ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ХИМИЧЕСКОЙ ТЕХНОЛОГИИ, 2025, том 59, № 2, с. 121–138

УДК 66.012-52

МЕТОД ПОСТРОЕНИЯ МОДЕЛЕЙ ДЛЯ ОЦЕНКИ ПОКАЗАТЕЛЕЙ КАЧЕСТВА ПРОДУКТОВ КОЛОННЫ ФРАКЦИОНИРОВАНИЯ В УСЛОВИЯХ МАЛОГО ОБЪЕМА ДАННЫХ АНАЛИТИЧЕСКОГО КОНТРОЛЯ

© 2025 г. А. А. Плотников, Д. В. Штакин, О. Ю. Снегирев, А. Ю. Торгашов*

Институт автоматики и процессов управления ДВО РАН, Владивосток, Россия

**e-mail: torgashov@iacp.dvo.ru*

Поступила в редакцию 09.12.2024
После доработки 19.03.2025
Принята в печать 31.03.2025

Рассматривается задача повышения точности моделей для оценки показателей низкотемпературных свойств, показателей воспламеняемости и противоизносных свойств целевых продуктов колонны фракционирования в условиях малого объема данных аналитического контроля. Для решения рассматриваемой задачи предложен метод построения моделей, в составе которого используется алгоритм расширения малой обучающей выборки по данным фракционного состава, отличающийся способом отбора дополнительных данных, учитывающим показатель разреженности, что позволило включить в обучающую выборку недостающее количество данных, и в итоге обеспечить повышение качества модели. Использование предложенного метода позволило повысить точность моделей в среднем на 18% в сравнении с известными методами и в среднем на 6% в сравнении с методом на основе расширения обучающей выборки без учета показателя разреженности. Результаты представлены на примерах построения моделей показателей качества предельной температуры фильтруемости, температуры вспышки, кинематической вязкости при 40°С и цетанового числа среднего дистиллята (фракции дизельного топлива) и температуры вспышки керосиновой фракции промышленной колонны фракционирования технологической установки гидрокрекинга.

Ключевые слова: математические модели для оценки показателей качества нефтепродуктов, ректификация, малая выборка, расширение выборки, разреженность, аналитический контроль

DOI: 10.31857/S0040357125020111 **EDN**: ndwfda

ВВЕДЕНИЕ

В современных системах управления химико-технологическими процессами широко применяются математические модели для оценки показателей качества (**ПК**) получаемых продуктов [1]. Данные модели применяются в режиме реального времени для оценки и прогнозирования труднодоступных ПК, например, низкотемпературных свойств нефтепродуктов, основываясь на значениях легкодоступных параметров технологического процесса (**ТП**), таких как значения температурного профиля, давления, расходов потоков в технологическом аппарате. Разработка и внедрение моделей для оценки ПК независимо от их типа требуют накопления качественных данных [2]. Следует отметить, что в реальных условиях некоторые ПК продуктов могут определяться сравнительно редко, что приводит к малому объему накопления данных аналитического контроля. Построение моделей в условиях малой выборки может часто приводить к переобучению модели [3], невозможности использования достаточного количества входных переменных [4] и выявления зависимостей между переменными [5], что негативно сказывается на их точности. Поэтому при работе с малыми обьемами данных необходимо адаптировать традиционные подходы для улучшения качества оценивания на тестовом сегменте данных [3]. Использование малого объема накопленных данных с меньшим возможным числом степеней свободы модели также приводит к необходимости построения более простой модели и ограничивает возможность извлечения зависимостей из данных [6]. Понятие малой выборки связано с количеством полезной информации, которую можно извлечь из данных в решаемой задаче [7]. В случае построения моделей для оце-

нивания показателей качества продуктов сложной ректификационной колонны, учитывая повторяемость технологических режимов объекта, авторами принято считать малой выборку, содержащую менее 100 наблюдений.

Известны работы, основной задачей которых для решения данной проблемы является разработка и модификация различных математических методов построения моделей [8]. Более широкое распространение получил подход на основе добавления сгенерированного набора данных (*VSG – Virtual Sample Generation*) [8, 9] к малой обучающей выборке (**ОВ**). Однако данный подход может быть неэффективным, если связи в синтетических данных отличаются от связей в реальных [10]. Сгенерированный набор данных может быть получен от откалиброванной строгой (англ. *rigorous*) модели, учитывающей физико-химические закономерности ТП [11]. Ввиду того, что использование строгих моделей не всегда возможно, в особенности для оценки низкотемпературных свойств, чаще применяются алгоритмы расширения ОВ, использующие методы машинного обучения. Так, в работе [5] генерирование наборов синтетических данных осуществляется с помощью смещения значений входных переменных путем добавления белого шума и последующим объединением нескольких выборок для увеличения итогового разнообразия данных. В работах [8, 12] описано получение синтетических данных с помощью генеративно-состязательных нейронных сетей. В работе [10] предложен метод расширения ОВ с использованием эволюционных алгоритмов для получения синтетических данных с сохранением нелинейных зависимостей. Работа [13] посвящена обзору непараметрических методов построения моделей с использованием функций ядер, отмечено преимущество данной группы методов при работе с малыми объемами данных.

При расширении ОВ следует отметить актуальность задачи определения достаточного количества дополнительных данных [10, 14], так как добавление избыточного количества синтетических наблюдений в ОВ может привести к снижению точности модели. С понятием малой выборки также связан термин разреженность данных (англ. *data sparsity*), используемый для обозначения широких интервалов между наблюдаемыми значениями в границах диапазона их изменения. Высокая разреженность данных также негативно сказывается на точности моделей [15], поэтому разреженность используется для обозначения недостаточного количества наблюдений при построении статистической модели.

В настоящей работе рассматривается промышленная колонна фракционирования установки гидрокрекинга, целевыми продуктами которой являются средний дистиллят и керосиновая фракция (**КФ**). В ходе аналитического контроля для указанных продуктов в среднем 1–2 раза в сутки определяются показатели испаряемости и значительно реже определяются показатели противоизносных свойств, воспламеняемости и показатели низкотемпературных свойств. Последние редкоизмеряемые показатели качества (**РПК**) регламентируются так же, как и часто измеряемые показатели (**ЧПК**) испаряемости, поэтому актуальной является задача повышения точности моделей для оценивания РПК в условиях малого обьема накопления данных. Данная задача рассматривается для предельной температуры фильтруемости (**ПТФ**), температуры вспышки, вязкости при 40°С, цетанового числа (**ЦЧ**) среднего дистиллята и температуры вспышки КФ.

В отличие от предыдущих работ, указанных выше, для получения синтетических данных в настоящей работе используется вспомогательная модель, в основе которой лежит зависимость между РПК и ЧПК. На основании исследований в [16–21], подтверждающих корреляции между РПК и ЧПК, выбраны следующие варианты со связями ПК: ПТФ среднего дистиллята и температурами конца кипения, температурой вспышки среднего дистиллята и началом кипения, вязкостью среднего дистиллята при 40°С и температурами выкипания в диапазоне 20–70 об. %, ЦЧ и температурами выкипания в диапазоне 10, 50 и 90 об. % и между температурой вспышки керосина и его началом кипения. Стоит отметить, что непосредственно применять данные корреляции невозможно ввиду необходимости использования всей кривой разгона или других РПК для расчета, а также необходимо учитывать возможные различия в распределении углеводородов (парафины, изопарафины, нафтены и ароматические соединения) при одинаковом фракционном составе.

Ранее в работе [22] предложен метод, в составе которого используется алгоритм расширения ОВ на основе вспомогательной модели оцениваемого ПК целевого продукта с учетом показателя разреженности. Значение показателя разреженности использовано для заполнения как можно большего числа пустых интервалов имеющейся ОВ в диапазоне изменения выходной переменной. В настоящей работе предлагается метод расширения ОВ, отличающийся от [22] заполнением интересующего диапазона синтетическими

данными (со вспомогательной модели) с учетом значения показателя разреженности с последующим отбором выборки по результатам тестирования на начальной известной ОВ. Полученный набор отобранных синтетических наблюдений далее предлагается добавлять к изначальной ОВ. Также приводится сравнение методов построения вспомогательной модели для получения синтетических данных с использованием робастной регрессии (**РР**), нейронных сетей прямого распространения (**НСПР**) и метода ортогональных проекций на скрытые структуры на основе ядра (англ. K-OPLS). При сравнении эффективности данных методов предлагается осуществлять проверку качества получаемых синтетических наблюдений путем использования их в качестве ОВ модели с последующим тестированием ее на начальной известной ОВ. Также результаты расширения ОВ с использованием предлагаемого метода приведены для пяти РПК.

ОПИСАНИЕ ТЕХНОЛОГИЧЕСКОГО ОБЪЕКТА И ПОСТАНОВКА ЗАДАЧИ

Описание технологического объекта. Объектом исследования является сложная колонна фракционирования (рис. 1), в которой происходит разделение подаваемого сырья (стабильного гидрогенизата гидрокрекинга) на тяжелую нафту, КФ, средний дистиллят, тяжелую дизельную фракцию и остаток гидрокрекинга.

Средний дистиллят в зависимости от режима работы установки может использоваться в качестве компонента масляной основы буровых растворов (**МОБР**), компонента зимнего или арктического дизельных топлив (**ДТЗ** и **ДТА** соответственно) путем смешения с КФ. В зависимости от целевого назначения среднего дистиллята диапазон изменения его ЧПК отличается, при этом некоторые РПК могут не определяться в ходе аналитического контроля.

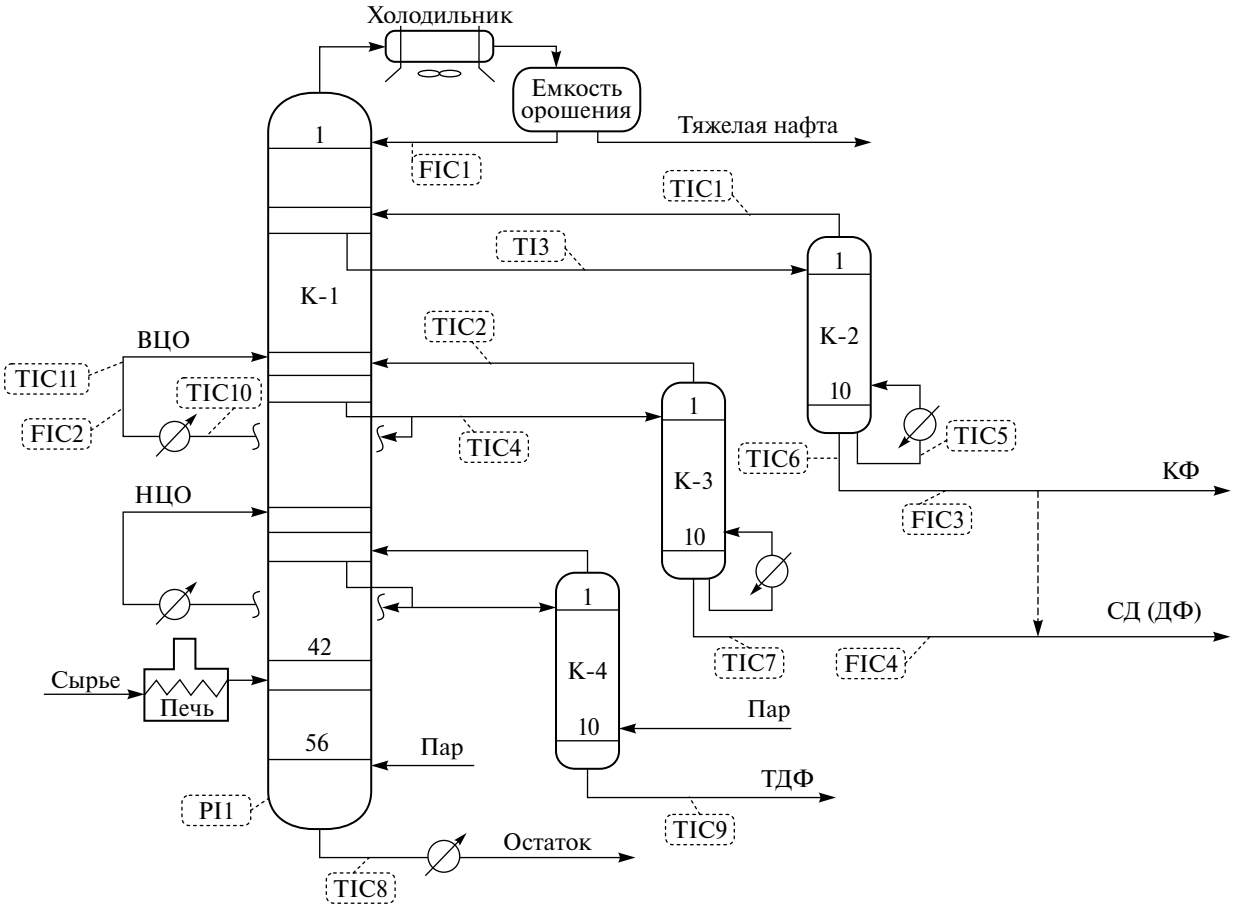

Рис. 1. Схема ТП фракционирования. К-1 – колонна фракционирования, К-2 – колонна отпаривания КФ, К-3 – колонна отпаривания среднего дистиллята, К-4 – колонна отпаривания тяжелого дизельного топлива, КФ – керосиновая фракция, СД – средний дистиллят, ДФ – дизельная фракция, ТДФ – тяжелая дизельная фракция, ВЦО – верхнее циркуляционное орошение, НЦО – нижнее циркуляционное орошение.

Таблица 1. Количество наблюдений в $ОВ_{исд}$ и $ТВ_{исд}$ для исследуемых РПК

ПК	Кол-во наблюдений в $ОВ_{исд}$, шт.	Кол-во наблюдений в $ТВ_{исд}$, шт.
ПТФ среднего дистиллята	34	33
$T_{всп}$ среднего дистиллята	46	45
Вязкость при 40°С среднего дистиллята	41	40
ЦЧ среднего дистиллята	105	105
$T_{всп}$ КФ	78	78

Таблица 2. Входные переменные модели для оценки ПТФ среднего дистиллята

№	Обозначение	Описание	Ед. изм.
1	TIC4	Температура среднего дистиллята из колонны К-1	°С
2	TIC5	Температура куба колонны К-2	°С
3	TIC10	Температура ВЦО	°С
4	FIC1	Расход орошения колонны К-1 нафтой	$м^3/ч$

Таблица 3. Входные переменные модели для оценки температуры вспышки в закрытом тигле среднего дистиллята

№	Обозначение	Описание	Ед. изм.
1	TIC1	Температура паров колонны К-2	°С
2	TIC2	Температура паров колонны К-3	°С
3	TIC7	Температура среднего дистиллята из колонны К-3	°С
4	FIC4	Расход среднего дистиллята	$м^3/ч$

Таблица 4. Входные переменные модели для оценки вязкости среднего дистиллята при 40°С

№	Обозначение	Описание	Ед. изм.
1	TIC11	Температура ВЦО	°С
2	FIC1	Расход орошения колонны К-1 нафтой	$м^3/ч$
3	FIC2	Расход орошения колонны К-1 средним дистиллятом	$м^3/ч$
4	FIC4	Расход среднего дистиллята	$м^3/ч$
5	PI1	Давление в нижней секции колонны К-1	МПа

Таблица 5. Входные переменные модели для оценки ЦЧ среднего дистиллята

№	Обозначение	Описание	Ед. изм.
1	TIC1	Температура паров колонны К-2	°С
2	TIC4	Температура бокового потока среднего дистиллята	°С
3	TIC8	Температура циркулирующего потока куба колонны К-1	°С
4	FIC2	Расход орошения колонны К-1 средним дистиллятом	$м^3/ч$

Таблица 6. Входные переменные модели для оценки температуры вспышки в закрытом тигле КФ

№	Обозначение	Описание	Ед. изм.
1	TIC1	Температура паров колонны К-2	°С
2	TI3	Температура бокового потока КФ	°С
3	TIC6	Температура КФ с куба колонны К-2	°С
4	TIC9	Температура ТДФ с куба колонны К-4	°С
5	FIC1	Расход орошения колонны К-1 нафтой	$м^3/ч$
6	FIC3	Расход КФ с куба колонны К-2	$м^3/ч$

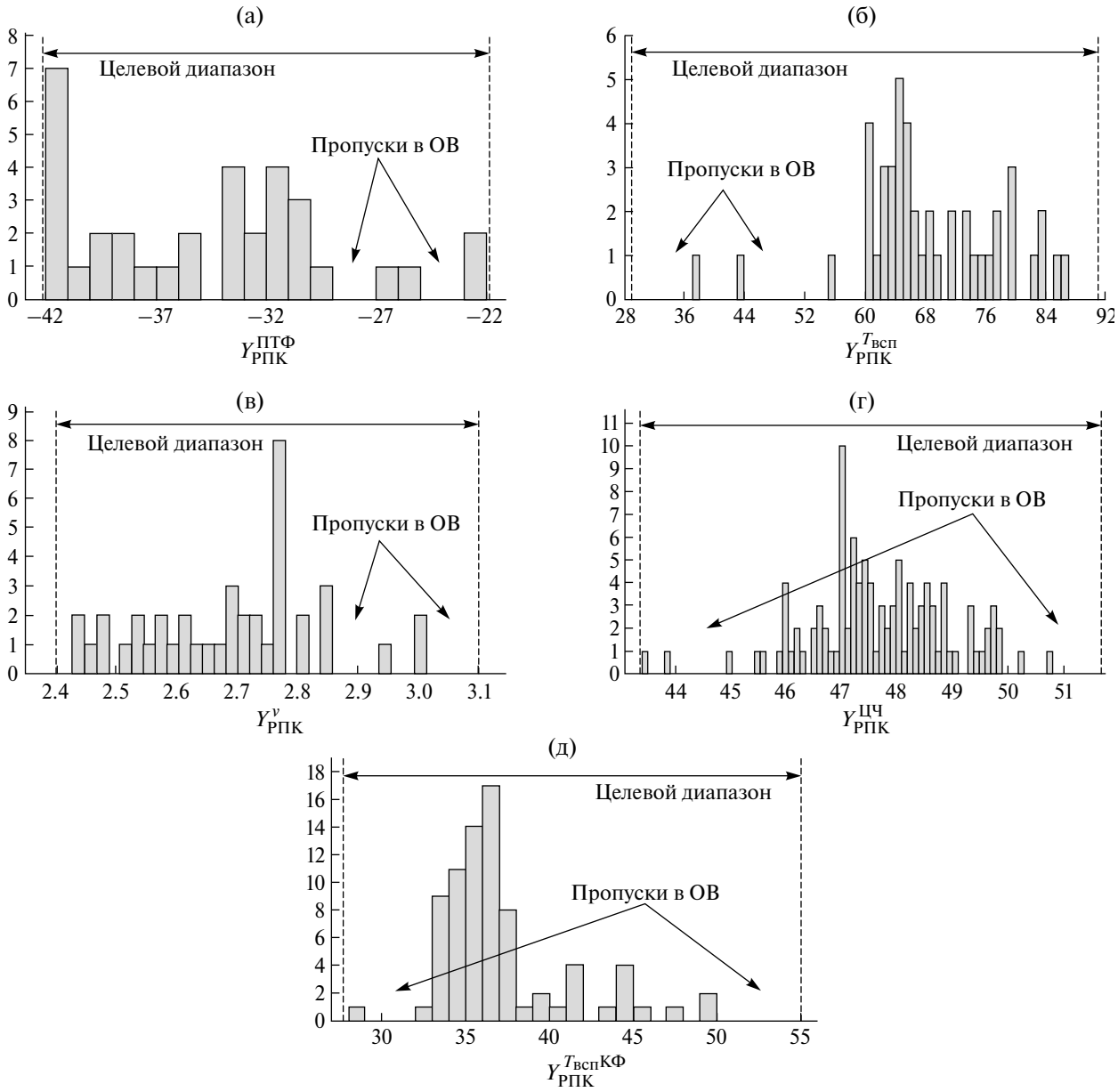

Рис. 2. Гистограммы распределения данных в ОВ: (а) ПТФ среднего дистиллята; (б) $T_{всп}$ среднего дистиллята; (в) вязкость при 40°С среднего дистиллята; (г) ЦЧ среднего дистиллята; (д) $T_{всп}$ керосиновой фракции.

Известный сегмент данных (ИСД). Рассматриваются оценки ПК для потоков среднего дистиллята и КФ. Доступные для анализа наблюдения по каждому РПК среднего дистиллята и КФ делятся последовательно на ОВ и тестовую выборку (**ТВ**). Деление осуществляется в равном соотношении по указанному времени и дате лабораторного анализа (ранний период – $ОВ_{ИСД}$, поздний период – $ТВ_{ИСД}$). Размеры полученных соответствующих обучающих и тестовых выборок представлены в табл. 1.

На рис. 2 представлены гистограммы распределения данных ОВ на целевом диапазоне для всех рассматриваемых РПК. Показаны пропуски в данных, заполнение которых позволит повысить точность оценок за счет учета зависимостей во всем интересующем диапазоне изменения выходной переменной.

В табл. 2–6 представлены входные переменные моделей для оценки ПТФ, температуры вспышки, вязкости при 40°С, ЦЧ среднего дистиллята и температуры вспышки КФ соответственно.

Постановка задачи. Задача заключается в разработке моделей для оценки ПТФ, температуры вспышки, вязкости при 40°С и ЦЧ среднего дистиллята и температуры вспышки КФ с более высокой точностью в условиях малой ОВ в сравнении с моделями, построенными с использованием ИСД.

ИСПОЛЬЗУЕМЫЕ МЕТОДЫ ПОСТРОЕНИЯ МОДЕЛЕЙ ДЛЯ ПОЛУЧЕНИЯ ОЦЕНОЧНЫХ ЗНАЧЕНИЙ РПК

Для построения моделей распространены следующие методы [23–25]: РР; НСПР; метод ортогональных проекций на скрытые структуры на основе ядра, также используемые в хемометрике. Точность моделей в условиях малой ОВ, построенных с использованием данных методов, как правило, не удовлетворяет требованиям в промышленных условиях и нуждается в адаптации и развитии подходов к моделированию. Поэтому в данной работе производится расширение ОВ с целью повышения точности моделей.

Робастная регрессия. Построение линейных регрессионных моделей широко распространено ввиду простоты их реализации и невысоких требований к вычислительным мощностям. Уравнение множественной линейной регрессии имеет вид:

$$\hat{y}_i = b_0 + \sum_{j=1}^{m} b_j x_{i,j}. \tag{1}$$

Свободный коэффициент b_0 и коэффициенты при переменных b_j могут быть оценены технологом на предмет их согласованности по знаку с физико-химическими принципами и скорректированы при построении модели, а также при построении линейной регрессионной модели просто интерпретировать полученный результат.

Также используется РР [23], вектор коэффициентов РР $\boldsymbol{b}_{RR}$ определяется в результате решения следующей системы уравнений:

$$\begin{cases} \sum_{i=1}^{N} \omega_i \left(y_i - \sum_{j=1}^{m} b_{RR,j} x_{i,j} \right) x_{i,1} = 0 \\ \vdots \\ \sum_{i=1}^{N} \omega_i \left(y_i - \sum_{j=1}^{m} b_{RR,j} x_{i,j} \right) x_{i,m} = 0 \,, \end{cases} \tag{2}$$

где m – количество входных переменных. Элементы выборки должны быть нормализованы на нулевое среднее и единичную дисперсию перед использованием в (2). Значения весовых коэффициентов ω_i итерационно определяются в соответствии с весовой функцией, начиная со всех $\omega_i = 1$. Здесь используется весовая функция Файера (англ. *Fair*) [26]:

$$\omega_i = 1 / \left(1 + |r_i|\right). \tag{3}$$

Вектор $\boldsymbol{r}$ значений для расчета весовых коэффициентов зависит от медианного значения модулей отклонений ошибок на предыдущем шаге итерации и определяется по формуле:

$$r_i = \frac{y_i - \sum_{j=1}^{m} b_{RR,j} x_{i,j}}{cA\sqrt{1-h_i}}, \tag{4}$$

$$A = \frac{\operatorname{med}\left(\left|\left(y - \hat{y}\right) - \operatorname{med}\left(y - \hat{y}\right)\right|\right)}{0.6745}. \tag{5}$$

Нейронная сеть прямого распространения. Нейронные сети получили широкое распространение в машинном обучении и хемометрике благодаря своей универсальности. Основными преимуществами нейронных сетей являются возможность работы с большими объемами данных, выявление глубоких зависимостей, что не всегда доступно при использовании других методов, и преобразование исходного сложноустроенного объекта в набор признаков. В данной работе используется однослойная НСПР. Оцениваемое значение выходной переменной вычисляется в соответствии с уравнением:

$$\hat{y}_i = W_2 \left(\varphi\left(W_1 \mathbf{x}_i + \mathbf{p}_1 \right) \right) + p_2, \tag{6}$$

$$\varphi\left(W_1 \mathbf{x}_i + \mathbf{p}_1 \right), \tag{7}$$

$$\mathbf{x}_i = \begin{pmatrix} x_1 & x_2 & \dots & x_m \end{pmatrix}^T, \tag{8}$$

$$\mathbf{p}_1 = \begin{pmatrix} p_1 & p_2 & \dots & p_k \end{pmatrix}^T, \tag{9}$$

$$W_1 = \begin{pmatrix} w_{1,1}^{(1)} & w_{1,2}^{(1)} & \dots & w_{1,m}^{(1)} \\ w_{2,1}^{(1)} & w_{2,2}^{(1)} & \dots & w_{2,m}^{(1)} \\ \dots & \dots & \dots & \dots \\ w_{k,1}^{(1)} & w_{k,2}^{(1)} & \dots & w_{k,m}^{(1)} \end{pmatrix}, \tag{10}$$

$$W_2 = \begin{pmatrix} w_1^{(2)} & w_2^{(2)} & \dots & w_k^{(2)} \end{pmatrix}. \tag{11}$$

В основе однослойной НСПР лежит теорема, доказанная Дж. Цыбенко в 1989 г. [24]. Основное утверждение теоремы – искусственная НСПР с одним скрытым слоем может аппроксимировать любую непрерывную функцию множества переменных с заданной точностью. Согласно работе

Цыбенко: пусть φ – любая непрерывная сигмоидная функция. Если дана любая непрерывная функция действительных переменных $f(x)$, определенная на пространстве $[0,1]^n$, то существуют векторы параметров w_i, α_i, θ_i и такая функция

$$Q(\mathbf{x}_i)=\sum_{j=1}^{k} w_i^{(2)}\varphi\left(\mathbf{w}_i^{(1)}\mathbf{x}_i+\mathbf{p}_i\right), \qquad (12)$$

что для всех $X\in[0,1]^N$ выполняется условие

$$\left|Q(\mathbf{x}_i)-f(\mathbf{x}_i)\right|<\varepsilon_i. \qquad (13)$$

K-OPLS. OPLS – метод множественной линейной регрессии путем построения ортогональных проекций на скрытые структуры. В ходе построения модели матрица входных данных X представляется в виде двух наборов скрытых переменных T_p и T_o:

$$X=T_pP_p^T+T_oP_o^T+E. \qquad (14)$$

Здесь T_p – y-прогнозная матрица вкладов (англ. *score matrix*), P_p^T – y-прогнозная матрица нагрузки (англ. *loading matrix*), T_o – соответствующая y-ортогональная матрица вкладов, P_o^T – соответствующая y-ортогональная матрица нагрузки, E – матрица остатков. Обе матрицы, y-прогнозная и y-ортогональная, описывают свойства смоделированных наблюдений, с помощью которых возможно выявление очевидных и неожиданных трендов, кластеров или выбросов в данных [25].

Kernel-OPLS – модифицированный метод OPLS, в котором добавлено преобразование матрицы X на основе функции ядра, позволяющее рассматривать матрицу ядра как скалярные произведения в пространстве признаков высокой размерности. Произведение XX^T заменяется на матрицу Грама K, где $K_{i,j}=\ker\left(x_i, x_j\right)$, что позволяет избежать отображения X в пространстве более высокой размерности. В качестве функции ядра $\ker(\cdot,\cdot)$ в данной работе использована функция Гаусса:

$$\ker(x,y)=\exp(-\|x-y\|^2/2\sigma^2), \qquad (15)$$

где σ – параметр настройки, нахождение оптимального значения которого позволяет снизить ошибку при обучении и тестировании модели. Среди преимуществ K-OPLS выделяют устойчивость к выбросам и мультиколлинеарности входных переменных. Подробное описание метода приведено в [27].

Критерии оценки результатов. В качестве критериев для оценки полученных результатов использованы: средняя абсолютная ошибка (**САО**) и коэффициент детерминации R^2, которые вычисляются по формулам, приведенным ниже:

$$\text{САО}=\frac{1}{N}\sum_{i=1}^{N}\left|y_i-\hat{y}_i\right|, \qquad (16)$$

$$R^2=1-\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2\Big/\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2. \qquad (17)$$

ПРЕДЛАГАЕМЫЙ МЕТОД ПОСТРОЕНИЯ МОДЕЛЕЙ ДЛЯ ОЦЕНКИ ПК

Структура метода. Структура предлагаемого метода построения моделей для оценки РПК представлена на рис. 3. Сущность метода заключается в расширении ОВ за счет использования оценок от вспомогательной модели, построенной на основании наблюдений ЧПК, например, фракционного состава, за период, соответствующий обучающей выборке известного сегмента данных.

Расширение ОВ происходит путем добавления к $\text{ОВ}_{\text{ИСД}}$ синтетических наблюдений, полученных от вспомогательной модели. Полученные синтетические наблюдения могут быть добавлены к $\text{ОВ}_{\text{ИСД}}$ полностью (добавление всех синтетических данных (**ВСД**)) или частично (добавление дополнительного сегмента данных (**ДСД**) после отбора из ВСД).

В условиях малого накопления данных аналитического контроля используется вспомогательная модель для получения оценок РПК при помощи ЧПК: $\widehat{Y}_{\text{РПК}}=G(Y_{\text{ЧПК}})$. Наблюдения ЧПК и РПК содержат в себе данные аналитического контроля ($Y_{\text{ЧПК}}$ и $Y_{\text{РПК}}$), наблюдения входных переменных ($X_{\text{ЧПК}}$ и $X_{\text{РПК}}$) в соответствующие моменты времени.

Таким образом, наблюдения расширенной ОВ (**РОВ**) будут содержать наблюдения аналитического контроля РПК $Y_{\text{РПК}}$ и выход вспомогательной модели $\dot{Y}_{\text{РПК}}$, а также соответствующие наблюдения входных переменных $X_{\text{РПК}}$ и $X_{\text{ЧПК}}$. Последние, $X_{\text{ЧПК}}$, соответствуют по времени синтетическим данным $\widehat{Y}_{\text{РПК}}$, включаемым в ДСД:

$$Y_{\text{РОВ}}=\left[Y_{\text{РПК}};Y_{\text{ДСД}}\right],\ \ X_{\text{РОВ}}=\left[X_{\text{РПК}};X_{\text{ДСД}}\right].$$

Показатель разреженности. Для получения ДСД предлагается алгоритм отбора синтетических наблюдений с учетом показателя разреженности данных S [22]. Основная задача применения данного алгоритма – формализация и упрощение процесса отбора наблюдений, а также предотвращение включения в ОВ избыточного количества наблюдений. Также показатель разреженности служит индикатором достаточ-

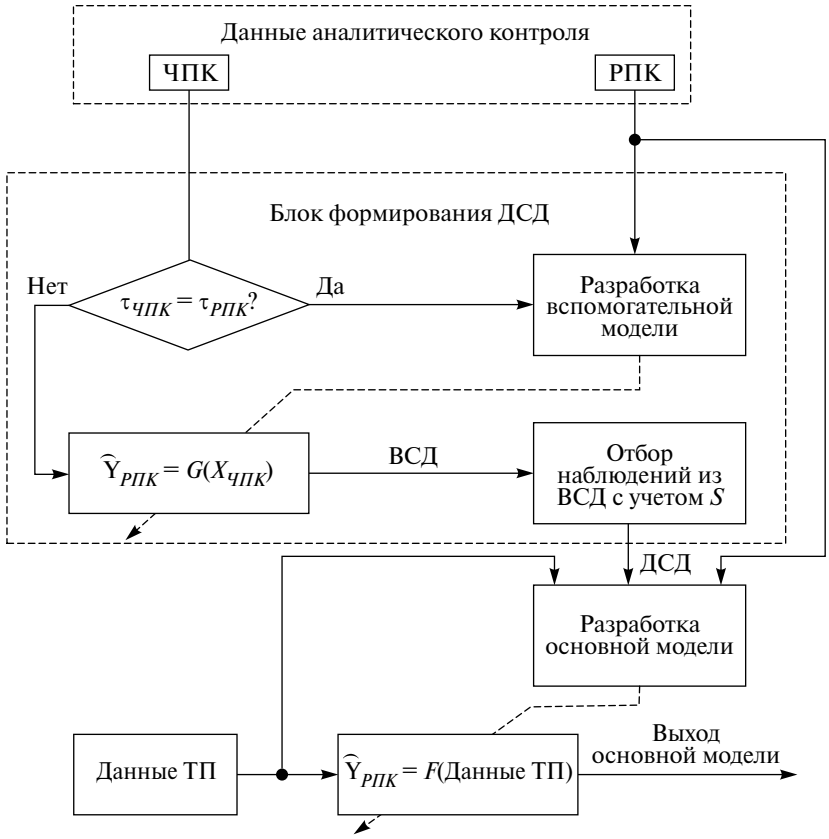

Рис. 3. Структура предлагаемого метода построения модели для оценки ПК в условиях малой выборки.

ности количества наблюдений выходной переменной в ОВ при построении модели.

Использование показателя разреженности S предполагает деление диапазона известных значений выходной переменной на n равных по длине интервалов и индикацию наличия хотя бы одного наблюдения в данном интервале. Показатель разреженности рассчитывается по формуле:

$$S = 1 - \frac{1}{n}\sum_{i=1}^{n} a_i, \qquad (18)$$

где $a_i = \begin{cases} 1, & \dim H_i \geq 1 \\ 0, & \dim H_i < 1 \end{cases}$ – индикатор того, что в i-м интервале есть хотя бы одно наблюдение; n – заданное количество интервалов; H_i – вектор, содержащий наблюдения выходной переменной, попадающих в i-й интервал:

$$H_i := x \in \left[L_l + \Delta \cdot (i-1), L_l + \Delta \cdot i \right], \; i = 1,...,n, \qquad (19)$$

где Δ – длина рассматриваемого интервала, которая вычисляется как:

$$\Delta = \left(L_h - L_l \right) / n, \qquad (20)$$

где L_h и L_l – верхняя и нижняя границы диапазона изменения выходной переменной соответственно.

Значения L_h и L_l можно задавать в соответствии с требованиями к ПК выходного продукта или в соответствии с интересующим сегментом выборки. Значения L_h и L_l выбираются в соответствии с максимальным и минимальным значениями $Y_{\text{РПК}}$ в $\text{ОВ}_{\text{ИСД}}$.

Показатель разреженности изменяется в интервале: $S \in [0,1)$, и его значение зависит от выбранного количества интервалов n. Определение задаваемого числа интервалов n осуществляется исходя из точности измерения ПК в ходе аналитического контроля.

Стоит отметить, что при проведении аналитического контроля ПК результат округляется до q-го разряда, который определяется в зависимости от метода проведения лабораторного анализа. Размерность значений (количество знаков после запятой) в синтетических данных определяется используемым программно-вычислительным комплексом и математическим методом построения вспомогательной модели и является достаточно большой величиной, избыточной для алгоритма расширения ОВ. Предлагается округление

значений синтетических наблюдений до q+1 разряда, что предотвращает включение всех синтетических наблюдений в ОВ при использовании $n \to +\infty$. В имеющихся данных по РПК температурные значения ПТФ и $T_{всп}$ среднего дистиллята, а также $T_{всп}$ керосиновой фракции представлены с точностью до целых ºС, предлагается округление значений выхода в синтетических данных до десятых долей ºС. ЦЧ среднего дистиллята определяется с точностью до десятых долей, в синтетических наблюдениях округляется до сотых долей. Вязкость в ходе аналитического контроля определяется с точностью до тысячных долей, в этом случае предлагается округление значений синтетических наблюдений до тысячных долей с целью уменьшения количества интервалов n, учитывая диапазон изменения выходной переменной в $ОВ_{ИСД}$.

Ниже представлен алгоритм построения моделей для оценки ПК продуктов колонны фракционирования на основе расширенной обучающей выборки с учетом показателя разреженности.

Алгоритм отбора наблюдений из ВСД с учетом показателя разреженности S	
На входе: $Y_{РПК}$, $X_{РПК}$ $\widehat{Y}_{РПК}$, $X_{ЧПК}$, ε_0	
На выходе: $Y_{ДСД}$, $X_{ДСД}$	
1	Формирование начальных значений $Y_{вДСД} = \widehat{Y}_{РПК}(1)$ и $X_{вДСД} = X_{ЧПК}(1)$
2	Вычисление $S_{вДСД\|1}$ для $Y_{вДСД}$
3	Инициализация счетчика $i = 2$ и $q = 1$
4	Делать пока $i \le \dim \widehat{Y}_{РПК}$:
5	Расчет $S_{вДСД\|i}$ для $[Y_{вДСД}; \widehat{Y}_{РПК}(i)]$
6	Если $S_{вДСД\|i} < S_{вДСД\|i-1}$, то:
7	$Y_{вДСД} = [Y_{вДСД}; \widehat{Y}_{РПК}(i)]$
8	$X_{вДСД} = [X_{вДСД}; X_{ЧПК}(i)]$
9	Если $\dim Y_{вДСД} \ge 10$, то:
10	Построение модели $Y_{вДСД} = F(X_{вДСД})$
11	Если $\lvert Y_{РПК} - F(X_{РПК}) \rvert < \varepsilon_{q-1}$, то:
12	$Y_{ДСД} = Y_{вДСД}$
13	$X_{ДСД} = X_{вДСД}$
14	$\varepsilon_j = \lvert Y_{РПК} - F(X_{РПК}) \rvert$
15	$q := q + 1$
16	$i := i + 1$
17	Возвращение $Y_{ДСД}$ и $X_{ДСД}$

Формирование временного дополнительного сегмента данных (**вДСД**) осуществляется в два этапа: сначала добавляется одно наблюдение синтетических данных с более поздним временем $\tau_{синт} \to \tau_{ТВ}$ для расчета начального значения $S_{вДСД}$, после чего добавление последующих наблюдений осуществляется по убыванию времени, и при условии уменьшения $S_{вДСД}$. Начиная с 10 включенных в вДСД наблюдений производится обучение модели на этой выборке и тестирование на $ОВ_{ИСД}$. Тестирование на $ОВ_{ИСД}$ предполагает выбор модели с наибольшей точностью. Полученный дополнительный сегмент данных содержит не более одного наблюдения в каждом из n интервалов и имеет наиболее высокую согласованность данных с $ОВ_{ИСД}$ по средней абсолютной ошибке при тестировании. Для сравнения эффективности предлагаемого алгоритма расширения ОВ с учетом разреженности используются следующие варианты ОВ: включение в ОВ только наблюдений из ИСД; к ИСД добавляются ВСД; к ИСД добавляется ДСД, то есть только отобранные с учетом S более поздние синтетические наблюдения.

Использование показателя разреженности в алгоритме расширения ОВ необходимо для заполнения как можно большего числа интервалов в диапазоне изменения выходной переменной $Y_{РПК}$. Основной целью заполнения интервалов в ходе отбора синтетических данных является добавление в ОВ для основной модели данных режимов функционирования технологического объекта, для которых отсутствуют измерения $Y_{РПК}$. Также учет показателя разреженности ограничивает включение в ОВ синтетических данных с дублирующимися значениями выхода, т. е. исключается дополнение данными близких технологических режимов.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Корреляционный анализ. В табл. 7–11 представлены корреляционные матрицы, показывающие линейную корреляцию между РПК и точками фракционного состава (**ФС**).

Стоит отметить, что результаты, представленные в табл. 7–11, не противоречат известным корреляциям [16–21] между РПК и точками ФС соответствующих продуктов колонны фракционирования. Поэтому в качестве входных переменных для вспомогательной модели при получении синтетических оценок использованы: для получения оценок ПТФ среднего дистиллята – 50, 90 и 95%-ные температуры выкипания; для $T_{всп}$ среднего дистиллята использована ТНК; для

Таблица 7. Коэффициенты корреляции между ПТФ среднего дистиллята и точками ФС

	ПТФ	ТНК	ФС10	ФС50	ФС90	ФС95
ПТФ	1	-0.0869	0.1723	0.6933	0.8064	0.7910
ТНК	-0.0869	1	0.7724	-0.0171	-0.1667	-0.2152
ФС10	0.1723	0.7724	1	0.3669	0.2514	0.2021
ФС50	0.6933	-0.0171	0.3669	1	0.8828	0.8673
ФС90	0.8064	-0.1667	0.2514	0.8828	1	0.9875
ФС95	0.7910	-0.2152	0.2021	0.8673	0.9875	1

Таблица 8. Коэффициенты корреляции между $T_{всп}$ среднего дистиллята и точками ФС

	ПТФ	ТНК	ФС10	ФС50	ФС90	ФС95
ПТФ	1	0.9138	0.7610	0.4063	0.3344	0.3451
ТНК	0.9138	1	0.7275	0.2901	0.2124	0.2141
ФС10	0.7610	0.7275	1	0.7926	0.7405	0.7303
ФС50	0.4063	0.2901	0.7926	1	0.9783	0.9734
ФС90	0.3344	0.2124	0.7405	0.9783	1	0.9969
ФС95	0.3451	0.2141	0.7303	0.9734	0.9969	1

Таблица 9. Коэффициенты корреляции между вязкостью среднего дистиллята при 40℃ и точками ФС

	ПТФ	ТНК	ФС10	ФС50	ФС90	ФС95
ПТФ	1	0.1500	0.3389	0.7004	0.5581	0.5058
ТНК	0.1500	1	0.8871	-0.2151	-0.4756	-0.4354
ФС10	0.3389	0.8871	1	-0.0096	-0.3611	-0.3534
ФС50	0.7004	-0.2151	-0.0096	1	0.7915	0.7421
ФС90	0.5581	-0.4756	-0.3611	0.7915	1	0.9703
ФС95	0.5058	-0.4354	-0.3534	0.7421	0.9703	1

Таблица 10. Коэффициенты корреляции между ЦЧ среднего дистиллята и точками ФС

	ПТФ	ТНК	ФС10	ФС50	ФС90	ФС95
ПТФ	1	-0.1124	0.3894	0.5791	0.5435	0.5066
ТНК	-0.1124	1	0.3576	-0.1498	-0.2631	-0.2440
ФС10	0.3894	0.3576	1	0.6613	0.2837	0.2300
ФС50	0.5791	-0.1498	0.6613	1	0.6929	0.6339
ФС90	0.5435	-0.2631	0.2837	0.6929	1	0.9908
ФС95	0.5066	-0.2440	0.2300	0.6339	0.9908	1

Таблица 11. Коэффициенты корреляции между $T_{всп}$ керосиновой фракции и точками ФС

	ПТФ	ТНК	ФС10	ФС50	ФС90	ФС98
ПТФ	1	0.7132	0.7482	0.5307	0.2631	0.2303
ТНК	0.7132	1	0.7717	0.3537	0.0173	0.0300
ФС10	0.7482	0.7717	1	0.8111	0.4939	0.4992
ФС50	0.5307	0.3537	0.8111	1	0.8519	0.7988
ФС90	0.2631	0.0173	0.4939	0.8519	1	0.8363
ФС98	0.2303	0.0300	0.4992	0.7988	0.8363	1

вязкости среднего дистиллята при 40℃ – 50 и 90%-ные температуры выкипания; для ЦЧ среднего дистиллята – 10, 50, 90 и 95%-ные температуры выкипания ввиду сравнительно низких значений коэффициентов корреляции; для $T_{всп}$ керосиновой фракции – ТНК и температура выкипания 10 об. % .

Вспомогательная модель. Построение вспомогательных моделей осуществляется тремя приведенными выше методами, используя $ОВ_{исд}$. Для обучения нейронных сетей использована НСПР, содержащая один скрытый слой, состоящий из одного нейрона, функция активации – гиперболический тангенс, и в качестве метода обучения

использован метод байесовской регуляризации [28]. Для получения синтетических оценок с помощью модели K-OPLS значение параметра σ подбиралось итерационно, критерий оптимальности – САО. Размеры выборок синтетических наблюдений по соответствующим РПК приведены в табл. 12.

В табл. 13 приведены рассчитанные критерии точности для построенных вспомогательных моделей.

Из полученных в табл. 13 результатов видно, что наименьшая ошибка при обучении на $ОВ_{ИСД}$ для ПТФ среднего дистиллята достигается при использовании K-OPLS; для $T_{всп}$ среднего дистиллята – при использовании НСПР; для вязкости среднего дистиллята при 40 ℃ – при использовании РР; для ЦЧ среднего дистиллята – при использовании K-OPLS; для $T_{всп}$ КФ – при использовании НСПР. Отличия значений САО и R^2 при сравнении методов в данном случае незначительны.

Для формирования ВСД используются отобранные наблюдения точек ФС, соответствующие условию $\tau_{ФС} \neq \tau_{РПК}$ и $\tau_{ФС} < \tau_{ТВ}$. Вычисление ошибок оценок моделей относительно значений лабораторных наблюдений в данном случае невозможно.

Проверка вспомогательных моделей. Для проверки качества вспомогательных моделей (насколько хорошо синтетические данные отражают зависимости ИСД) произведено построение моделей на основе полученных синтетических наблюдений и проведено тестирование их на $ОВ_{ИСД}$, использованной при построения вспомогательных моделей. Построение моделей проверки также осуществляется тремя указанными математическими методами. Рассчитанные критерии точности при обучении на синтетических данных и тестировании на $ОВ_{ИСД}$ для исследуемых РПК представлены в табл. 14–18.

Согласно результатам, приведенным в табл. 14–18, при различных методах построения модели проверки меньшая САО достигается при использовании в качестве обучающей выборки ВСД, полученного с помощью НСПР, сравнение осуществляется по столбцам. При сравнении по строкам полученных результатов для большинства ПК также меньшая САО достигается при использовании НСПР, поэтому этот метод используется далее для построения основных моделей. При этом меньшая ошибка также достигается при использовании пары вспомогательная модель – основная модель варианта НСПР – НСПР. Данный эффект объясняется тем, что наборы ВСД, полученные с помощью разных моделей, в основном отличаются границами диапазона изменения значений выходной переменной и распределением наблюдений внутри этого диапазона.

Таблица 12. Размеры выборок синтетических наблюдений для РПК

ПК	Кол-во синтетических наблюдений, шт.
ПТФ среднего дистиллята	617
$T_{всп}$ среднего дистиллята	630
Вязкость при 40°С среднего дистиллята	786
ЦЧ среднего дистиллята	512
$T_{всп}$ КФ	325

Построение основных моделей. Для сравнения эффективности использования критерия разреженности S при отборе синтетических наблюдений для обучения основных моделей использованы следующие варианты формирования ОВ:

– ОВ1 = [$ОВ_{ИСД}$];
– ОВ2 = [$ОВ_{ИСД}$; ВСД];
– ОВ3 = [$ОВ_{ИСД}$; ДСД].

В табл. 19 представлены выбранные параметры для расчета значения критерия разреженности S при работе алгоритма расширения ОВ для рассматриваемых РПК.

На рис. 4 представлены графики зависимостей САО при тестировании полученных моде-

Таблица 13. Критерии точности при обучении вспомогательных моделей

	Вспомогательная модель					
	РР		НСПР		K-OPLS	
РПК	R^2	САО	R^2	САО	R^2	САО
ПТФ среднего дистиллята	0.6531	2.8191	0.6453	2.9142	**0.6558**	**2.7858**
$T_{всп}$ среднего дистиллята	0.8351	2.8432	**0.8400**	**2.8166**	0.8390	2.8361
Вязкость среднего дистиллята при 40	**0.9084**	**0.0790**	0.8998	0.0860	0.8959	0.0950
ЦЧ среднего дистиллята	0.2589	0.7515	0.2744	0.7411	**0.3171**	**0.7343**
$T_{всп}$ КФ	0.6054	1.4897	**0.6406**	**1.4198**	0.5842	1.5464

Таблица 14. Критерии точности моделей проверки для ПТФ среднего дистиллята

	Модель проверки					
	PP		НСПР		K-OPLS	
Вспомогательная модель	R^2	САО	R^2	САО	R^2	САО
МНК	-0.1025	4.7660	-0.1033	4.7480	-0.0378	4.6889
НСПР	0.0690	4.4640	0.1751	4.1814	0.0925	4.5961
K-OPLS	0.0481	4.5614	**0.1893**	**4.1501**	0.0284	4.7498

Таблица 15. Критерии точности моделей проверки для $T_{всп}$ среднего дистиллята

	Модель проверки					
	PP		НСПР		K-OPLS	
Вспомогательная модель	R^2	САО	R^2	САО	R^2	САО
МНК	0.5010	5.9617	0.5707	4.3964	0.3932	5.4185
НСПР	0.5620	5.5148	**0.5982**	**4.0565**	0.4809	5.3253
K-OPLS	0.5349	5.7632	0.5862	4.2618	0.4555	5.3138

Таблица 16. Критерии точности моделей проверки для вязкости среднего дистиллята при 40℃

	Модель проверки					
	PP		НСПР		K-OPLS	
Вспомогательная модель	R^2	САО	R^2	САО	R^2	САО
МНК	-0.5954	0.4550	**-0.3041**	0.3816	-0.3843	0.4113
НСПР	-0.7258	0.4776	-0.3048	**0.3761**	-0.3735	0.4091
K-OPLS	-0.8388	0.4957	-0.3590	0.3795	-0.3763	0.4079

Таблица 17. Критерии точности моделей проверки для ЦЧ среднего дистиллята

	Модель проверки					
	PP		НСПР		K-OPLS	
Вспомогательная модель	R^2	САО	R^2	САО	R^2	САО
МНК	-0.1993	0.8552	-0.2508	**0.8518**	-0.2639	0.9085
НСПР	-0.1717	0.8566	-0.1280	0.8553	-0.1808	0.8804
K-OPLS	**-0.1149**	0.9104	-0.3661	1.0235	-1.1802	1.1356

Таблица 18. Критерии точности моделей проверки для ЦЧ КФ

	Модель проверки					
	PP		НСПР		K-OPLS	
Вспомогательная модель	R^2	САО	R^2	САО	R^2	САО
МНК	0.5912	1.6966	0.5817	1.7277	0.4810	1.8380
НСПР	0.5934	**1.6099**	**0.5935**	1.7069	0.4719	1.8270
K-OPLS	0.5618	1.8109	0.5625	1.7240	0.4794	1.8481

Таблица 19. Параметры для расчета значения критерия разреженности S

ПК	L_l	L_h	n	Δ
ПТФ среднего дистиллята	–42	–22	100	0.2
$T_{всп}$ среднего дистиллята	29	91	310	0.2
Вязкость при 40ºС среднего дистиллята	2.4	3.1	700	0.001
ЦЧ среднего дистиллята	43.4	51.7	415	0.02
$T_{всп}$ КФ	28	55	135	0.2

лей на $ОВ_{ИСД}$ от разреженности S обучающей выборки для рассматриваемых РПК.

В качестве ДСД выбирается набор наблюдений, для которого результаты тестирования модели на $ОВ_{ИСД}$ соответствуют минимуму глобального тренда зависимости САО от значения показателя разреженности. Это объясняется необходимостью заполнения как можно большего числа интервалов более поздними наблюдениями, время которых $\tau_{синт} \to \tau_{ТВ}$, без учета резкого изменения САО при тестировании на $ОВ_{ИСД}$, что характерно при малом числе наблюдений в ОВ. Здесь мы не рассматриваем вопрос отбора наблюдений, добавление которых в ОВ, помимо снижения значения показателя разреженности S, также снижает САО при тестировании на $ОВ_{ИСД}$. В этом случае добавление наблюдений при выполнении двух условий приводит к переобучению модели, а также не позволяет включить достаточное количество наблюдений в ОВ ввиду скачкообразного и резкого снижения САО при малом количестве наблюдений в ОВ.

В табл. 20 представлены рассчитанные критерии точности на тестовой выборке разработанных моделей для вариантов: без расширения ОВ (ОВ1), расширение путем добавления всех

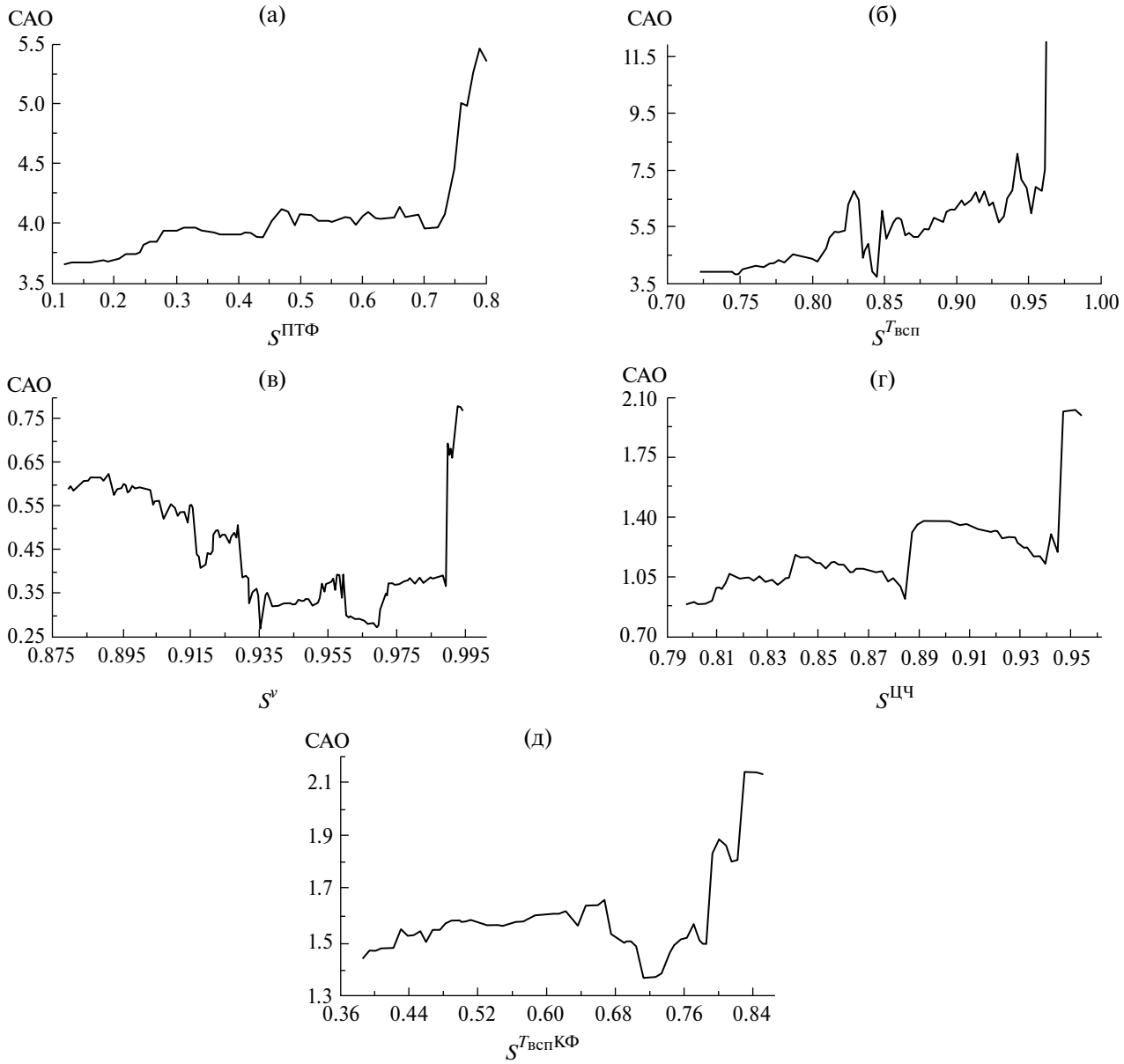

Рис. 4. Графики зависимостей САО при тестировании на $ОВ_{ИСД}$ от значения показателя разреженности обучающей выборки S: (а) ПТФ среднего дистиллята; (б) $T_{всп}$ среднего дистиллята; (в) вязкость при 40°С среднего дистиллята; (г) ЦЧ среднего дистиллята; (д) $T_{всп}$ керосиновой фракции.

Таблица 20. Критерии точности моделей на ТВ с разными вариантами ОВ для оценки ПК продуктов

ПК	Вариант ОВ	$N_{ов}$	R^2	САО	ΔСАО, %
ПТФ среднего дистиллята	ОВ1	34	0.1935	3.5663	0.00
	ОВ2	651	-0.0370	3.7096	4.02
	ОВ3	122	**0.3474**	**3.0291**	-15.06
$T_{всп}$ среднего дистиллята	ОВ1	46	0.4445	5.0140	0.00
	ОВ2	676	0.5052	4.5810	-8.64
	ОВ3	94	**0.5235**	**4.4433**	-11.38
Вязкость при 40°С среднего дистиллята	ОВ1	41	0.0358	0.1100	0.00
	ОВ2	827	0.0831	0.1047	-4.82
	ОВ3	172	**0.3311**	**0.0923**	-16.09
ЦЧ среднего дистиллята	ОВ1	105	-1.3906	1.6145	0.00
	ОВ2	617	**0.1303**	**0.9900**	-38.68
	ОВ3	140	-0.0915	1.0393	-35.63
$T_{всп}$ КФ	ОВ1	78	0.2462	1.6809	0.00
	ОВ2	403	0.3877	1.5225	-9.42
	ОВ3	103	**0.4246**	**1.4933**	-11.16

синтетических наблюдений (ОВ2) и расширение путем добавления к ОВ синтетических наблюдений, отобранных с учетом показателя разреженности S (ОВ3).

Для варианта ОВ2 – при добавлении всех синтетических наблюдений снижение САО достигается в большей степени за счет увеличения количества наблюдений в ОВ. Повышение ошибки в данном случае объясняется “качеством” полученных синтетических наблюдений ввиду того, что синтетические наблюдения получаются при использовании преимущественно точек ФС продуктов из других режимов работы колонны фракционирования. В случае для ПТФ среднего дистиллята границы изменения значений выходной переменной значительно отличаются в зависимости от целевого продукта. Поэтому добавление ВСД приводит к снижению точности получаемых моделей (табл. 20).

Отбор синтетических наблюдений с учетом показателя разреженности S позволяет выделить из ВСД сегмент данных, наиболее согласованный с известным $ОВ_{исд}$. Добавление ДСД к известным наблюдениям также может привести к снижению точности получаемой модели в сравнении с добавлением ВСД, как, например, при оценке ЦЧ среднего дистиллята (табл. 20). Данный эффект объясняется хорошей согласованностью получаемых синтетических наблюдений и $ОВ_{исд}$. Стоит отметить, что погрешность при измерении на “моторе” ЦЧ составляет $\Delta ЦЧ = \pm 2$ [29], поэтому для ЦЧ качество получаемых моделей ограничено погрешностью способа его определения в ходе аналитического контроля. При этом и в случае добавления ВСД, и при добавлении ДСД удается получить достаточно низкое значение САО модели, близкое к 1. Преимущество отбора синтетических наблюдений с учетом показателя разреженности S в данном случае заключается в возможности выделения наиболее согласованного с $ОВ_{исд}$ дополнительного сегмента данных, состоящего из 25 наблюдений, добавление которого также значительно повышает точность модели в сравнении с вариантом использования $ОВ_{исд}$.

Использование предлагаемого в данной работе метода направлено на повышение точности при оценке ПК продуктов колонны фракционирования за счет расширения ОВ включением отобранных синтетических наблюдений с учетом показателя разреженности данных S. Как видно из полученных результатов в табл. 20, использование предлагаемого метода расширения ОВ при оценке всех исследуемых ПК продуктов колонны фракционирования позволяет повысить точность получаемых моделей в сравнении с вариантом использования $ОВ_{исд}$.

ЗАКЛЮЧЕНИЕ

Предложен метод построения моделей для оценки ПК целевых продуктов колонны фракционирования технологической установки гидрокрекинга в условиях малого объема данных аналитического контроля. В основе метода лежит расширение ОВ за счет добавления ДСД, отбираемого с учетом показателя разреженности из ВСД. ВСД получены с помощью построения

вспомогательной модели на основе зависимости между РПК и ЧПК. Показана эффективность применения данного метода на примерах моделей для оценки показателей качества среднего дистиллята и КФ, а также показана эффективность применения НСПР для построения вспомогательной модели. Снижение САО моделей на тестовой выборке, в сравнении с использованием в качестве ОВ известного сегмента данных, составило: при оценке ПТФ среднего дистиллята – 15.1%, $T_{\text{всп}}$ среднего дистиллята – 11.4%, при оценке вязкости среднего дистиллята при 40°С – 16.1%, ЦЧ среднего дистиллята – 35.6% и $T_{\text{всп}}$ КФ – 11.2%. Показана эффективность использования критерия разреженности при отборе синтетических наблюдений, при этом снижение САО моделей на тестовой выборке, в сравнении с вариантом добавления в ОВ всех синтетических данных, составило: при оценке ПТФ среднего дистиллята – 18.3%, $T_{\text{всп}}$ среднего дистиллята – 2.7%, ЦЧ среднего дистиллята – 11.8% и $T_{\text{всп}}$ КФ – 1.9%.

Работа выполнена в рамках государственного задания ИАПУ ДВО РАН по теме № FWFW-2021-0003 (метод построения моделей для оценки ПК в условиях малой обучающей выборки) и FWFW-2025-0002 (реализация алгоритма отбора наблюдений из ВСД и его апробация на технологических данных).

ОБОЗНАЧЕНИЯ

a_i	индикатор того, что в i-м интервале есть хотя бы 1 наблюдение;
b_0	свободный коэффициент регрессии;
b_j	коэффициент при j-й входной переменной;
$\boldsymbol{b}_{RR}$	вектор коэффициентов робастной регрессии;
c	константа настройки (для Fair $c = 1.4$);
E	матрица остатков;
H_i	вектор, содержащий наблюдения выходной переменной, попадающих в i-й интервал;
h_i	i-е значение вектора $\boldsymbol{h}$;
$\boldsymbol{h}$	вектор, состоящий из диагональных элементов матрицы $X\left(X^T X\right)^{-1} X^T$;
k	количество нейронов в скрытом слое нейронной сети;
$\ker(\cdot,\cdot)$	функция ядра;
L_h	верхняя граница диапазона изменения выходной переменной;
L_l	нижняя граница диапазона изменения выходной переменной;
m	количество входных переменных;
med	медианное значение;
N	количество наблюдений;
n	заданное количество интервалов;
P_o^T	соответствующая y-ортогональная матрица нагрузки;
P_p^T	y-прогнозная матрица нагрузки (англ. loading matrix);
$\mathbf{p}_1$	вектор свободных членов скрытого слоя нейронной сети;
p_2	свободный член выходного слоя нейронной сети;
$\mathbf{p}_i$	вектор смещения для нейронов выходного слоя;
q	разряд числа для округления;
R^2	коэффициент детерминации;
$\boldsymbol{r}$	вектор значений для расчета весовых коэффициентов зависит от медианного значения модулей отклонений ошибок на предыдущем шаге итерации;
S	показатель разреженности;
T_o	соответствующая y-ортогональная матрица вкладов;
T_p	y-прогнозная матрица вкладов (англ. *score matrix*);
W_1	матрица весовых коэффициентов скрытого слоя нейронной сети;
$\mathbf{w}_2$	вектор весовых коэффициентов выходного слоя нейронной сети;
$\mathbf{w}_i^{(1)}$	вектор весов между входными нейронами и нейронами скрытого слоя;
$\mathbf{w}_i^{(2)}$	вектор масштабных коэффициентов между связями от нейронов скрытого слоя и выходным нейроном;
X	матрица входных переменных;
$x_{\mathbf{i}}$	i-я строка матрицы входных переменных X;
$x_{i,j}$	i-е наблюдение j-й входной переменной;
y_i	i-е наблюдение выходной переменной;
$\hat{y}_i$	оцениваемое значение i-го наблюдения выходной переменной;
$\bar{y}$	среднее значение выходной переменной;
Δ	длина рассматриваемого интервала;
$\boldsymbol{\varepsilon}$	вектор заданных ошибок;
ε_q	САО модели на $\text{ОВ}_{\text{ИСД}}$ в алгоритме формирования ДСД на q-м шаге;
σ	параметр настройки в методе K-OPLS;
τ	дата и время, указанные в данных аналитического контроля;
φ	функция активации скрытого слоя нейронной сети;
ω_i	значение весового коэффициента для i-го наблюдения.

ИНДЕКСЫ

i номер наблюдения;
h верхняя граница;
j номер входной переменной;
l нижняя граница;
o обозначение *y*-ортогональной матрицы;
p обозначение y-прогнозной матрицы;
q номер шага в алгоритме формирования ДСД;
RR робастная регрессия;
T операция транспонирования.

СПИСОК ЛИТЕРАТУРЫ

1. *Logunov P.L., Shamanin M.V., Kneller D.V., Setin S.P., Shunderyuk M.M.* Advanced process control: from a PID loop up to refinery-wide optimization // Autom. & Remote Control. 2020. V. 80. № 10. P. 1929.
2. *Iplik E., Aslanidou I., Kyprianidis L.* Hydrocracking: a perspective towards digitalization // Sustainability. 2020. V. 12. № 17. P. 1.
3. *Fortuna L., Graziani S., Sicilia M.G.* Comparison of soft-sensor design methods for industrial plants using small data sets // IEEE Transactions on Instr. And Meas. 2009. V. 58. № 8. P. 2444.
4. *Shaikhina T., Khovanova N.A.* Handling limited datasets with neural networks in medical applications: a small-data approach // Artificial Intel. In Med. 2016. V. 75. № 1. P. 1.
5. *Napoli G., Xibilia M.G.* Soft Sensor design for a Topping process in the case of small datasets // Comput. & Chem. Eng. 2010. V. 35. № 11. P. 2447.
6. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. М.: Финансы и статистика, 1986. С. 73.
7. *Гаскаров Д.В., Шаповалов В.И.* Малая выборка. М.: Статистика, 1978. С. 19.
8. *Zhu Q.X., Hou K.R., Chen Z.S., Gao Z.S., Xu Y., He Y.L.* Novel virtual sample generation using conditional GAN for developing soft sensor with small data // Eng. Appl. Artif. Intell. 2021. V. 106. № 2.
9. *Zhang X.H., Xu Y., He Y.L., Zhu Q.X.* Novel manifold learning based virtual sample generation for optimizing soft sensor with small data // ISA Transactions. 2021. V. 109. № 1. P. 229.
10. *Li D.C., Lin L.S., Peng L.J.* Improving learning accuracy by using synthetic samples for small datasets with non-linear attribute dependency // Decision Support Syst. 2014. V. 59. № 1. P. 286.
11. *Samotylova S.A., Torgashov A.Yu.* Application of a first principles mathematical model of a mass-transfer technological process to improve the accuracy of the estimation of the end product quality // Theor. Found. Chem. Eng. 2022. V. 56. № 3. P. 371. [*Самотылова С.А., Торгашов А.Ю.* Применение физически обоснованной математической модели массообменного технологического процесса для повышения точности оценивания качества конечного продукта // Теорет. основы хим. технологии. 2022. Т. 56. № 3. С. 371.]
12. *Bai X., Li S.* A virtual sample generation method based on manifold learning and a generative adversarial network for soft sensor models with limited data // J. of the Taiwan Inst. of Chem. Eng. 2023. V. 151. № 3.
13. *Liu Y., Xie M.* Rebooting data-driven soft-sensors in process industries: a review of kernel methods // J. of Proc. Control. 2020. V. 89. № 4. P. 58.
14. *He Y.L., Hua Q., Zhu Q.H., Lu S.* Enhanced virtual sample generation based on manifold features: applications to developing soft sensor using small data // ISA Transactions. 2021. V. 126. № 4. P. 1.
15. *Zhu Q.X., Chen Z.S., Zhang X.H., Rajabifard A., Xu Y., Chen Y.Q.* Dealing with small sample size problems in process industry using virtual sample generation: a kriging-based approach // Soft Computing. 2020. V. 24. № 1. P. 6889.
16. *Dinkov R., Stratiev D.* Investigation on diesel cold flow properties // Proc. 45th International Petroleum Conf. Bratislava, 2011. P. 1.
17. *Vrablik A., Velvarska R., Stepanek K., Psenicka M., Hidalgo J.M., Cerny R.* Rapid models for predicting the low-temperature behavior of diesel // Chem. Eng. Technology. 2019. V. 42. № 7. P. 735.
18. *Aleme H.G., Barbeira P.J.S.* Determination of flash point and cetane index in diesel using distillation curves and multivariate calibration // Fuel. 2019. V. 102. № 1. P. 129.
19. *Gorenkov A.F., Lifanova T.A., Klyuchko I.G.* Influence of jet fuel distillation range on quality indexes // Chem. & Technology of Fuels & Oils. 1985. V. 21. № 8. P. 37. [*Горенков А.Ф., Лифанова Т.А., Кличко И.Г.* Влияние диапазона перегонки реактивного топлива на показатели качества // Химия и Техн. Топлив и Масел. 1985. Т. 21. № 8. С. 37.]
20. *Aleme H.G., Assuncao R.A., Carvalho M.M.O., Barbeira P.J.S.* Determination of specific gravity and kinematic viscosity of diesel using distillation curves and multivariate calibration // Fuel Processing Techn. 2012. V. 102. № 1. P. 90.
21. *Shepherd J.E., Nyut C.D., Lee J.J.* Flash point and chemical composition of aviation kerosene (Jet A) // Explosion Dynamics Laboratory Report FM99-4. 1999. P. 1.
22. *Штакин Д.В., Снегирев О.Ю., Торгашов А.Ю.* Метод построения виртуальных анализаторов в условиях малой обучающей выборки для управления качеством целевых продуктов фракционатора установки гидрокрекинга // Автоматизация в пром. 2024. Т. 22. № 6. С. 7.
23. *Dumuochel W., O'Brien F.* Integrating a robust option into a multiple regression computing environment // Comp. and graphics in statistics. 1992. P. 41.

24. *Cybenko G.* Approximation by superpositions of a sigmoidal function // Math. of Control, Signals & Systems. 1989. V. 2. № 1. P. 303.
25. *Bylesjo M., Rantalainen M., Nicholson J.K., Holmes E., Trygg J.* K-OPLS package: Kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space // BioMed Central. 2008. V. 9. № 1. P. 1.
26. *Holland P.W., Welsch R.E.* Robust regression using iteratively reweighted least-squares // Communic. in Statistics – Theory & Methods. 1977. V. 6. № 9. P. 813.
27. *Rantalainen M., Bylesjo M., Cloarec O., Nicholson J.K., Holmes E., Trygg J.* Kernel-based orthogonal projections to latent structures (K-OPLS) // J. of Chemometrics. 2007. V. 21. № 7–9. P. 376.
28. *Hayrettin O.* Bayesian regularized neural networks for small *n* big *p* data // Artif. Neural Net. – Models & Appl. 2016. P. 27.
29. *Prak D.L., Cooke J., Dickerson T., McDaniel A., Cowart J.* Cetane number, derived cetane number, and cetane index: when correlations fail to predict combustibility // Fuel. 2021. V. 289. № 12. P. 1.

METHOD OF MODEL BUILDING FOR ESTIMATION OF QUALITY PARAMETERS OF FRACTIONATION COLUMN PRODUCTS UNDER CONDITIONS OF SMALL VOLUME OF ANALYTICAL CONTROL DATA

A. A. Plotnikov, D. V. Shtakin, O. Yu. Snegirev, A. Yu. Torgashov*

Institute of Automatics and Control Processes, Far East Branch, Russian Academy of Sciences, Vladivostok, Russia
**e-mail: torgashov@iacp.dvo.ru*

Abstract. The problem of improving the accuracy of models for estimating the low-temperature properties, flammability and anti-wear properties of the target products of the fractionation column under conditions of a small amount of analytical control data is considered. For the solution of the considered problem the method of model building is proposed, which includes the algorithm of expansion of a small training sample on the data of fractional composition, differing in the way of selection of additional data, taking into account the sparsity indicator, which allowed to include the missing amount of data in the training sample, and as a result to ensure the improvement of the model quality. The use of the proposed method improved the accuracy of the models by 18% on average compared to the known methods and by 6% on average compared to the method based on the expansion of the training sample without taking into account the sparsity index. The results are presented on examples of model building of quality indicators of filterability limit temperature, flash point, kinematic toughness at 40°C and cetane number of middle distillate (diesel fuel fraction) and flash point of kerosene fraction of industrial fractionation column of hydrocracking process unit.

Keywords: *mathematical models for assessing petroleum product quality indicators, rectification, small sample, sample expansion, sparsity, analytical control*