A hyperspectral index sensitive to subtle changes in the canopy chlorophyll content under arsenic stress

Xuqing Li, Xiangnan Liu*, Meiling Liu, Cuicui Wang, Xiaopeng Xia

School of Information Engineering, China University of Geosciences, Beijing 100083, China

ARTICLE INFO
Article history:
Received 28 May 2014
Accepted 30 October 2014

Keywords:
Canopy chlorophyll content
Random forest
Arsenic stress-sensitive index

ABSTRACT
Arsenic stress induces in subtle changes in the canopy chlorophyll content (CCC). Therefore, the establishment of a spectral index that is sensitive to subtle changes in the CCC is important for monitoring crop arsenic contamination in large areas by remote sensing. Experimental sites with three contamination levels were selected and were located in Chang Chun City, Jilin City, Jilin Province, China. Arsenic stress can induce small changes in the CCC, reflecting in the crop spectrum. This study created a new index to monitor the CCC. Then, the results from the index were compared with these from other indices and the random forest model, respectively. The final purpose of this study is to find an optimal index, which is sensitive to small changes in the CCC under arsenic stress for monitoring regional CCC in rice. The results indicate that the distribution of the CCC is aligned with the distribution of the arsenic stress level and that NVI ($R_{46}$, $R_{732}$, and $R_{792}$) is the best index for monitoring CCC. The correlation coefficient $R^2$ between the predicated values using NVI and the measured values of canopy chlorophyll content is 0.898, which performs better than the random forest model and other indices.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Arsenic is a threat to human health, and rice arsenic contamination is a global ecological problem. Several studies suggest that the arsenic ground water irrigation (Dahal et al., 2008), the use of agrochemicals and arsenic metal mineral resources (Liu et al., 2010) and the utilization of soil with arsenic contamination (Huang et al., 2006) is the main cause of arsenic pollution. Additionally, the ability of rice to accumulate arsenic is significantly higher than that of other dry land crops, such as wheat and corn (Kastori et al., 1998a,b).

Heavy metal stress can induce changes in canopy chlorophyll content. Leaf spectral characteristics are mainly governed by its cellular structure, water content, biochemical composition, and pigments content (Guyot et al., 1992). In summary, the metal stress induced a decrease in canopy chlorophyll content (Zengin and Munzuroglu, 2005) and variations of leaf internal structure (Slaton et al., 2001); these alter the Vis-IR radiation's reflectance (Horler et al., 1983; Schwaller et al., 1983; Milton et al., 1989a,b). Heavy metal stress induces variations in chlorophyll and cell structure, which can reflect on the spectral variations and can be estimated using vegetation index.

There is some research about how to use spectroscopy to monitor the heavy metal content in soil directly or use crop spectral to estimate heavy metal content in soil indirectly (Rathod et al., 2013). Some researchers estimated heavy metals in plant samples of Indian mustard, amaranth and rice (Font et al., 2004a,b,c, 2002, 2005) or detected Zn and Cu-induced stress in wheat and lettuce (Woodhouse et al., 1994); the spectral changes in leaves of maize and wheat plants as a result of soil contamination with hydrocarbons (Noomen et al., 2007); the visible reflectance of picking weed leaves closely followed the changes in pigment concentration due to Cd treatments (Rosso et al., 2005).

Although many previous chlorophyll indices have been proposed, they do not effectively estimate the arsenic stress level. It is therefore essential that measures be made to monitor the CCC under the stress of heavy metal. On the one hand, spectral characteristics of metal-contaminated soil have been mainly studied in the laboratory or field with studies of limited scope, so no general prediction mechanisms can be obtained, on the other hand, most remote sensing methods for vegetation estimation are only capable of estimation relatively severe plant stress. Some research about the subtle changes in canopy chlorophyll concentration (CCC) under the stress of heavy metal already exists in the farm environment. There are several common approaches to monitoring the CCC through remote sensing.

The new indices including a modified triangular vegetation index (MTIV12) and a modified chlorophyll absorption ratio index

* Corresponding author. Tel.: +86 10 82320506; fax: +86 10 8232095.
E-mail addresses: liuxqing@cugb.edu.cn (X. Li), liuxn@cugb.edu.cn (X. Liu).
http://dx.doi.org/10.1016/j.jag.2014.10.017
0303-2434/© 2014 Elsevier B.V. All rights reserved.
(MCAR12) were the best index to estimate green LAI (Haboudane et al., 2002). A vegetation index CCI (D-794/D-763) was presented and the model with CCI (D-794/D-763) as the independent variables was determined to be the best for chlorophyll content prediction of apple tree canopy (Pan). A developed conceptual model was used to estimate chlorophyll content remotely in maize and soybean, and it has good accuracy (Gitelson et al., 2003, 2005a,b). A spectral reflectance index was used to estimate genetic variation for leaf chlorophyll, and discovered that reflectance spectra index correlated with SPAD chlorophyll values (Rabar et al., 2006). The MERIS terrestrial chlorophyll index (MTCI) was created to evaluate chlorophyll using model spectra, field spectra and MERIS data (Dash and Curran, 2007; Dash et al., 2010).

Development of vegetation index has gone from normalized difference vegetation index, ratio vegetation index, reflectance derivatives or more complex band combination to many machine learning algorithms, such as PLSR (partial least square regression), PCR (principal component regression), MLR (multiple linear regression), and ANN (artificial neural networks), LUTF (look-up table method), AIF (the artificial intelligence algorithms), BA (Bayesian algorithms), SVM (support vector machine algorithm) (Kimes, 2000; Liang et al., 2004; Qu et al., 2008; Durbha et al., 2007) and so on, which develops calibration models.

A algorithm which combine 4-scale and PROSPECT to monitor leaf chlorophyll content through hyperspectral CASI imagery and to estimate different vegetation types' foliar chemistry from optical remote sensing data (Zhang et al., 2008; Croft et al., 2013). A chlorophyll estimation model based on the simulated canopy reflectance was built, which got good results when measured canopy reflectance and reflectance from Hyperion images the input parameter (Dong et al., 2009).

The chlorophyll content of irrigated land was much higher than dry land, and an estimate model of red-edge inflection points of canopy spectral reflectance and chlorophyll content of spring wheat at different lands was built (Jin et al., 2013). Hyperspectral remote sensing methods can be complex changes in the real-time detection of tiny creatures, better timeliness. Some previous studies are carried out under laboratory conditions (Milton et al., 1989a,b) rather than farm environment, and this is precisely one of the key issues in hyperspectral remote sensing for precision agriculture applications. Hyperspectral imaging at a 30-m resolution offers an opportunity for the operational estimate of the CCC at a regional scale and has been studied extensively in the past several decades. Model area application will most likely generate scale effects.

Heavy metal stress causes changes in the CCC, and these tiny changes can be quantitatively retrieved through crop spectral changes, the new index that is aimed at capturing the tiny changes. The purpose of the study is to determine the optimal index, which is very sensitive to subtle changes in the CCC under arsenic stress and can obtain higher accuracy than can the other models, such as the ratio vegetation index, the normalized difference vegetation index and even the random forest model. This paper selected an optimal index, scaled it up to the regional level and applied it to hyperspectral image date to produce regional maps after a linear change to reduce the scale effects.

2. Materials and methods

2.1. Materials

2.1.1. Study area and experimental design

Study area is located in Changchun City, Jilin Province (43° 51’ N, 125° 10’ E) (Fig. 1). The rice variety is Jilin Japonica 105. The area belongs to temperate continental climate with an average annual rain-fall of 522–615 mm and an average annual temperature of 5.8 °C. There were 150 samples in Changchun. Every elemental plot was 3 m × 3 m. Our experiment is the estimation of chlorophyll content in rice mature period, it has a roughly divided for heavy metal and arsenic pollution of arsenic stress through the estimation of chlorophyll content in rice mature period. All experimental data are sampled in three pollution levels in the same mature period, and in a similar growth environment, fertilization, water conditions. There were approximately 20 samples for validation in study area b.

2.1.2. Heavy metal content and CCC measurement

The arsenic content in the rice and soil was analyzed by the Chinese Academy of Agricultural Sciences (Table 1). The rice LCC was measured randomly 10 times and then averaged. In this paper, the chlorophyll concentration was calculated by the SPAD-502 (Minolta Corporation, New Jersey, USA) chlorophyll readings, and the equation was

\[ y = 0.996x - 1.52 \]

where \( y \) and \( x \) are chlorophyll concentration and SPAD-502 chlorophyll reading, respectively (unit: \( \mu g/cm^2 \)).

Plant canopy chlorophyll content (CCC, canopy chlorophyll content) is calculated by leaf chlorophyll content (LCC, leaf chlorophyll content) and leaf area index (LAI), namely: \( CCC = LCC \times LAI \). The chlorophyll content of the ground area of the canopy level units is the product of the chlorophyll content of the leaf level unit area and leaf area index.

2.1.3. Field spectral data

In this study, the rice canopy spectral reflectance was measured wavelength range of 350–2500 nm using an ASD Field Spec Pro spectrometer (Analytical Spectral Devices, Boulder, CO, USA), of which 350–1000 nm spectral sampling interval of 1.4 nm, 1000–2500 nm spectral sampling interval is 2 nm, measured before using standard whiteboard correction, bare field of fiber. We select Changchun area three pollution levels regional data in mature rice canopy spectral reflectance in September 30th, in Changchun area sampling nearly 10 plots, a paddy field, one is the road next to the paddy field, is a chemical plant near the paddy field, a total of 150 rice reflectivity pop data. Canopy spectra for ASD spectrometer measured parameters such as chlorophyll concentration, leaf area index and canopy spectra, synchronous determination of the test field, arsenic content of soil heavy metals. A regional study area is for the experimental area, B area is the verify area. Experimental zone sampling 150, each sample measured 10 spectral data, whichever is the mean as the final result of this sample.

In this study, a scene of Hyperion image which is from the NASA Earth Observing One (EO-1) satellite cover the city of Chang Chun, Jilin, Jilin Province, China was acquired on 7 October 2009, after a week of the mature period ground spectral data measuring with a dry paddy soil. The spectral resolution of the image is approximately 10 nm between 400 and 1100 nm, the spatial resolution is 30 m, each image is a 7.5 km × 100 km land area. The Hyperion data were pretreated through the following steps: bad line replacement, radiation calibration, and atmospheric correction and geometric calibration.

2.2. Methods

The traditional statistical method is to calculate the correlation between the CCC and variables to determine the optimal band or index. The optimal inversion model can also be determined through some intelligent algorithms, such as the machine learning
algorithms. In this study, we chose the random forest algorithm for comparison (Fig. 2).

2.2.1. Traditional statistical analyses

We first tested three CCC indexes, where each index derived the CCC as a function of an independent variable.

(1) The first and most simple index type is the reflectance value \((R)\) at the central wavelength \((x, \text{nm})\) of each band.

(2) The normalized difference spectral index (NDSI) is defined as 
\[
NDSI(x,y) = \frac{(y + x)}{(y - x)}
\]
the ratio spectral index (RSI) is defined as 
\[
RSI(x,y) = \frac{x}{y}
\]
where \(x, y\) are the reflectance \((R_i, R_j)\) values at \(i, j\) (nm).

(3) The NVI is defined as 
\[
NVI(x,y) = \frac{(y - x)}{(y - z)}
\]
where \(x, y\) and \(z\) are the reflectance \((R_i, R_j\) and \(R_k)\) values at \(i, j\) and \(k\) (nm), respectively, over the entire hyperspectral.

Among these values, the \(x\) and \(y\) values can also be derivatives of the reflectance.

---

Table 1

The location and heavy metal concentrations of the experiment sites.

<table>
<thead>
<tr>
<th>Geographical location</th>
<th>Arsenic content (mg kg(^{-1}))</th>
<th>Soil pollution level</th>
<th>Soil quality standard(^a) (mg kg(^{-1}))</th>
<th>Sample number</th>
</tr>
</thead>
<tbody>
<tr>
<td>125°10’25.3” E, 43°51’35” N</td>
<td>8.241482966</td>
<td>Clean soil</td>
<td>As &lt; 10</td>
<td>12</td>
</tr>
<tr>
<td>125°10’25.0” E, 43°51’34.8’ N</td>
<td>9.297108674</td>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>125°10’25.0” E, 43°51’34.8’ N</td>
<td>9.313181368</td>
<td>Clean soil</td>
<td>As &lt; 10</td>
<td>20</td>
</tr>
<tr>
<td>125°09’07.2” E, 43°51’37.0’ N</td>
<td>10.72323351</td>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>126°28’ E, 43°57’ N</td>
<td>16.17647059</td>
<td>Still clean soil</td>
<td>15 &lt; As ≤ 25</td>
<td>7</td>
</tr>
<tr>
<td>126°28’ E, 43°57’ N</td>
<td>16.44034918</td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>125°10’ E, 43°51’ N</td>
<td>18.41983318</td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>125°10’ E, 43°51’ N</td>
<td>19.35483871</td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>126°24’ E, 43°59’ N</td>
<td>20.40618956</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>126°37’ E, 43°55’ N</td>
<td>20.74416342</td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>126°24’ E, 43°59’ N</td>
<td>23.82958801</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>126°37’ E, 43°55’ N</td>
<td>34.23267327</td>
<td>Heavily polluted soil</td>
<td>30 &lt; As</td>
<td>20</td>
</tr>
</tbody>
</table>

\(^a\) Soil quality standard according to the Environment Monitoring Centre of China.
2.2.2. Random forest model

Efron (1979) proposed a general resampling technique called bootstrap, often used for statistical inference. A bootstrap sample usually generated by repeated sampling and replaces the original data. Each bootstrap sample has the same sample size: $n$ as the original data, each observed value is selected with probability $1/n$, it can occur many times or almost at most. For each of the bootstrap samples, a model is appropriate, and each of the parameters can be estimated. This process can be repeated many times, and provides an estimate of the sampling distribution, which can be used to infer the parameters. More about the description of bootstrap can be seen in the writings of Efron (1994).

Breiman use the process of “bootstrap aggregates” to implement the bootstrap method, which is bagging algorithm (Breiman, 1996a,b,c,d). Bagging is an elegant and simple integrated learning method. Bagging bootstrap process involves collecting samples and trained classifiers (such as decision trees), which predicted by aggregating or averaging each bootstrap prediction. Bagging need to use sub-samples which is selected randomly from the training data to construct a set of learning machine, including randomly selected data points and replaced. By selecting different sub-samples of random sampling for learning machine, integrated learning machine is not easy to fall into single or local optimum points. Thus each individual learning machine is training in a different random sub-sample of the whole dataset. The Bagging algorithms establish the integrated learning machine by selecting node randomly and replacing from the training dataset. Bagging integrated approach has been shown quite accurate, but their computation cost sometimes too high. To get maximum performance from bagging integrated approach, sometimes hundreds or even thousands of part learning machine integrated application. For even a small-scale problem, it is still a relatively large difficulty.

RF (random forest) (Breiman, 2001a,b,c) is a collection of tree-structured classifiers $\{h(x, \theta_k), k=1\ldots\}$, where the $x$ is an input vector and the $\{\theta_k\}$ are independent identically distributed random vectors, resulting in a classifier. New dataset is drawn by bagging, during this process generally two-thirds of bootstrap samples would be used, and the rest of samples would be treated as out-of-bag (OOB). Breiman used the OOB to estimate variable importance and the estimate error of the model. The RF regression algorithm is a bagging method that is based on the CART regression tree. RF employs recursive partitioning to divide the data into many homogenous subsets called regression ntrees and then averages the results of all of the trees. Each tree is independently grown to its maximum size based on a bootstrap sample from the training dataset without stopping the selection of the input variables at each node. In each tree, the RF uses randomness in the regression process by selecting a random subset of variables (mtry) to determine the random selection of the variable at each node (Breiman, 2001a,b,c). In each tree, the ensemble predicts the data that are not in the tree, and by calculating the difference in the mean square errors between the OOB (out of bag) data and the data that were used to grow the regression trees, the RF algorithm gives the OOB error of each variable. (Breiman, 2001a,b,c; Maidonald and Braun, 2006; Prasad et al., 2006; Palmer et al., 2007).

Modeling RF could balance the bias of different sets, and improve performance by tuning a few parameters; in addition, even the defaults present a high performance. Because of the Law of Large Numbers, RF do not over fit as supervised learning method. RF is an effective tool of dealing with big data, with fast learning ability and high computing speed. Moreover, it gives useful internal estimates of error, correlation and variable importance, and therefore, the results are explicable. There are three critical parameters: (a) the number of trees in the forest (ntree), (b) the minimum number of data points in each terminal node (node size), and (c) the number of features at each node (mtry). The default of mtry is one-third of independent variables, namely $mtry = \left\lfloor \frac{p}{3} \right\rfloor$ where $p$ is the
eigenvector of $x$. In this study, nodesize, ntree, and mtry are set to 5, 2000, 22, respectively.

The mse is the vector of mean square errors; the rsq is the "pseudo $R$-squared", pseudo multiple correlation coefficient, calculated by $1 − \text{mse/Var}(y)$ (Breiman, 1996a,b,c,d, 2001a,b,c). The formulas are as follows:

$$mse = \frac{1}{n} \sum_{i=1}^{n} (F(x_i) - y_i)^2$$

$$rsq = 1 - \frac{n - \sum_{i=1}^{n} (F(x_i) - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$  \hspace{1cm} (2)

Simply stated, the variable importance is evaluated based on how much worse the prediction would be if the data for those variables were permuted randomly (Prasad et al., 2006), and this can be used in the feature's selection by determining the importance of each variable in the regression process.

The vegetation indices were ranked according to their importance, and a forward selection function (Kohavi and John, 1997; Guyon and Elisseeff, 2003) was implemented to determine the least number of vegetation indices that can predict the chlorophyll concentration with greater accuracy. The forward selection procedure adds variables to the model one-by-one, and at each step the variable that is not in the model is included based on a probability threshold. The method initially begins with the most significant variable and continues adding variables until none of the remaining variables can satisfy the priority threshold.

The RF algorithm performs well when a large number of input variables are analyzed to build a model using a small number of samples for the output variable (Breiman, 2001a,b,c). The RF is a robust method against missing observations and can handle a large number of input variables as well as an unbalanced dataset. The method analyses both the categorical and continuous variables (Grimm et al., 2008).

2.3. Model accuracy evaluation

The RMSE represents a data point and its corresponding position on the regression line. The $R^2$ is the correlation between the predicted chlorophyll concentration and measured chlorophyll concentration. The higher the correlation value is, the stronger the linear relationships between the actual and predicted variables are.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (F(x_i) - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n}}$$  \hspace{1cm} (4)

3. Results and discussion

3.1. Spectral indices that were optimized for the CCC assessment based on ground-based measurements

Three experiment farmland levels with clean soil, light polluted soil and heavily polluted soil were used with safety, level 1 and level 2 corresponding crop conditions (Fig. 3). According to the spectral

![Fig. 3. Canopy reflectance curve of rice in sample sites.](image-url)
index under arsenic stress and the index structure, we can detect small changes in the CCC under arsenic stress.

The spectral band of visible light, as well as the “red edge” area, is an important indicator of the CCC and the health status description. The characteristic parameters of the red edge effect of the values are in Table 2, and with the increasing influence of arsenic metal elements, the red edge position has shown a slight “blue shift”, which is when the red edge width reduces, amplitude decreases, and peak area reduces; the absorption depth reduces, width reduces, and area decreases; and the red shoulder has a slight “blue shift” (Table 2).

As shown in Fig. 3, the three sample areas reflect the spectra of the safety, level 1, and level 2 crop conditions after continued removal in the rice area (Fig. 3). The green peak and red absorption depth have obvious differences. The green peak heights in the level 1 and level 2 samples were significantly higher than those of the safe sample, and the absorption valley depth was shallower. All of the calculated spectral characteristic parameters in the three samples are listed in Table 1.

1. The absorption peak position. The visible pigment substances in the leaves absorbed sunlight in the visible light wavebands, producing an extreme value approximately 680 to ~700 nm; therefore, the absorption valley appeared in the spectral reflectance curve. The location of the absorption valley is the spectral reflectance of the lowest point corresponding to the wavelength. The Hand Hz in Fig. 3 denotes the absorption depth of the green peaks and the absorption valley, respectively. The location of the absorption valley in the three samples is similar, in the vicinity of 679 nm, and the maximum change rate is 3 nm. This result shows that the location of the absorption valley is decided only by the properties of the absorbing materials and that the changes of the vegetation biochemical fraction will not affect the absorption mobile location.

2. The absorption depth. The absorption depth is the extent of the absorption in the wavelength. The absorption depth valley is decided by the vegetation leaf absorption of sunlight. The absorption capacity decreases, reflectivity increases, and the absorption depth of the spectral curve decreases, when rice is under metal stress. If the absorption solar capacity of the leaves is weak, the transmission rate is constant, the reflectivity increases, and the absorption depth is shallow. The reflectivity of the green peak and the absorption depth of the absorption peak are inversely changed.

3. The red edge position. The displacement of the red edge position (corresponding to Fig. 3R) is significant; when the pollution degree is enhanced, the red edge moves to a longer wavelength. The red edge position is a good indicator of arsenic stress.

Table 2

<table>
<thead>
<tr>
<th>Spectral indices</th>
<th>Safe</th>
<th>Level 1</th>
<th>Level 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample number</td>
<td>13</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>Green peak (nm)</td>
<td>557</td>
<td>569</td>
<td>632</td>
</tr>
<tr>
<td>Red edge (nm)</td>
<td>718.1666667</td>
<td>695.2105263</td>
<td>694.125</td>
</tr>
<tr>
<td>Red edge width (nm)</td>
<td>38.1666667</td>
<td>15.2105263</td>
<td>14.125</td>
</tr>
<tr>
<td>Red edge amplitude</td>
<td>0.00635167</td>
<td>0.005861579</td>
<td>0.004150475</td>
</tr>
<tr>
<td>Peak area of red edge</td>
<td>0.320861667</td>
<td>0.234473158</td>
<td>0.150663125</td>
</tr>
<tr>
<td>Absorption center (nm)</td>
<td>679.3</td>
<td>680.0526316</td>
<td>679</td>
</tr>
<tr>
<td>Absorption depth</td>
<td>0.783083333</td>
<td>0.569557895</td>
<td>0.3901625</td>
</tr>
<tr>
<td>Absorption width (nm)</td>
<td>83.4166667</td>
<td>54.26315789</td>
<td>43.8125</td>
</tr>
<tr>
<td>Absorption area</td>
<td>54.78135917</td>
<td>25.26086316</td>
<td>14.159825</td>
</tr>
<tr>
<td>Red peak (nm)</td>
<td>762</td>
<td>760</td>
<td>760</td>
</tr>
<tr>
<td>Near infrared absorption valley</td>
<td>980</td>
<td>935</td>
<td>935</td>
</tr>
<tr>
<td>Chlorophyll content (SPAD value)</td>
<td>35.02213333</td>
<td>22.0354</td>
<td>14.828625</td>
</tr>
</tbody>
</table>

(4) The absorption width and red shoulder position. The absorption width refers to the wavelength difference of the red edge position and the green peak (Fig. 3W). The experiments show that the absorption width is negatively correlated with the crop contaminated stress level: the greater the stress level, the smaller the absorption width. The red shoulder position (corresponding to Fig. 3R) is the band maximum light reflectance wavelength position. The experiments show that the red shoulder position is related to the degree of pollution; that is, the red shoulder position is the largest in unpoluted areas, due mainly to the maximum reflectivity in the maximum wavelength. This result is agreement with the red edge position, which also changes with the absorption width.

Fig. 4 shows the contour maps of R² between the CCC and the Rₐ (reflectance) or Dₐ (a first derivative). The most significant plots were R₉₆₉₃ and D₇₃₀, where the R² were, respectively, 0.59 and 0.82. These two bands correspond to the Red Valley location and the red shoulder position under arsenic stress, when the crop spectral curves in both locations are better reflected.

Fig. 5 shows that the NDSI (R₇₈₃, R₇₂₂) was the optimal point, with an R² of 0.88, the RSI (R₇₃₇₃, R₇₇₂₂) had the most best relationship with the CCC, with an R² of 0.87. The NDSI (D₇₃₂, D₇₀₂) was the most significant point, with an R² of 0.91, the RSI (D₇₄₂, D₇₂₀) had the most significant relationship with the CCC, with an R² of 0.90 (Fig. 6).

Fig. 7 shows the contour maps of R² between the CCC and the NVI (Rₐ, Rₗ, Rₜ₉₉) or the NVI (Dₙ, Dₗ, Dₚ₉₉) using the complete combination of three wavebands at i, j and k (nm), respectively, for the datasets from Chang Chun. un, the NVI (R₆₄₈, R₇₆₂, R₇₃₂) and NVI (D₆₄₈, D₇₅₂, D₇₇₁) were the most significant points, with an R² of 0.90 and 0.92, respectively.

3.2. Random forest model

3.2.1. RF regression models’ optimization

The optimum mtry values in Fig. 8 substantially differed from the default value (i.e., mtry = 1/3 of the total number of the variables). The results indicate that increasing the ntree value beyond 500 did not obviously affect the rsq and mse.

3.2.2. RF predictive models

Fig. 8 shows the optimal ntree and mtry values for the RF regression models that were developed in this study. Section (a1) is the relationship between the pseudo multiple correlation coefficient and the number of random forest trees. Section (b1) is the relationship between the tests set pseudo multiple correlation coefficients and the number of random forest trees. Section (a2)
Fig. 4. The relationship between CCC and reflectance or derivative.

is the relationship between the training set mean-square error vector and the number of random forest trees. Section (b2) is the relationship between the test set mean-square error vector and the number of random forest tree training sets. Section (a3) is the relationship between the predicted CCC and the measured CCC. Section (b3) is the relationship between the test set forecast CCC and the measured CCC.

The selected wavebands included seven indexes: TCARI/OSAVI, MCARI2, NDVI3A, MCARI2/OSAVI3, OSAVI, and OSAVI2 (Table 4). These indices are more highly correlated with the CCC than are the other indices. The figure shows that the accuracy of the Table 4.

This study presents a successful application of the machine-learning algorithm, RF. The predictive models that were developed in this study could be used to identify the canopy chlorophyll content when the models overestimate the CCC.

3.3. Various approaches comprehensive comparison

As shown in Table 3, it is the comparison of spectral indices' predictive ability for monitoring of canopy chlorophyll content, these indices which are more highly correlated with the CCC higher than 0.7 are the input variables of random forest model.

Fig. 5. A $R^2$ contour map between CCC and NDSI ($R_i, R_j$) and a $R^2$ contour map between CCC and RSI ($R_i, R_j$).
As shown in Fig. 9, the paddy areas are divided into no pollution, pollution level 1, and pollution level 2 and no rice based on the predicted CCC results. How can I divide it, because there is a relationship between the CCC and arsenic content. As shown in Fig. 10, the CCC estimated by NVI can diagnosis the space of As stress in rice.

3.4. CCC hyperspectral imagery for validation and mapping

This study set up a new index to estimate the chlorophyll concentration on the canopy scale; the scale effect will occur when the new index is applied to Hyperion image pixels due to the different measurements scales (Chehbouni et al., 2000). This study uses the liner transformation based on the input variable to reduce the scale effect, which is produced due to applying the new index to remote sensing image pixels from the rice canopy scale. Thus, the new index can estimate the rice chlorophyll concentration in high-precision and large areas on the remote sensing pixels scale. The scale transformation method is based on the input variables, namely the first linear transformation between the ASD and the Hyperion hyperspectral data. Then, the Hyperion hyperspectral data are included according to the linear transformation.
and the new index is used to build new Hyperion hyperspectral images according to the relationship between the index and the chlorophyll concentration. Finally, the chlorophyll concentration distribution remote sensing image is obtained. It is worth mentioning that because of the index structure, that the linear transformation step can be omitted without altering the result. As shown in Fig. 11, there is a scale transformation between Hyperion reflectance data and ASD reflectance data. There is no

![Graphs showing comparison of training set and validation set of random forest algorithm.](image)

**Table 3**
Comparison of spectral indices' predictive ability.

<table>
<thead>
<tr>
<th>Index</th>
<th>Formula</th>
<th>Reference</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blog(1/R737)</td>
<td>Blog(1/R737)</td>
<td>Yoder and Pettigrew-crosby (1995)</td>
<td>0.2059</td>
</tr>
<tr>
<td>TCARI2</td>
<td>3 × ((R750 – R705) − 0.2 × (R750 – R670) × (R750/R705))</td>
<td>Wu et al. (2008)</td>
<td>0.2945</td>
</tr>
<tr>
<td>TCARI</td>
<td>3 × ((R750 – R705) − 0.2 × (R750 – R670) × (R750/R705))</td>
<td>Haboudane et al. (2002)</td>
<td>0.3270</td>
</tr>
<tr>
<td>MCAIR/OSAVI</td>
<td>[(R705 – R670) − 0.2 × (R750 – R670)] × (R750/R670) × [1 + 0.16 × (R670 – R705)]×(R670 × R705 × 0.16)]</td>
<td>Daughtry (2001)</td>
<td>0.3707</td>
</tr>
<tr>
<td>EVI</td>
<td>2.5 × ((R750 – R705)/(R750 – (7.5 × R705)) + 1)</td>
<td>Huete et al. (1997)</td>
<td>0.4048</td>
</tr>
<tr>
<td>MCAIRI</td>
<td>((R705 – R670) − 0.2 × (R750 – R670) × (R750/R705))</td>
<td>Daughtry et al. (2000)</td>
<td>0.4057</td>
</tr>
<tr>
<td>TCARI2/OSAVI2</td>
<td>[3 × ((R750 – R705) − 0.2 × (R750 – R670) × (R750(R705))]×[(1 + 0.16 × (R750 – R705))×(R750 × R705 × 0.16)]</td>
<td>Wu et al. (2008)</td>
<td>0.5230</td>
</tr>
<tr>
<td>DDo</td>
<td>2.5 × (R710 – R710) – 0.2 × (R710 × R710) × (R705/R670)</td>
<td>Le Maire et al. (2008)</td>
<td>0.5966</td>
</tr>
<tr>
<td>MSAVI</td>
<td>0.5 × (2 × R630 + 1 – SQRT(2 × R630 + 1 + 2 × (R630 – R630))</td>
<td>Qi et al. (1994)</td>
<td>0.6528</td>
</tr>
<tr>
<td>RAVI</td>
<td>(R670 – R730)/(SQRT(R630 × R630))</td>
<td>Roujean and Breen (1995)</td>
<td>0.6851</td>
</tr>
<tr>
<td>Sum_Dr1A</td>
<td>Sum of first derivative reflectance between R625 and R725</td>
<td>Elvidge and Chen (1995)</td>
<td>0.6783</td>
</tr>
<tr>
<td>TCARI/OSAVI</td>
<td>[3 × ((R750 – R670) − 0.2 × (R750 – R670) × (R750/R670))]×[(1 + 0.16 × (R670 – R705))×(R670 × R705 × 0.16)]</td>
<td>Haboudane et al. (2002)</td>
<td>0.7134</td>
</tr>
<tr>
<td>MCAIRI2</td>
<td>[(R750 – R705) − 0.2 × (R750 – R670) × (R750/R705)]</td>
<td>Wu et al. (2008)</td>
<td>0.7329</td>
</tr>
<tr>
<td>NDVI3A</td>
<td>(R625 – R515)/(R625 × R515)</td>
<td>Gandia et al. (2004a, b)</td>
<td>0.7368</td>
</tr>
<tr>
<td>MTCI</td>
<td>(R750 – R705)/(R750 × R705)</td>
<td>Dash and Curran (2007)</td>
<td>0.7444</td>
</tr>
<tr>
<td>MCAIR2/OSAVI3</td>
<td>[(R750 – R670) − 0.2 × (R750 – R670) × (R750/R670)]×[(1 + 0.16 × (R750 – R705))×(R750 × R705 × 0.16)]</td>
<td>Wu et al. (2009a, b)</td>
<td>0.7461</td>
</tr>
<tr>
<td>OSAVI</td>
<td>(1 + 0.16) × (R670 – R710)/(R670 × R710 × 0.16)</td>
<td>Rondeaux et al. (1996)</td>
<td>0.7559</td>
</tr>
<tr>
<td>OSAVI2</td>
<td>(1 + 0.16) × (R710 – R750)/(R710 × R750 × 0.16)</td>
<td>Wu et al. (2008)</td>
<td>0.7568</td>
</tr>
<tr>
<td>NDVI</td>
<td>(R670 – R730)/(R670 × R730)</td>
<td>Tucker (1979)</td>
<td>0.7741</td>
</tr>
</tbody>
</table>
Table 4
Input variables of RF model.

<table>
<thead>
<tr>
<th>Index</th>
<th>Formulation</th>
<th>Reference</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCARI/OSAVI</td>
<td>$3 \times [(R_{700} - R_{670}) - 0.2 \times (R_{700} - R_{550}) \times (R_{700} / R_{670})]] \times [1 + 0.16 \times (R_{800} - R_{705}) / (1 + 0.16 \times (R_{800} + R_{705} + 0.16))]$</td>
<td>Haboudane et al. (2002)</td>
<td>0.7134</td>
</tr>
<tr>
<td>MCAVI2</td>
<td>$[(R_{750} - R_{705}) - 0.2 \times (R_{750} - R_{550}) \times (R_{750} / R_{705})]$</td>
<td>Wu et al. (2008)</td>
<td>0.7329</td>
</tr>
<tr>
<td>NDVII3A</td>
<td>$(R_{682} - R_{553}) / (R_{682} + R_{553})$</td>
<td>Gandia et al. (2004a,b)</td>
<td>0.768</td>
</tr>
<tr>
<td>MTCI</td>
<td>$(R_{754} - R_{718}) / (R_{718} - R_{681})$</td>
<td>Dash and Curran (2007)</td>
<td>0.7444</td>
</tr>
<tr>
<td>MCARI2/OSAVI3</td>
<td>$[((R_{750} - R_{705}) - 0.2 \times (R_{750} - R_{550}) \times (R_{750} / R_{705})] / [1 + 0.16 \times (R_{750} - R_{705}) / (R_{750} + R_{705} + 0.16)]$</td>
<td>Wu et al. (2009a,b)</td>
<td>0.7461</td>
</tr>
<tr>
<td>OSAVI</td>
<td>$1 \times (R_{800} - R_{705}) / (R_{800} + R_{705} + 0.16)$</td>
<td>Rondeaux et al. (1996)</td>
<td>0.7559</td>
</tr>
<tr>
<td>OSAVI2</td>
<td>$1 \times (R_{750} - R_{705}) / (R_{750} + R_{705} + 0.16)$</td>
<td>Wu et al. (2008)</td>
<td>0.7568</td>
</tr>
</tbody>
</table>

Fig. 9. The prediction results by using four indexes.

fixed relationship between the spectral reflectance based on single band, three band emission rate showed a linear relationship. NVI can be directly used for Hyperion image on CCC estimation and Arsenic pollution level division.

As shown in Fig. 12, it is the distribution of canopy chlorophyll content in area B which is estimated by NVI ($R_{640}$, $R_{732}$, $R_{752}$). According to the distribution of canopy chlorophyll concentration, we can do a rough division of arsenic pollution levels.

As shown in Fig. 12, it is a prediction results of pollution stress level of rice base on Hyperion data in area B. Due to this the fall range with the chlorophyll content of chlorophyll content corresponding to the pollution of heavy metals pollution levels in the characterization of stress level, and the level of pollution is divided, and large area coverage of this method will be extended to the effects of Hyperion (Fig. 12), in the range of Hyperion image on chlorophyll content for rice area 8–110 μg/cm². In Fig. 12, the CCC is corresponding with the pollution level. The results show that

Fig. 10. The relationship between the As stress and CCC estimated by NVI.
Fig. 11. Scale transformation of Hyperion image from ASD data.

Fig. 12. Distribution of canopy chlorophyll concentration in study area b.

Table 5
Prediction results of pollution stress level of rice based on Hyperion data.

<table>
<thead>
<tr>
<th>Pollution levels</th>
<th>Each sampling samples in actual pollution levels</th>
<th>NVI predict the sample number in each pollution levels</th>
<th>Discriminate accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1 pollution</td>
<td>1</td>
<td>1</td>
<td>90%</td>
</tr>
<tr>
<td>Level 2 pollution</td>
<td>10</td>
<td>9</td>
<td>89%</td>
</tr>
<tr>
<td>No pollution</td>
<td>6</td>
<td>7</td>
<td>83%</td>
</tr>
</tbody>
</table>
much of the study area is not polluted, and a few area is with level 1 pollution (Table 5). It has a similar accuracy of estimating CCC by NVI derived from ASD and Hyperion (Fig. 13).

4. Conclusion

The new index is based on the single band index, the normalized vegetation index, the ratio vegetation index and the three bands of the vegetation index. Compared to the related indices or models for the inversion of the chlorophyll concentration, the new index has the highest accuracy and is most sensitive to the small changes of chlorophyll concentration under arsenic stress. The spectral reflectance curve at the edge of the red area of the crops, including the red valley area to the entire range of the shoulder area, is obviously changed in rice under arsenic stress; the new index band values are also within this range.

This study attempted to reduce the scale effects result from different measurement scales based on the linear transformation of the input variables when using the new index to estimate the chlorophyll concentration, making the new index better when applied to a range of remote sensing images. Due to the particularity of the new index structure, the linear transformation is based on the input variables at the offset; therefore, the new index can be used directly with the remote sensing images to evaluate large areas of the rice chlorophyll concentration.

This study compared the new index with not only the related chlorophyll indices but also with the random forest algorithm of the machine learning algorithm. The results show that the new index has greater accuracy for estimating the CCC under the stress of arsenic and is more sensitive to arsenic stress.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (41371407). The authors wish to thank the anonymous reviewers for their constructive comments that helped improve the scholarly quality of the paper.

References
