УДК 681.51.015

ИДЕНТИФИКАЦИЯ СТАТИСТИЧЕСКИХ МОДЕЛЕЙ ТЕХНОЛОГИЧЕСКИХ ПРОЦЕССОВ С ЗАПОЛНЕНИЕМ ПРОПУСКОВ В ДАННЫХ

Л. А. Кузнецов, А. М. Корнеев, М. Г. Журавлева

Липецкий государственный технический университет

Исследована возможность применения стратегии *EM*-алгоритма к решению задачи построения линейных множественных моделей регрессии по массивам, содержащим неполную информацию о производственном процессе. Выполнено сравнение моделей, построенных по комплектным данным, и данным, пропуски в которых заполнялись безусловными и условными средними. Получена усовершенствованная модель зависимости одной из механических характеристик листового проката от набора факторов технологии, основанная на применении *EM*-алгоритма.

ВВЕДЕНИЕ

Построение статистических моделей зависимости выходных характеристик продукции от параметров технологии ее производства, предназначенных для решения задач прогнозирования и управления, предполагает использование массивов информации, которые часто содержат пропущенные значения. Если исключать наблюдения, содержащие хотя бы один пропуск, из-за недостаточного объема комплектных данных оценки параметров идентифицируемых моделей могут оказываться смещенными или искаженными. Поэтому предпочтительна обработка всей доступной информации и заполнение на ее основании пропущенных значений в информационном массиве.

Существующие методы заполнения пропусков [1, 2] не всегда применимы к решению задачи идентификации моделей зависимости между случайными величинами, изменяющимися в количественных шкалах. Один из способов анализа числовой информации заключается в заполнении безусловными средними или другими характеристиками частных выборочных распределений. Его целесообразно применять, когда исследуемые переменные независимы друг от друга, поэтому в рассматриваемом случае он малоэффективен (в работе это показано на практическом примере). Более результативны подходы, предполагающие использование существующей информации о связях. К простым относится метод заполнения пропусков условными средними по присутствующим значениям (метод Бака). Более надежна с точки зрения получения оптимальных оценок параметров итеративная стратегия *EM*-алгоритма (expectation-maximization algorithm), каждая итерация которого подразумевает выполнение двух шагов: шага E — вычисления математического ожидания и шага M — максимизации (метод Бака является частным случаем данного метода, реализующим его первые две итерации) [1]. Помимо оценивания параметров, с помощью *EM*-алгоритма для неполной многомерной нормальной выборки можно решать другие задачи статистического анализа, в том числе задачу построения множественной линейной регрессии. Рассмотрим подробнее методику *EM*-алгоритма в контексте идентификации моделей зависимости выходных свойств продукции, являющейся результатом сложного производственного процесса, от факторов технологии.

1. МАТЕМАТИЧЕСКОЕ ОПИСАНИЕ *EM*-АЛГОРИТМА

Пусть имеется k-мерная нормальная переменная $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k)$, причем $\boldsymbol{x} = (\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_r) \cup (\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_{k-r})$, где $\boldsymbol{t}_i = (t_{1i}, ..., t_{ni})^T$ — i-я технологическая величина, $i = 1, ..., r$, $\boldsymbol{s}_j = (s_{1j}, ..., s_{nj})^T$, — j-я выходная характеристика, $j = 1, ..., k - r$, полученные в результате пассивного эксперимента, n — число наблюдений. Она характеризуется параметрами $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, где $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_k)^T$ — вектор средних, $\boldsymbol{\Sigma} = (\sigma_{il})$, $i, l = 1, 2, ..., k$ — ковариационная матрица. Множество исходных данных удобно пред-

ставить в виде: $\boldsymbol{x} = (\boldsymbol{x}_{\text{набл}}, \boldsymbol{x}_{\text{проп}})$, где $\boldsymbol{x}_{\text{набл}}$ — подмножество наблюдаемых значений факторов технологии и выходных свойств, $\boldsymbol{x}_{\text{проп}}$ — подмножество пропущенных значений. Таким образом, $\boldsymbol{x}_{\text{набл}} = (\boldsymbol{x}_{\text{набл.1}}, \boldsymbol{x}_{\text{набл.2}}, ..., \boldsymbol{x}_{\text{набл.}n})$, где $\boldsymbol{x}_{\text{набл.}i}$ — множество переменных (технологических величин и (или) выходных характеристик) с присутствующими значениями в наблюдении i, $i = 1, 2, ..., n$.

Первая итерация *EM*-алгоритма предполагает вычисление начальных значений параметров. Если число комплектных наблюдений хотя бы на единицу превосходит число переменных, вектор $\boldsymbol{\mu}$ и матрицу $\boldsymbol{\Sigma}$ можно рассчитывать только по полным данным. В противном случае пропущенные данные заполняются с помощью одного из простых методов, например, безусловными средними. При этом по каждой переменной рассчитывается выборочное среднее по присутствующим значениям:

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n} x'_{ij}, \quad j = 1, 2, ..., k,$$

где n_j — число наблюдаемых значений переменной j,

$$x'_{ij} = \begin{cases} x_{ij}, \text{ если } x_{ij} \text{ наблюдается,} \\ 0, \text{ если } x_{ij} \text{ пропущено.} \end{cases}$$

Если на m-й итерации рассчитаны параметры $\boldsymbol{\theta}^{(m)} = (\boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)})$, на шаге E необходимо заменить пропущенные данные условными средними значениями по наблюдаемым в i-м опыте переменным $\boldsymbol{x}_{\text{набл.}i}$:

$$E\left(\sum_{i=1}^{n} x_{ij} \middle| \mathbf{x}_{\text{набл}}, \boldsymbol{\theta}^{(m)} \right) = \sum_{i=1}^{n} x_{ij}^{(m)}, \quad j = 1, 2, ..., k,$$

$$E\left(\sum_{i=1}^{n} x_{ij} x_{il} \middle| \mathbf{x}_{\text{набл}}, \boldsymbol{\theta}^{(m)} \right) = \sum_{i=1}^{n} x_{ij}^{(m)} x_{il}^{(m)} + c_{jli}^{(m)},$$
$$j, l = 1, 2, ..., k, \qquad (1)$$

где

$$x_{ij}^{(m)} = \begin{cases} x_{ij}, \text{ если } x_{ij} \text{ наблюдается,} \\ E(x_{ij} | \mathbf{x}_{\text{набл}}, \boldsymbol{\theta}^{(m)}), \text{ если } x_{ij} \text{ пропущено,} \end{cases}$$

а $c_{jli}^{(m)}$ — добавочные ковариации:

$$c_{jli}^{(m)} = \begin{cases} 0, \text{ если } x_{ij} \text{ или } x_{il} \text{ наблюдаются,} \\ \operatorname{cov}(x_{ij} | \mathbf{x}_{\text{набл}}, \boldsymbol{\theta}^{(m)}), \text{ если } x_{ij} \text{ и } x_{il} \text{ пропущены.} \end{cases}$$

Условные средние и добавочные ковариации $c_{jli}^{(m)}$ из формулы (1) вычисляются по текущим оценкам параметров посредством применения оператора свертки к приращенной ковариационной матрице [1] таким образом, что наблюдаемые в i-м опыте переменные рассматриваются как независимые, а те, в которых значение i пропущено — как отклики. Применение оператора свертки к некоторой симметричной матрице $\boldsymbol{A} = (a_{ij})$ по строке и столбцу p приводит к получению симметричной матрицы $\boldsymbol{B} = (b_{ij})$ той же размерности, что исходная, с элементами:

$$b_{pp} = -1/a_{pp},$$
$$b_{ip} = b_{pi} = a_{ip}/a_{pp}, \quad p \neq i,$$
$$b_{ij} = a_{ij} - a_{ip}a_{pj}/a_{pp}, \quad p \neq i, \quad p \neq j.$$

Преимущества применения оператора свертки, по сравнению с другими способами поиска оценок метода наименьших квадратов, состоят в простоте вычислений и удобном способе извлечения полной информации о регрессии из результирующей блочной матрицы.

Обозначим исходную k-мерную переменную $\boldsymbol{x}$, дополненную слева столбцом из единиц, через $\boldsymbol{X}$. Приращенная ковариационная матрица строится следующим образом. Вычисляется произведение $n^{-1}\boldsymbol{X}^T\boldsymbol{X}$, свертка которого по строке и столбцу j, $j = 1, ..., k$, позволяет получить регрессию всех остальных переменных на переменную с номером j. В общем случае свертка по нулевой строке и столбцу и первым r переменным позволяет получить многомерную регрессию переменных $\boldsymbol{x}_{r+1}, \boldsymbol{x}_{r+2}, ..., \boldsymbol{x}_k$ на $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_r$. Результатом такой свертки будет блочная матрица:

$$\operatorname{swp}[0, 1, ..., r] n^{-1}\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} -\boldsymbol{F} & \boldsymbol{G} \\ \boldsymbol{G} & \boldsymbol{H} \end{bmatrix}, \qquad (2)$$

где матрица $\boldsymbol{G}$, размерностью $(r+1) \times (k-r)$, j-й столбец которой содержит вычисленные методом наименьших квадратов свободный член и коэффициенты наклона регрессии $\boldsymbol{x}_{j+r}$ на $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_r$, $j = 1, 2, ..., k-r$; $\boldsymbol{H}$ — размерностью $(k-r) \times (k-r)$ остаточная ковариационная матрица $\boldsymbol{x}_{r+1}, \boldsymbol{x}_{r+2}, ..., \boldsymbol{x}_k$ при фиксированных $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_r$; элементы матрицы $\boldsymbol{F}$, размерностью $(r+1) \times (r+1)$, после умножения на соответствующие остаточные дисперсии и ковариации из матрицы $\boldsymbol{H}$, позволяют получить ковариационную матрицу коэффициентов регрессии (последнее справедливо для полных данных, в работе применялся метод оценивания данной матрицы для неполных данных [1, с. 176]).

На каждой итерации пропущенные значения заполняются новыми, более точными, по которым на шаге M заново вычисляются параметры:

$$\mu_j^{(m+1)} = n^{-1} \sum_{i=1}^{n} x_{ij}^{(m)}, \quad j = 1, 2, ..., k,$$

$$\sigma_{jl}^{m+1} = n^{-1} E\left(\sum_{i=1}^{m} x_{ij} x_{il} \middle| x_{\text{набл}} \right) - \mu_j^{(m+1)} \mu_l^{(m+1)},$$
$$j, l = 1, 2, ..., k.$$

Процесс сходится к некоторой оптимальной совокупности оценок параметров, по которым строится окончательное уравнение регрессии. В частности, для множества $(\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_r)$ факторов технологии и одного отклика $\boldsymbol{s}$ после завершения работы *EM*-алгоритма дополнительно проводится соответствующая свертка. При этом матрица $\boldsymbol{G}$ из матрицы (2) будет иметь размерность $(r+1) \times 1$, т. е. являться вектором, содержащим свобод-

ный член и оценки параметров регрессии $\boldsymbol{t}_1$, $\boldsymbol{t}_2$, ..., $\boldsymbol{t}_r$ на $\boldsymbol{s}$, а матрица $\boldsymbol{H}$ — состоять из единственного значения, остаточной дисперсии полученной модели регрессии.

2. ПРИМЕНЕНИЕ ПРОЦЕДУРЫ *EM*-АЛГОРИТМА К АНАЛИЗУ РЕАЛЬНЫХ ПРОИЗВОДСТВЕННЫХ ДАННЫХ С ПРОПУСКАМИ

Для экспериментального обоснования эффективности *EM*-алгоритма в работе была реализована следующая стратегия. На первом этапе в контексте поиска наилучшей модели зависимости исследовалось качество заполнения пропущенных значений безусловными средними, условными средними и с помощью процедуры *EM*-алгоритма. По комплектным данным (неполные наблюдения предварительно были удалены) рассчитывались параметры основной модели зависимости. Последняя использовалась в качестве контрольной (эталонной) для сравнения с тремя моделями, которые строились многократно по массиву комплектной информации, но с искусственно вводимыми пропусками. Индексы строки и столбца пропускаемого значения генерировались с помощью датчика псевдослучайных чисел по равномерному закону. На каждом этапе число пропущенных значений увеличивалось, осуществлялся поиск модели, наиболее близкой к контрольной. Для этого определялись отклонения вектора оценок параметров каждой из рассматриваемых моделей от соответствующего вектора для контрольной модели, рассчитывались евклидовы нормы разности. По оценкам параметров моделей оценивались качество предсказания ими исходного отклика по комплектным данным и адекватность, проверялась гипотеза о равенстве выборочных распределений откликов, прогнозируемых по каждой из трех моделей, отклику комплектной модели (по критериям χ^2, Колмогорова — Смирнова [3]). На основании полученной информации был выбран оптимальный способ заполнения пропущенных значений и построена результирующая модель. Далее подробно изложены основные моменты и результаты проведенного анализа массива реальных производственных данных.

Исследовался массив с пропусками, содержащий набор наиболее существенных переменных, характеризующих ход и результаты технологического процесса производства 309 рулонов стали марки 08Ю. Из исходного массива было выделено 254 комплектных наблюдения. Посредством разработанной программной реализации *EM*-алгоритма с использованием оператора свертки строились регрессионные модели зависимости выходной характеристики «глубина сферической лунки» от набора параметров сквозной технологии производства автолиста. По комплектным данным методом пошагового регрессионного анализа была построена следующая контрольная модель:

$$y = 11{,}543 - 4{,}264x_1 - 0{,}873x_2 - 0{,}002x_3 + 0{,}004x_4 - \\ - 0{,}020x_5 - 0{,}034x_6 - 0{,}035x_7 - 0{,}034x_8 + 0{,}002x_9, \quad (3)$$

где y — выходная характеристика, мм; x_1 и x_2 — массовые доли элементов химического состава, %, [C] и [Al] соответственно; x_3 — скорость полосы при горячей прокатке, м/мин; x_4 — температура конца горячей прокатки, °C; x_5 ... x_8 — обжатия по клетям стана холодной прокатки: соответственно, в 1-, 2-, 4- и 5-й клетях, %; x_9 — скорость полосы на выходе первой клети стана холодной прокатки, м/мин (свободный член измеряется в мм; коэффициенты при переменных имеют размерности, обратные размерностям соответствующих переменных,

Таблица 1

Фрагмент статистического исследования построенных моделей

m, %	Вид модели	*Norma*	*RSS*	*R*	*F*	χ^2	*p* (χ^2)	*D*	*p* (*D*)
0	(3)	–	61,200	0,820	49,897	–			
1	*A*	2,936	76,700	0,768	34,875	7,462	0,382	0,079	0,410
	B	0,205	61,200	0,820	49,828	0,529	0,999	0,028	1,000
	C	0,202	61,200	0,820	49,828	0,571	0,999	0,028	1,000
10	*A*	14,650	1170,000	–		149,051	0,000	0,724	0,000
	B	1,099	62,400	0,816	48,381	4,867	0,676	0,043	0,971
	C	0,825	62,000	0,817	48,909	1,351	0,987	0,039	0,989
25	*A*	20,342	3110,000	–		227,664	0,000	0,961	0,000
	B	3,539	90,000	0,720	26,096	19,107	0,008	0,079	0,410
	C	1,916	67,000	0,801	43,454	9,983	0,190	0,059	0,768
40	*A*	17,296	6580,000	–		242,561	0,000	0,984	0,000
	B	11,881	1440,000	–		172,694	0,000	0,838	0,000
	C	1,682	64,800	0,806	44,841	5,368	0,615	0,047	0,938

Примечания. *A* — модель, рассчитанная по данным, заполненным безусловными средними; *B* — по данным, заполненным условными средними; *C* — в результате работы *EM*-алгоритма; *m* — процент искусственно вводимых пропусков; *Norma* — норма разности векторов оценок параметров; *RSS* — остаточный квадрат; *R* — коэффициент множественной корреляции; *F* — значение статистики критерия Фишера; χ^2 — значение статистики критерия χ^2; *D* — значение статистики критерия Колмогорова — Смирнова. Доверительная вероятность обозначена символом "*p*". Число групп для расчета значения критерия χ^2 — восемь, получено по формуле Штургеса [4]. При проверке адекватности моделей регрессии использовалось критическое значение критерия Фишера $F_{\text{кр. } 9,\ 244,\ p\ =\ 0{,}99} = 2{,}481$.

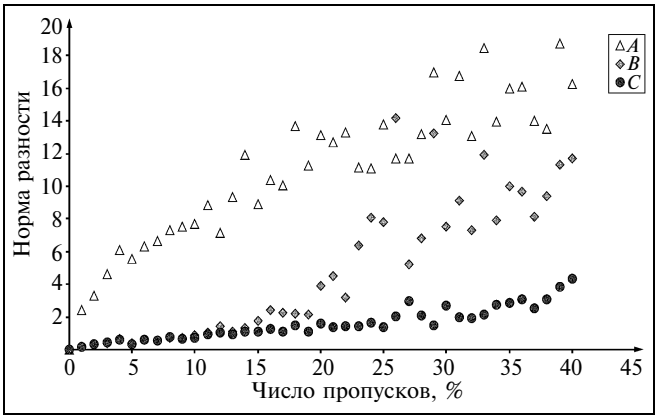

Рис. 1. Зависимость значений норм разностей векторов оценок параметров от числа пропусков

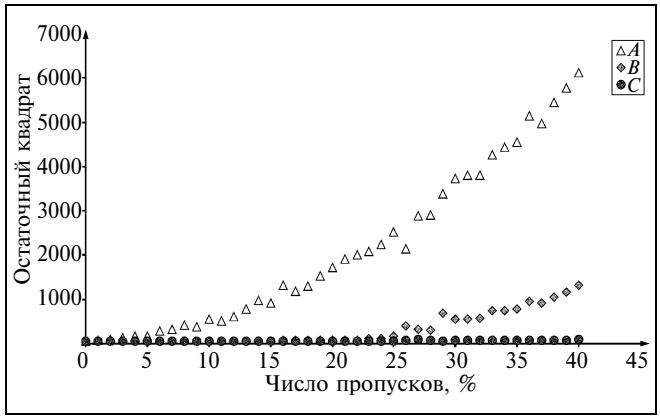

Рис. 2. Зависимость остаточных квадратов от числа пропусков

умноженные на мм). Общий квадрат для модели (3) составил 186,739.

В табл. 1 приведен фрагмент статистического исследования для различных моделей и нескольких случаев искусственно полученных данных с пропусками. В соответствии с результатами, модели вида *C* наиболее устойчивы к пропускам в данных и близки к контрольной модели. Рассчитанные по ним значения норм разностей векторов оценок параметров являются минимальными из соответствующих значений для моделей вида *B* и *C*; остаточные квадраты, коэффициенты корреляции оптимальны и менее всего отличаются от соответствующих значений для контрольной модели. Модели вида *A* наиболее неустойчивы. Уже при 1 % пропусков остаточный квадрат, рассчитанный по модели вида *A*, примерно в 1,25 раза превышает соответствующее значение для контрольной модели, а для большего количества пропусков модели вида *A* неадекватны. Модели вида *B* близки к моделям вида *C* при небольших процентах пропущенных значений, в рассматриваемом случае это справедливо для $m \leqslant 10$ %. Перечисленные выводы подтверждаются общими результатами исследования, которое состояло в искусственной генерации пропусков, числом от 1 до 40 %, с шагом 1 (при 40 % пропусков и выше скорость и результаты работы алгоритма начинают существенно зависеть от структуры пропусков). На каждом шаге набор из трех моделей вида *A*, *B* и *C* генерировался семь раз [4], структура пропусков изменялась случайным образом по равномерному закону, значения рассчитываемых характеристик (норм разностей векторов, остаточных квадратов и др.) усреднялись.

На рис. 1—2 для указанных в табл. 1 видов моделей приведены зависимости норм разностей векторов оценок параметров и остаточных квадратов от числа пропусков.

Как показывают рисунки, модели вида *A* характеризуются недопустимо большими значениями остаточных квадратов, а их векторы оценок параметров значительно отличаются от соответствующего вектора контрольной модели. Для числа пропусков от 30 % и выше модели вида *B* существенно неадекватны.

На практике модель должна быть пригодной для прогноза. Одним из существенных параметров оценки качества предсказания результатов по модели является коэффициент детерминации. На рис. 3 приведена зависимость последнего от числа пропусков, для наглядности значения ординат не усреднялись. Для моделей вида *A* значения коэффициентов детерминации характеризуются большим разбросом в диапазоне, примерно, от 2 до 6 % пропусков и близки к нулю, начиная с 7 %; для моделей вида *B* в диапазоне, примерно, от 18 до 29 % пропусков наблюдается большой разброс, а затем — нулевые значения рассматриваемых коэффициентов; для моделей вида *C* неустойчивость начинает проявляться при значениях около 35 % пропусков и выше.

Реализация первого этапа показала, что оптимальной из рассмотренных является методика *EM*-алгоритма. Поэтому на втором этапе с ее помощью были заполнены пропущенные значения в исходном массиве и построена результирующая модель:

$$y = 10{,}697 - 3{,}983x_1 - 0{,}783x_2 - 0{,}003x_3 + 0{,}004x_4 - \\ - 0{,}017x_5 - 0{,}030x_6 - 0{,}040x_7 - 0{,}027x_8 + 0{,}002x_9, \quad (4)$$

где использованы введенные для модели (3) обозначения. Некоторые характеристики контрольной (3) и результирующей (4) моделей приведены в табл. 2.

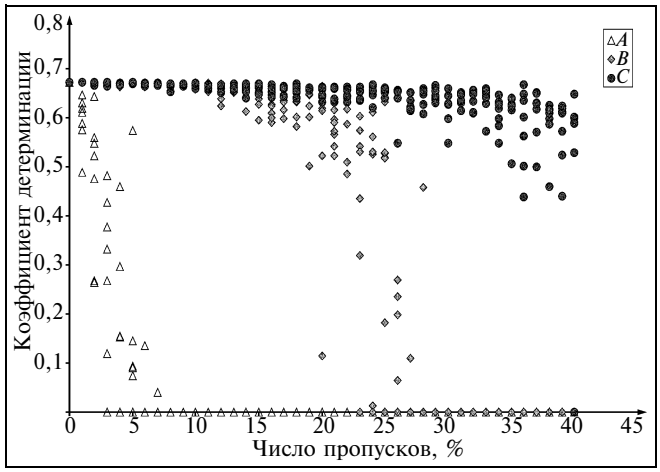

Рис. 3. Зависимость значений коэффициентов детерминации от числа пропусков

Таблица 2

Характеристики контрольной и результирующей моделей

<table>
<tr><td colspan="9">Контрольная модель: n = 254; RSS = 61,158; SS = 186,739; R = 0,820; F = 49,897; $p\ (F)_{9,\ 244}$ = 1,000</td></tr>
<tr><td colspan="9">Значения t-статистик для оценок параметров, $t_{кр.244,\ p\ =\ 0,99}$ = 2,596</td></tr>
<tr><td>t_1</td><td>t_2</td><td>t_3</td><td>t_4</td><td>t_5</td><td>t_6</td><td>t_7</td><td>t_8</td><td>t_9</td></tr>
<tr><td>−2,813</td><td>−1,241</td><td>−5,631</td><td>2,221</td><td>−2,329</td><td>−4,399</td><td>−3,673</td><td>−2,885</td><td>2,562</td></tr>
<tr><td colspan="9">Доверительные интервалы для значимых оценок параметров</td></tr>
<tr><td>x_1</td><td>x_2</td><td>x_3</td><td>x_4</td><td>x_5</td><td>x_6</td><td>x_7</td><td>x_8</td><td>x_9</td></tr>
<tr><td>−7,235</td><td rowspan="2">—</td><td>−0,003</td><td rowspan="2">—</td><td rowspan="2">—</td><td>−0,049</td><td>−0,054</td><td>−0,056</td><td rowspan="2">—</td></tr>
<tr><td>−1,293</td><td>−0,002</td><td>−0,019</td><td>−0,016</td><td>−0,011</td></tr>
<tr><td colspan="9">Результирующая модель: n = 309; m = 2,1 %; $Norma$ = 0,420; RSS = 71,375; SS = 231,389; R = 0,832; F = 63,445; $p\ (F)_{9,\ 299}$ = 1,000.</td></tr>
<tr><td colspan="9">Значения t-статистик для оценок параметров, $t_{кр.299,\ p\ =\ 0,99}$ = 2,592</td></tr>
<tr><td>t_1</td><td>t_2</td><td>t_3</td><td>t_4</td><td>t_5</td><td>t_6</td><td>t_7</td><td>t_8</td><td>t_9</td></tr>
<tr><td>−2,789</td><td>−1,336</td><td>−6,386</td><td>2,856</td><td>−2,128</td><td>−3,980</td><td>−4,264</td><td>−2,357</td><td>3,200</td></tr>
<tr><td colspan="9">Доверительные интервалы для значимых оценок параметров</td></tr>
<tr><td>x_1</td><td>x_2</td><td>x_3</td><td>x_4</td><td>x_5</td><td>x_6</td><td>x_7</td><td>x_8</td><td>x_9</td></tr>
<tr><td>−6,783</td><td rowspan="2">—</td><td>−0,003</td><td>0,001</td><td rowspan="2">—</td><td>−0,044</td><td>−0,058</td><td rowspan="2">—</td><td>0,001</td></tr>
<tr><td>−1,184</td><td>−0,002</td><td>0,007</td><td>−0,015</td><td>−0,021</td><td>0,004</td></tr>
<tr><td colspan="9">Примечание. В таблице использованы введенные ранее обозначения; n — число наблюдений; SS — общий квадрат.</td></tr>
</table>

Модели (3) и (4) адекватны и пригодны для прогнозирования и управления. Заполнение пропущенных значений привело к смещению границ доверительных интервалов для значимых оценок параметров, а также переходу отдельных параметров в разряд значимо отличающихся от нуля и, наоборот, с заданной доверительной вероятностью. Такие колебания обусловлены близостью к нулю некоторых оценок параметров (соответствующие значения t-статистик мало отличаются от критических) и оказывают незначительное влияние на конечный результат. В рассматриваемой модели на выходную характеристику в большей степени влияют скорость горячей прокатки и обжатия во второй и четвертой клетях стана холодной прокатки, об этом свидетельствуют значения соответствующих t-статистик для обеих моделей. В целом, модель (4) улучшает характеристики исходной модели, не изменяя ее структуру. На это указывает малое значение нормы разности векторов оценок параметров контрольной и результирующей моделей. Даже при небольшом общем проценте пропусков (2,1 %) заметно увеличилось значение статистики критерия Фишера (от 49,897 до 63,445) и незначительно вырос коэффициент множественной корреляции (от 0,820 до 0,832).

ЗАКЛЮЧЕНИЕ

Анализ данных с искусственно введенными пропусками показал максимальную устойчивость характеристик моделей, построенных путем реализации *EM*-алгоритма, к числу и структурам пропусков, по сравнению с характеристиками моделей, рассчитанных по данным, пропуски в которых заполнены с помощью безусловных средних и метода Бака. Применение *EM*-алгоритма к статистическому моделированию зависимости одного из показателей качества листовой стали от ряда факторов технологии ее производства по неполным данным показало целесообразность такого подхода для уточнения оценок параметров и совершенствования существующей модели.

ЛИТЕРАТУРА

1. *Литтл Р. Дж. А., Рубин Д. Б.* Статистический анализ данных с пропусками. — М.: Финансы и статистика, 1990. — 336 с.
2. *Бард Й.* Нелинейное оценивание параметров. — М.: Статистика, 1979. — 349 с.
3. *Гайдышев И.* Анализ и обработка данных: специальный справочник. — СПб.: Питер, 2001. — 752 с.
4. *Львовский Е. Н.* Статистические методы построения эмпирических формул. — М.: Высшая школа, 1988. — 239 с.

☎ *(4742) 46-53-54;*

e-mail: kuznetsov@stu.lipetsk.ru

Статья представлена к публикации членом редколлегии А. А. Дорофеюком. ❑