Промышленная ботаника, 2025. Вып. 25, № 4. С. 25–35.

УДК 528.852:004.896:581.524.4 DOI: 10.5281/zenodo.17800714

И.И. Стрельников, В.М. Остапко, Ю.В. Ибатулина

ТЕСТИРОВАНИЕ ПРИМЕНИМОСТИ БАЗОВЫХ МОДЕЛЕЙ ЭМБЕДДИНГА СПУТНИКОВЫХ ДАННЫХ ДЛЯ ДИСТАНЦИОННОЙ ИДЕНТИФИКАЦИИ ДОМИНАНТНЫХ ВИДОВ ТРАВЯНЫХ СООБЩЕСТВ

Федеральное государственное бюджетное научное учреждение «Донецкий ботанический сад»

В работе оценивается применимость эмбеддингов TESSERA для идентификации доминирующих видов в высокоразнообразных степных сообществах. Дистанционное зондирование традиционно эффективно для монокультур и лесов (точность 85–95 %), но дает слабые результаты в видово-богатых травяных сообществах из-за спектрального смешения. В исследовании использованы полевые данные 87 участков и 128-мерные эмбеддинги TESSERA, сгенерированные из годовых временных рядов Sentinel-1 и Sentinel-2. Модели Random Forest продемонстрировали высокую эффективность для трех из четырех анализируемых видов (*Festuca valesiaca* Schleich. ex Gaudin, *Stipa lessingiana* Trin. & Rupr, *Elymus repens* (L.) Gould) с ROC-AUC > 0,83, что существенно превосходит традиционные методы ($R^2 \leq 0,4$). Пространственный анализ подтвердил экологическую интерпретируемость прогнозов. Результаты открывают возможности для экономически эффективного мониторинга биоразнообразия на больших территориях.

Ключевые слова: TESSERA, Random Forest, ДЗЗ, определение видов

Цитирование: Стрельников И.И., Остапко В.М., Ибатулина Ю.В. Тестирование применимости базовых моделей эмбеддинга спутниковых данных для дистанционной идентификации доминантных видов травяных сообществ // Промышленная ботаника. 2025. Вып. 25, № 4. С. 25–35. DOI: 10.5281/zenodo.17800714

Введение

Оценка видового состава растительных сообществ представляет собой фундаментальную задачу экологического мониторинга, сохранения биоразнообразия и устойчивого управления природными ресурсами. Традиционные полевые методы исследования флоры, несмотря на высокую точность, обладают существенными ограничениями: они трудоемки, времязатратны и часто неприменимы для крупномасштабных или труднодоступных территорий [13, 14]. В условиях ускоренных антропогенных и климатических изменений особенно востребованы методы, обеспечивающие оперативную оценку состояния экосистем на больших территориях. Дистанционное зондирование Земли (далее – ДЗЗ) предлагает перспективное решение этой проблемы, позволяя осуществлять быструю, многократную и экономически эффективную оценку растительного покрова с полным пространственным охватом [11, 12].

Методы ДЗЗ уже доказали свою эффективность для картографирования видового состава в экосистемах с низким биоразнообразием. В лесных сообществах применение сверточных нейронных сетей и других методов глубокого обучения обеспечивает классификацию древесных пород с точностью 85–95 % и коэффициентами Каппа выше 0,85 [2, 7]. Аналогичные результаты достигаются в сельском хозяйстве при идентификации сельскохозяйственных культур [9, 16]. Однако в случае высокоразнообразных травяных сообществ возможности стандартных

спутниковых данных оказываются существенно ограниченными.

Идентификация видового состава в степных и луговых экосистемах с богатой флорой сталкивается с принципиальными трудностями. Увеличение видового разнообразия и пространственной неоднородности приводит к спектральному смешению сигналов от различных видов в пределах одного пикселя, что резко снижает точность классификации [3, 10]. Эмпирические исследования показывают слабую и неустойчивую связь между спектральными характеристиками стандартных спутниковых данных (например, Sentinel-2) и фактическим видовым составом естественных травяных сообществ [5, 13]. Даже при использовании передовых методов анализа, предсказание индексов разнообразия (таких как Шеннона или Симпсона) редко превышает уровни детерминации $R^2 = 0,4$ [3].

Новым перспективным направлением, способным преодолеть эти ограничения, являются методы машинного обучения на основе эмбеддингов – компактных векторных представлений сложных данных. В отличие от традиционного подхода, при котором исследователи вручную отбирают спектральные индексы или признаки, эмбеддинги автоматически извлекают и кодируют наиболее информативные паттерны из исходных данных [1]. Эта технология эффективно выявляет скрытые закономерности, такие как тонкие различия в сезонной динамике роста и развития разных сообществ, которые могут оставаться незамеченными при анализе отдельных спектральных каналов [15].

Примером реализации такого подхода является модель TESSERA, разработанная в 2025 г. [4]. Эта фундаментальная модель обрабатывает годовые временные ряды спутниковых данных Sentinel-1 и Sentinel-2, генерируя для каждого пикселя 128-мерные векторные представления с пространственным разрешением 10 м. TESSERA использует два параллельных энкодера на основе трансформеров, с помощью которых выполняет интеллектуальную компрессию исходной информации: примерно 1000 спектрально-временных измерений (61 снимок Sentinel-1 с двумя каналами и 73 снимка Sentinel-2 с двенадцатью спектральными каналами за год) сжимаются до 128 значений, сохраняющих ключевые биофизические и фенологические характеристики растительного покрова. Несмотря на очевидные преимущества, методы эмбеддинга представляют собой зарождающуюся технологию, применимость которой в ландшафтной экологии требует тщательной валидации. Хотя подобные подходы уже продемонстрировали высокую эффективность в других областях искусственного интеллекта, их потенциал для решения экологических задач, особенно в экосистемах с высоким биоразнообразием, остается неизученным.

Цель и задачи исследования

Целью исследования была оценка применимости эмбеддингов TESSERA для автоматической идентификации доминирующих видов в травяных сообществах степной зоны. Реализация этой цели позволит создать основу экономически эффективных методик мониторинга видового состава природных экосистем, где традиционные методы ДЗЗ демонстрируют ограниченную применимость. Для достижения цели были поставлены следующие задачи: 1) формирование и предобработка обучающего набора данных на основе полевых описаний степных сообществ и соответствующих им спутниковых эмбеддингов TESSERA; 2) подбор и оптимизация макропараметров классификационных моделей машинного обучения, адаптированных для работы со спецификой травяных сообществ; 3) комплексная оценка точности предложенного подхода и сравнительный анализ его эффективности.

Объекты и методики исследований

Сбор данных о видовом составе степных сообществ осуществляли в течение вегетационных периодов 2024–2025 гг. Обследования проводили на участках площадью 3×3 м², где фиксировали полный видовой состав и оценивали проективное покрытие каждого вида в процентах визуальным методом. Всего было обследовано 87 участков в пределах Донецкой Народной Республики, географическое распределение которых представлено на рисунке 1.

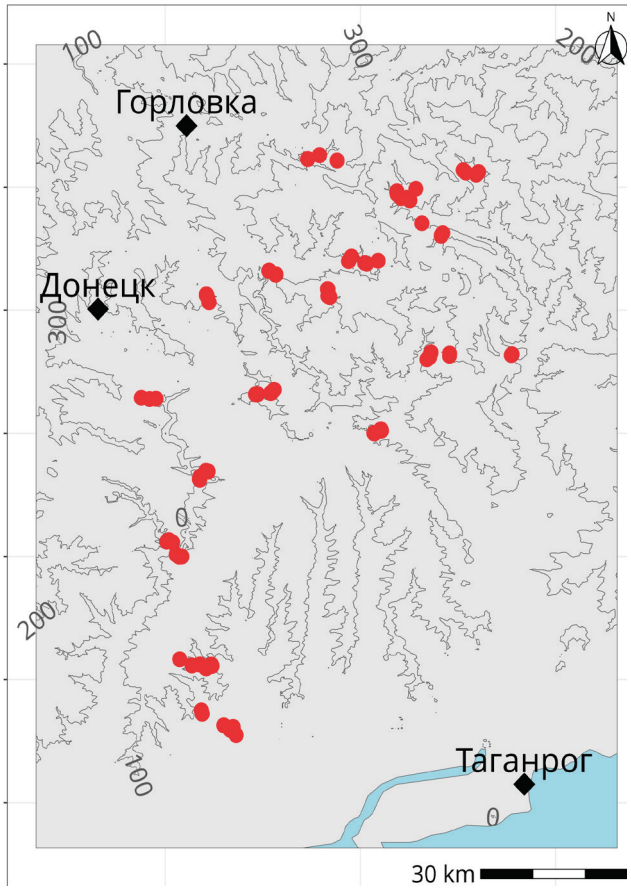

Рис. 1. Схема географического расположения обследованных участков природной травяной растительности
Fig. 1. Map showing the geographical location of the surveyed areas of natural grassland vegetation

Участки отбирались с учетом максимального разнообразия типов степной растительности в исследуемом регионе. Для последующего анализа вид считали доминирующим на участке при проективном покрытии более 10 %. Для классификации доминирующих видов по эмбеддингам TESSERA применили алгоритм случайного леса (Random Forest) – метод ансамблирования (объединения большого количества более простых древовидных моделей построенных на подвыборках данных), хорошо зарекомендовавший себя в задачах классификации биологических данных благодаря устойчивости к шуму и переобучению. В качестве обучающих данных использовали матрицу эмбеддингов TESSERA, а зависимой переменной служила бинарная оценка существенного присутствия вида на участке (1 – при покрытии >10 %, 0 – в противном случае). Учитывая несбалансированность классов в обучающих данных (количество участков с отсутствием вида существенно превышало количество участков с его присутствием), применяли балансировку классов методом апсэмплинга (upsampling), который заключается в искусственном увеличении численности наблюдений в миноритарном классе через повторную выборку имеющихся данных.

Оптимизацию гиперпараметров модели и оценку ее качества проводили с использованием процедуры повторной перекрестной проверки (3 повтора 5-кратной кросс-валидации) через пакет caret в среде R [8]. На каждом из 15 этапов обучения модель строилась на случайной выборке 80 % данных, а оставшиеся 20 % использовались для независимой оценки точности. Данный подход имитирует ситуацию, когда модель, обученная на имеющихся данных, применяется для предсказания на новых территориях, что обеспечивает объективную оценку ее прогностической способности.

Основной метрикой качества классификации служила площадь под кривой ошибок (ROC-AUC), характеризующая способность модели различать присутствие и отсутствие вида при различных пороговых значениях. Дополнительно анализировали чувствительность (sensitivity – доля правильно классифицированных участков с присутствием вида) и специфичность (specificity – доля правильно классифицированных участков без вида), а также общую точность (accuracy – доля всех правильных предсказаний в общей выборке):

$$\text{Чувствительность} = \frac{TP}{TP+FN},$$

$$\text{Специфичность} = \frac{TN}{TN+FP},$$

$$\text{Точность} = \frac{TP+TN}{TP+TN+FP+FN},$$

где TP – истинно положительные результаты, TN – истинно отрицательные, FP – ложно положительные, FN – ложно отрицательные.

Важным аспектом анализа стало определение оптимального порогового значения для конвертации прогнозируемых вероятностей в

бинарные предсказания. Стандартное значение в 0,5 демонстрировало несбалансированность метрик чувствительности и специфичности для всех моделей. Для решения этой проблемы применяли критерий Юдена, максимизирующий сумму чувствительности и специфичности, что позволило определить индивидуальные пороговые значения для каждого вида.

Обученные модели применяли ко всему набору эмбеддингов TESSERA для получения сплошных карт распространения доминирующих видов на исследуемой территории. На полученных бинарных растровых картах значение 1 соответствует пикселям с вероятностью присутствия вида выше порогового значения, а значение 0 – пикселям с низкой вероятностью существенного присутствия вида. Картирование проводили с исходным разрешением эмбеддингов TESSERA – 10 м, что обеспечило детализацию, достаточную для выявления пространственных паттернов в распределении доминантов степных сообществ.

Результаты исследований и их обсуждение

Анализ полевых данных. В ходе полевых исследований вегетационных периодов 2024–2025 гг. был получен подробный видовой состав 87 участков степной растительности. Выбор участков проведен с общей целью обеспечить высокую репрезентативность и представленность разных форм рельефа региона. Для оценки репрезентативности выделили все территории с природной и полуприродной растительностью в исследуемом регионе, после чего провели случайную выборку из 100000 точек для анализа распределения основных топографических параметров – аспекта и уклона склонов. Анализ показал, что в регионе доминируют южные и северные экспозиции склонов при минимальном представлении восточных и западных направлений. Распределение уклонов характеризуется гамма-распределением с модой около 2,4°, причем участки с уклоном более 15° составляют всего 0,1 % территории. Сравнение распределений топографических параметров в региональном масштабе и в выборке обследованных участков (рис. 2) подтверждает их сопоставимость: в исследовательской выборке сохраняется преобладание северных и южных экспозиций, а средний уклон составляет 2,8° (максимальный – 14,8°). Такое соответствие распределений, наряду с широким пространственным охватом точек наблюдений, позволяет сделать вывод о репрезентативности сформированной выборки и ее способности отражать основные варианты рельефа региона.

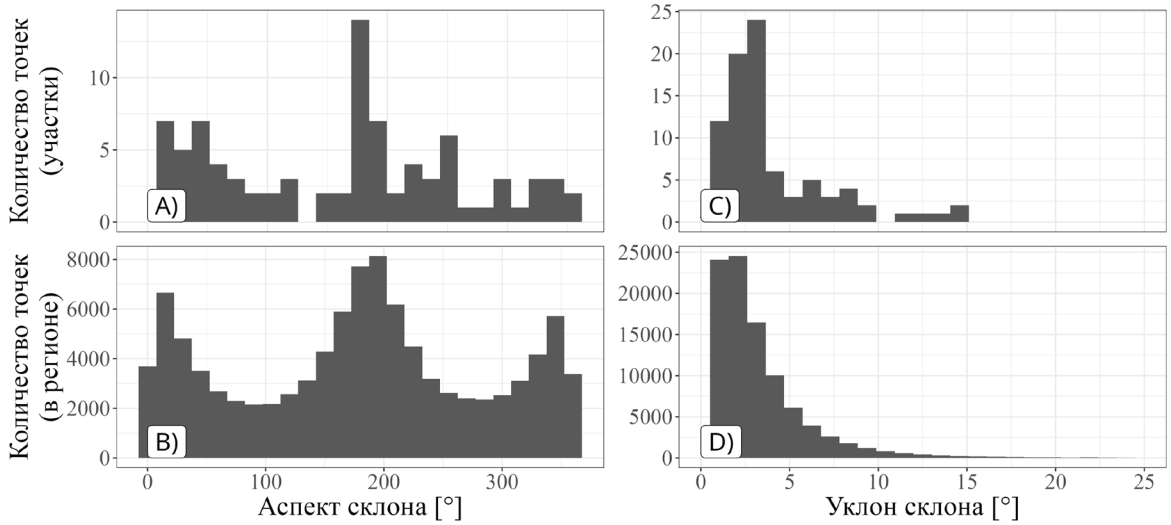

Рис. 2. Сравнение распределений встречаемости показателей аспекта (A, B) и уклона склонов (C, D) в региональном масштабе (B, D) и в выборке анализируемых участков (A, C)

Fig. 2. Comparison of relief aspect (A, B) and slope (C, D) distributions at the regional scale (B, D) and in the sample of analyzed plots (A, C)

Всего на обследованных территориях идентифицировано 243 вида растений, принадлежащих к 137 родам, 33 семействам и 19 порядкам. Анализ распределения видов по участкам выявил высокую степень неравномерности их представленности: 113 видов встречались менее чем на 10 участках из 87 обследованных, а 10 видов были обнаружены только на 8 участках.

Для построения моделей машинного обучения отобрали виды, удовлетворяющие двум ключевым критериям: достаточная представленность в обучающей выборке и экологическая значимость. При этом вид считали пригодным для анализа, если он присутствовал на более чем 15 участках с проективным покрытием более 10 %. Данным условиям соответствовали четыре доминирующих вида степной растительности: *Festuca valesiaca* Schleich. ex Gaudin (на 66 участках), *Stipa lessingiana* Trin. & Rupr. (на 16 участках), *Poa angustifolia* L. (на 16 участках) и *Elymus repens* (L.) Gould (на 18 участках). Такой отбор обеспечил баланс между статистической надежностью моделей и экологической значимостью прогнозируемых видов.

Характеристика эмбеддингов TESSERA. Предварительный анализ качества эмбеддингов TESSERA выявил их высокую информативность и взаимную независимость. Коэффициенты парной корреляции Пирсона между отдельными элементами 128-мерных векторов не превышали 0,8, что указывает на отсутствие избыточности в признаковом пространстве. Анализ дисперсии переменных показал отсутствие признаков с незначительной вариативностью, что подтверждает высокую информационную емкость эмбеддингов для конкретных участков исследования.

Визуальное сравнение изображений в натуральных цветах (Sentinel-2) и псевдоцветных композиций, построенных на основе первых трех компонент эмбеддингов TESSERA, продемонстрировало существенно большую детализацию пространственных паттернов растительного покрова в последнем случае (рис. 3). На участке возле Ольховского водохранилища (Харцызский городской совет) изображение TESSERA выявило четкие границы между различными типами степной растительности к северу и югу от водоема, которые визуально не различимы на RGB-композите Sentinel-2. Это свидетельствует о том, что эмбеддинги TESSERA сохраняют тонкие различия в структуре и динамике растительного покрова, недоступные для восприятия при использовании классических спектральных индексов.

Результаты моделей машинного обучения. В ходе сравнительного анализа алгоритмов машинного обучения (Random Forest, glmnet, xgboost) установлено, что метод случайного леса демонстрирует наилучшие результаты для всех четырех анализируемых видов. Модели glmnet характеризовались существенно меньшими значениями ROC-AUC (на 20–40 %

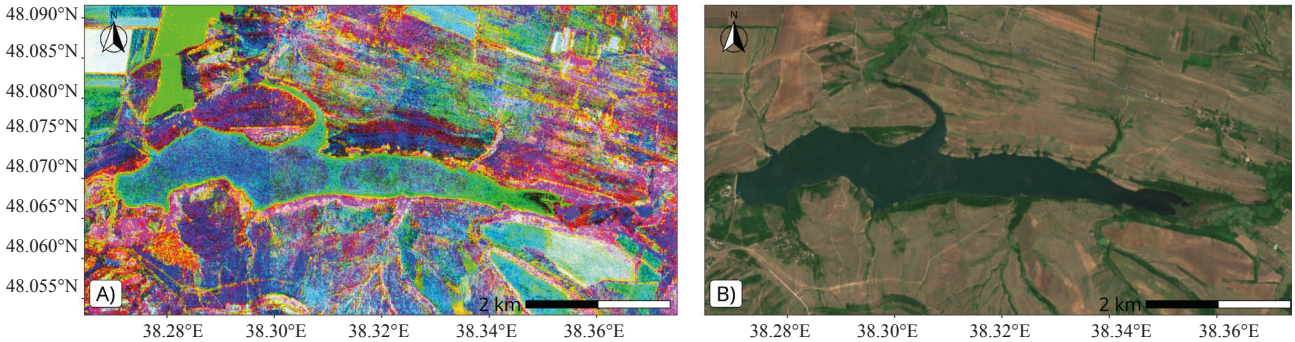

Рис. 3. Сравнение изображения территории вокруг Ольховского водохранилища (Харцызский городской совет) A) в псевдоцветах, полученных на основе первых трех компонент эмбеддингов TESSERA и B) в натуральных цветах (Sentinel-2)

Fig. 3. Comparison of images of the area around the Olkhovskoye Reservoir (Khartsyzsk City Council) A) in pseudo-colors obtained from the first three components of TESSERA embeddings and B) in natural colors (Sentinel-2)

ниже), что указывает на ограниченную способность линейных моделей выявлять сложные нелинейные зависимости в данных TESSERA. Алгоритм xgboost, напротив, проявлял признаки переобучения, проявляющиеся в нулевых значениях чувствительности или специфичности при независимой валидации. В таблице 1 приведены характеристики ROC-AUC и точности алгоритмов xgboost и glmnet.

Попытки дополнительного снижения размерности эмбеддингов через анализ главных компонент не привели к улучшению качества моделей. Напротив, во всех случаях наблюдалось снижение точности классификации, что подтверждает высокую информационную плотность и ортогональность исходных 128-мерных векторов, полученных от модели TESSERA.

Оптимизация гиперпараметра mtry алгоритма Random Forest (количество случайно выбираемых предикторов при построении каждого дерева) показала необходимость дифференцированного подхода: для *F. valesiaca* оптимальное значение составило 96, тогда как для остальных трех видов – 2. Средние метрики качества моделей, полученные в ходе 15-кратной перекрестной проверки, представлены в таблице 2.

Три из четырех построенных моделей показали высокие значения ROC-AUC (более 0,83), что свидетельствует о хорошей способности различать участки с присутствием и отсутствием вида (табл. 1). Исключение составила модель для *P. angustifolia* (ROC-AUC = 0,697), что обусловлено низкой чувствительностью (0,542) при относительно высокой специфичности (0,793). При этом точность моделей во всех случаях была выше 70 %. Данная модель демонстрирует консервативный характер предсказаний – она склонна к отрицательным прогнозам даже при реальном присутствии вида. Такое поведение может быть связано с экологической пластичностью *P. angustifolia*, который образует популяции в широком диапазоне условий среды, что затрудняет выявление устойчивых спектрально-временных паттернов в данных TESSERA.

Важно отметить, что полученная точность прогнозирования (около 80 % для большинства видов) является примечательным результатом в контексте задачи идентификации доминирующих видов в высокоразнообразных травяных сообществах. Предыдущие исследования показывали, что классические методы дистанционного зондирования (спектральные индексы, текстурные признаки) редко достигают коэффициентов детерминации более 0,4 при прогнозировании видового состава в аналогичных

Таблица 1. Средние метрики качества моделей xgboost и glmnet после оптимизации гиперпараметров и пороговых значений

Вид	xgboost		glmnet	
	ROC-AUC	Точность	ROC-AUC	Точность
F. valesiaca	0,809	0,762	0,666	0,674
S. lessingiana	0,785	0,709	0,494	0,533
P. angustifolia	0,611	0,716	0,474	0,483
E. repens	0,816	0,655	0,532	0,548

Таблица 2. Средние метрики качества моделей Random Forest после оптимизации гиперпараметров (mtry) и пороговых значений

Вид	mtry	ROC-AUC	Чувствительность	Специфичность	Точность	Порог
F. valesiaca	96	0,835	0,733	0,753	0,739	0,647
S. lessingiana	2	0,846	0,822	0,778	0,785	0,273
P. angustifolia	2	0,697	0,542	0,793	0,747	0,323
E. repens	2	0,848	0,771	0,784	0,782	0,264

условиях [3, 5]. Даже при учете ограничений настоящего исследования (небольшой объем обучающей выборки), результаты демонстрируют потенциал эмбеддингов TESSERA для решения сложных задач идентификации видового состава в природных сообществах.

Пространственный анализ прогнозов. Для оценки экологической интерпретируемости моделей была проведена классификация всей исследуемой территории с применением обученных моделей ко всем пикселям исходного набора данных TESSERA. Качество моделей оценивалось по трем критериям: 1) отсутствие случайных пространственных паттернов; 2) наличие логически обоснованных зон взаимного исключения видов; 3) соответствие предсказанных зон распространения известным экологическим предпочтениям видов.

Анализ карт распространения четырех видов в районе северного склона Ольховского водохранилища (рис. 4) выявил следующие закономерности. *Festuca valesiaca*, который был наиболее широко представлен в обучающей выборке (66 из 87 участков), демонстрирует предсказанное присутствие на большей части территории, что согласуется с его экологическим статусом широко распространенного степного вида. Прогнозы для других видов показывают четкие границы, коррелирующие с микрорельефом местности. В частности, *S. lessingiana*

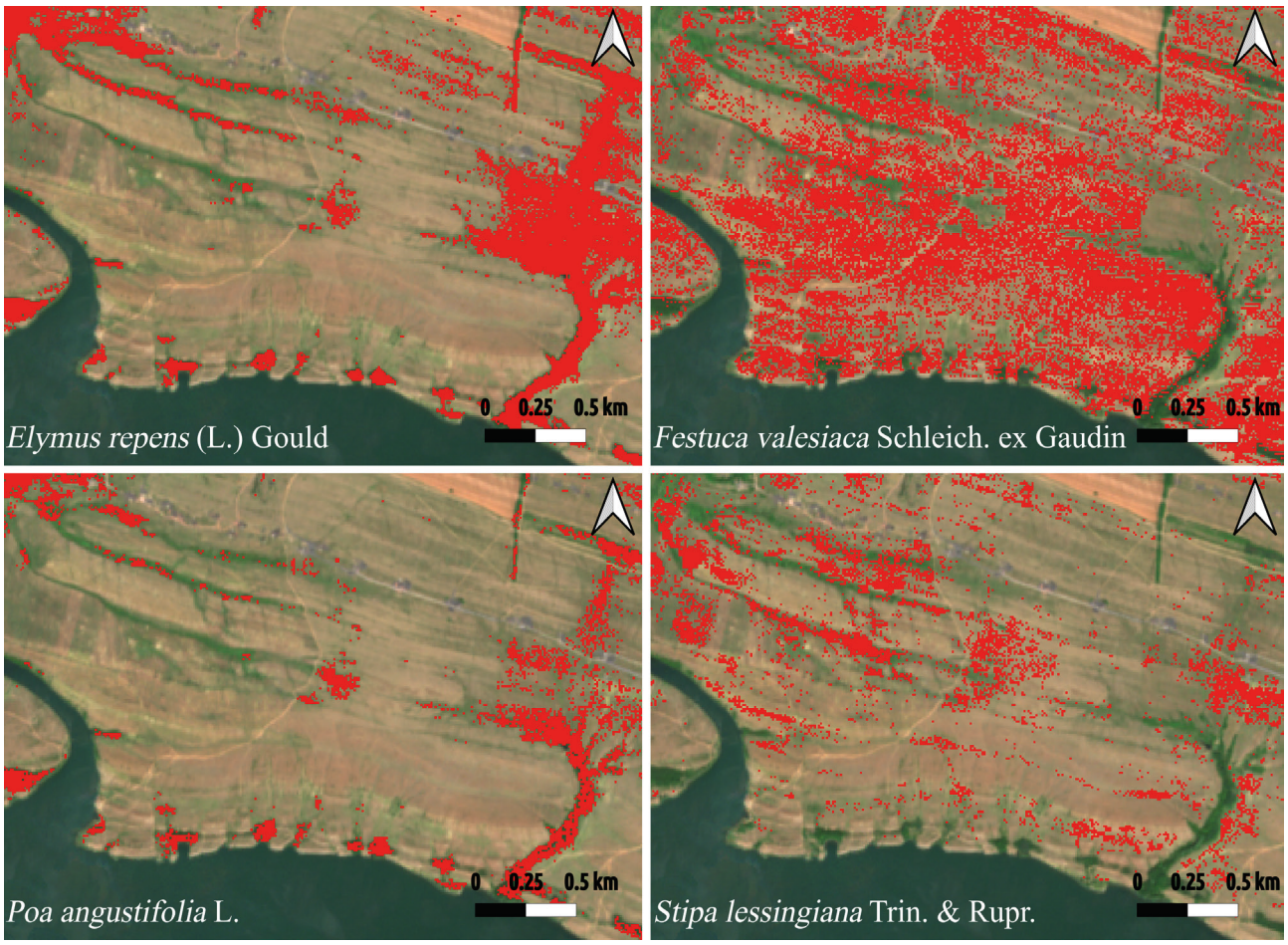

Рис. 4. Пространственные паттерны предсказанного распространения доминирующих видов (*Festuca valesiaca*, *Stipa lessingiana*, *Poa angustifolia*, *Elymus repens*) на участке северного склона Ольховского водохранилища (Харцызский городской совет), полученные с применением эмбеддингов TESSERA и моделей Random Forest.
Fig. 4. Spatial patterns of predicted distribution of dominant species (*Festuca valesiaca*, *Stipa lessingiana*, *Poa angustifolia*, *Elymus repens*) on the northern slope area of Olkhovsky Reservoir (Khartsyzk city council), derived using TESSERA embeddings and Random Forest models

преимущественно ассоциирован с локальными возвышенностями и крутыми склонами, что соответствует его ксерофильной экологической природе. Напротив, *P. angustifolia* и *E. repens* проявляют приуроченность к понижениям рельефа, особенно вдоль оврага в восточной части карты, где создаются более благоприятные гидрологические условия.

Пространственные паттерны показывают как зоны совместного распространения видов (например, *F. valesiaca* и *S. lessingiana*), так и территории взаимного исключения (первая пара видов редко встречается в тех же местах, что и вторая пара). Такая структура согласуется с известными экологическими различиями между видами и подтверждает экологическую интерпретируемость полученных моделей. Отсутствие хаотичного, «шумоподобного» распределения предсказаний указывает на устойчивость выявленных спектрально-временных паттернов и их репрезентативность для реальных сообществ.

Обсуждение результатов. Полученные результаты демонстрируют принципиальную возможность применения эмбеддингов TESSERA для автоматической идентификации доминирующих видов в высокоразнообразных степных сообществах. Достижение ROC-AUC более 0,8 для трех из четырех анализируемых видов представляет собой существенный прогресс по сравнению с традиционными методами дистанционного зондирования, которые показывают ограниченную эффективность в подобных условиях [3, 13].

Несмотря на впечатляющие результаты, необходимо учитывать ограничения данного исследования. Объем обучающей выборки (87 участков) относительно невелик для задач машинного обучения, что, вероятно, ограничивает потенциальную точность моделей. Кроме того, визуальная оценка проективного покрытия видов, используемая в качестве «золотого стандарта», сама содержит некоторую неопределенность, связанную с субъективностью экспертных оценок, что создает естественный предел точности для любых моделей, обучающихся на таких данных. Важно отметить дополнительные ограничения применимости TESSERA: для достижения высокой точности классификации требуется достаточное количество обучающих примеров для каждого анализируемого вида, так же модель может испытывать затруднения при различении таксонов с близкими экологическими и биофизическими характеристиками, что особенно актуально для родственных видов со схожими фенологическими паттернами.

Перспективы дальнейших исследований в этой области многообразны. Во-первых, необходим сбор более репрезентативных данных с использованием стандартизированных методик точечного отбора проб и, возможно, применения объективных методов оценки покрытия (например, фототочки с последующим анализом изображений). Во-вторых, перспективным направлением представляется совмещение эмбеддингов TESSERA с дополнительными источниками информации, такими как данные о рельефе, почвах или гидрологическом режиме. В-третьих, методология может быть расширена для решения задач совместного моделирования распределения видов (joint species distribution modelling), что позволит учитывать межвидовые взаимодействия и экологические взаимосвязи в сообществах.

Важно подчеркнуть, что даже при существующей точности (около 80 % для большинства видов) предложенный подход представляет практическую ценность для регионального экологического мониторинга. В условиях ограниченных ресурсов для полевых работ, возможность оперативной оценки видового состава на больших территориях с известным уровнем неопределенности может существенно улучшить процесс принятия решений в области сохранения биоразнообразия и управления природными ресурсами. Ошибки в 20 % случаев на отдельных участках могут нивелироваться при пространственном агрегировании данных для оценки состояния экосистем на более крупных территориях.

Таким образом, результаты исследования подтверждают, что эмбеддинги, генерируемые фундаментальными моделями дистанционного зондирования, открывают новые возможности для мониторинга видового состава высокораз-

нообразных травяных сообществ. Несмотря на определенные ограничения и необходимость дальнейшей оптимизации методологии, полученные результаты демонстрируют принципиальное превосходство этого подхода над классическими методами анализа спутниковых данных и создают основу для развития новых методов ландшафтной экологии и сохранения биоразнообразия.

Выводы

Исследование подтвердило возможность применения эмбеддингов TESSERA для идентификации доминирующих видов в высокоразнообразных степных сообществах. Модели на основе Random Forest продемонстрировали высокую эффективность для трех из четырех анализируемых видов (*Festuca valesiaca*, *Stipa lessingiana*, *Elymus repens*), достигнув ROC-AUC > 0,83. Даже при небольшом объеме обучающей выборки (87 участков) результаты существенно превосходят традиционные методы ДЗЗ ($R^2 \leq 0,4$).

Ключевое достижение – выявление способности 128-мерных векторов TESSERA кодировать скрытые биофизические и фенологические паттерны, позволяющие дифференцировать видовой состав в условиях спектрального смешения. Пространственный анализ подтвердил экологическую интерпретируемость моделей: предсказанные зоны распространения соответствуют известным экологическим предпочтениям видов.

Несмотря на ограничения, связанные с объемом данных, текущая точность моделей (~80 %) представляет практическую ценность для регионального экологического мониторинга. Технологии эмбеддингов открывают новые возможности в ландшафтной экологии, позволяя преодолеть ограничения классических подходов и переходить к оперативной оценке биоразнообразия на больших территориях.

Работа выполнена в рамках государственного задания ФГБНУ Донецкий ботанический сад по теме «Классификация почвенно-растительного покрова с помощью методов дистанционного зондирования Земли» (Регистрационный № 124101500495-0).

1. *Brown C.F.*, Kazmierski M.R., Pasquarella V.J., Rucklidge W.J., Samsikova M., Zhang C., Shelhamer E., Lahera E., Wiles O., Ilyushchenko S., Gorelick N., Zhang L.L., Alj S., Schechter E., Askay S., Guinan O., Moore R., Boukouvalas A., Kohli P. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data // arXiv:2507.22291. ArXiv, 2025.
2. *Deur M.*, Gašparović M., Balenovic I. Tree Species Classification in Mixed Deciduous Forests Using Very High Spatial Resolution Satellite Imagery and Machine Learning Methods // Remote Sensing. 2020. Vol. 12, Iss. 23: 3926.
3. *Fauvel M.*, Lopes M., Dubo T., Rivers-Moore J., Frison P., Gross N., Ouin A. Prediction of plant diversity in grasslands using Sentinel-1 and -2 satellite image time series // Remote Sensing of Environment. 2020. Vol. 237: 111536.
4. *Feng Z.*, Atzberger C., Jaffer S., Knezevic J., Sormunen S., Young R., Lisaius M.C., Immitzer M., Jackson T., Ball J., Coomes D.A., Madhavapeddy A., Blake A., Keshav S. TESSERA: Precomputed FAIR Global Pixel Embeddings for Earth Representation and Analysis // arXiv:2506.20380. ArXiv, 2025.
5. *Imran H.A.*, Gianelle D., Scotton M., Rocchini D., Dalponte M., Macolino S., Sakowska K., Pornaro C., Vescovo L. Potential and Limitations of Grasslands α-Diversity Prediction Using Fine-Scale Hyperspectral Imagery // Remote Sensing. 2021. Vol. 13, N 14: 2649.
6. *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. Brussels, Belgium: IEEE, 2021. P. 4704–4707.
7. *Kattenborn T.*, Eichel J., Wiser S., Burrows L., Fassnacht F., Schmidtlein S. Convolutional Neural Networks accurately predict cover fractions of plant species and communities in Unmanned Aerial Vehicle imagery // Remote

Sensing in Ecology and Conservation. 2020. Vol. 6, Iss. 4. P. 472–486.
8. *Kuhn M.* Building Predictive Models in R Using the caret Package // Journal of Statistical Software. 2008. Vol. 28, N 5. P. 1–26.
9. *Lake T.*, Runquist R. B., Moeller D. Deep learning detects invasive plant species across complex landscapes using Worldview2 and Planetscope satellite imagery // Remote Sensing in Ecology and Conservation. 2022. Vol. 8, Iss. 6. P. 875–889.
10. *Lopes M.*, Fauvel M., Ouin A. Girard S. Spectro-Temporal Heterogeneity Measures from Dense High Spatial Resolution Satellite Image Time Series: Application to Grassland Species Diversity Estimation // Remote Sensing. 2017. Vol. 9, N 10: 993.
11. *McPartland M.*, Falkowski M., Reinhardt J., Kane E., Kolka R., Turetsky M., Douglas T., Anderson J., Edwards J., Palik B., Montgomery R. Characterizing Boreal Peatland Plant Composition and Species Diversity with Hyperspectral Remote Sensing // Remote Sensing. 2019. Vol. 11, N 14: 1685.
12. *Pinto Ledezma J.*, Cavender-Bares J. Predicting species distributions and community composition using satellite remote sensing predictors // Scientific Reports. 2021. Vol. 11: 16448.
13. *Rocchini D.*, Boyd D.S., Féret J.-B., Foody G.M., He K.S., Lausch A., Nagendra H., Wegmann M., Pettorelli N. Satellite remote sensing to monitor species diversity: potential and pitfalls // Remote Sensing in Ecology and Conservation. 2016. Vol. 2, Iss. 1. P. 25–36.
14. *Schweiger A.*, Laliberté E. Plant beta-diversity across biomes captured by imaging spectroscopy // Nature Communications. 2022. Vol. 13: 2767.
15. *Sumbul G.*, Xu C., Dalsasso E., Tuia D. SMARTIES: Spectrum-Aware Multi-Sensor Auto-Encoder for Remote Sensing Images // arXiv:2506.19585. arXiv, 2025.
16. *Zhang J.*, Zhang Y., Zhou T., Sun Y., Yang Z., Zheng S. Research on the identification of land types and tree species in the Engebei ecological demonstration area based on GF-1 remote sensing // Ecological Informatics. 2023. Vol. 77: 102242.

Поступила в редакцию: 10.11.2025

UDC 528.852:004.896:581.524.4

TESTING THE APPLICABILITY OF FOUNDATIONAL SATELLITE DATA EMBEDDING MODELS FOR REMOTE IDENTIFICATION OF DOMINANT SPECIES IN HERBACEOUS COMMUNITIES

I.I. Strelnikov, V.M. Ostapko, Yu.V. Ibatulina

Federal State Budgetary Scientific Institution «Donetsk botanical garden»

This study evaluates the applicability of TESSERA embeddings for identifying dominant species in high-diversity steppe communities. Remote sensing has traditionally been effective for monocultures and forests (accuracy 85-95%), but yields poor results in species-rich grasslands due to spectral mixing. The research utilized field data from 87 plots and 128-dimensional TESSERA embeddings generated from annual time series of Sentinel-1 and Sentinel-2. Random Forest models demonstrated high efficiency for three of the four analyzed species (*Festuca valesiaca* Schleich. ex Gaudin, *Stipa lessingiana* Trin. & Rupr, *Elymus repens* (L.) Gould) with ROC-AUC >0.83, substantially outperforming traditional methods ($R^2 \leq 0.4$). Spatial analysis confirmed the ecological interpretability of predictions. Results open opportunities for cost-effective biodiversity monitoring across large territories.

Key words: TESSERA, Random Forest, remote sensing, species identification

Citation: Strelnikov I.I., Ostapko V.M., Ibatulina Yu.V. Testing the applicability of foundational satellite data embedding models for remote identification of dominant species in herbaceous communities // Industrial botany. 2025. Vol. 25, N 4. P. 25–35. DOI: 10.5281/zenodo.17800714