УДК 004.9

Технологии машинного обучения и обработки таксационных данных для визуализации геоинформационного моделирования лесных экосистем (Часть 2)

Machine learning technologies and taxation data processing for visualization of geoinformation modeling of forest ecosystems (Part 2)

Вагизов / Vagizov M.
Марсель Равильевич
(bars-tatarin@yandex.ru)
кандидат технических наук, доцент.
ФГБОУ ВО «Санкт-Петербургский государственный лесотехнический университет имени С. М. Кирова» (СПбГЛУ им. С. М. Кирова),
заведующий кафедрой информационных систем и технологий.
г. Санкт-Петербург

Потапов / Potapov A.
Артем Павлович
СПбГЛУ им. С. М. Кирова,
магистр кафедры информационных систем и технологий.
г. Санкт-Петербург

Ключевые слова: машинное обучение – machine learning; данные дистанционного зондирования – remote sensing data; анализ данных – data analysis; геоинформационное моделирование – geoinformation modeling; лесное хозяйство – forestry.

Методы геоинформационного моделирования лесных экосистем подразумевают не только визуализацию пространственных данных, но и данных, входящих в состав моделируемых объектов, к таким данным относятся таксационные данные, характеризующие качественный состав насаждений. В продолжение первой части статьи авторов, опубликованной в журнале "Информация и Космос" в 2022 г. № 3, приводятся результаты статистической обработки данных таксации на основе машинного обучения и оценки достоверности смоделированных, прогнозируемых показателей: высоты, диаметров, запаса древостоя.

Methods of geoinformation modeling of forest ecosystems imply not only the visualization of spatial data, but also the data of the objects being modeled, such data include taxation data characterizing the qualitative composition of plantations. In continuation of the first part of the article, the authors present the results of statistical processing of inventory data based on machine learning and the assessment of the reliability of modeled, predicted indicators: height, diameter, growing stock.

Введение

Задача геоинформационного моделирования состоит в сопровождении визуализации обрабатываемых данных, одним из способов визуализации данных являются графические инфограммы. Методы машинного обучения [1, 2] позволяют проводить анализ больших объёмов информации и моделировать данные древостоев на основе заложенных алгоритмов. После определения методики последовательного моделирования таксационных показателей [3] нам необходимо проверить точность полученных прогнозов моделей. Для этого необходимо провести анализ статистики и сравнить все пары выборок в разрезе – истинные-смоделированные (табл. 1). Тестирование библиотек машинного обучения выполнялось с применением дополнительного программного обеспечения: Microsoft Excel для предобработки данных; среда разработки на языке Python - Jupyter Notebook, используемые библиотеки: Pandas, Seaborn, NumPy.

Анализ таблицы с высот древостоя

Среднее значение истинной выборки – 20,2 м, а смоделированной – 20,5 м. При этом в обоих случаях медиана равна значению в 23 м, первый квартиль также равен между собой (15 метров), 3 квартиль имеет различие в 1 метр (26 м против 25 м). Также различается максимальное значение – 36 против 31 м. Стандартное отклонение почти равно – 6,8 против 6,7. Тест Колмогорова-Смирнова показывает, что обе выборки подчиняются одному закону распределения. В общем, таблица показывает, что две выборки имеют незначительные отличия моделировании данных в паре истинные-смоделированные. Далее перейдем к возрасту древостоя. Средние значения 82,1 лет и 83,9 лет очень близки между собой. Стандартное отклонение тоже похоже – 36,3 и 38,1. Все квартили во всех случаях отличаются на 10 лет (половину класса возраста). Тест Колмогорова-Смирнова также показывает однородность выборок. Выборки можно считать очень схожими между собой. По очереди идет три

анализируемых показателя, точность которых не высокая: бонитет, тип леса, тип условий местопроизрастания. Это связано с тем, что значения этих показателей категориальные, а не количественные, это означает, что при минимальной ошибке классификации значений будет качественно изменяться характеристика выдела, так, например, бонитет первого класса может сильно отличаться от бонитета второго класса, как и кисличник от черничника. Поэтому вычисление этих показателей можно считать экспериментальными. Также бонитеты обычно вычисляются по специальным таблицам, а типы леса и типы условий местопроизрастания устанавливаются глазомерным способом при наземной таксации древостоев.

Оставшийся важный параметр моделирования – запас древостоя. Данная величина является сложнейшей для вычисления в лесном хозяйстве, потому что даже при глазомерно-измерительном способе возможно допустить большую ошибку, как и в методах дистанционной таксации. Тем не менее попробуем оценить истинные и смоделированные значения запасов. Средние величины очень близки друг к другу – 64,9 кубометров против 69,2 кубометров на гектар. Стандартная ошибка гораздо больше в истинной выборке, это связано с тем, что запас является такой величиной, которая имеет настолько огромные колебания, что применяемые в исследовании алгоритмы методов машинного обучения не в состоянии их уловить. Первый квартиль сильно различается – 25 против 49 кубометров, второй квартиль – 47 и 64 кубометра. Третий квартиль очень схожий, 86 и 83 кубометров соответственно. Минимальные и максимальные значения различаются сильно, причины те же, что и в разнице стандартных отклонений – большие колебания показателей запаса.

Основываясь на статистических и теоретических исследованиях полученных данных результаты моделирования выглядят точными. Но для более полной проверки обратимся к визуализации данных, впоследствии данные графики могут быть интегрированы в состав геоинформационной модели выбранного участка леса. Сначала проанализируем распределение истинных и смоделированных значений высот (рис. 5.).

На рис. 1 мы видим, какое различие имеют две разные выборки, одна из которых полностью предсказана и показана на графике красным цветом (далее смоделированная выборка везде будет иметь красный цвет). Можно увидеть, что кривая распределения имеет схожий изгиб, но при этом смоделированные значения предсказывались с шагом пропуска примерно в один метр (что особенно видно в диапазоне 15–20 метров), и компенсация провалов происходила за счёт ближайших классов значений высоты – например, недобор на значении 16 метров компенсируется перебором на значении 17 метров. Это означает, что имела место ошибка, которая чаще всего равна 1–3 метра высоты. Чтобы дополнительно оценить степень ошибки, воспользуемся диаграммой, которая имеет название «ящик с усами» (рис. 2). Данный вид графиков позволяет отразить основные статистики выборки – минимальные и максимальные

Таблица 1

Основные статистики истинных и смоделированных выборок таксационных показателей

Stats	*H*	H_{pred}	*A*	A_{pred}	*B*	B_{pred}	*T*	T_{pred}	*U*	U_{pred}	*M*	M_{pred}
mean	20,2	20,5	82,1	83,9	2,5	2,4	3,1	2,1	3,5	3,3	64,9	69,2
std	6,8	6,7	36,3	38,1	0,9	0,8	2,3	1,7	2,4	2,3	60,6	32,3
min	4,0	4,0	10,0	10,0	1,0	1,0	1,0	1,0	1,0	1,0	0,0	15,0
25%	15,0	15,0	60,0	50,0	2,0	2,0	2,0	1,0	2,0	2,0	25,0	49,0
50%	23,0	23,0	90,0	80,0	2,0	2,0	2,0	1,0	3,0	2,0	47,0	64,0
75%	26,0	25,0	100,0	110,0	3,0	3,0	4,0	2,0	5,0	5,0	86,0	83,0
max	36,0	31,0	200,0	170,0	5,0	5,0	17,0	12,0	12,0	12,0	583,0	317,0

*где *H* – высота, м; *A* – возраст породы, лет; *B* – бонитет; *T* – тип леса; *U* – тип условий местопроизрастания; *M* – запас, м^3\га, приставка *p* – означает предсказанные значения; *Stats* – статистики, *mean* – среднее значение; *std* – стандартное отклонение; min – минимальное значение; 25% – 1 квартиль; 50% – медиана (2 квартиль); 75% – 3 квартиль; max – максимальное значение, подстрочная *pred* – спрогнозированная выборка.

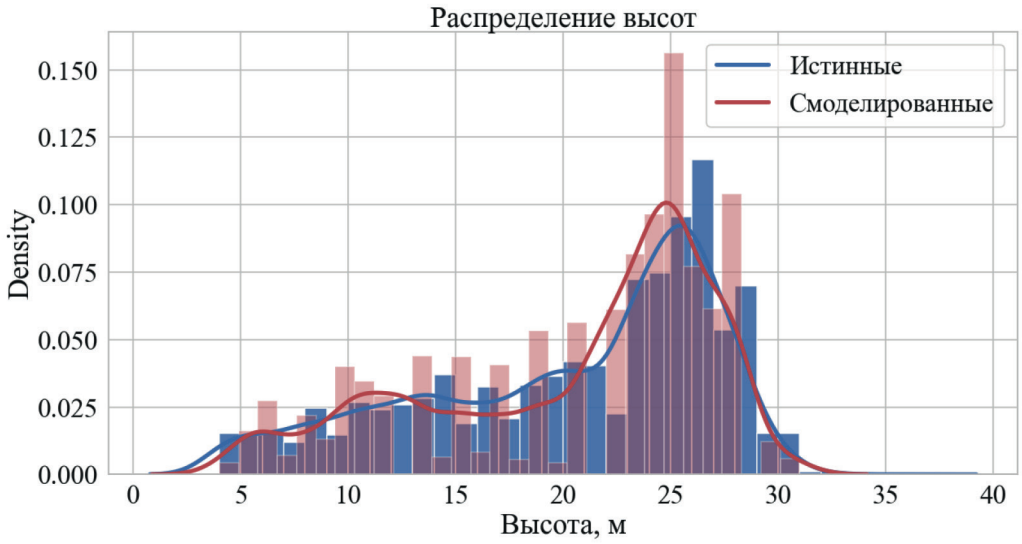

Рис. 1. Распределение истинных и спрогнозированных значений высот

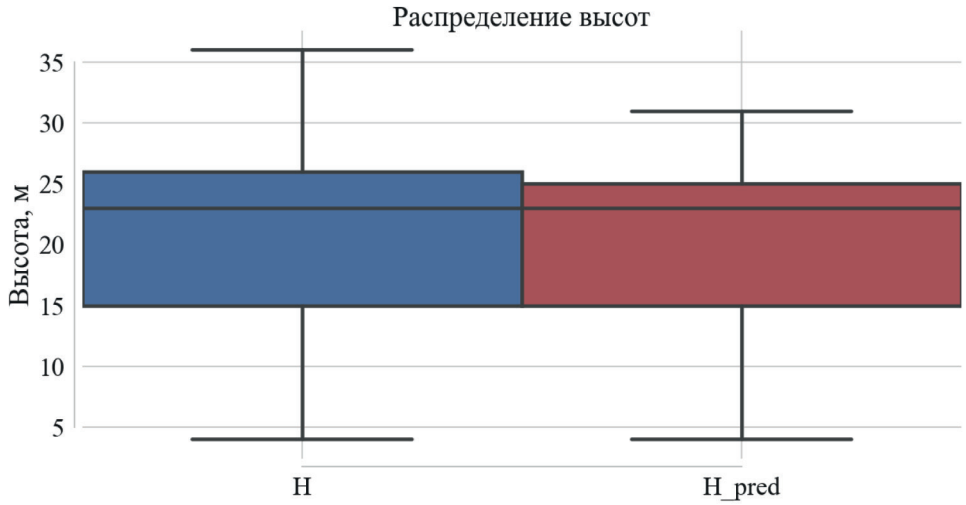

Рис. 2. Распределение высот в виде ящика с усами, где приставка _pred – спрогнозированная выборка

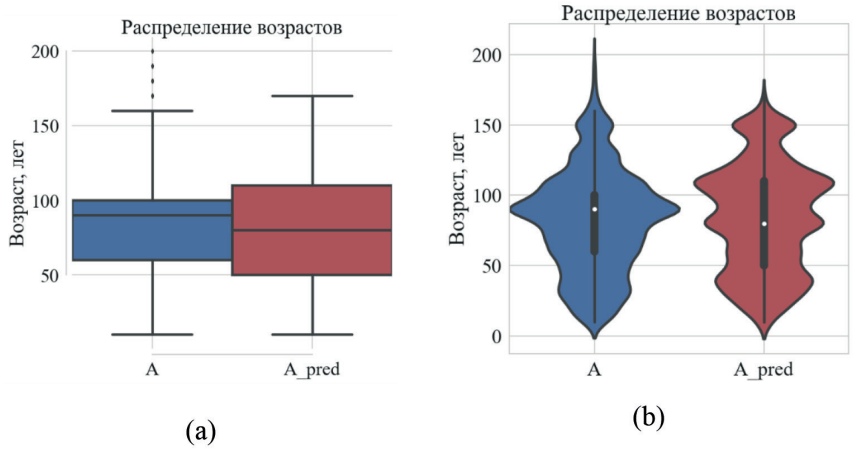

Рис. 3 (a) – график типа ящик с усами для возрастов, (b) – скрипичный график

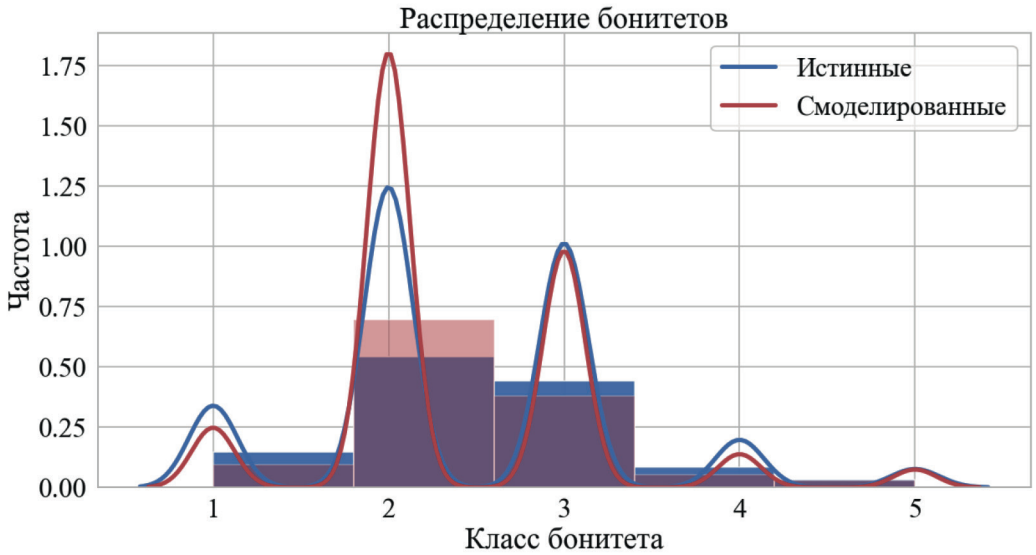

Рис. 4. Гистограмма распределения бонитетов

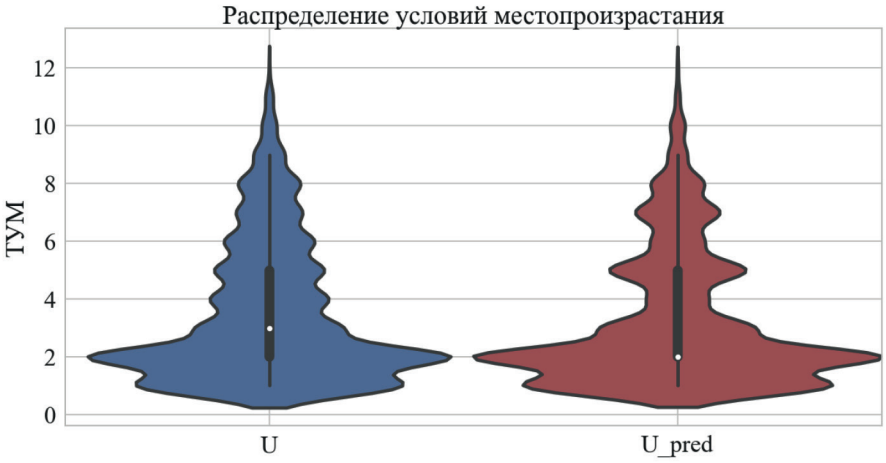

Рис. 5. Скрипичный график для типов условий местопроизрастания

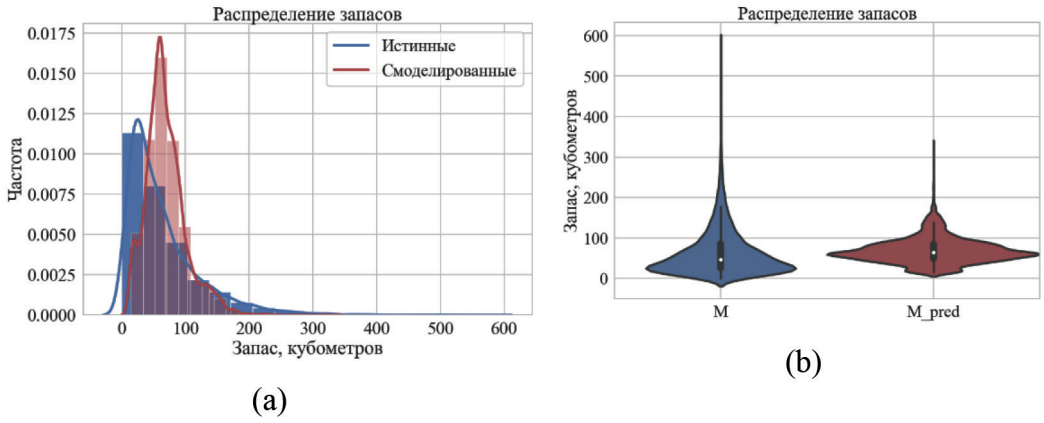

Рис. 6 (a) – гистограмма распределения для значений запасов, (b) – скрипичный график для запасов

значения, квартили, медиану, а также выбросы при их наличии.

На рис. 2 видно, что распределение двух выборок схожее. Медианы в этих случаях одинаковы, сильно различаются значения максимальной величины и совсем немного значение минимальной высоты и третий квартиль (на 1 метр).

Рисунки 3 (a) и 3 (b) показывают, что размах предсказанных возрастов больше, чем у истинных величин, также различается вид распределения – он больше похож на бимодальный, а не на мономодальный. Скрипичный график отражает основные статистики и распределение значений: чем больше формы значений похожи друг на друга, тем правильнее прогнозы используемого алгоритма.

На рис. 4. распределение бонитетов более правильное, в отличие от распределения возрастов. Несмотря на схожее с истинным распределение с определенной ошибкой, значения квартилей и медианы остается одинаковым, что означает, что величина ошибки остается небольшой.

На рис. 5 анализ распределения, алгоритм обработки данных реализован достаточно точно, но не идеально: видно, что шаг предсказания равен одному (имеются отклонения), примерно такие же, как и при моделировании высот элементов леса.

Самым сложным для дистанционной таксации традиционно является определение запасов древостоя (рис. 6). Это крайне затруднительный процесс, так как он требует большой точности основных показателей – диаметров, высот. На рис. 10 (a) и 10 (b) показано, как алгоритм справился с задачей прогноза запасов на основе входных данных. Видно, что в среднем алгоритм завышает запасы по выделам, из-за чего предсказанное распределение своим центром смещается вправо, также вырастает пик (меняются коэффициенты асимметрии и эксцесса). Для того, чтобы как-то компенсировать завышение смоделированного запаса, возможно применение конверсионных коэффициентов.

Нужно учитывать, что распределения истинных и прогнозных выборок могут быть практически одина-

Таблица 2

Величина ошибки прогнозов для разных этапов моделирования таксационных параметров, количественные признаки

Входной показатель	Por, D		Por, D, H		Por, D, H, A, B, T, U	
Выходной показатель	H		A		M	
Величина ошибки, %	Относительное количество ошибок					
	%	% накоп	%	% накоп	%	% накоп
0	26,41	26,41	3,39	3,39	1,00	1,00
1...5	28,03	54,44	21,10	24,49	4,93	5,93
6...10	23,95	78,39	18,38	42,87	5,59	11,52
11...15	10,81	89,20	15,54	58,41	5,58	17,10
16...20	7,55	96,75	14,04	72,46	5,39	22,48
21...25	1,82	98,57	9,65	82,11	5,43	27,92
26...30	0,66	99,23	5,24	87,35	5,41	33,33
31...35	0,54	99,77	3,45	90,79	4,84	38,17
36...50	0,23	100,00	4,99	95,79	16,22	54,39
51...75	-	-	2,50	98,28	25,89	80,28
76...100	-	-	0,96	99,24	12,65	92,93
101...200	-	-	0,69	99,93	4,98	97,92
201...300	-	-	0,07	100,00	1,09	99,01
>300	-	-	-	-	0,99	100,00

*где %накоп – накопительный процент

ковыми, но это не означает, что в каждом отдельно взятом случае прогноза отсутствует ошибка. Это может означать, что алгоритм прогнозирования выдает те же самые значения, характерные для истинной выборки, но для разных случаев этой выборки. Чтобы показать величины ошибки при каждом этапе моделирования, воспользуемся таблицами ошибок для каждого конкретного случая прогноза.

Также разделим количественные и качественные показатели. К первым относятся высоты, возраста и запасы насаждений. Ко вторым относятся класс бонитета, тип леса и ТУМ (тип условий местопроизрастания). Для них нужно оценивать количество безошибочных прогнозов, а остальные величины ошибок нас не интересуют, т.к. эти данные категорийные. Если использовать способы визуализации для этих показателей и знать, для какого процента случаев у нас имеются безошибочные прогнозы, то в совокупности можно понять, насколько точно алгоритм машинного обучения прогнозирует значения. Исходя из методики, представленной в работе [1], формируем таблицу ошибок.

Проанализируем таблицу 2. Если ссылаться на лесоустроительную инструкцию и сравнить допустимые ошибки при разных способах таксации, то можно примерно оценить точность предсказанных значений. К примеру, допустимая ошибка при измерении высот глазомерно-измерительным способом равна 8%, а просто глазомерным способом – 10%. Первому значению удовлетворяют в сумме примерно 72% значений, а второму значению – 78,39% значений. Если принять средний возраст равным 100 годам и характеристику, что большая часть насаждений – хвойные, то можно сказать следующее: в 82,11% случаев ошибка в определении может составлять максимум до 25 лет или чуть более, чем один класс возраста. При этом половина всех предсказанных значений имеют ошибку от 0 до 13 лет, а четверть – до 5 лет. В анализе данных запаса насаждений виден большой разброс значений и большие ошибки. Как говорилось выше, прогнозирование запасов требует введение конверсионных коэффициентов. Также стоит отметить, что если не использовать прогнозирование бонитетов, а находить их по бонитетным таблицам на основе уже предсказанной высоты, то можно добиться повышения точности определения, как бонитетов, так и запасов. Но в исследовании данная задача не рассматривалась: необходимо было испытать и опробовать алгоритм моделирования таксационных описаний, используя исключительно машинное обучение.

В таблице 3 видно, что классы бонитета безошибочно определяются в 64,66% случаев (ошибка совершается каждый третий прогноз). В этом показателе у нас всего 5 категорий (1–5 классы бонитетов). Для типов леса процент безошибочного определения меньше (39,23%), но при этом количество категорий составляет 17. Это означает, что из 100 выделов в 40 случаях тип леса моделируется абсолютно правильно. С анализом типов условий местопроизрастания можно сделать следующие выводы: процент безошибочных прогнозов наивысший (69,71%), при этом количество категорий равно 12 (больше почти в 2,5 раза, нежели у классов бонитета).

Интеграция визуализации данных машинного обучения в состав технологии геоинформационного моделирования лесных экосистем. Состав вспомогательных функций любой системы определяется её архитектурой. Геоинформационное моделирование позволяет анализировать данные, подлежащие статистическому анализу объектов моделирования, для такого типа анализа необходима система, сочетающая в себе как технологии визуализации, так и технологии машинной обработки информации, к такой системе можно отнести интеллектуальную геоинформационную систему, предложенную авторами в работе [4]. Пусть

Таблица 3

Величина ошибки прогнозов для разных этапов моделирования таксационных параметров, качественные признаки

Входной показатель	Por, D, H, A	Por, D, H, A, B	Por, D, H, A, B, T
Целевой показатель	B	T	U
Количество абсолютно точных прогнозов, %	64,66	39,23	69,71
Количество категорий	5	17	12

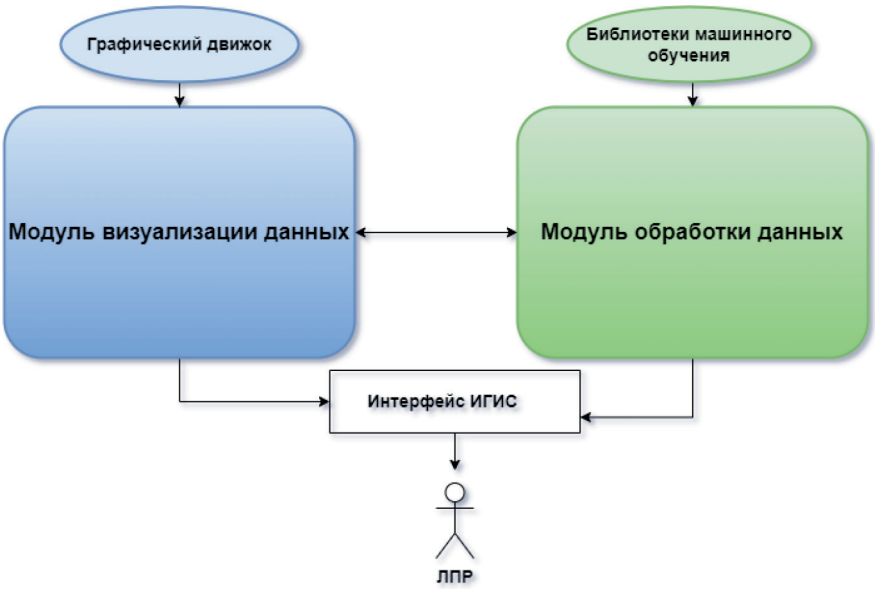

Рис. 7. Компоненты интерфейса и связь модулей интегрированных в состав ИГИС

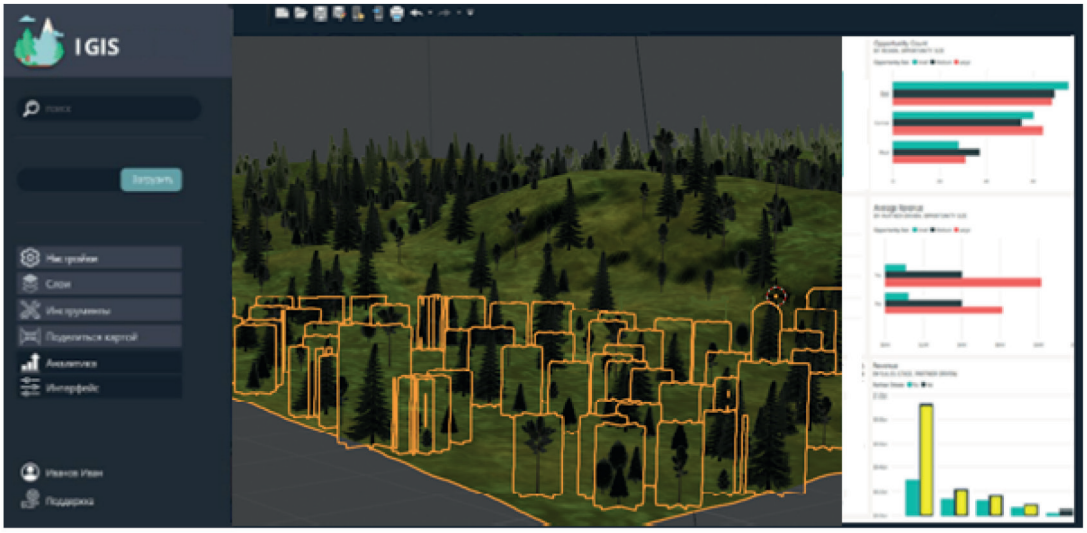

Рис. 8. Прототип интерфейса визуализации данных геоинформационного моделирования и обработки таксационной информации

в задаче проектирования интеллектуальной геоинформационной системы (ИГИС) необходимо отображение данных машинного обучения, для этого необходимо интегрировать прикладной модуль обработки информации как отдельного элемента в составной части модульной структуры ИГИС. Представим совокупность инструментальных компонентов интеллектуальной обработки информации (библиотеки, методики, модули) как функциональный компонент подсистемы ИГИС – «Модуль обработки данных» (рис. 7).

ЛПР – лицо, принимающее решение.

Вне зависимости от типа интегрируемой библиотеки для обработки массива данных (Pandas, ScikitLearn, NumPy) в составе модуля обработки данных, существует определённая функция подключения работы с данными. Реализация открытых библиотек в большинстве случаев используется посредством программирования на языке Python, который удобен для реализации web-приложений. [5] Геоинформационные данные, представляющие из себя отдельные объекты в модели, которые связываются программными средствами, используя языки визуального скриптинга, например, BluePrint [6, 7] на основе методов графического программирования. Данные методы основаны на описании взаимодействия различных объектов в моделируемой среде с помощью графовой системы [8, 9] построения взаимодействующих компонентов, когда структура данных гетерогенна. В результате интеграции модуля визуализации данных в состав ИГИС прототип представления информации, комплексирования данных геоинформационного моделирования и обработки таксационных данных при помощи методов машинного обучения выглядит следующим образом (рис. 8).

Выводы

В результате исследования продемонстрировано, какие возможности открывает машинное обучение в области анализа таксационных данных. В некоторых случаях точность не на высоком уровне, но при определении некоторых показателей она остается на приемлемом уровне (высоты, возраста, бонитеты, ТУМ). Если модернизировать методику прогноза, использовать большее количество данных и помимо машинного обучения различные таксационные таблицы, старые таксационные описания и результаты наземных исследований, то возможно добиться повышения точности прогноза для различных показателей и улучшить имеющиеся результаты. В качестве интеграции процессов обработки информации средствами машинного обучения в состав геоинформационного моделирования [10, 11] необходимо использовать модульную структуру проектируемой системы с преимущественным подходом прототипно-ориентированных языков программирования.

Литература

1. Братских, И. А. Искусственный интеллект, машинное обучение, глубокое обучение и нейронные сети / И.А. Братских // Правопорядок в России: проблемы совершенствования : сборник материалов XIV Всероссийской конференции (Москва, 11–14 февраля 2020 г.). – Москва : Московский университет МВД России имени В. Я. Кикотя, 2020. – С. 285–286.

2. Бутенко, А. И. Машинное обучение в инженерных исследованиях / А.И. Бутенко // Наука и Образование. – 2020. – Т. 3, № 4. – С. 28.

3. Потапов, А. П. Технологии машинного обучения и обработки таксационных данных для визуализации геоинформационного моделирования лесных экосистем (Часть 1) / А.П. Потапов, М.Р. Вагизов // Информация и Космос. – 2022. – № 3. – С. 125–129.

4. Разработка интеллектуальной геоинформационной системы для отрасли лесного хозяйства / М.Р. Вагизов, Е.П. Истомин, О.Н. Колбина [и др.] // Геоинформатика. – 2021. – № 3. – С. 4–13.

5. Ногоева, Г. Д. Создание веб приложений посредством библиотеки языка программирования Python / Г.Д. Ногоева, К.Д. Абыкеев // Вестник Кыргызстана. – 2021. – № 1-2. – С. 179–182.

6. Воробьева, Ю. А. Визуальное программирование на примере Blueprint / Ю.А. Воробьева // Лучшая студенческая статья 2018 : сборник статей XVIII Международного научно-исследовательского конкурса. В 2 частях (Пенза, 20 декабря 2018 г.) / Ответственный редактор Г.Ю. Гуляев. Том Часть 1. – Пенза : МЦНС «Наука и Просвещение», 2018. – С. 64–67.

7. Неизвестный, М. А. Преимущества визуальных языков програмирования на примере Blueprint в среде игрового движка Unreal Engine 4 / М.А. Неизвестный // Экономика и социум. – 2016. – № 8 (27). – С. 504–508.

8. Заяц, А. М. Графовые модели в анализе и синтезе беспроводных сенсорных сетей / А.М. Заяц // Информационные системы и технологии: теория и практика : научно-техническая конференция Института леса и природопользования СПбГЛТУ (Санкт-Петербург, 25 февраля 2022 г.). – Выпуск 14. – Санкт-Петербург : Санкт-Петербургский государственный лесотехнический университет имени С.М. Кирова, 2022. – С. 118–127.

9. Пилецкий, И. И. Графовые технологии в интеллектуальной системе комплексного анализа данных интернет-источников / И.И. Пилецкий, М.П. Батура, Л.Ю. Шилин // Доклады Белорусского государственного университета информатики и радиоэлектроники. – 2020. – Т. 18, № 5. – С. 89–97.

10. Булгаков, С. В. Геоинформационное моделирование и информационное взаимодействие / С. В. Булгаков // Конструкторское бюро. – 2018. – № 2. – С. 47–53.

11. Андреева, О. А. Геоинформационное массовое моделирование / О.А. Андреева, В.Я. Цветков, В.В. Ознамец // Информация и Космос. – 2020. – № 2. – С. 106–112.