УДК: 912.4 **DOI: 10.35595/2414-9179-2022-1-28-645-658**

В.А. Добрякова[1], Н.Н. Москвина[2], А.Б. Добряков[3], Л.Ф. Жегалина[4]

МОДЕЛИРОВАНИЕ СЕТИ ИССЛЕДОВАТЕЛЬСКИХ ПЛОЩАДОК ДЛЯ МОНИТОРИНГА ПОТОКОВ УГЛЕРОДА МЕТОДОМ RANDOM FOREST

АННОТАЦИЯ

Сети наблюдений за окружающей средой предоставляют информацию для понимания и прогнозирования пространственной и временной динамики биофизических процессов на Земле. Существует необходимость в оптимизации ресурсов для крупномасштабных мероприятий по мониторингу окружающей среды. В работе предложена, а затем протестирована пространственная структура сети исследовательских площадок для территории Тюменской области, сформированная на принципах ландшафтного подхода и с учетом минимизации издержек. Для выполнения работы было определено два тестовых набора из 40 и 105 точек. Оценка предложенного размещения выполнялась методом случайного леса (Random Forest, RF). Исследование выполнено в два этапа для каждого тестового набора. На первом проводилось обучение модели, изучались ее производительность и показатели дополнительной диагностики, на втором этапе обученная модель использовалась для прогнозирования в точки, сформированные на основе регулярной сетки, покрывающей всю территорию области (544 точки). В заключении выполнено сравнение полученных результатов с аналогичными, полученными для наборов точек того же объема, но сформированными случайным образом. В качестве прогнозируемой переменной выбран один из основных комплексных экологических показателей, связанный с выработкой углерода на данной территории – валовая первичная продуктивность далее GPP (Gross Primary Productivity). В качестве независимых переменных, характеризующих геосистемные процессы, отобран набор показателей, связанных с климатом, параметрами местности и изменчивостью почвенных ресурсов. Задача решалась с использованием инструмента «Классификация на основе леса и регрессия» (Forest-Based Classification and Regression, RF) из набора «Пространственная статистика – Моделирование пространственных отношений» программного комплекса ArcGIS Pro. В результате исследования получена высокая точность и достоверность прогноза для обоих подходов к размещению исследовательских площадок. Ландшафтный подход к выбору точек наблюдения показал свои преимущества перед случайным размещением.

КЛЮЧЕВЫЕ СЛОВА: мониторинг потоков углерода, случайный лес, валовая первичная продуктивность (GPP), ландшафтный подход

[1] Тюменский государственный университет, ул. Володарского, д.6, 625003, Тюмень, Россия, *e-mail:* **v.a.dobryakova@utmn.ru**

[2] Тюменский государственный университет, ул. Володарского, д.6, 625003, Тюмень, Россия, *e-mail:* **n.n.moskvina@utmn.ru**

[3] Уральское главное управление Центрального банка Российской Федерации, отделение по Тюменской области, ул. Володарского, д. 48, 625000, Тюмень, Россия, *e-mail:* **dobryakov_andrey@mail.ru**

[4] Балтийский федеральный университет им. И. Канта, ул. А. Невского, д. 14, 236016, Калининград, Россия, *e-mail:* **LZhegalina@kantiana.ru**

Valentina A. Dobryakova[1], **Natalya N. Moskvina**[2], **Andrey B. Dobryakov**[3], **Lilia F. Zhegalina**[4]

MODELING NETWORK OF RESEARCH SITES FOR MONITORING CARBON FLOWS BY RANDOM FOREST METHOD

ABSTRACT

Environmental observing networks provide information for understanding and predicting the spatial and temporal dynamics of Earth biophysical processes. The optimization of resources for large-scale environmental monitoring activities is required. The paper describes and then tests spatial structure of Tyumen region research sites network. The network is based on principles of landscape approach, taking into account cost minimization. At the baseline of research, two testing sets of 40 and 105 points were determined. Proposed locations were evaluated using Random Forest (RF) method. The study accomplished in two stages for each test set. At the first stage, the model was trained; its capacity and indicators of additional diagnostics were studied. At the second stage, the trained model was used to predict the points formed of regular grid covering entire territory of this region (544 points). In conclusion, the obtained results were compared with similar point sets of the same volume but generated randomly. Primary Productivity Gross (GPP) was chosen as predictable variable because it is one of the major complex environmental indicators associated with carbon production in this area. The ability of an area to absorb or produce carbon is one of the main parameters that determine climate processes. As independent variables characterizing geosystemic processes, a set of indicators associated with climate, terrain parameters, and variability of soil resources has been selected. The problem was solved using Forest-Based Classification and Regression tool from Spatial Statistics – Modeling Spatial Relationships toolkit of ArcGIS Pro software package. As the result of the study, a high forecast accuracy and reliability for both approaches to research sites locations was obtained. The study was based on open source data.

KEYWORDS: carbon flux monitoring, random forest, gross primary productivity (GPP), landscape approach

ВВЕДЕНИЕ

Глобальные климатические изменения оказывают негативное воздействие на человеческий потенциал, экономику и геосистемы всех стран мира, включая Россию. Исследование окружающей среды имеет решающее значение для принятия обоснованных экологических решений. Важнейшую роль в этой деятельности играет мониторинг происходящих процессов. Наблюдение за основными показателями окружающей среды (например, почвенным покровом, химическим составом атмосферы, температурой воздуха и др.) дает необходимые знания для разработки правильных действий, направленных на смягчение отрицательных климатических последствий. Именно для этой цели создается всемирная сеть экологических обсерваторий (EON) [*Villarreal* et al., 2019]. Поскольку создание таких обсерваторий – сложная, трудоемкая и дорогостоящая задача, возникает необходимость выработки оптимального подхода к размещению точек наблюдения [*Villarreal* et al., 2021].

[1] Tyumen State University, Volodarskiy str., 6, 625003, Tyumen, Russia, *e-mail:* **v.a.dobryakova@utmn.ru**
[2] Tyumen State University, Volodarskiy str., 6, 625003, Tyumen, Russia, *e-mail:* **n.n.moskvina@utmn.ru**
[3] The Ural head department of the Central bank of the Russian Federation, Volodarskiy str., 48, 625000, Tyumen, Russia, *e-mail:* **dobryakov_andrey@mail.ru**
[4] Immanuel Kant Baltic Federal University, Proletarskaya Str. 131, 236029, Kaliningrad, Russia, *e-mail:* **lzhegalina@kantiana.ru**

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Цель данного исследования – используя ландшафтный подход, наметить и оценить методом случайного леса (далее RF (Random Forest)) пространственную структуру исследовательских площадок для мониторинга окружающей среды.

Для изучения наземных запасов углерода и глобальных углеродных циклов большое значение имеет количественное картирование лесной надземной биомассы по данным дистанционного зондирования. При выборе алгоритма машинного обучения исследователи отмечают, что модель RF может служить эталоном для оценки лесной биомассы с использованием спутниковых изображений из нескольких источников данных. [*Han* et al., 2022].

Инструмент «Классификация на основе леса и регрессия» создает модели и строит прогнозы при помощи адаптации метода контролируемого машинного обучения случайного леса Лео Бреймана [*Breiman,* 2001]. На этот тип модели не влияет мультиколлинеарность, потому что это не линейная модель, и он может моделировать отношения между огромным количеством переменных-показателей и целевой переменной. Кроме того, метод не требует предварительного анализа значимости исходных показателей, избыточное число предикторов влияет только на продолжительность вычислительных процедур. Значимость переменных определяется в процессе обучения.

Алгоритм Random Forest имеет большие потенциальные преимущества: он непараметричен, нечувствителен к искажению данных, устойчив к большому количеству переменных входных данных. Путем разделения узла алгоритм ищет лучший атрибут среди случайных наборов, а не поиск основных характеристик, что приводит к высокой диверсификации, лучшему моделированию и меньшему количеству ошибок за счет подчеркивания снижения дисперсии [*Clewley* et al., 2017].

Изменение климата, ландшафтные и экологические исследования, глобальное картирование углерода являются приложениями данной ансамблевой модели. В многочисленных работах отмечено, что RF показывает наилучшую производительность при прогнозировании и моделировании [*Cutler* et al., 2007; *Rodriguez-Galiano* et al., 2011; *Baccini* et al., 2012; *Gomes* et al., 2019; *Han* et al., 2022; *Wang* et al., 2022]. К существенному улучшению моделирования запасов углерода приводит учет пространственного контекста [*Mascaro* et al., 2014].

Основные задачи проектирования и оценки размещения площадок для мониторинга:

1. Формирование тестовых наборов 40 (набор 1) и 105 точек (набор 2) на принципах геосистемного подхода и с учетом минимизации издержек.
2. Обучение модели на двух тестовых наборах.
3. Прогнозирование в точки, сформированные на основе регулярной сетки, покрывающей всю территорию области.
4. Сравнение результатов моделирования с аналогичными результатами для наборов точек того же объема, но размещенных по территории и сформированных случайным образом.

Основным инструментом исследования выбран инструмент «Классификация на основе леса и регрессия» (Forest-Based Classification and Regression, RF) из набора «Пространственная статистика – Моделирование пространственных отношений» программного комплекса ArcGIS Pro.

В качестве прогнозируемой переменной нами выбран один из основных комплексных экологических показателей, связанный с выработкой углерода на данной территории – валовая первичная продуктивность (GPP). Чистый экосистемный обмен представляет собой баланс между двумя компонентами: дыханием экосистемы и валовой первичной продукцией [*Aubinet* et al., 2012; *Kirschbaum* et al., 2001]. Способность территории поглощать или производить углерод является одним из главных параметров, определяющих климатические процессы [*Villarreal* et al., 2019]. Потенциал для обеспечения глобальных

измерений, связанных с GPP, показали исследования сезонного цикла фотосинтеза на основе данных газообмена с вышек ковариации со всего мира. Модель MODIS GPP основана на ежедневных метеорологических наблюдениях с башен вихревой ковариации, расположенных в центре каждого участка. Отмечено, что наземное масштабирование GPP может улучшить параметризацию эффективности использования света в алгоритмах спутникового мониторинга GPP [*Turner* et al., 2003].

При прогнозировании инструментом Random Forests с использованием рекурсивного исключения признаков, GPP был выбран одним из важных ковариат [*Gomes* et al., 2019].

Независимыми переменными, характеризующими геосистемные процессы, выбраны показатели, связанные с климатом, параметрами местности и изменчивостью почвенных ресурсов, что, собственно, является показателями функционирования геосистем.

Для понимания преимуществ или недостатков модели, рассчитанной на основе геосистемного подхода, выполнено сравнение с аналогичными результатами, полученными для наборов точек того же объема, но сформированных случайным образом.

Случайное распределение с точки зрения статистических процедур является очень правильным и при достаточно большом количестве точек дает хорошие результаты.

Для проведения исследования мы использовали следующие общедоступные данные на территорию Тюменской области:

– 19 биоклиматических предикторов для характеристики климатических условий: среднегодовые условия (среднегодовая температура, годовое количество осадков), среднегодовые сезонные условия (сезонность температуры) и внутригодовые сезонные условия (средняя температура самого засушливого квартала или осадки самого влажного квартала) температуры и осадков [*Fick* et al., 2017].

– Солнечное излучение (кДж м $^{-2}$ сут $^{-1}$), среднее по месяцам (04–10) за 1970–2000 гг. Эти показатели являются средними за 1970–2000 гг. Каждая загрузка представляет собой ZIP-файл, содержащий 19 файлов GeoTiff (.tif), по одному для каждого месяца переменных. Данные скачаны с пространственными разрешениями от 30 секунд (~ 1 км^2) с сайта WorldClim[1]. [*Fick* et al., 2017]

– Высота, уклон построены на основе цифрой модели рельефа (SRTM) с пространственным разрешением 1››, загруженной с сервиса USGS[2].

– Топографический индекс влажности рассчитан на основе данных о рельефе.

– Глобальная карта почвенного органического углерода[3].

– Общая плотность азота в почве[4].

– GPP – валовая первичная продуктивность, средние показатели по месяцам (04–10) за период 2006–2019 гг. GPP был получен с помощью приборов визуального спектрорадиометра (MODIS) с разрешением MODerate на спутниках NASA Terra и Aqua с использованием двунаправленной функции распределения отражения Nadir (BRDF), отрегулированной на отражения (NBAR) продукта в качестве входа в нейронные сети, которые использовались для глобального расширения GPP, оцененного по выбранным FLUXNET 2015 вихревых колоннам [*Joiner* et al., 2014].

– Глобальная карта землепользования/почвенного покрова (LULC), полученная из изображений ESA Sentinel-2 с разрешением 10 м. Создается каждый год на основе модели классификации земель с глубоким обучением [*Karra* et al., 2021].

Географическое положение потенциальных пунктов наблюдений определялось исходя из следующих принципов:

1. Измерения должны осуществляться в пределах существующих (запланированных) карбоновых полигонов.

[1] https://worldclim.org
[2] https://earthexplorer.usgs.gov/
[3] http://54.229.242.119/GSOCmap
[4] https://webmap.ornl.gov/ogc/dataset.jsp?ds_id=569

2. Должны включать территории научных полигонов и стационаров, площадки в пределах систем особо охраняемых территорий (ООПТ) и площадки сети FLUXNET в пределах Тюменской области.

3. При прочих равных обстоятельствах предпочтение отдается точкам, находящимся вблизи населенных пунктов и сложившейся дорожной сети.

4. Должны быть распределены по разным геосистемам. В качестве базовой карты для выбора использовалась карта земного покрова [*Friedl* et al., 2015].

5. Количество пунктов наблюдений должно быть достаточным для покрытия основных разностей природных комплексов (не менее 3–5 для каждой зональной / подзональной геосистемы).

Пункты 1–3 учитывают большую трудоемкость работ и высокую стоимость устанавливаемого оборудования, использование этих подходов должно дать снижение суммарных издержек. Используемый метод прогноза не позволил нам выбрать количество точек меньше 40, так как при уменьшении числа площадок инструмент «Классификация на основе леса и регрессия» начинает работать нестабильно, часто не может завершить операцию и даже при удачном завершении не формирует итоговые данные. В результате для тестового набора № 1 было решено определить именно 40 площадок для мониторинга.

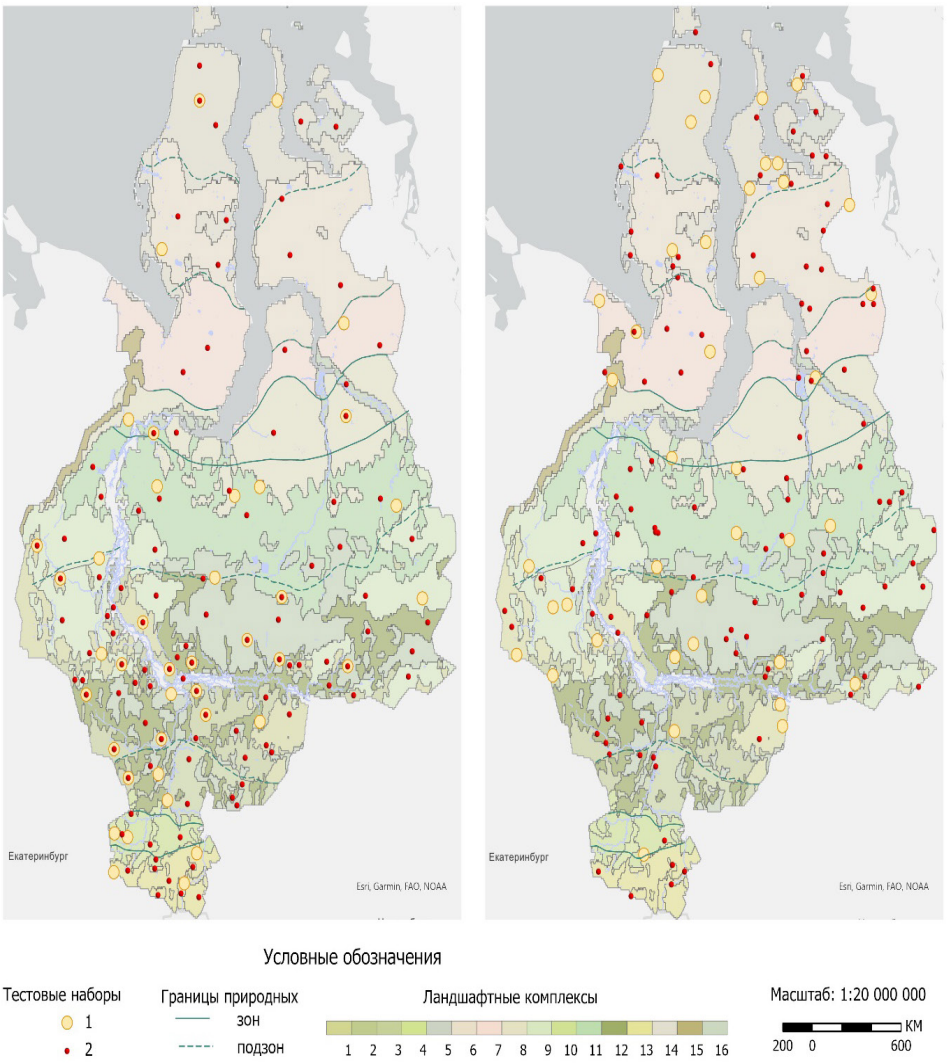

Рис. 1. Сеть потенциальных точек наблюдения.
Слева – выбор на основе ландшафтного подхода, справа – случайным образом
Fig 1. Network of potential observation sites.
Left – selection based on landscape approach, right – random

На территории Тюменской области по данным карты земного покрова [*Friedl* et al., 2015] выделено 14 типов ландшафтов (landcover). Так как карта земного покрова является мелкомасштабной, составленной на всю поверхность материков Земли, региональное уточнение проводилось по ландшафтной карте Тюменской области [Атлас Тюменской области, 1971] с опорой на данные ландшафтного районирования [*Козин*, 1996]. В результате в границах области определено 16 родов ландшафтных комплексов, отвечающих зональным и азональным характеристикам. Факторы функционирования в данных комплексах имеют существенные различия. Для регионального покрытия пунктами наблюдательной сети, с учетом пп. 4 и 5, намечено 105 мобильных точек, дополняющих данные 40 площадок для мониторинга. Каждая мобильная «точка» может представлять собой набор измерений по группам фаций или геохимическим трансектам. Их точное местоположение можно определять в соответствии с программами наблюдений региональных полигонов и стационаров.

Условные обозначения ландшафтных комплексов Тюменской области

1. Ледниковые и водно-ледниковые предгорные возвышенности с доминированием среднетаежных елово-кедровых лесов и их производных на подзолисто-элювиально-глеевых почвах (water).

2. Водно-ледниковые возвышенности с среднетаежными темнохвойно-березовыми с лиственницей кустарничково-зеленомошными лесами на подзолистых и дерново-подзолистых почвах (deciduous broad-leaved forest).

3. Водно-ледниковые возвышенности с южнотаежными сосновыми кустарничково-зеленомошными и лишайниковыми, долгомошно-сфагновыми лесами на подзолистых почвах (deciduous broad-leaved forest).

4. Озерно-аллювиальные равнины с доминированием подтаежных елово-кедровых лесов и их производных, в комплексе с сосновыми вейниковыми и травяно-кустарничковыми лесами на подзолисто-элювиально-глеевых почвах (deciduous broad-leaved forest).

5. Морские равнины с арктотундровыми лишайниково-моховыми с ивой и ерником сообществами на иллювиально-гумусовых почвах (closed shrub).

6. Морские и ледниково-морские возвышенные равнины с типично-тундровыми кустарниково-моховыми и лишайниково-моховыми с ивой и ерником сообществами на тундровых глеевых почвах (closed shrub).

7. Морские и ледниково-морские возвышенные равнины с доминированием южно-тундровых ерниковых моховых и лишайниковых сообществ, с участием угнетенной лиственницы на тундровых иллювиально-гумусовых почвах (closed shrub).

8. Ледниково-морские и озерно-аллювиальные плоские равнины с лиственничными, местами елово-лиственничными кустарничково-мохово-лишайниковыми редколесьями на слабоподзолистых почвах, в сочетании с ерниково-ивняковыми кустарничково-лишайниковыми и кустарничково-моховыми тундрами на тундровых слабооподзоленных почвах (closed shrub).

9. Ледниковые и водно-ледниковые возвышенности с сочетанием северо- и среднетаежных сосновых и их производных кустарничково-сфагновых и елово-кедровых с лиственницей кустарничково-зеленомошных лесов с доминированием иллювиально-железистых подзолов (open shrub).

10. Озерно-аллювиальные и аллювиальные равнины с сосновыми вейниковыми, травяно-кустарничковыми, березовыми и осиновыми злаково-разнотравными лесами на дерново-подзолистых и серых лесных почвах (open shrub).

11. Ледниково-морские плакорные равнины и озерно-аллювиальные речные долинные с сочетанием лиственничных, елово-лиственничных и сосново-лиственничных лишайниково-мохово-кустарничковых, зеленомошно-кустарничковых редкостойных лесов на подзолистых почвах (woody savannah).

12. Озерно-аллювиальные низины с комплексом сосновых и сосново-березовых долгомошно-сфагновых и кустарничково-сфагновых лесов и сосново-кустарничко-

во-сфагновых плоскобугристых и грядово-мочажинных болот на подзолистых иллювиально-железистых и торфяно-болотных почвах (woody savannah).

13. Лесостепные субаэральные распаханные равнины с агрофитоценозами на агроземах на месте разнотравно-злаковых остепненных лугов, в сочетании с осиново-березовыми и березово-осиновыми остепненными злаково-разнотравными лесными колками на серых лесных почвах (savannah).

14. Морские и ледниково-морские приокеанические низины с лишайниково-моховыми и кустарничково-моховыми, в сочетании с осоково-пушицево-моховыми заболоченными арктотундровыми группировками на тундрово-глеевых и болотно-тундровых почвах (savannah).

15. Ледниковые и водно-ледниковые возвышенные предгорье и низкогорье с лишайниково-моховыми и кустарничково-моховыми горными тундрами на тундрово-глеевых почвах

16. Преимущественно озерно-аллювиальные заболоченные низины с сочетанием сосново-сфагново-кустарничковых рямов и лишайниково-сфагновых плосковыпуклых болотно-озерных комплексов на болотных торфяных почвах (pastures).

Схема рабочего процесса для каждого тестового набора:

1. Обучение модели для прогнозирования (инструмент «случайный лес» запускается в режиме обучения), результатом процедуры обучения является формирование прогноза значений GPP.

2. Прогнозирование в контрольные точки, сформированные на основе регулярной сетки, покрывающей всю территорию Тюменской области. Для формирования этого набора выбраны центры правильных шестиугольников площадью 10 000 км2, образующих сплошное покрытие территории всей области (всего 544 точки).

3. Оценка расхождений между реальными значениями GPP и сформированными RF в каждой точке контрольного набора. По всей совокупности расхождений формировалась достоверность полученного прогноза и ряд дополнительных параметров, характеризующих выполненную процедуру.

Повторение действий, описанных в пунктах 1–3 для наборов точек того же объема (40 и 105), но сформированных случайным образом.

Сравнение результатов моделирования (достоверность прогноза). Использование двух наборов, по нашему мнению, должно дать объективную картину преимуществ или недостатков ландшафтного подхода.

Вычисления выполнены в программном комплексе ArcGIS Pro.

Исследование основано на общедоступной информации и программном обеспечении, поэтому эту схему с незначительными корректировками можно применять для любой территории.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Инструмент «Классификация на основе леса и регрессия» в процессе обучения и при последующем прогнозировании запускался со следующими параметрами: количество решающих деревьев – 500, количество запусков – 100, остальные параметры приняты по умолчанию. При обучении не менее 10 % данных исключаются из обучения и используются для предварительной оценки полученной модели. Мы пытались использовать максимальное число данных для обучения поэтому исключали именно 10 %.

В результате получены оценки: значимости предикторов и производительности моделей.

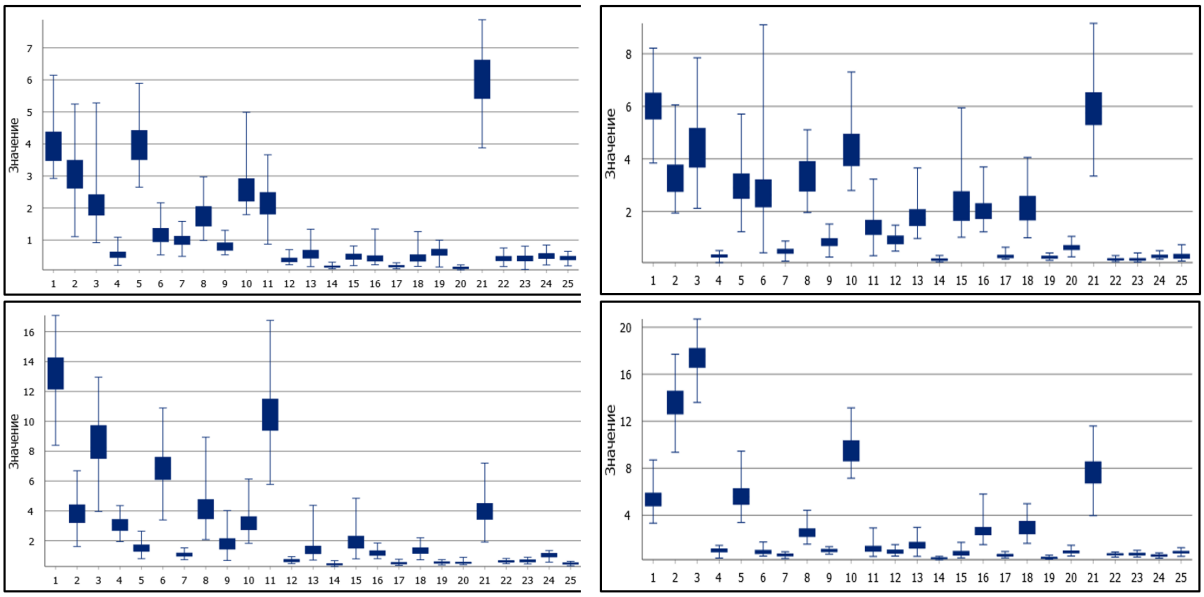

Рис. 2. Распределение значимости переменных, по результатам обучения набора 1(верхняя строка) и набора 2 (нижняя строка).

Левый столбик – выбор на основе ландшафтного подхода, правый – случайным образом. Исходные показатели (предикторы): 1 – Среднегодовая температура. 2 – Средний дневной диапазон. 3 – Изотермичность. 4 – Сезонность температуры. 5 – Максимальная температура самого теплого месяца. 6 – Минимальная температура самого холодного месяца. 7 – Годовой диапазон температуры. 8 – Средняя температура самого влажного квартала. 9 – Средняя температура самого засушливого квартала. 10 – Средняя температура самого теплого квартала. 11 – Средняя температура самого холодного квартала. 12 – Годовое количество осадков. 13 – Осадки самого влажного месяца. 14 – Осадки в самый засушливый месяц. 15 – Сезонность осадков. 16 – Осадки самого влажного квартала. 17 – Осадки самого засушливого квартала. 18 – Осадки самого теплого квартала. 19 – Осадки в самой холодной четверти. 20 – Гипсометрия. 21 – Солнечное излучение. 22 – Топографический индекс влажности. 23 – Уклон. 24 – Плотность азота в почве. 25 – Углерод в почве.

Fig 2. Variables significance distribution, according to the results of training set 1 (top line) and set 2 (bottom line).

The left column is a selection based on the landscape approach; the right is randomly. Initial indicators (predictors): 1 – Average annual temperature. 2 – Average daily range. 3 – Isothermal. 4 – Temperature seasonality. 5 – The maximum temperature of the warmest month. 6 – Minimum temperature of the coldest month. 7 – Annual temperature range. 8 – The average temperature of the wettest quarter. 9 – The average temperature of the driest quarter. 10 – The average temperature of the warmest quarter. 11 – The average temperature of the coldest quarter. 12 – Annual rainfall. 13 – Precipitation of the wettest month. 14 – Precipitation in the driest month. 15 – Seasonality of precipitation. 16 – Precipitation of the wettest quarter. 17 – Precipitation of the driest quarter. 18 – Precipitation of the warmest quarter. 19 – Precipitation in the coldest quarter. 20 – Hypsometry. 21 – Solar radiation. 22 – Topographic humidity index. 23 – Slope. 24 – Density of nitrogen in the soil. 25 – Carbon in the soil.

Наиболее значимыми переменными являются: 1 – Среднегодовая температура, 3 – Изотермичность, 21 – Солнечное излучение.

Значимость соответствует тому, сколько раз выполняется разбиение на основе переменной во всей модели леса. При разных прогонах модели (у нас – 100) значимость предикторов меняется, это отражает размах «усов» на ящичковой диаграмме (рис. 2). Сами «ящики» показывают диапазон, в который попадает 50 % деревьев. Длинные «усы» и «ящики» могут указывать на нестабильность модели.

В моделях на основе ландшафтного подхода список из 6 наиболее значимых показателей не меняется при переходе к большому набору точек, хотя изменения в их зна-

чимости происходят. В моделях на основе случайного выбора в тестовом наборе 1 видим большое число значимых переменных, в тестовом наборе 2 осталось только 6 значимых предикторов: 1, 2, 3, 5, 10, 21.

В целом точность полученной модели оценивается показателем R^2 (коэффициент детерминации), это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть построенным прогнозом.

Табл. 1. Результаты моделирования, значение R^2
Table 1. Simulation results, R^2 value

Режим	**Модель 1: выбор на основе ландшафтного подхода, количество точек**		**Модель 2: выбор случайным образом, количество точек**	
	105	**40**	**105**	**40**
Обучение	0,84	0,927	0,83	0,92
Прогноз	0,832	0,915	0,827	0,908

Таблица 1 показывает основные результаты моделирования в режиме обучения (верхняя строка) и прогнозирования (нижняя строка). Для обеих моделей мы видим высокую точность и достоверность. В целом отличия R^2 незначительны, тем не менее следует отметить, что моделирование на основе ландшафтного подхода на всех этапах дает все же более высокие показатели. Неожиданными являются результаты прогноза: достоверность на меньшем количестве точек заметно выше чем на большем. Это является настораживающим признаком и, как нам кажется, указывает на недостатки при обучении модели.

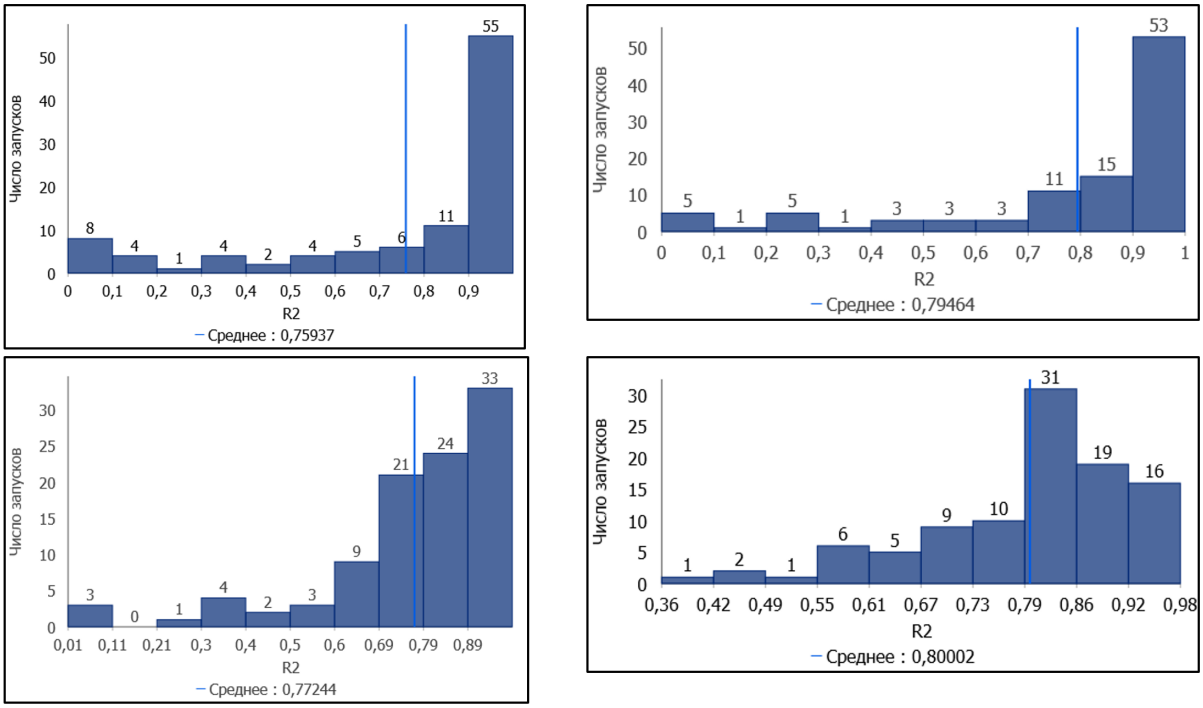

Рис. *3. Распределение R^2* по результатам прогнозирования, набора *1 (верхняя строка) и набора 2 (нижняя строка). Левый столбик – выбор на основе ландшафтного подхода, правый – случайным образом.*
Fig 3. Distribution of R^2 by prediction results, set 1 (top row) and set 2 (bottom row). The left column is a selection based on a landscape approach, the right column is random.

На рис. 3 представлена гистограмма изменения R^2 по данным 100 прогонов по результатам прогнозирования в 544 точки контрольного набора. Прогнозирование выполнялось на основании проведенного ранее обучения случайной и ландшафтной моделей на наборах 1 и 2.

Результаты распределений в обеих прогнозных моделях на наборе 1 весьма похожи: выраженные максимумы для R^2 от 0,9 до 1 (частоты 55 и 53, соответственно), остальные интервалы гистограммы дают незначительный вклад. Недостаток: сохраняются выбросы в области малых значений.

Результаты по набору 2 заметно отличаются от предыдущих. При ландшафтном подходе максимум вновь приходится на последний интервал от 0,9 до 1 (частота – 33), вклад других интервалов стал более заметным. При случайном подходе максимум приходится на интервал от 0,79 до 0,86 (частота – 31), к плюсам можно отнести то, что диапазон наблюдаемых значений R^2 теперь начинается с 0,36, то есть отсутствуют малые значения, которые сохранились в первом случае.

Наличие устойчивых выбросов малых значений R^2 у модели 1 (выбор на основе геосистемного подхода) указывает, возможно, на необходимость его уточнения. Размытость максимума при увеличении числа точек в обоих моделях является, скорее всего, следствием погрешности в определении показателей и, прежде всего, GPP, которая связана с отсутствием точных данных от наземных станций на территории России [*Pastorello* et al., 2020].

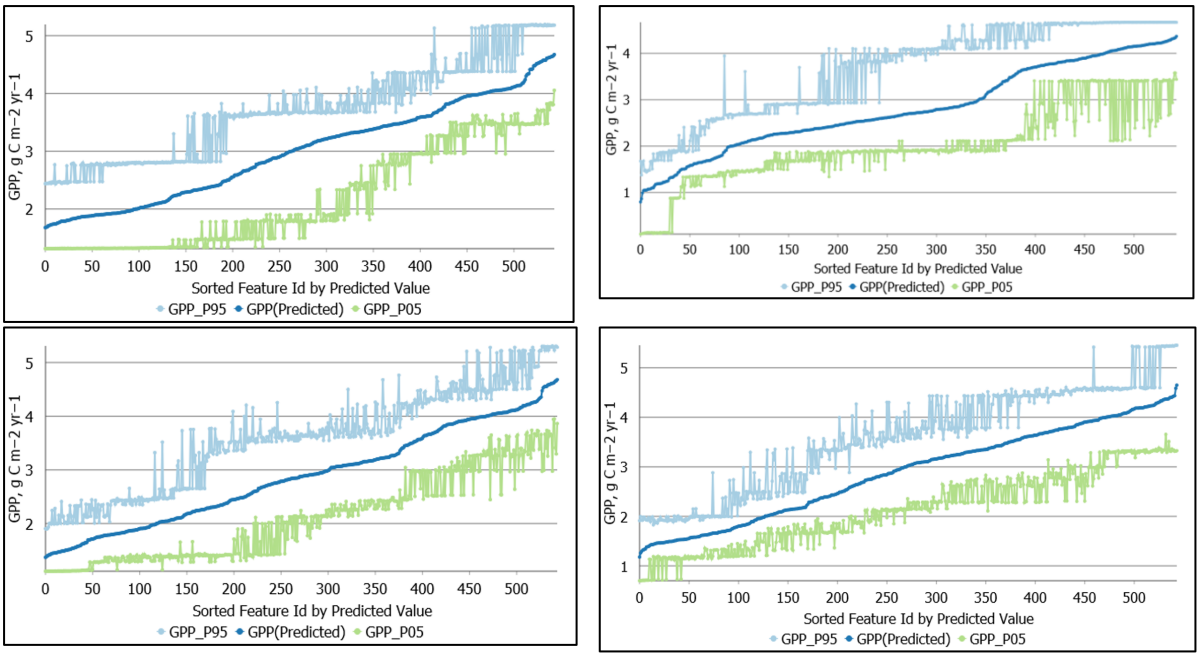

Рис. 4. Интервал прогнозирования целевой переменной (GPP) для 544 точек, по результатам обучения набора 1 (верхняя строка) и набора 2 (нижняя строка). Левый столбик – выбор на основе ландшафтного подхода, правый – случайным образом.

Fig 4. Target variable prediction interval (GPP) for 544 sites, based on the training results of set 1 (top line) and set 2 (bottom line). The left column is a selection based on the landscape approach, the right one is random.

На диаграмме показаны границы неопределенности прогноза, синяя линия является фактическим прогнозом. Значения прогноза GPP на точках контрольного набора изменяются от 1 до 4,5, кроме того, наблюдается высокий диапазон неопределенности прогноза. Значительная неопределенность прогноза является одним из признаков неустойчивости модели.

ВЫВОДЫ

1. Метод случайного леса действительно показал свои сильные стороны и прекрасно строит прогноз, если имеется достаточно данных для построения леса, содержащего большое число решающих деревьев. В нашем случае потребовалось не менее 40 точек, причем каждая точка описывалась 25 первичными показателями. Считается, что прогноз тем лучше и устойчивей, чем больше деревьев, мы при всех расчетах выбирали 500. Этого числа деревьев в данном случае оказалось достаточно для получения хорошего прогноза.

2. Ландшафтный подход к выбору точек наблюдения все-таки показал свои преимущества перед случайным размещением. Причем эти преимущества, хотя и незначительные, мы видели на каждом этапе расчетов.

3. Показатели точности прогноза оказались очень высокими, настолько высокими, что в будущем следует обязательно рассмотреть причину этого явления. Сравнивая наши результаты с результатами подобного прогноза, выполненного рядом авторов [Villarreal et al., 2021], мы видим у них значительно худшую достоверность. Можно предположить, что причиной такого результата для территории Тюменской области является отсутствие такой контрастности в данных и такого диапазона изменения как первичных, так и целевой переменной.

4. Модели демонстрируют следующие признаки неустойчивости и при обучении, и при прогнозировании в контрольные точки:

a. При обучении в разных прогонах меняется значимость переменных, на что указывает большой диапазон значений для переменных, давших наибольшее число разбиений леса.

b. При увеличении числа обучающих точек точность проверки прогноза падает и при обучении, и при прогнозировании в контрольные точки.

c. Сохраняются выбросы с малыми значениями достоверности.

Подытоживая, можно сказать, что главной проблемой исследования является то, что данные, полученные для расчетов из глобальных баз, приближенные, поскольку используют разного рода аппроксимации для показателей на территории Тюменской области. Недостаток точных данных является следствием отсутствия наземных станций, которые могли бы давать более точные данные для включения в глобальную сеть. Однако мы считаем, что, учитывая важность исследуемой территории и ряд оставшихся нерешенными вопросов, необходимо продолжить исследования в данном направлении.

БЛАГОДАРНОСТИ

Авторы благодарны А.Е. Пшеничникову за помощь в обработке и организации исходной информации.

ACKNOWLEDGEMENTS

The authors are grateful to A.E. Pshenichnikov for his help in processing and organizing the initial information.

СПИСОК ЛИТЕРАТУРЫ

1. Атлас Тюменской области. Выпуск 1. Москва-Тюмень: Главное Управление Геодезии и картографии, 1971.

2. *Козин В.В.* Природопользование на северо-западе Сибири: опыт решения проблем. Тюмень: ТГУ, 1996. 167 с.

3. *Aubinet M., Vesala T., Papale D.* Eddy Covariance: A Practical Guide to Measurement and Data Analysis. New York: Springer, 2012. 450 p. DOI: 10.1007/978-94-007-2351-1.

4. *Baccini A., Goetz S.J., Walker W.S., Laporte N.T., Sun M. et al.* Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. Nature Climate Change, 2012. V. 2(3). P. 182–185. DOI: 10.1038/nclimate1354.

5. *Breiman L.* Random Forests. Machine Learning, 2001. V. 45. P. 5–32. DOI: 10.1023/A:1010933404324.
6. *Clewley D., Whitcomb J., Akbar R., Silva A R., Berg A., Adams J.R. et al.* A method for upscaling in situ soil moisture measurements to satellite footprint scale using random forests. IEE Journal of Selected Topics in Applied Earth Observation and Removed Sensing, 2017. V. 10. No. 6. P. 2663–2673. DOI: 10.1109/JSTARS.2017.2690220.
7. *Cutler D.R., Edwards T.C., Beard K.H., Cutler A., Hess K.T.* Random forests for classification in ecology. Ecology, 2007. V. 88. No. 11. P. 2783–2792. DOI: 10.1890/07-0539.1.
8. *Fick S.E., Hijmans R.J.* WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology, 2017. V. 37. No. 12. P. 4302–4315. DOI: 10.1002/joc.5086.
9. *Friedl M., Sulla-Menashe D.* MCD12C1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 0.05Deg CMG V006. 2015. Web resource: https://lpdaac.usgs.gov/products/mcd12c1v006/ DOI: 10.5067/MODIS/MCD12C1.006.
10. *Gomes L.C., Faria R.M., de Souza E., Veloso G.V., Schaefer E.G.R., Filho E.I.F.* Modelling and mapping soil organic carbon stocks in Brazil. Geoderma, 2019. V. 340. P. 337–350. DOI: 10.1016/j.geoderma.2019.01.007.
11. *Han H., Wan R., Li B.* Estimating Forest Aboveground Biomass Using Gaofen-1 Images, Sentinel–1 Images and Machine Learning Algorithms: A Case Study of the Dabie Mountain Region. Remote Sensing, 2022. V. 14. No. 1. P. 176. DOI: 10.3390/rs14010176.
12. *Joiner J., Yoshida Y., Vasilkov A.P., Schaefer K., Jung M., Guanter L., Zhang Y., Garrity S., Middleton E.M., Huemmrich K.F., Gu L., Belelli Marchesini L.* The seasonal cycle of satellite chlorophyll fluorescence observations and its relationship to vegetation phenology and ecosystem atmosphere carbon exchange. Remote Sensing of Environment, 2014. V. 152. P. 375–391. DOI: 10.1016/j.rse.2014.06.022.
13. *Karra K., Kontgis C. Statman-Weil Z., Mazzariello J.C., Mathis M., Brumby S.P.* Global land use/land cover with Sentinel-2 and deep learning. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2021. P. 4704–4707 DOI: 10.1109/IGARSS47720.2021.9553499.
14. *Kirschbaum M.U.F., Eamus D., Gifford R.M., Roxburgh S.H., Sands P.J.* Definitions of Some Ecological Terms Commonly Used in Carbon Accounting. Net Ecosystem Exchange, 2001. P. 2–5.
15. *Mascaro J, Asner G.P., Knapp D.E., Kennedy-Bowdoin T., Martin R.E., Anderson C. et al.* A Tale of Two "Forests": Random Forest Machine Learning Aids Tropical Forest Carbon Mapping. PLoS ONE, 2014. V. 9(1). P. 1–9. DOI: 10.1371/journal.pone.0085993.
16. *Pastorello G., Trotta C., Canfora E. et al.* The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Scientific Data, 2020. V. 7 (225). DOI: 10.1038/s41597-020-0534-3.
17. *Rodriguez-Galiano V.F., Abarca-Hernandez F., Ghimire B., Chica-Olmo M., Atkinson P.M., Jeganathan C.* Incorporating Spatial Variability Measures in Land-cover Classification using Random Forest. Procedia Environmental Sciences, 2011. V. 3. P. 44–49. DOI: 10.1016/j.proenv.2011.02.009.
18. *Turner D.P., Ritts W.D., Cohen W.B., Gower S.T., Zhao M., Running S.W. et al.* Scaling Gross Primary Production (GPP) over boreal and deciduous forest landscapes in support of MODIS GPP product validation. Remote Sensing of Environment, 2003. V. 88. P. 256–270. DOI: 10.1016/j.rse.2003.06.005.
19. *Villarreal S., Guevara M., Alcaraz-Segura D., Vargas R.* Optimizing an environmental observatory network design using publicly available data. Journal of Geophysical Research: Biogeosciences, 2019. V. 124. No. 7. P. 1812–1826. DOI: 10.1029/2018jg004714.
20. *Villarreal S., Vargas R.* Representativeness of FLUXNET sites across Latin America. Journal of Geophysical Research: Biogeosciences, 2021. V. 126. No. 3. DOI: 10.1029/2020JG006090.
21. *Wang X., Liu C., Lv G., Xu J., Cui G.* Integrating Multi-Source Remote Sensing to Assess Forest Aboveground Biomass in the Khingan Mountains of North-Eastern China Using Machine-Learning Algorithms. Remote Sensing, 2022. V. 14. No. 4. DOI: 10.3390/rs14041039.

REFERENCES

1. Atlas of Tyumen region. Issue 1. Moscow-Tyumen: General Directorate of Geodesy and Cartography, 1971. (in Russian)
2. *Aubinet M., Vesala T., Papale D.* Eddy Covariance: A Practical Guide to Measurement and Data Analysis. New York: Springer, 2012. 450 p. DOI: 10.1007/978-94-007-2351-1.
3. *Baccini A., Goetz S.J., Walker W.S., Laporte N.T., Sun M. et al.* Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. Nature Climate Change, 2012. V. 2(3). P. 182–185. DOI: 10.1038/nclimate1354.
4. *Breiman L.* Random Forests. Machine Learning, 2001. V. 45. P. 5–32. DOI: 10.1023/A:1010933404324.
5. *Clewley D., Whitcomb J., Akbar R., Silva A R., Berg A., Adams J.R. et al.* A method for upscaling in situ soil moisture measurements to satellite footprint scale using random forests. IEE Journal of Selected Topics in Applied Earth Observation and Removed Sensing, 2017. V. 10. No. 6. P. 2663–2673. DOI: 10.1109/JSTARS.2017.2690220.
6. *Cutler D.R., Edwards T.C., Beard K.H., Cutler A., Hess K.T.* Random forests for classification in ecology. Ecology, 2007. V. 88. No. 11. P. 2783–2792. DOI: 10.1890/07-0539.1.
7. *Fick S.E., Hijmans R.J.* WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology, 2017. V. 37. No. 12. P. 4302–4315. DOI: 10.1002/joc.5086.
8. *Friedl M., Sulla-Menashe D.* MCD12C1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 0.05Deg CMG V006. 2015. Web resource: https://lpdaac.usgs.gov/products/mcd12c1v006/ DOI: 10.5067/MODIS/MCD12C1.006.
9. *Gomes L.C., Faria R.M., de Souza E., Veloso G.V., Schaefer E.G.R., Filho E.I.F.* Modelling and mapping soil organic carbon stocks in Brazil. Geoderma, 2019. V. 340. P. 337–350. DOI: 10.1016/j.geoderma.2019.01.007.
10. *Han H., Wan R., Li B.* Estimating Forest Aboveground Biomass Using Gaofen-1 Images, Sentinel–1 Images and Machine Learning Algorithms: A Case Study of the Dabie Mountain Region. Remote Sensing, 2022. V. 14. No. 1. P. 176. DOI: 10.3390/rs14010176.
11. *Joiner J., Yoshida Y., Vasilkov A.P., Schaefer K., Jung M., Guanter L., Zhang Y., Garrity S., Middleton E.M., Huemmrich K.F., Gu L., Belelli Marchesini L.* The seasonal cycle of satellite chlorophyll fluorescence observations and its relationship to vegetation phenology and ecosystem atmosphere carbon exchange. Remote Sensing of Environment, 2014. V. 152. P. 375–391. DOI: 10.1016/j.rse.2014.06.022.
12. *Karra K., Kontgis C. Statman-Weil Z., Mazzariello J.C., Mathis M., Brumby S.P.* Global land use/land cover with Sentinel-2 and deep learning. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2021. P. 4704–4707 DOI: 10.1109/IGARSS47720.2021.9553499.
13. *Kirschbaum M.U.F., Eamus D., Gifford R.M., Roxburgh S.H., Sands P.J.* Definitions of Some Ecological Terms Commonly Used in Carbon Accounting. Net Ecosystem Exchange, 2001. P. 2–5.
14. *Kozin V.V.* Nature management in the North-West of Siberia: experience in solving problems. Tyumen: Tyumen State University, 1996. 167 p. (in Russian)
15. *Mascaro J, Asner G.P., Knapp D.E., Kennedy-Bowdoin T., Martin R.E., Anderson C. et al.* A Tale of Two "Forests": Random Forest Machine Learning Aids Tropical Forest Carbon Mapping. PLoS ONE, 2014. V. 9(1). P. 1–9. DOI: 10.1371/journal.pone.0085993.
16. *Pastorello G., Trotta C., Canfora E. et al.* The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Scientific Data, 2020. V. 7 (225). DOI: 10.1038/s41597-020-0534-3.
17. *Rodriguez-Galiano V.F., Abarca-Hernandez F., Ghimire B., Chica-Olmo M., Atkinson P.M., Jeganathan C.* Incorporating Spatial Variability Measures in Land-cover Classification using Random Forest. Procedia Environmental Sciences, 2011. V. 3. P. 44–49. DOI: 10.1016/j.proenv.2011.02.009.

18. *Turner D.P., Ritts W.D., Cohen W.B., Gower S.T., Zhao M., Running S.W. et al.* Scaling Gross Primary Production (GPP) over boreal and deciduous forest landscapes in support of MODIS GPP product validation. Remote Sensing of Environment, 2003. V. 88. P. 256–270. DOI: 10.1016/j.rse.2003.06.005.
19. *Villarreal S., Guevara M., Alcaraz-Segura D., Vargas R.* Optimizing an environmental observatory network design using publicly available data. Journal of Geophysical Research: Biogeosciences, 2019. V. 124. No. 7. P. 1812–1826. DOI: 10.1029/2018jg004714.
20. *Villarreal S., Vargas R.* Representativeness of FLUXNET sites across Latin America. Journal of Geophysical Research: Biogeosciences, 2021. V. 126. No. 3. DOI: 10.1029/2020JG006090.
21. *Wang X., Liu C., Lv G., Xu J., Cui G.* Integrating Multi-Source Remote Sensing to Assess Forest Aboveground Biomass in the Khingan Mountains of North-Eastern China Using Machine-Learning Algorithms. Remote Sensing, 2022. V. 14. No. 4. DOI: 10.3390/rs14041039.