УДК 620.193.2

МОДЕЛЬ ПРОГНОЗА КОРРОЗИОННЫХ ПОТЕРЬ УГЛЕРОДИСТОЙ СТАЛИ ЗА ПЕРВЫЙ ГОД ЭКСПОЗИЦИИ НА ОСНОВЕ АЛГОРИТМА «СЛУЧАЙНЫЙ ЛЕС»

М.А. Гаврюшина*, Ю.М. Панченко и А.И. Маршаков

Институт физической химии и электрохимии им. А.Н. Фрумкина РАН
119071, Москва, Ленинский проспект, д. 31, корп. 4
**E-mail: maleeva.corlab@yandex.ru*

Аннотация

На основе алгоритма «случайный лес» (RF) получены две модели для прогноза первогодовых коррозионных потерь (K_1) углеродистой стали в открытой атмосфере в различных регионах мира. Первая модель RF_общая получена с использованием объединенных баз данных международных программ ISO CORRAG, MICAT, ECE/UN и испытаний на территории России и предназначена для оценки K_1 в различных типах атмосферы в различных регионах мира. Вторая модель RF_конт позволяет предсказать K_1 в континентальных районах мира. Проведено сравнение точности предсказаний K_1 по моделям RF и двум функциям «доза-ответ»: представленной в стандарте ISO 9223 и новой версии, разработанной ИФХЭ РАН для континентальных регионов. Показано, что достоверность обеих моделей RF существенно лучше, чем функций «доза-ответ», за исключением предсказаний коррозионных потерь стали на территории России с холодным климатом.

Ключевые слова: *малоуглеродистая сталь, модели, вероятная скорость коррозии, методы машинного обучения.*

Поступила в редакцию 12.02.2024 г.; После доработки 14.02.2024 г.; Принята к публикации 15.02.2024 г.

doi: 10.61852/2949-3412-2024-2-1-41-59

1. Введение

Коррозионные потери металлов в атмосфере могут варьироваться в больших интервалах в зависимости от агрессивности окружающей среды. По этой причине оправдан интерес к аналитическим и численным моделям, которые позволяют предсказывать массопотерю металлов в различных климатических регионах мира и

типах атмосферы. Наличие в атмосфере значительного числа агрессивных агентов, многостадийность, нелинейность и взаимное влияние физико-химических процессов, протекающих в тонком слое электролита на поверхности металла, делают задачу создания прогнозных моделей атмосферной коррозии очень трудной. Вместе с тем, для решения инженерных задач, таких как предсказание коррозионной стойкости материала конструкций, срока их службы, выбора средств антикоррозионной защиты, требуется разработка моделей, которые использовали бы минимальный набор параметров атмосферы. В идеале, для предсказания коррозионных потерь должны использоваться параметры, которые определяются на метеорологических станциях или на станциях, следящих за загрязнениями атмосферы, на всей территории Земного шара. В настоящее время этому требованию отвечают функции доза ответ (ФДО), которые позволяют предсказать массопотери металлов за первый год экспозиции (K_1) в зависимости от ограниченного числа климатических и аэрохимических параметров атмосферы. Величины K_1 необходимы для определения коррозионной агрессивности атмосферы [1] и для предсказаний долговременных коррозионных потерь в различных регионах мира без проведения натурных испытаний образцов металлов [2].

Модели для предсказания величин K_1 стандартных металлов в различных регионах мира описаны в международном стандарте (ФДО^{С}) [1]. Новая версия ФДО (ФДО^{Н}) для континентальных районов мира дана в [3, 4], численные коэффициенты ФДО^{Н} были уточнены в [5, 6]. ФДО^{С} были получены регрессионным анализом баз данных, которые включали экспериментальные коррозионные первогодовые потери типовых металлов ($K_1^{экс}$), метеорологические и аэрохимические параметры мест испытаний по программе ISO CORRAG [7] и проекту MICAT [8]. ФДО^{Н} – таким же методом, на основании данных программ ECE/UN [9] и российской [10]. Сопоставление величин K_1, рассчитанных в соответствии с этими ФДО, с величинами $K_1^{экс}$, показывает, что ошибка предсказаний довольно значительная [3−5]. В частности, отмечалось, что попытка разработать новые ФДО для прибрежных районов всего мира оказалась неудачной [6]. В связи с этим, представляется необходимым дальнейший поиск прогнозных моделей атмосферной коррозии металлов. В частности, для этого можно использовать методы машинного обучения.

Алгоритм случайного леса (RF) — один из популярных методов машинного обучения [11]. Алгоритм RF использовался для построения прогнозных моделей атмосферной коррозии малолегированных сталей [12, 13]. Скорости коррозии сталей, предсказанные RF моделью, искусственной нейронной сетью, методами регрессии опорных векторов и логистической регрессии, были сопоставлены с экспериментальными значениями, полученными в 10 местах экспозиции на территории Китая [12]. Оценка достоверности предсказаний скорости коррозии по таким статистическим показателям, как коэффициент детерминации (R^2), средняя абсолютная процентная ошибка (MAPE) и корень из среднеквадратичной ошибки (RMSE), показала преимущество RF модели [12]. RF модель, построенная на основе

базы данных, полученной в трех местах экспозиции сталей в открытой атмосфере и под навесом, также показала более точные предсказания скорости коррозии сталей по сравнению с другими методами машинного обучения [13]. В этом случае достоверность моделей оценивалась по величинам R^2 и средней абсолютной ошибке (MAE). Надо отметить, что RF модель, обученная по данным двух мест экспозиции, показала существенно большую ошибку предсказаний в третьем месте экспозиции, данные которого не использовались для обучения этой модели [13].

Алгоритм RF позволяет определить наиболее значимые параметры атмосферы, влияющие на коррозию металлов [12−14]. Это позволяет уменьшить число параметров во входных наборах, которые используются другими методами машинного обучения. Модель, в которой были объединены RF и алгоритм машинного обучения с учителем, была использована для предсказаний скорости коррозии углеродистой стали в 10 местах на территории Китая и показала высокую точность предсказаний [14]. Вместе с тем, достоверность RF моделей [12−14] не была проверена в различных регионах мира, то есть, в местах испытаний, результаты которых не были использованы при разработке этих моделей.

Целью настоящей работы является разработка RF модели на основании результатов первогодовых испытаний углеродистой стали по программам [7−10] и сопоставление величин K_1 стали, предсказанных по RF модели и функциям доза ответ [1, 5] в различных регионах мира.

2. Методика работы

2.1. Базы данных натурных коррозионных испытаний

Для разработки RF моделей использованы базы данных одногодовых экспозиций в каждом месте испытаний по программе ISO CORRAG [7] (далее БД ISO), проекту MICAT [8] (далее БД MICAT), по программе ECE/UN [9] (далее БД ECE/UN) и по российским программам [10] (далее БД RUS).

Из БД ISO использованы 234 набора, полученных в 41 местах за разные одногодовые испытания, включающих коррозионные потери стали, $K_1^{экс}$ (мкм) и соответствующие этому году среднегодовые значения параметров агрессивности атмосферы: температуры (T, °C) и относительной влажности воздуха (RH, %), концентрации SO_2 в воздухе ([SO_2], мкг/м^3) и осаждение хлоридов Cl^- ([Cl^-], мг/(м^2·сут)). Значения RH в отдельных местах приведены в соответствии с [15, 16]. Отсутствие данных концентрации SO_2 и осаждения хлоридов Cl^- в отдельных местах, в которых $K_1^{экс}$ имеют небольшие величины, заменили на фоновые значения, принятые условно: [SO_2] = 2,5 мкг/м^3 и [Cl^-] = 1,5 мг/(м^2·сут).

Из БД MICAT использованы 175 наборов данных, полученных в 66 местах испытаний. В набор данных входят величины $K_1^{экс}$, T, RH, [SO_2], [Cl^-] и годовое количество атмосферных осадков (*Prec*, мм/год).

БД ECE/UN состоит из 77 наборов данных, полученных в 27 континентальных местах испытаний. В набор данных входят величины $K_1^{экс}$, *T*, *RH*, [SO_2], [Cl^-] и *Prec*.

БД RUS состоит из 38 наборов данных, полученных в 32 местах испытаний. В набор данных входят величины $K_1^{экс}$, *T*, *RH*, [SO_2], [Cl^-] и *Prec*.

Для сопоставления ФДО^C и модели RF объединены БД ISO, БД MICAT, БД ECE/UN и БД RUS, в общую базу данных (БД_INT), которая включает 524 набора данных. Коды мест испытаний в соответствии с программами представлены в таблице 1.

Таблица 1. Код мест испытаний по разным программам, данные которых использованы при разработке RF модели

Программа	Код мест испытаний
ISO CORRAG [7]	ARG2, AGR5, CND1, CS1, CS2, CS3,D1, SF1, SF2,SF3, F1, F2, F3, F4, F 5, F 6, F 8, F 9, JAP1, JAP2, JAP3, N1, N2, N3, N4, N5, N6, E3, E4, S2, S3, UK1, UK2 , UK3, UK4, US1, US3, SU1, SU2, SU3, SU4
MICAT [8]	A1, A2, A3, A4, A5, A6, B1, B10, B11, B12, B2, B3, B4, B5, B6, B7, B8, B9, C H1, CH2, CH4, CO1, CO2, CO3, CR1, CR2, CR3, CR4, CU1, CU2, CU3, E1, E4, E5, E7, E8, EC1, EC2, EC3, EC5, M1, M2, M3, M4, PA1, PA2, PA3, PA4, PE2, PE3, PE4, PE5, PE6, PO1, PO2, PO3, U1, U2, U3, U4, U5, V1, V2, V3, V4, V5
ECE/UN [9]	CAN37, CS1, CS2, CS3, EST35, FIN4, FIN5, FIN6, GER10, GER11, GER12, G ER7, GER8, GER9, NL18, NL19, NL20, NOR21, NOR23, RUS34, SPA31, SPA3 3, SWE24, SWE25, SWE26, US38, US39
RUS [10, 6]	РФ1, РФ2, РФ3, РФ4, РФ5, РФ6, РФ7, РФ8, РФ9, РФ10, РФ11, РФ12, Армань, Апапельхино, Аян, Чумикан, о. Айон, Оха, Охотск, У-Хайрюзово, П-Камч, о. Байдуков, м. Шмидта, Невельск, м. Чаплина, м. Гамов, Владивосток, ДВКС, СКС, Никольское, м. Лопатка, ГЦКИ

Таблица 2. Параметры атмосферы и коррозионные потери массы, их символы, единицы измерения, интервалы среднегодовых (суммарных за год) значений для мест испытаний, включенных в БД_INT и БД_CON

Параметр	Символ	Единицы измерения	Интервал	
			БД INT	БД CON
Температура воздуха	*T*	°C	от -16,6 до +28,2	от -16.6 до +27.0
Относительная влажность воздуха	*RH*	%	от 33 до 98	от 33 до 98
Количество атмосферных осадков	*Prec*	мм/год	-	от 17 до 2624.0
Концентрация диоксида серы	[SO_2]	мкг/м^3	от 0,7 до 214,6	от 0,7 до 83.3
Скорость выпадения хлоридов	[Cl^-]	мг/(м^2·сут)	от 0,3 до 1093	-
Первогодовые коррозионные потери	$K_1^{экс}$	мкм	от 0.4 до 408,1	от 0.69 до 70.9

Для сопоставления ФДО^{H} и модели RF объединены данные, полученные в континентальных местах испытаний по проекту MICAT, программам ECE/UN и РФ, сформирована база данных БД_CON, которая включает 152 набора данных. В таблице 2 приведен интервал среднегодовых параметров атмосферы и первогодовых коррозионных поражений стали, для мест испытаний, включенных в БД INT и БД CON.

2.2. Функции доза-ответ

Для прогнозирования коррозионный потерь углеродистой стали за первый год для атмосфер, содержащих SO_2 и Cl^-, использованы два вида функций доза-ответ.

Стандартные ФДО (ФДО^{C}) [1]:

– при $T \leq 10$°C:

$$r_{corr} = 1{,}77 \cdot P_d^{0{,}52} \cdot \exp[0{,}02 \cdot RH + 0{,}150 \cdot (T - 10)] + 0{,}102 \cdot S_d^{0{,}62} \cdot \exp(0{,}033 \cdot RH + 0{,}04 \cdot T),$$

– при $T > 10$°C:

$$r_{corr} = 1{,}77 \cdot P_d^{0{,}52} \cdot \exp[0{,}02 \cdot RH - 0{,}054 \cdot (T - 10)] + 0{,}102 \cdot S_d^{0{,}62} \cdot \exp(0{,}033 \cdot RH + 0{,}04 \cdot T) \quad (1)$$

где r_{corr} (мкм) – скорость коррозии стали за первый год экспозиции; T – среднегодовая температура, °C; RH – среднегодовая относительная влажность воздуха, %; P_d и S_d - среднегодовые выпадения SO_2 и Cl^- соответственно, мг/(м^{2}сут).

Модель ФДО^{C} была разработана на основании данных программ ISO CORRAG, проекта MICAT и результатов испытаний в ряде мест на Дальнем Востоке России [14].

Новые ФДО (ФДО^{H}) [5]:

– при $T \leq 10$°C:

$$K_1 = 7{,}7 \cdot [SO_2]^{0{,}47} \cdot \exp[0{,}024 \cdot RH + 0{,}095 \cdot (T - 10) + 0{,}00035 \cdot Prec],$$

– при $T > 10$°C:

$$K_1 = 7{,}7 \cdot [SO_2]^{0{,}47} \cdot \exp[0{,}024 \cdot RH - 0{,}065 \cdot (T - 10) + 0{,}00035 \cdot Prec] \quad (2)$$

где K_1 (г/м^2) – коррозионная массопотеря стали за первый год экспозиции; $[SO_2]$ – среднегодовая концентрация SO_2 в воздухе, мкг/м^3; $Prec$ – среднегодовое количество атмосферных осадков, мм/год.

Для пересчета K_1, выраженной в г/м^2 (уравнение (2)), в мкм использована плотность углеродистой стали, равная 7,86 г/см^3. Для уравнения (1) сделан пересчет скорости осаждения SO_2 (P_d, мг/(м^{2}сут)) в концентрацию SO_2 в воздухе ($[SO_2]$, мкг/м^3) по соотношению: $P_d = 0{,}8\ [SO_2]$ [1].

2.3. Модель «случайный лес»

В машинном обучении принята следующая терминология: каждый набор данных в БД является объектом, что соответствует местам испытаний. Объект характеризуется признаками (входные данные для модели, то есть параметры атмосферы) и величиной прогноза (выходные данные, то есть, величина K_1).

Реализация алгоритма «случайный лес» проводилась при помощи библиотеки scikit-learn [17]. Обучение деревьев проводилось на основе обучающей выборки (train set), которая составляет 70% от всей базы данных. Каждое из деревьев получало на вход свою подвыборку, которая с помощью бутстрапа (bootstrap) получалась из исходной обучающей подвыборки. Бутстрап — один из популярных подходов к построению подвыборок. Он заключается в том, что из обучающей выборки длины L (длина выборки – количество принадлежащих ей объектов) выбирают с возвращением L объектов. При этом новая выборка также будет иметь длину L, но некоторые объекты в ней будут повторяться, а некоторые объекты из исходной выборки в нее не попадут.

Размер выборки был равен размеру обучающей выборки (т.е. часть данных дублировалась). Ветвление производили по случайно выбранным признакам (количество которых является гиперпараметром) до исчерпания данных. Ветвление производилось в согласии с критерием информативности (среднеквадратичной ошибкой) так, чтобы дисперсия значений в листе была минимальной.

В тестовой выборке (test set) (30% объектов БД) каждое из деревьев давало величину прогноза на основании признаков каждого объекта из этой выборки. В итоге прогнозом для объекта тестовой выборки становилось среднее значение прогноза по всем деревьям.

Значения глобальных гиперпараметров (число признаков для ветвления и число деревьев в лесу) подбирали с помощью функции GridSearchCV [18]: число деревьев в лесу от 50 до 600 с шагом 50, число признаков: от 1 до 5. Лучший набор соответствовал наименьшему значению средней абсолютной процентной ошибки MAPE по пяти тестовым выборкам.

Расчет важности признаков, показывающих влияние каждого из признаков на величину прогноза коррозионных потерь, проводился с помощью библиотеки scikit-learn [17].

2.4. Статистические критерии достоверности прогноза

Для оценки достоверности предсказаний моделей использовали статистические критерии, которые приведены ниже.

(1) Средняя абсолютная процентная ошибка (MAPE):

$$\mathrm{MAPE}(x,y) = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - y_i|}{|x_i|}100, \qquad (3)$$

где x_i и y_i – экспериментальное и прогнозное значения K_1, соответственно, N – количество объектов в БД. Чем меньше MAPE, тем меньше модель ошибается в прогнозе.

(2) Симметричная средняя абсолютная процентная ошибка (SMAPE):

$$\text{SMAPE}\,(\text{x},\text{y}) = \frac{2}{N}\sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i| + |y_i|}100 \qquad (4)$$

Преимущество SMAPE по сравнению с MAPE в том, что SMAPE учитывает возможную погрешность не только прогноза, но и экспериментального значения.

(3) Обобществленный коэффициент детерминации ($R^2_{нов}$) [19]:

$$R^2_{нов} = 1 - \frac{\sum_{i=1}^{n}(y_i - \frac{(yx)_{ср}}{(x^2)_{ср}}x_i)^2}{\sum_{i=1}^{n}(y_i - x_i)^2} \qquad (5)$$

где $(yx)_{ср} = \frac{1}{n}\sum_{i=1}^{n} y_i x_i$, $(x^2)_{ср} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

Коэффициент $R^2_{нов}$ показывает, насколько хорошо распределение точек с координатами x_i и y_i описывается функцией $y = x$. Значения $R^2_{нов}$ изменяются от 0 до 1; при $R^2_{нов} = 0$ все точки (x_i; y_i) попадают на биссектрису, то есть на прямую $y = x$. Увеличение коэффициента $R^2_{нов}$ показывает, что точки (x_i; y_i) наилучшим образом описываются прямой $y = ax$, где коэффициент $a \neq 1$.

Необходимость использования коэффициента $R_{нов}^2$ связана с тем, что стандартный коэффициент детерминации (R^2) не подходит для определения достоверности модели путем сравнения прогнозного и истинного значения [5, 19]. Коэффициент R^2 показывает, насколько хорошо линейная модель вида $y = ax + b$ описывает данные в сравнении с моделью $y = b$, но при $R^2 = 1$ условие $a = 1$ не обязано выполняться. Так, например, если прогноз будет всегда вдвое больше, чем ожидаемое значение, коэффициент R^2 будет в точности таким же, как и в случае, когда прогноз совпадает с ожидаемым значением.

(4) Процент удовлетворительных значений y (PSV):

$$\text{PSV} = (\text{M/N}) \cdot 100\% \qquad (6)$$

где M – число y_i, значения которых находятся в интервале от $0{,}67x_i$ до $1{,}5x_i$. На графике с координатами $y = K_1^{пр}$, $x = K_1^{экс}$ значения $K_1^{пр}$ должны находиться между линиями относительных ошибок $K_1^{пр}$, равных −33% и +50%, соответственно [5]. Этот интервал относительных ошибок прогноза соответствуют интервалу неопределенности расчета первогодовых коррозионных потерь стали по стандарту [1]. Чем ближе PSV к единице, тем большее число $K_1^{пр}$ лежит между линиями указанных относительных ошибок прогноза, и, следовательно, модель является более достоверной.

3 Результаты и их обсуждение

3.1 Получение моделей «случайный лес» на объединенных базах данных БД INT и БД CON

На основе объединенных баз данных в согласии со схемой, представленной в пункте 2.3, были получены две модели «случайный лес»: RF_общая и RF_континент. В дальнейшем эти модели будут применяться как к объединенным БД, так и к БД различных программ натурных испытаний.

При обучении моделей были рассчитаны значения важности признаков, которые представлены в таблице 3. Видно, что в случае RF_общая наибольшее влияние на величину коррозионных потерь оказывает скорость осаждения хлоридов, примерно одинаково влияют содержание SO_2 и температура воздуха, и меньшее значение имеет относительная влажность воздуха. В случае RF_континент величину коррозионных потерь в первую очередь определяет содержание SO_2 в воздухе, в гораздо меньшей степени влияет температура, и еще меньше – относительная влажность воздуха и количество осадков. Полученные значения важности признаков согласуются с известными представлениями о процессе атмосферной коррозии углеродистой стали [20]. Однако надо подчеркнуть, что данные значения характеризуют модель, обученную на конкретной базе данных.

Таблица 3. Важность признаков моделей, полученных на основе алгоритма «случайный лес»

Модель	Значение важности признака				
	SO_2	Cl	T	RH	Prec
RF_общая	0.187	0.402	0.238	0.173	-
RF_континент	0.617	-	0.179	0.101	0.103

3.2. Сравнение достоверности моделей RF_общая и ФДО^C

Величины $K_1^{пр}$ были рассчитаны в соответствии с моделью RF_общая и ФДО^C (уравнение (1)), используя БД INT. Сопоставляя предсказанные величины с экспериментальными значениями $K_1^{экс}$ (рисунок 1), были рассчитаны показатели достоверности этих моделей (таблица 4). Сплошная линия на рисунке 1 отвечает условию $K_1^{пр} = K_1^{экс}$.

Модель RF_общая была получена на 70% объектов БД INT, а ее оценка была проведена на оставшихся 30% объектов этой базы данных. ФДО^C была получена, используя наборы данных, большая часть которых входит в БД INT. Поэтому для корректного сравнения достоверности моделей RF и ФДО, значения $K_1^{пр}$ были рассчитаны по обеим моделям как для всей базы данных, так и на 30% БД (тестовая часть). Тестовая выборка, на которой определялись критерии достоверности моделей RF и ФДО, включала одни и те же объекты.

Как видно из таблицы 4, предсказания по модели RF_общая для всей БД являются более точными, чем для ФДО^{C}, так как коэффициенты $R^2_{нов}$, MAPE и SMAPE имеют меньшие значения, а PSV – большее. Применение обеих моделей к тестовым выборкам данных показывает, что RF_общая также имеет лучшие значения всех статистических критериев (таблица 4).

Модель RF_общая дает значения PSV = 90 и 75 % для всей базы данных и тестовой выборки, соответственно. Это означает, большая часть значений $K_1^{пр}$ лежит в интервале от $0{,}67K_1^{экс}$ до $1{,}5K_1^{экс}$. Функция «доза- ответ» может предсказать не более 60 % значений $K_1^{пр}$, которые попадают в интервал допустимых ошибок прогноза [1].

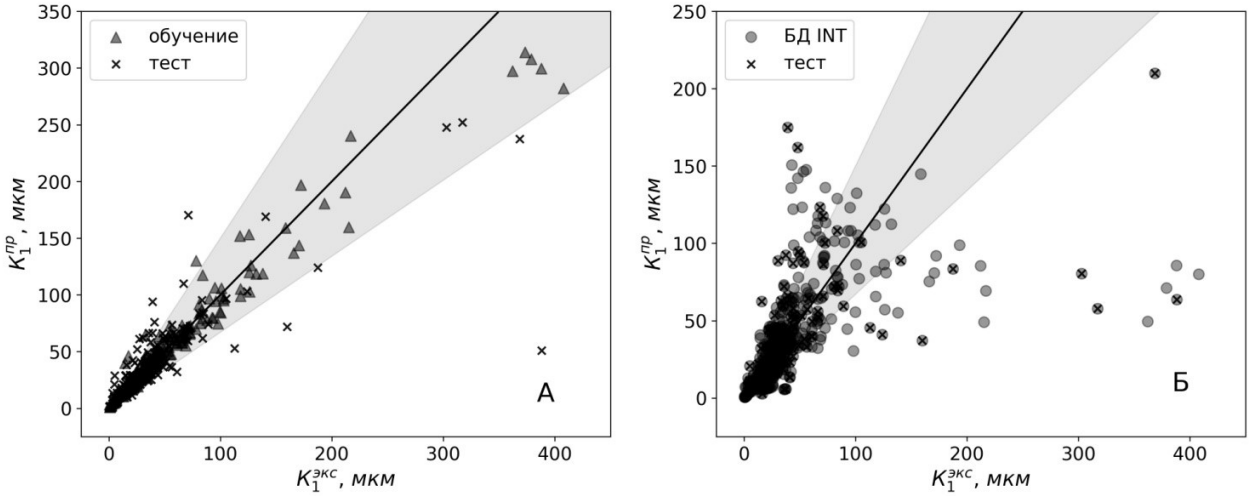

Рисунок 1. БД INT. Соответствие между экспериментальными и предсказанными значениями K_1: по RF_общая (А) и ФДО^{C} (Б). Линия соответствует $K_1^{пр} = K_1^{экс}$. Выделенная область показывает относительную ошибку предсказаний в интервале от −33% до + 50%.

Таблица 4. Значения статистических критериев достоверности модели RF_общая и ФДО^{C}, полученных на БД INT

Модель	Число объектов	Характер выборки	$R^2_{нов}$	MAPE, %	SMAPE, %	PSV, %
RF_общая	524	БД INT	0.27	22	18	90
	366	Обучение	0.30	15	13	96
	158	Тест	0.36	39	29	75
ФДО^{C}	524	БД INT	0.43	44	40	60
	158	Тест	0.50	44	40	60

3.3. Сравнение достоверности моделей RF_континент и ФДО^{H}

Сопоставление рассчитанных по модели «случайный лес» и функции «доза-ответ» значений первогодовых коррозионных потерь стали в континентальных местах испытаний (БД CON) с соответствующими экспериментальными величинами $K_1^{экс}$ показано на рисунке 2. Значения статистических критериев достоверности моделей

RF_континент и ФДО^{Н} были рассчитаны как для всей БД CON, так и для 30% тестовой выборки объектов этой базы данных (таблица 4). Как видно, в обоих случаях модель RF_континент дает более точный прогноз $K_1^{пр}$: меньшие значения $R^2_{нов}$, MAPE, SMAPE и большее значение PSV. При использовании RF_континент величина $R^2_{нов}$ близка к нулю, то есть, прогноз $K_1^{пр}$ дает наиболее симметричный разброс точек вокруг линии $K_1^{пр} = K_1^{экс}$ как при малых, так и при больших величинах $K_1^{пр}$. Это означает, что RF_континент наилучшим образом предсказывает коррозионные потери углеродистой стали, если рассматривать весь диапазон полученных экспериментальных данных (рисунок 2А).

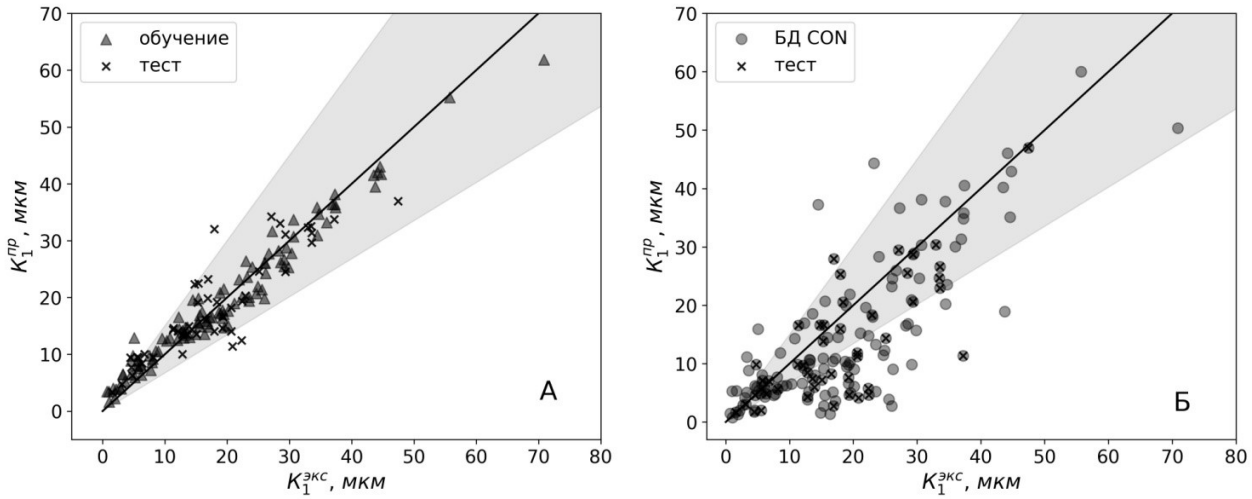

Рисунок 2. БД CON. Соответствие между экспериментальными и предсказанными значениями K_1 по RF_континент (А) и ФДО^{Н} (Б).

Таблица 5. БД CON. Значения статистических критериев достоверности модели RF_континент и ФДО^{Н}

Модель	Число точек	Характер выборки	$R^2_{нов}$	MAPE, %	SMAPE, %	PSV, %
RF_континент	152	БД CON	0.06	27	21	86
	106	Обучение	0.114	26	19	89
	46	Тест	0.026	29	25	78
ФДО^{Н}	152	БД CON	0.32	44	49	55
	46	Тест	0.40	37	47	57

Надо отметить, что ФДО^{Н} была разработана на основе данных, полученных в континентальных местах испытаний программ ECE/UN и РФ (всего 89 мест) и на этой базе данных показала большую достоверность предсказаний коррозионных потерь стали, чем ФДО^{С} [5]. Так, значения MAPE для ФДО^{Н} и ФДО^{С} равны 23% и 33,6 %, соответственно [5]. Включение в БД CON данных мест испытаний по проекту MICAT существенно увеличило MAPE предсказаний $K_1^{пр}$ по ФДО^{Н} (таблица 5), что возможно

связано как с расширением интервала значений параметров атмосферы объектов базы данных, так и с отдельными ошибками при определении $K_1^{экс}$ [3].

3.4. Оценка достоверности моделей «случайный лес» и ФДО на базах данных различных программ испытаний

Программы натурных испытаний типовых металлов ISO CORRAG, MICAT, ECE/UN и РФ, результаты которых были объединены в БД INT и БД CON, были проведены в разные годы в различных климатических регионах мира. Подход к выбору мест испытаний также был различным. Например, места испытаний в программе ECE/UN континентальные, которые вошли в соответствующую БД, а в проекте MICAT большинство мест испытаний было с приморской атмосферой. Естественно, что БД отдельных испытательных программ существенно отличаются, и достоверность предсказаний $K_1^{пр}$ может быть различна. Необходимо проверить достоверность моделей RF_общая и RF_континент в случае их применения к БД различных испытательных программ. Как и в случае объединенных баз данных (п.3.2 и 3.3), достоверность моделей RF сравнивали с точностью предсказаний $K_1^{пр}$, полученных при использовании ФДО^{С} и ФДО^{Н}. ФДО^{С} применяли для мест испытаний с любым типом атмосферы, ФДО^{Н} – только для континентальных мест.

БД ISO. В эту БД вошли места испытаний с любым типом атмосферы, но годовое количество осадков в них не измерялось. Для предсказаний K_1 использовали RF_общая и ФДО^{С}, статистические критерии достоверности моделей даны в таблице 6. Визуализация соответствия фактических и прогнозных значений K_1 представлена на рисунке 3. Видим, что ошибки MAPE и SMAPE для модели RF_общая примерно в 3 раза меньше, чем для ФДО^{С}. Более 90% предсказанных моделью RF значений $K_1^{пр}$ укладываются в допустимый интервал ошибок (PSV = 94 %, таблица 6). Величины $R^2_{нов}$ для обеих моделей примерно одинаковы (0.03 и 0.01 для модели RF_общая и ФДО^{С}, соответственно).

БД MICAT. Статистические критерии были рассчитаны для всей базы данных проекта MICAT (175 объектов) по моделям RF_общая и ФДО^{С} и только для континентальных мест испытаний (63 объекта) по моделям RF_континент и ФДО^{Н} (таблица 6). Визуализация соответствия фактических и прогнозных значений K_1 представлена на рисунке 4. Все критерии достоверности модели RF_общая для полной БД MICAT хуже, чем для БД ISO (таблица 6). Однако, модель RF_общая дает более точные предсказания K_1 в местах испытаний проекта MICAT, чем ФДО^{С}: ошибки MAPE и SMAPE примерно в 2 раза меньше, а коэффициент PSV – больше. Для континентальной выборки БД MICAT статистические критерии достоверности модели RF_континент также значительно лучше, чем для ФДО^{Н} (таблица 6). Как отмечалось выше, при разработке ФДО^{Н} не учитывались данные, полученные по проекту MICAT, поэтому модель дает значительную ошибку прогноза на этой БД.

БД ECE/UN. Поскольку в эту базу данных входят только данные, полученные в континентальных местах испытаний, то статистические критерии рассчитывались для RF_континент и ФДО^{Н}. Как видно на рисунке 5, модель RF_континент более точно предсказывает коррозионные потери стали, чем функция доза-ответ. Предсказанные моделью RF_континент значения лежат вблизи линии $K_1^{пр} = K_1^{экс}$ (рисунок 5А), что количественно выражается величиной $R^2_{нов}$, близкой к нулю, при этом 96% значений $K_1^{пр}$ попадают в интервал допустимых ошибок (таблица 6). Ошибка предсказаний K_1 по ФДО^{Н} заметно больше (рисунок 5Б и таблица 6).

БД RUS. Статистические критерии достоверности всех моделей были рассчитаны для полной базы данных (33 объекта) и континентальной выборки (12 объектов). Визуализация соответствия фактических и прогнозных значений K_1 представлена на рисунке 6. Для полной БД RUS точность прогноза K_1 достаточно низкая как при использовании RF_общая, так и ФДО^{С} (рисунок 6А, Б). Результат прогноза по модели RF_общая характеризуют большие значения MAPE и SMAPE, чем для ФДО^{С} (таблица 6). Число предсказанных K_1, которые имеют допустимую ошибку для обеих моделей отличается незначительно: коэффициент PSV = 68% и 76 % для RF_общая и ФДО^{С}, соответственно.

Для выборки континентальных объектов БД RUS предсказания K_1 по ФДО^{Н} точнее, чем по модели RF_континент (рисунок 6В, Г), что отражается в показателях достоверности моделей (Таблица 6). Так, число предсказанных по ФДО^{Н} значений K_1, которые имеют допустимую ошибку, равно 92%, что является лучшим результатом для функций «доза-ответ» для всех рассмотренных баз данных.

Надо отметить, что континентальная выборка БД RUS состоит из данных, полученных в холодном климате (интервал T от +3,6 до −16,6ºС) с небольшим количеством атмосферных осадков (Prec≤626 мм/год) и с низким содержанием агрессивных примесей в воздухе ([SO_2]≤10 мкг/м^3) [21]. Очевидно, что универсальные модели RF, разработанные на объединенных БД и предназначенные для предсказания K_1 в различных регионах мира, могут быть недостаточно точными при прогнозировании коррозионных потерь металла в конкретном климатическом регионе из-за малого числа объектов с экстремальными значениями признаков в объединенной базе данных.

Таблица 6. Значения статистических критериев достоверности моделей, примененных к различным базам данных

БД	Модель	Число точек	Признаки	$R^2_{нов}$	MAPE, %	SMAPE, %	PSV, %
ISO	RF_общая	234	T, RH, SO_2, Cl	0.03	15	13	96
	ФДО^C	234	T, RH, SO_2, Cl	0.01	42	37	60
MICAT	RF_общая	175	T, RH, SO_2, Cl	0.40	29	23	85
	ФДО^C	175	T, RH, SO_2, Cl	0.84	54	50	49
	RF_континент	63	T, RH, SO_2, Prec	0.12	33	26	81
MICAT	ФДО^H	63	T, RH, SO_2, Prec	0.51	69	75	32
ECE/UN	RF_континент	77	T, RH, SO_2, Prec	0.05	14	13	96
	ФДО^H	77	T, RH, SO_2, Prec	0.30	27	32	68
RUS	RF_общая	38	T, RH, SO_2, Cl	0.001	48	32	68
	ФДО^C	38	T, RH, SO_2, Cl	0.73	25	27	76
	RF_континент	12	T, RH, SO_2, Prec	0.42	75	43	42
	ФДО^H	12	T, RH, SO_2, Prec	0.03	22	19	92

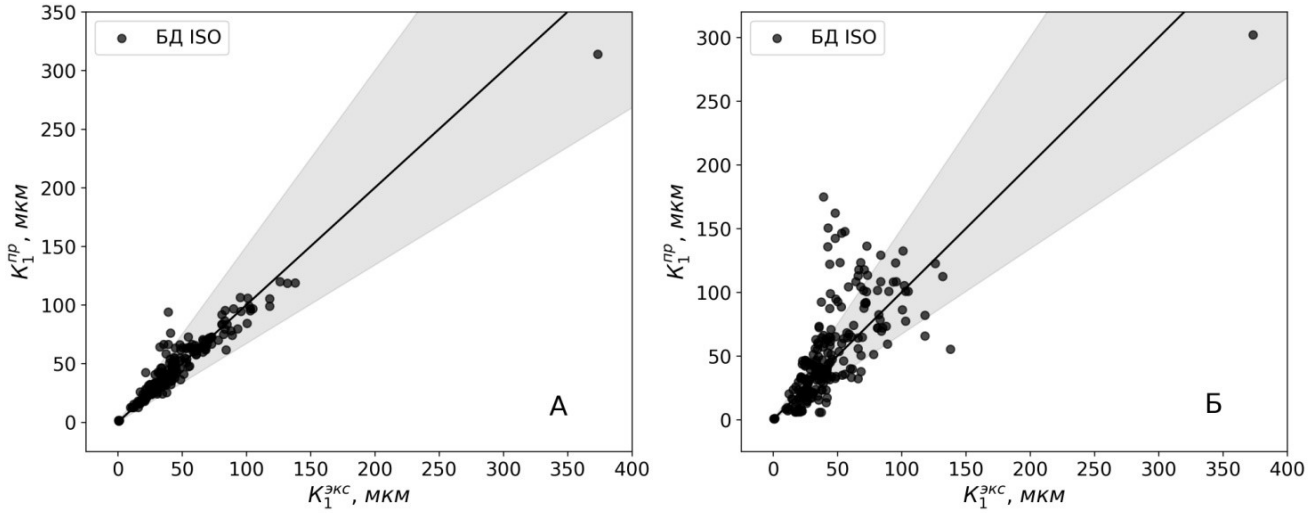

Рисунок 3. БД ISO. Соответствие между экспериментальными и предсказанными значениями K_1 по RF_общая (А) и ФДО^C (Б)

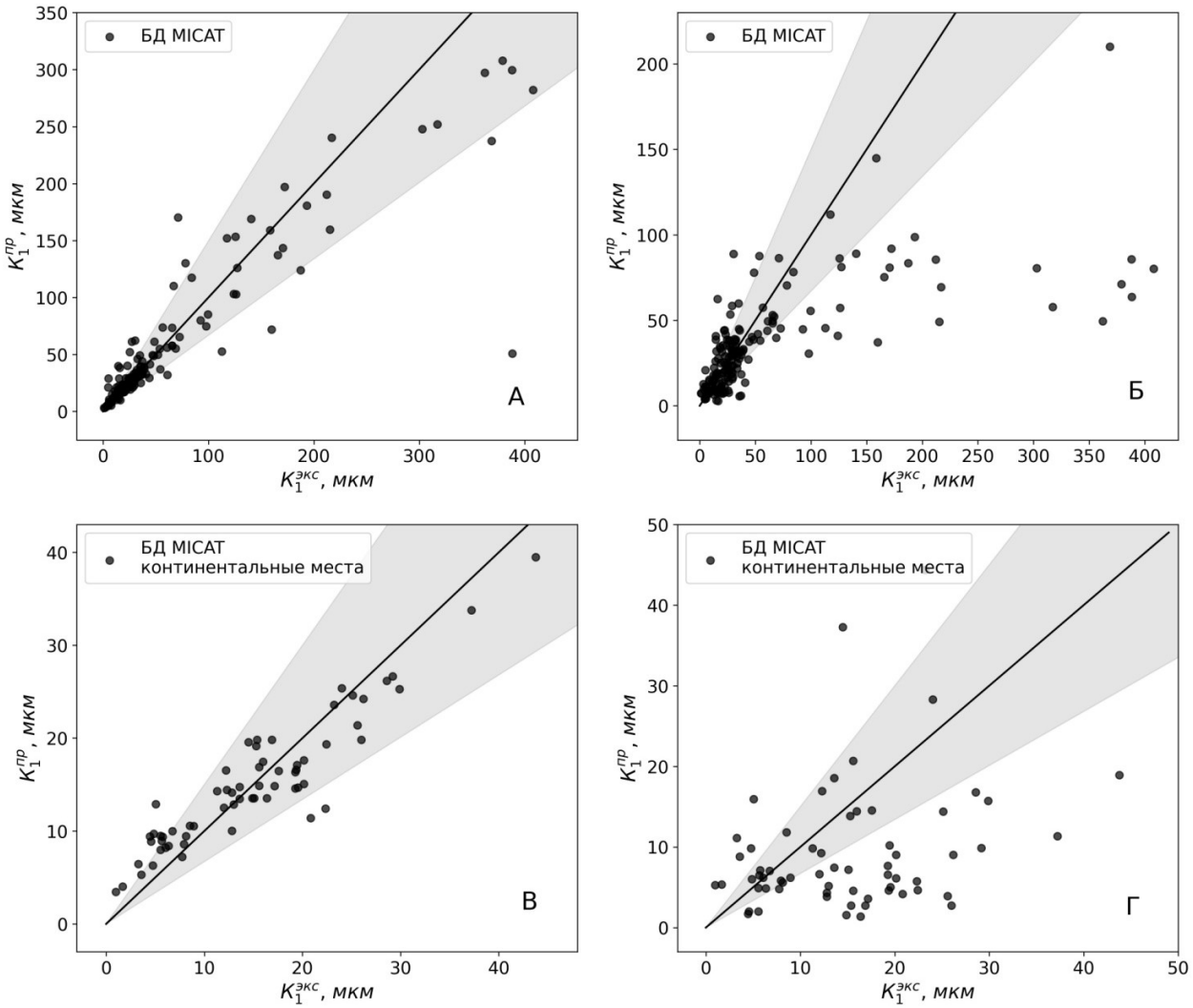

Рисунок 4. БД MICAT. Соответствие между экспериментальными и предсказанными значениями K_1 по RF_общая (А), ФДО^{С} (Б), RF_континент (В), ФДО^{Н} (Г)

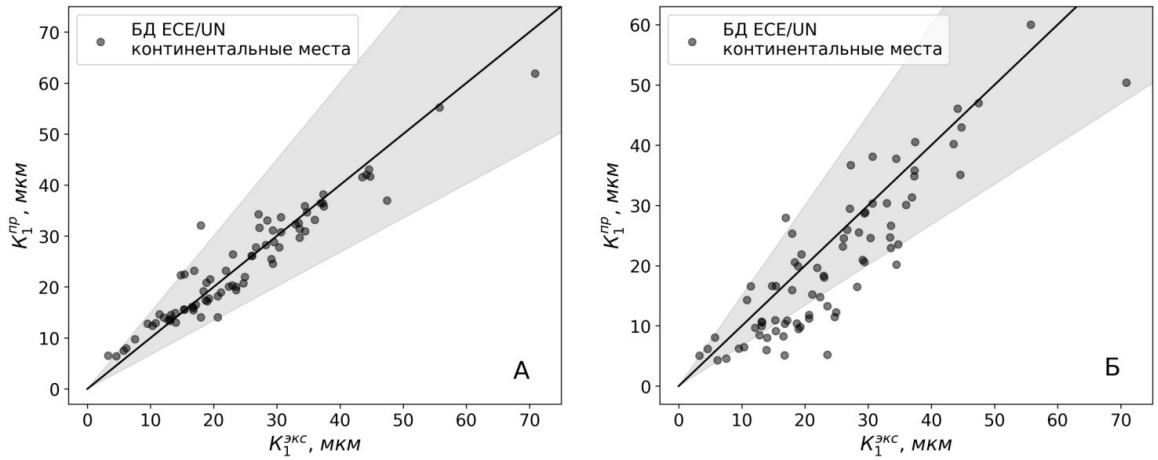

Рисунок 5. БД ECE/UN. Соответствие между измеренными и предсказанными значениями K_1 по RF_континент (А), ФДО^{Н} (Б)

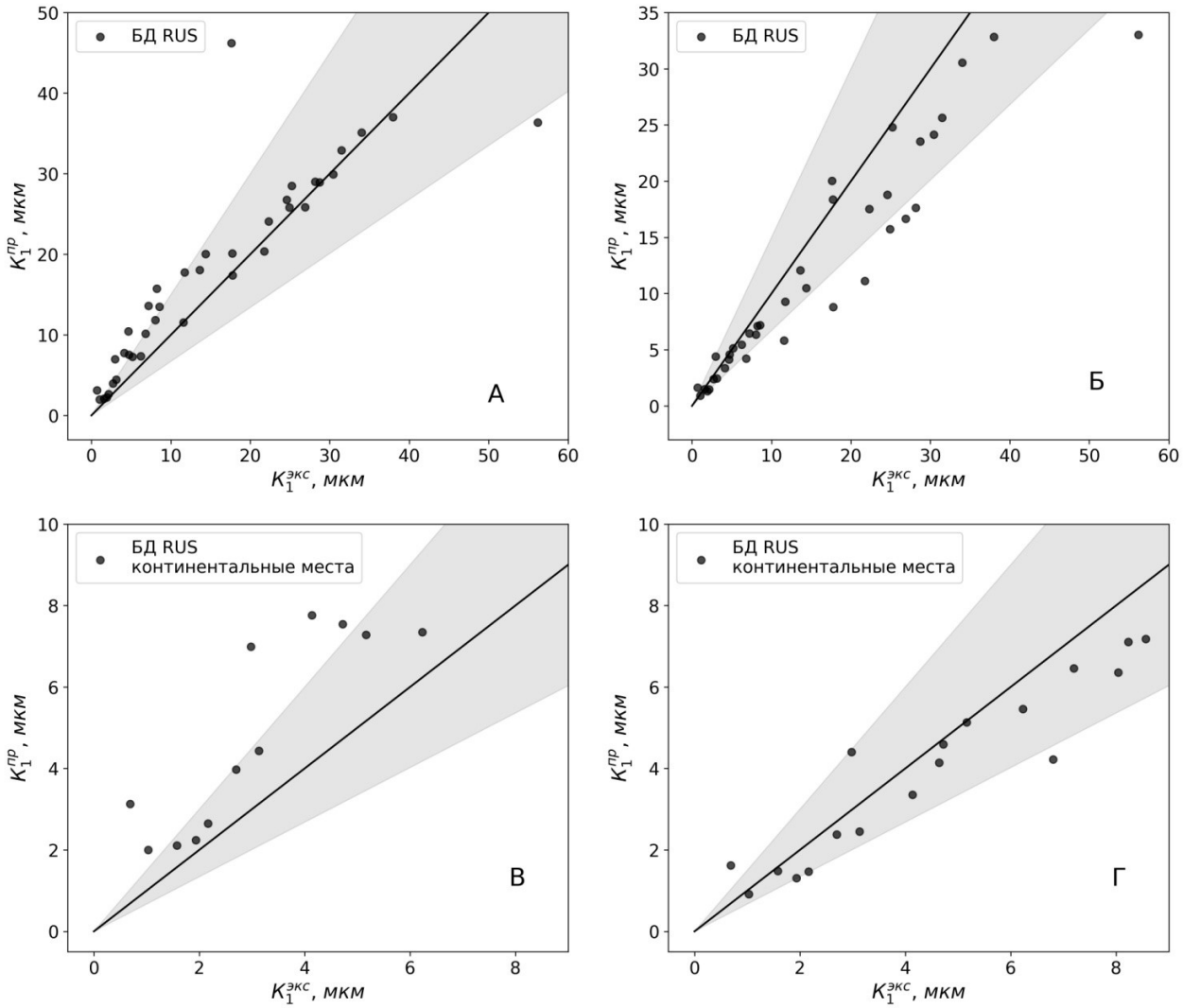

Рисунок 6. БД RUS. Соответствие между измеренными и предсказанными значениями по RF_общая (А), ФДО^{С} (Б), RF_континент (В), ФДО^{Н} (Г)

Заключение

1. С помощью алгоритма «случайный лес» получены две модели RF для предсказаний первогодовых коррозионных потерь (K_1) углеродистой стали в открытой атмосфере в различных регионах мира. Модель RF_общая получена на объединенной базе данных, которая включает данные программ ISO CORRAG, MICAT, ECE/UN и RUS, и предсказывает величины K_1 по значениям T, RH, [SO_2], [Cl]. Модель RF_континент получена на объединенной базе данных, которая включает данные, полученные в континентальных местах испытаний программ MICAT, ECE/UN и RUS, и предсказывает величины K_1 по значениям T, RH, $Prec$, [SO_2].
2. Достоверность предсказаний моделей RF оценивалась по совокупности статистических критериев: обобщенному коэффициенту детерминации $R^2_{нов}$, MAPE, симметричной средней относительной ошибке SMAPE и коэффициенту PSV.

Коэффициент PSV был предложен в этой работе и показывает долю предсказаний K_1, относительная ошибка которых не выходит за интервал ошибок, допустимый в соответствии со стандартом [1].

3. Проведено сравнение точности предсказаний K_1 по моделям RF и двум функциям «доза-ответ» (ФДО): ФДО стандарта [1] для всех типов атмосферы и новой версии ФДО [5] для не морской) атмосферы. Показано, что для объединенных баз данных достоверность обоих моделей RF лучше, чем ФДО.
4. Сопоставление точности прогнозов коррозионных потерь по моделям RF и ФДО, проведенное на базах данных отдельных испытательных программ, подтвердило, что модели RF дают более точный прогноз, за исключением предсказаний K_1 в местах испытаний по российской программе. Это исключение связано с недостаточным количеством наборов данных в местах испытаний с холодным климатом, включенных в объединенные базы данных.

Список литературы

1. *ISO 9223:2012(E). Corrosion of metals and alloys — Corrosivity of atmospheres — Classification, determination and estimation*, International Standards Organization, Geneve, 2012.
2. *ISO 9224:2012(E) Corrosion of metals and alloys —Corrosivity of atmospheres — Guiding values for the corrosivity categories*, 2012.
3. Yu.M. Panchenko, A.I. Marshakov, Prediction of First-Year Corrosion Losses of Carbon Steel and Zinc in Continental Regions, *Materials*, 2017, **10**, no. 4. 422. doi: 10.3390/ma10040422
4. Yu.M. Panchenko, A.I. Marshakov, L.A. Nikolaeva, VV. Kovtanyuk, Prediction of first-year corrosion losses of copper and aluminum in continental regions, *AIMS Materials Science*, 2018, **5**, no. 4, 624−649.doi: 10.3934/matersci.2018.4.624
5. Yu.M. Panchenko, A.I. Marshakov, L.A. Nikolaeva, T.N. Igonin, Evaluating the Reliability of Predictions of First-Year Corrosion Losses of Structural Metals Calculated Using Dose-Response Functions for Territories with Different Categories of Atmospheric Corrosion Aggressiveness, *Protection of Metals and Physical Chemistry of Surfaces*, 2020, **56**, no. 7. 1249–1263. doi: 10.1134/S207020512007014X
6. Yu.M. Panchenko, A.I. Marshakov, L.A. Nikolaeva, T.N. Igonin, Development of models for the prediction of first-year corrosion losses of standard metals for territories with coastal atmosphere in various climatic regions of the world, *Corrosion Engineering Science and Technology*, 2020, **55**, no. 8. 655−669. doi: 10.1080/1478422X.2020.1772535 https://doi.org/10.1080/1478422X.2020.1772535
7. D. Knotkova, P. Boschek, K. Kreislova, Results of ISO CORRAG Program: Processing of One Year Data in Respect to Corrosivity Classification, *ASTM Special Technical Publication*, West Conshohocken, PA, 1995, 38.

8. M. Morcillo, Atmospheric corrosion in Ibero-America. The MICAT project. In Atmospheric Corrosion, *ASTM Special Technical Publication*, Philadelphia, PA, USA, 1995, 257–275.
9. J. Tidblad, A.A. Mikhailov, V. Kucera, Acid Deposition Effects on Materials in Subtropical and Tropical Climates. Data Compilation and Temperate Climate Comparison, SCI Report 2000:8E, Swedish Corrosion Institute, Stockholm, Sweden, 2000, 1–34.
10. Ю.М. Панченко, Л.Н. Шувахина, Ю.Н. Михайловский, Атмосферная коррозия металлов в регионах Дальнего Востока, *Защита металлов*, 1982. **18**, № 4, 575–582.
11. L. Breiman, Random forests, *Machine Learning*, 2001, **45**, 5−32. doi: 10.1023/A:1010933404324
12. Y. Zhi, D. Fu, D. Zhang, T. Yang, X. Li, Prediction and Knowledge Mining of Outdoor Atmospheric Corrosion Rates of Low Alloy Steels Based on the Random Forests Approach, *Metals*, 2019, **9**. no. 3, 383. doi: 10.3390/met9030383
13. L. Yan, Y. Diao, K. Gao, Analysis of environmental factors affecting the atmospheric corrosion rate of low-alloy steel using random forest-based models, *Materials*, 2020, **13**, no. 15, 3266. doi: 10.3390/ma13153266
14. Y. Zhi, Z. Jin, L. Lu, T. Yang, D. Zhou, Z. Pei, D. Wu, D. Fu, D. Zhang, X. Li, Improving atmospheric corrosion prediction through key environmental factor identification by random forest-based model, *Corrosion Science*, 2021, **178**, no. 109084. doi: 10.1016/j.corsci.2020.109084
15. A.A. Mikhailov, J. Tidblad, V. Kucera, The classification system of ISO 9223 standard and the dose-response functions assessing the corrosivity of outdoor atmospheres, *Protection of Metals*, 2004, **40**, no. 6, 541−550. doi: 10.1023/B:PROM.0000049517.14101.68
16. J. Tidblad, V. Kucera, A.A. Mikhailov, D. Knotkova, Outdoor and Indoor Atmospheric Corrosion, *ASTM Special Technical Publication*, West Conshohocken, PA, USA, 2002, 73.
17. Scikit-learn, Machine Learning in Python, https://scikit-learn.org/stable/index.html
18. Scikit-learn, Sklearn.model_selection.GridSearchCV, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
19. Yu.M. Panchenko, A.I. Marshakov, I.V. Bardin, A.V. Shklyaev, Use of Statistical Analysis Methods for Estimating the Reliability of First-Year Carbon Steel and Zinc Corrosion Loss Predictions Calculated Using Dose-Response Functions, *Protection of Metals and Physical Chemistry of Surfaces*, 2019, **55**, no. 4, 753–760. doi: 10.1134/S2070205119040142
20. C. Leygraf, I.O. Wallinder, J. Tidblad, T. Graedel, Atmosheric Corrosion. Published by John Wiley & Sons, Inc., Hoboken, New Jersey, 2016, 397 p.
21. Yu. Panchenko, A. Marshakov, T. Igonin, L. Nikolaeva, V. Kovtanyuk, Corrosivity of atmosphere toward structural metals and mapping the continental Russian territory,

Corrosion Engineering, Science and Technology, 2019, **54**, no. 5, 369–378. doi: 10.1080/1478422X.2019.1594526

A MODEL FOR PREDICTING CORROSION LOSSES OF CARBON STEEL FOR THE FIRST YEAR OF EXPOSURE BASED ON THE RANDOM FOREST ALGORITHM

M.A. Gavryushina*, Yu.M. Panchenko and A.I. Marshakov

Frumkin Institute of Physical Chemistry and Electrochemistry of the Russian Academy of Sciences, 31 Leninsky Prospekt, 4, Moscow, 119071
**E-mail: maleeva.corlab@yandex.ru*

Abstract

Based on the random forest (RF) algorithm, two models have been obtained for predicting first-year corrosion losses (K_1) of carbon steel in an open atmosphere in various regions of the world. The first RF_general model was obtained using the combined databases of the international ISO CORRAG, MICAT, ECE/UN programs and tests in Russia and is designed to assess K_1 in various types of atmosphere in different regions of the world. The second RF_cont model allows you to predict K_1 in the continental regions of the world. The accuracy of K_1 predictions based on RF models and two dose-response functions was compared: the one presented in ISO 9223 standard and the new version developed by IPCE RAS for continental regions. It is shown that the reliability of both RF models is significantly better than the dose-response functions, with the exception of predictions of corrosion losses of steel in Russia with a cold climate.

Keywords*: low-carbon steel, models, probable corrosion rate, machine learning methods.*