УДК 681.3

Долгова Т. А., доцент; Бохан Е. В., студент.; Сергеенко Г. А., студент

РЕГРЕССИОННЫЙ ПОДХОД ПРИ ОПРЕДЕЛЕНИИ ЕМКОСТИ НАБОРНОЙ СТРОКИ

Regressive dependence of quantity of symbols in a line from its format, size and sets of a font is considered. Initial data are received by carrying out full factorial experiment. Recommendations for defining the capacity of a type-setting line are given.

Емкость наборной строки — число символов шрифта в строке определенного формата — является одним из основных факторов, определяющих объем издания. Расчет объема необходим для определения экономических показателей выпуска продукции. Кроме того, при выборе параметров верстки следует учитывать гигиенические требования к шрифтовому оформлению. Так, для определенных видов изданий, например для детской литературы, емкость строки формата 1 кв. не должна превышать заданных значений.

Для теоретического определения количества символов в строке следует перевести формат наборной строки из квадратов в миллиметры (1 кв. = 18,05 мм); полученную длину L разделить на среднеуточненную ширину (математическое ожидание ширины) знака для конкретной гарнитуры и кегля.

Среднеуточненную ширину E принято вычислять по формуле

$$E = \sum_{i=1}^{N} l_i p_i , \qquad (1)$$

где l_i — ширина знаков русского языка, мм; p_i — удельная частота встречаемости знаков; N — общее количество знаков, включая знаки препинания, цифры, прописные и строчные буквы и пробел. Суммарная вероятность появления всех знаков равна единице.

Методика определения объема издания, основанная на таких расчетах, позволяет оценить объем издания уже на начальной стадии его допечатной подготовки, что помогает решать задачу выбора оформления.

Компьютерные технологии внесли свои особенности в процесс формирования наборной строки. Значительно изменились метрические характеристики шрифтов [1].

Практически любая программа верстки позволяет увеличивать межбуквенное расстояние (трекинг) и изменять ширину символов. Этой возможностью иногда злоупотребляют. Такое «растягивание» для доведения емкости строки до рекомендуемого значения грубо нарушает рисунок шрифта и не способствует повышению удобочитаемости. Значение емкости строки конкретной гарнитуры предполагает использование стандартных параметров.

Расхождение между расчетным и практическим значением емкости происходит из-за ряда факторов. Невозможно в полной мере учесть правила деления слов на слоги при переносах и особенности выключки строк.

Число символов в строке может также изменяться при использовании различных программных пакетов, различных компьютерных платформ, операционных систем, драйверов выводных устройств. Поэтому число символов в строке можно считать величиной случайной и использовать для нее регрессионные модели.

Целью данной работы является построение математической модели для определения емкости строки набора в зависимости от различных параметров верстки: кегль, гарнитура, ширина полосы набора. В качестве метода планирования эксперимента использован полный факторный эксперимент.

В соответствии с планом эксперименты проводились для различных сочетаний верхнего и нижнего уровней выбранных факторов. В этом случае число опытов $n = 2^3 = 8$.

Для книжно-журнальных изданий кегль набора чаще всего изменяется в пределах от 8 до 12 пт, а формат полосы — от 3,5 до 7 кв. соответственно. В качестве исследуемых гарнитур выбраны CaslonC и Cyrillic University как примеры широкого и узкого текстового шрифта. Емкость первой гарнитуры примем за нижний уровень, второй — за верхний. Среднее количество знаков 10-го кегля в строке форматом 1 квадрат равно 8 для CaslonC и 12,5 для Cyrillic University.

Для сравнения эксперименты проводились на двух компьютерах AMD Sempron с одинаковой операционной системой Windows XP, но с разными моделями установленных принтеров. Использовался научно-технический текст, набранный в текстовом редакторе Word (как известно, этот пакет наиболее чувствителен к драйверу принтера).

На каждом компьютере для своего текста проводилась серия из 8 экспериментов, значения факторов для которых даны в табл. 1, для удобства восприятия второй фактор представлен не количественно, а названием гарнитуры.

В [2] показано, что достаточно точное определение среднеуточненной ширины символа шрифта обеспечивает подсчет числа знаков в 7 произвольно выбранных строках заданного формата. В соответствии с этой методикой и проводился подсчет емкости.

Таблица 1

План эксперимента

№ опыта	Значения факторов		
	Z_1, пт	Z_2	Z_3, кв.
1	8	CaslonC	3,5
2	12		
3	8	CyrillicUniversity	
4	12		
5	8	CaslonC	7
6	12		
7	8	CyrillicUniversity	
8	12		

Каждый опыт перепроверялся несколько раз, причем в параллельных опытах всякий раз использовались разные 7 строк анализируемого текста. Для каждого опыта найдены средние значения — $\bar{y}_i$. Число символов принято округлять в меньшую сторону. Но для построения регрессионного уравнения использовались значения с точностью до десятых, с тем, чтобы округлять конечные значения, полученные из регрессионных уравнений.

Полученные значения для каждой серии экспериментов (каждого компьютера) приведены в табл.2. На основе этих данных проведен регрессионный анализ.

Проверка однородности дисперсии проводилась по критерию Кохрена

$$G = \frac{S^2_{\max}}{\sum_{i=1}^{n} S_i^2}, \qquad (2)$$

где выборочные дисперсии для каждого i-го опыта, проверенного m раз, равны:

$$S_i^2 = \frac{\sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2}{m-1}. \qquad (3)$$

Расчетные значения критерия Кохрена $G = 0{,}226$ и $G = 0{,}438$. В первой серии экспериментов проводилось по 4 параллельных опыта, во второй — по 3 и табличные значения G_P взяты для соответствующих степеней свободы: $G_P(8;\ 4-1) = 0{,}231$ и $G_P(8;\ 3-1) = 0{,}816$. Для каждого компьютера выполняется условие $G < G_P$, т. е. дисперсии однородны.

Коэффициенты регрессионного уравнения рассчитаны по формуле

$$b_l = \frac{1}{n}\sum_{i=1}^{n} x_{li}\bar{y}_i, \qquad (4)$$

где x_{li} — безразмерные значения факторов, равные –1 и 1 для нижнего и верхнего уровня z_i соответственно.

Получены следующие коэффициенты регрессионного уравнения для 1-го компьютера:

$$b_0 = 55{,}18;\ b_1 = -11{,}85;\ b_2 = -11{,}4;$$
$$b_3 = 18{,}93;\ b_4 = 2{,}23;\ b_5 = 3{,}85;$$
$$b_6 = -3{,}65;\ b_7 = 0{,}73;$$

для 2-го компьютера:

$$b_0 = 58{,}04;\ b_1 = -12{,}04;\ b_2 = 12{,}46;$$
$$b_3 = 18{,}96;\ b_4 = -2{,}63;\ b_5 = -3{,}79;$$
$$b_6 = 3{,}88;\ b_7 = -0{,}88.$$

Проверка значимости коэффициентов проводилась по критерию Стьюдента:

$$t_j = \frac{|b_j|}{S_{b_j}}. \qquad (5)$$

Значения критериев для 1-го компьютера:

$$t_0 = 209{,}8;\ t_1 = 45{,}1;\ t_2 = 43{,}3;$$
$$t_3 = 71{,}9;\ t_4 = 8{,}5;\ t_5 = 14{,}6;$$
$$t_6 = 13{,}9;\ t_7 = 2{,}8;$$

для 2-го компьютера

$$t_0 = 223{,}24;\ t_1 = 46{,}32;\ t_2 = 47{,}92;$$
$$t_3 = 72{,}91;\ t_4 = 10{,}1;\ t_5 = 14{,}59;$$
$$t_6 = 14{,}9;\ t_7 = 3{,}37.$$

Таблица 2

Результаты экспериментов

№ опыта	1-й компьютер						2-й компьютер				
	Результаты параллельных опытов				$\bar{y}_i$	$\hat{y}_i$	Результаты параллельных опытов			$\bar{y}_i$	$\hat{y}_i$
1	34	35	35	34	34,5	33,8	37	37	37	37	36,1
2	22	22	23	21	22	22,7	23	25	24	24	24,9
3	54	54	54	52	53,5	54,2	58	58	57	57,7	58,5
4	34	36	35	35	35	34,3	38	37	38	37,7	36,8
5	71	70	71	70	70,5	71,2	72	74	73	73	73,9
6	46	47	46	46	46,3	45,6	48	48	49	48,3	47,5
7	108	107	109	107	107,8	107,1	113	112	113	112,7	111,8
8	72	72	72	71	71,8	72,5	74	75	73	74	74,9

Сравнение с табличными значениями $t_p(4-1) = 3{,}182$ и $t_p(3-1) = 4{,}303$ для каждой экспериментальной серии показало, что в обоих случаях седьмой коэффициент незначимый.

Полученные регрессионные уравнения имеют следующий вид:

– для 1-го компьютера

$$y = 55{,}18 - 11{,}85x_1 + 11{,}4x_2 + 18{,}93x_3 + \\ - 2{,}23x_1x_2 - 3{,}85x_1x_3 + 3{,}63x_2x_3; \quad (6)$$

– для 2-го компьютера

$$y = 58{,}04 - 12{,}04x_1 + 12{,}46x_2 + 18{,}96x_3 - \\ - 2{,}63x_1x_2 - 3{,}79x_1x_3 + 3{,}88x_2x_3. \quad (7)$$

Проверка адекватности регрессионного уравнения проводилась по критерию Фишера

$$F = \frac{S^2_{\text{ад}}}{S^2_{\text{воспр}}}, \quad (8)$$

для которого необходимые дисперсии адекватности и воспроизводимости вычисляются по формулам

$$S^2_{\text{ад}} = \frac{\sum_{i=1}^{n} (\bar{y}_i - \bar{y}_i)^2}{n-k}, \quad (9)$$

$$S^2_{\text{воспр}} = \frac{\sum_{i=1}^{n} S_i^2}{n}. \quad (10)$$

Теоретические значения емкости наборной строки $\bar{y}_i$, полученные подстановкой в регрессионные уравнения значений факторов, используемых в соответствующих опытах, приведены в табл. 2.

Для первой экспериментальной серии расчетный критерий Фишера $F = 28{,}354$ меньше табличного $F_p(4-1;\ 8-7) = 215{,}7$; для второй серии $F = 11{,}307$ также меньше табличного $F_p(3-1;\ 8-7) = 18{,}5$.

Таким образом, регрессионные уравнения отвечают условию адекватности.

Применим их для расчета емкости строки, набранной 10-м кеглем для формата 5 кв., используя те же гарнитуры. Перевод факторов в безразмерные координаты производится по следующей формуле

$$x_j = \frac{z_j - z_j^o}{\Delta z_j}, \quad (11)$$

где j — номер фактора, а значение нулевого уровня (центр плана) и интервала варьирования соответственно равны

$$z_j^o = \frac{1}{2}\left(z_j^{\max} + z_j^{\min}\right), \quad (12)$$

$$\Delta z_j = \frac{1}{2}\left(z_j^{\max} - z_j^{\min}\right). \quad (13)$$

Выбранное значение кегля соответствует центру плана и $x_1 = 0$; для формата полосы в соответствии с формулами (11)–(13) $x_3 = -0{,}14$; $x_2 = \pm 1$.

Полученные с помощью уравнений (6) и (7) значения для этих параметров вновь сравнивались с практическими результатами, которые, как и ранее, перепроверялись в параллельных опытах. Округленные практические и теоретические значения приведены в табл. 3 для каждого компьютера. Следует заметить, что расчетные значения нужно округлять по правилам математики. Как видно, погрешность не превысила 5%.

Но каждая из найденных математических зависимостей строилась и проверялась на конкретном компьютере. Уже первоначальный анализ экспериментальных данных из таблицы 2 позволяет увидеть разницу в результатах опытов для одних и тех же значений факторов.

Коэффициенты уравнений (6) и (7), как и следовало ожидать, тоже отличаются. Рассмотрим их подробнее.

Коэффициенты b_0, а соответственно и расчетное число знаков в центре плана ($x_i = 0$), отличаются примерно на 3.

Линейные эффекты первого и третьего факторов (кегль и формат наборной строки), а также их эффект взаимодействия в (6) и (7) очень близки.

Таблица 3

Емкость наборной строки (формат 5 кв., кегль 10 пт)

№ компьютера	Гарнитура	Среднее практическое значение	Расчетное значение (без округления)	Относительная погрешность
1	CaslonC	40	41 (41,2)	2,5%
	Cyrillic University	61	64 (63,8)	4,9%
2	CaslonC	42	43 (43,4)	2,4%
	Cyrillic University	65	67 (67,2)	3,1%

Разница между соответствующими коэффициентами b_1, b_3 и b_5 первого и второго уравнения составляет несколько сотых.

Если учесть, что $|x_j| \le 1$, то разница во вкладе этих слагаемых в вычисляемую емкость строки практически не заметна.

Коэффициент b_2 во втором уравнении больше почти на единицу. Абсолютное значение коэффициентов, характеризующих взаимодействие x_2 с остальными переменными, в этом уравнении также больше. При некотором сочетании факторов суммарный вклад x_2 во втором уравнении может быть больше на 1–2 знака.

Таким образом, выбор конкретной гарнитуры заметнее влияет на результат, полученный при использовании различного оборудования.

Как известно, шрифты форматов PostScript и TrueType хранятся в закодированном виде, а для воспроизведения символа нужного кегля специальные программы производят масштабирование контура символа и растрирование внутренней области в соответствии с разрешением устройства, на котором отображается текст.

Однако пересчет в абсолютные координаты точек символа происходит с округлением нецелых результатов. При этом могут проявляться некоторые специфические, особенно для малых кеглей, ошибки: нарушение пропорции символа, симметричности и т. д. Эти проблемы решаются с помощью специальных встроенных алгоритмов, которые в общем-то могут привести к небольшому отличию относительных размеров символов разных кеглей [4].

Для сравнения вычислим в соответствии с формулой (1) емкость шрифта для гарнитуры CaslonC, используя методику определения среднеуточненной ширины знака, рассмотренную в [3]. При этом ширину букв l_i найдем с помощью шрифтового редактора FontLab, где при описании символа шрифта используется координатная сетка 1000 × 1000 единиц, что реально соответствует 1 × 1 пт.

Для кегля 10 пт получим E = 2,2 мм. Следовательно, тогда по расчету для строки форматом 1 кв. (L = 18,05 мм) получается 8,2 знака, а для рассматриваемого формата 5 кв. — 41 знак. Как видно (см. табл. 3), это значение совпадает с емкостью строки, определенной с помощью регрессионного уравнения (6).

Если гарнитура и кегль будущего издания известны заранее, расчеты объемов удобнее производить на основании среднеуточненной ширины знака. Построение и использование регрессионного уравнения оправдано, когда встает необходимость выбора варианта полиграфического оформления.

Если на первоначальной стадии для нового или переиздаваемого произведения характеристики макета еще не известны, границы параметров верстки всегда могут быть определены по виду издания. Для типометрических характеристик текстовых шрифтов, стандартных форматов полос набора всегда можно установить нижние и верхние границы.

Серия экспериментов по определению емкости строки и построение регрессионного уравнения не требуют больших временных затрат. Исследование следует проводить на конкретном оборудовании в конкретной программе верстки, которые и будут использованы в дальнейшей работе над изданием.

Применение регрессионного уравнения позволит легко сравнивать объемы издания в печатных листах для различных вариантов оформления, параметры которого могут меняться довольно значительно. В качестве таких параметров можно использовать и отличные от рассмотренных выше. Например, учитывая интерлиньяж, можно сразу рассматривать емкость целой полосы набора. Такой регрессионный подход позволит оптимизировать экономические показатели выпуска.

Литература

1. Волкова Л. А., Ревякова О. Н. Исследование параметров компьютерных шрифтов // Полиграфия. — 2001. — № 1. — С. 42–43.

2. Ревякова О. Н. «Неопознанные» шрифты при моделировании изданий на донаборной стадии // Полиграфия. — 2003. — № 4. — С. 38–39.

3. Долгова Т. А. Учет вертикальной емкости при определении экономичности компьютерных шрифтов // Труды БГТУ. Сер. IX. Издат. дело и полиграфия. — 2003. — Вып. XI. — С. 43–48.

4. Ярмола Ю. А. Компьютерные шрифты — СПб.: BHV — Санкт-Петербург, 1994. — 208 с.