Redox Potentials, Basicities and Dipole Moments of the Diphenylamine Series As Analytical Reagents // Canad. J. Chem. 1999. Vol. 77, № 12. P. 2053–2058.

52. *Панкратов А. Н., Щавлев А. Е.* Протолитические, окислительно-восстановительные и полярные свойства реагентов ряда дифениламина : квантовохимическая оценка // Журн. аналит. химии. 2001. Т. 56, № 2. С. 143–150.
53. *Панкратов А.Н.* Строение продукта окисления дифениламина – родоначального представителя ряда аналитических редокс-реагентов // Журн. аналит. химии. 2001. Т. 56, № 2. С. 161–163.
54. *Минкин В. И., Симкин Б. Я., Миняев Р. М.* Теория строения молекул. Ростов н/Д : Феникс, 1997. 560 с.
55. *Ермаков А. И.* Квантовая механика и квантовая химия. М. : Изд-во Юрайт; ИД Юрайт, 2010. 555 с.
56. *Minkin V. I.* Glossary of Terms Used in Theoretical Organic Chemistry (IUPAC Recommendations 1999) // Pure and Appl. Chem. 1999. Vol. 71, № 10. P. 1919–1981.
57. Глоссарий терминов, используемых в теоретической органической химии (окончание) // Журн. орган. химии. 2001. Т. 37, вып. 7. С. 1105–1112.
58. *Цирельсон В. Г.* Квантовая химия. Молекулы, молекулярные системы и твердые тела. М.: БИНОМ. Лаб. знаний, 2010. 496 с.
59. *Жидомиров Г. М., Багатурьянц А. А., Абронин И. А.* Прикладная квантовая химия. Расчеты реакционной способности и механизмов химических реакций. М. : Химия, 1979. 296 с.
60. *Яновская Л. А.* Современные теоретические основы органической химии. М.: Химия, 1978. 360 с.
61. *Днепровский А. С., Темникова Т. И.* Теоретические основы органической химии. Строение, реакционная способность и механизмы реакций органических соединений. Л. : Химия. Ленингр. отд-ние, 1991. 560 с.
62. *Ахметов Н. С.* Общая и неорганическая химия. М. : Высш. шк., 2006. 743 с.
63. *Тодрес З. В.* Ион-радикалы в органическом синтезе. М. : Химия, 1986. 240 с.
64. *Camaioni D. M., Franz J. A.* Carbon-Hydrogen *vs*. Carbon-Carbon Bond Cleavage of 1,2-Diarylethane Radical Cations in Acetonitrile-Water // J. Org. Chem. 1984. Vol. 49, № 9. P. 1607–1613.

УДК 543.041

ИДЕНТИФИКАЦИЯ НЕФТЕЙ САМАРСКОЙ ОБЛАСТИ С ИСПОЛЬЗОВАНИЕМ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ И ФАКТОРНОГО ДИСКРИМИНАНТНОГО АНАЛИЗА

А. Л. Лобачев[1], Н. В. Фомина[1], Ю. Б. Монахова[2]

[1]Самарский государственный университет
E-mail: nkichimaeva@mail.ru
[2]Саратовский государственный университет
E-mail: yul-monakhova@mail.ru

Разработка методов идентификации месторождений нефти является приоритетной задачей нефтяной промышленности. В ходе исследования были определены следующие параметры для 2963 образцов нефти с пяти месторождений Самарской области: плотность, выход фракций при температуре 200 ºС и 300 ºС, массовая доля серы, содержание сероводорода, метил- и этилмеркаптанов, массовая концентрация хлористых солей и давление насыщенных паров. Матрица экспериментальных данных обработана с помощью хемометрического метода главных компонент (МГК) и факторного дискриминантного анализа (ФДА). Полученные модели позволяют определять месторождение образцов нефти с вероятностью практически 100%. Проведена проверка хемометрических моделей с помощью независимого тестового набора, которая показала достоверность и устойчивость моделей. Результаты проведенного анализа свидетельствуют о перспективности применения хемометрических методов для дискриминации образцов нефти различных месторождений Самарской области, а подобный подход может быть использован и для классификации образцов нефти из других регионов.

Ключевые слова: нефть, метод главных компонент, факторный дискриминантный анализ, классификация.

Identification of Oils from Samara Region Using Principal Component Analysis and Factor Discriminant Analysis

A. L. Lobachev, N. V. Fomina, Yu. B. Monakhova

Development of methods for identification of oils is of high priority in oil industry. The following parameters for 2963 oil samples from five oilfields in the Samara region were determined: density, fraction yield at 200 °C and 300 °C, the mass fraction of sulfur, hydrogen sulphide, methyl and ethyl mercaptan, the mass concentration of chloride salts, the saturated vapor pressure. The matrix of experimental data was analyzed using principal component analysis (PCA) and factorial discriminant analysis (FDA) methods. The models obtained are able to determine the oilfield of samples with probability of almost 100%. Chemometric models have been proved

by the independent test set validation, which showed the accuracy and stability of the models. The results of the analysis indicated the prospects of application of chemometric methods in the investigation of oil samples from Samara region and the developed approach can be used to discriminate oils from another regions.

Key words: oil, principal component analysis, factor discriminant analysis, classification.

Развитие новых отраслей науки и техники, анализ объектов природного и техногенного происхождения, занимающий ключевое место в экологических экспертизах, химической, нефтяной, пищевой промышленности, ставят перед аналитической химией задачу совершенствования методов качественного и количественного анализа. В зависимости от поставленных целей и задач выбираются схемы проведения анализа, этапность их реализации, методы исследований.

Наиболее простой и доступный метод идентификации – использование индивидуальных эталонных веществ или эталонных смесей. Необходимо лишь разложить анализируемую смесь при таких же условиях, при которых была разделена эталонная смесь. Но провести однозначную идентификацию таким образом можно только тогда, когда исследователь имеет необходимые эталонные вещества, причем компоненты смеси хорошо разделяются. Как правило, на практике данные условия не выполняются.

Нефть – весьма сложный объект анализа [1], для идентификации которого предпочтительным является использование математических методов моделирования экспериментальных данных. Особый интерес представляет определение месторождения, на котором добыт тот или иной образец нефти. Иногда для решения этой задачи используется комбинация аналитических методов. Так, авторами [2] на примере нефтей Ханты-Мансийского АО показано, что для идентификации источников нефтяных загрязнений по составу примесей могут быть использованы гамма-спектрометрический, атомно-абсорбционный, атомно-эмиссионный и рентгенофлуоресцентный методы анализа, а также индуктивно-связанная плазма с масс-спектрометрической или оптической регистрацией. Однако методы высокого разрешения также могут не давать 100%-ной надежности идентификации материала сложного состава.

Таким образом, на сегодняшний день основным путем решения задачи идентификации географического происхождения нефтей является прямой метод, который заключается в проведении полного качественного и количественного анализа состава материала и дальнейшего сравнения полученных данных с данными о материале сравнения. Однако идентификация и количественное определение абсолютно всех компонентов реальных объектов принципиально невозможны из-за отсутствия индивидуальных стандартных веществ (компонентов нефти). В общем случае причиной отсутствия стандартов могут быть дороговизна, малый срок годности, но чаще всего отсутствие его как такового. В любом случае данное принципиальное ограничение заставляет искать новые подходы в обработке и использовании полученной химиками-аналитиками информации на основе развития различных безэталонных аналитических методов, в частности, использования интегральной совокупности аналитических сигналов, обработанных с помощью различных статистических методов.

В настоящей работе изучена возможность использования метода главных компонент (МГК) и факторного дискриминантного анализа (ФДА) для определения месторождения нефти по ее стандартным показателям качества.

Экспериментальная часть

В работе проводилось определение стандартных характеристик проб нефти, отобранных с пяти месторождений Самарской области. Пробы отбирались в соответствии с ГОСТ 2517-85 «Нефть и нефтепродукты. Методы отбора проб», всего за период январь-декабрь 2011 г. с каждой группы месторождений было проанализировано от 350 до 725 проб нефти.

Определяли такие характеристики, как плотность (ареометрически по ГОСТ 3900-85, использовали ареометр для нефти АН соответствующего диапазона), выход фракций при температуре 200 ^{o}C и 300 ^{o}C (по ГОСТ 2177-99, на аппарате для разгонки нефти и нефтепродуктов DU-4), массовую долю серы (по ГОСТ Р 51947-2002, использовали энергодисперсионный рентгенофлуоресцентный анализатор OXFORD Lab-X-3500), содержание сероводорода, метил- и этилмеркаптанов (хроматографически по ГОСТ Р 50802-95, использовали комплекс хроматографический «Хроматэк-Кристалл-5000»), массовую концентрацию хлористых солей (титриметрически по ГОСТ 21534-76, использовали экстрактор хлористых солей ПЭ-8110, DE-8110, средства измерения в соответствии с требованиями ГОСТ), массовую долю воды (по ГОСТ 2477-65), давление насыщенных паров (по ASTM D 323-08, использовали автоматический анализатор давления насыщенных паров AutoReid). Данные характеристики предложены в качестве параметров для моделирования.

Набор данных составил выборку из 2963 образцов, из них 686 относятся к первому, 716 – ко второму, 725 – к третьему, 357 – к четвертому и 479 – к пятому месторождению нефти, различающихся по географическому положению.

Моделирование данных производили на основе программного комплекса MATLAB 2013b (The MathWorks, Natick, USA) с встроенной в него оболочкой для хемометрических расчетов SAISIR [3].

Для визуализации и поиска скрытых закономерностей в экспериментальных данных был использован МГК [4]. В качестве предварительной обработки использовано центрирование данных. Метод ФДА выбран для классификации месторождений нефти по географическому положению [5]. Наличие большой выборки объектов позволило применить метод проверки модели с помощью тестового набора. Для этого все образцы разделены случайным образом на обучающий (1975 образцов) и проверочный (988 образцов) наборы данных.

Результаты и их обсуждение

Метод главных компонент

Хемометрическое исследование многомерных данных независимо от природы сигналов (спектры, хроматограммы, дискретные данные) обычно начинают с применения метода главных компонент [6–9]. МГК дает возможность отделить содержательную часть данных от шума, что позволяет представить полезную информацию в более компактном виде, удобном для визуализации и интерпретации [5,6].

В нашем случае МГК применен для матрицы данных размером 2963·9 (9 различных характеристик для 2963 объектов). Установлено, что данные могут быть описаны тремя главными компонентами (ГК), которые в сумме объясняют 99.9% дисперсии данных (ГК1 – 97.6%, ГК2 – 1.5% и ГК3 – 0.8%). График счетов в пространстве ГК1–ГК2 показывает дискриминацию кластеров, соответствующих различным месторождениям нефти (рис. 1). Очевидно, что кластеры, отвечающие месторождениям 3–5, достаточно хорошо разделены друг от друга, однако кластеры групп месторождений 1 и 2 значительно перекрываются между собой. Следует отметить, что эти два кластера не разделены при рассмотрении третьей ГК, например, в пространстве ГК1–ГК3 или ГК2–ГК3.

Для выявления влияния девяти переменных на разделение групп месторождений был использован график нагрузок (рис. 2).

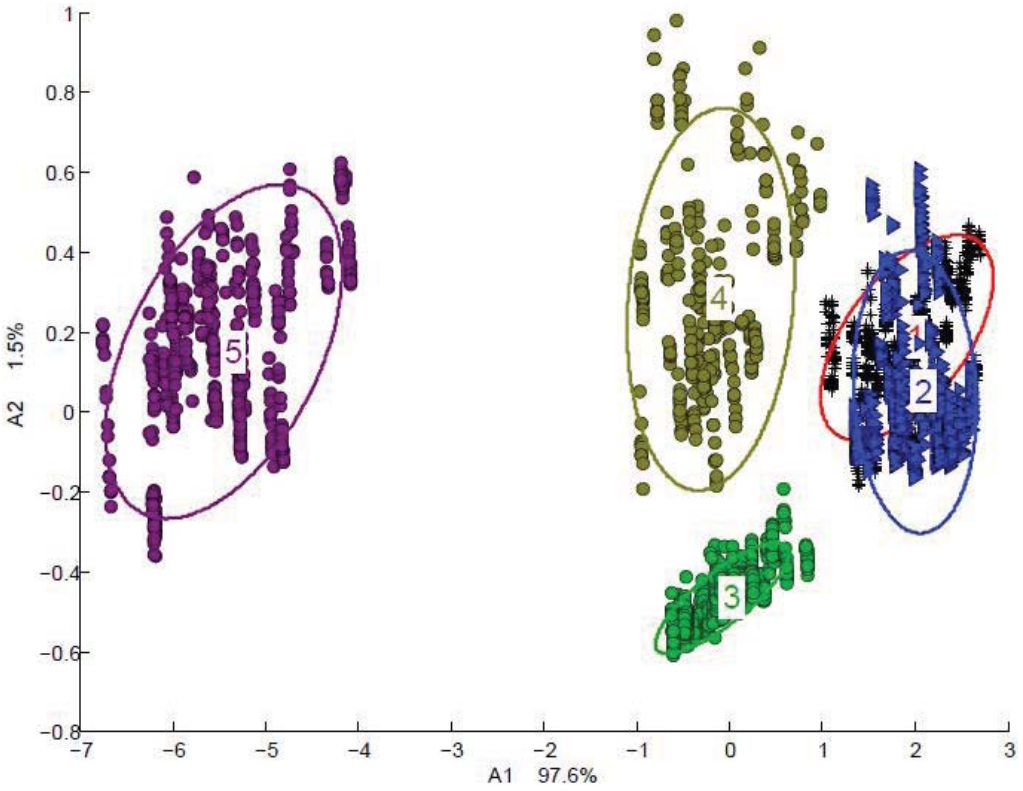

Рис. 1. График счетов в пространстве ГК1–ГК2 (эллипсы построены с вероятностью 95%, цифры указывают кластер каждого месторождения)

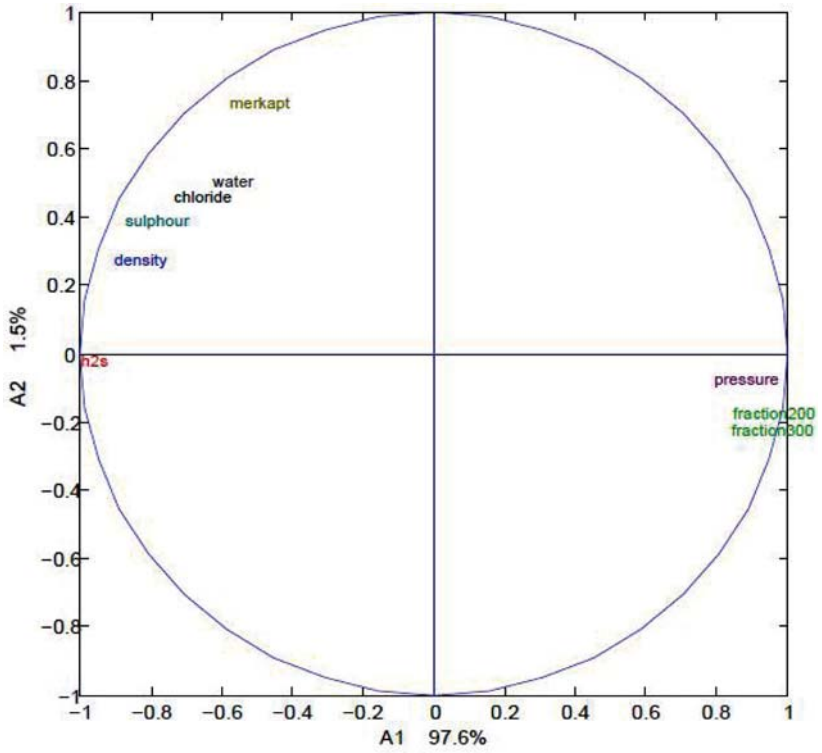

Рис. 2. График нагрузок в пространстве ГК1–ГК2: density – плотность при 20°C, chloride – концентрация хлорных солей, water – массовая доля воды, sulphour – массовая доля серы, pressure – давление насыщенных паров, h2s – масс. доля сероводорода, merkapt – массовая доля метил- и этилмеркоптанов в сумме, fraction 200 – выход фракций при 200°C, fraction 300 – выход фракций при 300°C

Видно, что такие параметры, как выход фракций при 200°C и 300°C, характеризуют месторождения 1 и 2, в то время как массовая доля сероводорода и значение плотности особенно значимы для месторождения 5.

Факторный дискриминантный анализ

Следующим этапом работы стало построение классификационной модели для установления принадлежности новых образцов к месторождению по географическому положению. С этой целью использовали факторный дискриминантный анализ – один из классификационных методов с обучением [5, 10, 11].

Обучающий набор образцов (1975 объектов) использован для построения модели классификации, с помощью которой новый образец может быть отнесен к конкретному месторождению. Результаты классификации в виде матрицы неточностей (*confusion matrix*) для обучающего набора данных приведены в табл. 1. Общий процент правильных классификаций с учетом всех 5 групп составил более 99%. Только 17 образцов были неправильно распознаны: 16 образцов, фактически принадлежащие ко 2-му месторождению, были классифицированы как объекты 4-го месторождения, а образец 1 из 4-го месторождения ложно отнесен к 1-му месторождению (см. табл. 1). Следует отметить, что дискриминантный анализ обеспечивает большую точность разделения групп месторождений, чем МГК, обеспечивая также полное разделение кластеров 1-го и 2-го месторождений.

Таблица 1

Матрица неточностей по результатам классификации объектов из обучающего набора методом ФДА

	1	**2**	**3**	**4**	**5**
1	456	0	0	0	0
2	0	454	0	16	0
3	0	0	493	0	0
4	1	0	0	224	0
5	0	0	0	0	331

Примечание. По горизонтали – фактические, по вертикали – предсказанные группы.

Очевидно, что построенная нами модель нуждается в полноценной проверке. Для этого выбран метод тест-валидации, так как объем выборки достаточен для проверки такого типа. В табл. 2 представлены результаты отнесения объектов из тестового набора к пяти группам месторождений. Как и в случае обучающего набора данных, наибольшая неопределенность

существует в отнесении между 2-м и 4-м месторождением (принадлежность 7 из 246 объектов второй группы была ложно предсказана). Процент точных предсказаний составил 97% для второй группы, в то время как средняя точность метода по всем группам достигла почти 100% (см. табл. 2).

Таблица 2

Матрица неточностей по результатам классификации объектов из проверочного набора методом ФДА

	1	2	3	4	5
1	230	0	0	0	0
2	0	239	0	7	0
3	0	0	232	0	0
4	0	1	0	131	0
5	0	0	0	0	148

Примечание. По горизонтали – фактические, по вертикали – предсказанные группы.

Заключение

Таким образом, нами показано, что хемометрические методы (МГК, ФДА) могут быть использованы для определения географического положения месторождения нефти на основании совместного моделирования численных значений 9 параметров, характеризующих качество нефти. Для идентификации неизвестного образца нефти использование указанного подхода требует проведения большого объема работы по многократному измерению каждого из идентификационных параметров. Однако, имея в виду хорошую точность предсказания, полученную для тестового набора, метод может быть рекомендован для рутинного контроля географического происхождения нефти Самарской области.

Работа выполнена в рамках государственного задания Минобрнауки России (проект № 4.1708.2014К).

Список литературы

1. *Вигдергауз М. С.* Аналитическая химия нефти. Куйбышев : Куйбыш. гос. ун-т, 1990. 27 с.
2. *Семенов В. А.* Экоаналитическая идентификация источников загрязнений нефтяными углеводородами // Разведка и охрана недр. 2005. № 5. С. 57–61.
3. *Cordella C. B. Y., Bertrand D.* SAISIR : A new general chemometric toolbox // Trends Anal. Chem. 2014. Vol. 54. P. 75–82.
4. *Wold S., Esbensen K., Geladi P.* Principal component analysis // Chemom. Intell. Lab. Syst. 1987. Vol. 2. P. 37–52.
5. *Benzecri J. P.* Analyse Discriminante et Analyse Factorielle // Les Cahiers de l'Analyse des Donnees. 1977. Vol. 2. P. 369–406.
6. *Родионова О. Е., Померанцев А. Л.* Хемометрика: достижения и перспективы // Успехи химии. 2006. Т. 75, № 4. С. 302–321.
7. *Monakhova Y. B., Kuballa T., Leitz J., Andlauer C., Lachenmeier D. W.* NMR spectroscopy as a screening tool to validate nutrition labeling of milk, lactose-free milk, and milk substitutes based on soy and grains // Dairy Sci. Technol. 2012. Vol. 92. P. 109–120.
8. *Macnaughtan Jr. D., Rogers L. B., Wernimont G.* Principal-component analysis applied to chromatographic data // Anal. Chem. 1972. Vol. 44. P. 1421–1427.
9. *Gergen I., Harmanescu M.* Application of principal component analysis in the pollution assessment with heavy metals of vegetable food chain in the old mining areas // Chem. Central J. 2012. Vol. 6. P. 156–162.
10. *Mouly P. P., Arzouyan C. R., Gaydou E. M., Estienne J. M.* Differentiation of citrus juices by factorial discriminant analysis using liquid chromatography of flavanone glycosides // J. Agric. Food Chem. 1994. Vol. 42. P. 70–79.
11. *Hammamia M., Rouissia H., Salaha N., Selmia H., Al-Otaibib M., Bleckerc C., Karoui R.* Fluorescence spectroscopy coupled with factorial discriminant analysis technique to identify sheep milk from different feeding // Food Chem. 2010. Vol. 122. P. 1344–1350.