Сравнение предиктивных методов БИК-спектроскопии для анализа качества нефтепродуктов

В.О. Кошевой, К.В. Вишневский, И.А. Пронченков, Е.А. Чернышева (ФГАОУ ВО РГУ нефти и газа (НИУ) им. И.М. Губкина), И.И. Салахов, А.В. Зурбашев (АО «ТАНЕКО»)

Использование спектроскопии ближнего инфракрасного диапазона (БИК) и методов математического моделирования является актуальным способом определения и прогнозирования свойств нефтяных и углеводородных смесей различного состава. Для прогнозирования физико-химических свойств дизельных топлив построены регрессионная модель ПЛС (PLS) и искусственная нейронная сеть ИНС (ANN). Показана возможность использования ИНС для решения практических задач хемометрического анализа с использованием БИК-спектроскопии. Приведено сравнение моделей ПЛС и ИНС по параметрам эффективности предсказания.

Ключевые слова: БИК-спектроскопия, дизельные топлива, хемометрика, регрессионный анализ, метод-ПЛС (PLS), искусственные нейронные сети (ИНС, ANN).

Введение

Для анализа нефтепродуктов применяются различные прямые инструментальные методы оценки их физико-химических свойств и характеристик, однако развитие электронно-вычислительной техники дало начало новым методам анализа — хемометрическим. Согласно Д. Массарту (D. Massart), хемометрика — это химическая дисциплина, применяющая математические, статистические и другие методы, основанные на формальной логике, для построения или выбора оптимальных методов измерения и планов эксперимента, а также для извлечения наиболее важной информации при анализе экспериментальных данных. Сейчас хемометрические способы анализа тесно связаны с методами, основанными на взаимодействии вещества с электромагнитным излучением, то есть спектроскопией. Наиболее ярко эта тенденция проявилась при использовании инфракрасной (ИК) спектроскопии, особенно в ближней области (БИК). Ближняя инфракрасная спектральная область охватывает диапазон длин волн в диапазоне 780...2500 нм (диапазон волновых чисел от 12800 $см^{-1}$ до 4000 $см^{-1}$). В спектрах БИК области представлены главным образом обертоны колебаний C–H, N–H, O–H и S–H и комбинации основных типов колебаний средней инфракрасной области.

Спектроскопия ближнего инфракрасного диапазона (БИК) является альтернативой стандартным методикам определения свойств различных химических веществ и их смесей. Основными преимуществами метода являются быстрота, дешевизна и недеструктивность, то есть возможность исследования образца без прямого воздействия на него. Метод доказал свою эффективность в лабораторной практике и в промышленном применении. Широкое распространение БИК-спектроскопия получила в фармацевтике, пищевой промышленности, а также в нефтепереработке и нефтехимии [1–2]. Перспективы применения метода в нефтяной промышленности выглядят многообещающими [3] из-за специфики анализируемых сред и задач анализа. Любое свойство продуктов нефтепереработки и нефтехимии определяется их химическим составом, который описывается большим числом углеводородов, индивидуальное определение которых не всегда представляется возможным. Сейчас доступны методики определения детального углеводородного состава бензиновых фракций (DHA), что позволяет рассчитать физико-химические свойства смеси, используя информацию баз данных о свойствах индивидуальных соединений. В подавляющем большинстве случаев идентификация каждого соединения не представляется возможной, однако применяя хемометрические методы и БИК-спектроскопию можно охарактеризовать всю систему в целом. Моделирование сложных и высокоразмерных данных БИК-спектроскопии позволяют исследователю найти связь между переменными и помочь в решении задач классификации и регрессии. Так, в [4] сообщается об использовании БИК спектроскопии для определения некоторых важных свойств биодизельного топлива: йодного числа, предельной температуры фильтруемости, кинематической вязкости и плотности. Для разработки калибровочных моделей между аналитическими и спектральными данными используется метод главных компонент (PCA) для качественного анализа спектров и проекции на латентные структуры PLS. Авторы указывают, что ИК-спектроскопия в сочетании с многомерной калибровкой является перспективным методом, применяемым для контроля качества биодизеля как в лабораторных, так и в промышленных условиях. Исследование [5] демонстрирует применение многомерной калибровки, основанной на методе генетических обратных наименьших квадратов (GILS), для спектроскопического определения в БИК области нескольких параметров дизельного топлива: цетановое число, температура кипения и замерзания, общее содержание ароматических веществ, вязкость и плотность. В [6] описано применение нейронных сетей и БИК-спектроскопии для предсказания свойств дизельного топлива. Исследуется производительность однослойной искусственной нейронной сети (ИНС), которая

обучалась с помощью метода Левенберга-Марквардта (SLM) и метода масштабируемых сопряженных градиентов (SCG). Предложенные методы сравнивались с многослойной сетью, которая обучалась на тех же алгоритмах. Результаты показывают, что использование ИНС позволяет прогнозировать свойства дизельного топлива с использованием БИК спектра без сокращения данных и достигать большей точности в прогнозировании.

Рассмотрим практическую реализацию методов проекции на латентные структуры и ИНС.

Объекты и методы исследования

Исследование проводилось для 200 образов товарного зимнего дизельного топлива, полученных на узле смешения АО «ТАНЕКО». Основными потоками на станции смешения являются керосиновые и дизельные фракции установок гидроочисток и гидрокрекинга. Соотношение сырьевых потоков определяется нуждами завода и потребностями рынка, что обеспечило широкую аналитическую базу для исследования. В основу исследования были положены анализы топлив, а именно, температура вспышки и помутнения, производимые центральной заводской лабораторией, наряду с полученными БИК-спектрами. Поточный спектрометр BRUKER MATRIX-F позволил получать спектры в режиме online строго в момент отбора проб на анализ.

Дальнейшая обработка спектров и создание калибровочных моделей происходила двумя путями:

1) в программном обеспечении OPUS PROCESS и OPUS QUANT II;

2) при помощи алгоритмов нормализации и ИНС, разработанных и написанных на языке программирования Python.

Программный пакет OPUS QUANT II предназначен для количественного анализа спектров, для которых характерно значительное перекрывание полос поглощения, наблюдаемое в многокомпонентных смесях. В программе реализован рассмотренный выше метод PLS. Отличительной особенностью программы QUANT II является автоматическое нахождение спектральных областей, наиболее коррелирующих с референтными данными.

Таблица 1. Показатели качества работоспособности моделей

Показатель \ свойство дизельного топлива	Температура вспышки		Температура помутнения	
Модель	№1	№2	№3	№4
Число главных компонентов, ед.	5	5	7	7
Среднеквадратичная погрешность перекрестной проверки (RMSECV)	3,01	1,94	4,19	2,15
Среднеквадратичная погрешность предсказания (RMSEP)	3,09	2,11	2,66	1,58
Коэффициент детерминации (R^2) калибровочного набора	69,19	88,6	86,64	96,42
Коэффициент детерминации (R^2) тестового набора	70,12	86,6	93,73	97,85

Проверка работоспособности полученной модели осуществлялась тестовым набором. Данный метод проверки используется для больших баз данных и позволяет оценить их применимость в реальных условиях.

Для оценки прогнозирующих свойств калибровочной модели были использованы следующие критерии.

• Коэффициент детерминации R^2, показывающий процент дисперсии между истинными и предсказанными значениями компонентов. R^2 приближается к 100% по мере того, как предсказанные значения приближаются к истинным.

• Среднеквадратичная погрешность перекрестной проверки (RMSECV). Используется при перекрестной проверке в качестве критерия оценки метода и позволяет снизить явные ошибки при создании модели путем исключения из обучающего набора наиболее отклоняющихся значений.

• Среднеквадратичная погрешность предсказания (RMSEP), используемая при проверке тестового набора. Служит для оценки работы модели на новых данных.

Экспериментальная часть

Реализацией метода PLS в ПО QUANT II получено четыре хемометрические модели, позволяющие предсказывать температуру вспышки (модель № 1, модель № 2 с исключениями) и помутнения (модель № 3, модель № 4 с исключениями). В качестве предварительной обработки спектральных данных использовался метод векторной нормализации. В процессе перекрестной проверки при построении моделей № 1 и № 3 были замечены и рекомендованы к исключению ПО QUANT II образцы, имеющие крайне большое значение отклонения. Модели № 2 и № 4 были построены после исключения шести образцов из калибровочного набора и одного из тестового, в результате чего среднеквадратичные погрешности сократились в 1,5...2,0 раза. Результаты оценки моделей представлены в табл. 1.

Отметим, что в ходе регрессионного анализа методом PLS для описания моделей, прогнозирующих температуру вспышки и помутнения, понадобилось разложение на 5 и 7 главных компонент соответственно. Увеличение числа главных компонент повышает сложность построенной модели, что говорит о большом числе факторов, влияющих на то или иное свойство [7]. На рис. 1–2 представлено графическое отображение работоспособности моделей (графики прогноз-измеренное) в ходе перекрестной проверки калибровочного набора и результаты тестового набора.

Для обучения ИНС использовались те же данные, что и для метода PLS: в качестве входных параметров сети использовался спектр, где каждая точка была значением активации нейрона входящего слоя.

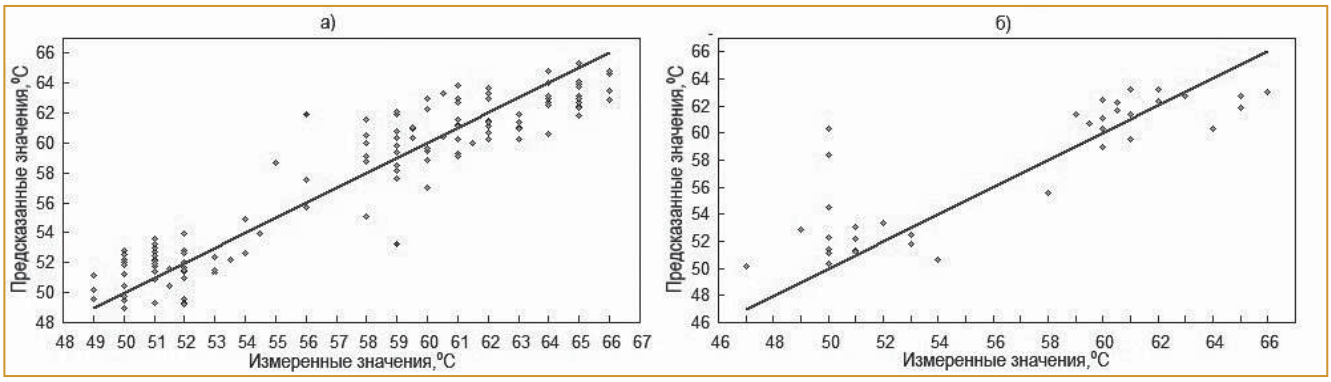

Рис. 1. Графики прогноз-измеренное калибровочного (а) и тестового набора (б) для температуры вспышки

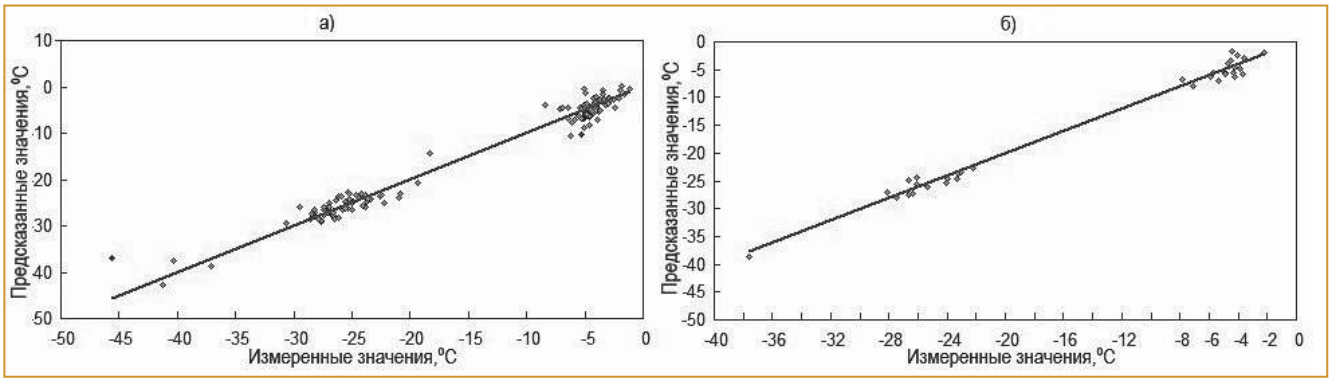

Рис. 2. Графики прогноз-истинное калибровочного (а) и тестового набора (б) для температуры помутнения

Каждый спектр представляет собой вектор длиной 2074 элемента. Выходным слоем сети является значение предсказываемого параметра — температуры вспышки или помутнения. Задача предсказания свойства по спектру топлива в терминологии ИНС носит название регрессионного анализа. В ходе такого анализа формируется значение активации нейрона выходящего слоя с линейной активацией.

Для предобработки спектральных данных использовался метод min-max нормализации. Указанный метод позволяет сократить утраченную информативность спектра по сравнению с другими методами, например, с производной.

Для построения модели была выбрана нейронная сеть глубокого обучения (DNN — Deep Neural Network) по 200 нейронов на каждом из трех скрытых слоев с функцией активации ReLU [8] для всех нейронов. В качестве алгоритма оптимизации сети был выбран Adam [9] с параметром обучения learning rate = 0,0001. Функцией оптимизации (loss) была выбрана среднеквадратичная ошибка (MSE). Данная конфигурация показала наилучший результат при обучении и валидации.

В процессе построения ИНС и конфигурации параметров сети используется разделение обучающего множества на три части: тренировочное множество — 60%, валидационное множество — 20% и тестовое множество — 20%. Процесс обучения происходил по эпохам. Перед началом каждой эпохи тренировочное множество перемешивается, затем каждый спектр подается на вход ИНС, после чего вычисляется значение loss-функции на основании ассоциированного со спектром анализа и происходит обратное распространение ошибки, в процессе которого

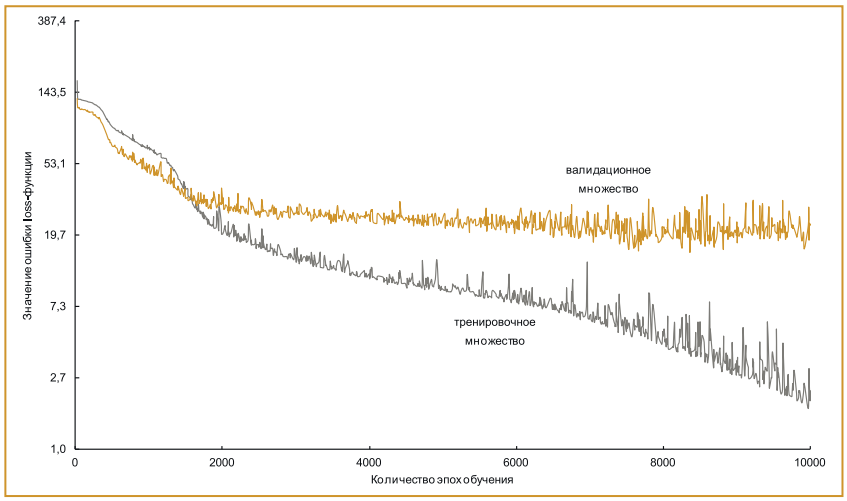

Рис. 3. Периодизация значений ошибки loss-функции для тренировочного и валидационного множеств

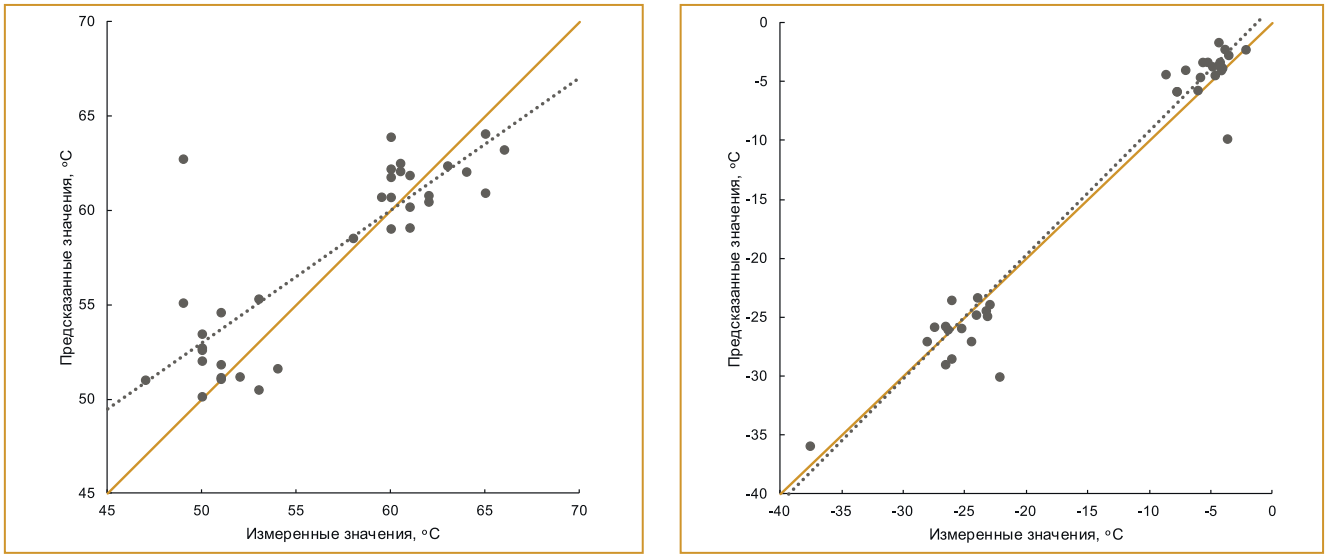

Рис. 4. Графики прогноз-измеренное для температуры вспышки и помутнения соответственно

сеть обучается. В конце эпохи на вход сети подается валидационное множество и вычисляется значение loss-функции на нем. В процессе создания модели обращается внимание на значения loss-функций тренировочного и валидационного множества. Характер их кривых на графике указывает насколько хорошо сеть описывает данные и позволяет избежать переобучения или недообучения.

Выбор числа эпох обучения зависит от изменения значения loss-функций в процессе обучения. По графику видно, что, начиная с 8000 эпох, значение loss-функции на валидационном множестве не уменьшается, а значение loss-функции на тренировочном продолжает падать, что говорит о переобучении сети, то есть она начинает слишком хорошо описывать данные, на которых обучается. Для устранения этого обучение длится 8000 эпох, графическое отображение представлено на рис. 3. Время обучения ИНС до 8000 эпох на графическом процессоре GPU составило 10 мин.

Конечным этапом является проверка модели на тестовом множестве. На рис. 4 представлены предсказанные значения температур помутнения и вспышки относительно измеренных. Для решения каждой задачи и определения соответствующего свойства дизельного топлива сеть обучалась заново с теми же параметрами.

Для детектирования экстремально «выбивающихся» из модели прогноза образцов была произведена проверка нормальности распределения ошибки RMSE на предмет выбросов. Диаграммы размаха представлены на рис. 5 и 6 для температуры вспышки и помутнения соответственно. Анализ нормальности проводился программно, визуализация представлена графически с помощью ПО Statistica.

Исключение экстремальных (Extremes) образцов из общего набора данных, выходящих за пределы ±2SD (стандартное отклонение) от Mean (среднего), позволило получить гораздо лучшие характеристики модели (рис. 7).

Основные параметры модели приведены в виде табл. 2.

Для оценки влияния числа образов в обучающей выборке на точность предсказания модели было создано шесть различных по мощности обучающих

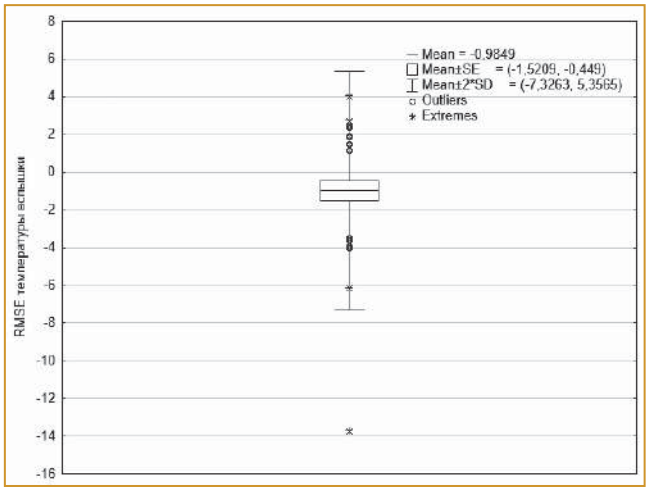

Рис. 5. Диаграмма размаха RMSE температуры вспышки

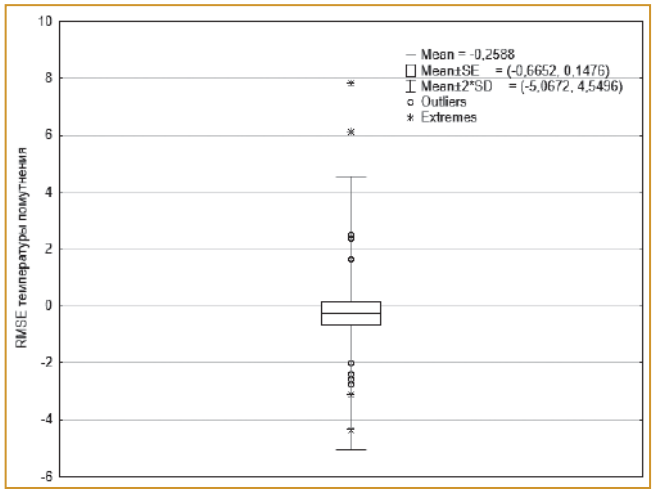

Рис. 6. Диаграмма размаха RMSE температуры помутнения

Таблица 2. Показатели качества моделей, полученных с помощью ИНС

Показатель \ свойство дизельного топлива	Температура вспышки		Температура помутнения	
Тестовая выборка	Исходная	Без экстремальных	Исходная	Без экстремальных
Среднеквадратичная погрешность предсказания (RMSEP)	3,28	2,04	2,38	1,59
Коэффициент детерминации (R^2) тестового набора	70,14	87,46	95,96	98,58

множеств. Параметры нейронной сети в процессе обучения не изменялись и соответствовали ее конечной версии. Влияние размера обучающего множества на среднеквадратичную погрешность (RMSE) модели, построенную для прогнозирования температуры помутнения, показано на рис. 8, на котором видно, что ошибка предсказания снижается с увеличением обучающего множества.

ИНС была написана на языке программирования Python 3.7.4 с использованием библиотеки Tensorflow 2.0 [10], которая позволяет декларативно описывать структуру сети, обучать модель и выполнять предсказания.

Обсуждение результатов

Производство товарных дизельных топлив осуществляется путем компаундирования различных продуктовых потоков в статических и динамических смесителях. Соотношение компонентов для достижения заданных свойств смеси подбирается эмпирическим путем ввиду того, что изменение свойств и показателей качества топлива в процессе компаундирования носит нелинейный характер. Определение этих свойств — рутинная операция, требующая достаточно большого количества времени. Установка поточных анализаторов, нацеленных на измерение единичных параметров не всегда возможна, ввиду требуемой пробоподготовки, сложности анализа и дороговизны оборудования, например в случае хроматографии. Число таких анализаторов часто равно числу определяемых параметров, что еще больше осложняет задачу контроля качества. Использование хемометрических моделей и одного поточного БИК-спектрометра позволяет измерять в режиме on-line сразу несколько свойств. Цикл таких измерений повторяется с периодом 1 мин. Это время необходимое для заполнения и термостатирования измерительной ячейки, измерения спектральных данных и прогностических расчетов. В этом случае наибольший интерес представляют именно методы математического моделирования. К таким методам относится классические, например, PLS, и бурно развивающиеся ИНС. Построенные в этой работе модели для прогнозирования температуры вспышки и помутнения хорошо согласуются со входными спектральными данными на перекрестной проверке и дают четкий прогноз этих показателей на тестовых наборах. В табл. 3 приведены сравнительные характеристики конечных моделей PLS и ИНС.

Обе модели показывают отличный результат, который укладывается в интервал воспроизводимости по стандартной методике ГОСТ 6356-75 и ГОСТ 5066-2018 соответственно.

ИНС имеют огромный потенциал в области анализа, когда влияние входящих данных различной природы на результат неравномерное, а их взаимосвязь не может быть описана классическими алгоритмами. Здесь следует отметить большую гибкость ИНС в регрессионных задачах прогноза. Сеть имеет множество параметров конфигурации, начиная от числа скрытых слоев, заканчивая выбором функции активации нейронов. От композиции параметров напрямую зависит качество модели и ее робастность. Устойчивость системы при описании входящих данных, полученных из иных источников, также может быть повышена за счет увеличения тренировочного набора образцами с других НПЗ. Такой подход предотвратит «переобучение» модели на однородных данных и придаст ИНС свойство универсальности. Проверка обеих моделей на данных спектрального анализа образцов другого потока показала, что ошибка RMSEP для ИНС меньше, чем у метода ПЛС. Изменение свойств тестовой выборки показывает большую устойчивость ИНС к отклонениям относительно тренировочного набора.

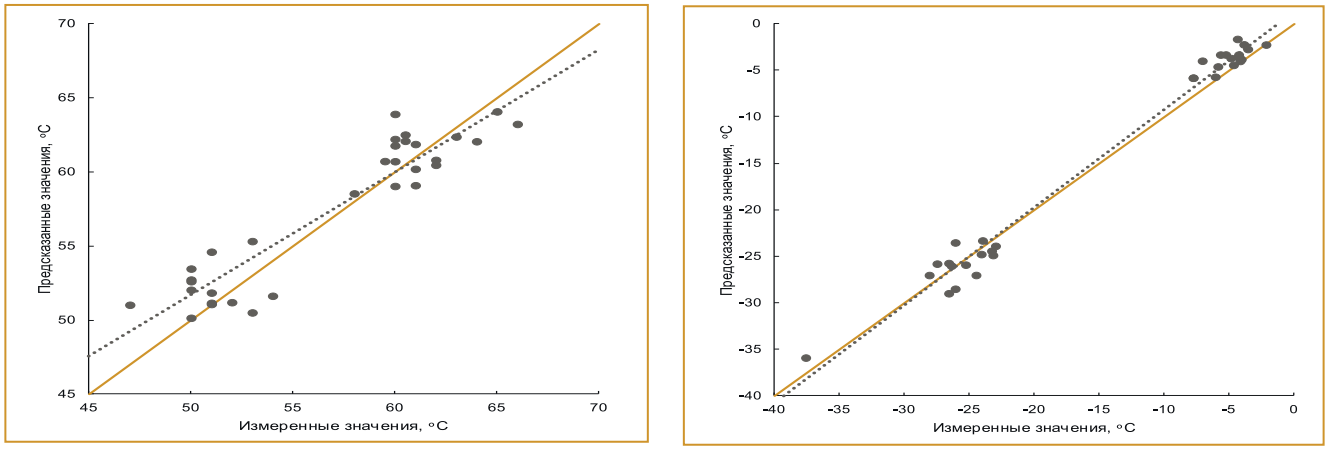

Рис. 7. Графики прогноз-измеренное для температуры вспышки и помутнения соответственно после исключения экстремальных образцов

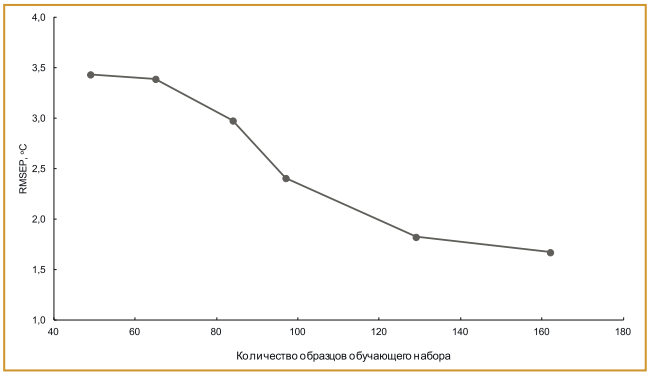

Рис. 8. Зависимость ошибки предсказания температуры помутнения от размера обучающего набора

Таблица 3. Сравнение точности предсказания моделей PLS и ИНС

Показатель	Температура вспышки		Температура помутнения	
Предсказываемый показатель	Метод PLS	ИНС	Метод PLS	ИНС
R^2, %	88,60	87,46	96,42	98,58
RMSEP, °C	2,11	2,04	1,58	1,59
Воспроизводимость по стандартной методике, °C	4*		3*	

* - при доверительном интервале 95%

В ходе исследования установлено, что ИНС лучше описывают косвенные характеристики нефтепродуктов, в том числе восприимчивость к различным добавкам и присадкам.

Отметим, что современное ПО позволяет обучать ИНС на огромном массиве данных за разумное время с использованием специальных технических решений, например, обучение с использованием графических ускорителей (GPU), а затем легко внедрить обученную модель в любую инфраструктуру. Важнейшей особенностью использования ИНС является наличие свободного программного обеспечения для их создания. Такое ПО бесплатно даже для коммерческого использования.

Заключение

В настоящий момент при приготовлении товарных нефтепродуктов используются многочисленные, не всегда успешно совместимые потоки с различными, постоянно изменяющимися физико-химическими характеристиками, анализ которых требует применения специального оборудования и методов, а также значительного времени исследования. Использование БИК-спектроскопии является актуальным способом определения и прогнозирования свойств нефтяных и углеводородных смесей различного состава. Метод доказал свою эффективность в лабораторной практике и в промышленном применении.

Представленные в работе предиктивные модели на основе методов ПЛС и ИНС глубокого обучения с использованием БИК-спектроскопии демонстрируют высокую точность определения свойств товарного дизельного топлива в режиме on-line. Выявлено, что модели, построенные на одинаковых исходных данных и обученные в идентичных условиях, показывают высокую степень корреляции данных для потока, на котором проходило обучение. В то же время отметим, что ИНС в большей степени способна на точный прогноз в случае существенного изменения углеводородного состава исследуемых топлив.

ИНС является мощным методом хемометрического моделирования на основе БИК-спектроскопии для прогнозирования свойств нефтепродуктов в условиях постоянного изменения состава и физико-химических характеристик сырьевых потоков.

Список литературы

1. *D. Wu, X. Chen, P. Shi, S. Wang, F. Feng, and Y.* He.Determination of α- linolenic acid and linoleic acid in edible oils using near-infrared spectroscopy improved by wavelet transform and uninformative variable elimination, Analytica Chimica Acta, №2,2009.
2. *D. Wu, Y. He, P. Nie, F. Cao and Y. Bao.* Hybrid variable selection invisible and near-infrared spectral analysis for non-invasive quality determination of grape juice, Analytica Chimica Acta, 2010, №2.
3. *R. Balabin, R. Safieva and E. Lomakina.* Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques, Analytica Chimica Acta, 2010, №1.
4. *P. Baptist,P. Felizardo, J.C.Menezes, M. J.N. Correia.* Multivariate near infrared spectroscopy models for predicting the iodine value, CFPP, kinematic viscosity at 40 °C and density at 15 °C of biodiesel, Talanta, 2008.
5. *D. Özdemir.* Near Iinfrared Spectroscopic Determination of Diesel Fuel Parameters Using Genetic Multivariate Calibration", Petroleum Science and Technology, 2008, №1.
6. *H. Ali Gamal Al-Kaf.* Predictive Model and Near Infrared Spectroscopy in Predicting the Diesel Fuel Properties, Master`s thesis, Universiti Tun Hussein Onn Malaysia, 2018.
7. *Pentland A. and Moghaddam B. and Starner T.* ViewbasedModular Eigenspaces for Face Recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern recognition, 1994, №1.
8. *G. Xavier, A. Bordes, Y. Bengio.* Deep spasrse rectifier neural networks, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics 2011.
9. *D.P. Kingma, J. Lei Ba.* Adam: A method for stochastic optimization, International Conference on Learning Representation, 2014.
10. Библиотека для машинного обучения [Электронный ресурс]. URL: https://www.tensorflow.org/ (дата обращения 20.11.2019).

Кошевой Виктор Олегович *– старший инженер-технолог АО ИПТ «Оргнефтехимзаводы»,*
Вишневский Кирилл Вячеславович *– руководитель группы ООО "ОЗОН ТЕХНОЛОГИИ",*
Пронченков Иван Александрович *– инженер,* ***Чернышева Елена Александровна*** *– канд. хим. наук, проф., зам. зав. кафедрой ФГАОУ ВО РГУ нефти и газа (НИУ) им. И.М. Губкина,*
Салахов Ильшат Илгизович *– канд. техн. наук, ген. директор,*
Зурбашев Алексей Владимирович *– зам. ген. директора по технической поддержке и качеству АО "ТАНЕКО".*
E-mail:koshevoy.vik@yandex.ru kvvishnevskiy@gmail.com