Вестн. Сам. гос. техн. ун-та. Сер. Физ.-мат. науки. 2025. Т. 29, № 1. С. 174–186
ISSN: 2310-7081 (online), 1991-8615 (print) https://doi.org/10.14498/vsgtu2120
EDN: ZDTCBW

УДК 519.24:543.5:544.3

Построение модели для прогнозирования двух- и трехкомпонентных неорганических систем в водных растворах спектральным анализом

К. Ю. Массалов[1], *Е. Ю. Мощенская*[2]

[1] Национальный исследовательский ядерный университет «МИФИ», Россия, 115409, Москва, Каширское шоссе, 31.
[2] Самарский государственный технический университет, Россия, 443100, Самара, ул. Молодогвардейская, 244.

Аннотация

Представлен алгоритм и разработанная на его основе программа, реализующая методы математического моделирования для анализа спектральных данных, построения прогностической модели и выбора оптимальных спектральных интервалов при проектировании мультисенсорных систем на основе светодиодов. Алгоритм прошел апробацию на реальных смесях водных растворов неорганических солей.

Для обработки экспериментальных данных применялись методы многомерной калибровки, включая PLS-регрессию и множественную линейную регрессию. Информативные длины волн определялись с использованием значений вектора Шепли, после чего методом перебора была найдена оптимальная комбинация спектральных интервалов.

Разработанная модель позволяет прогнозировать состав двух- и трехкомпонентных систем в водных растворах солей металлов с использованием ограниченного спектрального диапазона вместо полного видимого спектра. Проведенная кросс-валидация продемонстрировала сопоставимое качество новой модели по сравнению с полноспектральными аналогами, подтвердив ее адекватность и практическую применимость.

Ключевые слова: многомерная калибровка, PLS-регрессия, выбор спектральных интервалов, количественное определение ионов металлов, значения Шепли, хемометрика.

Математическое моделирование, численные методы и комплексы программ
Краткое сообщение

Образец для цитирования
Массалов К. Ю., Мощенская Е. Ю. Построение модели для прогнозирования двух- и трехкомпонентных неорганических систем в водных растворах спектральным анализом // *Вестн. Сам. гос. техн. ун-та. Сер. Физ.-мат. науки*, 2025. Т. 29, № 1. С. 174–186. EDN: ZDTCBW. DOI: 10.14498/vsgtu2120.

Сведения об авторах

Кирилл Юрьевич Массалов https://orcid.org/0009-0003-6214-7470
магистрант; каф. физика элементарных частиц; институт ядерной физики и технологий[1]; e-mail: kirill.massalov@yandex.ru

Елена Юрьевна Мощенская https://orcid.org/0000-0002-1070-3151
кандидат химических наук, доцент; доцент; каф. аналитической и физической химии[2]; e-mail: lmos@rambler.ru

Получение: 9 октября 2024 г. / Исправление: 11 февраля 2025 г. /
Принятие: 21 февраля 2025 г. / Публикация онлайн: 11 апреля 2025 г.

Введение. В современном высокотехнологичном мире наблюдается стремительное развитие инструментальных методов аналитической химии. Для сбора, хранения, обработки и интерпретации результатов анализов [1], а также для построения прогностических моделей и оптимизации условий экспериментов [2] требуются обширные информационные ресурсы, эффективные алгоритмы и специализированное программное обеспечение.

Особую актуальность приобретает обработка многомерных массивов данных, получаемых с помощью современных измерительных систем. Такие данные, как правило, представляются в виде матриц или тензоров, что облегчает организацию и анализ больших объемов информации [3].

В области спектроскопии эволюция методов сжатия данных прошла значительный путь: от селекции спектральных переменных до применения современных датчиков и сенсорных систем [4]. Это позволяет существенно повысить эффективность экспериментальных исследований. Видимая и ближняя инфракрасная спектроскопия в настоящее время занимает ведущее положение среди методов промышленного контроля качества [5].

Перспективным направлением является замена универсального спектрального анализа специализированными многомерными сенсорными системами, адаптированными к конкретным аналитическим задачам [6]. Теоретическое моделирование существенно упрощает исследование многокомпонентных систем [7,8], а прогресс в области вычислительных технологий обусловил неизбежность автоматизации научных экспериментов.

В данной работе решаются следующие актуальные задачи:
- анализ экспериментальных данных и построение прогностических моделей;
- разработка эффективных алгоритмов обработки данных;
- программная реализация предложенных методов.

Целью исследования является разработка методов и алгоритмов построения моделей для прогнозирования результатов количественного спектрального анализа.

1. Данные. В исследовании поставлена задача определения спектральных интервалов, оптимальных для прогнозирования количественного состава водных растворов смесей неорганических солей, а также для разработки мультисенсорной аналитической системы.

В качестве исходных данных использованы четыре набора данных (датасета), соответствующих растворам следующих ионов металлов: Ni^{2+} и Co^{2+}; Ni^{2+} и Cu^{2+}; Cu^{2+} и Co^{2+}; Ni^{2+}, Cu^{2+} и Co^{2+} (тройная система). Каждый датасет содержит спектральные данные в диапазоне 360–1100 нм и информацию о концентрациях соответствующих ионов металлов. Для обеспечения качества данных перед моделированием были исключены аномальные наблюдения с использованием метода итеративно взвешенных наименьших квадратов (IRLS) [9].

Метод основан на итеративной адаптации весовых коэффициентов наблюдений, пропорциональных величине соответствующих ошибок прогнозирова-

ния. Формально алгоритм можно описать следующим образом.

Пусть заданы исходные данные: $X \in \mathbb{R}^{n\times d}$ — матрица предикторов (выборка объектов), где n — количество наблюдений, d — количество признаков; $y \in \mathbb{R}^{n\times 1}$ — вектор откликов (целевых значений), σ_y — стандартное отклонение y; $N \in \mathbb{N}$ — максимальное допустимое количество итераций; $\varepsilon \in \mathbb{R}_+$ — порог сходимости алгоритма.

Введем следующие обозначения: $\Delta w = \|w^t - w^p\|_2$ — изменение весов на текущей итерации, где $w^t \in \mathbb{R}^{n\times 1}$ — весовые коэффициенты на текущей итерации, $w^p \in \mathbb{R}^{n\times 1}$ — на предыдущей итерации; j — счетчик итераций.

Алгоритм реализует следующую последовательность шагов.

1. Инициализация:
   - задание величин N и ε;
   - нормализация откликов: $y \leftarrow y/\sigma_y$;
   - присвоение начальных весов: $w_i^t \leftarrow 1 \ \forall\ i = \overline{1,n}$;
   - установка счетчика итераций: $j \leftarrow 0$;
   - задание изменения веса: $\Delta w \leftarrow 1$.
2. Итерационный процесс (пока $\Delta w > \varepsilon$ и $j < N$):
   - построение линейной регрессионной модели с весами w^t;
   - вычисление предсказанных значений $\hat{y}$;
   - обновление весовых коэффициентов:

$$w^p \leftarrow w^t;$$
$$w_i^t \leftarrow \frac{1}{|y_i - \hat{y}_i| + 1} \ \forall\ i = \overline{1,n};$$

   - вычисление изменения веса: $\Delta w = \|w^t - w^p\|_2$;
   - инкрементация счетчика: $j \leftarrow j + 1$.

Наблюдения с окончательными весами $w_i < \delta$ (где δ — заданный порог) классифицируются как промахи и исключаются из дальнейшего анализа.

После обработки методом IRLS датасеты были объединены и структурированы в три группы согласно присутствующим ионам металлов:

- растворы, содержащие ионы никеля Ni^{2+};
- растворы, содержащие ионы кобальта Co^{2+};
- растворы, содержащие ионы меди Cu^{2+}.

2. Входные данные модели. Для построения модели были исследованы два альтернативных подхода к формированию входных признаков на основе спектральных данных.

Метод усредненных интервалов. Исходный спектр разбивается на n непересекающихся интервалов. Для каждого интервала вычисляется среднее значение интенсивности. В результате формируется n-мерный вектор признаков.

Метод дискретизированных интервалов. Исходный спектр разбивается на n непересекающихся интервалов. В пределах каждого интервала производится дискретизация с шагом 5 нм. Каждое дискретное значение интенсивности используется как отдельный признак. Общее количество признаков составляет $n \cdot [w/5]$, где w — ширина интервала в нанометрах, $[\,\cdot\,]$ — операция взятия целой части числа.

Ширина интервалов была установлена равной 60 нм, что соответствует типичной спектральной ширине большинства светодиодов.

3. Выбор и обоснование моделей. В исследовании были использованы два метода регрессионного анализа: классическая линейная регрессия и метод проекции на латентные структуры (PLS). Количество компонент в PLS-регрессии было фиксировано и составляло пять.

Линейная регрессия выбрана на основании закона, устанавливающего линейную зависимость между интенсивностью поглощения и концентрацией вещества в растворе.

PLS-регрессия применена для случаев, когда предикторы демонстрируют высокую степень мультиколлинеарности, что характерно для второго подхода к формированию входных данных (метод дискретизированных интервалов).

Модель линейной регрессии в матричной форме представляется как

$$X\alpha = y,$$

где $X \in \mathbb{R}^{n\times d}$ — матрица объектов-признаков; $y \in \mathbb{R}^{n\times 1}$ — вектор целевых значений; $\alpha \in \mathbb{R}^d$ — вектор коэффициентов модели.

Оптимизация параметров модели выполняется методом наименьших квадратов:

$$Q(\alpha) = \|X\alpha - y\|_2^2 \to \min.$$

Аналитическое решение для данной постановки имеет вид

$$\hat{\alpha} = (X^\top X)^{-1} X^\top y,$$

где $^\top$ — операция транспонирования.

Алгоритм PLS [10] для матрицы предикторов $X \in \mathbb{R}^{n\times d}$ и вектора откликов $y \in \mathbb{R}^{n\times 1}$ реализуется следующей процедурой.

Проводится инициализация: $X_1 = X$, $y_1 = y$. Затем для каждой компоненты $k \in [1, K]$, где K — количество компонент, выполняются следующие шаги:

1) находятся весовые векторы $u_k \in \mathbb{R}^{d\times 1}$ и $v_k \in \mathbb{R}^{1\times 1}$ такие, что

$$\mathrm{Cov}(X_k u_k, y_k v_k) \to \max;$$

2) вычисляются счетные векторы: $\xi_k = X_k u_k$;
3) определяются коэффициенты регрессии:

$$\gamma_k^\top = (\xi_k^\top \xi_k)^{-1} \xi_k^\top X_k, \quad \delta_k^\top = (\xi_k^\top \xi_k)^{-1} \xi_k^\top y_k;$$

4) определяется дефляция матриц:

$$X_{k+1} = X_k - \xi_k \gamma_k^\top, \quad y_{k+1} = y_k - \xi_k \delta_k^\top.$$

Финальные коэффициенты модели вычисляются по формуле

$$\alpha = U(\Gamma^\top U)^{-1} \Delta^\top,$$

где U — матрица весовых векторов u_k; Γ — матрица коэффициентов γ_k; Δ — матрица коэффициентов δ_k.

Обе модели были реализованы с использованием библиотеки `scikit-learn` [11] для языка `Python`.

4. Выбор информативных признаков. Для идентификации наиболее значимых спектральных признаков в работе применялся метод Шепли. Значение Шепли для i-го признака вычисляется по формуле [12]:

$$\Phi(v)_i = \sum_{i \in K} \frac{(k-1)!(n-k)!}{n!} \big(v(K) - v(K \backslash i)\big),$$

где K — некоторое подмножество признаков; k — мощность подмножества K; n — полное число признаков; $v(K)$ — выход модели при наборе K; $v(K \backslash i)$ — выход модели при наборе K без признака i.

Значение $\Phi(v)_i$ количественно характеризует средний предельный вклад i-го признака во все возможные комбинации признаков. Признаки с большими абсолютными значениями $\Phi(v)_i$ оказывают наибольшее влияние на целевую переменную. Таким образом, на выборках с широким диапазоном целевой переменной дисперсия значений Шепли для значимых признаков возрастает.

Для практического вычисления значений Шепли использовалась библиотека `SHAP` [12] для `Python`. В качестве базовой модели применялась PLS-регрессия. Визуализация результатов представлена на рис. 1, где по оси ординат отложено среднеквадратичное отклонение значений Шепли для каждой длины волны.

После определения длин волн с максимальными значениями среднеквадратичного отклонения Шепли методом комбинаторного перебора находились оптимальные спектральные интервалы.

5. Результаты. В табл. 1 представлены максимальные значения коэффициента детерминации R^2, полученные в ходе кросс-валидации для моделей с различным количеством спектральных интервалов (от 1 до 4). Результаты демонстрируют прогностическую способность моделей в определении концентраций ионов никеля Ni^{2+}, кобальта Co^{2+} и меди Cu^{2+} в многокомпонентных системах. Значения R^2 приведены с указанием стандартного отклонения $\sqrt{D[R^2]}$. Анализ данных, представленных в табл. 1, позволяет провести сравнительную оценку четырех рассматриваемых подходов и выявить наиболее эффективные методы моделирования для каждого типа ионов.

Ионы никеля. Анализ значений Шепли (рис. 1, *a*) выявил два наиболее информативных спектральных диапазона: 370–410 нм и 590–650 нм. Расчеты показали, что оптимальным вариантом является использование PLS-регрессии без усреднения на двух интервалах — 625–685 нм и 705–765 нм — с полученным значением $R^2 = 0.992 \pm 0.005$ (см. табл. 1).

Следует отметить, что интервал 610–700 нм присутствует в большинстве лучших конфигураций, а использование диапазона 370–410 нм приводит к увеличению дисперсии ($\sqrt{D[R^2]} > 0.07$) по сравнению с оптимальной конфигурацией ($\sqrt{D[R^2]} \approx 0.005$).

Ионы кобальта. По данным анализа значений Шепли (рис. 1, *b*), наиболее значимым оказался диапазон 440–560 нм. Расчеты показали, что оптимальным вариантом является использование PLS-регрессии без усредне-

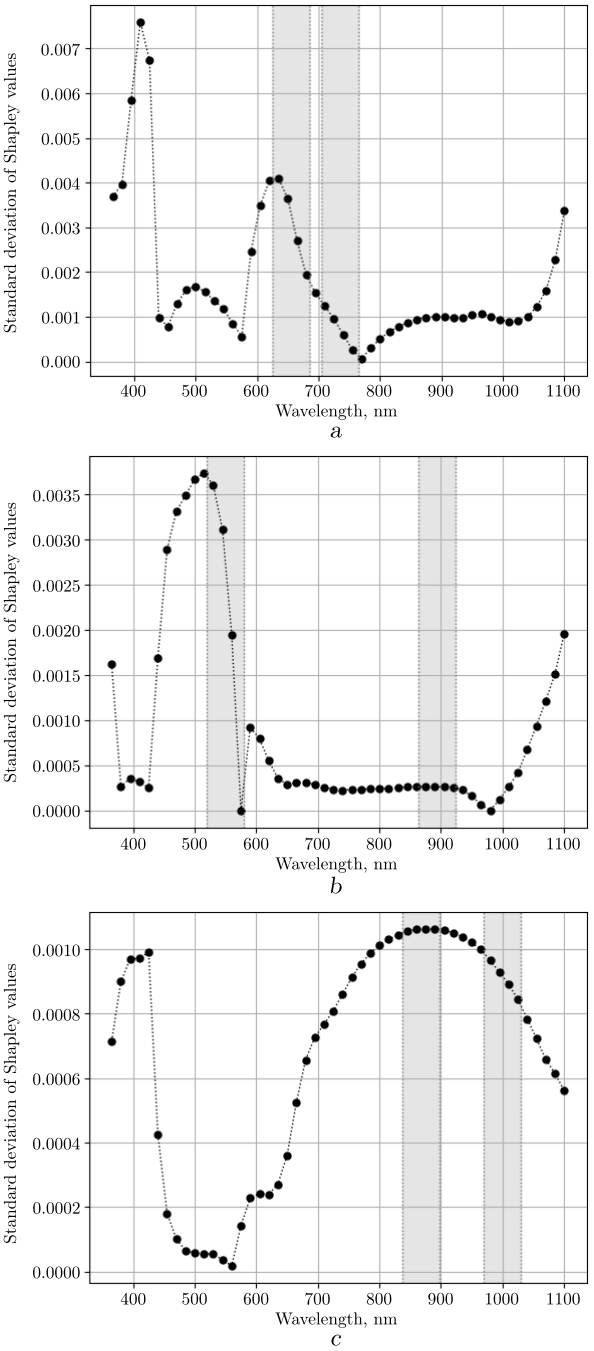

Рис. 1. Аппроксимация среднеквадратического отклонения значений Шепли и оптимальные интервалы для раствора с ионами никеля (*a*), кобальта (*b*) и меди (*c*)

[Figure 1. Approximation of standard deviation of Shapley values and optimum intervals for solution with nickel (*a*), cobalt (*b*), and copper (*c*) ions]

Таблица 1

Максимальное значение R^2 для разного количества интервалов n [Maximum R^2 value for different number of intervals n]

Me^{z+}	n	R^2			
		Averaged linear regression	Averaged PLS regression	Non-averaged linear regression	Non-averaged PLS regression
Ni^{2+}	1	0.640 ± 0.663	—	0.989 ± 0.016	0.989 ± 0.016
	2	0.978 ± 0.023	0.978 ± 0.023	0.991 ± 0.005	$\mathbf{0.992 \pm 0.005}$
	3	0.986 ± 0.016	0.984 ± 0.022	0.989 ± 0.012	0.992 ± 0.006
	4	0.989 ± 0.011	0.961 ± 0.083	0.987 ± 0.010	0.992 ± 0.006
Co^{2+}	1	0.978 ± 0.031	—	0.986 ± 0.017	0.987 ± 0.017
	2	0.978 ± 0.032	0.978 ± 0.032	0.985 ± 0.010	$\mathbf{0.989 \pm 0.009}$
	3	0.981 ± 0.026	0.978 ± 0.031	0.983 ± 0.018	0.989 ± 0.011
	4	0.987 ± 0.012	0.978 ± 0.030	0.974 ± 0.045	0.989 ± 0.010
Cu^{2+}	1	0.979 ± 0.029	—	0.986 ± 0.020	0.987 ± 0.015
	2	0.988 ± 0.017	0.988 ± 0.017	0.990 ± 0.010	$\mathbf{0.992 \pm 0.008}$
	3	0.988 ± 0.017	0.988 ± 0.017	0.990 ± 0.009	0.989 ± 0.014
	4	0.988 ± 0.018	0.987 ± 0.019	0.984 ± 0.013	0.989 ± 0.013

ния на двух интервалах — 519–579 нм, 864–924 нм — с полученным значением $R^2 = 0.989 \pm 0.009$ (см. табл. 1).

Ионы меди. Анализ значимости признаков (рис. 1, *c*) показал важность следующих диапазонов: 370–410 нм, 700–1000 нм. Расчеты показали, что оптимальным вариантом является использование PLS-регрессии без усреднения на двух интервалах — 837–897 нм, 970–1030 нм — с полученным значением $R^2 = 0.992 \pm 0.008$ (см. табл. 1).

6. Обсуждение результатов. На рис. 2 представлено сравнение модели на выделенных спектральных интервалах с моделью, использующей весь доступный спектральный диапазон.

Анализ значений средней абсолютной процентной ошибки в процентах (MAPE, [13]) демонстрирует, что для всех исследуемых ионов металлов (Ni^{2+}, Co^{2+}, Cu^{2+}) модель на выделенных интервалах показывает сопоставимое качество с моделью, использующей полный спектр.

Данные, представленные в табл. 1, позволяют сделать следующие выводы:
- оптимальные значения R^2 (выделены жирным) достигаются при использовании двух спектральных интервалов;
- переход от одного к двум интервалам приводит к увеличению качества (рост R^2) и повышению устойчивости модели (снижение $D[R^2]$);
- дальнейшее увеличение количества интервалов не дает существенного улучшения качества.

Проведенный анализ выявил важные закономерности:
- при использовании усреднения внутри интервалов PLS-регрессия уступает по эффективности линейной регрессии, т.к. происходит потеря информативных признаков при уменьшении размерности;
- при работе с неусредненными данными линейная регрессия демонстрирует худшие результаты по сравнению с PLS, т.к. увеличивается число мультиколлинеарных признаков.

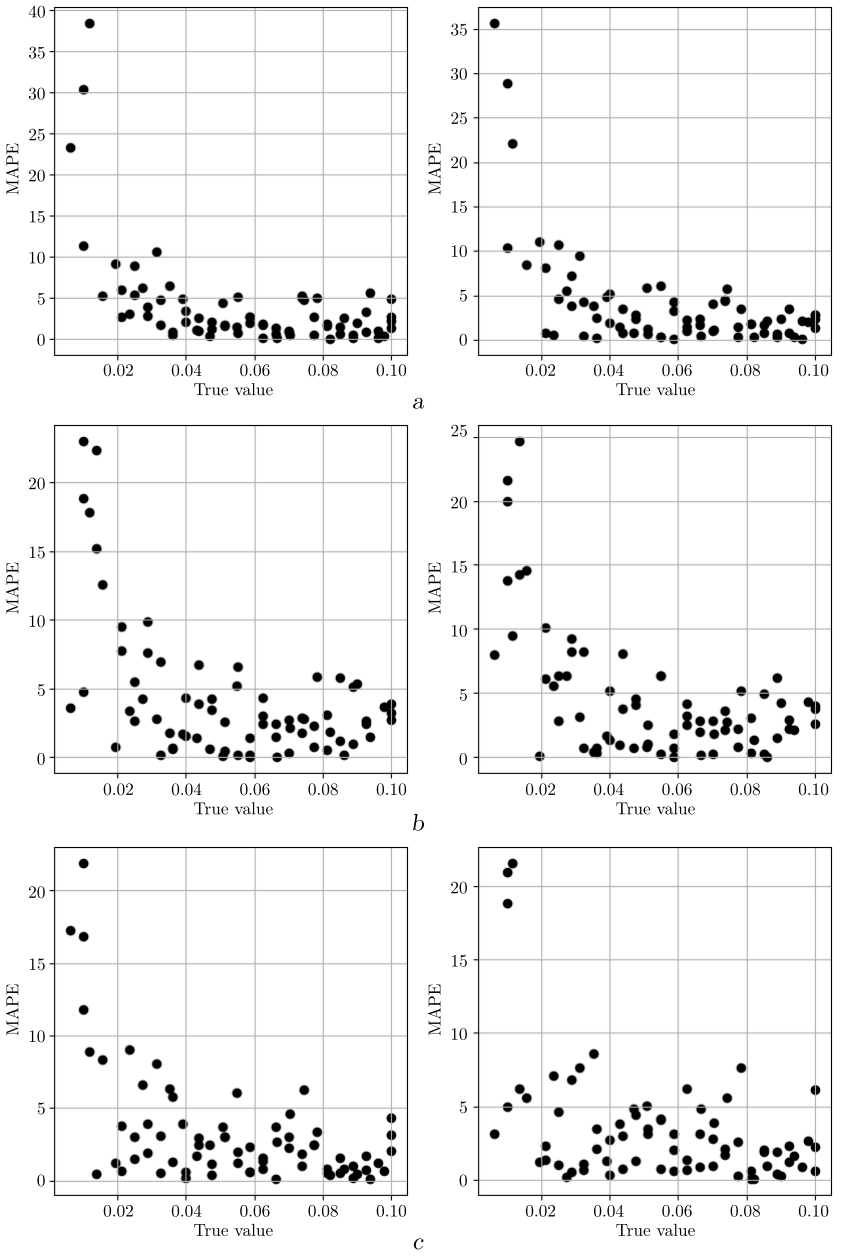

Рис. 2. Средняя абсолютная ошибка в процентах (MAPE) для модели, построенной на выделенных интервалах (слева), и для модели, построенной на всем спектре (справа), для ионов никеля (*a*), кобальта (*b*) и меди (*c*)

[Figure 2. Mean absolute percentage error (MAPE) for the model built using selected spectral intervals (left) and the model utilizing the full spectrum (right) for nickel (*a*), cobalt (*b*), and copper (*c*) ions]

Заключение. Представлена модель, позволяющая прогнозировать двух- и трехкомпонентные системы в водных растворах никеля, меди и кобальта путем спектрального анализа с использованием только части спектра. При помощи кросс-валидации подтверждена адекватность предложенной модели: проведена численная оценка ее качества и сравнение с моделью, использующей полный видимый спектр.

Конкурирующие интересы. Авторы заявляют об отсутствии каких-либо конфликтов интересов, связанных с подготовкой и публикацией данной статьи.

Авторский вклад и ответственность. Все авторы внесли равный вклад в разработку концепции исследования, проведение расчетов и анализ данных, подготовку и редактирование текста рукописи. Окончательная версия статьи была одобрена всеми соавторами, которые несут полную ответственность за представленные результаты.

Финансирование. Исследование выполнено без использования внешних источников финансирования.

Благодарность. Авторы благодарны заведующему кафедрой аналитической и физической химии А.Ю. Богомолову и его аспиранту А.М. Никитиной за предоставленные экспериментальные данные.

Библиографический список

1. Dubrovkin J. *Data Compression in Spectroscopy*. Cambridge Scholars Publ., 2022. 355 pp.
2. Родионова О.Е. Хемометрический подход к исследованию больших массивов химических данных // *Рос. хим. ж.*, 2006. Т. 50, № 2. С. 128–144. EDN: HTUUSZ.
3. Smilde A., Bro R., Geladi P. *Multi-Way Analysis: Applications in the Chemical Sciences*. Chichester: John Wiley & Sons, 2004. xiv+381 pp. DOI: https://doi.org/10.1002/0470012110.
4. Богомолов А. Ю. Оптические мультисенсорные системы в аналитической спектроскопии // *Рос. хим. ж.*, 2022. Т. 77, № 3. С. 227-247. EDN: MFSPES. DOI: https://doi.org/10.31857/S0044450222030033.
5. Bogomolov A. Multivariate process trajectories: capture, resolution and analysis // *Chemom. Intel. Lab. Syst.*, 2011. vol. 108, no. 1. pp. 49–63. DOI: https://doi.org/10.1016/j.chemolab.2011.02.005.
6. Galyanin V., Melenteva A., Bogomolov A. Selecting optimal wavelength intervals for an optical sensor: A case study of milk fat and total protein analysis in the region 400–1100 nm // *Sens. Actuat. B: Chem.*, 2015. vol. 218. pp. 97-104. EDN: UFYADR. DOI: https://doi.org/10.1016/j.snb.2015.03.101.
7. Мощенская Е. Ю., Стифатов Б. М. Моделирование диаграмм “состав-свойство” для системы “алюминий-кремний” // *Журн. Сиб. федер. ун-та. Химия*, 2023. Т. 16, № 1. С. 107–115. EDN: JWRAGD.
8. Мощенская Е. Ю., Стифатов Б. М. Исследование возможности применения методов теоретического моделирования для определения эвтектического состава бинарных сплавов // *Вестн. Тверск. гос. ун-та. Сер. Химия*, 2021. № 3. С. 105–122. EDN: JDZAEI. DOI: https://doi.org/10.26456/vtchem2021.3.12.
9. Holland P. W., Welsch R. E. Robust regression using iteratively reweighted least-squares // *Commun. Stat–Theor. M.*, 1977. vol. 6, no. 9. pp. 813–827. DOI: https://doi.org/10.1080/03610927708827533.
10. Wegelin J. A. *A Survey of Partial Least Squares (PLS) Methods, with Emphasis on the Two-Block Case*: Technical Report 371. Washington: Univ. of Washington, 2000. 44 pp. https://stat.uw.edu/research/tech-reports/survey-partial-least-squares-pls-methods-emphasis-two-block-case.

11. Pedregosa F., Varoquaux G., Gramfort A., et. al. Scikit-learn: Machine learning in Python // *J. Mach. Learn. Res.*, 2011. vol. 12. pp. 2825–2830.
12. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions / *Proc. Intern. Conf. Neural Inform. Proces. Systems*, 2017. pp. 4768–4777, arXiv: 1705.07874 [cs.AI]. DOI: https://doi.org/10.48550/arXiv.1705.07874.
13. de Myttenaere A., Golden B., Le Grand B., Rossi F. Mean Absolute Percentage Error for regression models // *Neurocomputing*, 2016. vol. 192. pp. 38-48. DOI: https://doi.org/10.1016/j.neucom.2015.12.114.

Vestn. Samar. Gos. Tekhn. Univ., Ser. Fiz.-Mat. Nauki
[J. Samara State Tech. Univ., Ser. Phys. Math. Sci.], 2025, vol. 29, no. 1, pp. 174–186
ISSN: 2310-7081 (online), 1991-8615 (print) https://doi.org/10.14498/vsgtu2120

MSC: 65C20, 62P99, 92E20

Development of a predictive model for two- and three-component inorganic systems in aqueous solutions using spectral analysis

K. Y. Massalov[1], *E. Y. Moshchenskaya*[2]

[1] National Engineering Physics Institute "MEPhI",
31, Kashirskoe shosse, Moscow, 115409, Russian Federation.
[2] Samara State Technical University,
244, Molodogvardeyskaya st., Samara, 443100, Russian Federation.

Abstract

This study presents an algorithm for analyzing spectral data through mathematical modeling, constructing prognostic models, and selecting optimal wavelength intervals for designing LED-based multisensor systems. The algorithm is implemented in Python and validated using experimental data from aqueous solutions of inorganic salts.

Key methodological aspects include:

– Application of multivariate calibration methods (PLS regression and multiple linear regression);
– Utilization of Shapley values to identify informative spectral wavelengths;
– Systematic enumeration to determine optimal wavelength intervals.

The developed model enables accurate prediction of two- and three-component systems in metal salt solutions using partial spectral data rather than full-spectrum analysis. Cross-validation demonstrates that:

– The model achieves comparable accuracy to full-spectrum approaches;
– The solution remains computationally efficient while maintaining predictive reliability.

The results confirm the model's adequacy for quantitative spectral analysis, particularly in resource-constrained environments where partial spectral data acquisition is advantageous.

Mathematical Modeling, Numerical Methods and Software Complexes
Short Communication

Please cite this article in press as:
Massalov K. Y., Moshchenskaya E. Y. Development of a predictive model for two- and three-component inorganic systems in aqueous solutions using spectral analysis, *Vestn. Samar. Gos. Tekhn. Univ., Ser. Fiz.-Mat. Nauki* [J. Samara State Tech. Univ., Ser. Phys. Math. Sci.], 2025, vol. 29, no. 1, pp. 174–186. EDN: ZDTCBW. DOI: 10.14498/vsgtu2120 (In Russian).

Authors' Details:

Kirill Y. Massalov https://orcid.org/0009-0003-6214-7470
Master's Student; Senior Researcher; Dept. of Elementary Particle Physics; Institute of Nuclear Physics and Engineering[1]; e-mail: kirill.massalov@yandex.ru

Elena Y. Moshchenskaya https://orcid.org/0000-0002-1070-3151
Cand. Chem. Sci., Associate Professor; Associate Professor; Dept. of Analytical and Physical Chemistry[2]; e-mail: lmos@rambler.ru

Keywords: multivariate calibration, PLS regression, spectral interval selection, metal ion quantification, Shapley values, chemometrics.

Received: 9th October, 2024 / Revised: 11th February, 2025 / Accepted: 21st February, 2025 / First online: 11th April, 2025

Competing interests. The authors declare no conflicts of interest related to the preparation and publication of this article.

Authors' contributions and responsibilities. All authors contributed equally to: the research concept development, calculations and data analysis, manuscript preparation and editing. The final version of the article was approved by all co-authors, who bear full responsibility for the presented results.

Funding. The study was conducted without external funding sources.

Acknowledgments. The authors are grateful to the Head of the Department of Analytical and Physical Chemistry A.Yu. Bogomolov and his graduate student A.M. Nikitina for providing the experimental data.

References

1. Dubrovkin J. *Data Compression in Spectroscopy*. Cambridge Scholars Publ., 2022, 355 pp.
2. Rodionova O. E. Chemometric approaches for analysis of large chemical data arrays, *Ros. Khim. Zh.*, 2006, vol. 50, no. 2, pp. 128–144 (In Russian). EDN: HTUUSZ.
3. Smilde A., Bro R., Geladi P. *Multi-Way Analysis: Applications in the Chemical Sciences*. Chichester, John Wiley & Sons, 2004, xiv+381 pp. DOI: https://doi.org/10.1002/0470012110.
4. Bogomolov A. Yu. Optical multisensor systems in analytical spectroscopy, *J. Anal. Chem.*, 2022, vol. 77, no. 3, pp. 277–294. EDN: YORSQC. DOI: https://doi.org/10.1134/S1061934822030030.
5. Bogomolov A. Multivariate process trajectories: capture, resolution and analysis, *Chemom. Intel. Lab. Syst.*, 2011, vol. 108, no. 1, pp. 49–63. DOI: https://doi.org/10.1016/j.chemolab.2011.02.005.
6. Galyanin V., Melenteva A., Bogomolov A. Selecting optimal wavelength intervals for an optical sensor: A case study of milk fat and total protein analysis in the region 400–1100 nm, *Sens. Actuat. B: Chem.*, 2015, vol. 218, pp. 97-104. EDN: UFYADR. DOI: https://doi.org/10.1016/j.snb.2015.03.101.
7. Moshchenskaya E. Yu., Stifatov B. M. Modeling "composition-property" diagrams for the "aluminum-silicon" system, *J. Sib. Fed. Univ. Chem.*, 2023, vol. 16, no. 1, pp. 107–115 (In Russian). EDN: JWRAGD.
8. Moshchenskaya E. Yu., Stifatov B. M. Investigation of the possibility of using theoretical modeling methods to determine the eutectic composition of binary alloys, *Vestn. Tversk. Gos. Univ., Ser. Khimiia*, 2021, no. 3, pp. 105–122 (In Russian). EDN: JDZAEI. DOI: https://doi.org/10.26456/vtchem2021.3.12.
9. Holland P. W., Welsch R. E. Robust regression using iteratively reweighted least-squares, *Commun. Stat–Theor. M.*, 1977, vol. 6, no. 9, pp. 813–827. DOI: https://doi.org/10.1080/03610927708827533.
10. Wegelin J. A. *A Survey of Partial Least Squares (PLS) Methods, with Emphasis on the Two-Block Case*, Technical Report 371. Washington, Univ. of Washington, 2000, 44 pp. https://stat.uw.edu/research/tech-reports/survey-partial-least-squares-pls-methods-emphasis-two-block-case.
11. Pedregosa F., Varoquaux G., Gramfort A., et. al. Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 2011, vol. 12, pp. 2825–2830.

12. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions, In: *Proc. Intern. Conf. Neural Inform. Proces. Systems*, 2017, pp. 4768–4777, arXiv: 1705.07874 [cs.AI]. DOI: https://doi.org/10.48550/arXiv.1705.07874.
13. de Myttenaere A., Golden B., Le Grand B., Rossi F. Mean Absolute Percentage Error for regression models, *Neurocomputing*, 2016, vol. 192, pp. 38-48. DOI: https://doi.org/10.1016/j.neucom.2015.12.114.