

# Выявление особенностей органического вещества нефтей и нефтегазоматеринских толщ путём сопоставления результатов геохимического анализа со статистическим анализом, основанным на методах машинного обучения (на примере одного из месторождений Западно-Сибирского нефтегазоносного бассейна)

К.О. Осипов\*, Э.А. Абля, Р.С. Сауткин, М.А. Большакова, А.А. Сулова, А.П. Антонов  
Московский государственный университет имени М.В. Ломоносова, Москва, Россия

Идеей данной работы является сравнение результатов двух видов анализа – геохимического и статистического – при изучении органического вещества экстрактов нефтегазоматеринских толщ (НГМТ) и образцов нефтей. Объектом исследования послужили НГМТ тутлеймской и тюменской свит и нефти викуловской свиты и юрско-доюрских отложений в западной части Западно-Сибирского бассейна. Среди методов статистического анализа НГМТ использовались метод главных компонент и метод случайного леса, а в качестве средств визуализации – тепловая карта корреляций. Метод главных компонент помог обнаружить явное различие органического вещества тутлеймской и тюменской НГМТ, а привлечение метода случайного леса и тепловой карты корреляций позволило не только выявить геохимические параметры, по которым наблюдаются наибольшие различия толщ, но и расшифровать геологические факторы, из-за которых это различие стало явным. Так, тюменская НГМТ имеет большую зрелость и относительно больший вклад наземной растительности относительно тутлеймской НГМТ. Тот же набор методов вместе с методом *k*-ближайших соседей применен для корреляции нефть-НГМТ. Нефти викуловской свиты по типу керогена, очевидно, происходят из тутлеймской НГМТ. Нефти юрских и доюрских пластов близки к тутлеймской НГМТ, хотя схожесть с тюменской НГМТ может достигать 30% в отдельных образцах. Нефти викуловских пластов менее зрелые, чем нефти из юрских и доюрских пород. Это может свидетельствовать о продолжающейся подпитке углеводородами юрских и доюрских отложений и о прекращении миграции в прошлом в вышележащие апт-альбские отложения (викуловская свита). Геохимический анализ также позволил выявить свойства НГМТ, отвечающие за зрелость и тип органического вещества. Он подтвердил выводы, сделанные на основе применения методов статистического анализа.

**Ключевые слова:** сопоставление исследований, геохимический анализ, методы статистического анализа, нефтегазоматеринская толща, корреляция нефть-нефтегазоматеринская порода

**Для цитирования:** Осипов К.О., Абля Э.А., Сауткин Р.С., Большакова М.А., Сулова А.А., Антонов А.П. (2022). Выявление особенностей органического вещества нефтей и нефтегазоматеринских толщ путём сопоставления результатов геохимического анализа со статистическим анализом, основанным на методах машинного обучения (на примере одного из месторождений Западно-Сибирского нефтегазоносного бассейна). *Георесурсы*, 24(2), с. 217–229. DOI: <https://doi.org/10.18599/grs.2022.2.20>

## Введение

Высокие темпы роста объема цифровых данных являются нормой современного мира. Эта тенденция не обходит стороной геологию, в частности, геохимию горючих ископаемых. Прежний во многом рутинный подход к анализу геохимических данных требует если не пересмотра, то альтернативы в виде оперативного, экспрессного анализа, позволяющего работать с большими данными, опираясь на современные научные знания в геохимии органического вещества (ОВ). Речь идёт о статистическом анализе, который обладает богатым выбором методов, каждый из которых по-своему уникален и имеет определенную область применимости.

Целью данной работы является сравнение методов статистического анализа с геохимическим анализом для выявления различий двух нефтегазоматеринских толщ, (НГМТ) и для корреляции нефть-НГМТ.

Хотя почти все из задействуемых в работе методов статистического анализа по отдельности давно используются в геохимии органического вещества (Peters и др., 2005), но их комплексное применение для многогранного изучения ОВ является нетривиальным и новым.

Данное исследование является частью процесса изучения органического вещества и углеводородов (УВ), начатого Н.Б. Вассоевичем и В.А. Успенским – основателями осадочно-миграционной теории происхождения нефти. На этапе создания и развития теории Николай Брониславович призывал к тщательной работе с результатами исследования вещества и уже тогда использовал элементы математической статистики, например, коэффициент ранговой корреляции Спирмена для установления связи УВ с НГМТ.

\* Ответственный автор: Константин Олегович Осипов  
e-mail: [k.osipov@oilmsu.ru](mailto:k.osipov@oilmsu.ru)

© 2022 Коллектив авторов

Контент доступен под лицензией Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>)

### Материалы и методы

Объектом исследования послужили НГМТ тутлеймской и тюменской свит и нефти викуловской свиты и юрско-доюрских отложений в западной части Западно-Сибирского бассейна (рис. 1). Исходный материал состоял из 50 экстрактов из тутлеймской и 40 из тюменской НГМТ, приблизительно 20 и 30 образцов нефтей из викуловской свиты и юрско-доюрских отложений соответственно. Для анализа НГМТ использовалось около 50 свойств и отношений свойств, а для проведения корреляции нефть-НГМТ – 35.

Геохимический анализ органического вещества нефтей и экстрактов НГМТ заключается в интерпретации величин геохимических параметров и их отношений по графикам, треугольным и звёздчатым диаграммам.

Для методов статистического анализа первым этапом работы с данными является их предварительная подготовка, которая зависит от используемых в работе методов (рис. 2). Первоначально данные хранятся в виде таблиц. В табличных данных могут встречаться выбросы, то есть anomalно низкие или anomalно высокие значения признаков. Используемые в работе методы линейной регрессии (для заполнения пустых ячеек), главных компонент чувствительны к таким значениям, и такие ячейки в таблице необходимо отсеивать.

В статье использовался ручной способ обнаружения аномалий по всем возможным парным графикам, хотя существуют автоматические способы обнаружения выбросов, как, например, поиск аномалий по одному

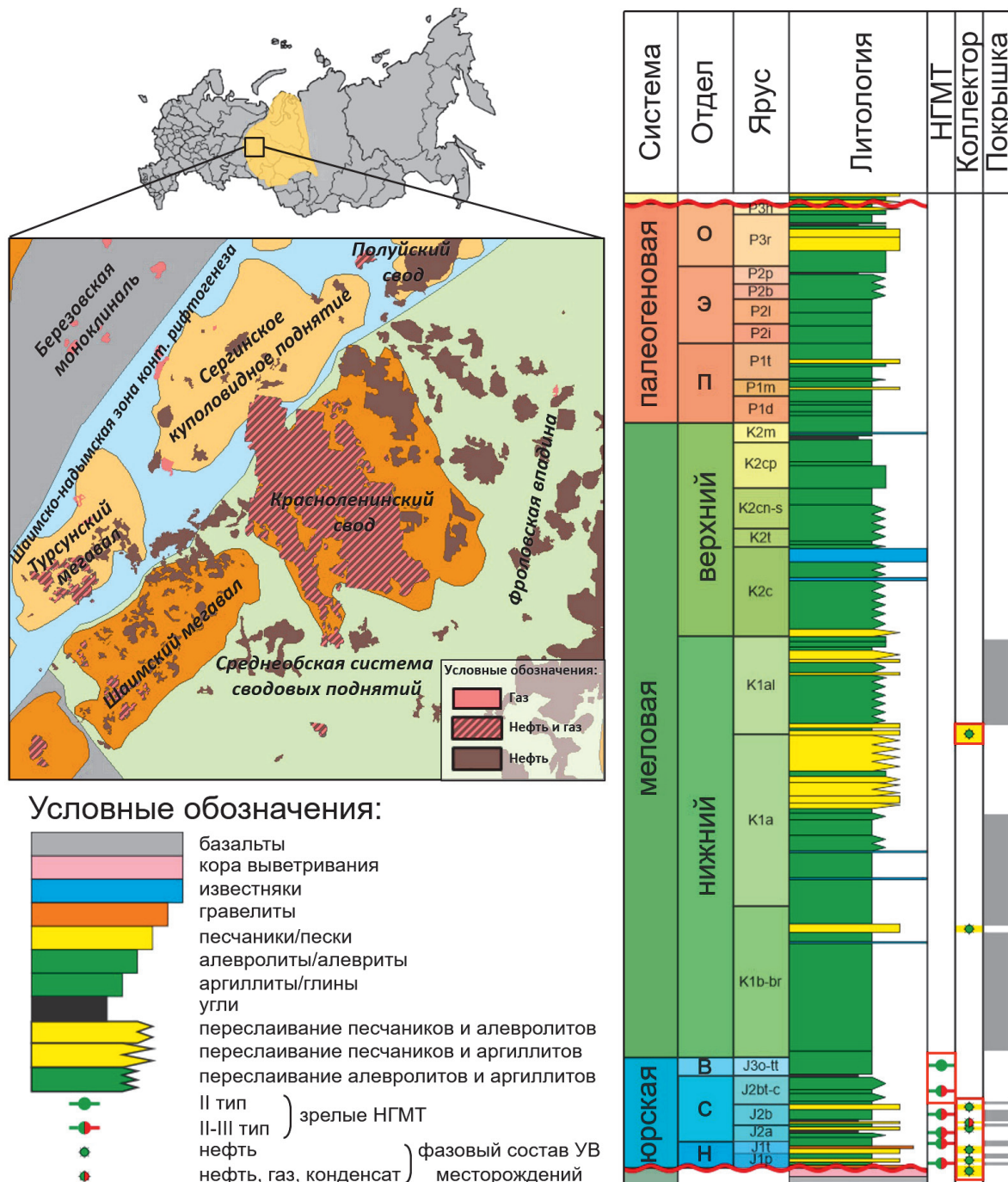


Рис. 1. Расположение района исследования (слева) на фрагменте тектонической карты Западно-Сибирского нефтегазоносного бассейна (Ступакова, 2011) и литолого-стратиграфическая колонка (справа) изучаемой территории (стилизованна на основе Атласа «Геология и нефтегазоносность...», 2004) с указанием в виде красных прямоугольников изучаемых НГМТ и нефтей в пластах коллекторов

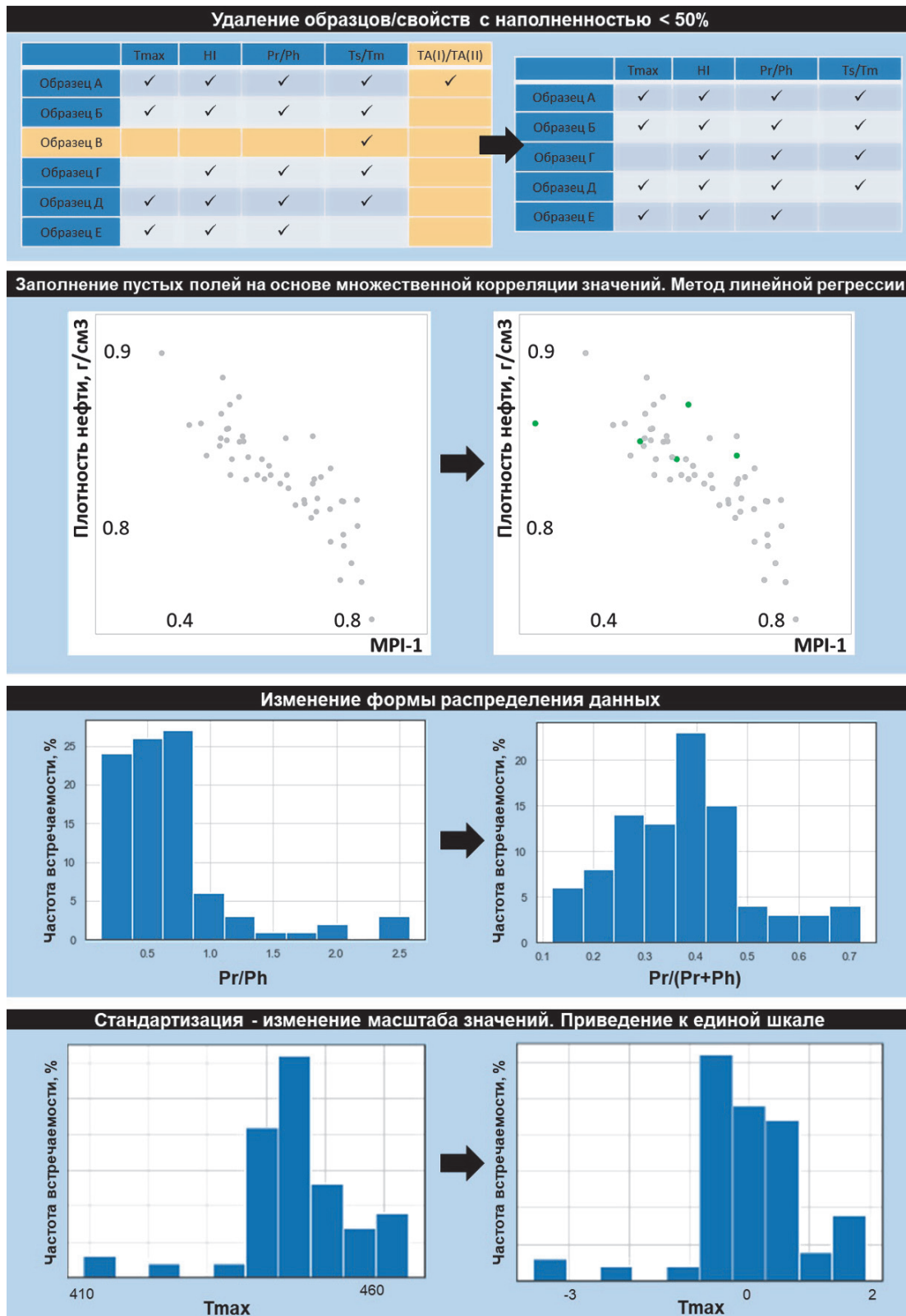


Рис. 2. Предварительная подготовка данных для применения метода главных компонент

признаку на основе квартилей или по нескольким признакам методом изолированного леса.

В табличных данных могут присутствовать пустые ячейки. Используемый в работе метод главных компонент не работает с данными, в которых присутствуют пропущенные значения. Образцы и геохимические параметры с высокой долей (обычно более 50 %) пустых полей являются непредставительными, и заполнение данных в них может привести к ухудшению качества работы

метода главных компонент. Такие образцы и свойства отсеиваются.

В оставшейся таблице всё еще могут присутствовать пустые поля. Способы их заполнения можно разделить на 2 группы – не учитывающие взаимосвязь признаков и учитывающие. Применение алгоритмов первой группы является нежелательным, поскольку они ухудшают корреляцию взаимосвязанных признаков. Примером алгоритмов из этой группы является заполнение медианным

значением признака. В работе применен метод из второй группы – метод линейной регрессии. В его основе лежит знание о взаимосвязи (множественной корреляции) признаков. То есть заполнение пустого поля производится согласно множествам трендов (линий корреляций или регрессий). При этом влияние остальных признаков (их вес) на заполняемое значение тем больше, чем выше коэффициент корреляции с признаком, в котором заполняется поле.

Метод главных компонент также чувствителен к масштабу признаков. Чем больше по модулю значения принимает признак, тем большую важность метод главных компонент придает ему. Если бы мы подали на вход таблицу без дальнейшей обработки, то метод главных компонент обращал бы большее внимание на HI (водородный индекс) со значениями от 50 до 700 мг уВ/г Сорг, и ничтожное влияние на его работу оказывали бы признаки с небольшими по модулю значениями, например,  $Pt/(Pt+Ph)$  (пристан/(пристан+фитан)) – от 0 до 1. Чтобы уравновесить влияние признаков их приводят к единому масштабу. Такая операция называется стандартизацией. Она удобна тем, что позволяет уравнивать между собой признаки, имеющие разные единицы измерения. Смысл алгоритма заключается в замене шкалы значений свойства, что иллюстрируется на рис. 2. Так, среднее значение  $T_{max}$  на новой шкале – 0, а величина стандартного отклонения – 1. Стандартизация предельно выполняется для каждого свойства.

Более сложным для понимания является необходимость изменения гистограммы распределения геохимических параметров. На рисунке 2 слева значения  $Pt/Ph$  концентрируются в узком диапазоне значений от 0 до 1, а при больших величинах редки. Это связано с самим отношением  $Pt/Ph$ : при  $Pt < Ph$ , значения укладываются в диапазон от 0 до 1, а при  $Pt > Ph$  – растягиваются в диапазоне от 1 до бесконечности. При различии  $Pt$  от  $Ph$  в 10 раз мы бы получили значения 0.1 и 10. Но 10 намного дальше от основной массы точек, чем 0.1. Перед использованием метода главных компонент нам необходимо уравнивать 0.1 и 10 относительно основной массы точек, иначе это явным образом ухудшит работу метода, так как образец с  $Pt/Ph = 10$  чрезвычайно выделяется относительно других, а образец с  $Pt/Ph = 0.1$  – нет. Для этого применен подход, когда в знаменатель помещается сумма делимого и делителя, то есть  $Pt/Ph$  преобразуется в  $Pt/(Pt+Ph)$ . Операция предельно выполняется для всех подобных отношений свойств вроде  $Pt/C17$ ,  $H29Ts/H29$ ,  $T24/H30$  и т.д.

Приведем ещё один пример: сравним размеры Гренландии и Африки в цилиндрической проекции Меркатора и в реальности (рис. 3). В проекции Меркатора они сопоставимы, а в действительности Гренландия в 14 раз меньше Африки. Как картографы выбирают для отображения объектов подходящую проекцию, минимизирующую искажения (в данном случае размер), так и статистики преобразуют гистограммы распределения свойств для более корректного сравнения их величин.

Перейдём к описанию используемых методов статистического анализа – метода главных компонент, случайного леса, k-ближайших соседей (рис. 4).

Метод главных компонент применяется, когда данных много, и существует потребность в упрощении их анализа. Смысл метода заключается в создании синтетических

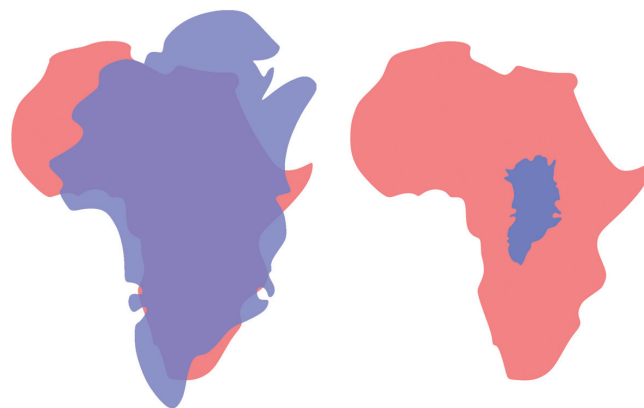


Рис. 3. Сравнение размеров Гренландии (сине-фиолетовая) с Африкой (кирпично-красная) в проекции Меркатора (слева) и в действительности (справа). Искажение размеров в проекции Меркатора явным образом проявляется на высоких широтах и обусловлено особенностями цилиндрической проекции

свойств, которые обобщают группы связанных между собой исходных свойств. Такие синтетические свойства еще называют главными компонентами. В целом, главные компоненты не интерпретируемы, но мы предполагаем, что каждый из них может отвечать за определенный геологический фактор, когда имеется сильная корреляция между главной компонентой и группой свойств, реагирующих на один фактор. В статье рассматриваются 3 геологических фактора – зрелость, тип органического вещества и состав НГМТ.

Следует отметить, что многие индивидуальные свойства зависимы также от биodeградации. Её проявление заключается в необычном изменении ОВ. По имеющимся образцам нефтей тяжелая степень биodeградации не наблюдается, поэтому исследование только этих 3 факторов (зрелость, тип, состав) является допустимым.

Чтобы понять, с каким геологическим фактором возможно ассоциировать каждую главную компоненту, в исследовании используется тепловая карта корреляций, призванная в виде интенсивности цвета запечатлеть степень взаимосвязи геохимических параметров и главных компонент (рис. 5). Рядом с тепловой картой корреляций находится таблица, показывающая какие факторы в теории (Peters et al., 2005a,b; Tissot, Velte, 1984) могут влиять на каждый признак.

До того, как приступить к корреляции нефтей с НГМТ необходимо определить, существуют ли различия между НГМТ, в каких свойствах это выражается, и с какими геологическими факторами связано. С решением всех этих вопросов может помочь метод случайного леса (рис. 4).

Для нашей задачи метод случайного леса оптимален, поскольку он не зависит от масштаба признаков, позволяя отображать на графиках геологические параметры в привычных для геохимиков шкалах значений. Поскольку мы знаем, на какие геологические факторы могут реагировать свойства, по которым толщи лучше всего отличаются, то мы способны раскрыть природу различий НГМТ и воспользоваться этим знанием как для расшифровки главных компонент, так и во время корреляции нефть-НГМТ (рис. 6).

Метод k-ближайших соседей способен количественно оценить схожесть образцов нефтей с экстрактами НГМТ (рис. 4, внизу) по всему набору геохимических параметров.

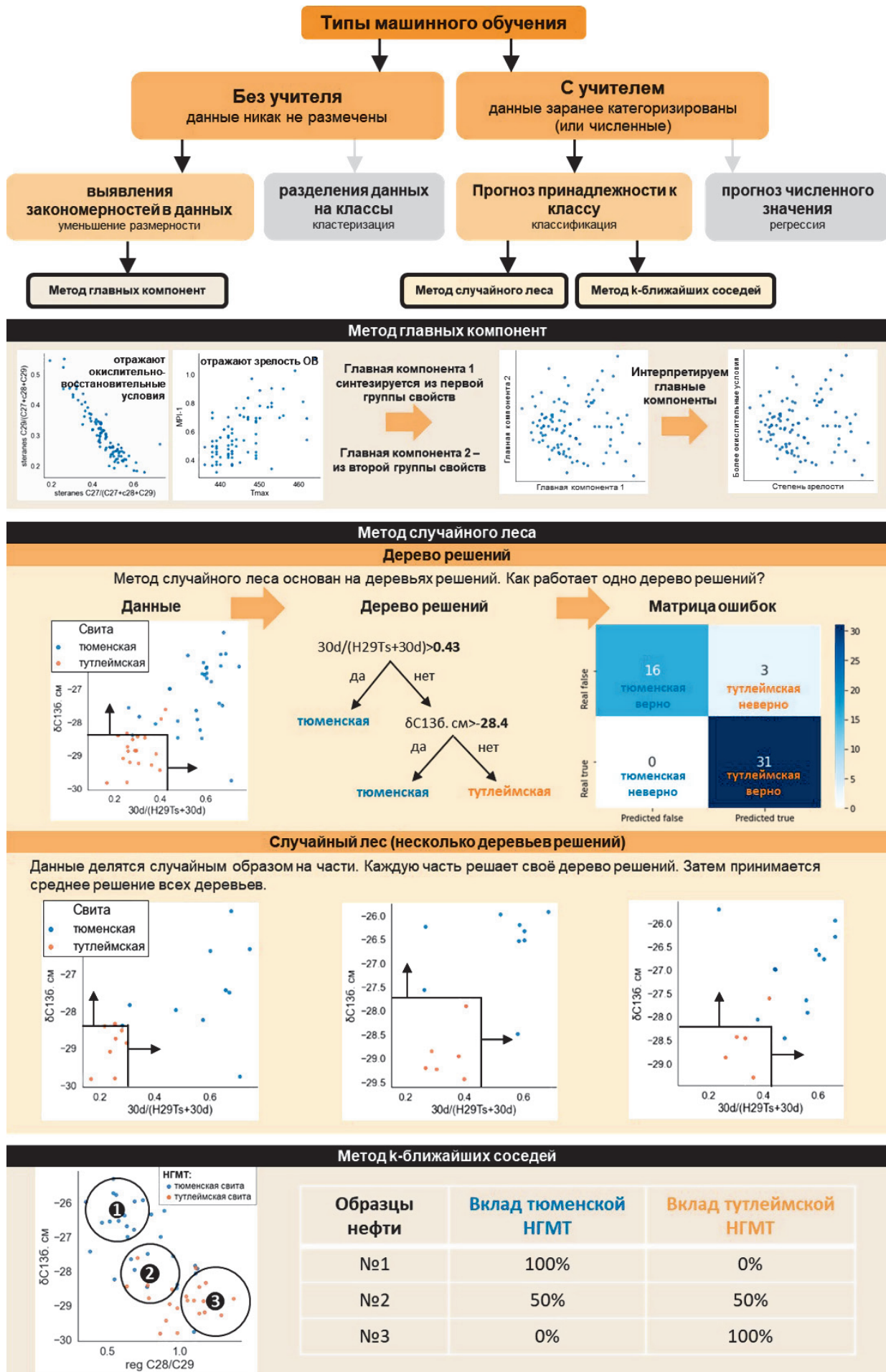


Рис. 4. Обзор применяемых в работе методов статистического анализа

Его смысл заключается в поиске ближайших к экстрактов нефтегазоматеринских толщ для каждого образца нефти. Среди ближайших к экстрактов часть из них (назовём это числом а) относятся к одной НГМТ, часть (назовём это числом b) – к другой. Отношения  $a/k$  и  $b/k$  определяют вероятность отнесения (схожести) образца нефти к одной или другой нефтегазоматеринской толще. Чтобы анализ был более достоверным, рассмотрено 2 случая: при  $k=7$  и  $k=11$ .

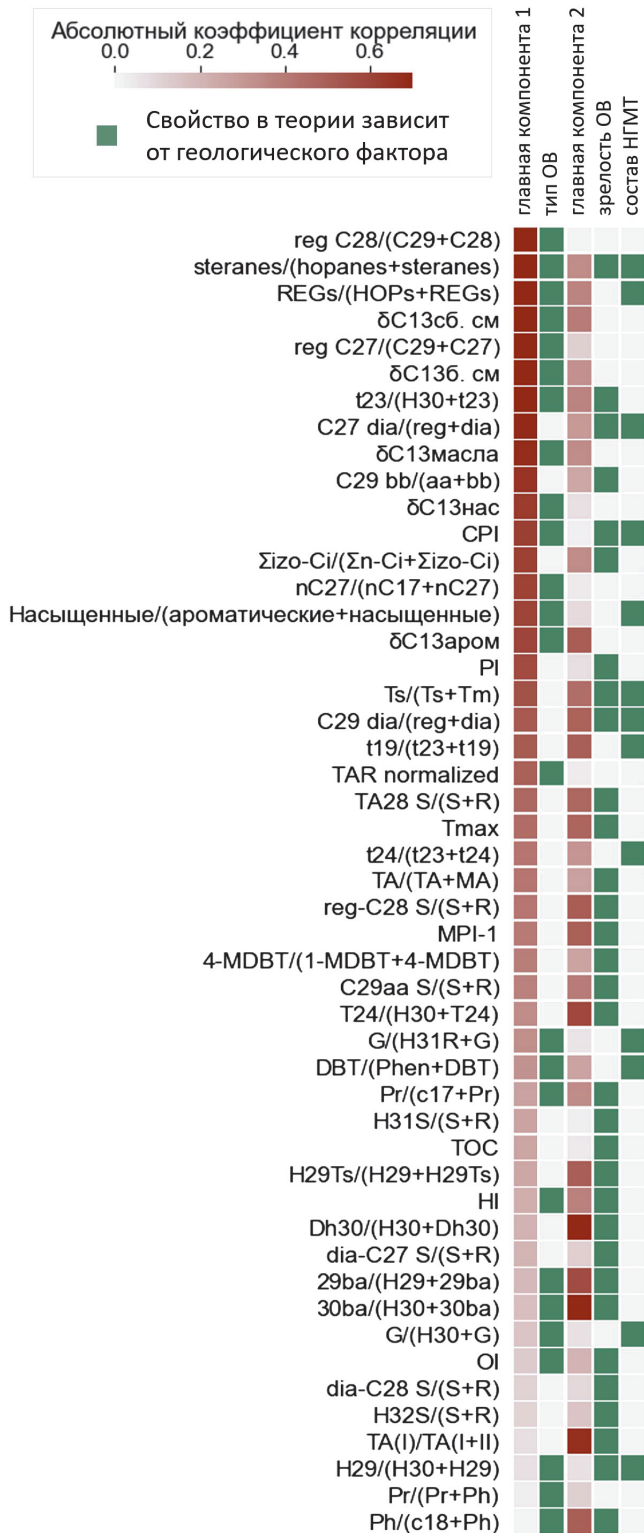


Рис. 5. Тепловая карта корреляций геохимических параметров и главных компонент и таблица теоретической зависимости свойств от зрелости, типа OB и состава НГМТ

## Результаты

### Статистический анализ органического вещества нефтей и экстрактов

Для проведения независимого и непредвзятого исследования изначально было принято решение проводить статистический анализ.

#### Выявление различий НГМТ

После предварительной подготовки данных были рассчитаны 2 главные компоненты, и по ним построен (рис. 7, слева) и далее проинтерпретирован (рис. 7, справа) график. Точками на графике показаны экстракты двух НГМТ. Облака точек НГМТ практически не пересекаются, что говорит о значительном различии тутлеймской и тюменской НГМТ. По тепловой карте корреляций (рис. 5) главная компонента 1 обобщает геохимические параметры, реагирующие на тип OB, а главная компонента 2 – параметры, чувствительные к зрелости НГМТ.

Тюменская и тутлеймская НГМТ различаются визуально по графику главных компонент. Для количественной оценки возможности разделения этих НГМТ применен метод случайного леса. Качество построенной модели оценивалось по общей точности, точности, полноте, F-мере и площади под ROC-кривой (рис. 8). Большинство из этих показателей за исключением ROC-кривой рассчитываются на основе матрицы ошибок. Чем выше эти показатели, в особенности F-мера и площадь под ROC-кривой, тем отчетливее различия между толщами. В нашем случае случайный лес обучался на 34 % данных (33 образца – 18 из тутлеймской и 15 из тюменской свит), а оценивал качество разделения на оставшихся 66 % (64 образца – 37 из тутлеймской и 27 из тюменской свит). Во избежание переобучения модели использовалась кросс-валидация по k блокам, в нашем случае  $k=5$ . Граничное значение для вероятности отнесения к той или иной НГМТ выбиралось исходя из максимизации F-меры (рис. 8 внизу) – средневзвешенной величины точности и полноты. F-мера стремится разом минимизировать 2 рода ошибок классификации – ложноположительные и ложноотрицательные. Общая точность на тестовой выборке составила 91 %, точность – 89 %, полнота – 89 %, F-мера – 89 %, площадь под ROC-кривой – 97 %, что свидетельствует о безусловном различии изучаемых НГМТ.

Метод случайного леса также помог выявить геохимические параметры, по которым наблюдаются различия НГМТ (рис. 6 слева). В основном, они реагируют на зрелость и тип OB. По графикам этих параметров (рис. 6 справа) НГМТ тутлеймской свиты является менее зрелой и имеет меньший вклад континентальной растительности по сравнению с тюменской НГМТ.

Комбинация тепловой карты корреляций и метода случайного леса позволила определить геологический смысл главных компонент (рис. 7 справа).

#### Корреляция нефть-НГМТ

Изучаемые НГМТ отличаются явным образом, что открыло окно возможностей для проведения корреляции нефть-НГМТ. Данные экстрактов и образцов нефтей прошли аналогичный этап обработки. Для анализа использовался тот же набор методов (рис. 9) плюс метод k-ближайших соседей (рис. 10).

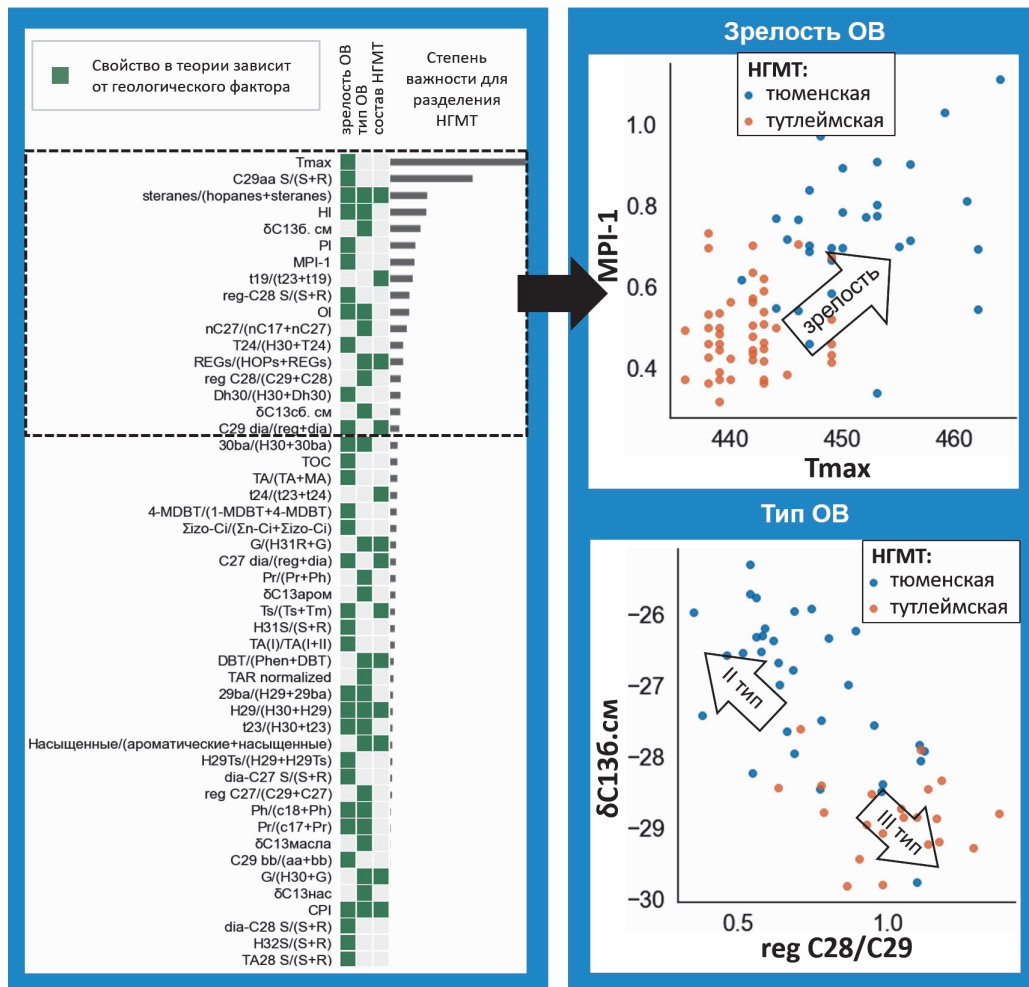


Рис. 6. Сортировка геохимических параметров по силе различий между тутлеймской и тюменской свитами (слева) инструментом важность признаков (feature importance) метода случайного леса. Этот инструмент оценивает вклад конкретного признака в качество предсказания метода случайного леса. Справа показаны графики отношений параметров, по которым наблюдаются различия двух НГМТ по зрелости, по типу ОВ

По тепловой карте корреляций (рис. 9 слева) главная компонента 1 коррелирует в основном со свойствами, реагирующими на тип ОВ, а главная компонента 2 – с геологическими параметрами, отвечающими за зрелость ОВ.

Метод случайного леса (рис. 9 по центру и справа) определил свойства, по которым отличаются лучше всего 4 группы образцов: НГМТ тутлеймской свиты, тюменской свиты, нефтей викуловской свиты, юрских и доюрских отложений. По графикам свойств нефти на качественном уровне больше приурочены к тутлеймской НГМТ по типу ОВ, хотя для юрских и доюрских пластов, вероятно, существует вклад из тюменской НГМТ. Нефти викуловской свиты обладают относительно меньшей зрелостью, чем нефти юрских и доюрских пластов, что может свидетельствовать о прекращении миграции УВ в пласты викуловской свиты в прошлом и о продолжении миграции в юрские и доюрские отложения по мере увеличения зрелости тутлеймской и тюменской НГМТ.

Тепловая карта корреляций и метод случайного леса позволили расшифровать оси главных компонент (рис. 9 внизу): главная компонента 1 уверенно ассоциируется с типом ОВ, а главная компонента 2 – со степенью зрелости. Чем выше значение главной компоненты 1, тем ближе к III типу ОВ, чем выше величина главной компоненты 2, тем выше зрелость отложений.

Количественная оценка принадлежности образцов нефтей к той или иной НГМТ производилась методом k-ближайших соседей с k=7 (рис. 10 слева) и k=11 (рис. 10 справа). Согласно его результатам ОВ нефтей викуловской толщи происходит из тутлеймской НГМТ, т.к. все ближайшие соседи к образцам нефтей викуловской свиты являются экстракты тутлеймской НГМТ. Схожесть нефтей юрских и доюрских пород с тюменской НГМТ достигает 30 % для отдельных образцов (2 ближайших соседа при k=7 или 3 ближайших соседа при k=11).

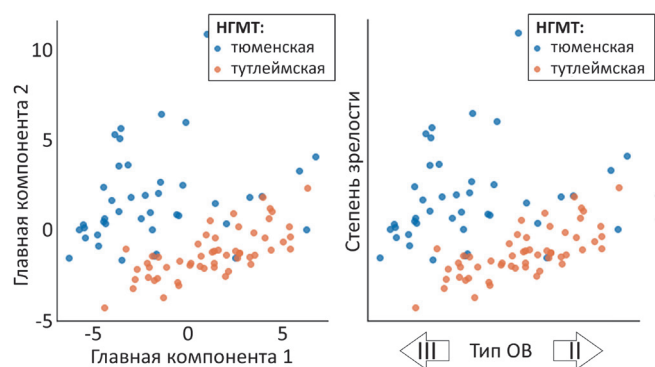


Рис. 7. График двух главных компонент до (слева) и после (справа) их расшифровки

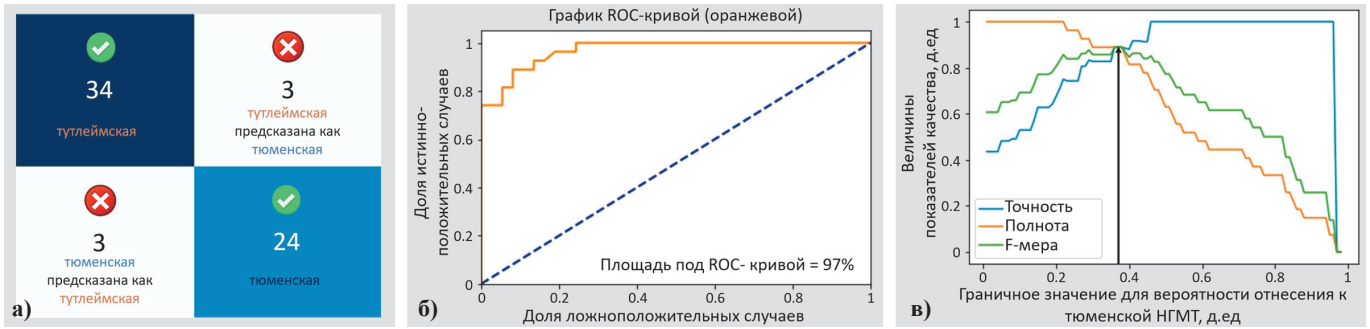


Рис. 8. а) Матрица ошибок, б) ROC-кривая и в) график зависимости F-меры, точности и полноты от граничного значения вероятности отнесения к тюменской НГМТ

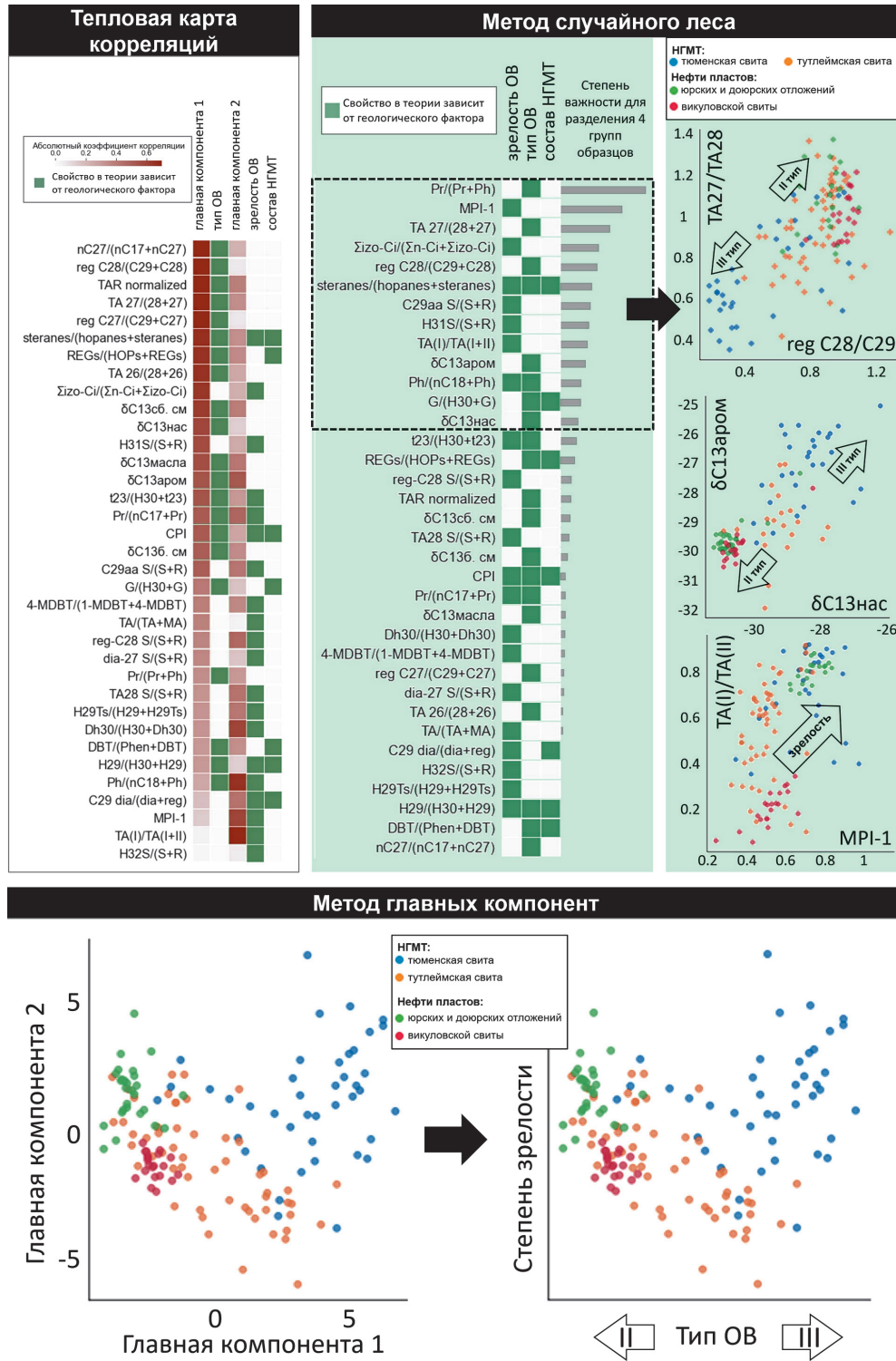


Рис. 9. Корреляция нефть-НГМТ с помощью тепловой карты корреляций, метода случайного леса и метода главных компонент



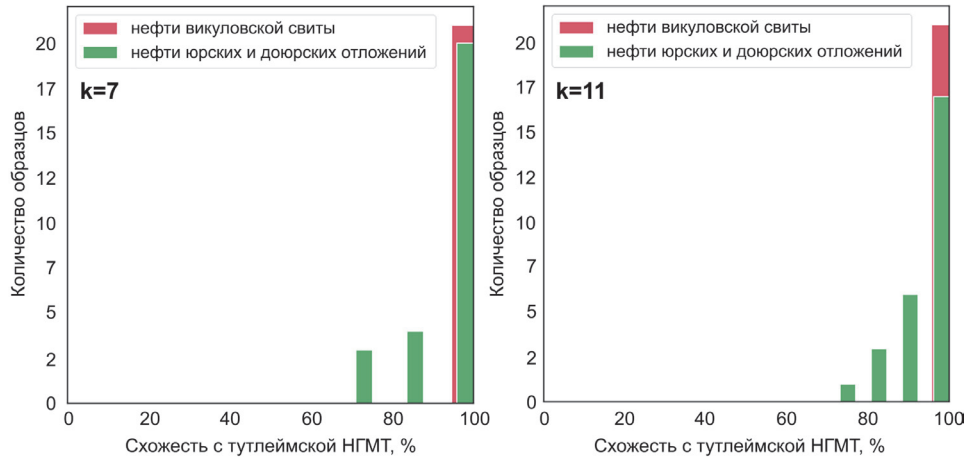


Рис. 10. Схожесть нефтей викуловских, юрских и доюрских отложений с тутлеймской НГМТ относительно тюменской, рассчитанная методом *k*-ближайших соседей: на левом графике гистограмма для *k*=7, на правом *k*=11

### Геохимический анализ органического вещества нефтей и экстрактов

При проведении геохимического анализа важным является нахождение среди множества геохимических параметров опорных, которые с высокой долей уверенности отвечают определенному геологическому фактору.

Так, после анализа всевозможных графиков выявлено, что зрелость лучше всего отражают MPI-1, T<sub>max</sub>, плотность нефти, TA(I)/TA(I+II), TA28S/(S+R) (рис. 11).

Тип ОВ отображают треугольные диаграммы регулярных стеранов, изостеранов, триароматических стероидов TA26-TA27-T28, гомогопанов H33-H34-H35, парные графики изотопного состава углерода δC13 насыщенные с δC13 ароматические и TA26/TA28 с reg-C27/reg-C29 (рис. 12).

Согласно рис. 11 НГМТ тюменской свиты имеет более высокую зрелость по сравнению с тутлеймской. ОВ нефтей юрских и доюрских отложений относительно более зрелое, чем ОВ нефтей викуловской свиты.

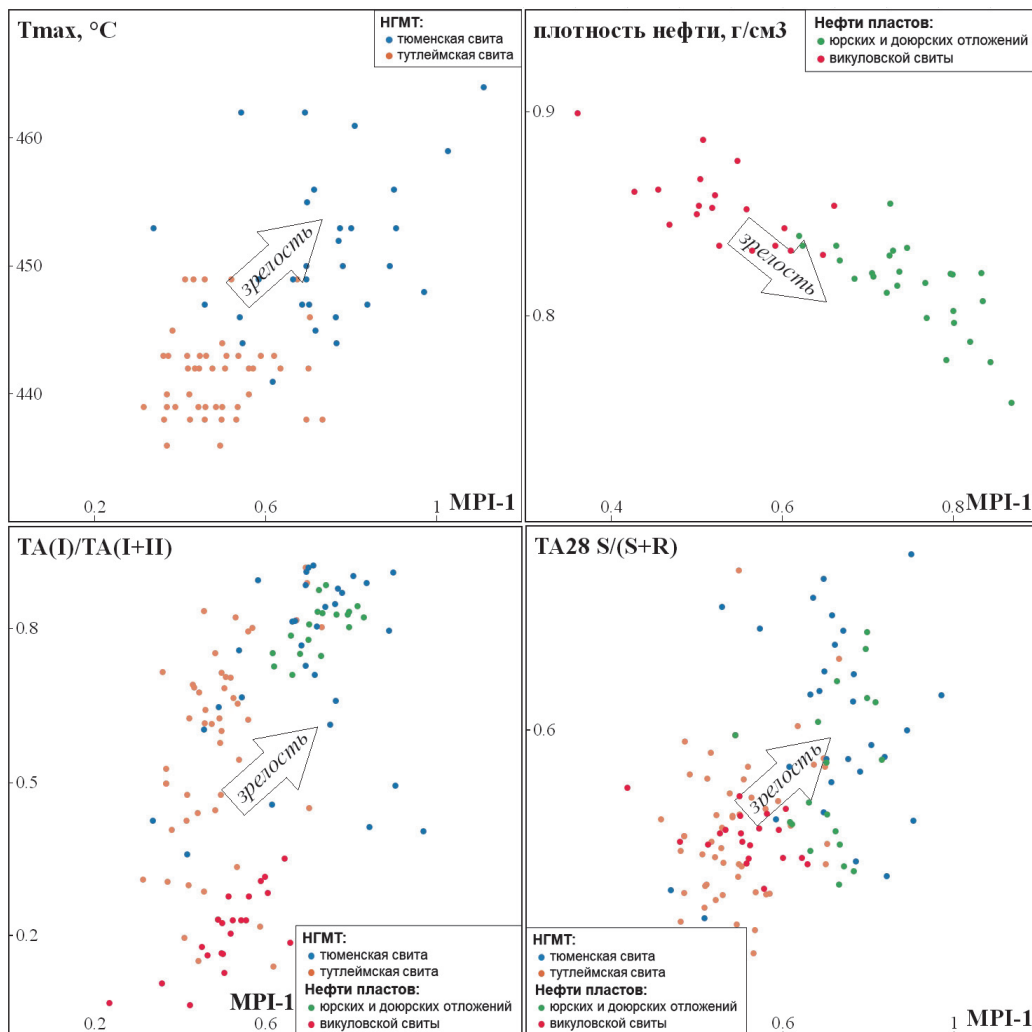


Рис. 11. Графики геохимических параметров, реагирующих на зрелость ОВ

Графики и треугольные диаграммы свойств, реагирующих на тип ОВ (рис. 12), отчетливо показывают, что органическое вещество тюменской свиты имеет больший вклад континентального ОВ по сравнению с остальными образцами. Это свидетельствует о близости нефтей к тутлеймской нефтегазоматеринской толще. В целом, имеющиеся графики не исключают наличие некоторого вклада тюменской НГМТ в нефтях юрских и доюрских отложений.

Результаты геохимического и статистического анализов сходятся.

### Обсуждение

В данной работе в основном наблюдается линейное изменение геохимических параметров по мере воздействия геологических факторов. Но не стоит забывать, что развитие состава ОВ зачастую импульсивно. Импульсивность связана с биогеохимическими барьерами, позволяющими геохимическим параметрам не реагировать на геологические факторы до определенного предела. Например,  $dia/reg$  C29 начинает отражать зрелость НГМТ при значениях  $>2$ , а до этих значений нет возможности судить о каком-либо её влиянии.

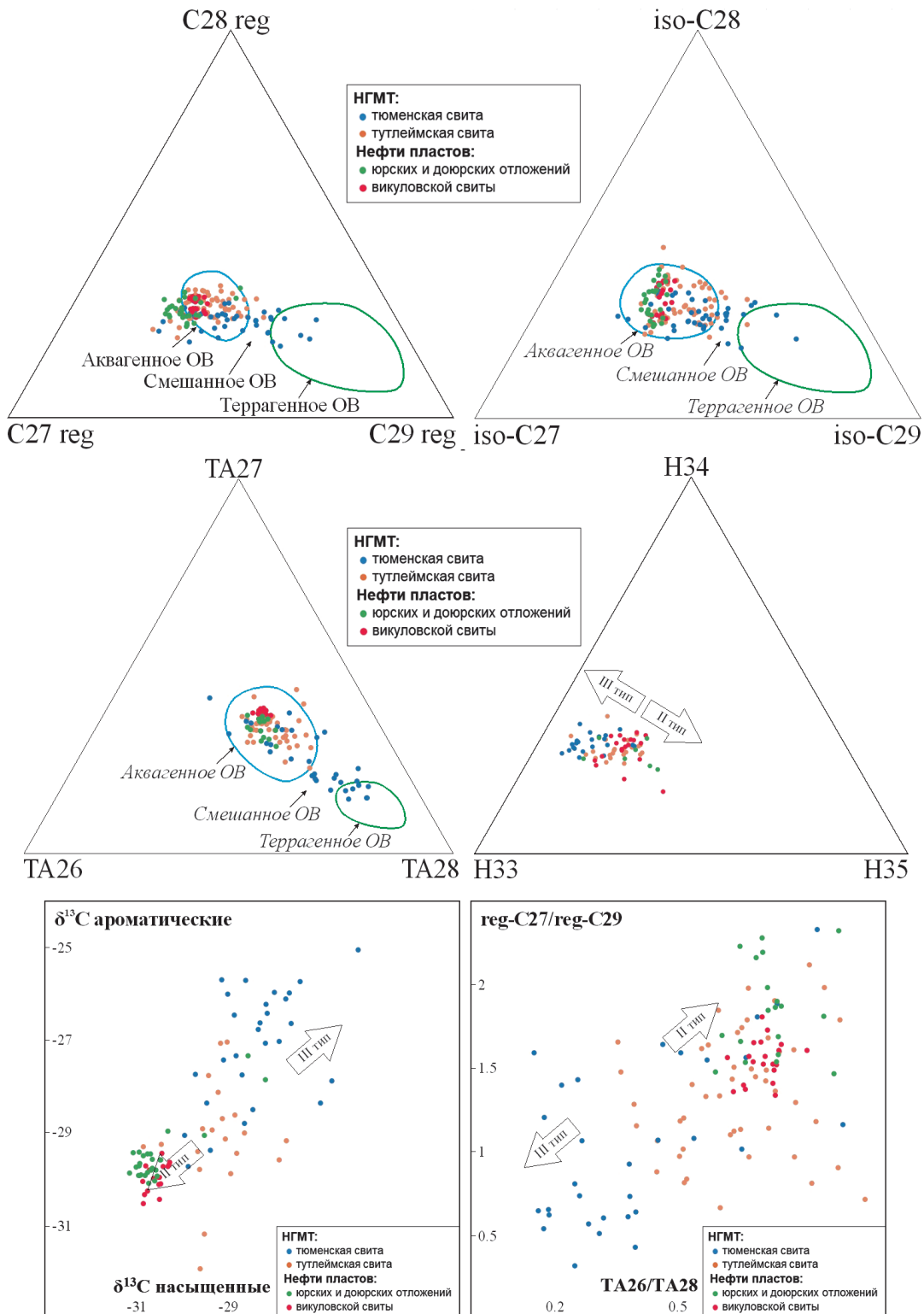


Рис. 12. Графики геохимических параметров, реагирующих на тип ОВ

Необходимо упомянуть, что каждый геохимический параметр имеет свои границы использования. Так, отношение нечетных *n*-алканов к ближайшим четным (OEP – odd even predominance) применимо в целом к палеозойским нефтям и ОВ, а изомеризация гопанов Н30-Н35 наблюдается с ростом зрелости в основном до градации катагенеза ПК3.

По мере развития науки некоторые теоретические знания устаревают. Так, в настоящее время статистический параметр (CV – canonical variable), рассчитываемый по изотопному составу ароматической и насыщенной фракций ОВ, не несёт полезной нагрузки.

Особенно важным при проведении анализа является понимание степени доверия к тем или иным данным. В частности, существуют трудности идентификации dia-C28, поэтому отношение dia/reg C28 считается менее надежным, чем dia/reg C29 или dia/reg C27. Другой пример, нормальные стераны часто с чем-то коэллинируют, а изостераны нет, следовательно, к последним доверие при прочих равных больше.

Хотя в ходе исследования не были выявлены недостатки методов статистического анализа, но следует отметить, что само по себе заполнение пустых ячеек на этапе обработки данных может ухудшать качество результатов. Поэтому необходимо тщательно выбирать методы заполнения пустых полей и в целом уделять большое внимание обработке данных перед применением методов статистического анализа.

### **Возможность автоматизации геохимических исследований с помощью статистического анализа**

Необходимая для проведения анализа единая таблица для НГМТ и нефтей собрана в ручном режиме из множества таблиц по каждому образцу. В целом, этот процесс можно автоматизировать, но для этого необходимо заранее договориться о единой внутренней структуре всех файлов.

Длительность расчётов, представленных в работе, зависит от вычислительных мощностей ПК, даже на слабом ноутбуке они заняли менее 10 минут. Необходимо понимать, что львиную долю времени занимают не столько расчёты, сколько разработка методики исследования и написание кода. Их наличие, а также возможность в автоматическом режиме собирать отдельные таблицы в единую позволяют производить подобный статистический анализ в день получения исходных данных.

Геохимический анализ не стоит на месте, какие-то знания устаревают, обнаруживаются новые геохимические параметры, в будущем, вероятно, появится возможность оценивать новые геологические факторы, о которых ранее не задумывались. Деятельность экспертов-геохимиков является творческим процессом, поскольку каждый образец ОВ является по-своему уникальным. Геохимик, помещающий данные в таблицу, понимает, каким образом они получены, можно ли им доверять, то есть он является первой инстанцией в вопросах качества входных данных.

В связи с вышеперечисленным, инструмент статистического анализа предлагается использовать в качестве экспресс оценки, а финальное слово оставлять за экспертами-геохимиками, работающими с веществом.

### **Заключение**

Геохимический и статистический анализы используют один и тот же набор входных данных, но работают с ними по-разному. Для проведения геохимического анализа нет необходимости устранять все аномальные значения, заполнять пустые поля в таблице, изменять масштаб и форму распределения геохимических параметров в отличие от статистического анализа.

Геохимический анализ имеет в значительной степени экспертную оценку, наработанную многими годами. Для применения методов статистического анализа требуется формализация этого опыта в виде таблиц. Например, теоретическое влияние геологических факторов на каждый геохимический параметр (рис. 5 зеленые ячейки).

Методам статистического анализа на вход подаются все имеющиеся геохимические параметры, и чем их больше, тем лучше. В то же время геохимический анализ направлен на поиск надёжных и отсеивание нерабочих свойств, то есть проявляется принцип отделения зёрен от плевел.

Геохимический анализ позволяет оценивать имеющиеся исходные данные, визуализируя их на парных графиках, треугольных или лепестковых диаграммах. С помощью методов статистического анализа создаются синтетические свойства на основе имеющихся, обычным является использование тепловой карты корреляции, а для оценки способности методов классификации к разделению нескольких групп образцов используется матрица ошибок, и строится ROC-кривая (кривая, которая наиболее часто используется для представления результатов бинарной классификации в машинном обучении).

В работе использовано 3 метода статистического анализа. Каждый из них имеет определенную область применимости (рис. 4), а потому их прямое сравнение друг с другом не имеет смысла. Даже при решении какой-либо одной задачи одни методы могут показывать большую точность на одном наборе данных и меньшую на другом. В работе намеренно используются базовые методы статистического анализа, поскольку в их основе лежат простые принципы, и чтобы показать осуществимость решения геохимических задач без использования более тяжелых технологий – ансамблей алгоритмов или нейронных сетей.

Хотя существует множество различий между статистическим и геохимическим анализами, они пришли к одним и тем же выводам:

- тутлеймская НГМТ менее зрелая и имеет меньший вклад континентального ОВ относительно тюменской НГМТ;
- ОВ нефтей викуловской свиты имеет явное сходство с тутлеймской НГМТ;
- нефти юрских и доюрских отложений имеют большую схожесть с тутлеймской, чем тюменской НГМТ (до 30% в некоторых образцах);
- нефти викуловской свиты менее зрелые по сравнению с нефтями юрских и доюрских отложений. Причина такого различия может состоять в прекращении миграции УВ в прошлом в вышележащие апт-альбские викуловские пласты и продолжающейся на сегодняшний день подпитке УВ в юрские и доюрские отложения.

Таким образом, современные методы исследований в ближайшем будущем позволят проводить автоматическую интерпретацию результатов уже в первые минуты после проведения исследований, при условии создания единых баз данных, в том числе и обезличенных. Прототипом такой базы данных может стать платформа <https://mygeomar.ru>.

Данная статья показывает, что интерпретацию сложных геолого-геохимических исследований можно внедрять непосредственно на дочерних обществах и корпоративных научно-исследовательских предприятиях нефтегазовых компаний, для оперативного решения актуальных производственных задач энергетической отрасли.

## Литература

Атлас «Геология и нефтегазоносность Ханты-Мансийского автономного округа» (2004). Ред. Э.А. Ахпателов, В.А. Волков, В.Н. Гончарова, В.Г. Елисеев, В.И. Карасев, А.Г. Мухер, Г.П. Мясникова, Е.А. Тепляков, Ф.З. Хафизов, А.В. Шпильман, В.М. Южакова. Екатеринбург: ИздатНаукаСервис, 148 с.

Ступакова А.В. (2011). Структура и нефтегазоносность Баренцево-Карского шельфа и прилегающих территорий. *Геология нефти и газа*, 6, с. 99–115.

Peters K.E., Walters C.C., Moldowan J.M. (2005a). The Biomarker Guide. Second Edition. I. Biomarkers and Isotopes in the Environment and Human History. New York: Cambridge University Press, 492 p. <https://doi.org/10.1017/CBO9780511524868>

Peters K.E., Walters C.C., Moldowan J.M. (2005b). The Biomarker Guide. Second Edition. II. Biomarkers and Isotopes in Petroleum Systems and Earth History. New York: Cambridge University Press, 704 p. <https://doi.org/10.1017/CBO9781107326040>

Tissot B.P., Welte D.H. (1984). Petroleum Formation and Occurrence. Second Revised and Enlarged Edition. Berlin: Springer-Verlag Berlin Heidelberg, 702 p. <https://doi.org/10.1007/978-3-642-87813-8>

*Статья поступила в редакцию 09.03.2022;*

*Принята к публикации 15.04.2022;*

*Опубликована 16.05.2022*

## Сведения об авторах

*Константин Олегович Осипов* – выпускник аспирантуры, научный сотрудник кафедры геологии и геохимии горючих ископаемых, Московский государственный университет имени М.В. Ломоносова

Россия, 119234, Москва, ул. Ленинские горы, д. 1

*Энвер Алексеевич Абля* – кандидат геол.-мин. наук, доцент кафедры геологии и геохимии горючих ископаемых, Московский государственный университет имени М.В. Ломоносова

Россия, 119234, Москва, ул. Ленинские горы, д. 1

*Роман Сергеевич Сауткин* – кандидат геол.-мин. наук, старший научный сотрудник кафедры геологии и геохимии горючих ископаемых, Московский государственный университет имени М.В. Ломоносова

Россия, 119234, Москва, Ленинские горы, д. 1

*Мария Александровна Большакова* – кандидат геол.-мин. наук, старший научный сотрудник кафедры геологии и геохимии горючих ископаемых, Московский государственный университет имени М.В. Ломоносова

Россия, 119234, Москва, ул. Ленинские горы, д. 1

*Анна Анатольевна Суслова* – кандидат геол.-мин. наук, ведущий научный сотрудник кафедры геологии и геохимии горючих ископаемых, Московский государственный университет имени М.В. Ломоносова

Россия, 119234, Москва, Ленинские горы, д. 1

*Алексей Петрович Антонов* – кандидат физ.-мат. наук, доцент кафедры математического анализа, руководитель НОЦ ПАО «НК «Роснефть» по цифровым технологиям в нефтегазовой отрасли на базе кафедры геологии и геохимии горючих ископаемых, Московский государственный университет имени М.В. Ломоносова

Россия, 119234, Москва, ул. Ленинские горы, д. 1

IN ENGLISH

ORIGINAL ARTICLE

# Identification of the features of the organic matter of oils and source rocks based on a comparison of the results of geochemical analysis with statistical analysis built upon machine learning methods: the case of one field of the West Siberian basin

*K.O. Osipov\*, E.A. Ablya, R.S. Sautkin, M.A. Bolshakova, A.A. Suslova, A.P. Antonov*

*Lomonosov Moscow State University, Moscow, Russian Federation*

*\*Corresponding author: Konstantin O. Osipov, e-mail: k.osipov@oilmsu.ru*

**Abstract.** The idea of this work is to compare the results of geochemical and statistical analyzes in the study of organic matter in extracts of oil and gas source rock. The object of the study were the samples of oil and gas source rocks of the Tutleim and Tyumen Formations, as well as oil samples of the Vikulov Formation and Jurassic and pre-Jurassic deposits in the western part of the West Siberian basin.

Among the methods of statistical analysis, the principal component method and the random forest method were used. A heat map of correlations was used as visualization tools.

The principal component method helped us to reveal a clear difference between the organic matter of the Tutleim and Tyumen source rocks. The random forest method and the heat map made it possible not only to identify the distinctive

geochemical properties for this strata, but also reveals the geological factors of their distinction. Thus, the organic matter of Tyumen source rock turned out to be more mature and it has relatively larger contribution of terrestrial organic matter compared to the Tutleim one. The same set of methods is applied to oil-source rock correlation. The oils of the Vikulov Formation obviously originate from the Tutleim source rock. Oils of Jurassic and pre-Jurassic reservoirs have a predominant contribution of the Tutleim source rock and some similarity (up to 30 %) with the Tyumen one. Oils from the Vikulov Formations are less mature than oils from the Jurassic and pre-Jurassic rocks. This may indicate the continued migration of hydrocarbons to Jurassic and pre-Jurassic deposits and the cessation of migration in the past to the overlying

Aptian–Albian Vikulov Formation. Geochemical analysis also revealed the geochemical properties responsible for the maturity of source rocks and the type of organic matter. It confirmed the conclusions made on the basis of the application of methods of statistical analysis.

**Keywords:** comparison of studies, geochemical analysis, methods of statistical analysis, identification of differences in source rocks, oil-source rock correlation

**Recommended citation:** Osipov K.O., Stoupakova A.V., Ablya E.A., Sautkin R.S., Bolshakova M.A., Suslova A.A., Antonov A.P. (2022). Identification of the features of the organic matter of oils and source rocks based on a comparison of the results of geochemical analysis with statistical analysis built upon machine learning methods: the case of one field of the West Siberian basin. *Georesursy = Georesources*, 24(2), pp. 217–229. DOI: <https://doi.org/10.18599/grs.2022.2.20>

### References

Geology and oil and gas potential of the Khanty-Mansiysk Autonomous Okrug. Atlas (2004). Ed. E.A. Akhpatelov, V.A. Volkov, V.N. Goncharova, V.G. Eliseev, V.I. Karasev, A.G. Mukher, G.P. Myasnikova, E.A. Teplyakov, F.Z. Khafizov, A.V. Shpil'man, V.M. Yuzhakova. Ekaterinburg: IzdatNaukaServis, 148 p. (In Russ.)

Peters K.E., Walters C.C., Moldowan J.M. (2005a). The Biomarker Guide. Second Edition. I. Biomarkers and Isotopes in the Environment and Human History. New York: Cambridge University Press, 492 p. <https://doi.org/10.1017/CBO9780511524868>

Peters K.E., Walters C.C., Moldowan J.M. (2005b). The Biomarker Guide. Second Edition. II. Biomarkers and Isotopes in Petroleum Systems and Earth History. New York: Cambridge University Press, 704 p. <https://doi.org/10.1017/CBO9781107326040>

Stoupakova A.V. (2011). Structure and petroleum potential of the Barents-Kara shelf and adjacent territories. *Geologiya nefi i gaza = Russian Oil and Gas Geology*, 6, pp. 99–115. (In Russ.)

Tissot B.P., Welte D.H. (1984). Petroleum Formation and Occurrence. Second Revised and Enlarged Edition. Berlin: Springer-Verlag Berlin Heidelberg, 702 p. <https://doi.org/10.1007/978-3-642-87813-8>

### About the Authors

*Konstantin O. Osipov* – PhD student, Researcher, Petroleum Geology Department, Lomonosov Moscow State University

1, Leninskie gory, Moscow, 119234, Russian Federation

*Enver A. Ablya* – PhD (Geology and Mineralogy), Associate professor, Petroleum Geology Department, Lomonosov Moscow State University

1, Leninskie gory, Moscow, 119234, Russian Federation

*Roman S. Sautkin* – PhD (Geology and Mineralogy), Senior Researcher, Petroleum Geology Department, Lomonosov Moscow State University

1, Leninskie gory, Moscow, 119234, Russian Federation

*Maria A. Bolshakova* – PhD (Geology and Mineralogy), Senior Researcher, Petroleum Geology Department, Lomonosov Moscow State University

1, Leninskie gory, Moscow, 119234, Russian Federation

*Anna A. Suslova* – PhD (Geology and Mineralogy), Leading Researcher, Petroleum Geology Department, Lomonosov Moscow State University

1, Leninskie gory, Moscow, 119234, Russian Federation

*Alexey P. Antonov* – PhD (Physics and Mathematics), Associate Professor of Mathematical Analysis Department, Head of Rosneft Research Center, Lomonosov Moscow State University

1, Leninskie gory, Moscow, 119234, Russian Federation

*Manuscript received 9 March 2022;  
Accepted 15 April 2022; Published 16 May 2022*