Голубев В.О. Литвинова Т.Е.

Реализация алгоритма построения статистической модели объекта по методу Брандона

Постановка задачи

Статистические модели создают на основании имеющихся экспериментальных данных, снятых на действующем объекте. Задачу формулируют следующим образом: по данной выборке объема *n* (т. е. по заданному числу опытов) построить модель и оценить адекватность ее реальному объекту.

В общем случае современный технологический процесс представляется в виде многомерного объекта, блок-схема которого приведена на рис. 1. На объект действуют вектор входных параметров $\overline{X}$, составляющие которого $\{x_1, x_2, ..., x_l\}$, и вектор управления $\overline{Z}$, составляющие которого $\{z_1, z_2, ..., z_k\}$. Выходные параметры $\{y_1, y_2, ..., y_p\}$ составляют вектор выходных параметров $\overline{Y}$. Общий вид статистической модели многомерного технологического объекта можно записать в виде системы алгебраических уравнений (1) или в векторной форме (2):

$$\begin{cases} y_1 = F_1(x_1, x_2, ..., x_m), \\ y_1 = F_2(x_1, x_2, ..., x_m), \\ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \\ y_p = F_p(x_1, x_2, ..., x_m); \end{cases} \quad (1)$$

$$\overline{Y} = F(\overline{X}), \quad (2)$$

где $\overline{X}$, $\overline{Y}$ – векторы входных и выходных параметров объекта.

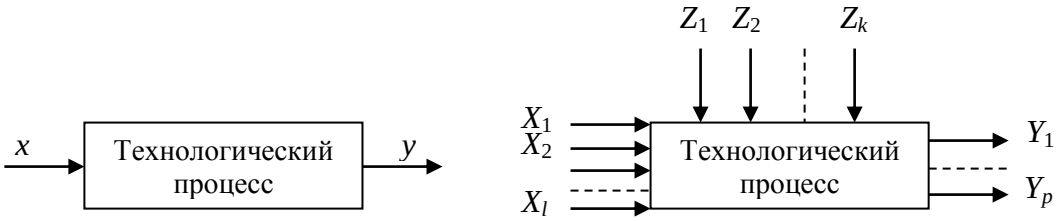
В системе (1) параметры управления учтены как входные параметры $x_{l+1}, x_{l+2}, ..., x_m$ ($m=l+k$).

Построение статистической модели одномерного технологического объекта

На практике часто возникает необходимость создания моделей для одномерного технологического объекта. Блок-схема такого объекта представлена на рис. 1, а модель описывается уравнением (3):

$$y = f(x), \quad (3)$$

Процесс построения статистических моделей состоит из нескольких этапов.



а *б*

Рис. 1. Блок-схема технологических процессов

Определение тесноты связи между переменными

О наличии или отсутствии связи между двумя переменными качественно можно судить по виду поля корреляции [1], а количественно — по величине выборочного коэффициента корреляции, определяемого по формуле

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}, \tag{4}$$

где $\bar{x}$ и $\bar{y}$ – средние значения:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\ ,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\ ,$$

n – число опытов; S_x и S_y – дисперсии.

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\ ,\quad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\ .$$

Для вычисления r_{xy} удобно использовать:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}, \tag{5}$$

Коэффициент корреляции по абсолютной величине не превышает единицы: $-1 \le r_{xy} \le 1$.

Чем ближе абсолютное значение коэффициента $|r_{xy}|$ к единице, тем сильнее линейная связь между величинами. Следует отметить, что коэффициент корреляции одинаково отмечает долю случайности и криволинейность связи между *x* и *y*. Зависимость *x* и *y* может быть близкой к функциональной, но существенно нелинейной; коэффициент корреляции при этом будет значительно меньше единицы.

Объективное определение тесноты связи может быть проведено в результате совместного анализа качественной и количественной оценок.

Выбор вида зависимости

Для определения вида зависимости (3) следует построить эмпирическую линию регрессии (рис. 3.2). Для это весь диапазон изменения x на поле корреляции разбивается на k равных интервалов Δx. Все точки, попавшие в j-й интервал Δx_j, относят к его середине x_j. Для этого определяв частные средние $\bar{y}_j$ для каждого интервала:

$$\bar{y} = \sum_{i=1}^{n_j} y_{j,i} \Big/ n_j \,, \tag{6}$$

где n_j – число точек в интервале Δx_j, причем $\sum_{i=1}^{n_j} n_j = n$ (n – объем выборки).

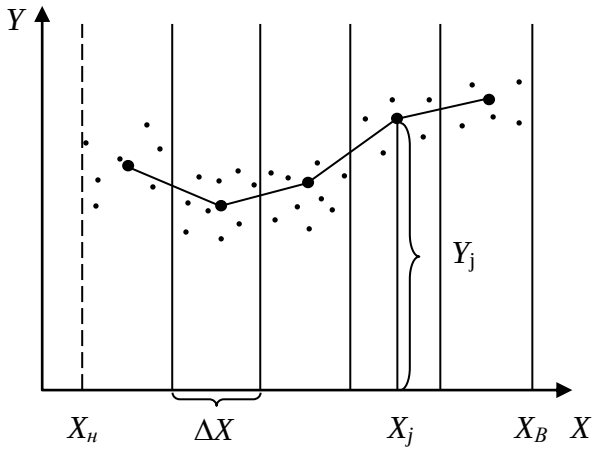


Рис. 3.2. Построение эмпирической линии регрессии

Затем последовательно соединяют точки (x_j, y_j) отрезками прямой. Полученная ломаная называемся эмпирической линией регрессии y по x. По виду эмпирической линии регрессии можно подобрать уравнение регрессии $y = f\{x\}$. На практике чаще всего используют уравнение регрессии в виде:

прямой – $\hat{y} = a_1 x + a_0$;

экспоненты – $\hat{y} = a_2 e^{a_1 x} + a_0$;

параболы – $\hat{y} = a_2 x^{a_1} + a_0$;

полинома n-й степени – $\hat{y} = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0$ и другие.

Уравнение регрессии в общем случае может быть записано как

$$\hat{y} = f(x, a_0, a_1, a_2, ..., a_m), \tag{7}$$

где $\hat{y}$ – расчетное значение функции y; a_j $(j = 0, 1, ..., m)$ – коэффициенты зависимости; x – значение аргумента.

Определение параметров зависимости

Задачу определения параметров уравнения регрессии (7) $a_0, ..., a_m$ решают обычно с использованием метода наименьших квадратов (МНК) [5, 7]. Согласно методу наилучшими значениями параметров будут те, при которых сумма квадратов отклонений расчетных величин $\hat{y}$ от экспериментальных будет наименьшей, т. е.

$$\Phi = \sum_{i=1}^{n} \left(y_{эi} - f(x_i, a_0, a_1, a_2, ..., a_m)\right)^2 = \min , \tag{8}$$

Условием минимума функции является равенство нулю ее частных производных по всем аргументам:

$$\partial\Phi / \partial a_0 = 0; \partial\Phi / \partial a_1 = 0; ...; \partial\Phi / \partial a_m = 0 . \tag{9}$$

В теории метода система (9) называется нормальной системой: для нее число уравнений равно числу неизвестных. Для рассматриваемой функции Φ (8) условием минимума будет

$$\begin{cases} \sum_{i=1}^{n} 2\left(y_i - f(x_i, a_0, ..., a_m)\right)\dfrac{\partial f(x_i, a_0, ..., a_m)}{\partial a_0} = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \sum 2\left(y_i - f(x_i, a_0, ..., a_m)\right)\dfrac{\partial f(x_i, a_0, ..., a_m)}{\partial a_m} = 0. \end{cases} \tag{10}$$

После преобразований нормальная система принимает вид

$$\begin{cases} \sum_{i=1}^{n} y_i \dfrac{\partial f(x_i, a_0, ..., a_m)}{\partial a_0} - \sum_{i=1}^{n} f(x_i, a_0, ..., a_m)\dfrac{\partial f(x_i, a_0, ..., a_m)}{\partial a_0}, \\ \sum_{i=1}^{n} y_i \dfrac{\partial f(x_i, a_0, ..., a_m)}{\partial a_1} - \sum f(x_i, a_0, ..., a_m)\dfrac{\partial f(x_i, a_0, ..., a_m)}{\partial a_1}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \sum_{i=1}^{n} y_i \dfrac{\partial f(x_i, a_0, ..., a_m)}{\partial a_m} - \sum_{i=1}^{n} f(x_i, a_0, ..., a_m)\dfrac{\partial f(x_i, a_0, ..., a_m)}{\partial a_m}. \end{cases} \tag{10a}$$

Решение системы алгебраических уравнений (а) позволит определить значения коэффициентов для конкретного вида зависимости $f(x_i, a_0, a_1, ..., a_m)$. Полученную зависимость называют уравнением регрессии.

Выражения для определения коэффициентов уравнений регрессии, для зависимостей $\hat{y} = f(x_i, a_0, a_1)$, часто встречающихся в практике при создании статистических моделей, могут быть получены по методу наименьших квадратов путем решения системы (а). В случае программной реализации, когда вид уравнения регрессии заранее не известен, вместо метода наименьших квадратов могут быть использованы

методы условной оптимизации, в частности, метод штрафных функций [6].

Проверка адекватности уравнения регрессии

Адекватность уравнения проверяют по критерию Фишера:

$$F = S_{ост}^2 / S_в^2, \tag{11}$$

где $S_{ост}^2$ – остаточная дисперсия, определяющая разброс экспериментальных данных относительно уравнения регрессии; $S_в^2$ – дисперсия воспроизводимости, определяющая величину случайной ошибки.

Значение $S_{ост}^2$ вычисляют по формуле [7]:

$$S_{ост}^2 = \frac{\sum_{i=1}^{n} \left(y_i - f(x_i, a_0, ..., a_m)\right)^2}{n-(m+1)}, \tag{12}$$

где $n-(m+1) = f_1$ – число степеней свободы, определяемое как разность количества опытных точек n и числа параметров $a_0, ..., a_m$, оцененных по этим же точкам.

Значение дисперсии воспроизводимости находят на стадии предварительного анализа экспериментальных данных [5]. Для этого используют зависимость

$$S_в^2 = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}{n-1}, \tag{13}$$

где $n-1 = f_2$ – число степеней свободы знаменателя.

Определив расчетное значение критерия Фишера по формуле (11), сравнивают его с табличным $F_т$. Если $F_т$ больше F для выбранных уровня значимости α и чисел степеней свободы f_1 и f_2, то уравнение регрессии адекватно. Математическая модель в виде уравнения регрессии может быть использована для практических целей (для расчета, решения задач оптимизации, управления и т. п.).

Если $F_т$ меньше F, то уравнение неадекватно. В этом случае нужно выбрать другой вид зависимости между величинами x и y и построить новую модель. В случае отсутствия данных для определения воспроизводимости процесса при определении адекватности модели на практике используют оценки адекватности – корреляционное отношение η и среднюю относительную ошибку ε:

$$\eta = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_{эi} - \hat{y})^2}{\sum_{i=1}^{n} (y_{эi} - \bar{y})^2}}; \tag{14}$$

$$\varepsilon = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_{эi} - \hat{y}}{y_{эi}}\right|,$$

где $y_э$, $\hat{y}$, $\bar{y}$ – экспериментальное, расчетное и среднее значения выходного параметра соответственно.

3.3.7. Построение статистической модели многомерного технологического объекта

Для построения модели многомерного технологического объекта (рис. 1) в настоящее время существуют несколько методов. Можно использовать метод множественной корреляции, метод группового учета аргументов [1], метод главных компонент [2], метод Брандона и др. Однозначно отдать предпочтение одному из методов нельзя, поскольку каждый из них связан с особенностями конкретного технологического объекта.

В работе для построения статистической модели использован метод Брандона.

Сущность метода заключается в следующем. Предполагается, что функция $F_i(x_1, x_2, ..., x_m)$ в системе (1) является произведением функций от входных параметров, т. е.

$$\hat{y}_i = \bar{y} f_1(x_1) f_2(x_2) \ldots f_m(x_m) \qquad (15)$$

или в более удобной форме:

$$\hat{y}_i = \bar{y}\prod_{k=1}^{m} f_k(x_k), \qquad (16)$$

где $\hat{y}$ – расчетное значение i-го выходного параметра; $\bar{y} = \sum_{j=1}^{n} y_{эi} \Big/ n$ – средняя величина экспериментальных значений i-го выходного параметра; n – количество опытов в исходной выборке.

При использовании метода Брандона большое значение имеет порядок следования функций в уравнении (15). Чем больше влияние оказывает фактор на выходной параметр, тем меньшим должен быть его порядковый номер в указанном уравнении. Поэтому задача построения модели разбивается на несколько этапов.

Ранжирование влияющих факторов

Оценить степень влияния k-го фактора на выходной параметр можно по величине частного коэффициента множественной корреляции [3]:

$$r_{yx_k / x_1, x_2, ..., x_m} = \frac{D_{1k}}{\sqrt{D_{11} D_{kk}}}, \qquad (17)$$

где $r_{yx_k / x_1, x_2, ..., x_m}$ – величина частного коэффициента корреляции, учитывающая влияние k-го фактора на выходной параметр y при условии, что влияние всех прочих факторов исключено; D – определитель матрицы, построенной из парных коэффициентов корреляции. Матрица имеет вид

$$R = \begin{bmatrix} 1 & r_{yx_1} & r_{yx_2} & r_{yx_3} & ... & r_{yx_m} \\ r_{yx_1} & 1 & r_{x_1x_2} & r_{x_1x_3} & ... & r_{x_1x_m} \\ r_{yx_2} & r_{x_2x_1} & 1 & r_{x_2x_3} & ... & r_{x_2x_m} \\ r_{yx_3} & r_{x_3x_1} & r_{x_3x_2} & 1 & ... & r_{x_3x_m} \\ r_{yx_4} & ... & ... & ... & ... & ... \\ r_{yx_m} & r_{x_mx_1} & r_{x_mx_2} & r_{x_mx_3} & ... & 1 \end{bmatrix}, \quad (18)$$

D_{1k} – определитель матрицы с вычеркнутыми первой строкой и k-м столбцом; D_{11}, D_{kk} – определитель матрицы с вычеркнутыми первой и k-й строками и k-ми столбцами соответственно.

При переходе от парных коэффициентов корреляции к частным может существенно измениться не только величина коэффициента корреляции, но и знак.

Порядок расположения влияющих факторов в уравнении (15) определяют в соответствии с убыванием величины частных коэффициентов корреляции. Следует иметь в виду, что коэффициент корреляции – чисто статистический показатель и не содержит предположения, что изучаемые величины находятся в причинно-следственной связи. Подобные предположения должны проверяться экспериментально.

Выбор вида зависимости и построение статистической модели

В уравнении (15) каждая из функций $f_1(x_1), ..., f_m(x_m)$ принимается либо линейной, либо нелинейной (степенной, показательной, экспоненциальной и т. д.). Прежде чем определять вид первой зависимости, следует представить исходные экспериментальные значения выходного параметра в каждом опыте $y_{эi}$ в безразмерной форме $y_{э0j}$:

$$y_{э0\,j} = y_{эj} / \bar{y},$$

где $\bar{y}$ – средняя величина выходного параметра.

Таким образом, исходными данными для поиска первой зависимости будут нормированные значения вектора выходных параметров $\bar{y}_0$ и опытные значения первого влияющего фактора. Поиск зависимости $\hat{y}_1 = f_1(x_1)$, где $\hat{y}_1$ – расчетные значения, осуществляется по той же методике, что и при построении модели одномерного технологического объекта (4, 6).

Выбрав зависимость $\widehat{y}_1 = f(x_1)$, определяют остаточный показатель $y_{э1}$ для каждого наблюдения:

$$y_{э1} = y_{э0} / f_1(x_1).$$

Предполагая, что $y_{эi}$ не зависит от x_1, а зависит от x_2, ..., x_m, выбирают зависимость от второго фактора. Исходные данные для поиска – остаточный показатель $y_{э1}$ и опытные значения второго фактора. Получив расчетную зависимость $\widehat{y}_2 = f_2(x_2)$, находят остаточный показатель $y_{эi}$ для каждого-наблюдения:

$$y_{э2} = y_{э1} / f_2(x_2).$$

Выполнив аналогичные действия для каждого *k*-го влияющего фактора, получают регрессионную зависимость для рассмотренного выходного параметра. Порядок расположения факторов для этой зависимости определен на этапе ранжирования и отличается от порядка в общем уравнении (15).

Совокупность зависимостей по каждому выходному параметру представляет собой статистическую модель многомерного технологического объекта.

Порядок синтеза статистической модели объекта с использованием ЭВМ

Использование ЭВМ в диалоговом режиме значительно ускоряет процесс синтеза модели по методу Брандона. Построение модели происходит в несколько этапов.

1. Ранжирование влияющих факторов:

а) определение коэффициентов парной корреляции;

б) построение исходной матрицы *D* для определения частных коэффициентов корреляции;

в) вычисление значений частных коэффициентов;

г) анализ результатов (проверка причинно-следственных связей; ранжирование факторов).

2. Выбор зависимостей выходных параметров от влияющих факторов:

а) получение одной или нескольких гипотез о виде расчетной зависимости;

б) идентификация параметров каждой расчетной зависимости методом МНК и определение оценок адекватности;

в) анализ результатов, дискриминация гипотез, окончательный выбор зависимости;

г) проверка адекватности модели.

Компьютерная реализация изложенного алгоритма синтеза статистической модели объекта методом Брандона выполнена в среде Matlab 7.0 (14 релиз). Интерфейс программы «Brandon_Analyze» в порядке выполнения расчета изображен на рис. 3–8. Листинг программы приведен в Приложении 1.

Программа дает возможность осуществлять контроль над ходом расчета на всех его стадиях. Результаты автоматического ранжирования факторов по результатам компьютерного расчета могут быть приняты, либо изменены по желанию пользователя. Предусмотрена возможность исключения из рассмотрения факторов, оказывающих незначимое влияние на функцию отклика (решение принимается исследователем). В стандартном пакете зависимостей содержится 20 наиболее распространенных функций, коэффициенты которых рассчитываются по аналитическим выражениям, полученным при дифференцировании исходного выражения. Пользователь имеет возможность дополнить пакет собственными функциями, чем удобно пользоваться в тех случаях, когда тип зависимости известен заранее, а задача состоит только в определении коэффициентов уравнения регрессии. В этом случае расчет выполняется с использованием оптимизационного модуля системы Matlab, предоставляющего расширенные возможности поиска решения. Затраты времени на составление модели не превышают 10 мин.

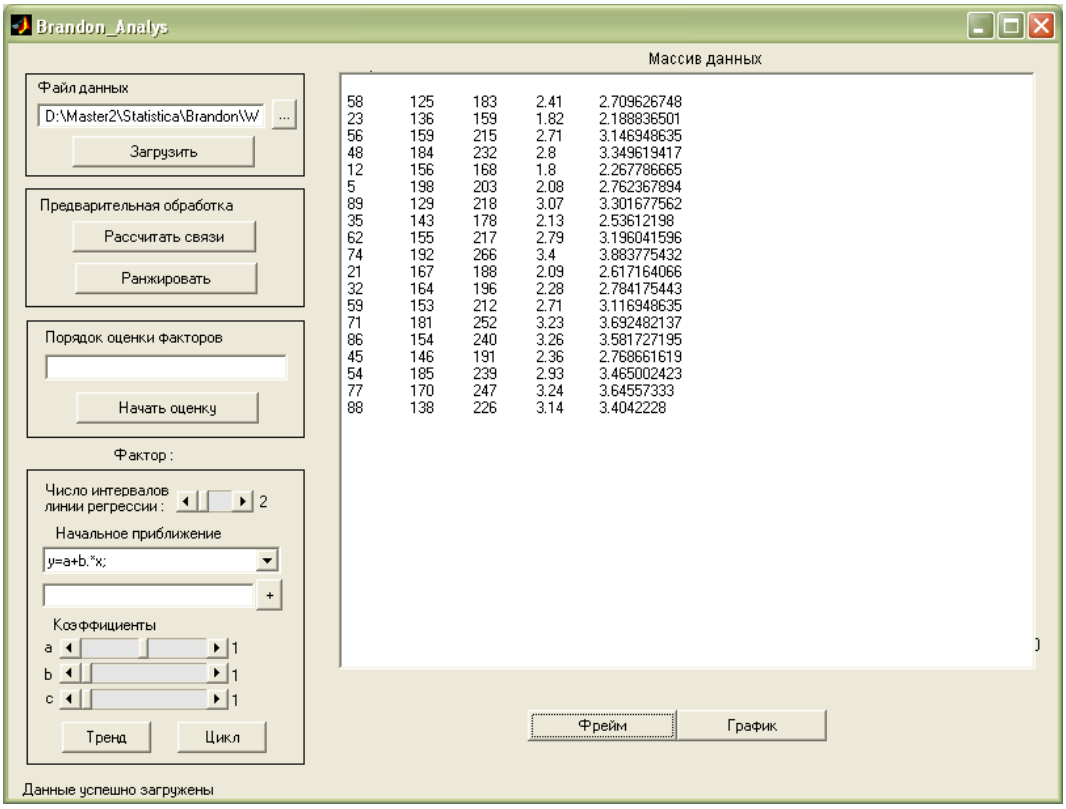


Рис. 3. Загрузка экспериментальных данных из файла

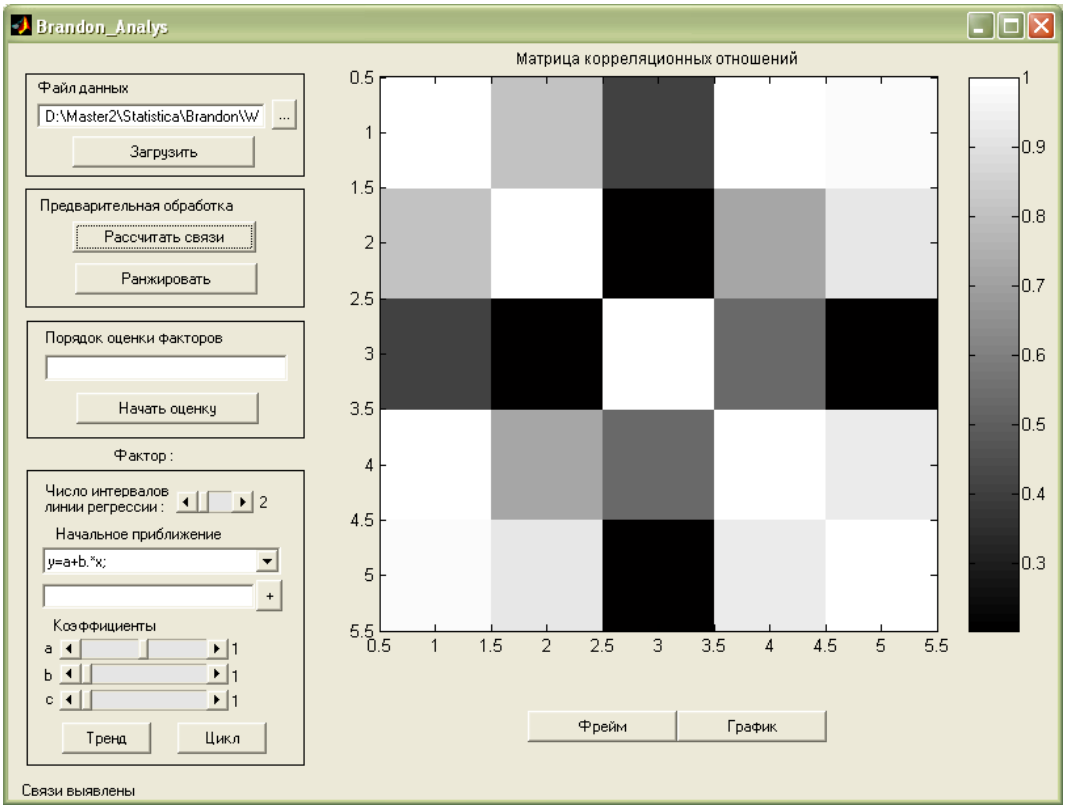


Рис. 4. Составление матрицы корреляционных отношений

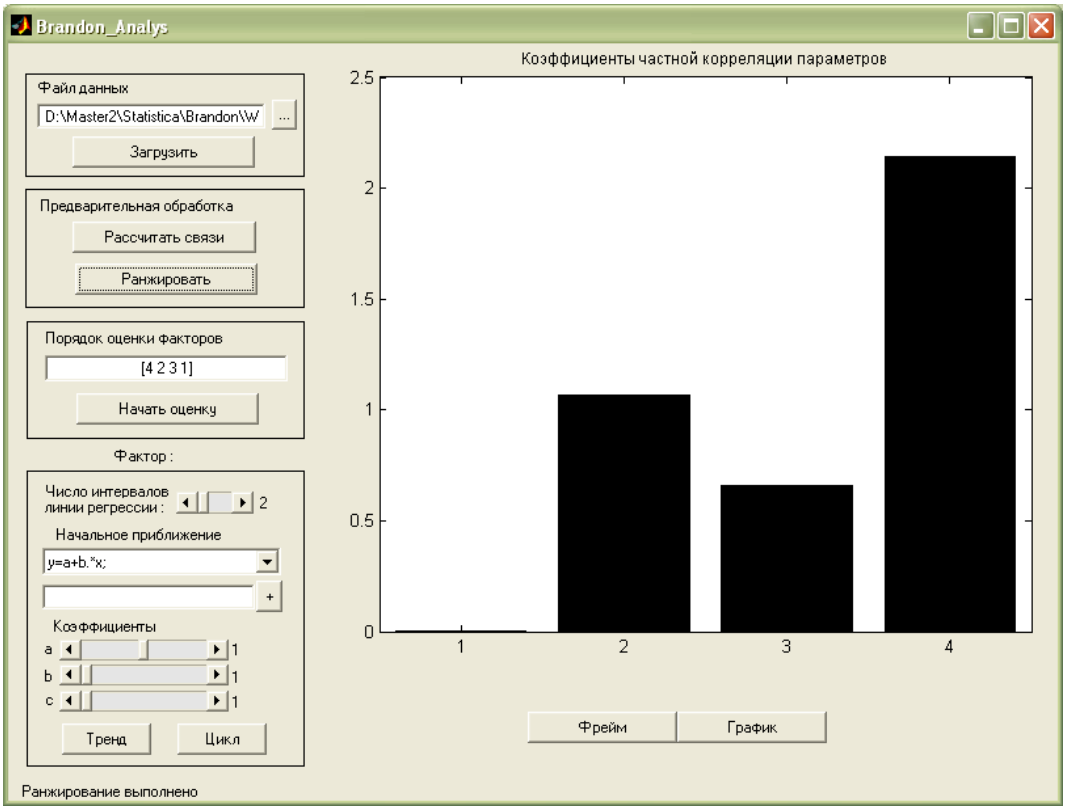


Рис. 5. Расчет коэффициентов частной корреляции параметров

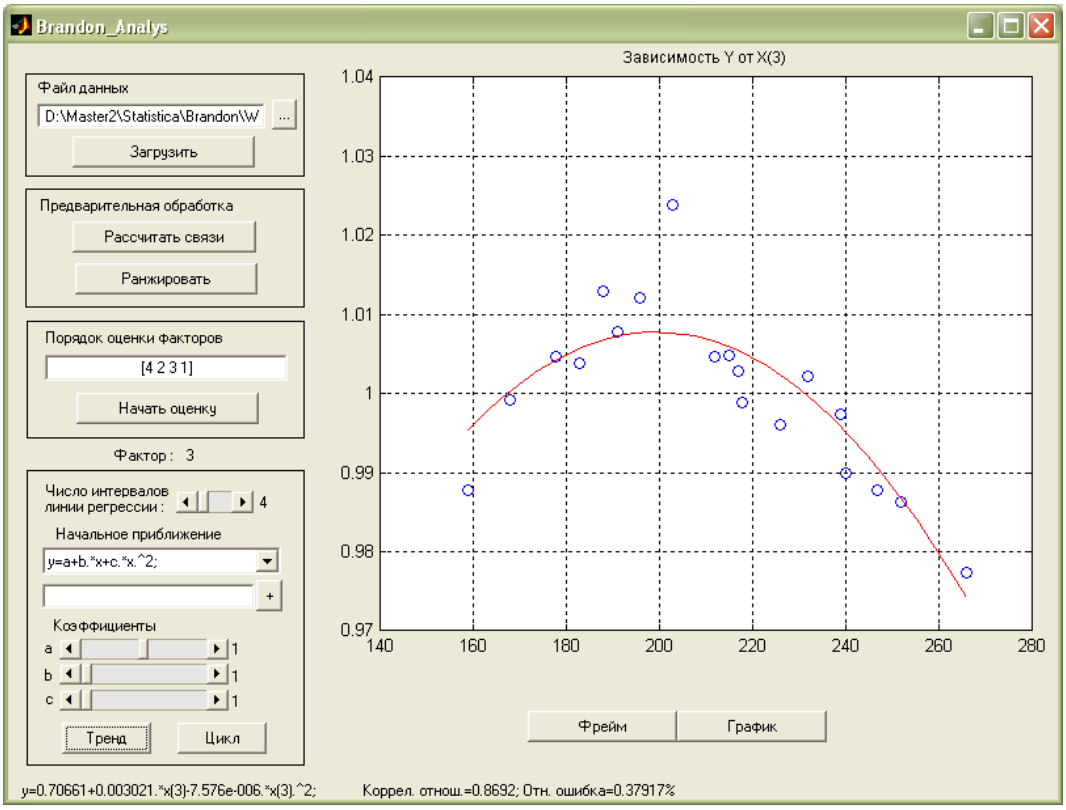


Рис. 6. Выбор вида зависимости и расчет коэффициентов частных регрессий

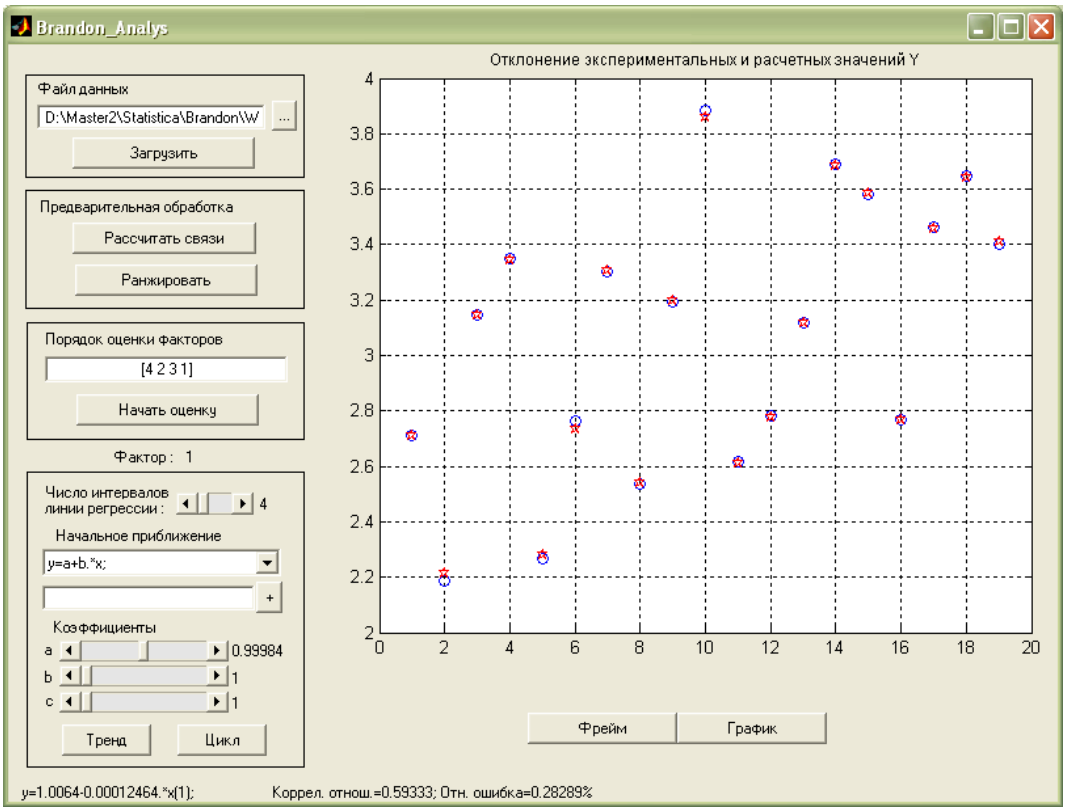


Рис. 7. Демонстрация точности совпадения экспериментальных и расчетных значений функции

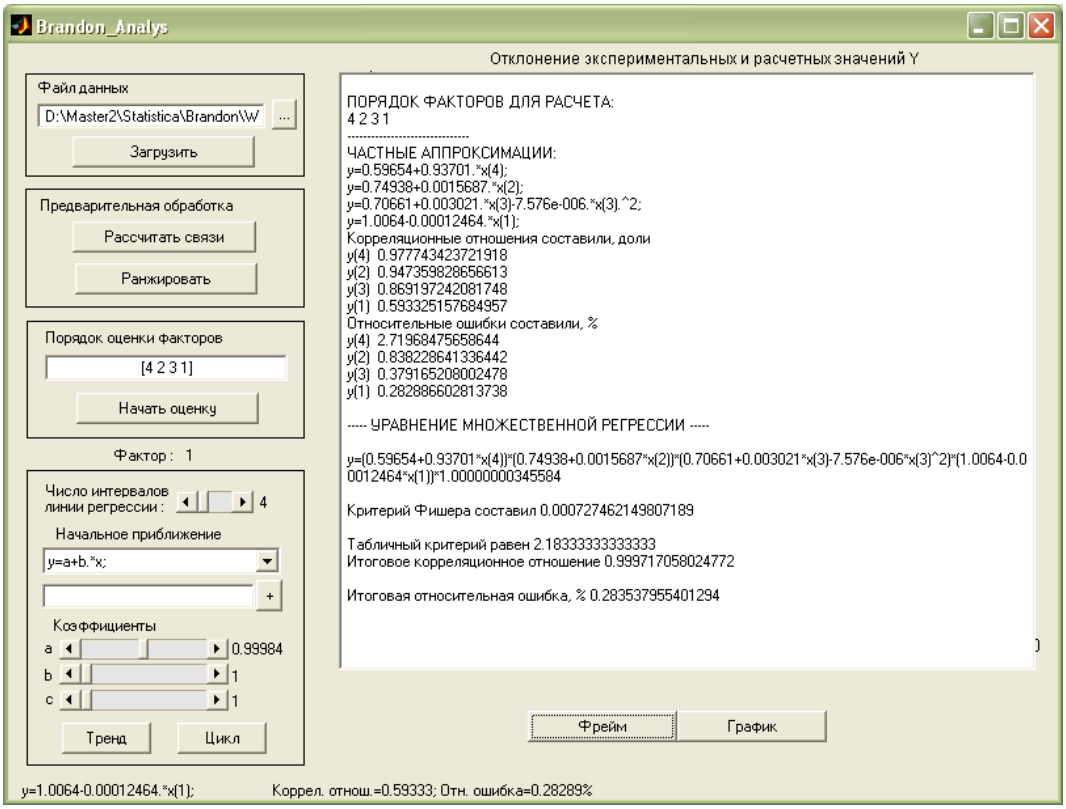


Рис. 8. Вывод результатов расчета во фрейм.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Ахназарова С.Л., Кафаров В.В.* Методы оптимизации эксперимента в химической технологии. М.: Высш. шк., 1985. – 326 с.
2. *Дубров А.М.* Обработка статистических данных методом главных компонент. М.: Статистика, 1978. – 164 с.
3. *Закгрейм А.Ю.* Введение в моделирование химико-технологических процессов. М.: Химия, 1982. – 288 с.
4. *Кафаров В.В.* Методы кибернетики в химии и химической технологии. М.: Химия, 1968. 379 с.
5. *Львовский Г.Н.* Статистические методы построения эмпирических формул. М.: Высш. шк., 1982. –224 с.
6. *Пантелеев А.В., Летова Т.А.* Методы оптимизации в примерах и задачах (учебное пособие). М.: Высшая школа, 2002. – 544 с.
7. *Стьюпер Э., Брюггер У., Джурс П.* Машинный анализ связи химической структуры и биологической активности. М.: Мир, 1982. – 232 с.

