Сравнительный анализ методов построения виртуальных анализаторов качества продуктов колонны фракционирования в условиях пропусков данных в обучающей выборке / А. А. Плотников, Д. В. Штакин, О. Ю. Снегирев и др. // Прикладная математика и вопросы управления. – 2024. – № 2. – С. 78–95. DOI 10.15593/2499-9873/2024.2.06

Библиографическое описание согласно ГОСТ Р 7.0.100–2018
Сравнительный анализ методов построения виртуальных анализаторов качества продуктов колонны фракционирования в условиях пропусков данных в обучающей выборке / А. А. Плотников, Д. В. Штакин, О. Ю. Снегирев, А. Ю. Торгашов. – Текст : непосредственный. – DOI 10.15593/2499-9873/2024.2.06 // Прикладная математика и вопросы управления / Applied Mathematics and Control Sciences. – 2024. – № 2. – С. 78–95.

пермский политех

ПРИКЛАДНАЯ МАТЕМАТИКА И ВОПРОСЫ УПРАВЛЕНИЯ
№ 2, 2024
https://ered.pstu.ru/index.php/amcs

Научная статья
DOI: 10.15593/2499-9873/2024.2.06
УДК 519.6

Сравнительный анализ методов построения виртуальных анализаторов качества продуктов колонны фракционирования в условиях пропусков данных в обучающей выборке

А.А. Плотников[2], Д.В. Штакин[1,2], О.Ю. Снегирев[1], А.Ю. Торгашов[1,2]

[1]Институт автоматики и процессов управления Дальневосточного отделения Российской академии наук, Владивосток, Российская Федерация
[2]Дальневосточный федеральный университет, Владивосток, Российская Федерация

О СТАТЬЕ

Получена: 15 марта 2024
Одобрена: 06 июня 2024
Принята к публикации: 06 июня 2024

Финансирование
Госбюджетная тема научных исследований ИАПУ ДВО РАН: FWFW-2021-0003.
Конфликт интересов
Авторы заявляют об отсутствии конфликта интересов.
Вклад авторов
равноценен.

Ключевые слова:
статистические модели, виртуальные анализаторы, сравнительный анализ, колонна фракционирования, керосиновая фракция, оценка качества.

АННОТАЦИЯ

Рассматривается сравнительный анализ методов построения виртуальных анализаторов с использованием робастной регрессии, гребневой регрессии, метода ортогональных проекций на скрытые структуры на основе ядра (англ. K-OPLS), метода чередующихся условных математических ожиданий (англ. ACE) и нейросетей прямого распространения. Данные модели в составе виртуальных анализаторов предназначены для оценки значений точек фракционного состава керосиновой фракции – продукта колонны фракционирования – в режиме реального времени. В ходе построения моделей рассмотрен вопрос усреднения значений входных переменных за определенный промежуток времени для привязки к значениям выходных переменных. В отличие от существующих работ, в данном исследовании обучение и тестирование моделей осуществляется на ограниченных по значениям выходной переменной сегментах данных, т.е. в условиях пропусков данных в обучающей выборки. Показано влияние ширины интервала усреднения значений входной переменной на точность оценки получаемых моделей. Также показано, что наименьшее значение средней абсолютной ошибки при оценке точек фракционного состава обеспечивают модели на основе нейронных сетей и K-OPLS при различных вариантах обучения и тестирования.

© **Плотников Александр Александрович** – магистрант ДВФУ e-mail: Plotnikov.aal@dvfu.ru.
Штакин Денис Владимирович – младший научный сотрудник ИАПУ ДВО РАН[1]; аспирант ДВФУ[2], e-mail: Dshtakin21@ya.ru.
Снегирев Олег Юрьевич – кандидат технических наук, научный сотрудник ИАПУ ДВО РАН e-mail: Snegirevoleg@iacp.dvo.ru, ORCID 0000-0002-0322-6609.
Торгашов Андрей Юрьевич – доктор технических наук, главный научный сотрудник ИАПУ ДВО РАН[1]; профессор ДВФУ[2], e-mail: Torgashov@iacp.dvo.ru, ORCID 0000-0003-3299-0742.

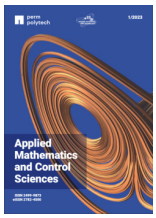
Perm Polytech Style: Plotnikov A.A., Shtakin D.V., Snegirev O.Yu., Torgashov A.Yu. The comparative analysis of the methods of constructing soft sensors for the quality estimation of fractionation column products with account to missing data in the training set. *Applied Mathematics and Control Sciences.* 2024, no. 2, pp. 78–95. DOI: 10.15593/2499-9873/2024.2.06

MDPI and ACS Style: Plotnikov, A.A.; Shtakin, D.V.; Snegirev, O.Yu.; Torgashov, A.Yu. The comparative analysis of the methods of constructing soft sensors for the quality estimation of fractionation column products with account to missing data in the training set. *Appl. Math. Control Sci.* **2024**, **2**, 78–95. https://doi.org/10.15593/2499-9873/2024.2.06

Chicago/Turabian Style: Plotnikov, Alexander A., Denis V. Shtakin, Oleg Yu. Snegirev, and Andrei Yu. Torgashov. 2024. "The comparative analysis of the methods of constructing soft sensors for the quality estimation of fractionation column products with account to missing data in the training set". *Appl. Math. Control Sci.* no. 2: 78–95. https://doi.org/10.15593/2499-9873/2024.2.06

perm polytech

APPLIED MATHEMATICS AND CONTROL SCIENCES
№ 2, 2024
https://ered.pstu.ru/index.php/amcs



Article
DOI: 10.15593/2499-9873/2024.2.06
UDC 519.6

The comparative analysis of the methods of constructing soft sensors for the quality estimation of fractionation column products with account to missing data in the training set

A.A. Plotnikov[2], D.V. Shtakin[1,2], O.Yu. Snegirev[1], A.Yu. Torgashov[1,2]

[1]Institute of Automation and Control Processes Far Eastern Branch of the Russian Academy of Sciences, Vladivostok, Russian Federation
[2]Far Eastern Federal University, Vladivostok, Russian Federation

ARTICLE INFO

Received: 15 March 2024
Approved: 06 June 2024
Accepted for publication: 06 June 2024

Funding
IACP FEB RAS: FWFW-2021-0003.
Conflicts of Interest
The authors declare no conflict of interest.
Author Contributions
equivalent.

Keywords:
statistical models, soft sensors, comparative analysis, fractionation column, kerosene fraction, quality estimating.

ABSTRACT

This article presents a comparative analysis of methods of constructing statistical models based on robust regression, ridge regression, kernel-based orthogonal projections to latent structures (K-OPLS), alternating conditional expectations (ACE) and direct distribution neural networks. These models are used for estimating the values of the points of fractional composition of the kerosene fraction, the product of the fractionation column. During the construction of models, the issue of meaning the values of input variables over a certain period was considered to link them to the values of output variables. Unlike the existing works, in this article, training and testing of models is carried out on segments of the data array limited in the values of the output variable. The training segment is formed from the general array by excluding observations whose values are limited by upper and lower limits. The excluded observations constitute the test sample. This paper shows the influence of the width of the interval of meaning the values of the input variable on the estimating accuracy of the resulting models. It is also shown that the lowest value of the mean absolute error for estimating the points of fractional composition is provided by models based on neural networks and K-OPLS for various training and testing options.

© **Alexander A. Plotnikov** – Student FEFU, e-mail: Plotnikov.aal@dvfu.ru.
Denis V. Stakin – Recearcher IACP FEB RAS, Postgaduate Student FEFU, e-mail: Dshtakin21@ya.ru.
Oleg Yu. Snegirev – CSc of Technical Sciences, recearcher IACP FEB RAS, e-mail: Snegirevoleg@iacp.dvo.ru, ORCID 0000-0002-0322-6609.
Andrei Yu. Torgashov – Doctor of Technical Sciences, principal recearcher IACP FEB RAS; Professor FEFU, e-mail: Torgashov@iacp.dvo.ru, ORCID 0000-0003-3299-0742.

Введение

В химической технологии нефтепереработка – один из самых нагруженных процессов с точки зрения компонентного состава входящих и выходящих из аппаратов потоков, что сильно усложняет управление технологическим процессом (ТП) и контроль качества выпускаемой продукции. Системы управления и мониторинга современных установок нефтепереработки позволяют отслеживать и изменять множество показателей качества продуктов ТП [1]. С целью обеспечения выпуска всех выделяемых фракций с требуемым составом осуществляется лабораторный контроль, периодичность которого редко превышает 1–2 раза в сутки. В ходе лабораторного анализа зачастую определяются точки фракционного состава (ФС): температура начала кипения и температуры выкипания определенных объемных долей смеси углеводородов (УВ) [2].

Применение виртуальных анализаторов (ВА) позволяет оценивать трудноизмеримые параметры ТП более оперативно, с меньшими экономическими затратами, а также исключает возможность влияния человеческого фактора [3]. Результаты оценки ВА, как правило, используются для решения задач оптимального управления и находят применение в составе систем усовершенствованного управления ТП (СУУТП) [4].

Модели в составе ВА отражают зависимости между трудно- и легкоизмеримыми показателями ТП (в основном давление, температура, расход), которые являются входными переменными. Точность ВА зависит от множества факторов, основными из которых являются корректность выбора входных переменных, размер используемого для построения модели массива данных, периодичность представленных наблюдений, диапазон изменения в них каждого параметра и метод построения математической модели [3; 5].

При выборе метода построения модели ВА в промышленности основной проблемой остается баланс между сложностью применяемой модели и точностью ее оценок. В настоящее время повышенный интерес вызывают ВА на основе нейронных сетей [3, 6–10], обеспечивающие высокую точность оценок, однако на практике зачастую используются методы множественной линейной регрессии ввиду простоты их реализации и технической поддержки в составе СУУТП [11].

В данной работе приводится сравнительный анализ методов построения математических моделей для ВА колонны фракционирования на примере керосиновой фракции. Описан метод определения оптимального промежутка усреднения значений параметров ТП, ранее не встречающийся в литературе. Тестирование полученных моделей проводилось в условиях пропусков в данных в обучающей выборке, так как в реальных условиях определенные режимы работы технологической установки могут часто не встречаться. Рассмотрены вопросы точности полученных моделей в условиях тестирования на массиве данных, не входящем в диапазон обучения по значению выходной переменной.

Описание технологического объекта

В настоящем исследовании рассматривается колонна фракционирования (рис. 1). В колонне фракционирования происходит разделение подаваемого сырья (смеси УВ) на нестабильную нафту, керосиновую (КФ), легкую (ЛДФ) и тяжелую (ТДФ) дизельные фракции и непроконвертированный остаток. Колонна фракционирования К1 имеет три боковые стриппинг-колонны: колонна отпаривания керосина К2, колонна отпаривания легкого дизельного топлива К3 и колонна отпаривания тяжелого дизельного топлива К4. Колонна фракционирования К1 имеет два потока циркуляционного орошения: верхнее циркуляционное орошения (ВЦО) и нижнее циркуляционное орошение (НЦО).

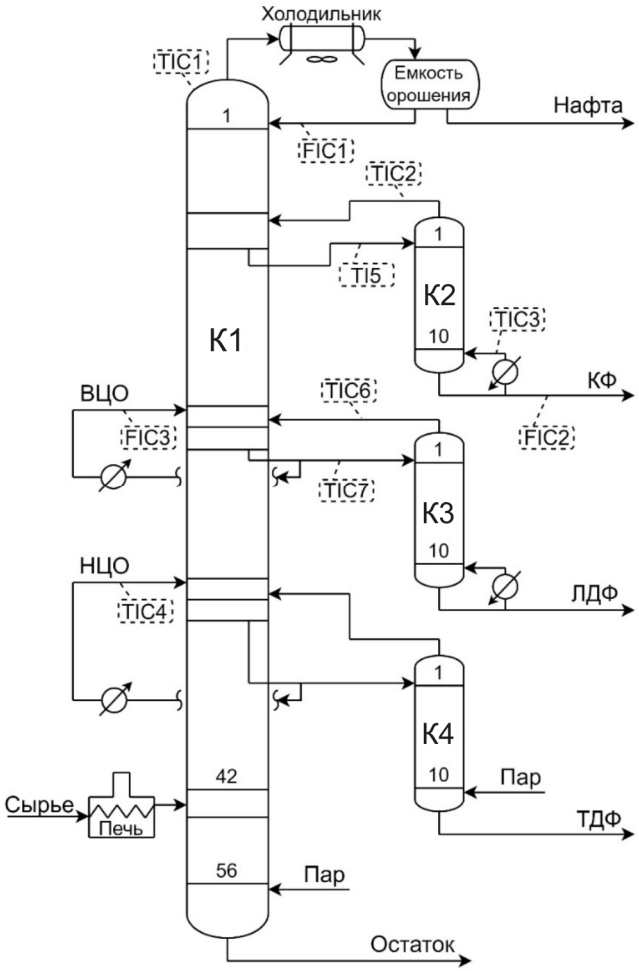


Рис. 1. Схема ТП фракционирования

Выходящий из колонны К2 поток керосиновой фракции в зависимости от режима работы установки может использоваться в качестве продуктового авиационного керосина или подаваться на смешение с легким дизельным топливом для получения летнего, зимнего или арктического дизельных топлив.

Трудноизмеримыми показателями, которые необходимо оценивать, являются точки ФС: температура начала кипения (ТНК) и температуры выкипания 10, 50, 90 и 98 % об. смеси УВ (ФС10, ФС50, ФС90, ФС98 соответственно) для потока керосиновой фракции.

На рис. 1 представлены технологические параметры ТП, используемые при построении ВА в качестве входных переменных. В данной работе из множества всех параметров технологической установки выбор входных переменных для построения ВА осуществляется в ходе корреляционного анализа, совмещенного с экспертным мнением технолога.

Отобранный набор наиболее информативных входных переменных для оценки точек ФС керосиновой фракции представлен в табл. 1.

Построение виртуальных анализаторов

Для использования промышленных данных для построения ВА необходимо провести их предобработку. В начале необходимо удалить наблюдения с потенциально некорректными значениями – выбросами, выходящими за диапазоны режимного функционирования колонны. Данные выбросы могут ухудшить качество ВА в период основных режимов функционирования технологического объекта.

Таблица 1

Общий набор входных переменных

№ п/п	Обозначение	Наименование	Единицы измерения
1	TIC1	Температура верхней секции колонны К1	°С
2	FIC1	Расход орошения колонны К1	$м^3/ч$
3	TIC2	Температура паров колонны отпаривания К2	°С
4	TIC3	Температура циркулирующего потока куба К2	°С
5	TIC4	Температура НЦО	°С
6	FIC2	Расход керосиновой фракции с куба К2	$м^3/ч$
7	TI5	Температура бокового потока КФ	°С
8	TIC6	Температура паров колонны отпаривания К3	°С
9	TIC7	Температура бокового потока ЛДФ	°С
10	FIC3	Расход ВЦО	$м^3/ч$

В табл. 2 представлены наборы входных переменных, сформированные из элементов общего набора, для каждого ВА по рассматриваемым точкам ФС.

Таблица 2

Набор входных переменных для ВА по каждой точке ФС

Точка ФС	Номер входной переменной из табл. 1
ТНК	1, 2, 3, 4, 5, 6, 7
ФС10	1, 2, 3, 4, 5, 6, 7
ФС50	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
ФС90	3, 4, 5, 6, 7, 8, 9, 10
ФС98	3, 4, 5, 6, 7, 8, 9, 10

Для построения ВА на основе статистических моделей осуществляется привязка значений параметров ТП, сохраненных с интервалом в 10 мин, к данным лабораторного анализа точек ФС, периодичность которого составляет 12–24 ч. Учитывая, что лабораторный анализ требует значительного количества времени, а также возможен факт некорректного внесения даты и времени в систему лабораторного контроля, производится усреднение значений по каждой из входных переменных за определенный промежуток времени (рис. 2):

$$X_{j,i} = \{x_{j,(i-\theta-\tau)} \dots x_{j,(i-\theta)}\}, \quad (1)$$

где $X_{j,i}$ – набор усредняемых значений j-й входной переменной для привязки к i-му наблюдению выходной переменной, i – номер наблюдения из данных лабораторного анализа, θ – сдвиг относительно соответствующего времени, τ – ширина времени окна усреднения.

В рамках данной работы проведено исследование, направленное на определение оптимальных значений параметров θ и τ для построения ВА по точкам ФС керосиновой фракции.

Построение моделей осуществлялось методом множественной линейной регрессии, при этом оценка точности моделей происходила на всем массиве данных для обучения. Критерием выбора оптимального варианта усреднения является значение средней абсолютной ошибки (САО) полученных моделей:

$$e = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \quad (2)$$

где e – средняя абсолютная ошибка, n – число наблюдений.

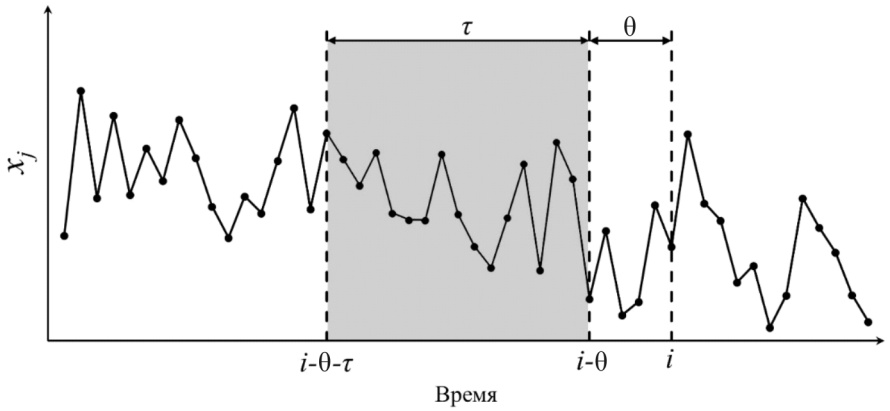


Рис. 2. Окно усреднения значений входной переменной
для привязки к значению выходной переменной

Массив данных вида «вход – выход» после удаления выбросов и привязки усредненных значений входных переменных к значениям выходной состоит из 688 наблюдений, из которых далее формируются обучающая выборка (ОВ) и тестовая выборка (ТВ).

Для проверки качества моделей в составе ВА в условиях пропуска данных в ОВ формирование ОВ и ТВ из общего набора осуществляется путем выбора диапазона изменения выходной переменной и включения в ОВ только наблюдений, входящих в заданный диапазон. Интересным и важным с точки зрения практики является случай, когда в ТВ присутствуют наблюдения, которые *вообще отсутствуют* в ОВ.

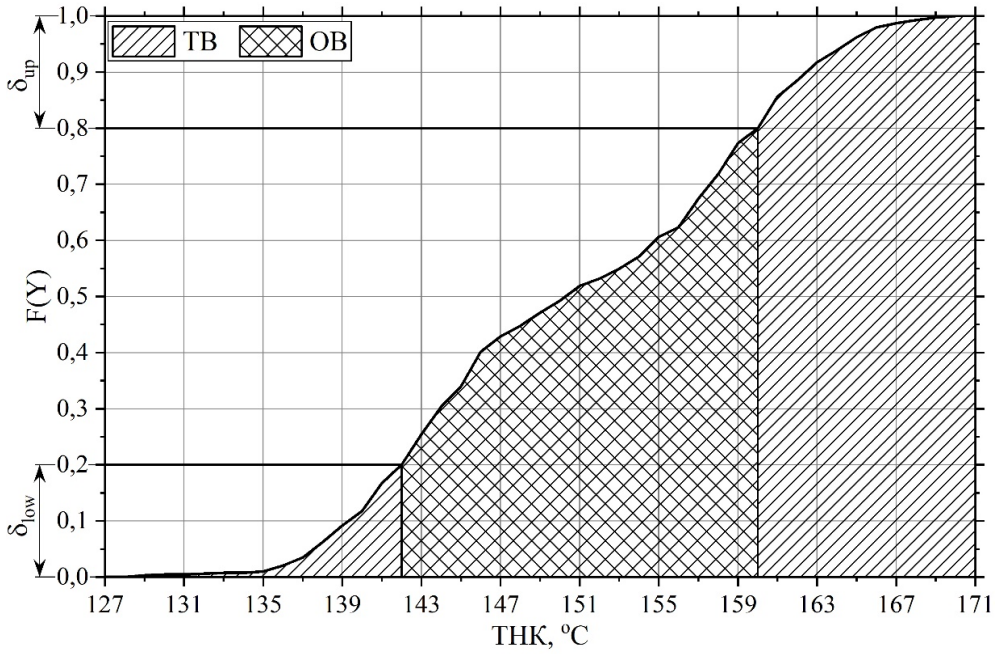


Рис. 3. График кумулятивной функции распределения значений
выходной переменной с указанием диапазона для ОВ и ТВ

В рамках данной работы предполагается, что диапазон данных выходной переменной в ОВ находится условно в середине от общего количества наблюдений, а для отражения меры пропусков данных на «хвостах» кумулятивной функции распределения выхода (рис. 3) вводится в рассмотрение параметр δ. Соответственно, в ТВ попадают *только* наблюдения с нижнего и верхнего краев всего диапазона. Граничные значения δ

зависят от каждого практического случая производственного задания и технологического режима фракционатора, и в ТВ присутствует определенное количество (доля) от всех наблюдений y нижней (δ_{low}) и верхней (δ_{up}) границ известного диапазона изменения выходной переменной. На рис. 3 показан пример нахождения границ диапазона данных входящих в ОВ с использованием кумулятивной функции распределения значений выходной переменной на примере ТНК керосиновой фракции.

По описанным выше сформированным выборкам для обучения и тестирования необходимо построить ВА для оценки ТНК, ФС10, ФС50, ФС90 и ФС98 с использованием следующих методов:

– множественная робастная регрессия (РР) с различными весовыми функциями;

– гребневая регрессия (ГР);

– метод ортогональных проекций на скрытые структуры на основе ядра (англ. Kernel-based Orthogonal Projections to Latent Structures, K-OPLS);

– метод чередующихся условных математических ожиданий (англ. Alternating conditional expectations, ACE);

– нейросети прямого распространения (НСПР).

Робастная регрессия

Один из базовых методов регрессионного анализа – метод наименьших квадратов (МНК), основным недостатком которого является высокая чувствительность к наблюдениям, значения которых сильно отклоняются от среднего [12]. Коэффициенты множественной линейной регрессии по МНК могут быть вычислены по формуле:

$$b_{\text{МНК}} = \left(X^{\text{T}}X\right)^{-1} X^{\text{T}}y, \tag{3}$$

где $b_{\text{МНК}}$ – матрица коэффициентов уравнения регрессии, полученная методом наименьших квадратов, X – матрица наблюдений входных переменных, X^{T} – транспонированная матрица, y – вектор измеренных значений выходной переменной.

Для повышения устойчивости к выбросам, минимизации ошибки оценки применяется робастная регрессия (РР). В этом случае в качестве множителя добавляется диагональная матрица весовых коэффициентов, определяемых весовой функцией (табл. 3):

$$b_{\text{РР}} = \left(X^{\text{T}}WX\right)^{-1} X^{\text{T}}Wy, \tag{4}$$

где $b_{\text{РР}}$ – матрица коэффициентов уравнения регрессии, полученная методом робастной регрессии, W – диагональная матрица весовых коэффициентов.

Таблица 3

Весовые функции для методов РР

Метод	Весовая функция	Стандартный весовой коэффициент
МНК	–	–
РР1	$\omega_1 = \lvert r\rvert \cdot \sin(r)/r,\ \lvert r\rvert < \pi$	1,339
РР2	$\omega_2 = \lvert r\rvert \cdot (1 - r^2)^2/r,\ \lvert r\rvert < 1$	4,685
РР3	$\omega_3 = 1/(1 + r^2)$	2,385
РР4	$\omega_4 = 1/(1 + \lvert r\rvert)$	1,400
РР5	$\omega_5 = 1/\max(1, \lvert r\rvert)$	1,345
РР6	$\omega_6 = \tanh(r)/r$	1,205
РР7	$\omega_7 = 1 \cdot \lvert r\rvert,\ \lvert r\rvert < 1$	2,795
РР8	$\omega_8 = \exp(-(r^2))$	2,985

Добавление матрицы весовых коэффициентов позволяет изменять веса значений наблюдений, зачастую – более высокий вес для наблюдений с меньшим отклонением и меньший вес для наблюдений с более высоким отклонением [13].

Основным преимуществом линейных регрессионных моделей является простая интерпретируемость полученных результатов – коэффициенты b_i для каждой входной переменной в общем уравнении множественной линейной регрессии могут быть оценены технологом на предмет соответствия физическому смыслу каждой из переменных, скорректированы и в дальнейшем использоваться в качестве модели ВА. Уравнение множественной линейной регрессии имеет вид:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ..., \tag{5}$$

где b_0 – свободный член регрессии, b_j – коэффициент предиктора, x_j – значение наблюдения предиктора.

В данной работе на каждом этапе сравнения точности оценки полученных моделей из моделей РР с различными весовыми коэффициентами выбиралась модель РР, обеспечивающая наименьшую САО на тестовой выборке.

Гребневая регрессия

При рассмотрении объекта химико-технологического производства очень часто возникает проблема мультиколлинеарности входных переменных, используемых в модели, так как большая часть из них представляет собой термодинамические параметры ТП, связанные уравнениями тепло- и массообмена. Одним из способов решения данной проблемы является применения гребневой регрессии:

$$b_{\text{ГР}} = \left(X^T X + \lambda I \right)^{-1} X^T y, \tag{6}$$

где $b_{\text{ГР}}$ – матрица коэффициентов уравнения регрессии, полученная методом гребневой регрессии; λ – параметр регуляризации (гребень регрессии); I – единичная матрица.

Модель ГР включает параметр, обеспечивающий смещение значений всех наблюдений, также называемый гребнем регрессии. При использовании гребневой регрессии параметры модели определяются в результате решения задачи оптимизации:

$$\min\left(\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} b_{\text{ГР},j} x_{j,i} - y_i\right)^2 + \lambda \sum_{j=1}^{m} b_{\text{ГР},j}^2\right), \tag{7}$$

где m – число входных переменных.

Регуляризация позволяет избежать переобучения модели путем сглаживания больших значений модулей коэффициентов $b_{\text{ГР},j}$.

В данной работе также исследовалась зависимость качества оценивания моделей от величины параметра гребня регрессии. Учитывая различные размерности входных переменных, подбор параметра смещения λ производился для исходной и нормированной матриц наблюдений. Существует множество различных способов определения оптимального параметра λ [14; 15], однако в ходе проведения исследования установлено, что применение параметра смещения λ в обоих случаях (с нормализацией данных и без) не дает значительного изменения с точки зрения качества оценивания.

Метод ортогональных проекций на скрытые структуры на основе ядра

K-OPLS – метод множественной линейной регрессии через построение ортогональных проекций на скрытые структуры на основе функции ядра. Среди преимуществ K-OPLS выделяют устойчивость к выбросам и мультиколлинеарности входных переменных [16]. Матрица входных данных X представляется следующим образом:

$$X = T_p P_p^{\mathrm{T}} + T_o P_o^{\mathrm{T}} + E, \tag{8}$$

где T_p – y-прогнозная матрица вкладов; P_p^{T} – y-прогнозная матрица нагрузки; T_{o} – соответствующая y-ортогональная матрица вклада; P_o^{T} – соответствующая y-ортогональная матрица загрузки; E – матрица остатков.

В данной работе в качестве функции ядра использовалась функция Гаусса:

$$k(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2). \tag{9}$$

Для функции Гаусса определялось оптимальное значение параметра σ, который обеспечивает наименьшую САО на обучающей и тестовой выборках.

Как правило, для всех построенных моделей на основе K-OPLS минимальные значения САО на обучении и тесте наблюдаются при достаточно близких значениях параметра σ (рис. 4).

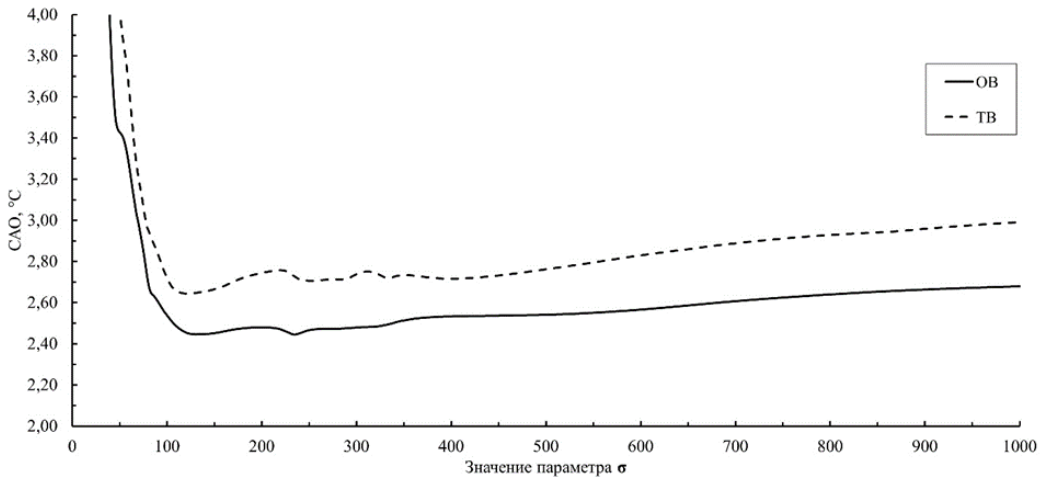


Рис. 4. График зависимости САО на обучающей и тестовой выборках при различных значениях параметра σ при оценке ТНК керосиновой фракции

Метод чередующихся условных математических ожиданий

ACE – один из непараметрических методов регрессионного анализа, в ходе которых значения входных и выходной переменных заменяются функциональными преобразованиями в процессе построения математической модели. Общий вид уравнения регрессии для ACE [17; 18]:

$$\theta(y) = \sum_{j=1}^{m} \varphi_j(x_j), \tag{10}$$

где $\theta(y)$ – оптимальные преобразования значений оцениваемой переменной; $\varphi_j(x_j)$ – оптимальные преобразования значений входных переменных.

В ходе определения функциональных преобразований итеративным подбором параметров минимизируется доля необъясненной дисперсии [17]:

$$\sum_{i=1}^{n}(\theta(y_i)-\sum_{j=1}^{m}\varphi_{j,i}(x_{j,i}))^2 \to \min. \qquad (11)$$

Пример построения оптимальных преобразований $\varphi_j(x_j)$ по ТНК керосиновой фракции показан на рис. 5.

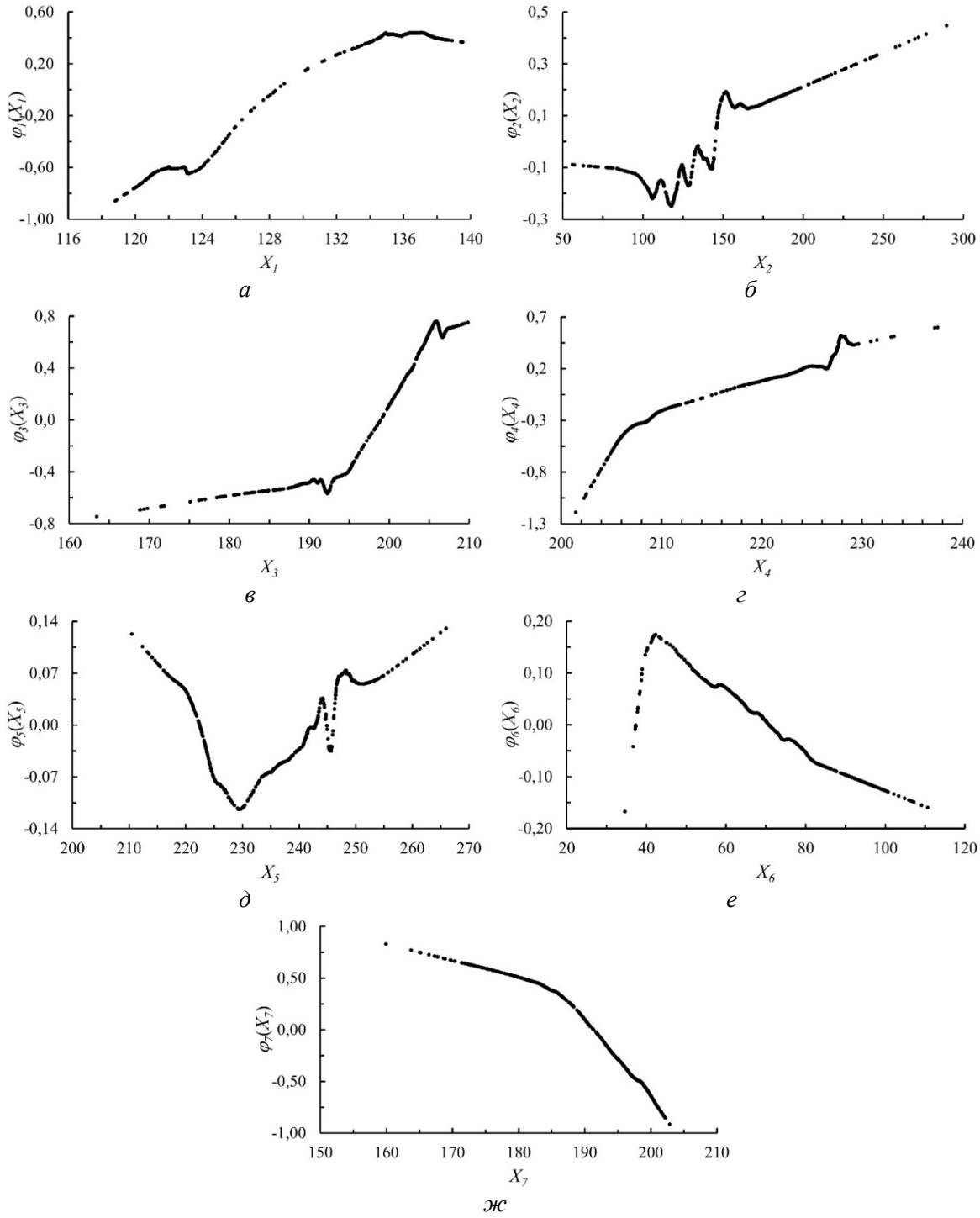


Рис. 5. Зависимости оптимальных преобразований по каждой входной переменной для ТНК керосиновой фракции (*а – ж*)

Обычно для аппроксимации используются экспоненциальные, логарифмические или полиномиальные зависимости [18].

Для зависимостей, полученных в данной работе, более применимы кусочные функции, поскольку некоторые участки в явном виде представляют собой прямые линии, для других более подходящими являются полиномиальные или экспоненциальные зависимости. Для снижения вычислительных затрат при использовании метода в данной работе применялась кусочно-линейная аппроксимация $\varphi_j(x_j)$ для каждого из наблюдений путем нахождения коэффициентов уравнения прямой, проходящей через две ближайшие точки для значения каждой из переменных.

Исходя из уравнения (10), оценка значения выходной переменной осуществляется с использованием обратной функции:

$$\hat{y} = \theta^{-1}(\sum_{j=1}^{m} \varphi_j(x_j)). \tag{12}$$

В качестве функции θ^{-1} использовалась модель линейной регрессии. На рис. 6 представлена зависимость значений выходной переменной y (на примере ТНК на ОВ) от суммы оптимальных преобразований $\sum_{j=1}^{m} \varphi_j(x_j)$, а также подобранная функция θ^{-1}.

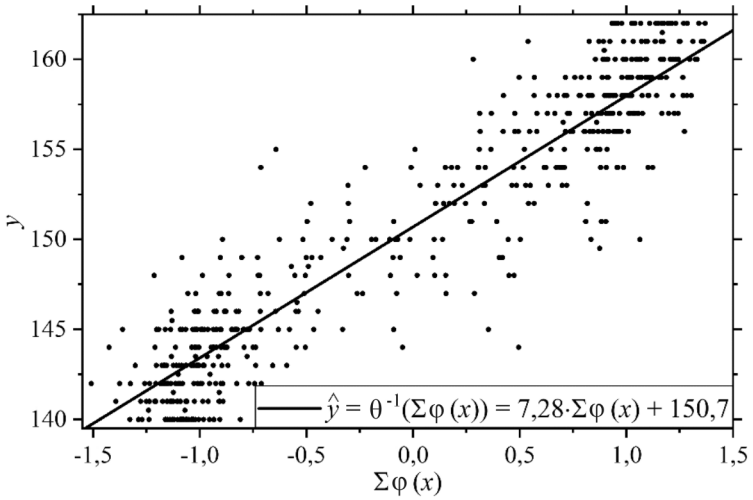


Рис. 6. Зависимость значения оцениваемой переменной y от суммы оптимальных преобразований входных переменных и функция θ^{-1} на примере ТНК на ОВ

Нейронные сети

Для сравнительного анализа с другими методами построения статистических моделей использовалась наиболее распространенная нейронная сеть прямого распространения с двумя промежуточными слоями нейронов, структура которой представлена на рис. 7.

Важным фактором, который следует учитывать при сравнении с другими методами, является структура нейронной сети – ее гиперпараметры [19]: метод обучения, функция активации, число нейронов в промежуточных слоях, от которых напрямую зависят результаты оценки, время и сходимость обучения, так как свойства нейросети также сильно зависят от весов связей между нейронами.

В данной работе проведен анализ влияния гиперпараметров на точность оценки моделей путем сравнения всевозможных комбинаций с различными числами нейронов в про-

межуточных слоях (в обоих слоях число нейронов устанавливалось одинаковое), функция активации и методами обучения.

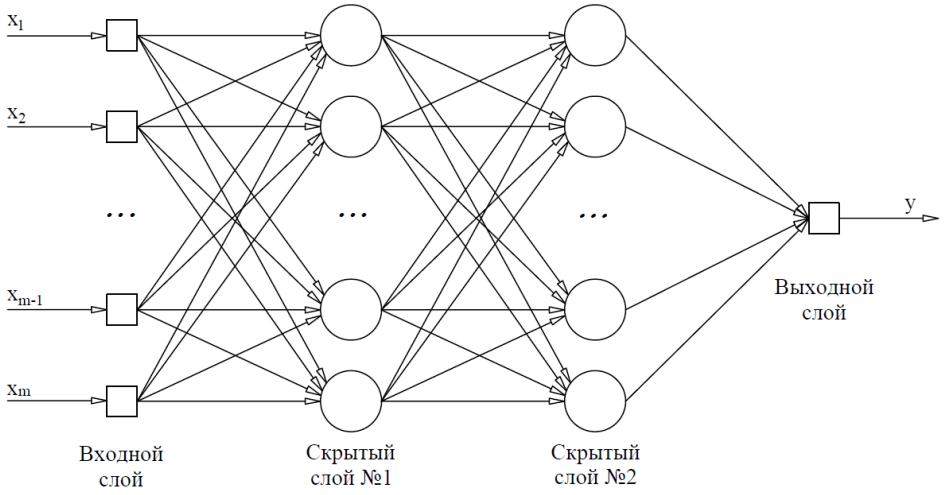


Рис. 7. Структура нейронной сети прямого распространения с двумя промежуточными слоями

В результате анализа выявлено, что для значительной части моделей наиболее предпочтительно использование числа нейронов в обоих скрытых слоях, равное числу входных переменных.

Анализ функционирования виртуальных анализаторов в условиях пропусков данных

При определении оптимальных значений параметров θ и τ для всех случаев рассматриваемый временной интервал составляет 10 ч (600 мин) от времени лабораторного анализа с шагом 10 мин. На рис. 8 представлены зависимости САО оценок ТНК от изменений значений ширины окна усреднения θ и сдвига относительно времени лабораторного анализа τ:

а) ширина изменяется ($\tau = \{0, 10 \ldots 600\}$), сдвиг фиксированный ($\theta = 0$);

б) ширина фиксированная ($\tau = 60$), сдвиг изменяется ($\theta = \{0, 10 \ldots 600\}$);

в) ширина фиксированная ($\tau = 120$), сдвиг изменяется ($\theta = \{0, 10 \ldots 600\}$);

г) ширина фиксированная ($\tau = 0$), сдвиг изменяется ($\theta = \{0, 10 \ldots 600\}$).

Для каждого варианта окна усреднения выбран диапазон, при котором обеспечивается наименьшее значение САО (табл. 4).

Определенный оптимальный промежуток времени усреднения для значений технологических параметров – значения в промежутке между двумя часами до и указанным временем лабораторного анализа – характерен для всех точек ФС.

Рассмотрены различные варианты интервалов (в зависимости от δ) пропусков данных выходной переменной в ОВ относительно общего 100%-ного набора наблюдений. В реальных условиях для промышленного фракционатора наблюдались следующие варианты распределения данных ОВ:ТВ

– вариант I – 80:20%; $\delta_{low} = \delta_{up} = 0,1$;

– вариант II – 60:40%; $\delta_{low} = \delta_{up} = 0,2$;

– вариант III – 40:60%; $\delta_{low} = \delta_{up} = 0,3$.

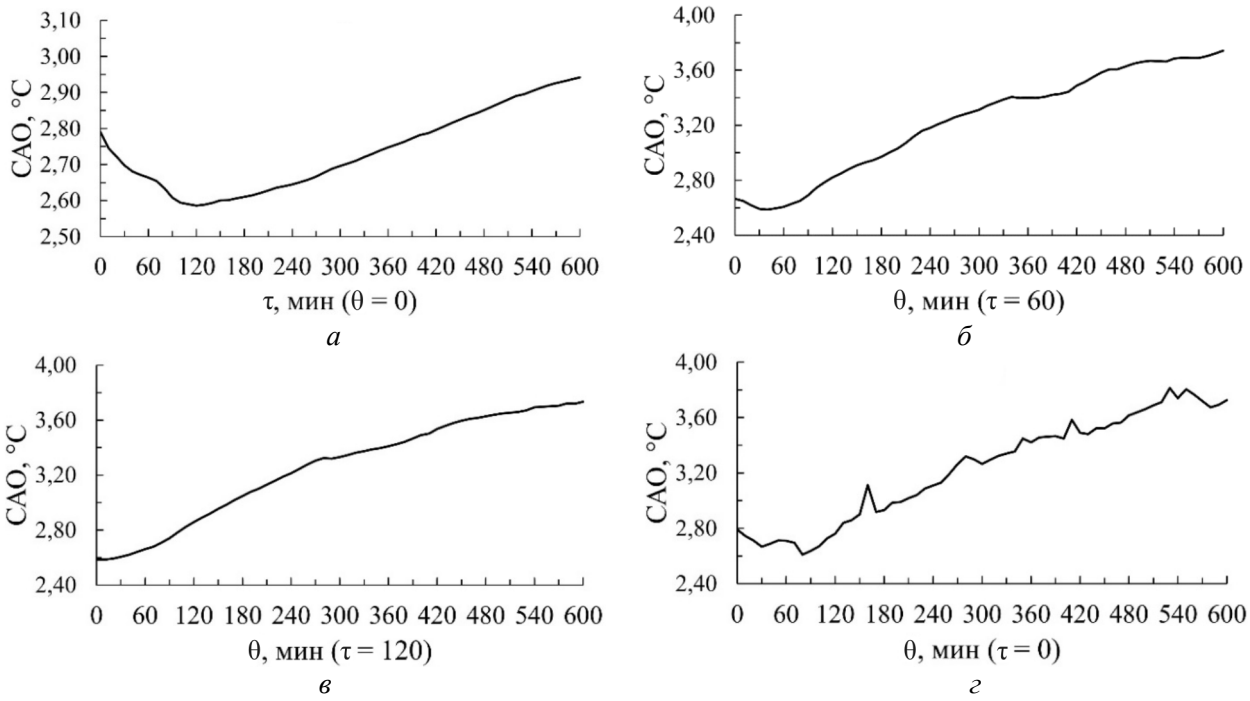


Рис. 8. Зависимость САО оценок ТНК от значений ширины окна усреднения θ и сдвига относительно времени лабораторного анализа τ

Таблица 4

Диапазон усреднения, обеспечивающий наименьшее значение САО оценок ТНК

Окно усреднения	Диапазон усреднения, мин	θ, мин	τ, мин	Наим. САО, °С
а	0 – 120	0	120	*2,5862*
б	40 – 100	40	60	2,5871
в	10 – 130	10	120	***2,5857***
г	80	80	0	2,6100

Таблица 5

Значение САО на ТВ для пяти точек ФС

Точка ФС	Вариант распределния данных в ОВ и ТВ	Метод построения математической модели (или метод обучения)				
		RR	ГР	K-OPLS	ACE	НСПР
ТНК	I	4,5116	4,7788	***2,6029***	5,1121	*2,8138*
	II	4,2941	4,7904	*2,5347*	5,0317	***2,1818***
	III	4,4828	5,1114	*2,4793*	5,3679	***2,1647***
ФС10	I	3,0168	3,2640	*1,7011*	3,8401	***1,6812***
	II	2,6051	2,9133	*1,5334*	3,5848	***1,3392***
	III	2,7650	3,2471	*1,4566*	3,8538	***1,3525***
ФС50	I	3,6872	4,1191	***1,7735***	4,4822	*1,9917*
	II	3,7267	4,4250	*1,9273*	4,9175	***1,7843***
	III	4,3889	5,3806	*1,8718*	5,8927	***1,3045***
ФС90	I	11,3999	12,0391	*4,1147*	7,6949	***3,6265***
	II	9,8615	11,5517	*4,9411*	9,4105	***3,8980***
	III	11,0105	12,1366	*5,7212*	12,7038	***2,7045***
ФС98	I	13,4600	14,5509	*4,3915*	10,9226	***3,3892***
	II	12,4834	14,1544	*5,7776*	11,2786	***3,7362***
	III	12,5697	13,8595	*6,9335*	12,2250	***3,1832***

Результаты сравнения различных моделей в составе ВА для точек ФС при рассматриваемых вариантах распеределения данных в ОВ и ТВ представлены в табл. 5, а также показаны на рис. 9.

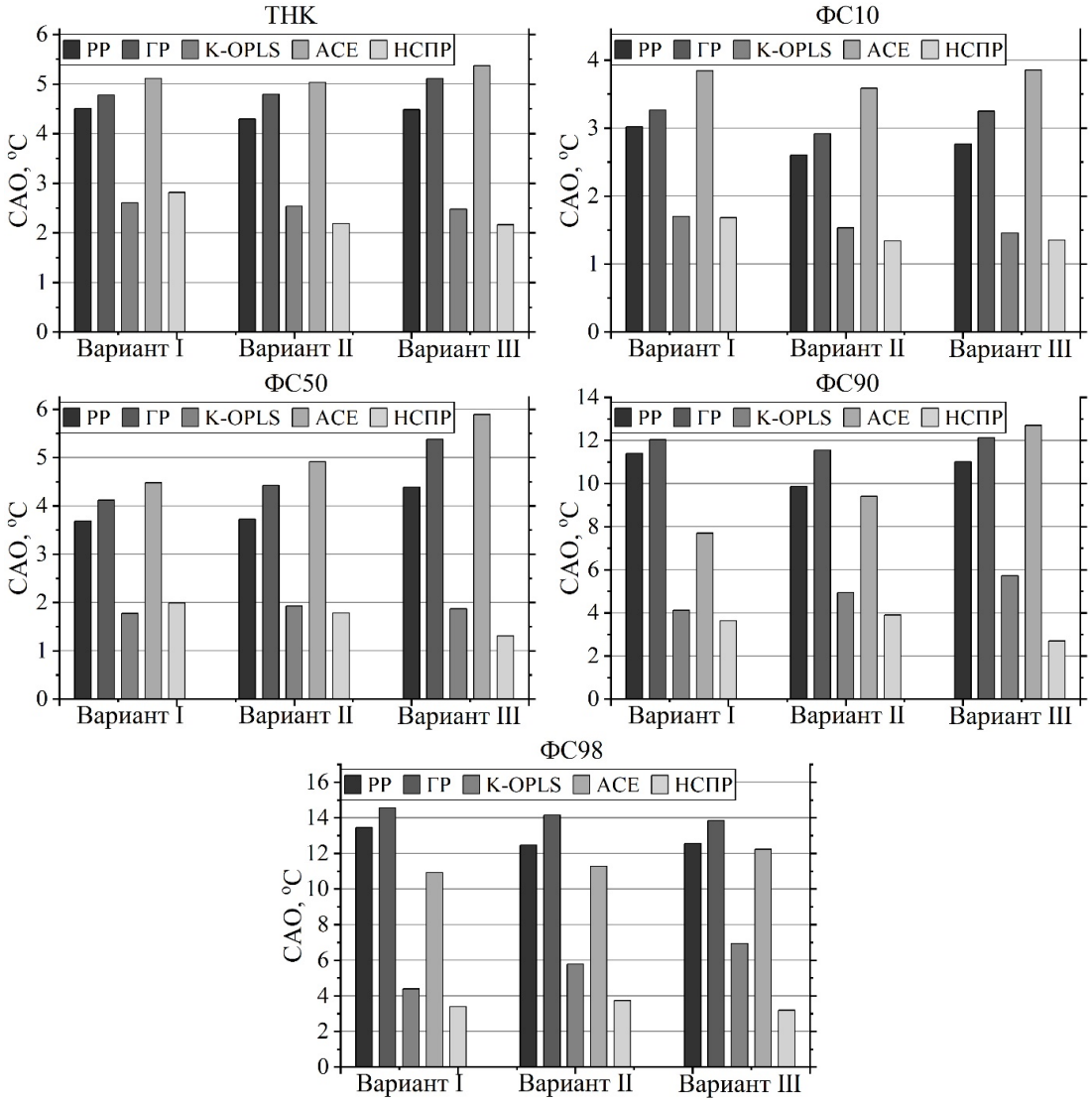


Рис. 9. Сравнение моделей в составе ВА для пяти точек ФС при различных вариантах пропусков данных в ОВ

Распространенным вариантом для оценки точности полученных математических моделей является коэффициент детерминации R^2. Однако значение САО при оценке точек ФС более применимо ввиду большей наглядности полученных результатов. Значения точек ФС определяются согласно стандарту ASTM D86, из которого следует предельно допустимая сходимость полученных результатов для всех точек ФС, приведенная в табл. 6 [20].

Таблица 6

Сходимость полученных результатов при определении точек ФС

Точка ФС	ТНК	ФС10	ФС50	ФС90	ФС98
ΔT, °С	3,5	4,1	4,1	3,3	3,3

Как видно из полученных результатов в табл. 5, температуры, определяющие содержания легких (ТНК и ФС10) и средних компонентов (ФС50) в смеси УВ, отличаются более

низкими значениями допусков ошибок оценок. Для ТНК, ФС10 и ФС50 высокую точность оценки показывают модели на основе НСПР и K-OPLS.

Большие значения ошибок при оценке содержания тяжелых компонентов (ФС90 и ФС98) характерны при использовании всех методов построения моделей. Однако для моделей на основе НСПР увеличение ошибки менее значительно, чем при использовании других рассмотренных методов.

Линейные регрессионные модели (РР и ГР) ввиду своей простоты не могут обеспечить высокую степень обобщения данных, поэтому показывают более низкую точность.

Для ACE важно, чтобы тестовое значение входной переменной находилось внутри диапазона значений ОВ. В этом случае для оценки значения используется известный участок графической зависимости $\varphi_j(x_j)$. В случае, если тестовое значение находится вне диапазона ОВ, увеличение ошибки данной оценки напрямую зависит от удаленности этого значения.

K-OPLS относится к методам, способным учитывать нелинейные зависимости в данных [16] на основе функции ядра, поэтому показывает высокую точность при оценке температур начала и середины кипения. Однако в нашем случае эффективности метода недостаточно для поиска всех связей между входными и выходной переменными при оценке температур конца кипения.

НСПР показывают высокую точность оценки всех точек ФС и также демонстрируют высокую степень обобщения в условиях пропусков в данных, соответствующих определенному технологическому режиму. Использование метода ACE оказалось не предпочтительным. При увеличении сегмента пропуска данных в ОВ, т.е. при возрастании δ_{low} или δ_{up}, повышается риск нахождения значений ТП для каждого из наблюдений за пределами диапазона получаемых зависимостей оптимальных преобразований.

Заключение

При проведении сравнительного анализа моделей в составе ВА в условиях пропусков данных в ОВ в условиях рассмотренных практических вариантов распределения данных в ОВ и ТВ выявлено, что модели K-OPLS и НСПР показали высокую точность оценки выходной переменной в сравнении с РР, ГР и ACE. Так, для модели K-OPLS в составе ВА значение САО в среднем меньше для ТНК на 42,7 %, для ФС10 на 44 %, для ФС50 на 52,5 % для ФС90 на 54 % и для ФС98 на 55,3 % в сравнении с распространенной на производстве линейной регрессией. Для модели НСПР в составе ВА значение САО в среднем меньше для ТНК на 46,2 %, для ФС10 на 48 %, для ФС50 на 56 % для ФС90 на 68 % и для ФС98 на 73,2 % в сравнении с распространенной на производстве линейной регрессией.

Таким образом, наилучший результат в условиях пропуска данных в ОВ при оценке точек фракционного состава обеспечивают модели на основе НСПР.

Также рассмотрен вопрос усреднения значений входных переменных для привязки к значению выходной переменной, определенный оптимальный промежуток усреднения для всех точек ФС соответствует ширине окна в τ = 120 мин и сдвигу относительно времени лабораторного анализа θ = 0 мин.

Список литературы

1. Kadlec, P. Data-driven soft sensors in the process industry / P. Kadlec, B. Gabrys, S. Strandt // Computers and Chemical Engineering. 2009. – Vol. 33, iss. 4. – P. 795–814. DOI: 10.1016/j.compchemeng.2008.12.012

2. Iplik, E. Hydrocracking: a perspective towards digitalization / E. Iplik, I. Aslanidou, K. Kyprianidis // Sustainability (Switzerland). 2020. – Vol. 12, iss. 17. – 26 p. DOI: 10.3390/su12177058

3. The role of artificial intelligence-driven soft sensors in advanced sustainable process industries: A critical review / Y.S. Perera, D.A.A.C. Ratnaweera, C.H. Dasanayaka, C. Abeykoon // Engineering Applications of Artificial Intelligence. 2023. – Vol. 121, iss. 105988. – 24 p. DOI: 10.1016/j.engappai.2023.105988

4. King, M. Process control. A practical approach / M. King // Wiley. – 2016. – 2 Ed. – 623 p.

5. Data-driven prediction of product yields and control framework of hydrocracking unit / Z. Pang, P. Huang, C. Lian, C. Peng, X. Fang, H. Liu // Chemical Engineering Science. 2024. – Vol. 283, iss. 119386 – 10 p. DOI: 10.1016/j.ces.2023.119386

6. A layer-wise data augmentation strategy for deep learning networks and its soft sensor application in an industrial hydrocracking Process / X. Yuan, C. Ou, Y. Wang, C. Yang, W. Gui // IEEE Trans Neural Netw Learn Syst. 2021. – Vol. 32, iss. 8. – P. 3296–3305. DOI: 10.1109/TNNLS.2019.2951708

7. Rani, A. Development of soft sensor for neural network based control of distillation column / A. Rani, V. Singh, J.R.P. Gupta // ISA Transactions. 2013. – Vol. 52, iss. 3. – P. 438–449. DOI: 10.1016/j.isatra.2012.12.009

8. Wang, Y. A two-layer ensemble learning framework for data-driven soft sensor of the diesel attributes in an industrial hydrocracking process / Y. Wang, D. Wu, X. Yuan // Journal of Chemometrics. 2019. – Vol. 33, iss. 12. – 14 p. DOI: 10.1002/cem.3185

9. Popoola, L.T. A Review of an expert system design for crude oil distillation column using the neural networks model and process optimization and control using genetic algorithm framework / L.T. Popoola, G. Babagana, A.A. Susu // Advances in Chemical Engineering and Science. 2013. – Vol. 3, iss. 2. – P. 164–170. DOI: 10.4236/aces.2013.32020

10. Soft-sensor design for a crude distillation unit using statistical learning methods / A. Urhan, N.G. Ince, R. Bondy, B. Alakent // Computer Aided Chemical Engineering. 2018. – Vol. 44. – P. 2269–2274. DOI: 10.1016/B978-0-444-64241-7.50373-6

11. Kano, M. The state of the art in advanced chemical process control in Japan / M. Kano, M. Ogawa // IFAC Proceedings Volumes. 2009. – Vol. 7, iss. 1. – P. 10–25. DOI: 10.3182/20090712-4-TR-2008.00005

12. Hinich, M.J. A simple method for robust regression / M.J. Hinich, P.P. Talwar // Journal of the American Statistical Association. 1975. – Vol. 70, iss. 349. – P. 113–119. DOI: 10.1080/01621459.1975.10480271

13. Дрейпер, Н. Прикладной регрессионный анализ / Н. Дрейпер, Г. Смит. – М.: Финансы и статистика, 1986. – Кн. 2. – 351 с.

14. Alheety, M.I. Choosing ridge parameters in the linear regression model with AR(1): A comparative simulation study / M.I. Alheety, B.M.G. Kibria // International Journal of Statistics and Economics. – 2011. – Vol. 7, iss. 11. – 18 p.

15. Khalaf, G. Choosing ridge parameter for regression problems / G. Khalaf, G. Shukur // Communications in statistics – Theory and Methods. 2005. – Vol. 34, iss. 5. – P. 1177–1182. DOI: 10.1081/STA-200056836

16. K-OPLS package: kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space / M. Bylesjo, M. Rantalainen, J. K. Nicholson, E. Holmes, J. Trygg // BMC Bioinformatics. 2008. – Vol. 9, iss. 106. – 7 p. DOI: 10.1186/1471-2105-9-106

17. Correlating Bacharach Opacity in Fuel Oil Exhaust. Prediction of the Operating Parameters that Reduce It / M. Blanco, J. Coello, S. Maspoch, A. Puigdomenech, X. Peralta, J.M. Gonzalez, J. Torres // Oil & Gas Science and Technology. – 2000. – Vol. 55, iss. 5. – P. 533–541. DOI: 10.2516/ogst: 2000040

18. Wang, D. Identifying nonlinear relationships in regression using the ACE Algorithm / D. Wang, M. Murphy // Journal of Applied Statistics. 2005. – Vol. 32, iss. 3. – P. 243–258. DOI: 10.1080=02664760500054517

19. Li, Yang. On hyperparameter optimization of machine learning algorithms: theory and practice / Yang Li, Abdallah Shami. // Neurocomputing. 2020. – Vol. 415. – P. 295–316. DOI: 10.1016/j.neucom.2020.07.061

20. ASTM D86 – 23. Standard test method for distillation of petroleum products at atmospheric pressure // American National Standard. ASTM International. – 2023. – 22 p.

References

1. Kadlec P., Gabrys B., Strandt S. Data-driven soft sensors in the process industry. *Computers and Chemical Engineering*, 2009, vol. 33, iss. 4, pp. 795–814. DOI: 10.1016/ j.compchemeng.2008.12.012.

2. Iplik E., Aslanidou I., Kyprianidis K. Hydrocracking: a perspective towards digitalization. *Sustainability (Switzerland)*, 2020, vol. 12, iss. 17, 26 p. DOI: 10.3390/su12177058.

3. Perera Y.S., Ratnaweera D.A.A.C., Dasanayaka C.H., Abeykoon C. The role of artificial intelligence-driven soft sensors in advanced sustainable process industries: a critical review. *Engineering Applications of Artificial Intelligence*, 2023, vol. 121, iss. 105988, 24 p. DOI: 10.1016/ j.engappai.2023.105988.

4. King M. Process control. A practical approach. Wiley. Whitehouse consulting, Isle of Wight, UK, 2016, 623 p. ISBN: 9781119157755.

5. Pang Z., Huang P., Lian C., Peng C., Fang X., Liu H. Data-driven prediction of product yields and control framework of hydrocracking. *Chemical Engineering Science*, 2024, vol. 283, iss. 119386, 10 p. DOI: 10.1016/j.ces.2023.119386.

6. Yuan X., Ou C., Wang Y., Yang C., Gui W. A layer-wise data augmentation strategy for deep learning networks and its soft sensor application in an industrial hydrocracking process. *IEEE Trans Neural Netw Learn Syst*., 2021, vol. 32, iss. 8, pp. 3296–3305. DOI: 10.1109/ TNNLS.2019.2951708.

7. Rani A., Singh V., Gupta J.R.P. Development of soft sensor for neural network based control of distillation column. *ISA Transactions*, 2013, vol. 52, iss. 3, pp. 438–449. DOI: 10.1016/ j.isatra.2012.12.009.

8. Wang Y., Wu D., Yuan X. A two-layer ensemble learning framework for data-driven soft sensor of the diesel attributes in an industrial hydrocracking process. *Journal of Chemometrics*, 2019, vol. 33, iss. 12, 14 p. DOI: 10.1002/cem.3185.

9. Popoola L.T., Babagana G., Susu A.A. A review of an expert system design for crude oil distillation column using the neural networks model and process optimization and control using genetic algorithm framework. *Advances in Chemical Engineering and Science*. 2013. – Vol. 3, Iss. 2. – P. 164–170. DOI 10.4236/aces.2013.32020.

10. Urhan A., Ince N.G., Bondy R., Alakent B. Soft-sensor design for a crude distillation unit using statistical learning methods. *Computer Aided Chemical Engineering*, 2018, vol. 44, pp. 2269–2274. DOI: 10.1016/B978-0-444-64241-7.50373-6.

11. Kano M., Ogawa M. The state of the art in advanced chemical process control in Japan. *IFAC Proceedings Volumes*, 2009, vol. 7, iss. 1, pp. 10–25. DOI: 10.3182/20090712-4-TR-2008.00005

12. Hinich M.J., Talwar P.P. A simple method for robust regression // *Journal of the American Statistical Association*, 1975, vol. 70, iss. 349, pp. 113–119. DOI: 10.1080/01621459.1975.10480271

13. Draper N.R., Smith H. Applied regression analysis. Second edition. Wiley, 1981, 736 p.

14. Alheety M.I., Kibria B.M.G. Choosing ridge parameters in the linear regression model with AR(1): A comparative simulation study. *International Journal of Statistics and Economics*, 2011, vol. 7, iss. 11, 18 p. URL: https://www.researchgate.net/publication/309736495.

15. Khalaf G. Choosing ridge parameter for regression problems. *Communications in statistics – Theory and Methods*, 2005, vol. 34, iss. 5, pp. 1177–1182. DOI: 10.1081/STA-200056836.

16. 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. *BMC Bioinformatics*, 2008, vol. 9, iss. 106, 7 p. DOI: 10.1186/1471-2105-9-106.

17. Blanco M., Coello J., Maspoch S., Puigdomenech A., Peralta X., Gonzalez J.M., Torres J. Correlating bacharach opacity in fuel oil exhaust. prediction of the operating parameters that reduce it. *Oil & Gas Science and Technology*, 2000, vol. 55, iss. 5, pp. 533–541. DOI: 10.2516/ogst: 2000040.

18. Wang D., Murphy M. Identifying nonlinear relationships in regression using the ACE algorithm. *Journal of Applied Statistics*, 2005, vol. 32, iss. 3, pp. 243–258. DOI: 10.1080=02664760500054517.

19. Yang Li, Abdallah Shami. On hyperparameter optimization of machine learning algorithms: theory and practice // *Neurocomputing*. 2020. – Vol. 415. – P. 295-316. DOI 10.1016/j.neucom.2020.07.061.

20. ASTM D86 – 23. Standard test method for distillation of petroleum products at atmospheric pressure. American National Standard. ASTM International, 2023, 22 p.