

## Большие данные как инструмент для построения прогнозной модели износа валков прокатных станов

Н.В. ВАСИЛЬЕВА<sup>1</sup> ✉, Э.Р. ФЕДОРОВА<sup>1</sup>, А.С. КОЛЕСНИКОВ<sup>2</sup>

<sup>1</sup> Санкт-Петербургский горный университет, Санкт-Петербург, Россия

<sup>2</sup> Некоммерческое акционерное общество «Южно-Казахстанский университет им. М.Ауэзова», Шымкент, Казахстан

✉ vasilyeva\_nv@pers.spmi.ru; ☎ +7-921-441-51-78

**Ссылка для цитирования оригинальной статьи:** Vasilyeva, N.; Fedorova, E.; Kolesnikov, A. Big Data as a Tool for Building a Predictive Model of Mill Roll Wear. *Symmetry* 2021, 13, 859. <https://doi.org/10.3390/sym13050859>

**Аннотация:** Анализ больших данных становится повседневной задачей как для компаний по всему миру, так и для российских компаний. Благодаря технологическому прогрессу и снижению стоимости хранения данных, компании сегодня могут собирать и хранить большие объемы разнородных данных. Важный этап извлечения знаний и практической пользы из таких данных – задача, с которой в конечном итоге столкнутся все компании, стремящиеся сохранить свою конкурентоспособность и место на рынке. В статье рассматривается подход к изучению металлургических процессов с использованием анализа большого массива данных оперативного контроля. На примере сталепрокатного производства рассматривается разработка прогнозной модели на основе обработки большого массива данных оперативного контроля. Целью работы является разработка прогностической модели износа валков прокатного стана на основе большого массива данных оперативного контроля, содержащего информацию о времени заполнения и выгрузки валков, сортаменте проката, материале валков, времени нахождения валка в работе. Была проведена предварительная подготовка данных для моделирования, включающая удаление выбросов, нехарактерных и случайных результатов измерений (промахов), а также пробелов в данных. Корреляционный анализ данных показал, что наибольшее влияние на износ валков прокатного стана оказывают размеры и марки прокатываемых стальных листов, а также материал, из которого изготовлены валки. На основе обработки большого массива данных оперативного контроля были построены различные прогнозные модели технологического процесса. Адекватность моделей оценивалась по величине средней квадратической ошибки (СКО), коэффициенту детерминации ( $R^2$ ) и значению коэффициента корреляции Пирсона ( $R$ ) между расчетными и экспериментальными значениями износа валков прокатного стана. Кроме того, адекватность моделей оценивалась по симметрии значений, предсказанных моделью, относительно прямой  $\hat{Y}_{predicted} = Y_{actual}$ . Линейные модели, построенные с использованием метода наименьших квадратов и перекрестной валидации, оказались неадекватными (коэффициент детерминации  $R^2$  не превышает 0,3) объекту исследования. На основе той же базы данных оперативного контроля были построены следующие регрессии: многомерная линейная регрессия, многомерная модель Лассо, многомерная модель Риджа и многомерная модель ElasticNet. Однако и эти модели оказались неадекватными объекту исследования. Проверка этих моделей на симметричность показала, что во всех случаях имеет место занижение прогнозируемых значений. Также были построены модели, использующие композицию алгоритмов. Рассмотрены методы случайного леса и градиентного бустинга. Оба метода были признаны адекватными объекту исследования (для модели случайного леса коэффициент детерминации составил  $R^2 = 0,798$ , для модели градиентного бустинга –  $R^2 = 0,847$ ). Однако алгоритм градиентного усиления признан более предпочтительным благодаря высокой точности по сравнению с алгоритмом случайного леса. Контрольные данные на симметричность относительно прямой  $\hat{Y}_{predicted} = Y_{actual}$  показали, что в случае разработки модели случайного леса наблюдается тенденция к занижению прогнозируемых значений (расчетные значения располагаются ниже прямой линии). В случае разработки модели градиентного бустинга прогнозируемые значения располагаются симметрично относительно прямой  $\hat{Y}_{predicted} = Y_{actual}$ . Поэтому модель градиентного усиления предпочтительнее. Прогнозная модель износа валков мельницы позволит рационально использовать валки с точки зрения минимизации общего износа валков. Таким образом, предложенная модель позволит перераспределить имеющиеся рабочие валки между клетями с целью снижения общего износа валков.

**Ключевые слова:** износ валков прокатного стана; математическая модель; коэффициент корреляции; большие данные; прокатный стан; стальной прокат.

## Big Data as a Tool for Building a Predictive Model of Mill Roll Wear

Natalia V. VASILYEVA<sup>1</sup> ✉, Elmira R. FEDOROVA<sup>1</sup> and Alexandr S. KOLESNIKOV<sup>2</sup>

<sup>1</sup> Saint Petersburg Mining University, Saint Petersburg, Russia

<sup>2</sup> Non-Profit Joint Stock Company «M. Auezov South Kazakhstan University», Shymkent, Kazakhstan

✉ vasilyeva\_nv@pers.spmi.ru; ☎ +7-921-441-51-78

**Abstract:** Big data analysis is becoming a daily task for companies all over the world as well as for Russian companies. With advances in technology and reduced storage costs, companies today can collect and store large amounts of heterogeneous data. The important step of extracting knowledge and value from such data is a challenge that will ultimately be faced by all

\* Перевод подготовлен для повышения интереса русскоязычной аудитории к данной статье.

© 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

© Санкт-Петербургский горный университет императрицы Екатерины II (перевод), 2024.

companies seeking to maintain their competitiveness and place in the market. An approach to the study of metallurgical processes using the analysis of a large array of operational control data is considered. Using the example of steel rolling production, the development of a predictive model based on processing a large array of operational control data is considered. The aim of the work is to develop a predictive model of rolling mill roll wear based on a large array of operational control data containing information about the time of filling and unloading of rolls, rolled assortment, roll material, and time during which the roll is in operation. Preliminary preparation of data for modeling was carried out, which includes the removal of outliers, uncharacteristic and random measurement results (misses), as well as data gaps. Correlation analysis of the data showed that the dimensions and grades of rolled steel sheets, as well as the material from which the rolls are made, have the greatest influence on the wear of rolling mill rolls. Based on the processing of a large array of operational control data, various predictive models of the technological process were designed. The adequacy of the models was assessed by the value of the mean square error (MSE), the coefficient of determination ( $R^2$ ), and the value of the Pearson correlation coefficient ( $R$ ) between the calculated and experimental values of the mill roll wear. In addition, the adequacy of the models was assessed by the symmetry of the values predicted by the model relative to the straight line  $Y_{\text{predicted}} = Y_{\text{actual}}$ . Linear models constructed using the least squares method and cross-validation turned out to be inadequate (the coefficient of determination  $R^2$  does not exceed 0.3) to the research object. The following regressions were built on the basis of the same operational control database: Linear Regression multivariate, Lasso multivariate, Ridge multivariate, and ElasticNet multivariate. However, these models also turned out to be inadequate to the object of the research. Testing these models for symmetry showed that, in all cases, there is an underestimation of the predicted values. Models using algorithm composition have also been built. The methods of random forest and gradient boosting are considered. Both methods were found to be adequate for the object of the research (for the random forest model, the coefficient of determination is  $R^2 = 0.798$ ; for the gradient boosting model, the coefficient of determination is  $R^2 = 0.847$ ). However, the gradient boosting algorithm is recognized as preferable thanks to its high accuracy compared with the random forest algorithm. Control data for symmetry in reference to the straight line  $Y_{\text{predicted}} = Y_{\text{actual}}$  showed that, in the case of developing the random forest model, there is a tendency to underestimate the predicted values (the calculated values are located below the straight line). In the case of developing a gradient boosting model, the predicted values are located symmetrically regarding the straight line  $Y_{\text{predicted}} = Y_{\text{actual}}$ . Therefore, the gradient boosting model is preferred. The predictive model of mill roll wear will allow rational use of rolls in terms of minimizing overall roll wear. Thus, the proposed model will make it possible to redistribute the existing work rolls between the stands in order to reduce the total wear of the rolls.

**Keywords:** big data; rolling mill; rolled steel; rolling mill roll wear; mathematical model; correlation coefficient.

## 1. Введение

Металлургическая промышленность является одной из ведущих отраслей российской экономики. Продукция, выпускаемая этой отраслью, используется в строительстве, машиностроении, химической промышленности и многих других отраслях [1-3].

Производство металлопроката – одна из важнейших статей российского экспорта. Деформируя металл в пространстве между вращающимися валками, можно получить практически любой вид металлопродукции из стали и других сплавов. Этот процесс называется металлопрокатом. Одной из основных проблем проката является износ валков, деформирующих металл.

В данной работе под износом понимаются качественные и количественные изменения поверхности валков, вызванные физико-химическими процессами, а также механическими воздействиями одного тела на другое [4-6].

Современные тенденции развития металлургии характеризуются разработкой и внедрением информационных систем и технологий, в основе которых лежат компьютеры и компьютерные сети с богатейшим программным обеспечением, а также системы управления базами данных и компьютерные системы поддержки принятия решений, методологической основой которых является теория систем и системный анализ.

Научно-технический прогресс создает предпосылки для повышения качества управления за счет использования вычислительной техники, математических методов обработки данных, теории управления и автоматизации управления. Все это нашло конкретное воплощение в автоматизированных системах управления. Благодаря развитию информационных технологий (ИТ) появились современные программные продукты и системы управления базами данных (СУБД) для решения задач управления производством. Современная программная и микропроцессорная техника позволяет создавать высокоуровневые системы управления с включением мощных алгоритмов управления.

Актуальность работы обусловлена тем, что построение линейных и многомерных регрессионных моделей на основе большого массива данных не дает качественного результата, так как не позволяет учесть сложные и многосвязные зависимости между входными переменными. В этом случае лучше всего проявляют себя композиционные модели, устойчивые к переобучению, шуму и выбросам. Однако при меньшем количестве данных, которые можно описать простой моделью, целесообразнее использовать многомерную регрессию.

Целью работы является разработка прогностической модели износа валков прокатного стана на основе большого массива данных оперативного контроля, содержащих информацию о времени заполнения и выгрузки валков, сортаменте проката, материале валков, времени нахождения валков в работе.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Подготовить данные для моделирования (отфильтровать и агрегировать данные).
2. Провести корреляционный анализ данных для выявления факторов, оказывающих наибольшее влияние на износ валков прокатного стана.
3. Построить различные модели для прогнозирования износа валков мельницы (линейные модели, многомерные модели и интеллектуальные модели). Проверить их адекватность и определить наиболее точную.

Прогнозная модель износа валков мельницы позволит рационально использовать валки с точки зрения минимизации общего износа валков. Таким образом, предложенная модель позволит перераспределить имеющиеся рабочие валки между клетями с целью снижения общего износа валков.

## 2. Теоретическая основа

В технической литературе данные о долговечности и износе прокатных валков встречаются крайне редко. Величина и характер износа рабочих валков зависят от многих факторов. Основными из них являются: усилие, температурный и скоростной режимы прокатки, свойства и количество прокатываемого металла, твердость и диаметр валков. Однако изучить индивидуальное влияние каждого фактора на износ валков крайне сложно [7, 8].

Наличие большого количества факторов затрудняет получение зависимостей, которые бы учитывали их и позволяли рассчитать износ валков.

Исходя из обзора литературы, износ связывают с количеством (длиной) прокатываемых полос и описывают эту зависимость с помощью эмпирических уравнений, коэффициенты которых определяются экспериментально на каждом прокатном стане. Основным недостатком данных зависимостей является то, что они учитывают влияние небольшого количества факторов и не могут быть использованы при изменении условий прокатки.

Существующие теоретические методы основаны на определении пути трения в очаге деформации и контактных напряжений или на расчете работы деформации. Они достаточно сложны и длительны, и часто дают большую погрешность [9].

Поэтому для оценки износа прокатных валков удобнее использовать методы статистического анализа и математического моделирования, которые позволяют использовать статистические данные, накопленные в процессе эксплуатации, для оценки состояния и прогнозирования дальнейшего поведения валков. Под методами статистического анализа и математического моделирования здесь понимается определенный вычислительный алгоритм, реализованный на компьютерах и упрощенно имитирующий функционирование объектов.

Статистический анализ делится на три последовательных этапа [10]:

- статистическое наблюдение, т.е. сбор первичного статистического материала;
- обобщение и развитие результатов наблюдения, т.е. их обработка;
- анализ полученных общих материалов.

С развитием технологий Big Data и IoT поиск зависимостей между параметрами технологического процесса может дать компании больший эффект, чем просто методы статистического анализа.

Технологии Big Data и анализа данных позволяют сделать следующее [11-13]:

- находить закономерности, проявляющиеся в массовых явлениях под действием закона больших чисел;
- систематизировать и классифицировать данные на основе сходств и различий;
- анализировать общий материал, выявлять закономерности и связи в изучаемых фактах, рассчитывать обобщающие показатели (общие, относительные и средние величины, а также статистические коэффициенты).

### 3. Объект и постановка задачи

Данные оперативного контроля технологического процесса характеризуются различным происхождением и измеряются в различных количественных и качественных отложениях. Приведение данных оперативного контроля к виду, пригодному для разработки модели технологического процесса, является необходимым условием эффективности процесса моделирования [14].

Исходные данные представлены на пяти листах (рис. 1) в файле Microsoft Office Excel. Данные содержат информацию о материале валков (500 строк), рабочем процессе валков за 9 месяцев работы прокатного стана (18 080 строк), поставщиках валков (25 строк), сортаменте проката (269 968 строк).

|   | A                | B                | C            | D                  | E           | F |
|---|------------------|------------------|--------------|--------------------|-------------|---|
| 1 | дата_завалки     | дата_вывалки     | номер_клетки | положение_в_клетки | номер_валка |   |
| 2 | 01.10.2018 18:15 | 01.10.2018 21:46 | 12           | низ                | 199         |   |
| 3 | 01.10.2018 18:15 | 01.10.2018 21:46 | 12           | верх               | 445         |   |
| 4 | 01.10.2018 18:15 | 01.10.2018 21:46 | 11           | верх               | 284         |   |
| 5 | 01.10.2018 18:15 | 01.10.2018 21:46 | 10           | низ                | 172         |   |
| 6 | 01.10.2018 18:15 | 01.10.2018 21:46 | 11           | низ                | 275         |   |

(a) Лист 5

|   | A               | B               | C            | D                  | E           | F           | G |
|---|-----------------|-----------------|--------------|--------------------|-------------|-------------|---|
| 1 | дата_завалки    | дата_вывалки    | номер_клетки | положение_в_клетки | номер_валка | износ       |   |
| 2 | 01.01.2018 0:14 | 01.01.2018 3:32 | 8            | верх               | 222         | 0,684804912 |   |
| 3 | 01.01.2018 0:14 | 01.01.2018 3:32 | 8            | низ                | 260         | 0,614913529 |   |
| 4 | 01.01.2018 0:14 | 01.01.2018 3:32 | 9            | верх               | 410         | 1,135412151 |   |
| 5 | 01.01.2018 0:14 | 01.01.2018 3:32 | 9            | низ                | 57          | 0,516086546 |   |
| 6 | 01.01.2018 0:14 | 01.01.2018 3:32 | 10           | верх               | 273         | 0,657682871 |   |

(b) Лист 2

Рис. 1. Фрагмент файла с исходными данными

В качестве исходных данных для моделирования рассматривались: минуты (время прокатки партии продукции); номер клетки (задается цифрой); положение клетки (верхнее или нижнее); номер и материал валка (в закодированном виде каждый из параметров); количество листов, прокатываемых определенным валком; калибр, ширина и вес листа; сортность проката; износ валков.

Колонка «положение клетки стана» является проблемной, так как содержит текстовые данные («верх» – «низ»). Для удобства они кодируются цифрами 0 и 1.

Чтобы правильно подготовить данные для разработки прогностической модели, необходимо сначала выяснить, какие типы данных представлены в исходном файле, и проверить их на целостность. Проще всего удалить «пустые» значения, но если их много, имеет смысл заменить недостающие данные каким-либо числом, например, средним арифметическим по всему столбцу.

В результате проверки было установлено, что в столбцах нет пробелов. Кроме того, было обнаружено, что в некоторых строках указан нулевой износ валков после прокатки стали. Такие записи следует игнорировать, поскольку, даже если подобные «промахи» не являются ошибками, а представляют собой редкие исключения, их все равно вряд ли можно использовать [15–17].

Вычисление разницы между временем заполнения и временем выгрузки позволяет получить время работы валков для одной партии проката. Анализируя время прокатки рулонов с диапазонами заполнения и разгрузки валков, указанными в листе «Валки», можно рассчитать средний вес, ширину, калибр и количество рулонов, прокатываемых через эти валки. Полученные характеристики можно использовать для построения моделей.

Для определения влияния каждого исследуемого фактора на износ валков были рассчитаны критерии корреляции Пирсона ( $R$ ), характеризующие линейное влияние факторов, и построена матрица кросс-корреляции. При незначительном значении коэффициента некоторые признаки можно не учитывать при построении моделей (табл. 1).

Таблица 1

Диаграмма корреляции признаков

|                  | Минуты      | Номер клетки        | Положение клетки    | Номер валка | Материал валка | Листы    | Калибр | Ширина | Марки стали | Вес | Износ |
|------------------|-------------|---------------------|---------------------|-------------|----------------|----------|--------|--------|-------------|-----|-------|
| Минуты           | <b>1</b>    |                     |                     |             |                |          |        |        |             |     |       |
| Номер клетки     | 0.0037      | <b>1</b>            |                     |             |                |          |        |        |             |     |       |
| Положение клетки | 0.00012     | $-4 \times 10^{-5}$ | <b>1</b>            |             |                |          |        |        |             |     |       |
| Номер валка      | 0.0089      | 0.0086              | 0.0041              | <b>1</b>    |                |          |        |        |             |     |       |
| Материал валка   | 0.009       | 0.0021              | $3 \times 10^{-5}$  | $-0.014$    | <b>1</b>       |          |        |        |             |     |       |
| Листы            | <b>0.87</b> | 0.0034              | $-9 \times 10^{-5}$ | 0.0033      | 0.0046         | <b>1</b> |        |        |             |     |       |

|             | Минуты       | Номер<br>клетки | Положение<br>клетки   | Номер<br>валка | Материал<br>валка | Листы        | Калибр       | Ширина      | Марки<br>стали | Вес      | Износ    |
|-------------|--------------|-----------------|-----------------------|----------------|-------------------|--------------|--------------|-------------|----------------|----------|----------|
| Калибр      | <b>-0.34</b> | -0.0011         | $4.2 \times 10^{-5}$  | -0.003         | -0.0069           | -0.025       | <b>1</b>     |             |                |          |          |
| Ширина      | <b>-0.27</b> | 0.00049         | -0.00085              | -0.0026        | -0.0035           | -0.016       | <b>0.35</b>  | <b>1</b>    |                |          |          |
| Марки стали | <b>0.12</b>  | 0.002           | -0.00046              | 0.00047        | -0.012            | <b>0.16</b>  | <b>0.1</b>   | -0.032      | <b>1</b>       |          |          |
| Вес         | <b>-0.16</b> | -0.00079        | $-4.1 \times 10^{-5}$ | -0.0091        | -0.0045           | <b>-0.23</b> | -0.0021      | <b>0.14</b> | -0.087         | <b>1</b> |          |
| Износ       | <b>0.28</b>  | <b>-0.35</b>    | 0.0011                | -0.0029        | 0.063             | <b>0.17</b>  | <b>-0.23</b> | -0.037      | 0.094          | -0.021   | <b>1</b> |

(значимые коэффициенты выделены жирным шрифтом).

Проверка значимости коэффициентов корреляции по тесту Стьюдента показала, что значимыми являются коэффициенты корреляции, абсолютное значение которых превышает 0,1, то есть должно выполняться условие  $|R| \geq 0,1$ .

Из полученных данных следует, что положение валка в клетке ( $R = 0,0011$ ) и серийный номер валка ( $R = -0,0029$ ) не оказывают линейного влияния на износ валков. Кроме того, серийный номер валка (от 1 до 500) не является технологическим параметром и носит исключительно информационный характер. Положение валка в клетке (верхнее или нижнее) также носит исключительно информационный характер. Эти признаки не будут учитываться при построении будущей модели.

Несмотря на то, что такие эксплуатационные параметры, как материал валков, ширина, вес и марка проката, также не удовлетворяют условию  $|R| \geq 0,1$ , было принято решение не исключать эти параметры из рассмотрения.

Таким образом, следующим этапом исследования является разработка прогнозной модели износа валков прокатного стана на основе большого массива данных оперативного контроля, содержащих информацию о времени заполнения и выгрузки валков, сортаменте проката, материале валков и времени нахождения валков в работе [18].

#### 4. Алгоритм

Алгоритм разработки прогнозной модели износа валков прокатного стана на основе большого массива данных оперативного контроля представлен на рисунке 2.

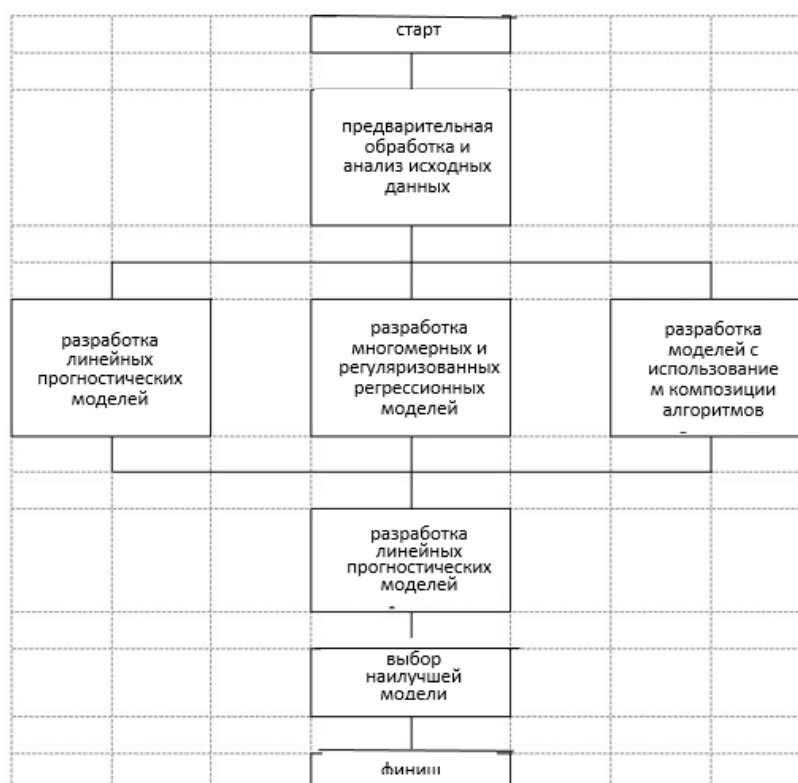


Рис. 2. Алгоритм разработки прогностической модели износа валков мельницы

#### 4.1. Использование больших данных для разработки линейных прогнозных моделей

Для разработки линейной прогностической модели используются перекрестная валидация (CV) и метод наименьших квадратов.

Суть метода наименьших квадратов заключается в том, что сумма квадратов отклонений экспериментальных значений от сглаживающей кривой сводится к минимуму:

$$\sum_{i=1}^N [y_i - \varphi(x_i)]^2 = \min$$

где  $y_i$  и  $x_i$  – значения экспериментальных данных в  $i$ -м эксперименте,  $N$  – количество экспериментов,  $\varphi(x)$  – желаемая линейная регрессия  $y$  на  $x$  в виде  $\varphi(x) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_kx_k$ , и  $k$  – число факторов.

Суть метода CV заключается в том, что весь массив данных оперативного контроля разбивается на определенное количество подвыборок (блоков). Один из блоков используется для тестирования модели (проверки модели на адекватность исследуемому процессу), остальные – для обучения. Затем тестовый блок используется для обучения, а следующий блок выбирается для тестирования. Схема кросс-валидации представлена на рисунке 3 (открытые блоки – блоки обучения модели, заполненный блок – тестовая подвыборка). Данный метод позволяет получить несмещенную оценку вероятности ошибки в прогностической модели и предотвратить оптимистическую переоценку качества вышеупомянутой.



Рис. 3. Схема работы «Перекрестной валидации»

#### 4.2. Использование больших данных для разработки многомерных и регуляризованных регрессионных моделей

Суть регуляризации заключается в наложении дополнительных ограничений на различные параметры или добавлении априорной информации, что позволяет уменьшить ошибку модели при увеличении ее сложности [19, 20].

На основе одной и той же базы данных оперативного управления были построены: многомерная регрессия с L1-регулятором (Лассо), многомерная регрессия с L2-регулятором (Ридж) и многомерная регрессия со смешанным регуляризатором (ElasticNet).

Регуляризация – это способ уменьшить сложность модели, чтобы предотвратить переобучение или исправить некорректно поставленную задачу. Обычно это достигается путем добавления некоторой априорной информации к постановке задачи.

Суть регуляризации L1 заключается в том, чтобы выбрать из всего массива факторов лишь небольшое количество наиболее важных, задающих тренд, и удалить все остальные, которые являются просто шумом. Таким образом, регуляризация L1 направлена на уменьшение размерности модели.

Регуляризация L2 направлена на уменьшение размерности пространства за счет запрета непропорционально больших весовых коэффициентов, что предотвращает переобучение модели.

Разработка многомерной регрессии с использованием как L1, так и L2 регуляризации называется смешанным регуляризатором (ElasticNet) и учитывает эффективность обоих методов: уменьшение размерности модели и уменьшение размерности факторного пространства.

### 4.3. Композиция алгоритмов для разработки моделей на основе больших данных

Основной метод композиции алгоритмов заключается в объединении большого количества моделей в одну композицию. Итоговое качество полученной модели будет значительно улучшено за счет того, что отдельные модели будут исправлять ошибки друг друга.

В данном исследовании рассматриваются такие методы, как случайный лес и градиентный бустинг.

Метод случайного леса является одним из наиболее профессиональных и качественных методов машинного обучения. Ключевая идея этого метода для поиска регрессионных зависимостей заключается в усреднении результатов нескольких моделей, построенных независимо друг от друга на случайных подвыборках одного массива данных. Таким образом, набор низкоточных алгоритмов при объединении в одну композицию дает впечатляющий результат, несмотря на значительную долю случайности, представленную в этом методе.

Преимуществом метода случайного леса является его устойчивость к переборщикам. Поскольку все алгоритмы разрабатываются независимо друг от друга, увеличение их количества в композиции не усложняет итоговую модель [24, 25].

В данном исследовании алгоритм случайного леса использует снижение размерности пространства признаков с помощью анализа главных компонент (АГК). Использование техники снижения размерности пространства признаков позволяет представить исходный набор данных в терминах меньшего количества переменных и при этом уменьшить объем вычислительных ресурсов, необходимых для обеспечения работы модели.

Метод градиентного наращивания. Отличие этого метода от предыдущего заключается в том, что в данном алгоритме при построении композиции все модели не являются независимыми, а следуют друг за другом. Более того, каждый последующий алгоритм старается исправить и компенсировать ошибки предыдущего. Поэтому на получение правильного ответа уходит меньше времени.

В данном исследовании градиентный бустинг использует метод градиентного спуска для минимизации функции ошибки прямо в этих последовательных моделях. Такой подход позволяет расширить круг задач, решаемых этим алгоритмом, а также часто приводит к повышению точности предсказания.

### 4.4. Оценка качества модели

Качество модели оценивается с помощью средней квадратичной ошибки (МКО) между прогнозируемым и фактическим износом валков, коэффициента корреляции ( $R$ ) между фактическими и прогнозируемыми значениями износа валков прокатного стана и коэффициента детерминации ( $R^2$ ) между фактическими и прогнозируемыми значениями износа валков прокатного стана.

Коэффициент детерминации наглядно показывает, насколько построенная модель точнее среднего значения целевой переменной, и соответствует следующему выражению:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \approx 1 - \frac{MSE}{VAR(y)}$$

где  $y_i$  – фактическое значение износа валков,  $\hat{y}_i$  – прогнозируемое моделью значение износа валков, а  $\bar{y}$  – среднее значение износа валков по исходным данным. Если коэффициент детерминации  $R^2$  равен 1, то рассчитанные по модели значения износа валков прокатного стана в точности повторяют фактические значения, что свидетельствует об адекватности математической модели объекту исследования. Если коэффициент детерминации  $R^2$  близок к нулю, то это означает, что модель несовершенна и лучше взять среднее значение  $y$ . Модели признаются адекватными, если коэффициент детерминации  $R^2 \geq 0,7$ .

## 5. Результаты

На рисунке 4 показаны результаты сравнения фактического и прогнозируемого износа валков для различных моделей.

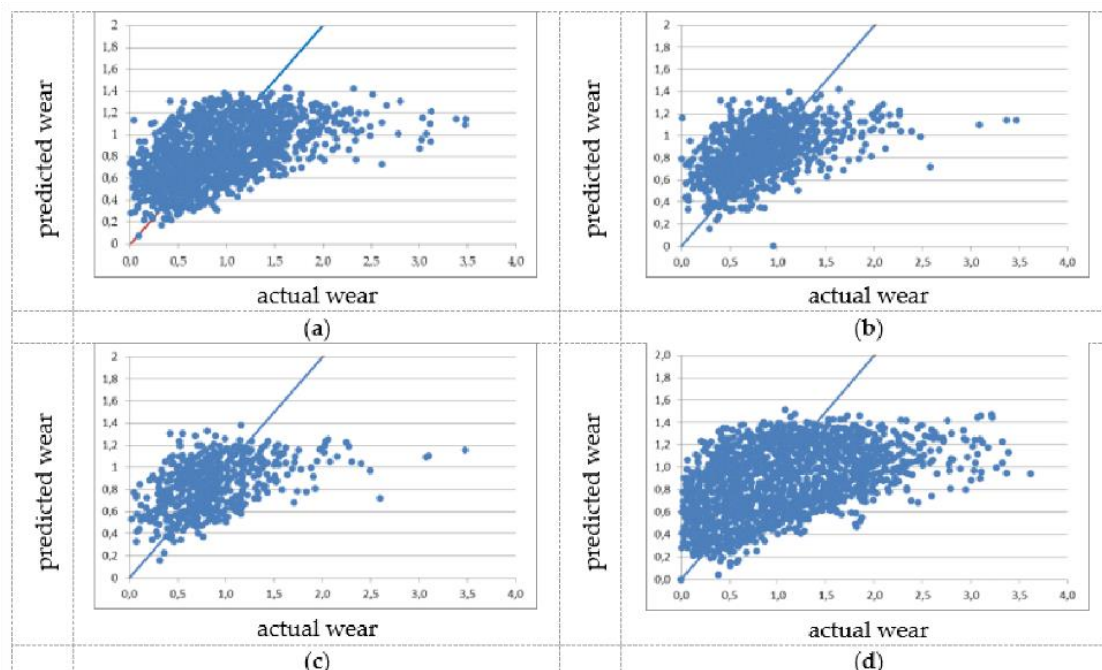


Рис. 4. Сравнение прогнозируемого и фактического износа валков прокатного стана:  
(a) 30 % тестовый образец; (b) 10 % тестовый образец; (c) 5 % тестовый образец; (d) перекрестная валидация

Для наглядности можно сравнить модели, построенные с кросс-валидацией и без нее.

Вместо кросс-валидации весь массив данных оперативного контроля делится на обучающую и тестовую выборки путем смешивания всех признаков и выбора определенного процентного соотношения между обучающей и тестовой выборками. В качестве модели используется линейная регрессия, найденная методом наименьших квадратов.

Анализ графиков (рисунок 4) на предмет симметрии относительно прямой линии  $Y_{\text{predicted}} = Y_{\text{actual}}$  показывает, что во всех случаях имеет место недооценка прогнозируемых значений. При реальных значениях износа 0–4 прогнозируемые значения не превышают 0–1,6.

В этом случае качество модели меняется в зависимости от количества данных, выбранных для обучения модели и тестовой валидации. Большее количество данных на тест уменьшает объем обучающих данных и приводит к снижению точности модели, и наоборот [26, 27].

Результаты оценки адекватности полученных моделей приведены в таблице 2.

Таблица 2

Оценка адекватности многомерной и регуляризованной регрессионных моделей.  
МКО – средняя квадратическая ошибка

| Модель                | Оценка | СКО   | $R^2$ | $R$   |
|-----------------------|--------|-------|-------|-------|
| 30 % тестовый образец |        | 0,109 | 0,257 | 0,513 |
| 10 % тестовый образец |        | 0,112 | 0,257 | 0,505 |
| 5 % тестовая выборка  |        | 0,108 | 0,253 | 0,503 |
| Кросс-валидация       |        | 0,113 | 0,256 | 0,507 |
| Лассо                 |        | 0,113 | 0,253 | 0,504 |
| Гребень               |        | 0,113 | 0,256 | 0,506 |
| ElasticNet            |        | 0,113 | 0,256 | 0,507 |
| Случайный лес         |        | 0,021 | 0,798 | 0,933 |
| Градиентное усиление  |        | 0,021 | 0,847 | 0,927 |

Таким образом, результаты этого анализа свидетельствуют о незначительных различиях в результатах моделирования. Все модели нельзя считать пригодными для прогнозирования величины износа валков в прокатном стане. Поэтому необходимо выбрать другой тип зависимости [28, 29].

Введение регуляризатора в линейную или многомерную модель не привело к повышению точности прогнозирования износа валков прокатного стана. Наглядно видно, что предложенные модели предсказывают значение целевого параметра не точнее, чем среднее арифметическое значение износа валков прокатного стана.

Исходя из полученных данных, можно утверждать, что в данном случае требуется либо переосмысление или интеллектуализация исходных данных, либо использование более сложных моделей [30].

Сравнение прогнозируемых методом случайного леса и фактических значений износа валков прокатного стана представлено на рисунке 5а. Сравнение прогноза по методу градиентного усиления, и фактических значений износа валков прокатного стана показано на рисунке 5б.

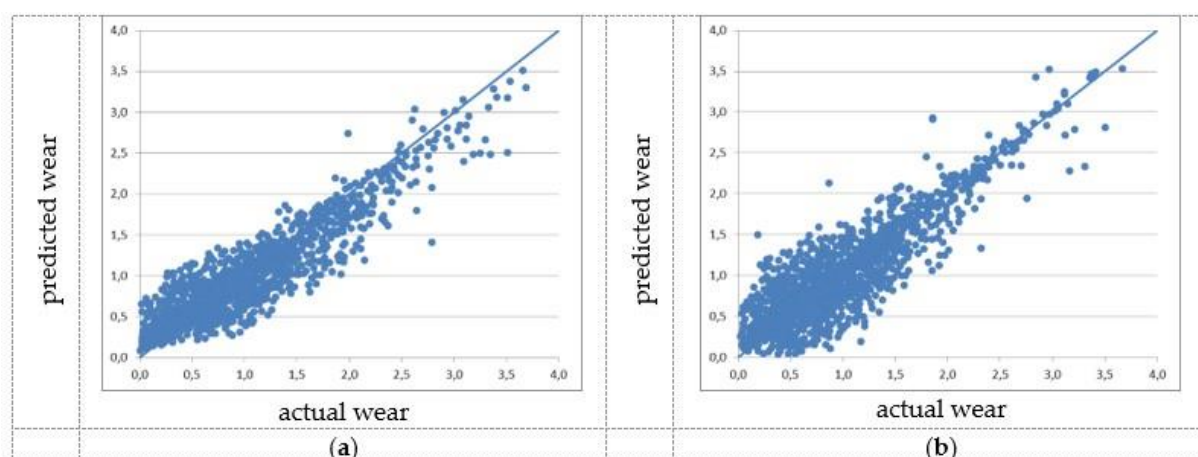


Рис. 5. Сравнение прогнозируемого и фактического износа валков прокатного стана: (а) по модели случайного леса; (б) по модели градиентного усиления

Результаты оценки адекватности модели случайного леса и модели градиентного усиления значительно превосходят предыдущие модели (табл. 2).

По сравнению с линейной, многомерной и регуляризованной моделями среднеквадратичная ошибка (СКО) уменьшилась примерно в пять раз, а коэффициенты детерминации и корреляции приблизились к единице. То есть модель случайного леса можно признать адекватной объекту исследования и использовать для прогнозирования степени износа валков прокатного стана в сталелитейной промышленности.

По коэффициенту детерминации  $R^2$  градиентный бустинг является более точной моделью по сравнению с моделью случайного леса (коэффициент детерминации ближе к единице). Среднеквадратичные ошибки обеих моделей равны, но, как видно из рисунка 5, при использовании метода градиентного усиления наблюдается большее количество совпадений прогнозируемого и фактического износа, чем при использовании метода случайного леса.

Анализ графиков (Рис. 5) на предмет симметрии относительно прямой линии  $Y_{\text{predicted}} = Y_{\text{actual}}$  показывает, что при построении модели случайного леса наблюдается тенденция к занижению прогнозируемых значений. Очевидно, что большинство значений расположены ниже прямой линии (рис. 5а). В случае разработки модели градиентного усиления прогнозируемые значения располагаются симметрично относительно прямой линии  $Y_{\text{predicted}} = Y_{\text{actual}}$  (рис. 5б). Таким образом, модель градиентного усиления является более предпочтительной.

При необходимости, проведя дополнительную оптимизацию модели, можно добиться еще большего снижения ошибки прогноза [31]. Таким образом, модель градиентного бустинга предпочтительнее.

## 6. Выводы

На основании проведенного исследования можно сделать следующие выводы.

1. Гипотеза об использовании большого объема производственных данных (Big Data) для поиска статистически значимых зависимостей оказалась полностью подтвержденной [32]. Данные оперативного контроля являются неисчерпаемым источником информации. Извлечение полезной информации из Big Data – важная производственная задача [33].

2. Для повышения точности моделей необходимо заранее подготовить статистический материал (убрать выбросы, «странности» и случайные результаты измерений; отфильтровать данные; выделить различные режимы работы и рассмотреть их отдельно) и выбрать подходящий тип математической зависимости. Качество разработанных моделей напрямую зависит от качества подготовки учебного материала [34].

3. Анализ корреляционных зависимостей данных показал, что наиболее значимыми факторами, влияющими на износ валков, являются размеры и марки прокатываемых стальных листов. Кроме того, не последнюю роль играет материал, из которого изготовлены валки.

4. Построение линейных и многомерных регрессионных моделей на основе большого массива данных не дает качественного результата, так как не позволяет учесть сложные и многосвязные зависимости между входными переменными. Лучше всего работают композиционные модели, устойчивые к перестройке, шуму и выбросам. Однако при небольшом объеме данных, которые можно описать простой моделью, целесообразнее использовать многомерную регрессию.

5. Таким образом, прогностическая модель износа валков прокатного стана позволит рационально использовать валки с точки зрения минимизации общего износа валков. Предложенная модель позволит перераспределить имеющиеся рабочие валки между клетями с целью снижения общего износа валков.

**Вклад авторов:** концептуализация, Н.В.; методология, Н.В.; программное обеспечение, Е.Ф.; валидация, Н.В. и Е.Ф.; обработка данных, А.К.; визуализация, А.К. Все авторы прочитали и согласились с опубликованной версией рукописи.

**Финансирование:** данное исследование не получало внешнего финансирования.

**Заявление Совета по институциональному надзору:** не применимо.

**Заявление об информированном согласии:** не применимо.

**Заявление о доступности данных:** не применимо.

**Конфликты интересов:** авторы не заявляют о конфликте интересов.

## ЛИТЕРАТУРА

1. Abrosimov, A.A.; Shelyago, E.V.; Yazynina, I.V. Justification of Representative Data Volume of Porosity and Permeability Properties for Obtaining Statistically Reliable Petrophysical Connections. *J. Min. Inst.* 2018, 233, 487–491. <http://doi.org/10.31897/PMI.2018.5.487>
2. Sharikov, Y.V.; Snegirev, N.V.; Tkachev, I.V. development of a control system based on predictive mathematical model of the C5–C6 isomerization process. *J. Chem. Technol. Metall.* 2020, 55, 335–344.
3. Kadyrov, E.D.; Koteleva, N.I. Introducing neural-network algorithms into an automated system designed to control metallurgical processes. *Metallurgist* 2011, 54, 799–802. <http://doi.org/10.1007/s11015-011-9377-7>
4. Demidovich, V.B.; Chmilenko, F.V.; Rastvorova, I.I. Utilization of induction heating in the line of continuous casting–continuous rolling of steel. *Acta Technica CSAV* 2015, 60, 107–118.
5. Zhukovskiy, Y.L.; Korolev, N.A.; Babanova, I.S.; Boikov, A.V. The prediction of the residual life of electromechanical equipment based on the artificial neural network. *IOP Conf. Ser. Earth Environ. Sci.* 2017, 87, 032056. <http://doi.org/10.1088/1755-1315/87/3/032056>
6. Oprea, G.; Andrei, H. Power quality analysis of industrial company based on data acquisition system, numerical algorithms and compensation results. In *Proceedings of the 2016 International Symposium on Fundamentals of Electrical Engineering, ISFEE, Bucharest, Romania, 30 June–2 July 2016*; p. 7803232.
7. Galkin, V.; Kolytyn, A. Investigation of probabilistic models for forecasting the efficiency of proppant hydraulic fracturing technology. *J. Min. Inst.* 2021, 246, 650–659. <http://doi.org/10.31897/PMI.2020.6.7>
8. Vasilyeva, N.V.; Koteleva, N.I.; Fedorova, E.R. Real-time control data wrangling for development of mathematical control models of technological processes. *J. Phys. Conf. Ser.* 2018, 1015, 032067. <http://doi.org/10.1088/1742-6596/1015/3/032067>
9. Bazhin, V.Y.; Kulchitskiy, A.A.; Kadrov, D.N. Complex control of the state of steel pins in Soderberg electrolytic cells by using computer vision systems. *Tsvetnye Met.* 2018, 27–32. <http://doi.org/10.17580/tsm.2018.03.04>
10. Utekhin, G. Use of statistical techniques in quality management systems. In *Proceedings of the 8 International Conference Reliability and Statistics in Transportation and Communication–2008, Riga, Latvia, 17–20 October 2018*; pp. 329–334.

11. Boikov, A.V.; Savelev, R.V.; Payor, V.A.; Erokhina, O.O. The control method concept of bulk material behaviour in the pelletizing drum for improving the results of DEM-modeling. *CIS Iron Steel Rev.* 2019, 17, 10–13. <http://doi.org/10.17580/cisir.2019.01.02>
12. Leonidovich, Z.Y.; Urievich, V.B. The development and use of diagnostic systems and estimation of residual life in industrial electrical equipment. *Int. J. Appl. Eng. Res.* 2015, 10, 41150–41155.
13. Milyuts, V.G.; Tsukanov, V.V.; Pryakhin, E.I.; Nikitina, L.B. Saint Petersburg Mining University Development of Manufacturing Technology for High-Strength Hull Steel Reducing Production Cycle and Providing High-Quality Sheets. *J. Min. Inst.* 2019, 239, 536–543. <http://doi.org/10.31897/pmi.2019.5.536>
14. Thombansen, U.; Purrio, M.; Buchholz, G.; Hermanns, T.; Molitor, T.; Willms, K.; Schulz, W.; Reisgen, U. Determination of process variables in melt-based manufacturing processes. *Int. J. Comput. Integr. Manuf.* 2016, 29, 1147–1158. <http://doi.org/10.1080/0951192X.2016.1145799>
15. Servin, R.; Arreola, S.A.; Calderón, I.; Perez, A.; Miguel, S.M.S. Effect of Crown Shape of Rolls on the Distribution of Stress and Elastic Deformation for Rolling Processes. *Metals* 2019, 9, 1222. <http://doi.org/10.3390/met9111222>
16. Li, H.-J.; Xu, J.-Z.; Wang, G.-D.; Shi, L.-J.; Xiao, Y. Development of strip flatness and crown control model for hot strip mills. *J. Iron Steel Res. Int.* 2010, 17, 21–27. [http://doi.org/10.1016/S1006-706X\(10\)60067-2](http://doi.org/10.1016/S1006-706X(10)60067-2)
17. Zhao, N.; Cao, J.; Zhang, J.; Su, Y.; Yan, T.; Rao, K. Work roll thermal contour prediction model of nonoriented electrical steel sheets in hot strip mills. *J. Univ. Sci. Technol. Beijing Miner. Met. Mater.* 2008, 15, 352–356. [http://doi.org/10.1016/S1005-8850\(08\)60066-3](http://doi.org/10.1016/S1005-8850(08)60066-3)
18. Turk, R.; Fajfar, P.; Robic, R.; Perus, I. Prediction of hot strip mill roll wear. *Metalurgija* 2002, 41, 47–51.
19. Taimasov, B.T.; Sarsenbayev, B.K.; Khudyakova, T.M.; Kolesnikov, A.S.; Zhanikulov, N.N. Development and testing of low-energy intensive technology of receiving sulfate-resistant and road Portland cement. *Eurasian Chem. Technol. J.* 2017, 19, 347–355. <http://doi.org/10.18321/ectj683>
20. Abdulaev, E.K.; Makharatkin, P.N.; Kuzhelev, A.I.; Grudin, N.N. Assessment of technical condition of gearbox-motor-wheels and tires according to heating wear criterion when transporting building materials. *IOP Conf. Series Mater. Sci. Eng.* 2020, 775, 012001. <http://doi.org/10.1088/1757-899X/775/1/012001>
21. Bolobov, V.; Chupin, S.; Binh, L.T. On the Wear Intensity Ratio of a Striker under Dynamic and Static Conditions. *IOP Conf. Ser. Earth Environ. Sci.* 2020, 459, 062085. <http://doi.org/10.1088/1755-1315/459/6/062085>
22. Krasnyy, V.; Maksarov, V.V.; Maksimov, D. Improving the Wear Resistance of Piston Rings of Internal Combustion Engines when Using Ion-Plasma Coatings. *Key Eng. Mater.* 2020, 854, 133–139. <http://doi.org/10.4028/www.scientific.net/KEM.854.133>
23. Ratra, R.; Gulia, P. Big Data Tools and Techniques: A Roadmap for Predictive Analytics. *Int. J. Eng. Adv. Technol.* 2019, 9, 4986–4992.
24. Thillaieswari, B. Comparative Study on Tools and Techniques of Big Data Analysis. *Int. J. Adv. Netw. Appl.* 2017, 8, 61–66.
25. George, G.; Lavie, D. Big data and data science methods for management research. *Acad. Manag. J.* 2016, 59, 1493–1507. <http://doi.org/10.5465/amj.2016.4005>
26. Maratea, A.; Petrosino, A.; Manzo, M. Extended Graph Backbone for Motif Analysis. In Proceedings of the 18th International Conference on Hybrid Systems: Computation and Control, Seattle, WA, USA, 14–16 April 2015; pp. 36–43.
27. Nguyen, T.L. A Framework for Five Big V's of Big Data and Organizational Culture in Firms. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 5411–5413.
28. Kaur, N.; Singh, G. A Review Paper On Data Mining And Big Data. *Int. J. Adv. Res. Comput. Sci.* 2017, 8, 407–409.
29. Rao, J.N.; Ramesh, M. A Review on Data Mining & Big Data, Machine Learning Techniques. *Int. J. Recent Technol. Eng.* 2019, 7, 914–916.
30. Kaisler, S.; Armour, F.; Espinosa, J.A.; Money, W. Big data: Issues and challenges moving forward. System sciences (HICSS). In Proceedings of the 2013 46th Hawaii International Conference on System Sciences, Maui, HI, USA, 7–10 January 2013; pp. 995–1004.
31. Dean, J.; Ghemawat, S. Mapreduce: Simplified data processing on large clusters. *Commun. ACM* 2008, 51, 107–113. <http://doi.org/10.1145/1327452.1327492>
32. Katal, A.; Wazid, M.; Goudar, R.H. Big data: Issues, challenges, tools and Good practices. In Proceedings of the 2013 Sixth International Conference on Contemporary Computing (IC3), Noida, India, 8–10 August 2013; pp. 404–409.
33. Wu, X.; Zhu, X.; Wu, G.-Q.; Ding, W. Data mining with big data. *IEEE Trans. Knowl. Data Eng.* 2014, 26, 97–107. <http://doi.org/10.1109/tkde.2013.109>
34. Lindell, Y.; Pinkas, B. Privacy Preserving Data Mining. *J. Cryptol.* 2002, 15, 177–206. <http://doi.org/10.1007/s00145-001-0019-2>