**Реализация алгоритма N-модальной проекции на латентные структуры для определения растворенного органического вещества флуоресцентным методом**

***Ферубко А.О., Крылов И.Н.***

*Студент, 6 курс специалитета*

*Московский государственный университет имени М.В. Ломоносова, Москва, Россия*

*E-mail:ferubko1999@yandex.ru*

Нахождение в Мировом океане органического углерода в количестве порядка 1012 тонн играет важную роль в глобальном биогеохимическом цикле углерода, оказывающем значительное влияние на климат и разнообразие живых организмов. В связи с этим изучение данного цикла является актуальной и важной научной задачей. Существенная часть органического углерода в водах представлена окрашенным растворенным органическим веществом (РОВ). Поскольку содержание РОВ в морской воде невелико, применение флуоресцентных методов представляется наиболее перспективным для их прямого определения. При этом использование флуоресценции возбуждение-испускание позволяет учитывать разнообразие РОВ, однако требует применения методов машинного обучения для проведения количественного анализа [1]. Классические методы, такие как метод главных компонент или проекция на латентные структуры, не позволяют учитывать трилинейную структуру данных спектров возбуждения-испускания, поэтому в данной работе была поставлена задача реализовать алгоритм N-модальной проекции на латентные структуры (N-PLS1) и применить его для количественного определения РОВ.

Алгоритм N-PLS1 с добавлением L2-регуляризации по Тихонову был реализован с использованием Python 3 и опубликован в индексе пакетов PyPi[[1]](#footnote-1)[[2]](#footnote-2) под лицензией GPL-3.0. Для выбора количества компонентов модели (сингулярных векторов) использовали норму производной n-го порядка, чтобы избежать включения шума в модель и переобучения. В качестве предобработки, из всех спектров удаляли сигнал Рамановского и Рэлеевского рассеяния воды, и образовавшуюся область интерполировали при помощи монотонной кусочно-кубической интерполяции, потому что они не удовлетворяли условию трилинейности сигнала. Также проводили центрирование всех наборов данных. Для валидации модели применяли кросс-валидацию с решётчатым поиском[[3]](#footnote-3) и многократным перемешиванием данных, что позволило добиться равномерного распределения значений в обучающем наборе. Оптимизацию модели проводили по коэффициенту регуляризации и числу компонент разложения.

С помощью разработанной модели содержание РОВ определяли как для синтетических и стандартизированных наборов данных «Fluorescence» и «Dorrit fluorescence data» с сайта Копенгагенской хемометрической группы, так и для набора образцов, имитирующего морские воды, для приготовления которых использовались различные по происхождению гуминовые вещества и аминокислоты. Для всех наборов данных RMSE/диапазон концентраций не превосходит 0.13, а коэффициент детерминации больше 0.7. Таким образом, показано, что предложенная модель позволяет проводить прямое определение РОВ в морских водах в условиях сильной вариабельности их спектров флуоресценции.

**Литература:**

1. *Smilde, A.* Multi-Way Analysis with Applications in the Chemical Sciences. Multi-Way Analysis with Applications in the Chemical Sciences / A. Smilde, R. Bro, P. Geladi. – WILEY. – Chichester, UK: John Wiley & Sons, Ltd, 2004. – 369 p.

1. https://pypi.org/project/npls/ [↑](#footnote-ref-1)
2. https://pypi.org/project/signal-noise/ [↑](#footnote-ref-2)
3. https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.GridSearchCV.html [↑](#footnote-ref-3)