Вавиловский журнал генетики и селекции. 2021;25(1):64-70
DOI 10.18699/VJ21.008

Английский текст https://vavilov.elpub.ru/jour

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

Е.А. Урбанович[1] ✉, Д.А. Афонников[2, 3], С.В. Николаев[2, 4]

[1] Новосибирский государственный технический университет, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[4] Московская государственная академия ветеринарной медицины и биотехнологии – МВА им. К.И. Скрябина, Москва, Россия
✉ e.urbanovich98@gmail.com

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 «автоматически» строить решающие правила для получения результата (модель предсказания), а исследователю (для повышения качества предсказания) остаются модификация предикторов и выбор множества параметров метода. В статье приведены результаты построения решающих правил алгоритмом случайного леса (random forest) для предсказания суммарной концентрации хлорофиллов *a* и *b* по спектрам отражения листьев растений в видимом и инфракрасном (ИК) диапазонах длин волн. Набор данных взят из открытых источников. Они включали 276 образцов листьев 39 видов растений. При этом 181 образец получен при анализе листьев белого клена (*Acer pseudoplatanus* L.). Спектр отражения представлен в диапазоне 400–2500 нм с шагом 1 нм. Обучение происходило на 85 % образцов *A. pseudoplatanus* L., оценка качества предсказания – на оставшихся 15 % образцов этого вида (валидационная выборка). Построено шесть моделей на основе алгоритма случайного леса с разными предикторами. Подбор управляющих параметров осуществляли при помощи перекрестной проверки на пяти разбиениях. Предикторами первой модели выступали имеющиеся значения по спектру отражения без какой-либо обработки с нашей стороны. После проведения анализа этой модели были выбраны диапазоны длин волн предикторов для оставшихся пяти моделей. Лучшие предсказания имеют модели с разностной производной спектра отражения в видимом диапазоне длин волн. Модель с первой производной спектра отражения в диапазоне 400–800 нм с шагом 1 нм брали для сравнения с моделью других авторов. Этой моделью выступает функциональная зависимость с двумя неизвестными параметрами, подбираемыми методом наименьших квадратов и двумя коэффициентами отражения, выбор которых описывается в настоящей статье. Сравнение результатов предсказаний модели с применением алгоритма случайного леса проводили как на валидационной выборке клена, так и на выборке из других видов растений. В первом случае предсказания метода на основе случайного леса имели меньшую оценку среднеквадратического отклонения. Во втором случае предсказания этого метода были с большой ошибкой при малых значениях хлорофилла, в то время как сторонний метод имел приемлемые предсказания. В статье приводятся анализ результатов и рекомендации по применению этого метода машинного обучения для оценки количественного содержания хлорофиллов в листьях.
Ключевые слова: случайный лес; дистанционные методы; оптика листа растения; пигменты.

Для цитирования: Урбанович Е.А., Афонников Д.А., Николаев С.В. Определение количественного содержания хлорофиллов в листьях по спектрам отражения алгоритмом случайного леса. *Вавиловский журнал генетики и селекции*. 2021;25(1):64-70. DOI 10.18699/VJ21.008

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

E.A. Urbanovich[1] ✉, D.A. Afonnikov[2, 3], S.V. Nikolaev[2, 4]

[1] Novosibirsk State Technical University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
[4] Moscow State Academy of Veterinary Medicine and Biotechnology – MVA named after K.I. Skryabin, Moscow, Russia
✉ e.urbanovich98@gmail.com

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 Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):64-70. DOI 10.18699/VJ21.008

Введение

Пигменты – низкомолекулярные соединения, которые придают окрашивание органам растений и играют в их жизни важную роль, выполняя фотосинтетические, защитные и метаболические функции. У наземных растений наиболее известными пигментами являются хлорофиллы (обеспечивают зеленую окраску органов растений и играют важнейшую роль в фотосинтезе), каротиноиды (придают красную и желтую окраску, также участвуют в фотосинтезе), антоцианы (обеспечивают фиолетовую окраску, выполняют защитные функции), а также ряд других соединений (Croft, Chen, 2018). Фотосинтетические пигменты, хлорофиллы и каротиноиды, привлекают наибольшее внимание исследователей, они имеют разные спектры поглощения и выполняют в процессе фотосинтеза разные функции, что обуславливается структурными различиями между молекулами этих веществ.

Хлорофилл в растениях представлен молекулами двух типов, *a* и *b*, которые имеют структурные отличия и различаются по своим светопоглощающим свойствам (Du et al., 1998). Это позволяет фотосинтезирующим организмам собирать солнечный свет на различных длинах волн, чтобы максимизировать энергию света, доступную для фотосинтеза. Изменение концентраций фотосинтетических пигментов тесно связано с физиологическим состоянием растений. Например, при увядании листьев растений происходит быстрое снижение концентрации хлорофиллов по сравнению с каротиноидами, тем самым увеличивается отношение содержания каротиноидов к хлорофиллам, что вызывает появление у листьев окраски красных и желтых оттенков (Croft, Chen, 2018). Содержание пигментов, в частности хлорофиллов *a* и *b*, таким образом, может служить индикатором состояния растений в ходе нормального роста и при развитии инфекций, а также стресса, фотосинтетической активности, нарушения метаболизма и т. д. (Młodzińska, 2009). Потребности в определении физиологического состояния растений часто возникают в ходе решения многих научных и практических задач, поэтому методы оценки содержания пигментов в органах и тканях растений постоянно развиваются и совершенствуются.

Количественную и качественную информацию о пигментах можно получить с использованием химических методов (Lichtenthaler, 1987; Porra et al., 1989; Wellburn, 1994). Однако для многих задач более удобный подход – применение дистанционных методов на основе спектров отражения света от листа растения (Horler et al., 1983; Curran et al., 1990; Gitelson et al., 2001, 2003). Отражательная способность листа в оптическом и инфракрасном (ИК) диапазонах волн (400–2500 нм) зависит от различных биохимических и физических факторов, включая содержание хлорофилла и других пигментов листьев, азота, воды, а также от внутренней структуры листьев и особенностей их поверхности (Croft, Chen, 2018). Для растительных пигментов характерно поглощение электромагнитного излучения в видимом (400–700 нм) и ближнем ИК (1300–2500 нм) диапазонах длин волн. Поглощение компонентами листа в ближней инфракрасной области в диапазоне 750–1300 нм низкое, так как в этом интервале длин волн происходит интенсивное отражение от компонентов внутренней структуры листьев. Таким образом, коэффициент отражения в ближнем ИК-диапазоне зависит и от концентрации ферментов, и от структуры листа. Все это позволяет применять методы дистанционного наблюдения как в видимом, так и ближнем ИК-диапазоне длин волн для мониторинга физиологического состояния растений (Merzlyak et al., 2003; Alt et al., 2020).

Один из подходов к оценке содержания хлорофиллов по спектру отражения заключается в подборе эмпирических зависимостей (индексов) между коэффициентами отражения на определенных длинах волн, выбор которых – также важная часть метода, и содержанием хлорофиллов (Horler et al., 1983; Curran et al., 1990; Gitelson et al., 2001,

2003; Suo et al., 2010; Nikolaev et al., 2018). Успех такого «классического» подхода прямо зависит от глубины нашего понимания физики процесса.

В настоящее время в задачах предсказания характеристик биологических объектов часто применяют методы машинного обучения (Doktor et al., 2014; Feng et al., 2020). Их достоинство в том, что обычно сложную нелинейную зависимость от многих переменных можно аппроксимировать с необходимой точностью методами машинного обучения. В простых случаях на вход программы данные подаются без какой-либо обработки, тем не менее точность предсказываемого параметра будет достаточно высокой. Для каждого метода машинного обучения имеются свои способы улучшения точности предсказания, например при помощи варьирования управляющих воздействий. Существуют также способы преобразования входных данных, позволяющие улучшить результат. Так, при анализе спектров расчет производной дает возможность устранить аддитивные компоненты и выделить такие характерные особенности спектра, как положения максимумов, минимумов и точек.

Целью нашего исследования была разработка метода машинного обучения с использованием алгоритма случайного леса для предсказания суммарной концентрации хлорофиллов *a* и *b* в листьях растений по значениям спектров отражения в видимом и инфракрасном диапазонах длин волн. Проведена оценка точности предсказания в сравнении с результатами, полученными по аналитической функциональной зависимости, определены преимущества и недостатки обоих подходов.

Материалы и методы

Экспериментальные данные. Характеристики спектров отражения листьев при различных концентрациях в них хлорофиллов *a* и *b* были загружены из базы данных EcoSIS (ecosis.org), набор angers2003 (Jacquemound et al., 2003; Féret et al., 2008). Рассматривали 276 образцов листьев 39 видов растений. При этом 181 образец был получен при анализе листьев белого клена (*Acer pseudoplatanus* L.). Данные по спектру отражения представлены в диапазоне 400–2500 нм с шагом 1 нм. Для этого использован спектрорадиометр ASD FieldSpec; концентрации пигментов определены по методу Лихтенхелера и представлены в единицах измерения мкг/см^2 (см. детали в (Jacquemound et al., 2003; Féret et al., 2008)).

Математическая постановка задачи. Пусть есть генеральная совокупность $R_\lambda^{\text{ген}}$ всех возможных коэффициентов отражения листьев растений для заданных длин волн λ и $Chl^{\text{ген}}$ – значения суммы концентрации хлорофиллов *a* и *b*, соответствующие $R_\lambda^{\text{ген}}$. Мы имеем R_λ – подвыборку из $R_\lambda^{\text{ген}}$, и *Chl* – значения суммы концентрации хлорофиллов *a* и *b*, соответствующие R_λ. Требуется по набору (R_λ, *Chl*) построить функционал f: $R_\lambda^{\text{ген}} \rightarrow Chl^{\text{ген}}$. Причем, так как этот идеализированный функционал невозможно реализовать, то получится аппроксимирующий функционал: $\widetilde{f}$: $R_\lambda \rightarrow \widehat{Chl}$.

Построение модели предсказания методом случайного леса. Для построения функционала был выбран метод случайного леса (random forest, RF) (Breiman, 2001; Hastie et al., 2009). Он позволяет получить точность предсказания целевой функции, как правило, выше, чем в случае методов линейной регрессии. Идея алгоритма заключается в применении ансамбля решающих деревьев. Каждое дерево решений в этом ансамбле задает кусочно-постоянную функцию, которая получается при минимизации функции потерь (например, среднего квадрата отклонения). Алгоритм сочетает в себе две основные идеи: метод бэггинга Бреймана (Breiman, 1996) и метод случайных подпространств, предложенный T.K. Ho (1998). В его работе использована реализация метода случайного леса из библиотеки sklearn (scikit-learn.org) языка Python.

Для предсказания концентраций хлорофилла методом случайного леса были взяты несколько моделей, которые отличались наборами входных данных. Каждый набор характеризовался, во-первых, интервалом длин волн, интенсивность отражения на которых принималась во внимание. Всего было рассмотрено несколько наборов интервалов: 400–2450, 400–800 нм и комбинированный набор из двух интервалов 500–600 и 680–740 нм. Во-вторых, модели отличались типом входных данных. К ним относились значения интенсивности спектров отражения на определенных длинах волн (тип данных base), значения первых производных спектральных кривых для этих же длин волн (тип данных der), значения вторых производных (тип данных der2). Ряд моделей базировался лишь на одном типе данных, в других были совместно несколько типов данных. Такие комбинации отмечали знаком суммирования (например, base+der).

Было рассмотрено шесть моделей, они обозначены как RF-(X-Y)-Z, где (X–Y) – интервалы длин волн, Z – тип модели данных: RF-(400–2450)-base (интенсивности спектра в интервалах длин волн 400–2450 нм); RF-(400–800)-base (интенсивности спектра в интервалах длин волн 400–800 нм); RF-(400–800)-base+der (интенсивности спектра и первые производные в интервалах длин волн 400–800 нм); RF-(400–800)-der (первые производные в интервалах длин волн 400–800 нм); RF-(400–800)-der+der2 (первые и вторые производные в интервалах длин волн 400–800 нм); RF-(500–600; 680–740)-base+der+der2 (интенсивности, первые и вторые производные в интервалах длин волн 500–600 и 680–740 нм).

В качестве аппроксимации производной спектральных кривых выступала разностная производная первого порядка с единичным приращением, которую вычисляли по формуле: $D_i = R_i - R_{i-1}$. При таком расчете для первого значения нет производной. Для упрощения во всем тексте разностная производная именуется просто как производная. Вторую производную рассчитывали как производную от производной спектральной кривой.

При настройке алгоритма случайного леса выбраны следующие управляющие параметры:

- max_depth: [2, 3, 4, 5, 6] – максимальная глубина дерева;
- max_features: [2, 7, sqrt, log2, auto] – число признаков, по которым ищется разбиение (auto – все признаки);
- n_estimators: [5, 10, 15, 30, 40] – число деревьев в ансамбле случайного леса;
- random_state: 20200605.

Указанные параметры алгоритма подбирали при помощи перекрестной проверки на пяти выборках одинакового размера, полученных из предварительно пере-

мешанной случайным образом исходной тренировочной выборки. Четыре подвыборки служили для обучения модели, а пятая – для ее тестирования. Для определения наилучших управляющих параметров результаты тестирования (средний квадрат отклонения целевого показателя – *mse*) были усреднены между моделями с одинаковыми управляющими параметрами (т. е. полученными во время перекрестной проверки) и отсортированы. Управляющие параметры, для которых усредненное *mse* – минимальное, являются наилучшими. В качестве итоговой модели выбирается одна из пяти моделей с лучшими управляющими параметрами, имеющая минимальное *mse* при тестировании среди моделей, полученных по методу перекрестной проверки.

Максимальная глубина деревьев выбрана равной 6, что дает $2^6 = 64$ интервала разбиения пространства параметров, при том, что длина выборки, используемая для построения модели, равна 123. Увеличение глубины могло привести к переобучению. Количество деревьев в лесу (до 40) может показаться избыточным для 123 значений выборки, но параметры каждого из решающих деревьев подбирали на разных подпространствах (так как применяется метод случайных подпространств), а размерность признаков всегда была больше количества элементов в выборке.

Следует отметить, что алгоритм, реализованный в библиотеке sklearn, позволяет получить информативность каждого из признаков модели и отобрать из них наиболее информативные для полученных решающих правил (Breiman, 2001; Hastie et al., 2009; Louppe et al., 2013).

Построение эмпирических функциональных зависимостей. В качестве функционала $\tilde{f}: R_\lambda \to \widehat{Chl}$ мы дополнительно выбрали эмпирическую зависимость из работы (Gitelson et al., 2003) (метод GGM, названный нами по фамилиям авторов), представленную выражением

$$\widehat{Chl} = \alpha \cdot \left(\frac{1}{R_\lambda} - \frac{1}{R_{\mathrm{NIR}}}\right) \cdot R_{\mathrm{NIR}} + \beta, \qquad (1)$$

где $\widehat{Chl}$ – суммарная концентрация хлорофиллов *a* и *b*; R_λ – коэффициент отражения на длине волны λ; R_{NIR} – коэффициент отражения в ближнем инфракрасном диапазоне (например, на длине волны 800 нм); α и β подбираются таким образом, чтобы минимизировать выбранную функцию потерь. A.A. Gitelson с коллегами (2003) рекомендуют выбирать в качестве предикторов длины волн из диапазона $\lambda \in [525; 555] \cup [695; 725]$. По мнению авторов, достоинство этого алгоритма в том, что коэффициент R_{NIR} «корректирует» влияние структуры ткани растения на спектр отражения и позволяет распространить найденную функцию на растения с различающимся строением листа.

Сравнение методов предсказания концентрации хлорофилла. Выборка белого клена из набора данных angers2003 была поделена случайным образом на обучающую и валидационную в соотношении 85:15. Для примененных в настоящей работе методов предсказания алгоритмом случайного леса (RF) и функциональной зависимости (GGM) оптимальные параметры подбираются на обучающей выборке. Проверка качества алгоритмов проводится на валидационной выборке, представленной белым кленом, и на выборке образцов, не относящихся к клену. В качестве метрик для оценки точности предсказания концентраций хлорофилла были: *mse*, средняя абсолютная ошибка (*mae*) и коэффициент детерминации R^2. Формулы для расчета метрик следующие:

$$mse = \frac{1}{n}\sum_1^n (x_i - \hat{x}_i)^2,$$

$$mae = \frac{1}{n}\sum_1^n |x_i - \hat{x}_i|,$$

$$R^2 = 1 - \frac{\sum_1^n (x_i - \hat{x}_i)^2}{\sum_1^n (x_i - \bar{x}_i)^2},$$

где x – истинные значения; $\hat{x}$ – предсказанные значения; n – количество образцов; $\bar{x}$ – математическое ожидание для истинных значений. С точки зрения оптимизации, *mae* и R^2 эквивалентны. Коэффициент детерминации R^2 удобен тем, что это безразмерная величина обычно в интервале [0; 1], значение $R^2 < 0$ показывает, что среднее арифметическое $\bar{x}$ имеет лучший результат, чем предсказания построенной модели.

Результаты

Подбор параметров для метода функциональной зависимости. Для предсказания методом GGM на обучающей выборке образцов мы подбирали коэффициенты α и β уравнения (1), а также значения λ так, чтобы максимизировать значение R^2. В качестве длины волны в ближнем инфракрасном диапазоне выбрано значение $\lambda_{\mathrm{NIR}} = 800$ нм. Для получения коэффициентов α и β взяли линейную модель на основе метода наименьших квадратов (класс LinearRegression из пакета sklearn.linear_model). Для каждого $\lambda \in [400; 800]$ с шагом 1 нм был найден конкретный вид кривой GGM. Коэффициенты детерминации R^2 для предсказаний полученных моделей представлены на рис. 1. Наибольший коэффициент детерминации достигался на длине волны $\lambda = 705$ нм. Результат согласуется с рекомендованным диапазоном $\lambda \in [525; 555] \cup [695; 725]$ (Gitelson et al., 2003). Метод RF сравнивают с полученной на этой длине волны ($\lambda = 705$ нм) моделью GGM.

Результаты построения алгоритма на основе метода случайного леса. Характеристики точности предсказания концентраций хлорофилла (значения параметров *mse*, *mae*, R^2) для всех шести моделей на тестовой выборке образцов приведены в таблице. Методы RF-(400–800)-der и RF-(400–800)-der+der2 продемонстрировали высокую

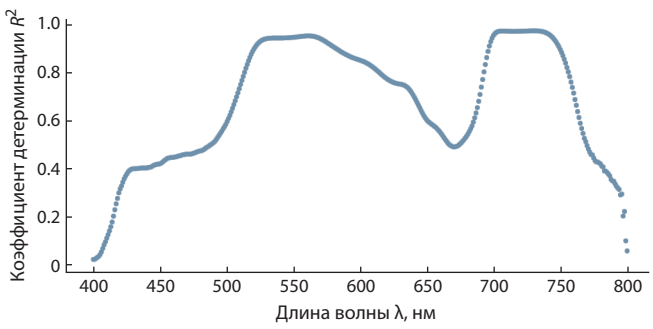

Рис. 1. Коэффициенты детерминации полученных моделей GGF при $\lambda \in [400; 800]$, которые рассчитывали на обучающей выборке.

Результаты работы модели случайного леса, обученной на различных наборах входных признаков

№ п/п	Модель случайного леса	Кол-во входных признаков	*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

Примечание. Цифрами при описании признака указан диапазон длин волн. Дополнительные характеристики признаков: base – спектр отражения; der – значения первой производной спектра; der2 – значения второй производной спектра. Курсивом выделены значения, которые имеют наихудшую точность, подчеркнутым полужирным шрифтом – наилучшую.

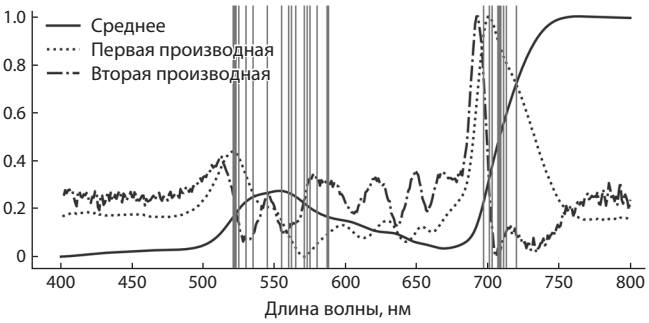

Рис. 2. Характеристики спектра отражения образцов пигментов белого клена, на которых производилось обучение моделей.

Линиями показаны: среднее значение интенсивности спектра отражения R_λ (ось *Y*) для различных длин волн (ось *X*); значение первой производной от средней интенсивности; значение второй производной. Значения производных нормированы на интервал [0; 1]. Вертикальными линиями отмечены длины волн, интенсивности спектра для которых вносят наибольший вклад в точность предсказания модели RF-(400–2450)-base.

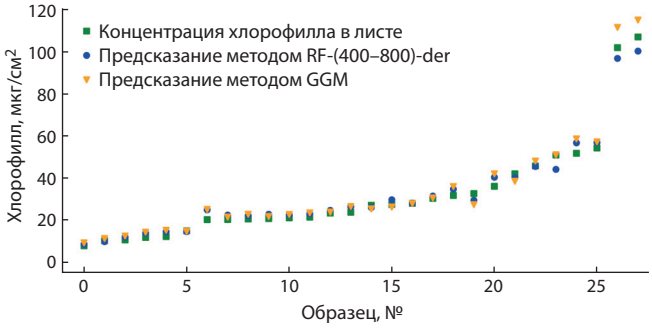

Рис. 3. Сравнение истинных и предсказанных значений концентрации хлорофилла в тканях листьев белого клена для верификационной выборки образцов.

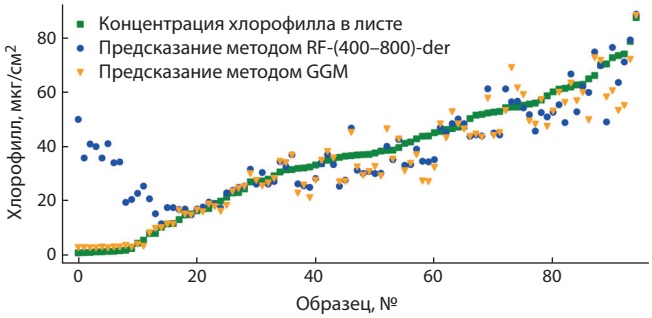

Рис. 4. Сравнение истинных и предсказанных значений концентрации хлорофилла в тканях листьев выборки образцов, не относящихся к белому клену.

точность предсказаний. В качестве наилучшего из них был отобран метод RF-(400–800)-der как имеющий меньшее количество входных параметров.

Отбор длин волн, коэффициенты отражения для которых брали в качестве входных признаков для предсказания концентраций хлорофилла методом случайного леса, осуществляли на основе первой модели (RF-(400–2450)-base). Это связано с тем, что сначала не было известно, нужен ли весь спектр, или только его часть, и какая именно. Как было указано ранее, алгоритм RF позволяет оценить информативность признаков, на которых происходило обучение. После настройки управляющих параметров модели RF-(400–2450)-base мы брали полученные параметры, чтобы заново обучить модели на пяти тренировочных выборках (из перекрестной проверки). Для этих пяти моделей мы выделили по 10 признаков с наибольшим вкладом в предсказание. Результаты показаны на рис. 2: вертикальными линиями представлен объединенный набор длин волн, интенсивности спектра для которых вносят наиболее значимый вклад в точность предсказания (26 длин волн из $10 \cdot 5 = 50$ возможных, если бы значения не пересекались). Интересно, что наиболее значимые признаки лежат в видимом диапазоне, большинство из этих признаков находится в диапазоне длин волн 500–600 и 680–740 нм. На основании этого нами были сформированы длины волн входных признаков для оставшихся пяти моделей предсказания методом случайного леса (см. выше).

Сравнение точности методов RF и GGM. Результаты сравнения методов предсказания концентраций хлорофилла методами RF-(400–800)-der и GGM и их экспериментально измеренные значения при разных значениях концентраций представлены на рис. 3 и 4. Для образцов белого клена (вида, взятого для подгонки параметров) метод RF-(400–800)-der показывает лучший результат по сравнению с методом GGM: $\sqrt{mse_{\mathrm{RF}}} = 3.01$ мкг/см^2 против $\sqrt{mse_{\mathrm{GGM}}} = 3.21$ мкг/см^2. При тестировании методов

на выборке листьев растений из других видов преимущество у метода функциональной зависимости GGM: $\sqrt{mse_{\mathrm{GGM}}}$ = 6.31 мкг/см^2 против $\sqrt{mse_{\mathrm{RF}}}$ = 12.97 мкг/см^2. Метод GGM демонстрирует высокую точность при малых концентрациях хлорофилла, в то время как метод RF на этих значениях показывает большую ошибку. Однако на интервале концентраций хлорофилла выше 20 мкг/см^2 алгоритм RF-(400–800)-der имееет лучший результат: $\sqrt{mse_{\mathrm{RF}}}$ = 5.91 мкг/см^2 против $\sqrt{mse_{\mathrm{GGM}}}$ = 7.01 мкг/см^2.

При дальнейшем анализе выяснилось, что для образцов, у которых концентрация хлорофилла меньше 7 мкг/см^2, коэффициенты отражения R_{550} (максимум спектра отражения) и R_{680} (минимум спектра отражения) визуально значительно отличны от всех остальных (рис. 5, точки в верхней правой четверти). Предсказания для этих образцов имеют значительную ошибку. Тем не менее не удалось выяснить, с чем связаны различия в спектре отражения: данные образцы не отличаются от остальных ни поверхностной плотностью листа, ни эквивалентной толщиной воды для листа (leaf equivalent water thickness) (Jacquemound et al., 2003).Шесть из десяти видов растений из этих образцов имеют также образцы с нормально предсказанными значениями. Дальнейший анализ причин аномального спектра затруднен, так как данные взяты из открытых источников, а сами измерения проводили более 17 лет назад.

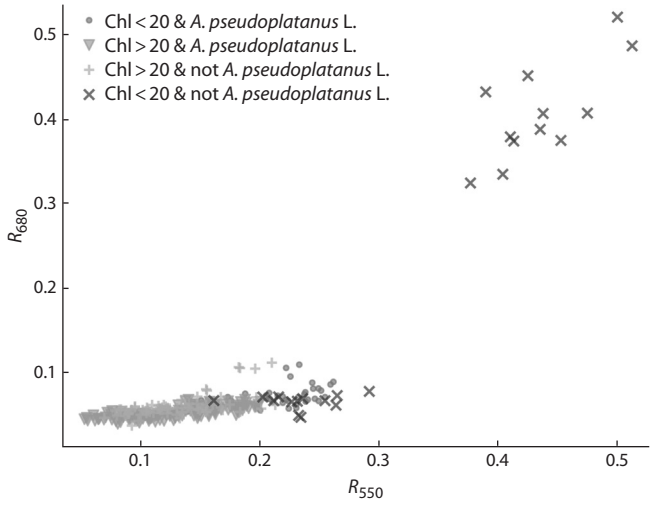

Рис. 5. Диаграмма рассеяния коэффициентов отражения R_{680} от R_{550} с выделенными категориями по концентрации хлорофилла (менее/более 20 мкг/см^2) и по виду растения (*A. pseudoplatanus* L. или др.).

Обсуждение

Во многих работах по применению спектров отражения для оценки концентраций пигментов задействуют нейронные сети (Golhani et al., 2018), в то же время в исследовательских задачах по машинному обучению также распространены методы, основанные на деревьях решений. Мы задействовали метод деревьев решений для предсказания концентраций хлорофилла в листьях растений и сравнили результаты с методом функциональной зависимости. Нами обнаружены диапазоны спектра, интенсивность отражения в которых наиболее сильно влияет на точность предсказания методом случайного леса.

Диапазон 690–750 нм в литературе называется красным краем фотосинтеза (Curran et al., 1990; Gitelson et al., 2003; Croft, Chen, 2018), а окрестность 550 нм, где находится максимум спектра отражения хлорофилла, известна как зеленый край (green edge) (Gitelson et al., 2003). Как видно из рис. 2, в нашем исследовании эти области содержат наиболее важные предикторы для метода случайного леса. Выбор в качестве входных признаков более узкого диапазона длин волн видимого спектра (400–800 нм) по сравнению с полными исходными данными (400–2450 нм) повысил качество модели. Объяснением является то, что после разделения выборки на подпространства некоторые из них оказываются менее пригодными для обучения, и обученные на этих значениях деревья вносят ошибку в суммарный результат. Наибольшего эффекта удалось добиться с применением производных спектральных зависимостей.

Метод случайного леса RF хорошо проявил себя при работе с образцами белого клена, в то время как функциональная зависимость GGM отлично показала себя при работе с разными видами растений. Это связано с бóльшей обобщающей способностью метода GGM, так как он имеет меньшее количество настраиваемых параметров. Вместе с тем более низкая точность методов RF на образцах из других видов растений частично объясняется с небольшим размером обучающей выборки и тем, что в ней представлен лишь один вид. Так, например, лучшие результаты метода случайного леса достигались при глубине деревьев, равной 5 или 6, а для этого требуется минимум 32 или 64 объекта обучающей выборки, в то время как для функционального метода (1) требуется минимум две точки (желательно, точку при малых значениях хлорофилла и точку – при больших значениях). По-видимому, эту особенность метода RF можно будет устранить с помощью бóльшего количества обучающих данных с образцами из разных видов растений.

Тем не менее процедура отбора параметров для метода RF показала, что наиболее значимые для предсказания признаки лежат в видимой области, однако влияние структуры растения в этом методе не принималось во внимание. Наряду с этим в функциональной зависимости (1) структура ткани растения учитывается членом R_{NIR}. Если эксперимент проводится с разными видами растений (см. рис. 4), то при малых значениях хлорофилла структура растения начинает играть значительную роль.

Интересно, что оба метода работают в диапазоне $\lambda \in [525; 555] \cup [695; 725]$. Они работают на спаде производной спектра отражения, что демонстрирует рис. 2.

Слово «случайный» в названии метода «случайный лес» может привести к мысли, что при смене случайного параметра, используемого алгоритмом, можно получить кардинально другие результаты. Полагаем, что при обоснованно выбранных управляющих параметрах, разумном разбиении на обучающую и проверочную выборки такая вероятность невелика. В нашем случае для каждого набора входных признаков строили по 625 моделей (перебор из множества 125 сочетаний управляющих параметров, и по 5 моделей на перекрестной проверке для каждого сочетания). К тому же из приведенной выше

таблицы следует, что методы RF-(400–800)-base+der, RF-(400–800)-der, RF-(400–800)-der+der2 имеют близкие результаты (и, что важно, имеют *mse* меньше, по сравнению с методом GGM), это косвенно подтверждает, что результаты радикально не изменятся.

Заключение

Метод случайного леса – один из алгоритмов построения функциональных зависимостей методами машинного обучения. Поэтому его можно применять для массового автоматического построения функций, связывающих наблюдаемые признаки с искомым в задачах мониторинга. Результаты настоящей работы показали, что использовать алгоритм случайного леса (и ему подобные) в задаче определения содержания хлорофилла в листе растения целесообразно, если имеется большая выборка, минимум 32 элемента, представленная широким диапазоном концентрации хлорофилла, при этом структура ткани листа меняется слабо (к примеру, применение алгоритма только на тех растениях, на которых он был обучен). В остальных случаях лучше отдать предпочтение методам, основанным на эмпирических зависимостях (как рассмотренный здесь метод GGM).

Список литературы / 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 Selektsii = 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.

Croft H., Chen J. Leaf pigment content. In: Liang S. (Ed.). Comprehensive Remote Sensing. Oxford, UK: Elsevier, 2018;117-142. DOI 10.1016/B978-0-12-409548-9.10547-0.

Curran P.J., Dungan J.L., Gholz H.L. Exploring the relationship between reflectance red edge and chlorophyll content in slash pine. *Tree Physiol*. 1990;7:33-48. DOI 10.1093/treephys/7.1-2-3-4.33.

Doktor D., Lausch A., Spengler D., Thurner M. Extraction of plant physiological status from hyperspectral signatures using machine learning methods. *Remote Sens*. 2014;6(12):12247-12274. DOI 10.3390/rs61212247.

Du H., Fuh R.-C. A., Li J., Corkan L.A., Lindsey J.S. PhotochemCAD: A computer-aided design and research tool in photochemistry. *Photochem. Photobiol*. 1998;68:141-142. DOI 10.1111/j.1751-1097.1998.tb02480.x.

Feng X., Zhan Y., Wang Q., Yang X., Yu C., Wang H., He Y. Hyperspectral imaging combined with machine learning as a tool to obtain high-throughput plant salt-stress phenotyping. *Plant J*. 2020;101(6):1448-1461. DOI 10.1111/tpj.14597.

Féret J.-B., François C., Asner G.P., Gitelson A.A., Martin R.E., Bidel L.P.R., Ustin S.L., le Maire G., Jacquemoud S. PROSPECT-4 and 5: advances in the leaf optical properties model separating photosynthetic pigments. *Remote Sens. Environ*. 2008;112:3030-3043. DOI 10.1016/j.rse.2008.02.012.

Gitelson A.A., Gritz Y., Merzlyak M.N. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol.* 2003;160(3):271-282. DOI 10.1078/0176-1617-00887.

Gitelson A.A., Merzlyak M.N., Chivkunova O.B. Optical properties and nondestructive estimation of anthocyanin content in plant leaves. *Photochem. Photobiol.* 2001;74(1):38-45. DOI 10.1562/0031-8655(2001)074<0038:OPANEO>2.0.CO;2.

Golhani K., Balasundram S.K., Vadamalai G., Pradhan B. A review of neural networks in plant disease detection using hyperspectral data. *Inf. Process. Agric.* 2018;5:354-371. DOI 10.1016/j.inpa.2018.05.002.

Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag, 2009. DOI 10.1007/978-0-387-84858-7.

Ho T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 1998;20(8):832-844. DOI 10.1109/34.709601.

Horler D.N.H., Dockray M., Barber J. The red edge of plant leaf reflectance. *Int. J. Remote Sens*. 1983;4:273-288. DOI 10.1080/01431168308948546.

Jacquemound S., Bidel L., Francois C., Pavan G. ANGERS Leaf Optical Properties Database. 2003. Data set. Available online [ecosis.org] from the Ecological Spectral Information System (EcoSIS), 2003.

Keskitalo J., Bergquist G., Gardeström P., Jansson S. A cellular timetable of autumn senescence. *Plant Physiol*. 2005;139:1635-1648. DOI 10.1104/pp.105.066845.

Lichtenthaler H.K. Chlorophylls and carotenoids: Pigments of photosynthetic biomembranes. *Methods Enzymol*. 1987;148:350-382. DOI 10.1016/0076-6879(87)48036-1.

Louppe G., Wehenkel L., Sutera A., Geurts P. Understanding variable importances in forests of randomized trees. *Adv. Neural Inf. Process. Syst.* 2013;26:431-439.

Merzlyak M.N., Gitelson A.A., Chivkunova O.B., Solovchenko A.E., Pogosyan S.I. Application of reflectance spectroscopy for analysis of higher plant pigments. *Rus. J. Plant Physiol*. 2003;50(5):704-710. DOI 10.1023/A:1025608728405.

Młodzińska E. Survey of plant pigments: molecular and environmental determinants of plant colors. *Acta Biol. Crac. Ser. Bot.* 2009;51(1):7-16.

Nikolaev S.V., Urbanovich E.A., Shayapov V.R., Orlova E.A., Afonnikov D.A. A method of evaluating the absorption spectrum of wheat leaf by the spectrum of diffuse reflection. *Sibirskii Vestnik Sel'skokhozyaistvennoi Nauki = Siberian Herald of Agricultural Science*. 2018;48(5):68-76. DOI 10.26898/0370-8799-2018-5-9. (in Russian)

Porra R.J., Thompson W.A., Kriedemann P.E. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls *a* and *b* extracted with four different solvents: Verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. *BBA – Bioenergetics.* 1989;975:384-394. DOI 10.1016/S0005-2728(89)80347-0.

Suo X.-M., Jang Y.-T., Yang M., Li S.-K., Wang K.-R., Wang C.-T. Artificial neural network to predict leaf population chlorophyll content from cotton plant images. *Agric. Sci. China*. 2010;9(1):38-45.

Wellburn A.R. The spectral determination of chlorophylls *a* and *b*, as well as total carotenoids, using various solvents with spectrophotometers of different resolution. *J. Plant Physiol.* 1994;144:307-313. DOI 10.1016/S0176-1617(11)81192-2.

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

Благодарности. Работа поддержана грантом РФФИ № 17-29-08028 и бюджетным проектом № 0259-2021-0009.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 15.10.2020. После доработки 14.12.2020. Принята к публикации 15.12.2020.