Determination of the quantitative content of chlorophylls in leaves by reflection spectra using the random forest algorithm

E.A. Urbanovich, D.A. Afonnikov, S.V. Nikolaev

Abstract. Determining the quantitative content of chlorophylls in plant leaves by their reflection spectra is an important task both in monitoring the state of natural and industrial phytocenoses, and in laboratory studies of normal and pathological processes during plant growth. The use of machine learning methods for these purposes is promising, since these methods allow inferring the relationships between input and output variables (prediction model), and in order to improve the quality of the prediction, a researcher may modify predictors and selects a set of method parameters. Here, we present the results of the implementation and evaluation of the random forest algorithm for predicting the total concentration of chlorophylls \( a \) and \( b \) from the reflection spectra of plant leaves in the visible and infrared wavelengths. We used the reflection spectra for 276 leaf samples from 39 plant species obtained from open sources. 181 samples were from the sycamore maple (\( Acer \) pseudoplatanus L.). The reflection spectrum represented wavelengths from 400 to 2500 nm with a step of 1 nm. The training set consisted of the 85 % of \( A. \) pseudoplatanus L. samples, and the performance was evaluated on the remaining 15 % samples of this species (validation sample). Six models based on the random forest algorithm with different predictors were evaluated. The selection of control parameters was performed by cross-checking on five partitions. For the first model, the intensity of the reflection spectra without any transformation was used. Based on the analysis of this model, the optimal ranges of wavelengths for the remaining five models were selected. The best results were obtained by models that used a two-point estimation of the derivative of the reflection spectrum in the visible wavelength range as input data. We compared one of these models (the two-point estimation of the derivative of the reflection spectrum in the range of 400–800 nm with a step of 1 nm) with the model by other authors (which is based on the functional dependence between two unknown parameters selected by the least squares method and two reflection coefficients, the choice of which is described in the article). The comparison of the results of predictions of the model based on the random forest algorithm with the model of other authors was carried out both on the validation sample of maple and on the sample from other plant species. In the first case, the predictions of the method based on a random forest had a lower estimate of the standard deviation. In the second case, the predictions of this method had a large error for small values of chlorophyll, while the third-party method had acceptable predictions. The article provides the analysis of the results, as well as recommendations for using this machine learning method to assess the quantitative content of chlorophylls in leaves.

Key words: random forest; remote methods; leaf optics; pigments.

For citation: Urbanovich E.A., Afonnikov D.A., Nikolaev S.V. Determination of the quantitative content of chlorophylls in leaves by reflection spectra using the random forest algorithm. Vavilovskii Zhurnal Genetiki i Seleksii = Vavilov Journal of Genetics and Breeding. 2021;25(1):64-70. DOI 10.18699/VJ21.008

Определение количественного содержания хлорофиллов в листьях по спектрам отражения алгоритмом случайного леса

Е.А. Урбанович, Д.А. Афонников, С.В. Николаев

Аннотация. Определение количественного содержания хлорофиллов в листьях растений по их спектрам отражения – важная задача как при мониторинге состояния естественных и промышленных фитоценозов, так и в лабораторных исследованиях нормальных и патологических процессов в ходе роста растения. Применение для этих целей методов машинного обучения является перспективным, поскольку они позволяют «автомати-
**Introduction**

Pigments are low-molecular-weight compounds that give color to plant organs and play an important role in their life, performing photosynthetic, protective and metabolic functions. In terrestrial plants, the most well-known pigments are chlorophylls (which provide the green color of plant organs and play a crucial role in photosynthesis), carotenoids (which give red and yellow color, also participate in photosynthesis), anthocyanins (which give a purple color, perform protective functions), as well as a number of other compounds (Croft, Chen, 2018). Photosynthetic pigments, chlorophylls and carotenoids, attract the most attention from researchers; they have different absorption spectra and perform different functions in the process of photosynthesis, which is due to structural differences between the molecules of these substances.

Chlorophyll in plants is represented by two types of molecules, \( \text{a} \) and \( \text{b} \), which have structural differences and differ in their light-absorbing properties (Du et al., 1998). It allows photosynthetic organisms to collect sunlight at different wavelengths to maximize the light energy available for photosynthesis. Changes in the concentrations of photosynthetic pigments are closely related to the physiological state of plants. For example, when the leaves of plants wither, there is a rapid decrease in the concentration of chlorophylls compared to carotenoids, thereby increasing the ratio of carotenoids to chlorophylls causes the leaves to turn red and yellow (Croft, Chen, 2018). The content of pigments, in particular chlorophylls \( \text{a} \) and \( \text{b} \), can thus serve as an indicator of the state of plants during normal growth and during the development of infections, as well as stress, photosynthetic activity, metabolic disorders, etc. (Modzińska, 2009). The need to determine the physiological state of plants often arises in the course of solving many scientific and practical problems, so methods for assessing the content of pigments in plant organs and tissues are constantly being developed and improved.

Quantitative and qualitative information about pigments can be obtained using chemical methods (Lichtenthaler, 1987; Porra et al., 1989; Wellburn, 1994). However, for many tasks, a more convenient approach is to use remote methods based on the light reflection spectra from the plant leaf (Horler et al., 1983; Curran et al., 1990; Gitelson et al., 2001, 2003). The reflectivity of the leaf in the optical and infrared (IR) wavelengths (400–2500 nm) depends on various biochemical and physical factors, including the content of chlorophyll and other leaf pigments, nitrogen, water, as well as on the internal structure of the leaves and the characteristics of their surface (Croft, Chen, 2018). Plant pigments are characterized by the absorption of electromagnetic radiation in the visible (400–700 nm) and near-IR (1300–2500 nm) wavelength ranges. The absorption of the leaf components in the near-infrared region in the range of 750–1300 nm is low, since in this wavelength range there is an intense reflection from the components of the internal structure of the leaves. Thus, the reflection coefficient in the near-IR range depends on both the concentration of enzymes and the structure of the leaf. All these facts make it possible to use remote observation methods in both the visible and near-infrared wavelength ranges to monitor the physiological state of plants (Merzlyak et al., 2003; Alt et al., 2020).

One of the approaches to estimating the content of chlorophylls from the reflection spectrum is to select empirical dependencies (indices) between the reflection coefficients at certain wavelengths, the choice of which is also an important part of the method, and the content of chlorophylls (Horler et al., 1983; Curran et al., 1990; Gitelson et al., 2001, 2003; Sua et al., 2010; Nikolaev et al., 2018). The success of such a “classical” approach directly depends on the depth of our understanding of the physics of the process.

Currently, machine learning methods are often used to predict the characteristics of biological objects (Doktor et
To predict the chlorophyll concentrations by the random forest method, several models that differed in the input data sets were taken. First, each set was characterized by an interval of wavelengths, the intensity of reflection at which was taken into account. In total, several sets of intervals were considered: 400–2450, 400–800 nm, and a combined set of two intervals of 500–600 and 680–740 nm. Second, the models differed in the type of input data. These included the values of the intensity of the reflection spectra at certain wavelengths (base data type), the values of the first derivatives of the spectral curves for the same wavelengths (der data type), and the values of the second derivatives (der2 data type). Some models were based on only one data type, while others shared multiple data types. Such combinations were marked with a summation sign (for example, base+der).

In this paper, six models have been considered. They are designated as RF-(X–Y)-Z, where (X–Y) – intervals of wavelengths, Z – type data model: RF-(400–2450)-base (the intensity of the spectrum in intervals of wavelengths 400–2450 nm); RF(400–800)-base (the intensity of the spectrum in intervals of wavelengths 400–800 nm); RF(400–800)-base+der (intensity spectrum and the first derivative in the intervals of wavelengths 400–800 nm); RF(400–800)-der (first derivative in the intervals of wavelengths 400–800 nm); RF(400–800)-der+der2 (first and second derivatives in the interval of wavelengths 400–800 nm); RF(500–600; 680–740)-base+der+der2 (intensities, first and second derivatives in the wavelength ranges 500–600 and 680–740 nm).

As an approximation of the derivative of the spectral curves, the first-order finite difference with a change equal to 1 was used, which was calculated by the formula $D_i = R_i - R_{i-1}$. In this calculation, there is no derivative for the first value. For simplicity, the finite difference is referred to the derivative throughout the text. The second derivative was calculated as the derivative of the derivative of the spectral curve.

When configuring the random forest algorithm, the following control parameters were selected:

- max_depth: [2, 3, 4, 5, 6] – the maximum depth of the tree;
- max_features: [2, 7, sqrt, log2, auto] – the number of features that the partition is searched for (auto – all features);
- n_estimators: [5, 10, 15, 30, 40] – the number of trees in the random forest ensemble;
- random_state: 20200605.

The specified parameters of the algorithm were selected by cross-checking on five samples of the same size obtained from a randomly mixed initial training sample. Four subsamples were used for training the model, and the fifth one was used for testing it. To determine the best control parameters, the test results (mean square deviation of the target indicator – mse) were averaged between models with the same control parameters (i.e., obtained during cross-validation) and sorted. The control parameters for which the average mse is the minimum are the best. As the final model, one of the five models with the best control parameters is selected, which has the minimum mse when tested among the models obtained by the cross-validation method.

The maximum depth of the trees is chosen to be 6, which gives $2^6 = 64$ intervals for partitioning the parameter space, despite the fact that the sample length taken to build the model
Определение количественного содержания хлорофиллов в листьях по спектрам отражения

Е.А. Урбанович, Д.А. Афонников
С.В. Николаев
2021

is 123. The depth increasing could lead to overfitting. The number of trees in the forest (up to 40) may seem redundant for 123 sample values, but the parameters of each of the decision trees were selected on different subspaces (since the random subspace method is used), and the dimension of the features was always greater than the number of elements in the sample.

It should be noted that the algorithm implemented in the sklearn library allows us to obtain the informativeness of each of the model features and select the most informative ones for the obtained decision rules (Breiman, 2001; Hastie et al., 2009; Louppe et al., 2013).

**Construction of empirical functional dependencies.** As a functional of \( \tilde{f}: R_2 \to \tilde{C}hl \) we additionally chose an empirical dependence from the work (Gitelson et al., 2003) (the GGM method, which we named after the authors’ surnames), represented by the expression

\[
\tilde{C}hl = \alpha \left[ \frac{1}{R_{\lambda}} \right] R_{\text{NIR}} + \beta,
\]

where \( \tilde{C}hl \) is the total concentration of chlorophylls \( a \) and \( b \); \( R_{\lambda} \) is the reflection coefficient at the wavelength \( \lambda \); \( R_{\text{NIR}} \) is the reflection coefficient in the near-infrared range (for example, at a wavelength of 800 nm); \( \alpha \) and \( \beta \) are selected in such a way as to minimize the selected loss function. A.A. Gitelson and co-authors (2003) recommend choosing wavelengths from the range \( \lambda \in [525; 555] \cup [695; 725] \). According to the authors, the advantage of this algorithm is that the \( R_{\text{NIR}} \) coefficient “corrects” the influence of the plant tissue structure on the reflection spectrum and allows us to extend the found function to plants with different leaf structure.

**The comparison of methods for predicting the concentration of chlorophyll.** The sycamore maple sample from the angers2003 data set was randomly divided into a training and a validation sample in the ratio of 85 : 15. For the methods used in this work for predicting the random forest algorithm (RF) and functional dependence (GGM), the optimal parameters are selected on the training sample. The quality control of the algorithms is carried out on a validation sample represented by a sycamore maple and on a sample of non-maple samples. The following metrics were used to evaluate the accuracy of predicting chlorophyll concentrations: \( mse \), mean absolute error (mae), and determination coefficient \( R^2 \). The formulas for calculating metrics are as follows:

\[
mse = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2,
\]

\[
mae = \frac{1}{n} \sum_{i=1}^{n} |x_i - \hat{x}_i|,
\]

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2},
\]

where \( x \) is the true values; \( \hat{x} \) is the predicted values; \( n \) is the number of samples, and \( \overline{x} \) is the mathematical expectation for the true values. In terms of optimization, \( mae \) and \( R^2 \) are equivalent. The coefficient of determination \( R^2 \) is convenient because it is a dimensionless value usually in the range \([0; 1]\), the value of \( R^2 < 0 \) shows that the arithmetic mean \( \overline{x} \) has a better result than the predictions of the constructed model.

**Results**

**Selection of parameters for the functional dependence method.** For the GGM prediction on the training sample, we selected the coefficients \( \alpha \) and \( \beta \) of equations (1), as well as the values \( \lambda \) so as to maximize the value of \( R^2 \). The value \( \lambda_{\text{NIR}} = 800 \text{ nm} \) is selected as the wavelength in the near-infrared range. To get the coefficients \( \alpha \) and \( \beta \), we took a linear model based on the least squares method (the LinearRegression class from the sklearn.linear_model package). For each \( \lambda \in [400; 800] \) with a step of 1 nm, a specific type of GGM curve was found. The coefficients of determination \( R^2 \) for the predictions of the obtained models are shown in Fig. 1. The highest coefficient of determination was achieved at the wavelength \( \lambda = 705 \text{ nm} \). The result is consistent with the recommended range \( \lambda \in [525; 555] \cup [695; 725] \). (Gitelson et al., 2003). The RF method is compared with the GGM model obtained at this wavelength \( \lambda = 705 \text{ nm} \).

**Results of constructing an algorithm based on the random forest method.** The characteristics of the accuracy of the prediction of chlorophyll concentrations (the values of the \( mse \), \( mae \), \( R^2 \) parameters) for all six models in the test sample are shown in the table. The RF-(400–800)-der and RF-(400–800)-der+der2 methods demonstrated high prediction accuracy. As the best of them, the RF-(400–800)-der method was selected as having a smaller number of input parameters.

The selection of wavelengths, the reflection coefficients for which were taken as input features for predicting chlorophyll concentrations by the random forest method, was carried out on the basis of the first model (RF-(400–2450)-base). This is due to the fact that at first it was not known whether the entire spectrum was needed, or only a part of it was necessary, and which one. As mentioned earlier, the RF algorithm allows you to evaluate the information content of the features the training took place on. After configuring the control parameters of the RF-(400–2450)-base model, we took the obtained parameters to re-train the models on five training samples (from cross-validation). For these five models, we identified 10 features with the greatest contribution to the prediction. The results are shown in Fig. 2: the vertical lines represent the combined set of wavelengths, the spectrum intensities for which make the most significant contribution to the prediction accuracy (26 wavelengths out of 10 · 5 = 50 possible if the values did not intersect). Interestingly, the most significant features lie in

«Fig. 1. Determination coefficients of the obtained GGF models at \( \lambda \in [400; 800] \), which were calculated on the training sample.»
Determination of the quantitative content of chlorophylls in leaves by reflection spectra

E.A. Urbanovich, D.A. Afonnikov, S.V. Nikolaev

Results of a random forest model trained on different sets of input features

| No. | Random forest model | Number of input features | mse  | mae  | $R^2$ |
|-----|---------------------|--------------------------|------|------|-------|
| 1   | RF-(400–2450)-base  | 2051                     | 30.5 | 3.7  | 0.945 |
| 2   | RF-(400–800)-base   | 401                      | 26.6 | 3.8  | 0.952 |
| 3   | RF-(400–800)-base+der| 401 + 400 = 801          | 10.1 | 2.4  | 0.981 |
| 4   | RF-(400–800)-der    | 400                      | 9.1  | 2.4  | 0.984 |
| 5   | RF-(400–800)-der+der2| 400 + 399 = 799          | 8.9  | 2.3  | 0.984 |
| 6   | RF-(500–600; 680–740)-base+der+der2| 101 + 100 + 99 + 61 + 60 + 59 = 380| 10.5 | 2.7  | 0.981 |

Note: The numbers in the description of the feature indicate the range of wavelengths. Additional characteristics of the features: base – reflection spectrum; der – values of the first derivative of the spectrum; der2 – values of the second derivative of the spectrum. The values with the worst accuracy are shown in italics, and the values with the best accuracy are highlighted in bold.

Fig. 2. Characteristics of the reflection spectrum of sycamore maple pigment samples used for model training.

The lines show: the average value of the intensity of the reflection spectrum $R_\lambda$ (Y-axis) for different wavelengths (X-axis); the value of the first derivative of the average intensity; the value of the second derivative. The values of the derivatives are normalized to the interval [0; 1]. Vertical lines indicate the wavelengths whose spectrum intensities make the greatest contribution to the prediction accuracy of the RF-(400–2450)-base model.

The lines show: the average value of the intensity of the reflection spectrum $R_\lambda$ (Y-axis) for different wavelengths (X-axis); the value of the first derivative of the average intensity; the value of the second derivative. The values of the derivatives are normalized to the interval [0; 1]. Vertical lines indicate the wavelengths whose spectrum intensities make the greatest contribution to the prediction accuracy of the RF-(400–2450)-base model.

Fig. 3. Comparison of true and predicted values of chlorophyll concentration in sycamore maple leaf tissues for validation sampling.

Fig. 4. Comparison of true and predicted values of chlorophyll concentration in leaf tissues of not related to sycamore maple samples.

Comparison of the accuracy of the RF and GGM methods. The results of the comparison of the methods for predicting chlorophyll concentrations by the RF-(400–800)-der and GGM methods and their experimentally measured values at different concentrations are shown in Fig. 3 and 4. For sycamore maple samples (the type taken to fit the parameters), the RF-(400–800)-der method shows a better result compared to the GGM method: $\sqrt{\text{mse}_{\text{RF}}}$ = 3.01 $\mu g/cm^2$ versus $\sqrt{\text{mse}_{\text{GGM}}}$ = 3.21 $\mu g/cm^2$. When testing the methods on a sample of plant leaves from other species, the GGM functional dependence method has an advantage $\sqrt{\text{mse}_{\text{GGM}}}$ = 6.31 $\mu g/cm^2$ versus $\sqrt{\text{mse}_{\text{RF}}}$ = 12.97 $\mu g/cm^2$. The GGM method shows high accuracy at low concentrations of chlorophyll, while the RF method shows a large error at these values. However, in the range of chlorophyll concentrations above 20 $\mu g/cm^2$, the RF-(400–800)-der algorithm has the best result: $\sqrt{\text{mse}_{\text{RF}}}$ = 5.91 $\mu g/cm^2$ versus $\sqrt{\text{mse}_{\text{GGM}}}$ = 7.01 $\mu g/cm^2$. Further analysis revealed that for samples with a chlorophyll concentration of less than 7 $\mu g/cm^2$, the reflection coefficients $R_{\text{CSM}}$ (maximum of the reflection spectrum) and $R_{\text{CSM}}$ (minimum of the reflection spectrum) are visually significantly different from all the others (Fig. 5, points in the upper right quarter). The predictions for these samples have a significant error. However, it was not possible to find out what the differences in the reflection spectrum are related to: these samples do not differ from the rest either in the surface density of the leaf or in the equivalent water thickness for the leaf (Jacquemound et al., 2003). Six out of ten plant species from these samples also have samples with normally predicted values. Further analysis of the causes of the anomalous spectrum is difficult, since the data are taken from open sources, and the measurements themselves were carried out more than 17 years ago.

Further analysis revealed that for samples with a chlorophyll concentration of less than 7 $\mu g/cm^2$, the reflection coefficients $R_{\text{CSM}}$ (maximum of the reflection spectrum) and $R_{\text{CSM}}$ (minimum of the reflection spectrum) are visually significantly different from all the others (Fig. 5, points in the upper right quarter). The predictions for these samples have a significant error. However, it was not possible to find out what the differences in the reflection spectrum are related to: these samples do not differ from the rest either in the surface density of the leaf or in the equivalent water thickness for the leaf (Jacquemound et al., 2003). Six out of ten plant species from these samples also have samples with normally predicted values. Further analysis of the causes of the anomalous spectrum is difficult, since the data are taken from open sources, and the measurements themselves were carried out more than 17 years ago.
Определение количественного содержания хлорофиллов в листьях по спектрам отражения

Е.А. Урбанович, Д.А. Афонников
С.В. Николаев

2021
25 • 1

References

Alt V.V., Gurova T.A., Elkin O.V., Klimenko D.N., Maximov L.V., Pestunov I.A., Dubrovskaya O.A., Genaev M.A., Erst T.V., Genaev K.A., Komyshev E.G., Khlestkin V.K., Afonnikov D.A. The use of Specim IQ, a hyperspectral camera, for plant analysis. Vavilovskii Zhurnal Genetiki i Selektsi = Vavilov Journal of Genetics and Breeding. 2020;24(3):259-266. DOI 10.18699/VJ19.587. (in Russian)

Breiman L. Bagging predictors. Mach. Learn. 1996;24:123-140. DOI 10.1023/A:1018054314350.

Breiman L. Random forests. Mach. Learn. 2001;45(1):5-32. DOI 10.1023/A:1010933404324.
Determination of the quantitative content of chlorophylls in leaves by reflection spectra

E.A. Urbanovich, D.A. Afonnikov
S.V. Nikolaev

ORCID ID
E.A. Urbanovich orcid.org/0000-0003-0602-3097
D.A. Afonnikov orcid.org/0000-0001-9738-1409

Acknowledgements. The work was supported by Russian Foundation for Basic Research No. 17-29-08028 and budget project No. 0259-2021-0009.

Conflict of interest. The authors declare no conflict of interest.

Received October 15, 2020. Revised December 14, 2020. Accepted December 15, 2020.