Веб-приложение многомерной регрессии на основе метода наименьших квадратов и программной библиотеки конструируемых базисов

Р.В. Костенко, Ю.В. Клинаев

Саратовский государственный технический университет имени Гагарина Ю.А.

Аннотация: Рассмотрены алгоритмы построения базисных функций, методы оптимизации вычислений и механизмы L1/L2 регуляризации, на основе которых разработано веб-приложение для выполнения многомерной регрессии с использованием метода наименьших квадратов и программной библиотеки конструируемых базисов. Разработанное программное обеспечение обеспечивает эффективную аппроксимацию многомерных данных с возможностью визуализации результатов в двухмерной и трехмерной проекциях. Практическая ценность работы заключается в создании программного инструмента, находящегося в открытом доступе сети, для анализа и моделирования сложных многомерных зависимостей.

Ключевые слова: аппроксимация, метод наименьших квадратов, базисные функции, многомерная регрессия, L1/L2-регуляризация, веб-приложение, многомерный эллиптический параболоид.

Введение

Одними из важных и эффективных инструментов математического моделирования являются численные алгоритмы аппроксимации и интерполяции результатов экспериментальных исследований.

Эксплуатация современного инженерного оборудования приводит к необходимости решения задач оптимального управления на основе данных измерений датчиками значений многочисленных физических и технологических параметров технологического процесса. Поэтому, задачи анализа многомерных массивов данных в плане их аппроксимации аналитическими зависимостями, представляются актуальными и в практическом аспекте.

Применение моделей регрессии для анализа многомерных массивов и построения их математических моделей как функций многих переменных влечёт за собой значительный рост вычислительных затрат и неизменно возникающих при этом сложностей математического характера, приводящих

к необходимости конструирования, в принципе, любой системы базисных функций, но удовлетворяющих требуемому критерию оптимизации для конкретной задачи аппроксимации.

Целью настоящей работы является разработка программного обеспечения многомерной регрессии на основе метода наименьших квадратов и программной библиотеки конструируемых базисных функций.

Анализ существующих программных решений для аппроксимации данных показал, что большинство из них имеют ограничения при работе с многомерными данными или предоставляют только фиксированный набор базисных функций.

Теоретические основы множественной линейной регрессии изложены в [1, 2], алгоритмы квадратичной регрессии с мультипликативными компонентами — [3, 4], программные реализации данных методов — [5, 6]. Особое место среди известных разработок по этому направлению принадлежит одной из самых ранних фундаментальных работ [7], и более поздняя – [8], в которых представлен полный текст на языке FORTRAN реализации программного комплекса восстановления зависимостей по разработанным авторами численным алгоритмам метода структурной минимизации риска для аппроксимации многомерных эмпирических данных. Развитие этих идей в современных методах машинного обучения для автоматической обработки данных продолжается и в настоящее время [9, 10].

Среди онлайн-ресурсов можно выделить planetcalc.ru, bl2.ru и mathhelpplanet.com, однако они не поддерживают работу с данными, имеющими более одной входной переменной [4, 5, 11].

Методология

В основе разработанного программного обеспечения лежит обобщенная модель метода наименьших квадратов для многомерных данных.

Минимизируемая функция потерь имеет вид многомерного эллиптического параболоида (рис.1):

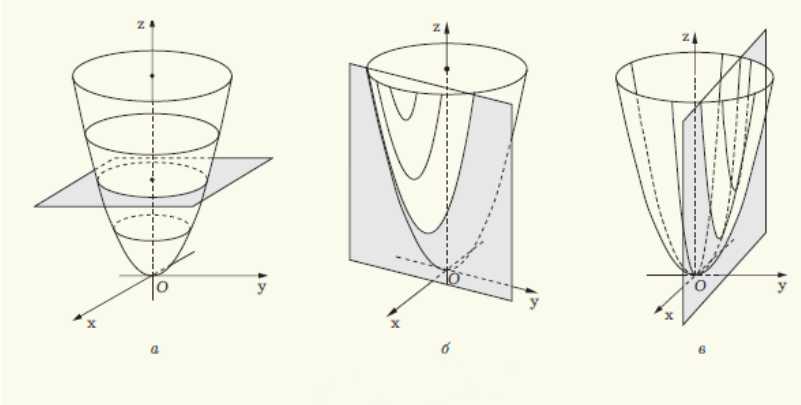

Рис. 1. – Эллиптический параболоид

Суть метода наименьших квадратов – свести к минимуму значение функционала:

$$S(a_1, a_2 \dots, a_m) = \sum_{i=1}^{n}\left(x_{1_i} * a_1 + x_{2_i} * a_2 \dots + x_{m_i} * a_m - F_i\right)^2 \to min \quad (1)$$

где n – число наблюдений, m – число базисов, x_{mi} – значение m-го базиса для i-го наблюдения, F_i – наблюдаемое выходное значение, a_m – подбираемые коэффициенты.

Для улучшения обобщающей способности модели и повышения численной стабильности реализованы механизмы регуляризации L_1 (2) [12] (регуляризация L_1, - «LASSO», позволяет получить разреженные модели за счет добавления штрафа, основанного на абсолютном значении коэффициентов) и L_2 (3) [12] (регуляризация L_2, - «RidgeRegression», предполагает использование меньших и более равномерно распределенных весов добавлением штрафа, основанного на квадрате коэффициентов).

L_1: $$S(a_1, a_2 \dots, a_m) = \sum_{i=1}^{n}\left(x_{1_i} * a_1 + x_{2_i} * a_2 \dots + x_{m_i} * a_m - F_i\right)^2 + L_1 * (a_1 + a_2 + a_3 + \cdots + a_m) \to min \quad (2)$$

L_2: $$S(a_1, a_2 \dots, a_m) = \sum_{i=1}^{n}\left(x_{1_i} * a_1 + x_{2_i} * a_2 \dots + x_{m_i} * a_m - F_i\right)^2 + L_2 * (a_1^2 + a_2^2 + a_3^2 + \cdots + a_m^2) \to min \quad (3)$$

Программная реализация включает следующие ключевые компоненты:

1. Механизм построения базисных функций.

2. Получение предвычислений для оптимизации работы.

3. Алгоритм формирования матрицы и решения системы нормальных уравнений с использованием WebAssembly.

4. Визуализации результатов в 2D и 3D проекциях.

Для разработки использованы следующие технологии:

1. Фреймворк Vue3 с TypeScript.

2. Material Design фреймворк Vuetify.

3. WebAssembly для ресурсоёмких вычислений.

4. Библиотеки визуализации: amCharts 5, THREE.js.

Архитектура программного обеспечения

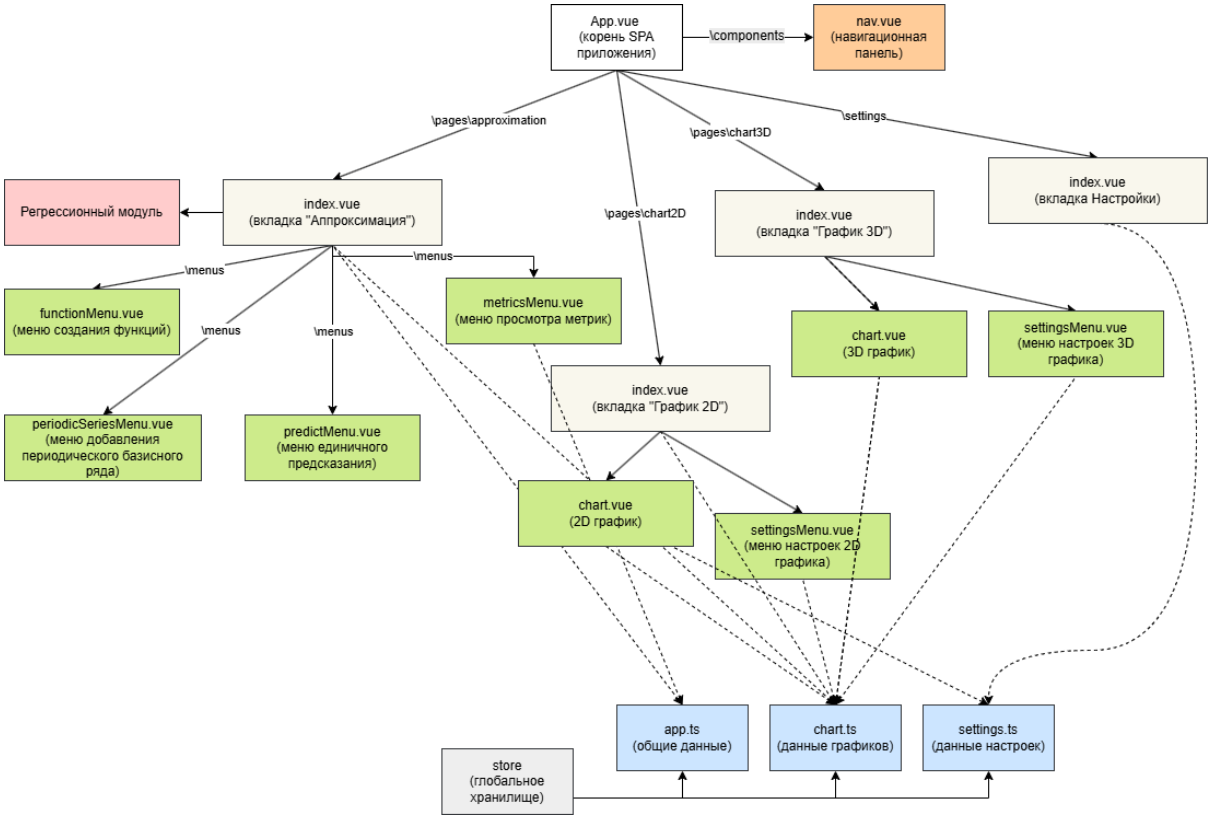

Рис. 2. – Файловая структура ПО многомерной регрессии

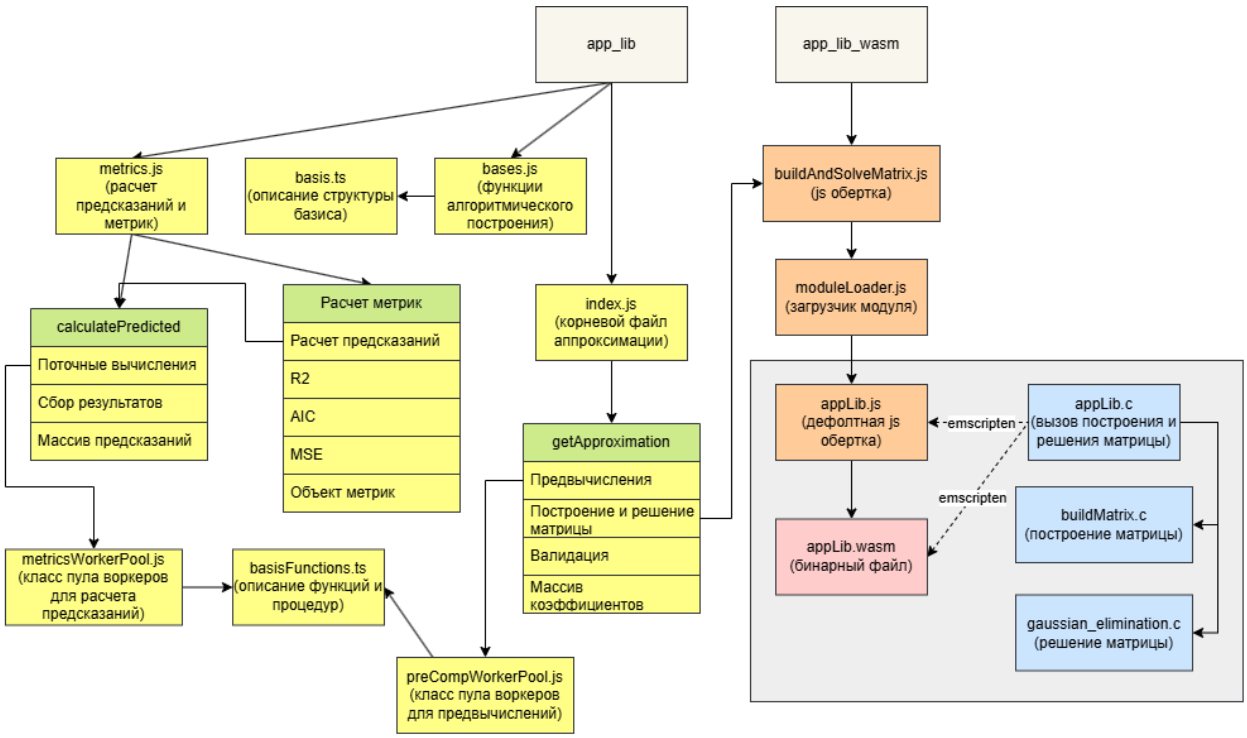

Рис. 3. – Файловая структура регрессионного модуля

Алгоритмы построения базисов

Для упрощения создания базисных функций реализован рекурсивный алгоритм, позволяющий быстро генерировать большое количество базисов из записей вида: **<глубина> <функция> ^ <степень> / <выходная функция> ^ <выходная степень>.**

x^5 для a, b, c, d переменных:

$a^5 * b^1 * c^1 * d^1$
$a^4 * b^2 * c^1 * d^1$
$a^4 * b^1 * c^2 * d^1$
$a^4 * b^1 * c^1 * d^2$
$a^3 * b^3 * c^1 * d^1$
$a^3 * b^2 * c^2 * d^1$
$a^3 * b^2 * c^1 * d^2$
…

Рис. 4. – Перебор «четвёрок» для пятой степени

Данный алгоритм перебирает все возможные коллекции от единиц, двоек … <глубина>, варьируя свободную степень между переменными.

Количество таких сочетаний можно оценить как C_{n-1}^{m-1}, где n – степень в записи перед функцией, а m – количество элементов (переменных).

Для тригонометрических функций реализовано построение рядов вида sin(kx), cos(kx) с заданным шагом увеличивающих значение k.

Оптимизация вычислений

Для повышения производительности программы реализована стратегия предвычислений, которая существенно снижает вычислительную сложность формирования матрицы системы нормальных уравнений.

Таблица № 1

Определение данных и предвычислений (набора переменных-столбцов)

Переменная столбец	**x_1**	**x_2**	**x_3**	**x_m**	**$f_1(x_1, x_2)$**	**$f_2(x_1, x_3)$**	**$f_2(x_1, x_2, x_3)$**	**f_m**
№1	$x_{1(1)}$	$x_{2(1)}$	$x_{3(1)}$	$x_{m(1)}$	$f_1(x_{1(1)}, x_{2(1)})$	$f_1(x_{1(1)}, x_{3(1)})$	$f_1(x_{1(1)}, x_{2(1)}, x_{3(1)})$	$f_m(\ldots args_1)$
№2	$x_{1(2)}$	$x_{2(2)}$	$x_{3(2)}$	$x_{m(2)}$	$f_1(x_{1(2)}, x_{2(2)})$	$f_1(x_{1(2)}, x_{3(2)})$	$f_1(x_{1(2)}, x_{2(2)}, x_{3(2)})$	$f_m(\ldots args_2)$
№3	$x_{1(3)}$	$x_{2(3)}$	$x_{3(3)}$	$x_{m(3)}$	$f_1(x_{1(3)}, x_{2(3)})$	$f_1(x_{1(3)}, x_{3(3)})$	$f_1(x_{1(3)}, x_{2(3)}, x_{3(3)})$	$f_m(\ldots args_3)$
№4	$x_{1(4)}$	$x_{2(4)}$	$x_{3(4)}$	$x_{m(4)}$	$f_1(x_{1(4)}, x_{2(4)})$	$f_1(x_{1(4)}, x_{3(4)})$	$f_1(x_{1(4)}, x_{2(4)}, x_{3(4)})$	$f_m(\ldots args_4)$
№5	$x_{1(5)}$	$x_{2(5)}$	$x_{3(5)}$	$x_{m(5)}$	$f_1(x_{1(5)}, x_{2(5)})$	$f_1(x_{1(5)}, x_{3(5)})$	$f_1(x_{1(5)}, x_{2(5)}, x_{3(5)})$	$f_m(\ldots args_5)$
№ i	$x_{1(i)}$	$x_{2(i)}$	$x_{3(i)}$	$x_{m(i)}$	$f_1(x_{1(i)}, x_{2(i)})$	$f_1(x_{1(i)}, x_{3(i)})$	$f_1(x_{1(i)}, x_{2(i)}, x_{3(i)})$	$f_m(\ldots args_i)$

Асимптотическая сложность сокращена с $O(b^2 \cdot N \cdot f \cdot log_2(p))$ до $O(b \cdot N \cdot (b + f \cdot log_2(p)))$, где b – количество базисов, N – количество наблюдений, f – среднее количество функций в базисе.

Где:

b – количество представлений базисов;

f – среднее количество функций в представлениях базисов.

N – количество наблюдений в данных;

B – асимптотика самого алгоритмически сложного базиса.

Поскольку сложность базисов заранее неизвестна, определим B как асимптотику наиболее алгоритмически сложного базиса, что позволит получить верхнюю оценку.

Таблица № 2

Оценка асимптотической сложности различных этапов вычисления

Тип операции	**Сложность «O»**
Предвычисления	$b * f * B * N$
Построение матрицы	$N * b^2$
Решение матрицы	b^3
Итоговая сложность вычислений	$b * (b * (N + b) + B * N)$
Расчет метрик	$b * B * N$

Решение системы нормальных уравнений реализовано с использованием технологии WebAssembly через метод Гаусса с частичным выбором ведущего элемента матрицы системы (на строчку n перемешается строка с наибольшим по модулю элементом в столбце n среди нерассмотренных строк, начиная с первой), что обеспечило улучшенную численную стабильность решения системы. Позже, сравнивая такое решение с обычным, было опытным путем выявлено существенное улучшение показателей метрик.

Пользовательский интерфейс

Разработан интуитивно понятный пользовательский интерфейс, обеспечивающий удобную работу с программой (рис.5-7). Основные функциональные возможности интерфейса:

1. Загрузка данных в формате XLSX.
2. Конструирование базисных функций.
3. Выполнение аппроксимации.
4. Визуализация результатов в 2D и 3D.
5. Настройка параметров регуляризации.

На рис.5 представлен интерфейс главной страницы приложения.

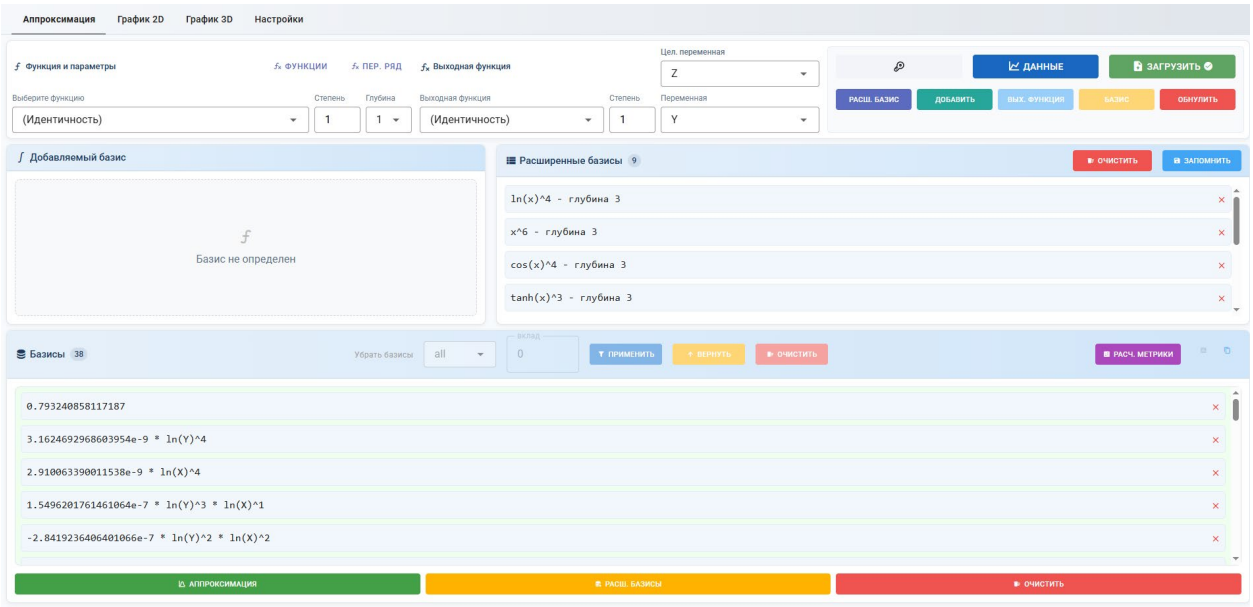

Рис. 5. – Стартовая вкладка «Аппроксимация»

Рис.6 содержит трехмерную проекцию с двумя графиками: реальные значения (черные данные) и предсказанные (синие данные).

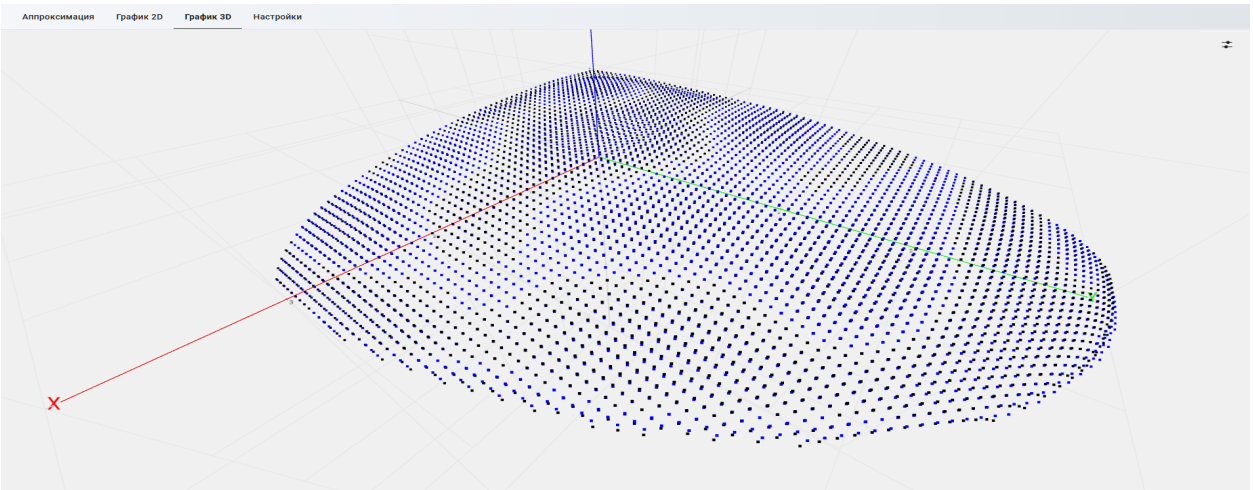

Рис. 6. – Вкладка «3D - график»

Рис.7 содержит трехмерную проекцию с двумя графиками: реальные значения (черные данные) и предсказанные (синие данные).

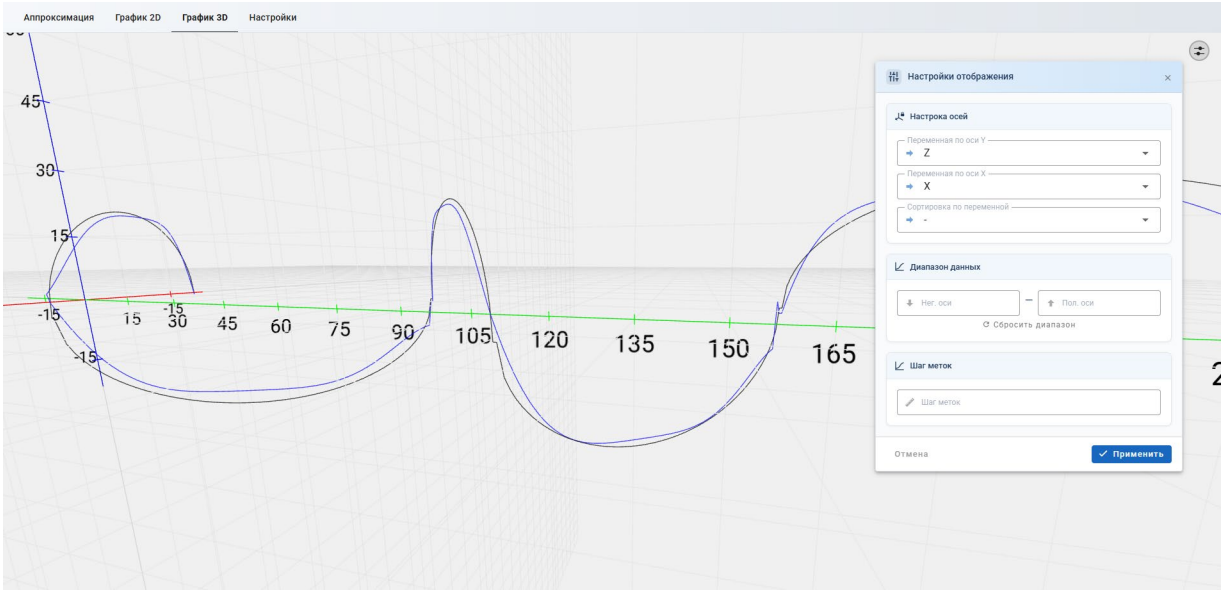

Рис. 7. – Вкладка «3D - график»

Метрики качества аппроксимации

Для оценки качества полученных моделей реализованы три основные метрики:

Среднеквадратичная ошибка MSE [13] – (4):

$$MSE = \frac{1}{n} * \sum_{i=1}^{n} (F_{real_i} - F_{predicted\_i})^2, \quad (4)$$

где F_{real_i} / $F_{predicted\_i}$ – наблюдаемая / предсказываемая величина i-ого наблюдения.

Коэффициент детерминации R^2 [14] – (5):

$$R^2 = 1 - \frac{\frac{1}{n} * rss}{\frac{1}{n} * tss}, \quad (5)$$

где $\frac{1}{n} * rss$ – среднеквадратичная ошибка, $\frac{1}{n} * tss$ – средняя дисперсия.

Информационный критерий Акаике AIC [13] – (6):

$$AIC = n * \ln\left(\frac{rss}{n}\right) + 2 * k, \quad (6)$$

где k – количество базисных представлений.

Публикация программного обеспечения

Разработанное программное обеспечение опубликовано в сети интернет по адресу: datapprox.com, что делает его доступным для широкого круга пользователей без необходимости установки дополнительного программного обеспечения.

Заключение

Новизна результатов проведённого исследования заключается в разработке веб-приложения, предоставляющего пользователю возможность создавать и использовать разнообразные базисные функции для аппроксимации многомерных данных.

В ходе исследования разработано программное обеспечение для выполнения многомерной регрессии на основе метода наименьших квадратов и библиотеки конструируемых базисов, а именно:

1. Реализован алгоритм построения базисных функций, позволяющий генерировать широкий спектр базисов из компактных записей.
2. Разработаны механизмы L_1 и L_2 регуляризации, повышающие устойчивость модели аппроксимации.
3. Применены методы оптимизации вычислений, существенно улучшающие производительность.
4. Создан интуитивно понятный пользовательский интерфейс с возможностью визуализации результатов.
5. Программное обеспечение реализовано в виде веб-приложения, доступного через интернет.

Литература

1. Виленкин С.Я. Сборник научных программ на Фортране. Вып. 1. Москва: Статистика, 1974. 316 с.

2. Васильев А.Н. Научные вычисления в Microsoft Excel. Москва: Вильямс, 2004. 512 с.

3. Половко А.М., Бутусов П.Н. Интерполяция. Методы и компьютерные технологии их реализующие. Санкт-Петербург: БХВ-Петербург, 2004. 320 с.

4. Bl2. Интерполяция функций. URL: bl2.ru/mathematics/interpolation.html.

5. Математический форум Math Help Planet. Методы интерполяции и регрессии. URL: mathhelpplanet.com/static.php?p=metody-interpolyatsii-i-regressii.

6. Клинаев Ю.В., Терин Д.В. Методы и технологии компьютерных вычислений в математическом моделировании. Саратов: Изд-во СГТУ, 2010. 208 с.

7. Вапник В.Н., Глазкова Т.Г., Кощеев В.А., Михальский А.И., Червоненкис А.Я. Алгоритмы и программы восстановления зависимостей. Москва: Наука, 1984. 816 с.

8. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. Москва: Наука, 2019. 448 с.

9. Беспалова Н.В., Корчагин С.А., Сердечный Д.В. Анализ алгоритмов машинного обучения используемых для обработки текстовых документов // Инженерный вестник Дона, 2025, №5. URL: ivdon.ru/ru/magazine/archive/n5y2025/10037.

10. Корчагин С.А., Сердечный Д.В., Окунев А.И., Андриянов Н.А. Методы машинного обучения для автоматической обработки документов // Инженерный вестник Дона, 2025, №3. URL: ivdon.ru/ru/magazine/archive/n3y2025/9914.

11. Planetcalc. Аппроксимация функций. URL: planetcalc.ru/8735.

12. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2009. 745 p.

13. Kuhn M., Johnson K. Applied Predictive Modeling. New York: Springer, 2013. 600 p.

14. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. Москва: ЮНИТИ, 1998. 1022 с.

References

1. Vilenkin S. Ja. Sbornik nauchnyh programm na Fortrane. Vyp. 1. [Collection of scientific programs in Fortran. Vol. 1. Statistics]. Moskva: Statistika, 1974. 316 p.

2. Vasiliev A.N. Nauchnye vychislenija v Microsoft Excel [Scientific computing in Microsoft Excel]. Moskva: Vil'yams, 2004. 512 p.

3. Polovko A.M., Butusov P.N. Interpoljacija. Metody i komp'juternye tehnologii ih realizujushhie [Interpolation. Methods and computer technologies for their implementation]. Sankt-Peterburg: BKhV-Peterburg, 2004. 320 p.

4. Bl2. Function Interpolation. URL: bl2.ru/mathematics/interpolation.html.

5. Mathematical forum Math Help Planet. Methods of Interpolation and Regression. URL: mathhelpplanet.com/static.php?p=metody-interpolyatsii-i-regressii.

6. Klinaev Yu.V., Terin D.V. Metody i tehnologii komp'juternyh vychislenij v matematicheskom modelirovanii [Methods and technologies of computer calculations in mathematical modeling]. Saratov: Izd-vo SGTU, 2010. 208 p.

7. Vapnik V.N., Glazkova T.G., Koshcheev V.A., Mikhalskiy A.I., Chervonenkis A.Ya. Algoritmy i programmy vosstanovlenija zavisimostej [Algorithms and programs for dependency recovery]. Moskva: Nauka,1984. 816 p.

8. Vapnik V.N. Vosstanovlenie zavisimostej po jempiricheskim dannym [Restoring dependencies from empirical data]. Moskva: Nauka, 2019. 448 p.

9. Bespalova N.V., Korchagin S.A., Serdechny D.V. Inzhenernyj vestnik Dona, 2025, №5. URL: ivdon.ru/ru/magazine/archive/n5y2025/10037.

10. Korchagin S.A., Serdechny D.V., Okunev A.I., Andriyanov N.A. Inzhenernyj vestnik Dona, 2025, №3. URL: ivdon.ru/ru/magazine/archive/n3y2025/9914.

11. Planetcalc. Function Approximation. URL: planetcalc.ru/8735.

12. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2009. 745 p.

13. Kuhn M., Johnson K. Applied Predictive Modeling. New York: Springer, 2013. 600 p.

14. Ajvazjan S.A., Mhitarjan V.S. Prikladnaja statistika i osnovy jekonometriki [Applied statistics and basic econometrics]. Moskva: JuNITI, 1998. 1022 p.

Дата поступления: 27.05.2025

Дата публикации: 25.08.2025