



ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ В ПРОГНОЗЕ РЕЗУЛЬТАТА ПИРОЛИТИЧЕСКОГО АНАЛИЗА

ЛЕ ТХИ НЫТ СЫОНГ, А.В. БОНДАРЕВ, Л.И. БОНДАРЕВА, А.С. МОНАКОВА*, А.В. БАРШИН

ФГАОУ ВО «Российский государственный университет нефти и газа
(национальный исследовательский университет) имени И. М. Губкина»
65, Ленинский проспект, г. Москва 119991, Россия

АННОТАЦИЯ

Введение. Геохимические исследования органического вещества в нефтематеринских породах играют важную роль для оценки нефтегазонакопления на любой территории. Особенно важную роль эти исследования играют при прогнозе нетрадиционных ресурсов и запасов нефти и газа (т.н. сланцевые УВ). Пиролитические исследования по методу Rock-Eval для пород, насыщенных органическим веществом, рекомендовано проводить на образцах до и после экстракции их хлороформом. Однако экстракция — трудоемкий и длительный процесс, а нагрузка на лабораторное оборудование и время, необходимое для анализа, при этом удваивается.

Цель. Получить рабочую модель прогноза пиролитических параметров после экстракции образцов без проведения самой экстракции.

Материалы и методы. В этой работе алгоритмы регрессии машинного обучения применяются для прогнозирования одного из параметров пиролиза экстрагированных образцов на основе результатов пиролитического анализа экстрагированных и неэкстрагированных образцов. Для разработки модели прогнозирования были протестированы и сопоставлены 5 различных алгоритмов регрессии машинного обучения, включая множественную линейную регрессию, полиномиальную регрессию, опорную векторную регрессию, дерево решений и случайный лес.

Результаты. Результат прогнозирования демонстрирует, что взаимосвязь между параметрами до и после экстракции является сложной и нелинейной. Также была оценена производительность этих алгоритмов. Некоторые методы показали свою несовместимость с поставленными задачами, другие показали хорошие и удовлетворительные результаты. Аналогичные алгоритмы возможно применить для прогноза всех геохимических параметров образцов после экстракции.

Заключение. Наилучшим методом машинного обучения для данной задачи оказался метод случайного леса (Random forest).

Ключевые слова: машинное обучение, регрессия, Rock-Eval, пиролиз, кероген, нефтяные сланцы

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Финансирование: исследование не имело спонсорской поддержки.

Благодарности: авторы выражают благодарность и глубокую признательность, к.г.-м.н., ведущему научному сотруднику Сколковского института науки и технологий (Сколтех) Елене Владимировне Козловой за проведение пиролитического анализа и ценные замечания при работе над данной статьей.

Для цитирования: Ле Тхи Ныт Сыонг, Бондарев А.В., Бондарева Л.И., Монакова А.С., Баршин А.В. Применение алгоритмов машинного обучения в прогнозе результата пиролитического анализа. *Известия высших учебных заведений. Геология и разведка*. 2020;63(6):8—19. <https://doi.org/10.32454/0016-7762-2020-63-6-8-19>

Статья поступила в редакцию 01.04.2021
Принята к публикации 18.01.2022
Опубликована 21.02.2022

* Автор, ответственный за переписку

APPLICATION OF MACHINE LEARNING ALGORITHMS IN PREDICTING PYROLYTIC ANALYSIS RESULT

LE THI NHUT SUONG, ALEXSANDR V. BONDAREV, LIANA I. BONDAREVA, ALEKSANDRA S. MONAKOVA*,
ANDREY V. BARSHIN

National University of Oil and Gas "Gubkin University"
65 Leninskiy ave., Moscow 119991, Russia

ABSTRACT

Introduction. Geochemical studies of organic matter in oil source rocks play an important role in assessing oil and gas accumulation in any territory. These studies play a particularly important role in forecasting unconventional resources and oil and gas reserves (so-called shale hydrocarbons). It is recommended to carry out pyrolytic studies by the Rock-Eval method for rocks saturated with organic matter on samples before and after their extraction with chloroform. However, extraction is a laborious and time-consuming process, and the load on laboratory equipment and the time required for analysis is doubled.

Aim. To get a working model for predicting pyrolytic parameters of extracted samples, without carrying out extraction analysis.

Materials and methods. In this paper, machine learning regression algorithms are applied for predicting one of the pyrolysis parameters of extracted samples based on the pyrolytic analysis results of the extracted and non-extracted samples. To develop the prediction model, 5 different machine learning regression algorithms were applied and compared, including multiple linear regression, polynomial regression, support vector regression, decision tree, and random forest.

Results. The prediction result showcases that the relationship between the parameters before and after extraction is complex and non-linear. Some methods have shown their incompatibility with the assigned tasks, others have shown good and satisfactory results. Those algorithms can be applied to predict all geochemical parameters of extracted samples.

Conclusions. The best machine learning algorithm for this task is the Random forest.

Keywords: machine learning, regression, Rock-Eval, pyrolysis, kerogen, oil shale

Conflict of interest: the authors declare no conflict of interest.

Financial disclosure: no financial support was provided for this study.

Acknowledgments: the authors express their gratitude and deep gratitude to Elena Vladimirovna Kozlova, Ph.D., leading researcher at the Skolkovo Institute of Science and Technology (Skoltech), for the pyrolytic analysis and valuable comments while working on this article.

For citation: Le Thi Nhut Suong, Bondarev A.V., Bondareva L.I., Monakova A.S., Barshin A.V. *Proceedings of higher educational establishments. Geology and Exploration*. 2020;63(6):8—19. <https://doi.org/10.32454/0016-7762-2020-63-6-8-19>

Manuscript received 01 April 2021
Accepted 18 January 2022
Published 21 February 2022

* Corresponding author

Введение

Геохимические исследования ОВ в материнских породах (особенно экспресс-анализ — пироллиз Rock-Eval) на сегодня играют важную роль в прогнозе нефтегазоносности любой территории. Причем справедливо это как для относительно новых нефтегазоносных провинций и районов (с точки зрения оценки объемов генерации и историко-генетического метода прогноза), так и для уже изученных (для доразведки и прогноза нетрадиционных ресурсов УВ).

В отложениях, около $\frac{1}{3}$ матрицы которых занято органическим веществом, необходимо разделение генерирующего вещества и уже сгенерированных углеводородов. Хорошо показала себя методика разделения керогена и битумоида по сравнению результатов пиролитических исследований исходного порошка породы и экстрагированного органическим растворителем [1—3]. После экстракции мы получаем истинные параметры S_2 и T_{max} , водородного индекса, которые можно использовать как входные параметры для бассейнового моделирования. Экстракция проводится хлороформом в аппарате Сокслета до полного растворения битумоида (до 3 баллов свечения под люминесцентной лампой).

Экстракция представляет из себя трудоемкий и времязатратный процесс. Например, для пород баженовской свиты Западно-Сибирской нефтегазоносной провинции исчерпывающая экстракция продолжается месяц-полтора. Для этого нужно достаточное количество растворителя, кроме того, загруженность лабораторного оборудования возрастает в два раза. Прогноз параметров керогена после экстракции без проведения самой экстракции уменьшает затраты и время, необходимое для анализа. Описанная проблема является типичной для «машинного обучения» (“machine learning”). Раздел машинного обучения условно разделяется на задачи Классификации и Регрессии.

Задачей **классификации** является разбиение массива данных на блоки по схожести нескольких параметров (например, выделение пачек и подслоев в толще, в том числе продуктивных интервалов и/или интервалов повышенного содержания ОВ).

Задачей **регрессии** является числовой прогноз параметров, исходя из набора начальных данных. В рамках данной работы прогнозировался один, наиболее важный, по мнению авторов, параметр после экстракции — S_2 . Для прогноза нескольких параметров лучше

подойдут многослойные регрессионные модели нейронных сетей.

Исследования выполнены по более чем 300 образцам из трех скважин одного месторождения, расположенного в центральной части Западно-Сибирской провинции в пределах Фроловской впадины.

Данные представляют из себя результаты анализа Rock-Eval, одних и тех же образцов как до экстракции, так и после. Кроме того, имеются данные о молекулярном составе пород по разрезу скважин, однако использование этих данных не показало значительного увеличения точности моделей, более того, большинство методов показало худшие результаты сходимости. Авторы связывают это с несовпадением глубин взятия образцов на геохимический и молекулярный анализ. Иногда разница глубин достигала метра и более, что в условиях резкой анизотропии по вертикали приводит к отсутствию корреляции. А чувствительные к входному набору параметров регрессионные модели машинного обучения пересыщались маловажными входными параметрами, среди которых автоматическому алгоритму было трудно выявить весомые входные параметры (которых оказалось немного).

Входные параметры для обучения моделей представляют собой стандартный набор результатов анализа Rock-Eval, а именно:

- S_0 — количество свободных (газообразных) УВ (термическое испарение до 90 °C);
- S_1 — количество УВ — продуктов термического испарения до температуры 300 °C;
- S_2 — сумма УВ, 1) образованных в процессе крекинга керогена, также называется остаточным потенциалом — обозначает то количество УВ, которое может выделяться в процессе дальнейшего катагенеза исследуемых пород, и 2) продуктов термического испарения асфальтеновых компонентов.

Сумма пиков S_0 , S_1 , S_2 рассматривается как генетический потенциал породы;

- S_3 (S_3') — количество образованного в течение пироллиза углекислого газа CO_2 (CO), измеряется в $mgCO_2/g$ ($mgCO/g$) породы и дает представление о содержании кислорода в ОВ;
- S_4 (S_4') — количество образованного в процессе окисления углекислого газа CO_2 (CO);
- S_5 — количество образованного в процессе окисления углекислого газа CO_2 при высоких температурах, образованного из минерального углевода;
- T_{max} — температура, при которой образование УВ при термодеструкции (крекинге) керогена достигает максимального значения, °C.

По результатам измерений рассчитываются следующие параметры и коэффициенты:

- TOC — общее содержание органического углерода в породе (весовые %);
- HI (Hydrogen Index) — водородный индекс. Параметр, характеризующий генерационный потенциал ОБ;
- OI (Oxygen Index) — кислородный индекс;
- OSI (Oil Saturation Index) — индекс нефтенасыщенности;
- PI (Production Index) — индекс продуктивности.

Процесс анализа образцов происходит по следующей схеме (рис. 1.).

Для решения поставленных задач авторами статьи были опробованы следующие методы машинного обучения.

1. Multiple Linear Regression (Множественная линейная регрессия) [10, 11, 14, 20]

Множественная линейная регрессия (МЛР), также известная как множественная регрессия, — это статистический метод, который использует несколько независимых переменных для прогнозирования результирующей переменной. Цель такой регрессии — смоделировать линейную связь между независимыми переменными и ответной (зависимой) переменной.

По сути, множественная регрессия — это расширение обычной регрессии методом наименьших квадратов (МНК), которая включает более одной независимой переменной.

Формула множественной линейной регрессии:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip},$$

где y_i — зависимая переменная; x_i — объясняющая переменная; β_i — вес переменной (от англ. bias — смещение).

2. Polynomial Regression (Полиномиальная регрессия) [15, 21, 23]

Полиномиальная регрессия — это форма линейной регрессии, в которой связь между независимой переменной x и зависимой переменной y моделируется как полином n -й степени. Так же как и множественная линейная регрессия, полиномиальная регрессия определяет зависимость результирующей переменной от нескольких независимых переменных.

Формула полиномиальной регрессии:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \dots + \beta_n x_{in}^n,$$

где y_i — зависимая переменная; x_i — объясняющая переменная; β_i — вес переменной (от англ. bias — смещение).

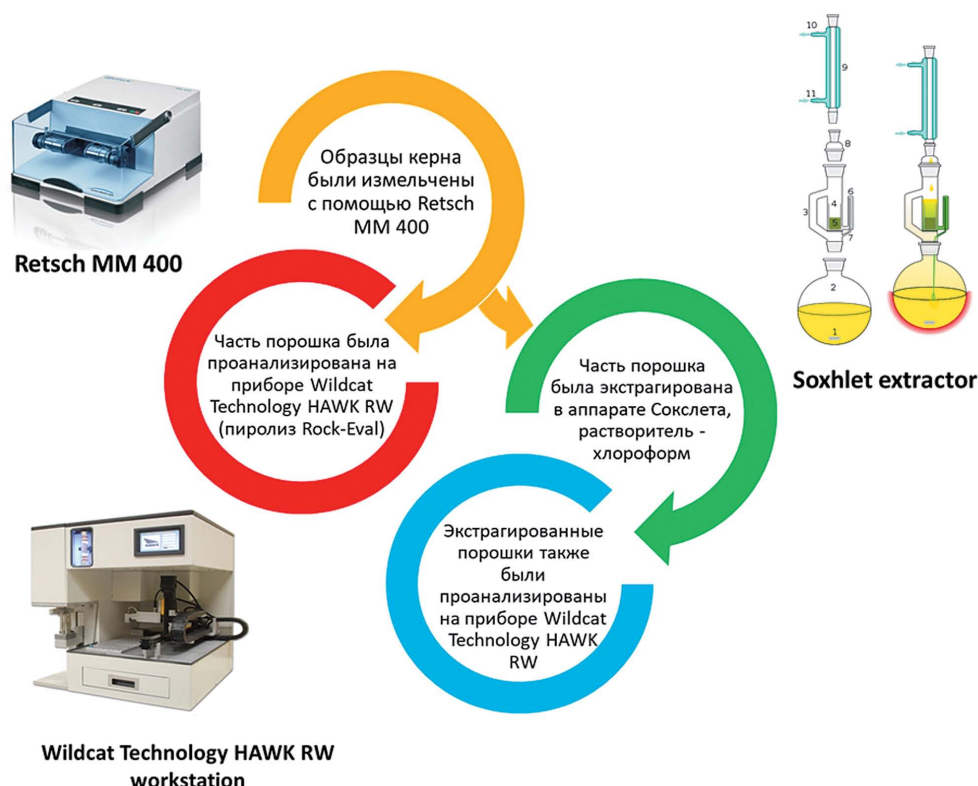


Рис. 1. Схема геохимического анализа образцов
 Fig. 1. Scheme of geochemical analysis of samples

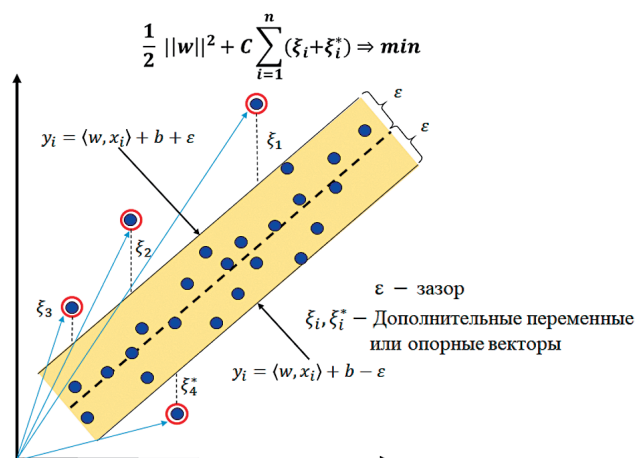


Рис. 2. Схема модели одномерной регрессии опорных векторов

Fig. 2. Schematic of a one-dimensional support vector regression model

3. Support vector regression (Опорная векторная регрессия) [7—9, 20]

Опорная векторная регрессия представляет собой метод анализа данных типа «обучение с учителем» (Supervised Learning), разработанный Вапником с коллегами в AT&T Bell Laboratories [8]. Суть метода заключается в том, что этот метод переводит исходные векторы в пространство более высокой размерности. И в этом пространстве он ищет разделяющую гиперплоскость с максимальным зазором. Чем больше разница

или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

4. Decision tree (Дерево решений) [5, 12, 13, 18]

Дерево решений — это модель машинного обучения с учителем, используемая для прогнозирования цели путем изучения правил принятия решений из их характеристик. Как следует из названия, мы можем думать об этой модели как о разбивке наших данных путем принятия решения, основанного на серии вопросов.

5. Random forest (Случайный лес) [4, 6, 16, 17, 19, 22, 24]

Случайный лес — это метод, который объединяет прогнозы из нескольких алгоритмов машинного обучения вместе, чтобы делать более точные прогнозы, чем любая отдельная модель. Алгоритм выдает множество прогнозов методами «дерева решений» и потом выводит среднее значение из всех. Забегая вперед, алгоритм случайного леса замечательно справляется со своей задачей, это — один из лучших алгоритмов по точности, но при этом он совершенно ничего не объясняет. То есть это такой черный ящик, который выдает хорошие прогнозы, но абсолютно не рассказывает о том, на чем основана на самом деле зависимость, без дополнительного скрупулезного анализа.

Как было указано выше, основной задачей экстракции (а значит, и машинного обучения) является истинное значение параметра S_2 , который характеризует количество продуктов крекинга

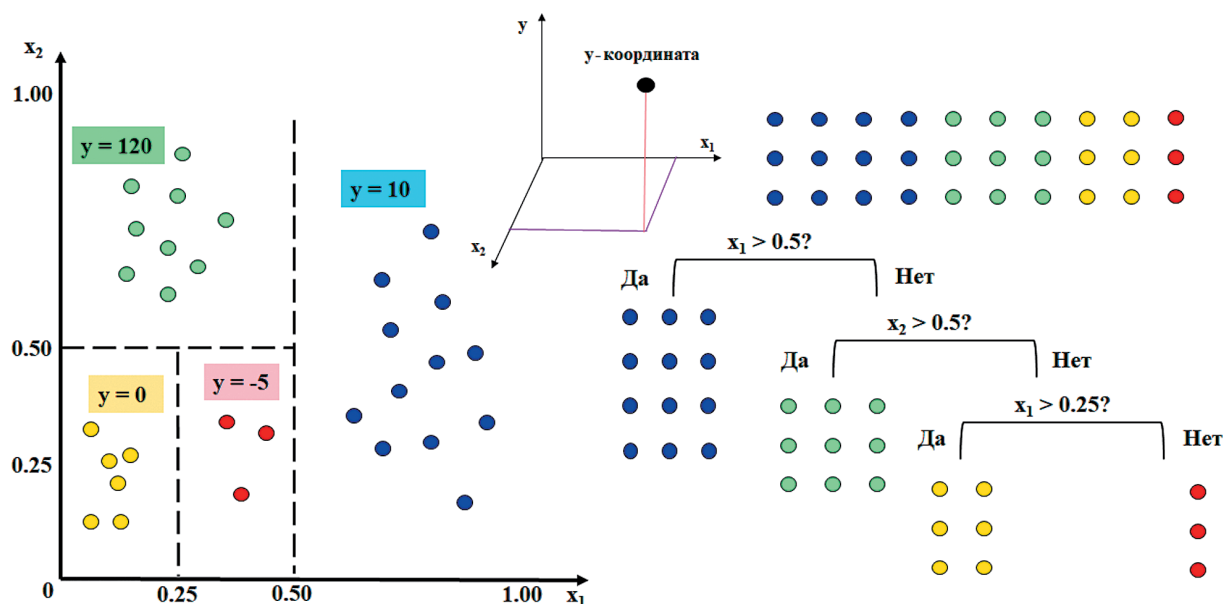


Рис. 3. Схема модели регрессии «дерево решений»

Fig. 3. Scheme of the regression model "decision tree"

керогена (нерастворимого органического вещества). Для создания перечисленных выше моделей использовалась скважина X2 (80% данных), а проверка моделей выполнялась как на скважинах X1, X3 так и по X2 (20% данных).

В итоге из описанных выше для создания и обучения модели были выбраны следующие параметры: $S_0, S_1, S_2, S_3, S_3', S_4, S_4', S_5, T_{\max}$. А параметры ТОС, НІ, ОІ, РІ, ОSІ, не использовались для обучения и проверки моделей регрессии как детерминированные от уже имеющихся параметров. Одним из условий использования перечисленных методов является независимость входных переменных.

Результаты прогноза каждого алгоритма оцениваются коэффициентом детерминации (R^2 — **R-квадрат**). Его рассматривают как универсальную меру зависимости одной случайной величины от множества других. Обычно он изменяется в пределах от 0 до 1. В случае если коэффициент детерминации отрицательный — это означает, что выбранная функция $y = f(i)$ описывает зависимость хуже, чем $y = y_{\text{ср}}$, т.е. использовать данную функцию (Модель регрессии) нельзя.

$$R^2 = \frac{RSS}{TSS},$$

где R^2 = коэффициент детерминации; RSS = сумма квадратов откл. от среднего (residual sum of squares); TSS = общая сумма квадратов (total sum of squares).

Результаты

Сводная таблица коэффициента корреляции результатов с проверочным блоком данных приведена ниже (табл. 1).

Для прогноза параметров S2ex прежде всего была применена самая простая регрессия — множественная линейная регрессия. Несмотря на то что это простейший метод, он хорошо подходит для работы с нашим набором данных. Сопоставление результатов расчета модели

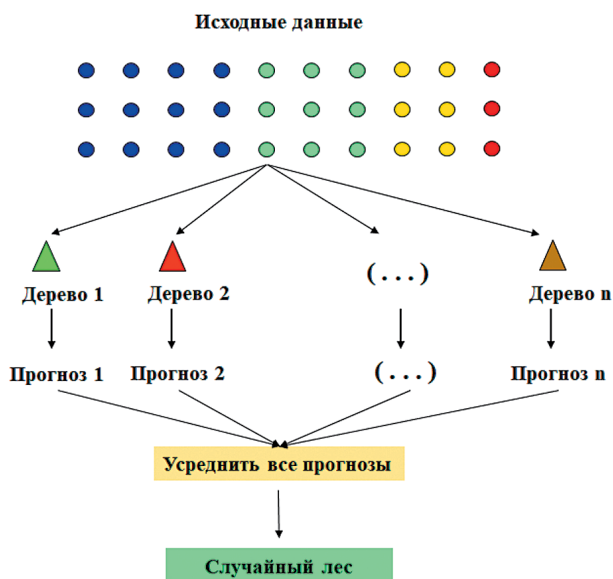


Рис. 4. Схема модели регрессии «случайного леса»
Fig. 4. Scheme of the “random forest” regression model

и фактически измеренных S2ex показано на рисунке 5.

Негативной стороной этого метода является то, что он характеризует данные как функцию, поэтому он может дать нам отрицательное значение параметра S2ex, чего быть не может. С новым набором данных, отличающихся от учебного набора, модель может дать нам прогноз, который сильно отличается от реальных значений. Примером может служить скважина X1, которая находится далеко от исходной скважины. В ней есть некоторые значения, которые сильно отличаются от фактических значений (от 75 до 115 мг разницы). И из-за нескольких этих точек оценка детерминации сильно снижается (R^2 скважины X1 равен -0.76). Благодаря этому мы также можем увидеть, насколько строго R^2 оценивает наши модели. Для улучшения модели нам нужно больше данных из разных регионов. Тем не менее этот метод

Таблица 1. Оценка результатов точности моделей регрессии
Table 1. Evaluation of the results of the accuracy of the regression models

Метод регрессии	Коэффициент детерминации R^2		
	X2 (обучающий)	Скважина X1	Скважина X3
Случайный лес	0,95	0,88	0,96
Опорная векторная регрессия	0,89	0,50	0,87
Дерево решений	0,90	0,77	0,91
Полиномиальная регрессия	-0,84	-15,54	0,03
Множественная линейная регрессия	0,96	-0,76	0,97

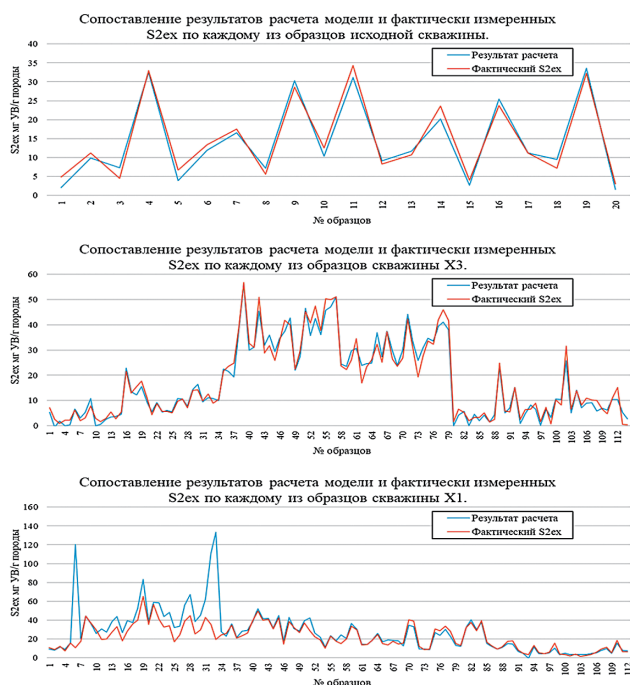


Рис. 5. Сопоставление результатов расчета модели и фактически измеренных S2ex методом множественной линейной регрессии

Fig. 5. Comparison of the results of the model calculation and actually measured S2ex by the method of multiple linear regression

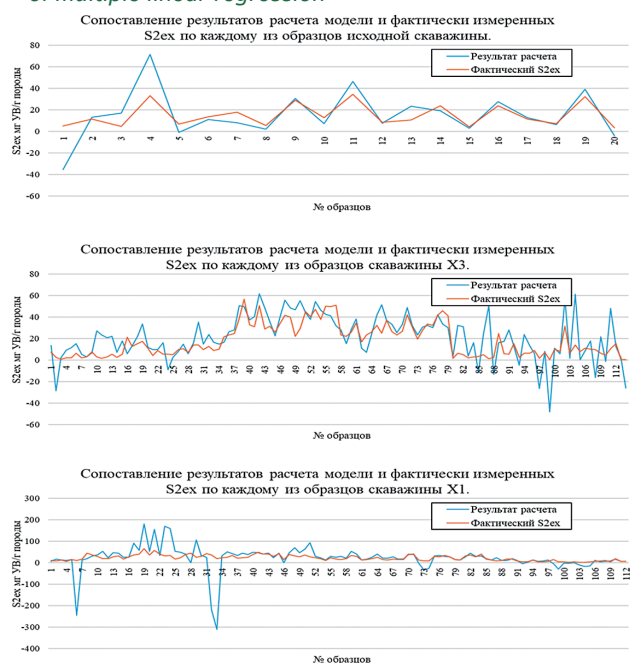


Рис. 6. Сопоставление результатов расчета модели и фактически измеренных S2ex методом полиномиальной регрессии

Fig. 6. Comparison of the results of the calculation of the model and actually measured S2ex by the method of polynomial regression

достойн прогнозирования значения S2ex соседних скважин, имеющих аналогичные характеристики разреза и фациальные условия.

Следующий алгоритм, который был применен, — полиномиальная регрессия. Этот метод дает нам худший результат прогноза даже в обучающей модели ($R^2 = -0,84$), и, очевидно, результат прогнозирования скважин X1, X3 не может быть лучше. R^2 , соответственно, составляет $-15,54$ и $0,03$. К модели применялись разные степени полинома “n”, но чем больше степень, тем худший результат прогноза мы получили. Результат предсказания полиномиальной регрессии ($n = 4$) показан на рисунке 6. Мы можем сделать вывод, что взаимосвязь между всеми пиролитическими параметрами до экстракции и S2ex не может быть смоделирована как полином n -й степени.

Опорная векторная регрессия также применяется для прогнозирования значения S2ex. Этот метод не дает самый лучший результат, но его результат является приемлемым. Результат представлен на рисунке 7. Прогнозируемые результаты обучающей скважины и скважины X3 достаточно хорошие, R^2 составляет $0,89$ и $0,87$ соответственно, но прогнозируемый результат скважины X1 недостаточно хорош ($R^2 = 0,5$). Этот алгоритм не дает

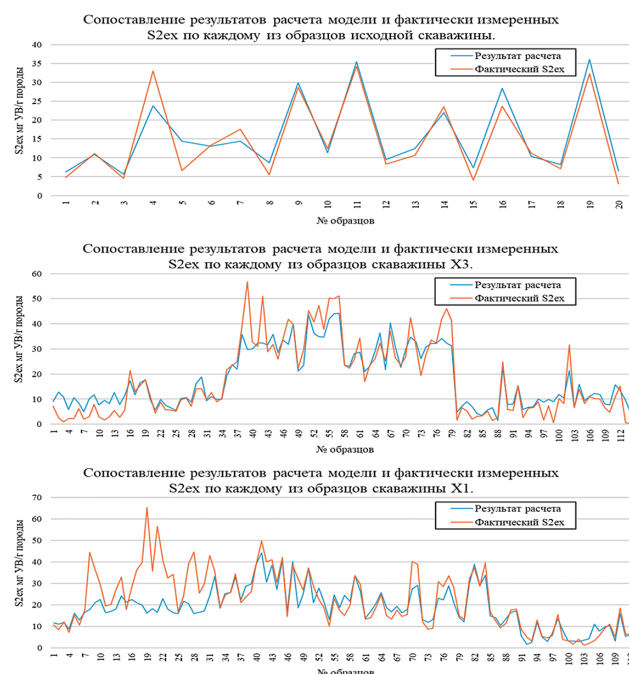


Рис. 7. Сопоставление результатов расчета модели и фактически измеренных S2ex методом опорной векторной регрессии

Fig. 7. Comparison of the model calculation results and actually measured S2ex by the support vector regression method

нам отрицательного значения S2ex, что говорит о его применимости и удовлетворительных результатах. Этот метод характеризует наш набор данных с помощью гиперплоскости или функции, а взаимосвязь между параметрами до и после экстракции является достаточно сложной, поэтому не может характеризоваться функцией.

Переходим к новым инструментам интеллектуального анализа данных и предсказательной аналитики: алгоритмам дерева решений и случайного леса.

Дерево решений организовано в виде иерархической структуры, состоящей из узлов принятия решений по оценке значений определенных переменных для прогнозирования результирующего значения. Использование деревьев классификации приводит к символическому обозначению класса. Геологический разрез состоит из разных групп пород, имеющих близкие характеристики, т.е. близкое значение S2ex, поэтому алгоритм дерева решений был создан для анализа наших данных. Как и предполагалось, этот метод дает очень хорошие результаты (рис. 8). Значение R^2 исходной скважины и скважин X1, X3, соответственно, 0,90, 0,77 и 0,91.

Алгоритм дерева решений дает нам хорошие прогнозы, но это не самый лучший способ. Метод, который дает нам лучший результат, — случайный лес (рис. 9). Этот метод объединяет прогнозы из нескольких алгоритмов машинного обучения вместе, чтобы делать более точные прогнозы, чем любая отдельная модель. Значение R^2 исходной скважины и скважин X1, X3 составляет 0,95, 0,88 и 0,96 соответственно, что является очень хорошим результатом для такого маленького набора данных.

Прогнозируемый результат S2ex всех моделей также сравнивался со значением S2, потому что в экстрагированном образце значение S2 должно быть меньше. Результат представлен на рисунке 10.

Как видно на рисунке 10, у методов множественной линейной, регрессии, опорной векторной регрессии и дерева решений есть прогнозные значения S2ex больше значения S2, что по определению не может быть верным. Метод случайного леса не только дает хорошие результаты, но и его прогнозы всегда меньше значения S2. Исходя из всех проверок, можем сделать вывод, что алгоритм

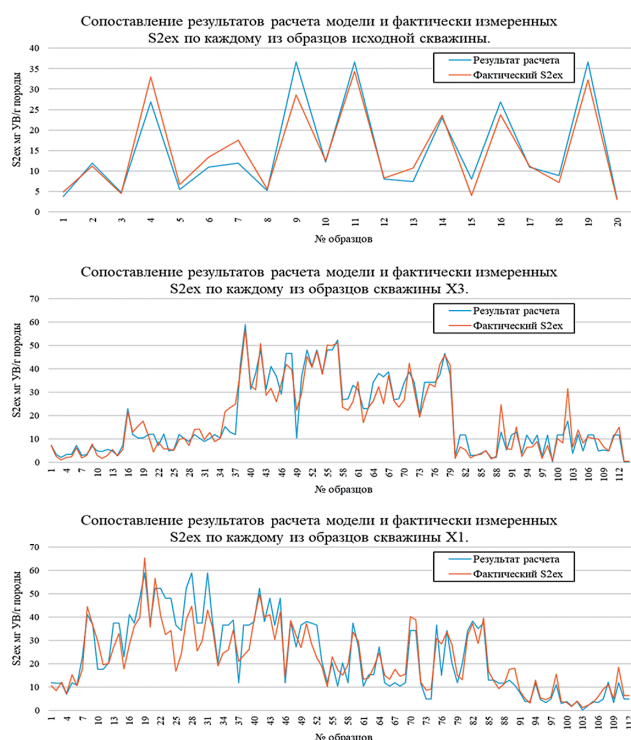


Рис. 8. Сопоставление результатов расчета модели и фактически измеренных S2ex методом дерева решений

Fig. 8. Comparison of the model calculation results and actually measured S2ex by the decision tree method

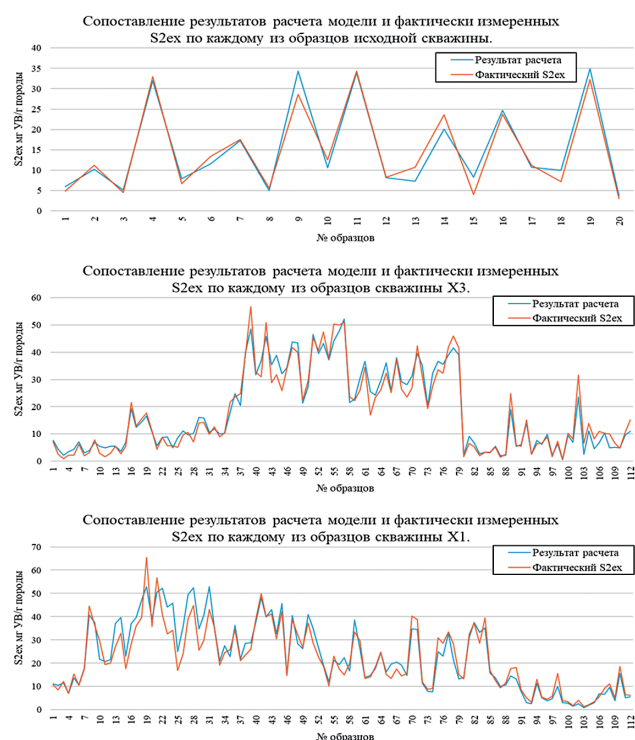


Рис. 9. Сопоставление результатов расчета модели и фактически измеренных S2ex методом случайного леса

Fig. 9. Comparison of the model calculation results and actually measured S2ex by the random forest method

Множественная линейная регрессия



Опорная векторная регрессия



Дерево решений



Случайный лес



Рис. 10. Сопоставление результатов расчета моделей (А — множественная линейная регрессия; Б — опорная векторная регрессия; В — дерево решений; Г — случайный лес) и значения S2 по каждому из образцов скважины X3

Fig. 10. Comparison of the results of calculating models (A — multiple linear, regression; Б — support vector regression; В — decision tree; Г — random forest) and S2 values for each of the samples of well X3

случайного леса — самый хороший метод прогноза для наших данных.

Заключение

Несмотря на то что количество данных для обучения модели ограничено, результат превзошел все ожидания. Следует ожидать увеличения точности по мере расширения блока входных данных для обучения модели. Наилучший результат показали методы случайного леса, опорной векторной регрессии и дерева решений. Алгоритмы случайного леса и дерева решений рекомендуются для прогнозирования результата анализа любых скважин, а множественная линейная регрессия — для всех скважин, близких к исходной скважине.

Однако следует понимать, что близость расположения скважин обеспечивает фаціальную однородность пород, и в случае увеличения площади исследования возможно уменьшение точности.

Результат прогнозирования также демонстрирует, что взаимосвязь между параметрами до и после экстракции является сложной и нелинейной. Доказательство тому — значение R^2 множественной линейной регрессии, а полиномиальная регрессия отрицательна, что означает, что модель не работает. Однако случайный лес и дерево решений дают нам хорошую производительность. С помощью тех же алгоритмов мы можем применить для прогноза все геохимические параметры по глубине или использовать их для данных ГИС.

ЛИТЕРАТУРА

1. Бондарев А.В., Данцова К.И., Баршин А.В., Минлигалиева Л.И. Моделирование катагенетического преобразования силурийских нефтегазоматеринских толщ Южного Предуралья на основе статистической обработки результатов анализа Rock-eval // Труды Российского государственного университета нефти и газа им. И.М. Губкина. 2020. № 1 (298). С. 29—37.
2. Ле Тхи Ныт Сыонг, Бондарев А.В., Баршин А.В., Бондарева Л.И. Применение методов машинного обучения для прогноза геохимических параметров анализа Rock-EVAL после экстракции хлороформом // Сборник докладов Региональной научно-технической конференции «Губкинский университет в решении вопросов нефтегазовой отрасли России», 22—23 октября 2020 г. — М.:

- Издательский центр РГУ нефти и газа (НИУ) имени И.М. Губкина, 2020 — 484 с. ISBN 978-5-91961-332-9
- Монакова А.С., Осипов А.В., Бондарев А.В., Минлигалиева Л.И. Геохимическая характеристика нефтематеринских пород силурийского возраста южного сегмента Предуральяского прогиба (г. Кувандык) // В сб.: Новые идеи в науках о Земле. Материалы XIV Международной научно-практической конференции: в 7 т. 2019. С. 69—70.
 - Arlot, Sylvain and Robin Genuer. "Analysis of purely random forests bias." ArXiv abs/1407.3939 (2014)
 - Ben-Haim Y. Information-gap theory: decisions under severe uncertainty. Academic Press, London, 2001.
 - Biau G. Analysis of a random forests model. Journal of Machine Learning Research. 2012. Vol. 13. P. 1063—1095.
 - Chih-Chung Chang, Chih-Jen Lin. LIBSVM: a library for support vector machines, 2001.
 - Cortes C., Vapnik V. Support-Vector Networks // MachineLearning. Vol. 20. P. 273—297.
 - Lane D.M. Online Statistics Education: A Multimedia Course of Study (<http://onlinestatbook.com/>). Rice University.
 - Draper N.R., Smith H. Applied Regression Analysis. Wiley-Interscience. 1998.
 - Glantz Stanton A., Slinker B.K. Primer of Applied Regression and Analysis of Variance. McGraw-Hill, 1990.
 - Kamiński B., Jakubczyk M., Szufel P. A framework for sensitivity analysis of decision trees // Central European Journal of Operations Research, 2017. P. 135—159. <https://doi.org/10.1007/s10100-017-0479-6>. PMC 5767274. PMID 29375266
 - Karimi K., Hamilton H.J. Generation and Interpretation of Temporal Decision Rules // International Journal of Computer Information Systems and Industrial Management Applications, 2011. Vol. 3.
 - Rouaud M. Probability, Statistics and Estimation Propagation of Uncertainties in Experimental Measurement. 2013.
 - Nefedova A.S., Osipov A.V., Ermolkin V.I. The Source Rock Generation Potential of Lower Permian Artinskian Age in Southern Part of Pre-Ural Fore Deep // In: Geomodel 2016 — 18th Science and Applied Research Conference on Oil and Gas Geological Exploration and Development. 2016.
 - Ho Tin Kam. Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14—16 August 1995. P. 278—282.
 - Liaw A. and Wiener M. Classification and regression by random forest. Researchgate, 2002. Vol. 2. P. 18—22.
 - Quinlan J.R. Simplifying decision trees // International Journal of Man-Machine Studies. 1987. P. 221—234. [https://doi.org/10.1016/S0020-7373\(87\)80053-6](https://doi.org/10.1016/S0020-7373(87)80053-6)
 - Scornet E. Random forests and kernel methods. 2015.
 - Smola A., Schoelkopf B. A tutorial on support vector regression: Tech. Rep. 1998.
 - Tissot B.P., Welt, D.H. Petroleum formation and occurrence. Springer, Berlin, 1984.
 - Yan Xin. Linear Regression Analysis: Theory and Computing, World Scientific. 2009. P. 1—2.
 - Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, Chih-Jen Lin, Training and testing low-degree polynomial data mappings via linear SVM // Journal of Machine Learning Research. P. 1471—1490.
 - Zhu R., Zeng D., Kosorok M.R. Reinforcement Learning Trees // Journal of the American Statistical Association. 2015. P. 1770—1784. <https://doi.org/10.1080/01621459.2015.1036994>
 - Bondarev Alexander V., Dantsova Kristina I., Barshin Andrey V., Minligalieva Liana I. Modeling maturity of organic matter in source rocks of silurian oil and gas source strata of southern urals based on statistical processing of rock-eval results // Proceedings of Gubkin Russian State University of oil and gas. 2020. № 1 (298). pp. 29—37.
 - Le Thi Nhut Suong, Bondarev A., Barshin A., Bondareva L. Application of machine learning methods for prediction of geochemical parameters of Rock-Eval analysis after chloroform extraction // Collection of reports of the IV regional scientific and technical conference "Gubkin University in solving issues of the oil and gas industry of Russia", dedicated to the 90th anniversary of Gubkin University and the Faculty of Economics and Management. 2020. 484 p. ISBN 978-5-91961-332-9.
 - Monakova A., Osipov A., Bondarev A., Minligalieva L. Geochemical characteristics of oil-bearing rocks of the Silurian age of the southern segment of the Pre-Ural foredeep (Kuvandyk) // New ideas in Earth sciences. Materials of the XIV International Scientific and Practical Conference. 2019. Vol. 7. pp. 69—70.
 - Arlot, Sylvain and Robin Genuer. "Analysis of purely random forests bias." ArXiv abs/1407.3939 (2014)
 - Ben-Haim Y. Information-gap theory: decisions under severe uncertainty. Academic Press, London, 2001.
 - Biau G. Analysis of a random forests model. Journal of Machine Learning Research. 2012. Vol. 13. P. 1063—1095.
 - Chih-Chung Chang, Chih-Jen Lin. LIBSVM: a library for support vector machines, 2001.
 - Cortes C., Vapnik V. Support-Vector Networks // MachineLearning. Vol. 20. P. 273—297.
 - Lane D.M. Online Statistics Education: A Multimedia Course of Study (<http://onlinestatbook.com/>). Rice University.
 - Draper N.R., Smith H. Applied Regression Analysis. Wiley-Interscience. 1998.

REFERENCES

ГЕОЛОГИЯ И РАЗВЕДКА МЕСТОРОЖДЕНИЙ УГЛЕВОДОРОДОВ / GEOLOGY AND PROSPECTING FOR HYDROCARBON RESERVES

11. Glantz Stanton A., Slinker B.K. Primer of Applied Regression and Analysis of Variance. McGraw-Hill, 1990.
12. Kamiński B., Jakubczyk M., Szufel P. A framework for sensitivity analysis of decision trees // Central European Journal of Operations Research, 2017. P. 135—159. <https://doi.org/10.1007/s10100-017-0479-6>. PMC 5767274. PMID 29375266
13. Karimi K., Hamilton H.J. Generation and Interpretation of Temporal Decision Rules // International Journal of Computer Information Systems and Industrial Management Applications, 2011. Vol. 3.
14. Rouaud M. Probability, Statistics and Estimation Propagation of Uncertainties in Experimental Measurement. 2013.
15. Nefedova A.S., Osipov A.V., Ermolkin V.I. The Source Rock Generation Potential of Lower Permian Artinskian Age in Southern Part of Pre-Ural Fore Deep // In: Geomodel 2016 — 18th Science and Applied Research Conference on Oil and Gas Geological Exploration and Development. 2016.
16. Ho Tin Kam. Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14—16 August 1995. P. 278—282.
17. Liaw A. and Wiener M. Classification and regression by random forest. Researchgate 2002. Vol. 2. P. 18—22.
18. Quinlan J.R. Simplifying decision trees // International Journal of Man-Machine Studies. 1987. P. 221—234. [https://doi.org/10.1016/S0020-7373\(87\)80053-6](https://doi.org/10.1016/S0020-7373(87)80053-6)
19. Scornet E. Random forests and kernel methods. 2015.
20. Smola A., Schoelkopf B. A tutorial on support vector regression: Tech. Rep. 1998.
21. Tissot B.P., Welt, D.H. Petroleum formation and occurrence. Springer, Berlin, 1984.
22. Yan Xin. Linear Regression Analysis: Theory and Computing, World Scientific. 2009. P. 1—2.
23. Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, Chih-Jen Lin, Training and testing low-degree polynomial data mappings via linear SVM // Journal of Machine Learning Research. P. 1471—1490.
24. Zhu R., Zeng D., Kosorok M.R. Reinforcement Learning Trees // Journal of the American Statistical Association. 2015. P. 1770—1784. <https://doi.org/10.1080/01621459.2015.1036994>

ВКЛАД АВТОРОВ / AUTHOR CONTRIBUTIONS

Ле Тхи Нхыт Сьонг — разработала методику исследования, изложенную в статье, подготовила текст статьи для публикации.

Бондарев А.В. — выступил научным руководителем проведения исследования, изложенного в статье.

Бондарева Л.И. — оказала помощь при подготовке научной статьи.

Монакова А.С. — оказывала всестороннюю помощь при проведении научного исследования, корректировала текст статьи, окончательно утвердила публикуемую версию статьи.

Баршин А.В. — корректировал графику, приведенную в статье.

Le Thi Nhut Suong — developed the research methodology outlined in the article, prepared the text of the article for publication.

Aleksandr V. Bondarev — acted as the scientific supervisor of the research described in the article.

Liana I. Bondareva — assisted in the preparation of a scientific article.

Aleksandra S. Monakova — provided comprehensive assistance in conducting scientific research, corrected the text of the article, finally approved the published version of the article.

Andrey V. Barshin — corrected the schedule given in the article.

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Ле Тхи Нхыт Сьонг — студентка ФГАОУ ВО «Российский государственный университет нефти и газа (национальный исследовательский университет) имени И. М. Губкина»

65, Ленинский проспект, г. Москва 119991, Россия

e-mail: lethinhutsuong@gmail.com

тел.: +7 (977) 586-98-45

ORCID: <https://orcid.org/0000-0002-5088-0419>

Le Thi Nhut Suong — student, National University of Oil and Gas “Gubkin University”.

65 Leninskiy ave., Moscow 119991, Russia.

e-mail: lethinhutsuong@gmail.com

tel.: +7 (977) 586-98-45

ORCID: <https://orcid.org/0000-0002-5088-0419>

Бондарев Александр Владимирович — кандидат геолого-минералогических наук, доцент кафедры поисков и разведки нефти и газа ФГАОУ ВО «Российский государственный университет нефти и газа (национальный исследовательский университет) имени И. М. Губкина»
65, Ленинский проспект, г. Москва 119991, Россия
e-mail: jcomtess@yandex.ru
тел.: +7 (499) 507-88-88
ORCID: <https://orcid.org/0000-0001-8221-1052>
Scopus ID: 56308173600
SPIN-код: 6559-1469

Бондарева Лиана Ильясовна — старший преподаватель кафедры поисков и разведки нефти и газа ФГАОУ ВО «Российский государственный университет нефти и газа (национальный исследовательский университет) имени И. М. Губкина»
65, Ленинский проспект, г. Москва 119991, Россия
e-mail: liana_abril@mail.ru
тел.: +7 (499) 507-84-32
ORCID: <https://orcid.org/0000-0003-3986-858X>
Scopus ID: 57209737387
SPIN-код: 1584-1518

Монакова Александра Сергеевна* — кандидат геолого-минералогических наук, доцент кафедры поисков и разведки нефти и газа ФГАОУ ВО «Российский государственный университет нефти и газа (национальный исследовательский университет) имени И. М. Губкина»
65, Ленинский проспект, г. Москва 119991, Россия
тел.: +7 (916) 849-57-04
ORCID: 0000-0003-3854-4436
Scopus ID: 8574084700
SPIN-код: 5619-7973

Баршин Андрей Витальевич — преподаватель кафедры поисков и разведки нефти и газа ФГАОУ ВО «Российский государственный университет нефти и газа (национальный исследовательский университет) имени И. М. Губкина»
65, Ленинский проспект, г. Москва 119991, Россия
e-mail: barshinsp@gmail.com
тел.: +7 (915) 127-32-11
ORCID: <https://orcid.org/0000-0002-5207-1121>
Scopus ID: 57221607978
SPIN-код: 3618-5049

Alexsandr V. Bondarev — Cand. of Sci. (Geol.-Min.), Assoc. Prof., National University of Oil and Gas “Gubkin University”.
65 Leninskiy ave., Moscow 119991, Russia
e-mail: jcomtess@yandex.ru
tel.: +7 (499) 507-88-88
ORCID: <https://orcid.org/0000-0001-8221-1052>
Scopus ID: 56308173600
SPIN-code: 6559-1469

Liana I. Bondareva — senior lecturer, National University of Oil and Gas “Gubkin University”.
65 Leninskiy ave., Moscow 119991, Russia
e-mail: liana_abril@mail.ru
tel.: +7 (499) 507-84-32
ORCID: <https://orcid.org/0000-0003-3986-858X>
Scopus ID: 57209737387
SPIN-code: 1584-1518

Aleksandra S. Monakova* — Cand. of Sci. (Geol.-Min.), Assoc. Prof., National University of Oil and Gas “Gubkin University”.
65 Leninskiy ave., Moscow 119991, Russia
tel.: +7 (916) 849-57-04
ORCID: 0000-0003-3854-4436
Scopus ID: 8574084700
SPIN-code: 5619-7973

Andrey V. Barshin — lecturer, National University of Oil and Gas “Gubkin University”.
65 Leninskiy ave., Moscow 119991, Russia
e-mail: barshinsp@gmail.com
tel.: +7 (915) 127-32-11
ORCID: <https://orcid.org/0000-0002-5207-1121>
Scopus ID: 57221607978
SPIN-code: 3618-5049

* Автор, ответственный за переписку / Corresponding author