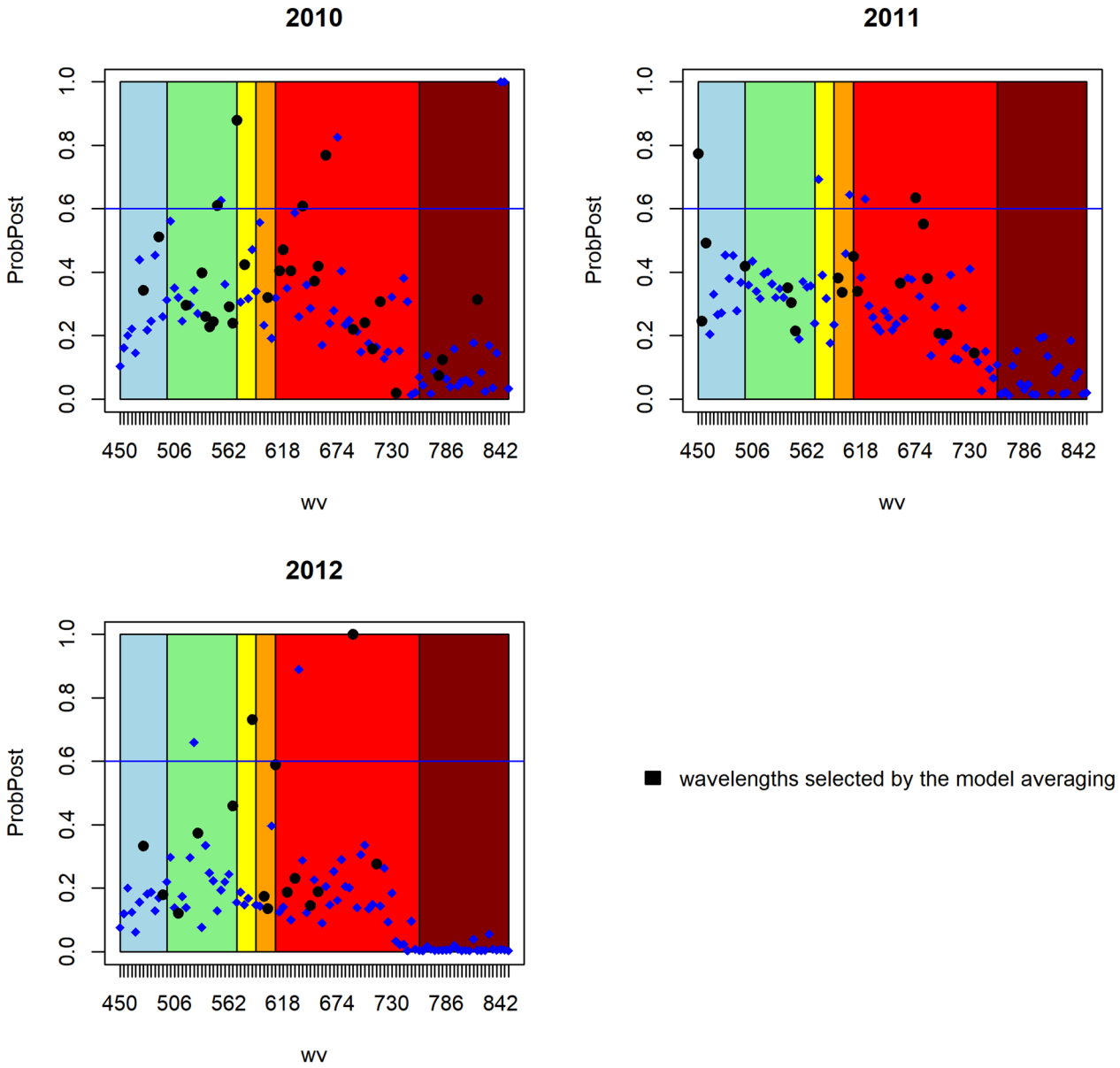

Fig. 5 Posterior probability of the spectral bands of each year of the model without spatial correlation

a posteriori probability of the spectral bands selected by the most probable model, and above the blue horizontal line, which represents the posterior probability of 0.6, are the important spectral bands.

It can also be observed that in general the posterior probability of the spectral bands is greater in the visible spectrum from green to red, that is, in the range of 500 to 690 nm, regardless of the fitted model or the year of the experiment. However, there are no specific or recurring wavelengths that can be considered in general to directly relate them to the behavior of the phosphorus content in wheat grain.

As mentioned, the way to select the best model was through the DIC. In Table 1 it can be seen the CAR model was the selected model, since the value of the DIC obtained is the smallest for all scenarios considered and is even lower if we consider a model that includes only selected bands.

To measure the predictive power of the models, cross validation was used. Table 2 shows the Pearson's correlation coefficient calculated between $y_{testing}$ and $\widehat{y}_{testing}$ in each of the models used, for each year, for each wavelength selection. Consistently, it is observed that the CAR model on average is the one that has the

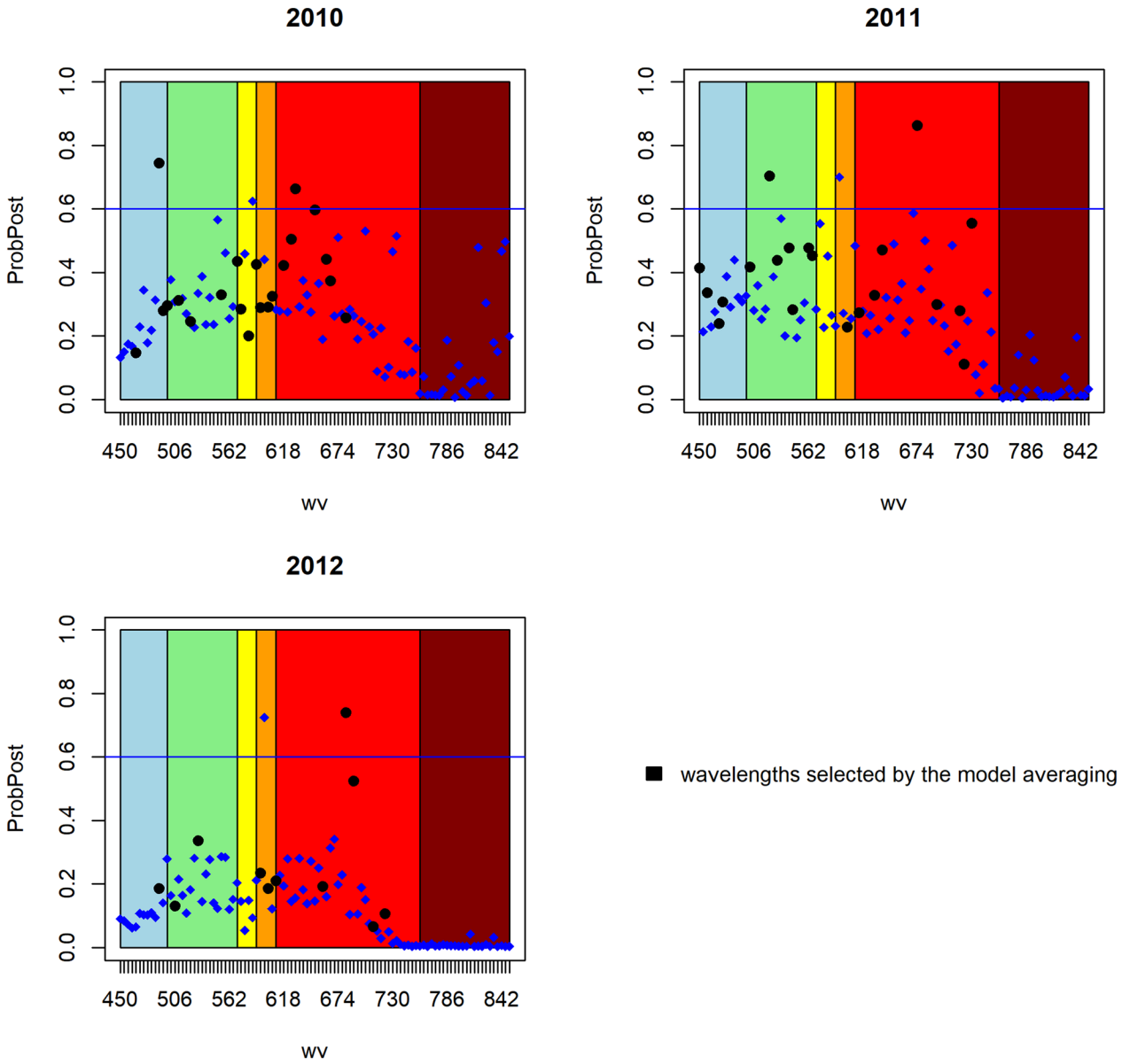

Fig. 6 Posterior probability of the spectral bands of each year of the geospatial model

best predictive power when presenting the highest correlations.

Once the CAR model was selected to model the P content in the wheat grain, we proceeded to select from the literature some vegetative indexes that have been used to monitor the nutrient content in plants (N, P, K, S), as done in [13] and with these evaluate the prediction of the CAR model. The same 5 subsets of cross-validation tests were used to calculate the vegetation indices in Table 3 and obtain the correlation coefficient for each year. In general, for the three years, the results with the CAR model are better than with the vegetative indices, since their average correlation remains constant above 0.6 or more, rather what stands out is the fact that the vegetative indices alone cannot be trusted because is clearly observed that we can have both good results as in the case of year 2011 and 2012, but very bad in the year 2010 (Table 3).

Conclusion

Under the model selection criterion, DIC, it was concluded that the best of the adjusted models was the CAR. With the result of the cross-validation, it was concluded

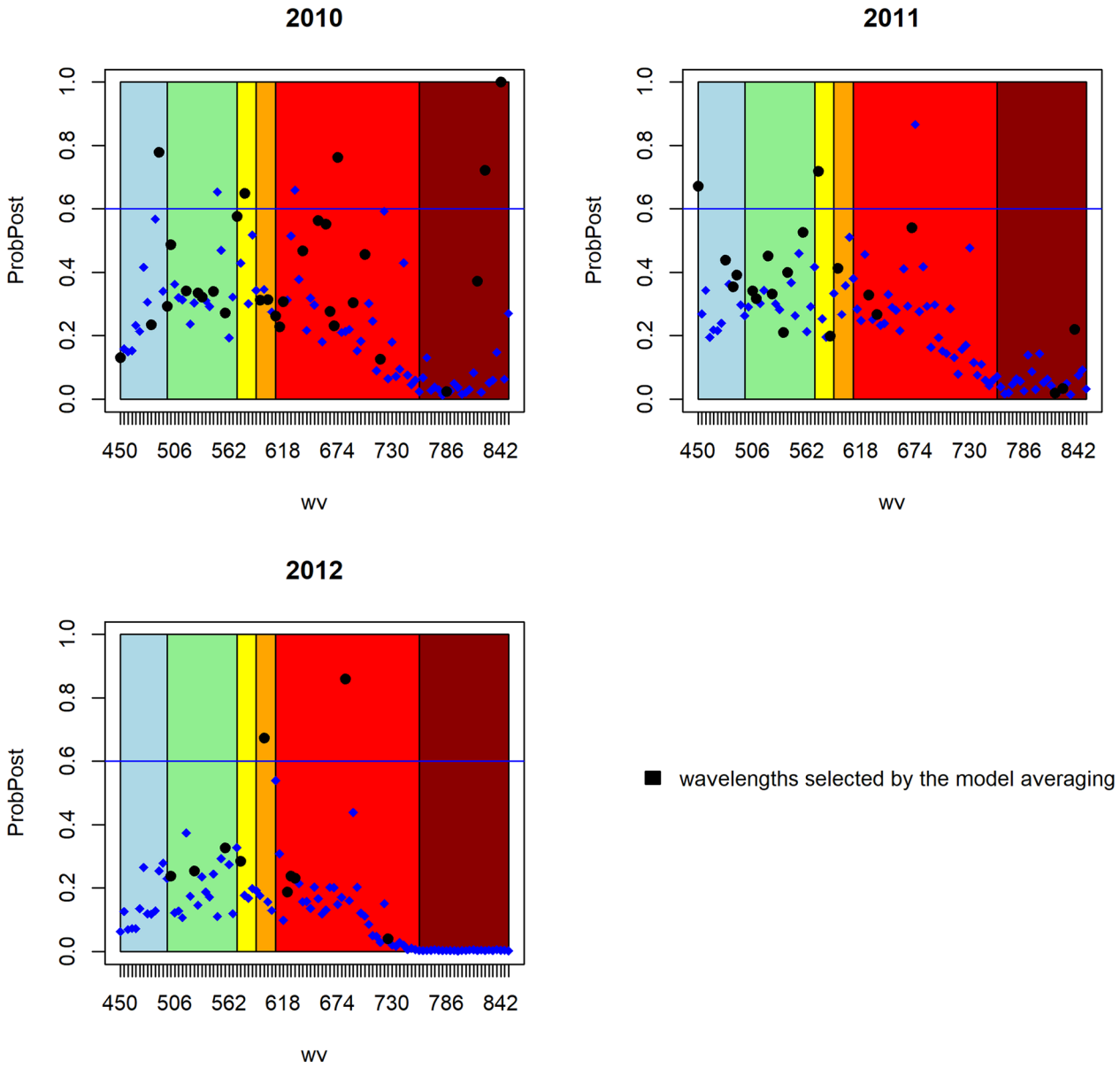

Fig. 7 Posterior probability of the spectral bands of each year of the CAR model

Table 1 DIC values

	Selection criterio								
	All bands			$p(\beta_j \neq 0 \mid data) \geq 0.6$			Avgmod		
Model	2010	2011	2012	2010	2011	2012	2010	2011	2012
SCE	590.67	534.12	886.98	335.74	333.37	552.68	466.71	434.38	731.89
Geospatial	438.87	461.98	288.03	485.25	314.71	548.48	542.73	392.09	493.97
CAR	242.97	226.02	121.83	254.92	233.35	162.73	239.87	211.82	102.87

Table 2 Average correlations in cross-validation

	Selection criterio								
	All bands			$p(\beta_j \neq 0 \mid data) \geq 0.6$			Avgmod		
Model	2010	2011	2012	2010	2011	2012	2010	2011	2012
SCE	0.642	0.677	0.652	0.523	0.389	0.635	0.533	0.630	0.634
Geospatial	0.647	0.736	0.749	0.499	0.639	0.630	0.595	0.671	0.733
CAR	0.655	0.709	0.814	0.524	0.621	0.647	0.601	0.680	0.823

Table 3 Average correlations in cross-validation of vegetative indexes

Index	Expression	Year		
		2010	2011	2012
Simple Ratio [4]	$\frac{R_{800-900}}{R_{650-700}}$	0.338	0.738	0.666
Normalized difference vegetation index [5]	$\frac{R_{800}-R_{680}}{R_{800}+R_{680}}$	0.356	0.685	0.527
Green normalized difference vegetation index [11]	$\frac{R_{800-900}-R_{540-560}}{R_{800-900}+R_{540-560}}$	0.443	0.674	0.687
Soil adjusted vegetation index [12]	$\frac{1.5R_{800-900}-R_{650-700}}{R_{800-900}+R_{650-700}+0.5}$	0.353	0.692	0.530
Optimized soil adjusted vegetation index [19]	$\frac{1.16R_{800}-R_{670}}{R_{800}+R_{670}+0.16}$	0.352	0.671	0.566
	CAR model avg	0.601	0.680	0.823

R corresponds to the reflectance at corresponding subscripted wavelength (nm)

that the model with the best predictive power coincided with the best-adjusted model, that is, the CAR.

This was expected given that the spatial dependence in a CAR model can separate and clarify the structural and functional components, with the structural ones we understand the correlation that is determined by physical proximity (close neighbors) the functional ones refer to the correlation that it is affected by dispersion, landscape characteristics and other variables of interest that are taken into account by the CAR model. These desirable characteristics, which cannot be included with the simple calculation of an index, enhance the use of the CAR model.

Also, the a posteriori probability obtained by the implementation of variable selection in all the evaluated models is observed and a repetitive pattern that would determine that with the spectral bands selected as the most probable and lead to an index built to determine the content of phosphorus in the wheat grain was not found, however, if it could be concluded that the range of the light spectrum that goes from 500 to 690 nm, is the one that most likely directly intervenes when predicting the phosphorus content in the grain. This information is useful to recommend making a good calibration of the sensor with which reflectance readings will be taken in the range found.

Comparing the vegetative indices with the CAR model, it was observed that the average correlation of the CAR model exceeded that of the vegetative indices by 23.26%, − 1.2% and 22.78% for the year 2010, 2011 and 2012 respectively; therefore, the use of the proposed methodology outperformed the vegetative indices in prediction.

Therefore, the use of this methodology is not only useful to reduce dimensionality, even when there are multicollinearity problems, but also with the posterior probabilities obtained, the importance and/or inclusion of some band in the prediction model can be decided. It is also possible to take into account the inclusion of spatial variability in the model, with which the model (1) was surpassed in prediction by up to 19%.

In this research, this methodology is proposed to predict the phosphorus content in wheat grain, reflecting with the results as a good alternative, however, if the reflectance information and the disposition of the experiment in the field are available, it is suggested to evaluate it with other variable responses of interest such as yield for example and in other crops.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13007-023-00980-9.

Additional file 1: Appendix: Conditional distributions.

Additional file 2: Hyperspectral data.

Additional file 3: R codes.

Acknowledgements
Not applicable.

Author contributions
IOM designed the experiment and provided supervision of the data collection. FR led the hyperspectral data processing. IOM, FR, JB, PPR, FT, SPE, DHVP data interpretation and critically reviewed the manuscript. RAPG drafted the manuscript. RAPG and CVC develop the method and data analysis. All authors read and approved the final manuscript.

Funding
We are thankful for the financial support provided by the Bill & Melinda Gates Foundation [INV-003439, BMGF/FCDO, Accelerating Genetic Gains in Maize and Wheat for Improved Livelihoods (AG2MW)], the USAID projects [USAID Amend. No. 9 MTO 069033, USAID-CIMMYT Wheat/AGGMW, AGG-Maize Supplementary Project, AGG (Stress Tolerant Maize for Africa], and the CIMMYT CRP (maize and wheat).

Availability of data and materials
The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Texcoco, México. [2]Integrated Development Program, International Maize and Wheat Improvement Center (CIMMYT), Texcoco, México. [3]Socioeconomics, Statistics and Informatics Department, Colegio de Postgraduados, Texcoco, México. [4]Lincoln Agritech Ltd, Lincoln University, Canterbury, New Zealand.

Received: 25 August 2022 Accepted: 31 December 2022
Published online: 20 January 2023

References

1. Acevedo, E., Silva, P., and Silva, H. (2002). Bread Wheat: Improvement and Production, Wheat growth and physiology. FAO Plant Production and Protection Series. Food and Agriculture Organization of the United Nations.
2. Banerjee S, Carlin PB, Gelfand AE. Hierarchical Modeling and Analysis for Spatial Data. Boca Raton: CRC Press Taylor and Francis Group; 2015.
3. Barbieri M, Berger O. Optimal predictive model selection. Ann Stat. 2004;32:870–97.
4. Birth GS, McVey GR. Measuring the color of growing turf with a reflectance spectrophotometer. Agron J. 1968;60:640–3.
5. Blackburn GA. Quantifying chlorophylls and caroteniods at leaf and canopy scales: an evaluation of some hyperspectral approaches. Remote Sens Environ. 1998;66:273–85.
6. Siddhartha C, Greenberg E. Understanding the metropolis-hastings algorithm. Am Stat. 1995;49(4):327–35.
7. De Boor C. A practical guide to splines. Berlin: Springer; 1978.
8. Geman S, Geman D. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell. 1984;6:721–41.
9. Genz A., Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. 2020. mvtnorm: multivariate normal and t distributions. R package version 1.1–1. https://CRAN.R-project.org/package=mvtnorm.
10. Geweke J. Variable selection and model comparison in regression. Bayesian statistics. 1996;5:609–20.
11. Gitelson A, Kaufman Y, Merzlyak M. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens Environ. 1996;58:289–98.
12. Huete A. A soil-adjusted vegetation index (SAVI). Remote Sens Environ. 1988;25:295–309.
13. Mahajan GR, Sahoo RN, Pandey RN, Gupta VK, Kumar D. Using hyperspectral remote sensing techniques to monitor nitrogen, phosphorus, sulphur and potassium in wheat (triticum aestivum l.). Precision Agric. 2014;15:499–522.
14. Walli GM. Bayesian variable selection in normal regression models. Austria: Johannes Kepler University Linz (master degree thesis); 2010.
15. Martin AD, Quinn KM, Park JH. "MCMCpack: Markov Chain Monte Carlo in R." J Stat Software. 2011;42(9):22.
16. Meisner CA, Acevedo E, Flores D, Sayre K, Ortiz-Monasterio JI, Byerlee D. Wheat production and grower practices in the yaqui valley, sonora, Mexico. Texcoco: CIMMYT; 1992. p. 6.
17. Mutanga O, Kumar L. Estimating and mapping grass phosphorus concentration in an african savanna using hyperspectral image data. Int J Remote Sens. 2007;28:4897–911.
18. R Core Team. 2019. R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Australia. http://www.R-project.org/.
19. Rondeaux G, Steven M, Baret F. Optimization of soil-adjusted vegetation indices. Remote Sens Environ. 1996;55:95–107.
20. Spiegelhalter DJ, van der Linde A. Bayesian measures of model complexity and fit (with discussion). J Royal Stat Soc Serie B. 2002;64:583–639.

Publisher's Note