## Лекция 7. Построение и использование статистических моделей

В соответствии с предложенным алгоритмом моделирования на основании статистических моделей после получения набора дискретных экспериментальных данных, например, по матрице, представленной в табл. 3.1, следует этап построения статистической модели. Эта процедура представляет собой решение задачи аппроксимации (описания) дискретной зависимости ($y_1$, $x_1$; $y_2$, $x_2$; …; $y_n$, $x_n$) некоторой непрерывной функцией $y = f(x)$. В общем случае различают три варианта аппроксимации:

1) интерполяция или экстраполяция, которые заключаются в нахождении новых значений функции по известному набору дискретных значений внутри (интерполяция) или за пределами (экстраполяция) заданного интервала независимой переменной. Этот вид аппроксимации основан на том, что аппроксимирующая функция проходит через все рассматриваемые точки (и известные, и определяемые). Например, при линейной аппроксимации все точки принадлежат одной прямой. На практике мы часто используем линейную интерполяцию для определения справочных данных в некоторых промежуточных точках, а линейную экстраполяцию – для графической обработки экспериментальных данных, например при определении ширины запрещенной зоны полупроводника по спектрам оптического поглощения;

2) регрессия (или в более общем смысле – сглаживание), при которой аппроксимирующая функция максимально приближена к описываемым дискретным значениям, но не обязательно проходит через все точки данного набора значений. В качестве функции регрессии при многомерном моделировании зачастую используется полиномиальная зависимость первого порядка в общем виде

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n,$$

где y – выходная переменная; $x_i$ – входные переменные; $b_i$ – коэффициенты уравнения регрессии.

3) фильтрация, которая является более общим случаем процедуры сглаживания и предполагает описание дискретных данных с учетом погрешности измерений и исключением шумовой компоненты.

При моделировании ХТП с использованием статистических моделей построение конкретной модели фактически сводится к заданию общего вида аппроксимирующей зависимости и дальнейшему определению частного вида соответствующего уравнения. Рассмотрим основные этапы построения статистической модели.

1. Выбор общего вида регрессионной зависимости. В рассматриваемом примере для получения набора дискретных экспериментальных данных применялся полнофакторный эксперимент по плану I порядка, предполагающий использование полиномиальных регрессионных уравнений в виде полинома первой степени. Для трехфакторного эксперимента такое уравнение в кодированных переменных имеет вид

$$y = b_0 + b_1Z_1 + b_2Z_2 + b_3Z_3.$$

2. Расчет коэффициентов $b_j$ уравнения регрессии, число которых равно k + 1, с помощью статистической обработки полученного набора дискретных эмпирических значений по формулам

$$b_0 = \frac{\sum_{i=1}^{N} y_i}{N}; \quad b_j = \frac{\sum_{i=1}^{N} x_{ij}y_i}{N},$$

где N – общее число опытов в матрице эксперимента ($N = N_\Sigma = 8$).

Для примера с учетом значений кодированных переменных $x_1$, $x_2$, $x_3$ и полученных значений выходной переменной $y_i$ определим коэффициент $b_2$

$$b_2 = \frac{+1,20 + 3,11 - 2,43 - 3,71 + 0,82 + 2,91 - 2,09 - 3,51}{8} = -0,462.$$

3. Оценка значимости коэффициентов $b_j$ уравнения регрессии в кодированных переменных с использованием t-критерия Стьюдента, который позволяет проверить гипотезу о значимости рассчитанных коэффициентов, то есть отражает существенность или несущественность влияния j-того фактора на выходную переменную с учетом ошибки данного эксперимента.

$$t_j = \frac{|b_j|\sqrt{N}}{\sqrt{S_y^2}},$$

где $S_y^2$ – дисперсия воспроизводимости экспериментальных данных $y_i$ по переменной y, характеризующая ошибку эксперимента.

В случае однородной дисперсии воспроизводимости, когда ошибка эксперимента близка по всем факторам, ее можно оценить по формуле

$$S_y^2 = \frac{\sum_{i=1}^{N_y}(y_i - \bar{y})^2}{f_y},$$

где y – среднее значение выходной переменной в серии опытов по проверке воспроизводимости; $N_y$ – число опытов по проверке воспроизводимости значений выходной переменной (опыты проводятся в отдельной серии при одинаковых значениях факторов, например в центре плана эксперимента, то есть в точке $x_1 = T = 1250$ °C, $x_2 = C = 0,15$ моль/л, $x_3 = K_G = 150$); $f_y$ – число степеней свободы в данном эксперименте, равное числу связей независимых наблюдений ($f_y = N_y - 1$).

Например, если в серии из трех опытов $N_y = 3$ по проверке воспроизводимости данных эксперимента в вышеуказанной точке получены значения скорости роста $y_1 = 2,82$; $y_2 = 2,49$; $y_3 = 2,95$, то при $y_{ср} = 2,75$

$$S_y^2 = \frac{(2,82 - 2,75)^2 + (2,49 - 2,75)^2 + (2,95 - 2,75)^2}{3 - 1} = 0,056.$$

Условием значимости коэффициентов $b_j$ является превышение рассчитанного значения критерия Стьюдента над критическим табличным значением $t_{табл}$, то есть соотношение $t_j > t_{табл}$. Табличное значение $t_{табл}$ определяется по таблицам критических значений статистических критериев при соответствующем значении $f_y$ и выбранном значении уровня значимости α, отражающего уровень строгости испытаний объекта. На практике наиболее часто применяется уровень значимости $\alpha = 0,05$, при котором учитывается вероятность отбраковывания 5 % исследуемых образцов или опытов. В нашем примере для коэффициента $b_2$ рассчитанное значение критерия Стьюдента составляет

$$t_2 = \frac{|-0,462|\sqrt{8}}{\sqrt{0,056}} = 5,5 \succ 4,3.$$

То есть условие значимости коэффициента $b_2$ выполняется, поскольку критическое значение $t_{табл} = 4,3$ (для случая $f_y = 2$ и $\alpha = 0,05$) меньше рассчитанного значения $t_2 = 5,5$.

Окончанием данного этапа построения статистической модели является определение частного вида уравнения регрессии с учетом только значимых коэффициентов и анализ влияния выбранных факторов на выходную переменную. В нашем примере коэффициент $b_3$ оказался незначимым, поскольку $t_3 = 1,7 < 4,3$ (табл. 3.2).

Таблица 3.2

**Данные проверки значимости коэффициентов уравнения регрессии на примере моделирования процесса газофазной эпитаксии кремния**

| Характеристика | Значения для $j$, равного | | | |
|---|---|---|---|---|
| | 0 | 1 | 2 | 3 |
| $b_j$ | 2,472 | –0,837 | –0,462 | 0,140 |
| $t_j$ | 29,5 | 10,0 | 5,5 | 1,7 |
| $t_{табл}$ | 4,3 | 4,3 | 4,3 | 4,3 |
| Значимость | + | + | + | – |

Таким образом, окончательно уравнение регрессии в кодированных переменных принимает частный вид:

$$y = 2,472 - 0,837Z_1 - 0,462Z_2.$$

Анализ уравнения позволяет констатировать, что увеличение первого (температура) и второго (концентрация) фактора приводит к уменьшению скорости роста эпитаксиальной пленки в изученном диапазоне значений этих факторов. Причем более существенным является влияние температуры, а влияние отношения расходов компонентов газовой смеси (переменная $Z_3$) при данной ошибке эксперимента является несущественным.

Следующим этапом моделирования с использованием статистических моделей является оценка адекватности полученного уравнения регрессии, то есть проверка гипотезы об адекватности полученной модели объекту моделирования. Проверка адекватности модели может быть проведена с помощью расчета F-критерия Фишера, отражающего сравнение двух дисперсий: дисперсии воспроизводимости результатов эксперимента $S_y^2$ с дисперсией адекватности значений функции, рассчитанных по модели, значениям функции, полученным в эксперименте, $S_{ад}^2$.

$$F = \frac{S_{ад}^2}{S_y^2},$$

где $S_y^2$ – дисперсия воспроизводимости эксперимента; $S_{ад}^2$ – дисперсия адекватности модели эксперименту:

$$S_{ад}^2 = \frac{\sum_{i=1}^{N_y} (y_i - y_i^{рас})^2}{f_{ад}},$$

где $y_i$ – значения выходной переменной, полученные в эксперименте; $y_i^{рас}$ – значения выходной переменной, рассчитанные по полученной модели; $f_{ад}$ – число степеней свободы, то есть независимых связей, которое соответствует адекватности модели ($f_{ад} = N – d$, где d – число значимых коэффициентов уравнения регрессии, то есть в примере $f_{ад} = 8 – 3 = 5$).

Условием адекватности уравнения регрессии является превышение табличного значения критерия Фишера $F_{табл}$ над рассчитанным значением F, то есть соотношение $F_{табл} > F$. Табличное значение $F_{табл}$ определяется по таблицам критических значений статистических критериев при соответствующих значениях $f_y$, $f_{ад}$ и выбранном значении уровня значимости α. В рассматриваемом примере при $f_y = 2$, $f_{ад} = 5$ и $\alpha = 0,05$ табличное значение $F_{табл} = 19,3$. Если рассчитанное значение F оказывается большим, чем $F_{табл}$, то процедура моделирования возвращается на этап выполнения эксперимента при условии отсутствия ошибок на этапе статистической обработки результатов экспериментов.

Следующим этапом алгоритма моделирования с использованием статистических моделей является собственно расчет, называемый часто симуляцией, промежуточных значений выходной переменной в пределах поля экспериментальных значений входных переменных по полученному адекватному уравнению регрессии. На данном этапе осуществляется переход от уравнения регрессии в кодированных переменных к уравнению статистической модели в натуральных переменных. Для этого вместо кодированных безразмерных переменных $Z_i$ в уравнение модели подставляют натуральные переменные. В нашем примере эта подстановка приводит к выражению

$$v_p = 2,472 - 0,837\frac{T - 1250}{25} - 0,462\frac{C - 0,15}{0,05}.$$

По уравнению можно рассчитать скорость роста эпитаксиальной пленки кремния (мкм/мин) в любой промежуточной точке для температурного диапазона 1225–1275 °С и концентрации тетрахлорида кремния в газовом потоке 0,1–0,2 моль/л. Важно строго придерживаться использованной размерности величин. Например, использование полученной статистической модели для определения скорости роста пленки при температуре T = 1238 °С и концентрации 0,18 моль/л дает следующий результат:

$$v_p = 2,472 - 0,837\frac{1238 - 1250}{25} - 0,462\frac{0,18 - 0,15}{0,05} =$$
$$= 2,472 + 0,402 - 0,277 = 2,597 \text{ (мкм/мин).}$$

Напомним, что полученная в примере статистическая модель эпитаксии пленок кремния не отражает влияние соотношения расходов компонентов газовой смеси на скорость роста монокристаллической пленки. Также не следует забывать, что в рамках моделирования с использованием статистических моделей некорректно проводить экстраполяцию данных за пределы экспериментального поля значений входных переменных. Например, в рассматриваемом случае нельзя рассчитывать по уравнению скорость роста эпитаксиальной пленки кремния для условий:

$$1225 \text{ °C} > T > 1275 \text{ °C};$$
$$0,1 \text{ моль/л} > C > 0,2 \text{ моль/л}.$$

Заключительным этапом моделирования с использованием статистических моделей может являться универсальная процедура оптимизации исследованного технологического процесса. При этом важно помнить, что глобальная оптимизация возможна для случая, когда выбранный критерий оптимизации, то есть выходная оптимизируемая переменная,

экстремально зависит от одного (одномерная оптимизация) или нескольких (многомерная оптимизация) оптимизирующих факторов.