УДК 519.67, 551.581.1

Моделирование потенциального ареала обитания растений методами машинного обучения*

А.В. Ваганов[1,2], В.Ф. Зайков[1], О.С. Кротова[3], А.И. Мусохранов[3], З.В. Покалякин[3], Л.А. Хворова[3]

[1]Южно-Сибирский ботанический сад (Барнаул, Россия)
[2]Сахалинский филиал Ботанического сада-института ДВО РАН (Южно-Сахалинск, Россия)
[3]Алтайский государственный университет (Барнаул, Россия)

Modeling a Potential Plant Habitat Using Machine Learning Methods

A.V. Vaganov[1,2], V.F. Zaikov[1], O.S. Krotova[3], A.I. Musokhranov[3], Z.V. Pokalyakin[3], L.A. Khvorova[3]

[1]South-Siberian Botanical Garden (Barnaul, Russia)
[2]Sakhalin Branch of the Botanical Garden Institute of the Far Eastern Branch of the Russian Academy of Sciences (Yuzhno-Sakhalinsk, Russia)
[3]Altai State University (Barnaul, Russia)

Статья посвящена моделированию потенциального ареала обитания вида *Pulsatilla turczaninovii* Kryl. et Serg. (Прострел Турчанинова). Моделирование экологических ниш растений — процесс построения моделей с использованием современных компьютерных алгоритмов и биоклиматических данных для прогнозирования ареала обитания видов растений. Результатом моделирования является модель, с помощью которой можно картографировать территорию произрастания или проживания видов, прогнозировать ареал или анализировать влияние окружающей среды на виды.

Для построения эффективных моделей прогнозирования экологических ниш растений требуются данные как о присутствии видов, так и об их отсутствии на той или иной территории. Точки отсутствия видов (или фоновые точки) не регистрируются в базах данных, но могут быть сгенерированы с использованием разных подходов.

В данной статье описывается реализация трех подходов к выбору точек псевдо-отсутствия видов на определенной территории и представлен результат моделирования потенциального ареала обитания вида *Pulsatilla turczaninovii* Kryl. et Serg. с помощью алгоритма случайного леса — наиболее популярного способа построения ансамблей деревьев решений. Программная реализация модели осуществлена на высокоуровневом языке программирования Python.

Ключевые слова: экологическая ниша, биологический вид, точки псевдо-отсутствия, биоклиматические характеристики, язык программирования Python, модели машинного обучения, RandomForest.

The article is devoted to modeling the potential distribution area of the species *Pulsatilla turczaninovii* Kryl. et Serg. Plant ecological niche modeling is the process of building models using modern computer algorithms and bioclimatic data to predict the distribution range of plant species. The result of the simulation is a model that can be used to map the area of growth or residence of species, predict the range or analyze the impact of the environment on species.

Data are required on both the presence of species and their absence in a particular territory to build effective models for predicting ecological niches of plants. View absence points (or background points) are not registered in databases, but can be generated using different approaches.

This article describes the implementation of three approaches to selecting pseudo-absence points of species in an operationally divided territory and presents the result of modeling the potential distribution area of the species *Pulsatilla turczaninovii* Kryl. et Serg. using the random forest algorithm — the most popular way to build ensembles of decision trees. The software implementation of the model is carried out in the high-level Python programming language.

Key words: ecological niche, biological species, pseudo-absence points, bioclimatic characteristics, Python programming language, machine learning models, RandomForest.

DOI: 10.14258/izvasu(2022)4-13

*Работа поддержана средствами программы развития ФГБОУ ВО «Алтайский государственный университет» «Приоритет-2030».

Введение

Одной из важных задач ботаники является оценка пространственного распределения объектов растительного мира [1–4]. Объективность такой оценки возможна только при комплексном подходе, объединяющем различные прикладные и фундаментальные направления ботаники, математики, информационных технологий и возможности ГИС. Современные глобальные геоинформационные технологии, методы машинного обучения и возможности прогнозного моделирования все более широко используются в биологических науках для выявления закономерностей распространения и расчета потенциальных ареалов обитания растений [5, 6]. При должном подходе данные методы можно применить для мониторинга и оценки растительных ресурсов хозяйственно-ценных растений.

Статья посвящена моделированию потенциального ареала обитания вида *Pulsatilla turczaninovii* Kryl. et Serg. (Прострел Турчанинова). Вид *Pulsatilla turczaninovii* Kryl. et Serg. является раннецветущим декоративным растением. Благодаря наличию лекарственных свойств у растения сырье используется в практике народной медицины. *P. turczaninovii* является редким растением и внесено в некоторые региональные Красные книги Российской Федерации. Исследование фондов ведущих Гербариев Евразии (LE, MW, B, VLA, NS (NSK), ALTB, PE), специальных литературных источников и личных наблюдений авторов в природе позволило детализировать современный ареал вида. *Pulsatilla turczaninovii* занимает территорию Западной и Восточной Сибири, Дальнего Востока, Китая и Монголии, не выходя за пределы Северной Азии.

Процесс моделирования потенциального ареала распространения вида растений включает в себя несколько этапов:

1) выбор точек псевдо-отсутствия вида;

2) построение и оптимизация модели;

3) визуализация и оценка результатов моделирования.

Данные о присутствии биологического вида на определенной территории представляют собой набор географических координат, называемый точками регистрации видов в пространстве. Точки регистрации видов, как правило, присутствуют на ботанических этикетках гербарных листов.

Для построения эффективных моделей прогнозирования экологических ниш требуются данные как о присутствии видов, так и об их отсутствии на той или иной территории. Точки отсутствия видов (или фоновые точки) не регистрируются в базах данных, но могут быть сгенерированы с использованием разных подходов. В статье рассмотрены три подхода к выбору точек псевдо-отсутствия вида. На каждом из сформированных наборов данных обучен алгоритм случайного леса, вследствие чего получены три модели. Точечные данные о распространении вида *Pulsatilla turczaninovii Kryl. et Serg.* на исследуемой территории, которые покрывают большой временной период, взяты из глобальной информационной системы о биоразнообразии GBIF — Global Biodiversity Information Facility [7].

1. Выбор точек псевдо-отсутствия вида

Всесторонний сравнительный анализ, приведенный в статье [8], показал, что результаты моделирования чувствительны к методу выбора точек псевдо-отсутствия вида и зависят от количества выбранных точек. Набор данных точек регистрации вида *P. turczaninovii* включает 122 точки. Для проведения исследования используется 1220 точек псевдо-отсутствия.

В ходе проведенного исследования на языке программирования Python были реализованы три подхода к выбору точек псевдо-отсутствия:

1) случайный отбор из всех точек в исследуемой области, исключая имеющиеся точки присутствия;

2) случайный выбор любой точки, расположенной по меньшей мере на один градус широты или долготы от любой точки присутствия;

3) случайный выбор точек из всех точек за пределами подходящей области, оцененной на основе биоклиматических показателей.

Первый подход является наиболее популярным, но обладает серьезным недостатком, заключающимся в том, что точки псевдо-отсутствия могут совпадать с местами, где фактически встречается исследуемый вид, и данные о ложном отсутствии могут оказывать негативное влияние на модели распространения видов (рис. 1).

Второй подход, называемый буферной методикой, состоит в том, что выбор точек отсутствия происходит за пределами определенного радиуса вокруг каждой точки присутствия. Для решения поставленной задачи использовалась буферная зона, равная 1° по широте (рис. 2).

Третий подход заключается в выделении территориальных областей, на которых значения биоклиматических показателей близки к тем значениям, какие имеют известные ареалы обитания вида. Любое место с условиями окружающей среды, подобными тем, в которых обитает вид, включается в потенциальный ареал для этого вида. Точки псевдо-отсутствия выбираются за пределами таких территорий. Схожесть территорий по биоклиматическим условиям определялась следующим образом: если значение определенного показателя для точки попадает в интерквартильный размах (Q_3–Q_1), вычисленный для показателя по всем точкам присутствия, то точка определяется как потенциальная точка присутствия вида и исключается из множества возможных точек отсутствия вида (рис. 3).

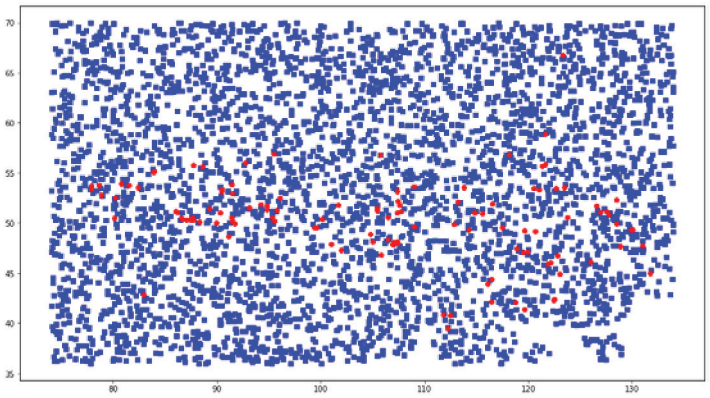

Рис. 1. Случайный отбор точек в исследуемой области, исключая имеющиеся точки присутствия

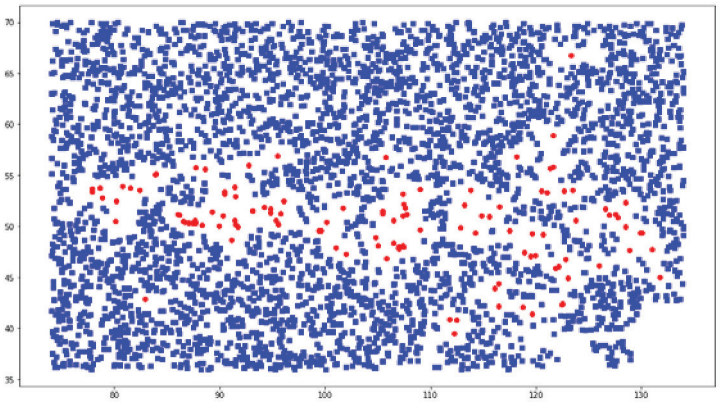

Рис. 2. Случайный отбор точек, расположенных по меньшей мере на один градус широты от любой точки присутствия

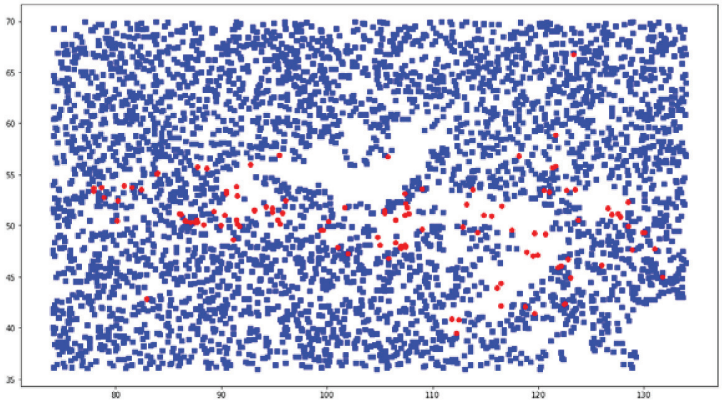

Рис. 3. Случайный отбор точек из всех точек за пределами подходящей области, оцененной на основе биоклиматических показателей

Для реализации данного подхода использовались 19 биоклиматических характеристик из базы данных WorldClim [9], отражающих годовые тренды температуры и осадков, сезонность и лимитирующие факторы. Авторами были получены три набора данных, каждый из которых включает в себя 122 точки присутствия и 1220 точек псевдо-отсутствия вида, выбранных с использованием трех описанных подходов.

2. Алгоритм метода RandomForest

Для построения моделей на полученных наборах данных авторами выбран алгоритм случайного леса.

Случайный лес (RandomForest) — один из наиболее популярных способов объединения деревьев принятия решений в ансамбли [10]. Ансамблевое обучение базируется на идее объединения множества моделей машинного обучения с целью получить более мощную модель, чем каждая из моделей по отдельности. Основная идея случайного леса заключается в том, что каждое дерево может довольно хорошо решать поставленную задачу, но с большой вероятностью оно переобучается на части данных. Если построить большое количество деревьев, которые хорошо работают и переобучаются с разной степенью, то это поможет уменьшить переобучение путем усреднения их результатов.

Алгоритм случайного леса можно описать следующими шагами [11]:

Шаг 1. Необходимо извлечь из исходного набора данных бутстрап-выборку размера n. При использовании бутстрапа из исходной выборки размером l берется случайный объект и записывается в обучающую выборку. Следующий объект также берется случайным образом из исходной выборки размером l. Так повторяется n раз, где n — желаемый размер обучающей выборки.

Шаг 2. Каждое дерево решений обучается на одной конкретной бутстрап-выборке. При этом в каждом узле дерева:

1) случайным образом отбирается s признаков бесповторным способом;

2) происходит расщепление узла с помощью признака, который обеспечивает наилучшее расщепление согласно целевой функции. Целевая функция состоит в максимизации прироста информации при каждом расщеплении:

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \to \max,$$

где f — признак, по которому выполняется расщепление, D_p — набор данных p-го родительского узла, D_j — набор данных дочернего j-го узла, I — критерий расщепления, N_p — общее число объектов в p-ом родительском узле, N_j — число объектов в дочернем j-ом узле.

Для бинарных деревьев решений обычно используются следующие критерии расщепления.

• Энтропия:

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) \log_2 p(i|t),$$

$p(i|t)$ — доля объектов, которая принадлежит классу i для отдельно взятого узла t. Энтропия равна 0, если все объекты в узле принадлежат одному и тому же классу, и энтропия максимальна, если классы распределены равномерно.

• Мера неоднородности Джини — критерий, минимизирующий вероятность ошибочной классификации:

$$I_G(t) = -\sum_{i=1}^{c} p(i|t)\,(1 - p(i|t)) = 1 - \sum_{i=1}^{c} p(i|t)^2.$$

• Ошибка классификации:

$I_E(t) = 1 - \max\{p(i|t)\}$.

Шаг 3. Шаги 1 и 2 повторяются k число раз, где k — это количество деревьев в лесу.

Шаг 4. Для назначения объекту метки класса ответы деревьев агрегируются на основе большинства голосов.

Построить модель случайного леса для классификации можно с помощью класса RandomForestClassifier() модуля sklearn.ensemble языка программирования Python. Количество деревьев задается параметром n_estimators, критерий расщепления и максимальную глубину каждого дерева можно задать с помощью параметров criterion и max_depth, параметр max_features определяет количество случайно выбранных признаков, рассматриваемых для расщепления.

3. Моделирование экологических ниш растений

Задача моделирования экологических ниш заключается в обнаружении связей между местонахождением видов в природе и факторами окружающей среды [12, 13]. Таким образом, входными переменными для таких моделей являются биоклиматические переменные, характеризующие местность, в которой произрастает вид *P. turczaninovii*. Необходимо подчеркнуть, что речь идет лишь о моделировании вероятностного распределения климатических условий, благоприятных для произрастания вида. При этом не учитываются биологические особенности, конкурентные способности вида и прочие факторы.

Важным этапом предварительной обработки данных является отбор признаков. Нередко причиной, по которой включение тех или иных признаков в модель может привести к неудовлетворительным результатам, является мультиколлинеарность — явление, при котором наблюдается сильная корреляция между признаками. В машинном обучении мультиколлинеарность приводит к переобучению модели, избыточные коэффициенты увеличивают сложность модели и время ее обучения. Целесообразно включать в модель переменные, коэффициент корреляции между которыми не превышает значения 0,7 [14–16]. Таким образом, из 19 переменных были отобраны 5 переменных: bio1 — среднегодовая температура, bio2 — суточные колебания температуры, bio7 — среднегодовая амплитуда колебания температуры, bio12 — среднегодовые осадки.

На следующем этапе исследования были построены три модели на основе алгоритма случайного леса с использованием отобранных признаков для каж-

дого из трех наборов данных. Каждая модель включает в себя 100 деревьев. В качестве критерия расщепления использовалась мера неоднородности Джини. Для оценки результатов моделирования выбраны следующие метрики оценки качества классификации: кривая ошибок (ROC) и показатель AUC, известный как площадь под ROC-кривой.

Кривая ошибок, или ROC-кривая — графический метод оценки качества работы бинарного классификатора и выбора порога для разделения классов. ROC-кривая описывает взаимосвязь между двумя величинами: чувствительностью модели и ее специфичностью. Диагональ ROC-графика можно интерпретировать как случайное угадывание метки класса. Модели классификации, которые попадают ниже диагонали, считаются хуже случайного угадывания. Количественную интерпретацию ROC дает показатель AUC — площадь под ROC-кривой. AUC принимает значения от 0 до 1. Предполагается, что чем значение AUC ближе к 1, тем лучше качество модели.

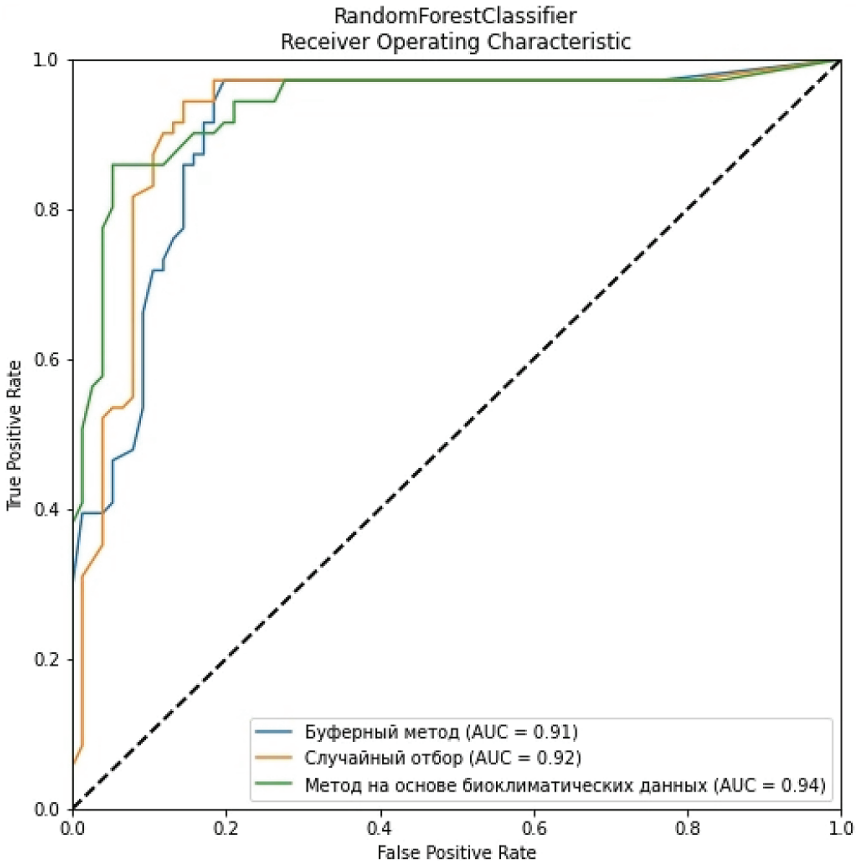

Рис. 4. ROC-кривые для построенных моделей

На рисунке 4 изображены ROC-кривые для построенных моделей: синяя кривая соответствует модели, обученной на наборе данных, в котором точки псевдо-отсутствия выбирались с помощью буферной методики (AUC=0.91), оранжевая кривая соответствует модели, обученной на наборе данных, в котором точки псевдо-отсутствия были выбраны случайным образом из всех точек в исследуемой области, исключая имеющиеся точки присутствия (AUC=0.92), зеленая кривая соответствует модели, обученной на наборе данных, в котором точки псевдо-отсутствия выбирались за пределами подходящей для обитания вида области, оцененной на основе биоклиматических показателей (AUC=0.94).

Результаты моделирования представлены на рисунках 5–7. Красный цвет отражает большую вероятность присутствия вида на территории. Чем светлее цвет, тем меньше вероятность присутствия вида.

Интерпретация полученных результатов осуществлялась специалистами в области ботаники. Каждая из построенных моделей отражает распространение вида *Pulsatilla turczaninovii* Kryl. et Serg. Наиболее точно, по мнению специалистов, отражает распространение вида карта, представленная на рисунке 5.

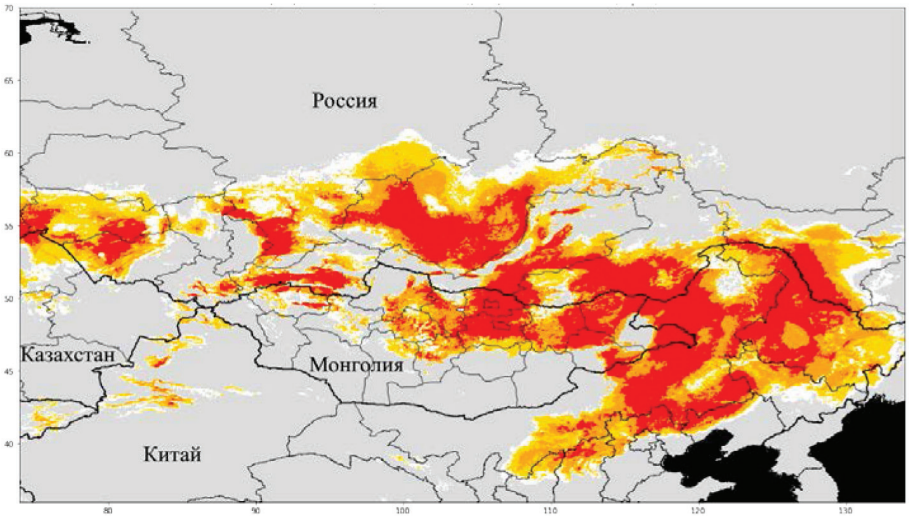

Рис. 5. Потенциальный ареал *P. turczaninovii,* полученный в результате обучения модели на наборе данных, в котором точки псевдо-отсутствия вида выбирались на основе биоклиматических показателей

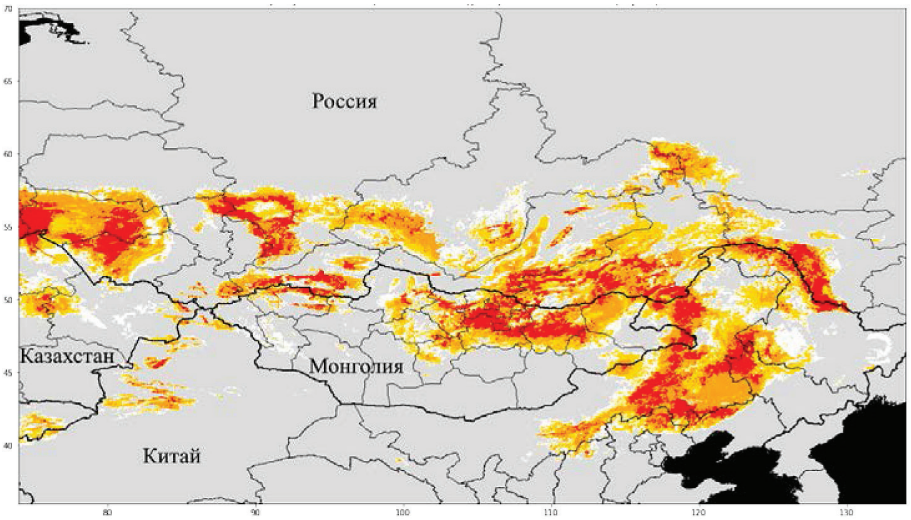

Рис. 6. Потенциальный ареал *P. turczaninovii,* полученный в результате обучения модели на наборе данных, в котором точки псевдо-отсутствия вида выбирались с помощью буферной методики

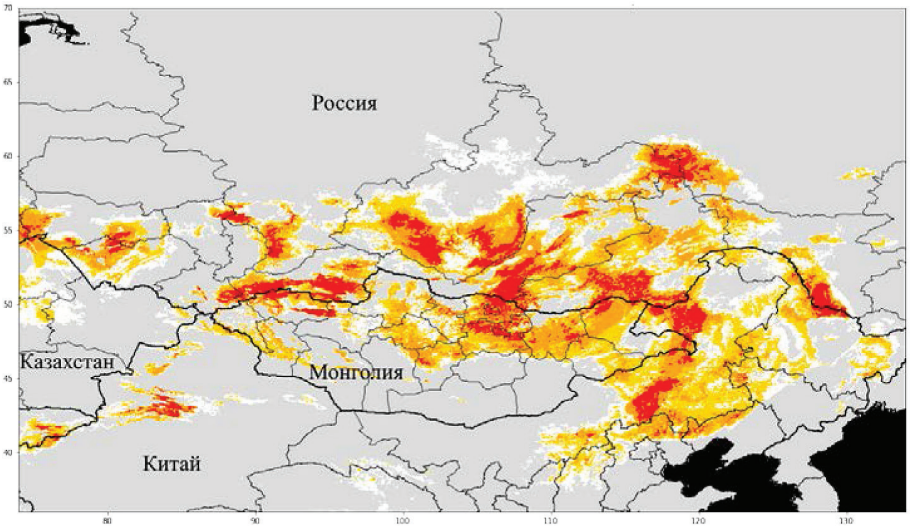

Рис. 7. Потенциальный ареал *P. turczaninovii,* полученный в результате обучения модели на наборе данных, в котором точки псевдо-отсутствия вида выбирались случайным образом

Заключение

В результате проведенного исследования был уточнен современный ареал распространении вида *Pulsatilla turczaninovii* Kryl. et Serg. и определены факторы, в наибольшей степени ограничивающие распространение вида. Полученные данные могут служить опорой для поиска новых местонахождений вида. Настоящая работа дополняет исследования по мониторингу состояния растительности на территории Большого Алтая.

Исходя из полученных результатов, можно сделать вывод, что результат моделирования напрямую зависит от метода выбора точек псевдо-отсутствия вида. Наилучшим образом подходит метод генерации на основе биоклиматических данных.

Необходимо отметить, что с ростом объема данных для моделирования при ограниченности мощностей персональных компьютеров все большее развитие получают виртуальные лаборатории. Так, для биоклиматического моделирования сегодня пользуется большой популярностью виртуальная лаборатория Biodiversity and Climate Change Virtual Laboratory, BCCVL [17]. BCCVL — англоязычная версия виртуальной лаборатории. Процесс обработки анализа данных и моделирования в BCCVL занимает достаточно много времени. Поэтому авторами разрабатывается уникальный ИТ-продукт — виртуальная лаборатория для решения задач цифровой инвентаризации биоты Алтая, биоклиматического моделирования, исследования глобального биоразнообразия регионов Большого Алтая. Разрабатываемая коллективом АлтГУ и партнерами виртуальная лаборатория опирается на передовые цифровые и интеллектуальные инструменты и включает в себя разработку алгоритмов и программ, баз данных, компьютерных методов и моделей для обработки, анализа и визуализации биологических данных для более эффективной работы с ними.

Библиографический список

1. Guisan A., Thuiller W. Predicting species distribution: offering more than simple habitat models // Ecology letters. 2005. Vol. 8. № 9.

2. Зайков В.Ф., Ваганов А.В., Шмаков А.И. Климатическое моделирование потенциального ареала Pulsatilla turczaninovii Kryl. et Serg. (Ranunculaceae) на территории Евразии // Теоретическая и прикладная экология, 2022. № 1.

3. Ваганов А.В., Покалякин З.В., Хворова Л.А. Комплексное решение задач оценки растительных ресурсов методами ГИС и климатического моделирования // Проблемы ботаники Южной Сибири и Монголии. 2021. Т. 20, № 1.

4. Макунина Н.И., Егорова А.В., Писаренко О.Ю. Построение потенциальных ареалов растительных сообществ с целью ботанико-географического районирования (на примере лесов Тувы) // Сибирский экологический журнал. 2020. № 4. DOI: 10.1134/s1995425520040095.

5. Дудов С.В. Моделирование распространения видов по данным рельефа и дистанционного зондирования на примере сосудистых растений нижнего горного пояса хр. Тукурингра (Зейский заповедник, Амурская область) // Журнал общей биологии. 2016. Т. 77. № 2.

6. Elith J., Leathwick J.R. Species distribution models: ecological explanation and prediction across space and time // Annual Rev. Ecol. Evol. Systematics. 2009. Vol. 40. DOI: 10.1146/annurev.ecolsys.110308.120159.

7. Global Biodiversity Information Facility (GBIF) Occurrence Download (accessed: 20.12.2019). DOI: 10.15468/dl.4khq61.

8. Barbet-Massin M., Jiguet F., Albert C.H., Thuiller W. Selecting pseudo-absences for species distribution models: how, where and how many? https://besjournals.onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2011.00172.x.

9. WorldClim. https://www.worldclim.com/node/1 (дата обращения: 30.04.2022).

10. Thuiller W., Lafourcade B., Engler R., Araújo M.B. BIOMOD — a platform for ensemble forecasting of species distributions // Ecography. 2009. Vol. 32. № 3.

11. Полетаева Н.Г. Классификация систем машинного обучения // Вестник Балтийского федерального ун-та. 2020. № 1.

12. Anderson R. P., Lew D., Peterson A. T. Evaluating predictive models of species' distributions: criteria for selecting models // Ecological Modelling. 2003. Vol. 162. DOI: 10.1016/s0304-3800(02)00349-6.

13. Barthlott W., Biedinger N., Braun G., Feig F., Kier G., Mutke J. Terminological and Methodological Aspects of the Mapping and Analysis of the Global Biodiversity // Acta Bot. Fennica, 1999. Vol. 162.

14. Austin M. Species distribution models and ecological theory: A critical assessment and some possible new approaches // Ecol. Model. 2007. Vol. 200. DOI: 10.1016/j.ecolmodel.2006.07.005.

16. Brown J.L. SDMtoolbox: a python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses // Methods in Ecology and Evolution. 2014. Vol. 5. № 7. DOI: 10.1111/2041-210X.12200.

16. Корзников К.А. Климатическое моделирование местообитания Kalopanax septemlobus и Phellodendron amurense var. sachalinense в островном секторе Дальнего Востока России // Известия РАН. Серия биологическая. 2019. № 6. DOI: 10.1134/S0002332919040088.

17. Hallgren W., Beaumont L., Bowness A., Chambers L., Graham E., Holewa, H., Laffan S., Mackey B., Nix H., Price J., Vanderwal J., Warren R., Weis G. The Biodiversity and Climate Change Virtual Laboratory: Where ecology meets big data // Environmental Modelling and Softwareю 2016. № 76. DOI: 1016/j.envsoft.2015.10.025.