SUPPLEMENTARY MATERIALS – ДОПОЛНИТЕЛЬНЫЕ МАТЕРИАЛЫ

ASSESSMENT OF THE INFLUENCE OF ENVIRMENTAL FACTORS ON CARBON STOCK IN FOREST SOILS OF BRYANSK POLES'E

ОЦЕНКА ВЛИЯНИЯ ФАКТОРОВ СРЕДЫ НА ЗАПАСЫ ОРГАНИЧЕСКОГО УГЛЕРОДА В ЛЕСНЫХ ПОЧВАХ БРЯНСКОГО ПОЛЕСЬЯ

A.I. Kuznetsova, E.A. Gavrilyuk, A.V. Gornov, A.P. Geraskina, E.V. Ruchinskaya, A.D. Nikitina

А.И. Кузнецова, Е.А. Гаврилюк, А.В. Горнов, Е.В. Ручинская, А.П. Гераськина, А.Д. Никитина

Eurasian Soil Science

Почвоведение

Геоботанические методы. Для выявления состава основных типов леса на территории заповедника заложены геоботанические пробные площади размером 400 м2. На каждой площади выполнено геоботаническое описание. Для каждого вида определяли проективное покрытие, используя шкалу Ж. Браун-Бланке (Braun-Blanquet, 1964 цит. по: Миркин и др., 1989). На всех площадках выявлен флористический состав с учётом ярусной структуры. Латинские названия сосудистых растений даны по (Черепанов, 1995). Названия групп типов леса даны по определителю типов леса Европейской России (cepl.rssi.ru/bio/forest/index.htm).

Методы определения макрофауны. Количественный учет почвенных беспозвоночных проведен методом раскопки и ручного разбора почвенно-зоологических проб размером 25х25см, глубиной 30 см. Беспозвоночные зафиксированы в растворе этанола (дождевые черви – в 95%; многоножки, личинки насекомых и моллюски – в 70%). Биомасса макрофауны определена путем взвешивания зафиксированных беспозвоночных животных с наполненным кишечником. Идентификация дождевых червей проведена до вида (Всеволодова-Перель, 1997), представителей других групп – до семейств и родов (Гиляров, 1975; Лихарев, Раммельмейер, 2013;Локшина, 1969; Плавильщиков, 1994).

Организация SFFS и настройка SVM.

Одним из наиболее простых и универсальных методов определения значимости переменных является т.н. прямой последовательный поиск признаков (Sequential Forward Feature Selection, SFFS). Общая схема подобного поиска включает следующие этапы:

1) На первой итерации строится серия регрессионных моделей с использованием каждого признака из общего набора в качестве единственной независимой переменной и определяется их эффективность по заданному критерию.

2) Признак, на основе которого была получена самая эффективная модель, переходит на следующую итерацию, как наиболее информативный.
3) На второй итерации строится серия моделей с использованием попарных сочетаний признака, перешедшего из первой итерации, со всеми оставшимися переменными из набора.
4) Снова оценивается эффективность моделей, и лучшая пара признаков переходит на следующую итерацию.
5) Последовательный отбор и увеличение количества признаков продолжается до тех пор, пока не будет построена модель на полном наборе переменных.

В результате получается ранжированная в порядке убывания значимости последовательность признаков, а изменение показателя эффективности по итерациям характеризует вклад каждой новой добавленной переменной в модель.

Мы использовали программную реализацию SFFS и сопутствующий функционал из *R*-фреймворка *mlr3* (*Lang* et al., 2019). Поскольку при большом количестве независимых переменных SFFS требует большое количество вычислительных и временных ресурсов, мы использовали сокращенную версию поиска – расчеты прекращались, если эффективность лучшей модели на текущей итерации была ниже, чем за 10 итераций до нее.

Для построения регрессионных моделей в процессе SFFS применяли метод опорных векторов (Support Vector Machine, SVM), в частности, его программную реализацию *LIBSVM* (Chang, Lin, 2011), доступную в среде R через пакет e1071 (Meyer et al., 2023).

В отличие от классических линейных моделей, SVM позволяет получать надежные результаты при числе признаков большем, чем количество измерений, а также он не накладывает дополнительных условий на характер распределения значений в выборке и относительно устойчив к выбросам. В то же время, использование современных ансамблевых (случайные леса, градиентный бустинг) и нейросетевых методов для моделирования в нашем случае нецелесообразно из-за малого объема обучающих данных.

Стоит отметить, что для достижения наилучших результатов SVM требует настройки, как минимум, двух ключевых параметров. Мы использовали вариант SVM с оптимизацией через значение параметров ν (*ню*) и *C (cost),* и линейной функцией в качестве ядра для упрощения настройки и сохранения возможности интерпретации получаемых результатов по аналогии с классической линейной регрессией. Автоматическая оптимизация выполнялась путем простого перебора сочетаний различных значений настроечных параметров: для ν проверялись значения от 0.1 до 0.9 с шагом через 0.1, а для C – набор экспоненциальных значений вида $1Ep$, где p – целое число в диапазоне [-4; 4]. Таким

образом, суммарно проверялся 81 вариант настройки алгоритма (девять значений для ν и девять для C).

Поскольку реализация SFFS требует последовательного построения большого числа регрессионных моделей, а оптимизация SVM для каждой из них увеличивает время расчетов кратно количеству проверяемых вариантов настройки, мы разбили процесс определения значимости признаков на несколько этапов:

1) Настройка параметров SVM для модели с полным набором переменных.
2) Первый прогон SFFS с фиксированными параметрами SVM, определенными на предыдущем этапе.
3) Определение оптимального набора переменных, обеспечивших наилучшую эффективность моделирования на предыдущем этапе. При сравнении эффективности моделей на этом этапе использовалась скорректированная версия коэффициента детерминации – т.н. adjusted R^2 – которая штрафует случайный характер роста значения стандартного коэффициента детерминации по мере увеличения числа переменных. Скорректированный коэффициент детерминации R^2_{adj} рассчитывался по формуле («Wherry Formula-1» в Yin, Fan, 2001):

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

где n – количество измерений, p – количество переменных, R^2 – стандартный коэффициент детерминации.
4) Второй прогон SFFS для оптимального набора переменных, определенного на предыдущем этапе, но с индивидуальной настройкой параметров SVM для каждого сочетания признаков.

По результатам второго прогона оценивается итоговая (максимальная) эффективность моделей и на ее основе окончательно определяется оптимальное сочетание наиболее значимых признаков.

Список литературы

1 *Миркин Б.М., Розенберг Г.С., Наумова Л.Г.* Словарь понятий и терминов современной фитоценологии. М.: «Наука», 1989. 223 с. *23

2 *Черепанов С.К.* Сосудистые растения России и сопредельных государств (в пределах бывшего СССР) / С. К. Черепанов. - 2-е изд. - Санкт-Петербург: Мир и семья-95, 1995. 990 с. *30

3 *Заугольнова Л.Б., Мартыненко В.Б.* Определитель типов леса Европейской России. URL: http://www.cepl.rssi.ru/bio/forest/index.htm (дата обращения 02.12.2022). *15

4 *Всеволодова-Перель Т.С.* Дождевые черви фауны России. Кадастр и определитель. М.: Наука, 1997. 101 с. *4

5 *Гиляров М.С.* (ред). Методы почвенно-зоологических исследований. М.: Наука, 1975. 304 с. *9

6 *Лихарев И.М., Раммельмейер Е.С.* Наземные моллюски фауны СССР. М.: Рипол Классик, 2013. Т. 43. 511 с. *19

7 *Локшина И.Е.* Определитель двупарноногих многоножек (Diplopoda) равнинной части Европейской территорий СССР. М.: Наука, 1969. 78 с. *20

8 *Плавильщиков Н.Н.* Определитель насекомых: краткий определитель наиболее распространенных насекомых европейской части России. М.: Топикал, 1994. 544 с. *24

9 *Lang M., Binder M., Richter J., Schratz P., Pfisterer F., Coors S., Au Q., Casalicchio G., Kotthoff L., Bischl B.* mlr3: A modern object-oriented machine learning framework in R // Journal of Open Source Software. 2019. https://doi.org/10.21105/joss.01903 *40

10 *Chang C.-C., Lin C.-J.* LIBSVM: A library for support vector machines // ACM Transactions on Intelligent Systems and Technology. 2011. V. 2(3). P.1–27. *32

11 *Meyer D., Dimitriadou E., Hornik K., Weingessel A., Leisch F.* e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-13, 2023. https://CRAN.R-project.org/package=e1071. *43

12 Yin P., Fan X. Estimating R^2 Shrinkage in Multiple Regression: A Comparison of Different Analytical Methods // The Journal of Experimental Education. 2021. Vol. 69(2). P. 203-224. DOI: 10.1080/00220970109600656