

**ПРИМЕНЕНИЕ РЕГУЛЯРИЗАЦИИ К ИДЕНТИФИКАЦИИ МАТЕМАТИЧЕСКОЙ МОДЕЛИ  
ДЕСТРУКЦИИ ФЕНОЛЬНЫХ СОЕДИНЕНИЙ В ПРОМЫШЛЕННЫХ СТОЧНЫХ ВОДАХ  
ПО ТЕХНОЛОГИИ ADVANCED OXIDATION PROCESSES**

© 2015 Л.Э. Шейнкман, Д.В. Дергунов, Л.В. Котлеревская, Н.А. Антоненко

Тульский государственный университет

Статья поступила в редакцию 30.09.2015

Для эффективного управления процессом очистки сточных вод промышленных предприятий от фенольных соединений на основе усовершенствованных окислительных процессов предложена экспоненциальная модель снижения концентрации фенольных соединений (на примере Бисфенола-А) при воздействии физико-химических факторов (УФ излучения,  $\lambda = 365$  нм,  $\text{H}_2\text{O}_2$ ,  $\text{Fe}^{3+}$ ). Модель идентифицирована в программных средах Statistica и MathCad на основе метода наименьших квадратов и является статистически значимой результатам эксперимента, проведенного на основе оптимального планирования. Для стабилизации неустойчивых параметров модели применен метод регуляризации А.Н. Тихонова. Полученная регуляризованная модель, устанавливающая зависимость степени фотохимического разложения фенольных соединений в водной среде от параметров процесса, является статистически значимой и обладает лучшими прогнозными свойствами, чем исходная зависимость, идентифицированная по методу наименьших квадратов. Статистическая значимость регуляризованной модели определялась с использованием F-критерия Фишера, t-критерия Стьюдента и 95%-ных доверительных интервалов для регуляризованных параметров модели. Значение коэффициента детерминации,  $R^2 = 0,9995$ , показывает, что вариация физико-химических факторов при применении регуляризации на 99,95% объясняет вариацию концентрации фенольного соединения (ВРА). Значимость коэффициента детерминации подтверждается с 99%-ной вероятностью по критерию Фишера. Модель фотохимического разложения фенольных соединений в водной среде, полученная на основе применения метода регуляризации может давать более устойчивые прогнозные значения уровней содержания фенольных соединений в промышленных сточных водах.

**Ключевые слова:** Регуляризация, некорректные задачи, моделирование, фенол, сточные воды, очистка, усовершенствованные окислительные процессы.

**ВВЕДЕНИЕ**

Сточные воды ряда отраслей промышленности (химической, фармацевтической, металлургической, целлюлозно-бумажной, горно-перерабатывающей и др.) вносят значительный вклад в загрязнение поверхностных и подземных водных объектов фенольными и трудноокисляемыми органическими соединениями. В технологических схемах очистки промышленных вод от фенольных соединений на многих предприятиях используется обработка стоков коагулянтами и флокулянтами. Однако эти методы имеют ряд недостатков, приводящих к повышенному образованию осадка и солесодержанию.

Современным и высокоэффективным методом удаления органических загрязнителей, содержащихся в промышленных сточных водах, яв-

ляются процессы, основанные на использовании свободных радикалов в качестве окислителей, использующие технологию усовершенствованных окислительных процессов (Advanced Oxidation Processes – AOP) [1]. Достоинством AOP-процессов является легкость регулирования их скорости в широких пределах с изменением интенсивности светового потока, быстрое и полное их приостановление с прекращением облучения.

Для управления процессом очистки сточных вод от фенольных соединений и идентификации математической модели на основе применения фотохимических методов при совместном действии окислителей перекиси водорода, хлорида железа (III) и ультрафиолетового излучения длиной волны 365 нм проводились экспериментальные исследования с применением жидкостной и газовой хроматографии [2].

Управление процессами очистки сточных вод осуществляется на основе математических моделей, адекватно описывающих законы фотохимического разложения фенольных соединений в водной среде, представленные в зашифрованном виде экспериментальными данными. Для выявления этих законов и идентификации математических моделей применяется аппарат математической статистики, а именно регрессионный

*Шейнкман Леонид Элардович, доктор технических наук, профессор, профессор кафедры «Аэрология, охрана труда и окружающей среды». E-mail: eliard@yandex.ru*

*Дергунов Дмитрий Викторович, кандидат технических наук, инженер. E-mail: dmitrov83@mail.ru*

*Котлеревская Людмила Викторовна, кандидат технических наук, доцент. E-mail: eliard@yandex.ru*

*Антоненко Наталья Александровна, аспирант. E-mail: eliard@yandex.ru*

анализ, основанный на методе наименьших квадратов (МНК) [3]. Однако, алгебраическая задача МНК, применительно к оценке параметров математических моделей, описывающих процесс фотохимического разложения фенольных соединений в водной среде, является некорректно поставленной. Решение такой задачи может не существовать или не иметь единственного решения или может отсутствовать непрерывная зависимость решения от исходных данных, т.е. небольшим погрешностям исходных данных могут соответствовать существенные погрешности получаемых решений, что в конечном итоге приводит к неустойчивости оценок параметров идентифицируемой модели [4, 5]. Однако, при управлении процессом очистки сточных вод на стадии моделирования, варьированием уровнями факторов, необходимо получать адекватные прогнозные значения уровней остаточной концентрации фенольного соединения, содержащегося в воде по окончании очистки. Таким образом, целью работы является построение устойчивых решений при идентификации параметров математической модели, устанавливающей зависимость уровня снижения концентрации фенольных соединений в водной среде при воздействии фотохимических факторов.

### ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА

При проведении экспериментальных исследований, связанных с оценкой воздействия физико-химических факторов на уровень раз-

ложения фенольных соединений в водной среде, в качестве зависимой переменной выбран показатель «содержание остаточной концентрации ВРА на выходе процесса». На уровень остаточной концентрации фенольного загрязнителя в процессе фотодеградациии влияют следующие факторы: исходная концентрация ВРА в модельном растворе  $x_1 - C_{ВРА}$ , мг/л; концентрация перекиси водорода  $x_2 - C_{H_2O_2}$ , мг/л; концентрация активатора, содержащего ионы железа (III)  $x_3 - C_A$ , г/л; время облучения  $x_4 - t$ , ч.

При проведении опытных исследований каждый из факторов варьировался на двух уровнях – минимальном (кодированное значение – 1) и максимальном (кодированное значение +1). Границы пространства планирования заданы m-мерным кубом со сторонами  $-1 \leq x_j \leq +1, j = 1, 2, \dots, m$  [6]. Матрица планирования эксперимента приведена в табл. 1 [2].

### МОДЕЛИРОВАНИЕ И ОБРАБОТКА РЕЗУЛЬТАТОВ

Будем считать, что влияние, оказываемое воздействием физико-химических факторов, на уровень снижения концентрации фенольного соединения в водной среде можно описать моделью вида [7]:

$$y = \eta + \varepsilon = \sum_{i=1}^n \beta_i f_i(x_1, x_2, \dots, x_m) + \varepsilon, \quad (1)$$

где:  $f_i(x_1, x_2, \dots, x_m)$  – произвольные функции факторов (регрессоров);  $\beta_1, \beta_2, \dots, \beta_k$  – коэффициенты

Таблица 1. Матрица полного факторного эксперимента

Опыты	Планирование									Переменная состояния y
	z <sub>0</sub>	кодированный масштаб				натуральный масштаб				
		z <sub>1</sub>	z <sub>2</sub>	z <sub>3</sub>	z <sub>4</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>4</sub>	
1	+1	-1	-1	-1	-1	0,05	100	1	1	0,036
2	+1	+1	-1	-1	-1	0,1	100	1	1	0,071
3	+1	-1	+1	-1	-1	0,05	200	1	1	0,035
4	+1	+1	+1	-1	-1	0,1	200	1	1	0,068
5	+1	-1	-1	+1	-1	0,05	100	2	1	0,033
6	+1	+1	-1	+1	-1	0,1	100	2	1	0,064
7	+1	-1	+1	+1	-1	0,05	200	2	1	0,031
8	+1	+1	+1	+1	-1	0,1	200	2	1	0,061
9	+1	-1	-1	-1	+1	0,05	100	1	2	0,032
10	+1	+1	-1	-1	+1	0,1	100	1	2	0,062
11	+1	-1	+1	-1	+1	0,05	200	1	2	0,03
12	+1	+1	+1	-1	+1	0,1	200	1	2	0,059
13	+1	-1	-1	+1	+1	0,05	100	2	2	0,028
14	+1	+1	-1	+1	+1	0,1	100	2	2	0,056
15	+1	-1	+1	+1	+1	0,05	200	2	2	0,027
16	+1	+1	+1	+1	+1	0,1	200	2	2	0,053

модели;  $\varepsilon$  – ошибка эксперимента.

Матрица регрессоров  $F$  (при планировании эксперимента  $F$  – расширенная матрица эксперимента) представляет собой следующую  $(N \times k)$  – матрицу:

$$F = \begin{pmatrix} f_{11} & f_{12} & \dots & f_{k1} \\ f_{12} & f_{22} & \dots & f_{k2} \\ \dots & \dots & \dots & \dots \\ f_{1N} & f_{2N} & \dots & f_{kN} \end{pmatrix}.$$

В каждой строке  $F$  записаны значения функций  $f_{iu}$  из модели (1) при данном наблюдении ( $u = 1, 2, \dots, N$ ), т.к. в регрессионном анализе предварительно предполагается некоторая структура модели, функции  $f_{iu}$  известны.

Исходя из закона действующих масс [8], которому подчиняется гомогенный фото-каталитический процесс разложения органических соединений, зависимость концентрации фенольного соединения от факторов процесса математически можно описать следующим выражением:

$$\eta = \beta_1 x_1^{\beta_2} x_2^{\beta_3} x_3^{\beta_4} x_4^{\beta_5}, \quad (2)$$

где:  $\eta$  – уровень остаточной концентрации ВРА в момент времени  $t$ , мг/л;  $x_1$  – исходная концентрация ВРА, мг/л;  $x_2$  – концентрация перекиси водорода, мг/л;  $x_3$  – концентрация хлорида железа (III), г/л;  $x_4$  – время процесса очистки, ч;  $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$  – параметры модели.

Коэффициенты  $\beta$  в модель (2) входят нелинейно, но при линеаризации путем логарифмирования по натуральному основанию, правой и левой частей уравнения (2), получим:

$$\ln \tilde{\eta} = \ln \tilde{\beta}_1 + \beta_2 \ln \tilde{x}_1 + \beta_3 \ln \tilde{x}_2 + \beta_4 \ln \tilde{x}_3 + \beta_5 \ln \tilde{x}_4,$$

где в соответствии с (1):  $\eta = \ln \tilde{\eta}$ ,  $\beta = \ln \tilde{\beta}_1$ ,  $f_1 = 1$ ,  $f_2 = \ln \tilde{x}_1$ ,  $f_3 = \ln \tilde{x}_2$ ,  $f_4 = \ln \tilde{x}_3$ ,  $f_n = f_5 = \ln \tilde{x}_4$ .

Однако при таком преобразовании случайное возмущение (ошибка эксперимента) входит в модель мультипликативно и имеет логнормальное распределение, т.е.  $y = \tilde{\eta}\varepsilon$ , и после логарифмирования это дает:  $\ln y = \ln \tilde{\eta} + \ln \varepsilon$  [7].

После линеаризации и введении новых переменных, выражение (2) примет вид:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4, \quad (3)$$

где: предикторные переменные  $X_1, X_2, X_3, X_4$  и отклик  $Y$  являются логарифмическими функциями:  $Y = \ln y$ ,  $X_1 = \ln x_1$ ,  $X_2 = \ln x_2$ ,  $X_3 = \ln x_3$ ,  $X_4 = \ln x_4$ ;  $b_0, b_1, b_2, b_3, b_4$  – параметры модели.

Обычно в задачах обработки данных матрица эксперимента и вектор отклика известны не точно, т.е. с погрешностями. Но даже при точно заданной входной матрице конечная разрядность ЭВМ, наличие ошибок округления и их накопление в процессе вычисления коэффициентов регрессии приводят к неопределенности представления матрицы эксперимента на этапах вы-

числительной процедуры. Поэтому погрешности в матрице данных принципиально неустранимы, т.е. задача определения регрессионных коэффициентов по методу наименьших квадратов неустойчива к ошибкам в исходных данных [4].

Одно из основных предположений классической процедуры регрессионного анализа заключается в том, что матрица регрессоров  $F$  имеет полный ранг, а информационная матрица  $G = F^T F$  – невырождена [7].

При вырожденности информационной матрицы нельзя вычислить коэффициенты регрессионной модели по формуле:

$$b = (F^T F)^{-1} F^T Y, \quad (4)$$

поскольку не существует матрица  $C = (F^T F)^{-1}$ . Действительно, элементы матрицы  $C$  находят следующим образом:

$$c_{ij} = (-1)^{i+j} |G^{ij}| / |G|.$$

Как известно, линейным преобразованием можно любую матрицу привести к диагональному виду. Пусть  $V$  будет ортогональной матрицей, столбцы которой – ортонормированные собственные векторы матрицы  $F^T F$ , т.е.

$$(FV)^T FV = V^T F^T FV = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k),$$

где  $\lambda_1, \lambda_2, \dots, \lambda_k$  – собственные числа матрицы  $F^T F$ .

Ортогональность матрицы  $V$  означает, что  $V^T V = I$  и скалярное произведение любых двух столбцов матрицы  $V$  равно

$$V_i^T V_j = \sum_{l=1}^k V_{il} V_{jl} = \begin{cases} 1 & \text{при } i = j, \\ 0 & \text{при } i \neq j. \end{cases}$$

Из линейной алгебры известно, что если  $\lambda_i$  – собственное число  $F^T F$ , то

$$F^T F V_i = \lambda_i V_i,$$

а также

$$|G| = |F^T F| = \lambda_1 \lambda_2 \dots \lambda_k.$$

Следовательно, информационная матрица  $G$  вырождена, если хотя бы одно из её собственных чисел  $\lambda_i$  равно нулю. Тогда выполняется условие

$$F^T F V_i = 0,$$

что возможно только при  $F V_i = 0$ , когда строки матрицы  $F$  ортогональны к  $V_i$  и существует зависимость:

$$V_{1i} f_1 + V_{2i} f_2 + \dots + V_{ki} f_k = 0. \quad (5)$$

Уравнение (5) означает, что существует линейная зависимость между столбцами матрицы  $F$ , т.е. хотя бы один из них можно выразить как линейную комбинацию остальных. Если матрица  $F^T F$  имеет  $p$  нулевых собственных чисел, существует  $p$  независимых между собой уравнений типа (5). Если столбцы матрицы  $F$  рассматривать как векторы, то  $p$  из них будут коллинеарны, т.е. будут иметь совпадающие направления. Такое

явление называют *строгой мультиколлинеарностью*. Оно означает, что ранг матрицы  $F^T F$  равен  $k - p$  (числу линейно независимых столбцов  $F$ ). Иногда уравнение (5) выполняется лишь приближительно. Это явление называется *мультиколлинеарностью* и имеет отрицательные последствия для оценивания регрессионных коэффициентов.

В случае мультиколлинеарности одно или несколько собственных чисел  $G = F^T F$  очень близки к нулю, поэтому близок к нулю и определитель информационной матрицы  $|G|$ . А так как обращение информационной матрицы связано с делением на  $|G|$ , который при сильной мультиколлинеарности будет очень малым числом, малые ошибки в вычислениях могут вызвать существенные различия в рассчитываемых элементах  $G = F^T F$ , а значит и в оцениваемых коэффициентах регрессии.

Отрицательные последствия мультиколлинеарности заключаются в неустойчивости оценок коэффициентов модели, численной неустойчивости процедуры оценивания, вызванной ошибками машинного округления и накоплением ошибок.

В линейной алгебре для исследования мультиколлинеарности прибегают к числам обусловленности. Для информационной матрицы плана  $G = F^T F$  число обусловленности может быть определено как отношение максимального и минимального значений собственных чисел  $\lambda$ , т.е.  $P = \text{Cond}(F^T F) = \lambda_{\max} / \lambda_{\min}$ . Чем больше число обусловленности, тем сильнее ошибка в исходных данных сказывается на решении линейной системы, т.е. число обусловленности характеризует точность решения [7].

Определяя в системе MathCad вектор собственных чисел информационной матрицы Фишера ( $F^T F$ ), получим:  $\lambda\{F^T F\} = (0,058; 526,312; 1,922; 1,922; 1,922)$ ; максимальное и минимальное собственные числа матрицы  $F^T F$  равны соответственно  $-\lambda_{\max} = 526,312$ ;  $\lambda_{\min} = 0,0584$ . Число обусловленности информационной матрицы, в смысле евклидовой нормы, определенное через отношение собственных чисел составляет  $P(F^T F) = 9 \cdot 10^5$ . Сильная мультиколлинеарность проявляется при  $\text{Cond}(F^T F) \geq 10^5 - 10^6$  [7].

### РЕГУЛЯРИЗАЦИЯ

Учитывая, что алгебраически задача определения оценок неизвестных параметров множественной линейной регрессии сводится к вычислению псевдорешений  $\tilde{\varphi}$  системы линейных алгебраических уравнений (СЛАУ):

$$A\varphi = f, \quad (6)$$

где:  $A \in R^{m \times n}$ ,  $m > n$ ,  $\text{rank} A = n$ ,  $\varphi \in R^n$ ,  $f \in R^m$ , то наиболее эффективным способом «борьбы» с мультиколлинеарностью является идея регуляризации псевдорешений, изложенная в работах А.Н. Тихонова [4, 5, 9, 10].

Согласно методу регуляризации А.Н. Тихонова, решение  $\tilde{\varphi}$  определяется из минимизации функционала

$$M[\tilde{\varphi}] = \|A\tilde{\varphi} - f\|_2^2 + \alpha \|\tilde{\varphi}\|_2^2,$$

то есть регуляризованное псевдорешение СЛАУ (6) определяется как решение задачи

$$\tilde{\varphi}_\alpha = \arg \min_{x \in R^n} \left\{ \|A\tilde{\varphi} - f\|_2^2 + \alpha \|\tilde{\varphi}\|_2^2 \right\}, \quad (7)$$

где:  $\|\bullet\|_2$  – евклидова норма вектора;  $\alpha$  – параметр регуляризации.

Решение регуляризованной задачи (7) эквивалентно решению системы уравнений:

$$(A^T A + \alpha I) \tilde{\varphi}_\alpha = A^T f,$$

где:  $A$  – матрица СЛАУ;  $f$  – вектор-столбец правой части СЛАУ;  $T$  – символ транспонированная;  $I$  – единичная матрица.

Регуляризация А.Н. Тихонова, в решении регрессионных задач, представляющая попытку преодолеть последствия плохой обусловленности матрицы ( $F^T F$ ), интерпретирована рядом исследователей как подход, названный гребневой регрессией (ридж-регрессией) [7, 11, 12].

Процедура построения *гребневой регрессии*, впервые предложенная в 1962 году А.Э. Хоэрлом, предназначена для работы в случаях, когда существование приближенных зависимостей между столбцами матрицы  $F$  приводит к тому, что матрица  $F^T F$  близка к вырожденной, из-за чего оценки параметров становятся неустойчивыми и, как правило, обладают большими стандартными ошибками. При использовании *метода гребневой регрессии* для стабилизации МНК-оценок, определяемых по формуле (4) при плохой обусловленности информационной матрицы, регуляризация связана с добавлением некоего положительного числа  $\tau$  (параметра регуляризации) к диагональным элементам матрицы  $F^T F$  [11].

При условии выбора подходящего значения параметра регуляризации  $\tau$ , можно существенно уменьшить квадратичную ошибку МНК-оценок за счет их небольшого смещения. Выбор параметра регуляризации Хоэрл, Кеннард и Белдвин [11, 12] предложили осуществлять следующим образом:

$$\tau = \frac{m \cdot SS_e}{n - m - 1} \cdot \frac{1}{b^{*T} b^*}, \quad (8)$$

где:  $m$  – количество параметров (без учета свободного члена) в исходной модели регрессии;  $SS_e$  – остаточная сумма квадратов, полученная по исходной модели регрессии без корректировки на мультиколлинеарность;  $b^*$  – вектор-столбец коэффициентов регрессии, преобразованный по формуле:

$$b_j^* = b_j \sqrt{\sum (X_j - \bar{X}_j)^2}, \quad (9)$$

где:  $b_j$  – параметр при переменной  $X_j$  в исходной

модели регрессии, определенный по МНК;  $\bar{X}_j$  – среднее значение  $j$ -ой независимой переменной.

После выбора величины  $\tau$  формула для оценки регуляризованных параметров регрессии будет иметь вид:

$$b_\tau = (F^T F + \tau I)^{-1} F^T Y, \quad (10)$$

где:  $I$  – единичная матрица;  $F$  – матрица регрессоров;  $Y$  – вектор значений зависимой переменной.

Оценки параметров регрессии, полученные по формуле (10), будут смещенными. Однако, т.к. определитель матрицы  $(F^T F + \tau I)$ , больше определителя матрицы  $(F^T F)$ , дисперсия оценок параметров регрессии уменьшится, что положительно повлияет на прогнозные свойства модели [12].

В работах [7, 13] показано, что регуляризованные оценки можно записать еще и следующими двумя способами:

$$b_\tau = \left[ I + \tau (F^T F)^{-1} \right]^{-1} b = Zb, \quad (11)$$

или

$$b_\tau = \left[ I - \tau (F^T F + \tau I)^{-1} \right] b = (I - \tau W) b, \quad (12)$$

где  $Z = \left[ I + \tau (F^T F)^{-1} \right]^{-1}$ , а  $W = (F^T F + \tau I)^{-1}$ .

Если рассматривать математическое ожидание обоих выражений (12), выясняется, что регуляризованные оценки оказываются смещенными и смещение их равно:

$$E(b_\tau) - \beta = -\tau (F^T F + \tau I)^{-1} \beta. \quad (13)$$

Очевидно, что смещение зависит как от параметра регуляризации  $\tau$ , так и от неизвестных значений регрессионных коэффициентов  $\beta$ .

За счет смещения можно достичь уменьшения квадратичной ошибки оценок  $M(\tau)$ , которая разложима на две составляющие:

$$M(\tau) = \sum_{i=1}^k (b_{\tau_i} - \beta_i)^2 = E(\hat{b}_\tau - \beta)^T (\hat{b}_\tau - \beta) = \gamma_1(\tau) + \gamma_2(\tau). \quad (14)$$

Первая составляющая:

$$\gamma_1(\tau) = \sigma^2 \sum_{i=1}^k (\lambda_i / (\lambda_i + \tau)^2), \quad (15)$$

где  $\lambda_i$  – собственные числа матрицы  $F^T F$ . Она равна сумме дисперсий оценок коэффициентов.

Второй член

$$\gamma_2(\tau) = \tau^2 \beta (F^T F + \tau I)^{-2} \beta = \sum_{i=1}^k \left[ \alpha_i^2 / ((\lambda_i / \tau) + 1)^2 \right], \quad (16)$$

где:  $\alpha = V\beta$ , а  $V$  – ортогональная матрица, столбцы которой – собственные векторы  $F^T F$ . Тогда  $V^T F^T F V = \Lambda$ , где  $\Lambda$  – диагональная матрица с собственными числами на диагонали  $\lambda_1, \lambda_2, \dots, \lambda_k$ .

При сравнении (13) и (16), видно, что  $\gamma_2(\tau)$  равна квадрату смещения оценок регрессионных коэффициентов. Из (15) и (16) следует, что с ростом параметра регуляризации  $\tau$  сумма дис-

персий оценок коэффициентов  $\gamma_1(\tau)$  монотонно убывает, в то время как квадрат смещения оценок монотонно возрастает. Очевидно, квадратичная ошибка оценки имеет минимум при  $\gamma_1(\tau) = \gamma_2(\tau)$ , что видно из (14).

Доказано [13], что всегда существует такое значение  $\tau$ , при котором  $M(\tau) < M(0)$ ,  $M(0)$  – квадратичная ошибка, полученная МНК-оцениванием при  $\tau = 0$ .

Таким образом, при выборе подходящего значения параметра регуляризации, уменьшается квадратичная ошибка оценок [7].

## ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Для идентификации оценок коэффициентов регрессионной модели (3) по МНК в соответствии с формулой (4) использовались пакеты прикладных программ Statistica v.6.1 и MathCad 14.0. Анализ экспериментальных данных, проводимый в Statistica и MathCad позволил определить вектор коэффициентов модели  $b_{\text{МНК}} = (-0,080788; 0,971642; -0,069869; -0,156662; -0,203063)$  и идентифицировать модель вида:

$$C_{\text{ост}} = \exp \left( \begin{matrix} -0,080788 + 0,971642 \ln C_{\text{BPA}} - 0,069869 \ln C_{\text{H}_2\text{O}_2} \\ -0,156662 \ln C_{\text{A}} - 0,203063 \ln t \end{matrix} \right), \quad (17)$$

где:  $C_{\text{ост}}(t)$  и  $C_0$  – остаточная и начальная концентрации фенольного загрязнителя соответственно, мг/л;  $C_{\text{H}_2\text{O}_2}$  – концентрация перекиси водорода, мг/л;  $C_{\text{A}}$  – концентрация хлорида железа (III), г/л;  $t$  – время, ч.

Значение коэффициента детерминации, близкого к единице ( $R^2=0,999$ ), показывает, что вариация факторов на 99,9% объясняет вариацию остаточной концентрации фенольного соединения при фотохимическом воздействии. Значимость коэффициента детерминации подтверждается с 99%-ной вероятностью (уровень значимости  $\alpha=0,01$ ), в связи с тем, что критическое значение критерия Фишера  $F_{\text{кр}}$  со степенями свободы  $\nu_1 = 4$ ,  $\nu_2 = 11$  не превышает расчетного значения для регрессионной модели ( $F_{\text{кр}}(0,01;4,11)=5,67 < F=7292,13$ ). На основе  $t$ -критерия Стьюдента установлена значимость оценок параметров модели (17) на уровне  $\alpha = 0,01$ , свободного члена на уровне  $\alpha = 0,05$  (модули  $t$ -критерия  $t_1=164,7$ ;  $t_2=11,8$ ;  $t_3=26,6$ ;  $t_4=34,4$  превышают критическое значение  $t(0,99;11)=2,72$ , а модуль  $t$ -критерия для свободного члена  $t_0=2,4 > t(0,95;11)=1,79$ ). О 99%-ной достоверности модели (17) экспериментальным данным свидетельствует принадлежность оценки дисперсии ошибок наблюдений 99%-ному доверительному интервалу ( $S^2=6,7 \cdot 10^{-5} \in (2,75 \cdot 10^{-5}; 28,25 \cdot 10^{-5})$ ). Статистика Дарбина-Уотсона  $DW=2,02$  и сериальная корреляция остатков  $\rho_{\text{cor}} = -0,092$ , подтверждают высокую степень адекватности модели (17) результатам эксперимента [2].

Учитывая, что остаточная сумма квадратов в исходной модели регрессии (17) составляет

$SS_e = 735,4402 \cdot 10^{-6}$  и квадратный корень из суммы квадратов  $\sqrt{\sum (X_j - \bar{X}_j)^2} = 1,3863$ , то вектор преобразованных коэффициентов регрессии, определяемый по формуле (9), будет иметь вид  $b_j^* = \{1,346987; -0,096859; -0,217181; -0,281506\}$ ,  $j = 1, \dots, 4$ . Величина параметра регуляризации, определяемая по формуле (8) будет иметь значение, равное  $\tau = 1,371 \cdot 10^{-4}$ .

Точность оценивания определяется по величине квадратичной ошибки оценки коэффициентов (СКОК):

$$СКОК = \sum_{i=1}^k (\hat{b}_i - b_i)^2,$$

где:  $\hat{b}_i$  – оценки значений коэффициентов модели, полученные методом регуляризации;  $b_i$  – значения коэффициентов модели, полученные по МНК.

Дополнительно к определенному по формуле (8) параметру регуляризации, зададимся рядом иных значений величины  $\tau$ . Результаты оценивания параметров модели (3) методом регуляризации при различных значениях  $\tau$  приведены в табл. 2.

Из табл. 2 видно, что наименьшая квадратичная ошибка оценок получается при  $\tau = 10^{-5}$ . Однако, при очень малых значениях параметра регуляризации её эффект незначителен, а при больших  $\tau$  влияние данных падает по сравнению с влиянием члена  $\tau l$ . Авторы работы [14] для того, чтобы не сильно «испортить» задачу, рекомендуют выбирать параметр регуляризации весьма малым, порядка  $10^{-3} - 10^{-9}$ .

Оценка параметра регуляризации  $\tau = 1,371 \cdot 10^{-4}$ , принятая в соответствии с рекомендациями Хоэрла, Кеннарда и Белдвина [11, 12] дает приемлемый минимальный уровень квадратической

ошибки, хотя это понятие субъективное [7], т.е. приемлемый выбор параметра регуляризации основывается на гребневом следе и варьируется в данном случае вокруг оптимального значения  $1,371 \cdot 10^{-4}$ . При использовании пакетов прикладных программ Statistica и MathCad идентифицирован вектор регуляризованных оценок коэффициентов регрессии  $b_r = \{0,081205; 0,971509; 0,069860; 0,156641; 0,203035\}$ , и регуляризованная модель снижения концентрации фенольного соединения в виде:

$$C_{ост} = \exp \left( \begin{matrix} -0,081205 + 0,971509 \ln C_{BPA} - 0,06986 \ln C_{H_2O_2} - \\ -0,156641 \ln C_A - 0,203035 \ln t \end{matrix} \right), \quad (18)$$

где:  $C_{ост}(t)$  и  $C_0$  – остаточная и начальная концентрация фенольного загрязнителя соответственно, мг/л;  $C_{H_2O_2}$  – концентрация перекиси водорода, мг/л;  $C_A$  – концентрация хлорида железа (III), г/л;  $t$  – время, ч.

Значение коэффициента детерминации,  $R^2 = 0,9995$ , показывает, что вариация факторов при применении параметра регуляризации  $\tau$  на 99,95% объясняет вариацию остаточной концентрации фенольного соединения (BPA) по модели (18). Значимость коэффициента детерминации подтверждается с 99%-ной вероятностью (уровень значимости  $\alpha=0,01$ ), в связи с тем, что критическое значение критерия Фишера  $F_{кр}$  со степенями свободы  $v_1 = 4$ ,  $v_2 = 11$  не превышает расчетного значения для регуляризованной регрессионной модели ( $F_{кр}(0,01; 4, 11) = 5,67 < F = 5348,417$ ). На основе t-критерия Стьюдента установлена значимость оценок регуляризованных параметров модели  $b_{jr}$  на уровне  $\alpha = 0,01$ , свободного члена  $b_{or}$  на уровне  $\alpha = 0,05$ . Уровни значений статистики t-критерия Стьюдента для регуляризованных параметров модели (18) со-

Таблица 2. Результаты оценивания параметров модели

Коэффициенты модели МНК	$b_1$	$b_2$	$b_3$	$b_4$	$СКОК$
		0,971642	-0,069869	-0,156662	-0,203063
$\tau = 1,37 \cdot 10^{-4}$	0,971509	-0,069860	-0,156641	-0,203035	$1,9 \cdot 10^{-8}$
$10^{-4}$	0,971545	-0,069862	-0,156647	-0,203043	$1,008 \cdot 10^{-8}$
$2 \cdot 10^{-4}$	0,971448	-0,069855	-0,156631	-0,203022	$4,047 \cdot 10^{-8}$
$5 \cdot 10^{-4}$	0,971157	-0,069834	-0,156584	-0,202961	$2,529 \cdot 10^{-7}$
$10^{-5}$	0,971632	-0,069868	-0,156661	-0,203061	$1,06 \cdot 10^{-10}$
$10^{-3}$	0,970672	-0,069799	-0,156506	-0,202860	$1,011 \cdot 10^{-6}$
$1,836 \cdot 10^{-3}$	0,969862	-0,069741	-0,156375	-0,202691	$3,406 \cdot 10^{-6}$
$1,2 \cdot 10^{-2}$	0,960121	-0,069041	-0,154805	-0,200655	$1,427 \cdot 10^{-4}$
$1,7 \cdot 10^{-2}$	0,955400	-0,068701	-0,154044	-0,199668	$2,835 \cdot 10^{-4}$
$10^{-2}$	0,962022	-0,069177	-0,155111	-0,201052	$9,947 \cdot 10^{-5}$
$1,55 \cdot 10^{-1}$	0,841249	-0,060493	-0,135639	-0,175812	0,018
$10^{-1}$	0,883311	-0,063517	-0,142420	-0,184603	$8,386 \cdot 10^{-3}$

ставляют:  $t\{b_{0r}\} = 2,0880$ ;  $t\{b_{1r}\} = 141,0815$ ;  $t\{b_{2r}\} = 10,1449$ ;  $t\{b_{3r}\} = 22,7472$ ;  $t\{b_{4r}\} = 29,4845$ . Модули значений t-критерия Стьюдента регуляризованных параметров  $b_{jr}$  превышают критическое значение t-критерия со степенью свободы  $\nu = 11$  на уровне значимости  $\alpha = 0,01 - t(0,99;11) = 2,72$ . Для свободного члена регуляризованной модели модуль расчетного значения t-критерия Стьюдента превышает критическое значение при уровне значимости  $\alpha = 0,05$  ( $t\{b_{0r}\} = 2,0880 > t(0,95;11) = 1,7959$ ). Таким образом, регуляризованное значение оценки свободного члена, равное  $b_{r0} = 0,081205$  значимо на уровне  $\alpha = 0,05$ . Стандартная ошибка оценки модели (18) составляет 0,00955; стандартная ошибка оценки свободного члена – 0,038892. Статистика Дарбина-Уотсона  $DW=2,019$  и сериальная корреляция остатков  $\rho_{cor} = -0,091273$ , подтверждают высокую степень адекватности регуляризованной модели (18) результатам эксперимента.

Доверительные интервалы для истинных значений параметров  $\tilde{b}_j$  множественной регрессионной модели в матричной форме имеют вид [15]: 
$$\tilde{b}_j - t_{1-\frac{\alpha}{2}}(N-m-1) \cdot \sqrt{S^2 \left( (F^T F)^{-1} \right)_{jj}} < b_j < \tilde{b}_j + t_{1-\frac{\alpha}{2}}(N-m-1) \cdot \sqrt{S^2 \left( (F^T F)^{-1} \right)_{jj}},$$
  $j=1, \dots, m$ , где:  $t_{1-\frac{\alpha}{2}}(N-m-1)$  – квантиль распределения Стьюдента порядка  $(1 - \alpha/2)$  с  $(N - m - 1)$  степенями свободы;  $S^2 \left( (F^T F)^{-1} \right)_{jj}$  – оценка дисперсии ошибки наблюдений (остаточная дисперсия);  $\left( (F^T F)^{-1} \right)_{jj}$  – диагональный элемент корреляционной матрицы;  $N$  – число наблюдений (строк матрицы планирования);  $m$  – число факторов модели;  $\alpha$  – уровень значимости.

Учитывая, что корень квадратный из дисперсии есть стандартные отклонения оценок параметров регрессии  $\tilde{b}_j$  от истинных значений  $b_j$ , т.е. стандартные ошибки определения параметров регрессии  $b_j$ , то доверительный интервал упрощается к виду:

$$\tilde{b}_j \pm S\{\tilde{b}_j\} \cdot t_{1-\frac{\alpha}{2}}(N-m-1), \quad (19)$$

где:  $S\{\tilde{b}_j\}$  – стандартные ошибки отклонения МНК-оценок регрессии  $\tilde{b}_j$  от истинных значений  $b_j$ ;  $t_{1-\frac{\alpha}{2}}(N-m-1)$  – квантиль распределения

Стьюдента порядка  $(1 - \alpha/2)$  с  $(N - m - 1)$  степенями свободы;  $N$  – число строк матрицы планирования;  $m$  – число параметров модели.

Вероятностным калькулятором системы Statistica при доверительной вероятности  $p = 1 - \alpha = 0,95$  ( $\alpha = 0,05$ ) определен квантиль распределения Стьюдента:  $t_{0,975}(11) = 2,200985$ .

Доверительные интервалы (95%-ные) для истинных значений параметров регрессии  $\tilde{b}_j$ , построенные на основании МНК-оценок  $b_{jMНК}$  в соответствии с (19) имеют вид:

- для  $b_0$ :  
 $-0,080788 \pm 2,200985 \cdot 0,0333 = (-0,154081; -0,007495)$ ;
- для  $b_1$ :  
 $0,971642 \pm 2,200985 \cdot 0,005898 = (0,958661; 0,984623)$ ;
- для  $b_2$ :  
 $-0,069869 \pm 2,200985 \cdot 0,005898 = (-0,082851; -0,056888)$ ;
- для  $b_3$ :  
 $-0,156662 \pm 2,200985 \cdot 0,005898 = (-0,169643; -0,143681)$ ;
- для  $b_4$ :  
 $-0,203063 \pm 2,200985 \cdot 0,005898 = (-0,216044; -0,190082)$ .

95%-ные доверительные интервалы для регуляризованных коэффициентов регрессии определялись в соответствии с формулой (19) с учетом того, что уровни стандартных ошибок отклонения регуляризованных оценок регрессии от истинных значений, оцененные в Statistica составляют:  $S\{b_{0r}\} = 0,038896$ ;  $S\{b_{1r}\} = S\{b_{2r}\} = S\{b_{3r}\} = S\{b_{4r}\} = 0,006887$ .

Доверительные интервалы для истинных значений параметров регрессии, построенные на основании регуляризованных оценок  $b_{jr}$ , принимают вид:

- для  $b_0(r)$ :  
 $-0,081206 \pm 2,200985 \cdot 0,038896 = (-0,166815; 0,004403)$ ;
- для  $b_1(r)$ :  
 $0,971509 \pm 2,200985 \cdot 0,006887 = (0,956351; 0,986667)$ ;
- для  $b_2(r)$ :  
 $-0,069860 \pm 2,200985 \cdot 0,006887 = (-0,085018; -0,054702)$ ;
- для  $b_3(r)$ :  
 $-0,156641 \pm 2,200985 \cdot 0,006887 = (-0,171799; -0,141483)$ ;
- для  $b_4(r)$ :  
 $-0,203035 \pm 2,200985 \cdot 0,006887 = (-0,218193; -0,187877)$ .

Область 95%-х доверительных интервалов для истинных оценок параметров модели, построенных на основе регуляризованных оценок коэффициентов регрессии  $b_{jr}$  пересекается с областью 95%-х доверительных интервалов, построенных для МНК-оценок коэффициентов, т.е. области доверительных интервалов являются смещенными, как и регуляризованные оценки параметров модели. Регуляризованные оценки являются значимыми, т.к. принадлежат 95%-м доверительным интервалам, построенным для значений этих коэффициентов.

## ВЫВОДЫ

Обобщая приведенные выше результаты исследований, можно сделать вывод о том, что для идентифицированной по методу наименьших квадратов экспоненциальной модели снижения концентрации фенольных соединений в водной среде под действием физико-химических факторов проведена регуляризация, направленная на стабилизацию неустойчивых к ошибкам, ко-

эффициентов модели. Регуляризованные оценки параметров модели получены методом гребневой регрессии с использованием приемлемого значения параметра регуляризации, выбранного способом, предложенным иностранными специалистами в области решения некорректных задач.

Полученная регуляризованная модель, устанавливающая зависимость степени разложения фенольных соединений в водной среде от параметров процесса (начальной концентрации фенольного соединения, концентрации перекиси водорода, концентрации хлорида железа (III), времени ультрафиолетового воздействия на систему) является статистически значимой и обладает лучшими прогнозными свойствами, чем исходная модель, идентифицированная по методу наименьших квадратов. Таким образом, регуляризованная модель фотохимического разложения фенольных соединений может давать более устойчивые прогнозные значения уровней содержания фенольных соединений в промышленных сточных водах на этапах планирования и проведения водоочистных мероприятий.

#### СПИСОК ЛИТЕРАТУРЫ

1. Кармазинов, Ф.В., Костюченко, С.В., Кудрявцев, Н.Н., Храменков, С.В. Ультрафиолетовые технологии в современном мире. – Долгопрудный, Издательский Дом «Интеллект», 2012. – 392 с.
2. Соколов, Э.М., Шейнкман, Л.Э., Дергунов, Д.В. Исследование снижения концентрации фенольных

- соединений в водных средах с использованием математического моделирования // Вестник Южного научного центра РАН. – 2013. – Т. 9, №2. – С. 23 – 31.
3. Марков, Ю.Г., Маркова, И.В. Математические модели химических реакций. – СПб.: Издательство «Лань», 2013. – 192 с.
4. Тихонов, А.Н., Уфимцев, М.В. Статистическая обработка результатов экспериментов. – М.: Издательство МГУ, 1988. – 174 с.
5. Тихонов, А.Н., Арсенин, В.Я. Методы решения некорректных задач. – М.: Наука, 1979. – 285 с.
6. Ермаков, С.М., Жиглявский, А.А. Математическая теория оптимального эксперимента. – М.: Наука, 1987. – 320 с.
7. Вучков, И., Бояджиева, Л., Солаков, Е. Прикладной линейный регрессионный анализ. – М.: Финансы и статистика, 1987. – 240 с.
8. Эммануэль, Н.М., Кнорре, Д.Г. Курс химической кинетики. – М.: Высшая школа, 1984. – 463 с.
9. Тихонов, А.Н. Решение некорректно поставленных задач и метод регуляризации // Доклады АН СССР. – 1963. – Т. 151, №3. – С. 501 – 504.
10. Тихонов, А.Н. О регуляризации некорректно поставленных задач // Доклады АН СССР. – 1963. – Т. 153, №1. – С. 45 – 52.
11. Дрейнер, Н., Смит, Г. Прикладной регрессионный анализ. – М.: Издательский дом «Вильямс», 2007. – 912 с.
12. Елисеева, И.И. Эконометрика. – М.: Издательство Юрайт, 2014. – 449 с.
13. Hoerl, A., Kennard, R. Ridge regression: biased estimation for nonorthogonal problems// Technometrics. 1970. №12. – P. 55 – 67.
14. Ватульян, А.О., Беляк, О.А., Сухов, Д.Ю., Явруян, О.В. Обратные и некорректные задачи. – Ростов-на-Дону, Издательство Южного федерального университета, 2011. – 232 с.
15. Афифи, А., Эйзен, С. Статистический анализ. Подход с использованием ЭВМ. – М.: Мир, 1982. – 488 с.

#### APPLICATION OF REGULARIZATION TO THE IDENTIFICATION OF THE MATHEMATICAL MODEL OF THE DEGRADATION OF PHENOLIC COMPOUNDS IN INDUSTRIAL WASTEWATER TECHNOLOGY ADVANCED OXIDATION PROCESSES

© 2015 L.E. Sheinkman, D.V. Dergunov, L.V. Kotlerevskaya, N.A. Antonenko

Tula State University

For effective management of wastewater treatment process industrial enterprises of phenolic compounds on the basis of advanced oxidation processes proposed exponential model to reduce the concentration of phenolic compounds (for example, Bisphenol-A) under the action of physical and chemical factors (UV radiation,  $\lambda = 365$  nm,  $H_2O_2$ ,  $Fe^{3+}$ ). Model identified in the Statistica software environments and MathCad based on the method of least squares and is statistically significant results of the experiment carried out on the basis of optimal planning. For stabilization of unstable parameters of the model, the method of regularization A.N. Tikhonov. The resulting regularized model establishing the degree of photochemical degradation of phenolic compounds in the aquatic environment of the process parameters is statistically significant and has the best forward-looking features than the original relationship, identified by the method of least squares. The statistical significance of a regularized model was determined using Fisher's F-test, t-Student's t test and 95% confidence intervals for the regularized model parameters. The coefficient of determination,  $R^2 = 0,9995$ , shows that the variation of physical and chemical factors in the application of regularization to explain 99,95% of the variation of concentration of phenolic compounds (BPA). The significance of the coefficient of determination is confirmed with 99% probability by the Fisher test. Model photochemical decomposition of phenolic compounds in an aqueous medium, obtained by applying the regularization method may provide more stable levels of predicted values of phenolic compounds in industrial waste water.

**Keywords:** Regularization, incorrect problems, modeling, phenol, waste water, treatment, advanced oxidation processes.

Leonid Sheinkman, Doctor of Technical Sciences, Professor at the Aerology, Labor Protection and the Environment Department. E-mail: eliard@yandex.ru

Dmitriy Dergunov, Candidate of Technical Sciences, Engineer. E-mail: dmitrov83@mail.ru

Lyudmila Kotlerevskaya, Candidate of Technical Sciences, Associate Professor. E-mail: eliard@yandex.ru

Natalya Antonenko, Graduate Student.

E-mail: eliard@yandex.ru.