

Системное оптимальное планирование регрессионного эксперимента

С.Г. Радченко, д.т.н., профессор
Национальный технический университет Украины

УДК 519.242.5:519.233.5

Сформулированы системные требования к структуре многофакторной статистической модели и плану эксперимента. Определены классы планов, для которых выполняются сформулированные требования. Показано, что использование разработанного подхода позволило оптимально решать прикладные системные задачи.

Постановка проблемы

Классическая теория планирования эксперимента для построения статистических моделей, линейных по параметрам и, в общем случае, нелинейных по факторам, разработала методологию постановки многофакторных экспериментов в предположении, что структура модели и критерии ее оптимальности заданы. План эксперимента конструируется в соответствии с этой информацией. Практика решения прикладных задач по новым мало изученным техническим и технологическим системам показала, что структура модели часто исследователю не известна. Постулирование структуры модели не гарантирует успешное решение задачи. Необходимо разработать методы выбора плана эксперимента и получения структуры модели с гарантированным обеспечением ее адекватности, устойчивости, статистической эффективности, семантической. Выполнение указанных критериев для получаемых моделей будем определять как системное оптимальное планирование эксперимента. В доступных публикациях такое решение проблемы не известно.

Анализ публикаций по теме исследований

Авторы учебного пособия [1, с. 9] отмечают, что «...задача аппроксимации эмпирических зависимостей как на компьютере, так и без него, начиная с ее постановки, во многих отношениях строго не решается, остается неопределенной и неоднозначной».

Д.т.н. Налимов В.В., к.т.н. Голикова Т.И. подчеркивают, что матрица независимых переменных X появляется только после того, как модель задана [2, с. 31].

Д.т.н. Новицкий П.В., к.т.н. Зограф И.А. обращают внимание на то, что планирование эксперимента возможно только после установления вида математической модели исследуемого явления [3, с. 277].

В работе [2, с. 120] поставлена задача получения робастных планов экспериментов, «...т. е. планов малочувствительных к возможному изменению моделей хотя бы в пределах некоторого их класса». Однако поставленная задача не была решена.

В работе [4, с. 184] отмечается, что план эксперимента, будучи оптимальным для определенной модели, может оказаться «наихудшим из возможных», если модель в действительности окажется иной.

Помимо выбора устойчивого (робастного) плана эксперимента, т.е. плана, позволяющего устойчиво выделять различные структуры моделей, необходимо разработать общий метод получения устойчивых структур, не известных исследователю для конкретно решаемой задачи.

К.т.н. Адлер Ю.П. считает, что «никакие изошрения не позволяют получить единственный и однозначный ответ» при определении уравнения регрессии по предложенным исходным данным. «Автоматизация задач регрессионного анализа зашла в тупик» [5, с. 5-6]. При рассмотрении разнообразных шаговых и комбинированных методов регрессионного анализа не формулируются конкретные алгоритмы выделения структур уравнения регрессии, а предла-

гается более 11 процедур, «состоящих в последовательном добавлении к регрессии или, наоборот, удалении из нее отдельных носителей» [5, с. 125...131, 142...143]. Обращается внимание на «проблему коллинеарности» в регрессионных задачах.

В работе [6, с. 22, 175...176] отмечается, что «не существует стандартных приемов и методов, которые образовывали бы строгую теоретическую базу для решения этой важнейшей задачи» – «задачи правильного определения структуры модели».

Авторы книги [7, с. 255] «отмечают важность выбора такой функции регрессии, чтобы она имела «физический смысл, т. е. открывала какую-то объективную закономерность... Поэтому выбор типа регрессионной зависимости является самой острой проблемой в любом исследовании».

Аналогичные суждения высказываются и в других публикациях по статистическим методам исследования.

Другая важная проблема связана с разработкой метода устойчивого оценивания статистических моделей в условиях мультиколлинеарности факторов.

Цель статьи

Изложение методов получения устойчивых структур моделей, не известных исследователю, и устойчивых многофакторных планов экспериментов, которые обеспечивают комплекс необходимых критериев качества многофакторных регрессионных моделей.

Основная часть

Сформулируем основные критерии, которым должны соответствовать структура статистической модели и многофакторный план эксперимента.

Структура статистической модели, линейной по параметрам и, в общем случае, нелинейной по факторам должна соответствовать следующим требованиям:

1. Структурная группа коэффициентов многофакторного уравнения регрессии не известна исследователю. Задание исследователем структуры уравнения регрессии, в общем случае, будет ошибочной по причине сложности, новизны, эмергентности (несводимости свойств системы к сумме свойств ее отдельных подсистем) систем.
2. Структуры моделей выбираются из структуры модели полного факторного эксперимента с ортогональными или близкими к ортогональным структурными эффектами. Иное множество эффектов, в общем случае, не образует достаточное множество совместно ортогональных эффектов и, следовательно, получаемая модель будет не адекватной и не устойчивой.
3. Выбор структуры модели должен быть формализованным. В противном случае исследователь будет вынужден использовать метод проб и ошибок, что не гарантирует получение адекватной модели.
4. Возможность формализованного отображения в математической модели произвольной (но конечной) по сложности реальной действительности при условии правильного выбора степени полинома по каждому фактору. Необходимо выбрать множество главных эффектов и их взаимодействий со свойствами семантичности, т.е. интерпретации коэффициентов в терминах предметной (моделируемой) области знаний.
5. Доступность, простота и надежность фактического получения адекватной структуры при решении задач на потоке. Выполнение такого требования возможно при использовании соответствующего программного обеспечения, которое реализует новую технологию регрессионного анализа.

Многофакторный план эксперимента должен соответствовать следующим критериям:

1. Обеспечение соответствия эффектов модели критерию ортогональности или (для случая дробного факторного эксперимента) близости к нему. Невыполнение этого требования приводит, в общем случае, к получению неадекватных и неустойчивых моделей, а коэффициенты моделей будут смещенные.
2. Соответствие плана эксперимента критериям D -, A -, E -, G -, Q -оптимальности. Не выполнение критериев приведет к получению не эффективных в статистическом смысле моделей.
3. План эксперимента должен обеспечивать за счет ортогональности и нормирования эффектов максимально возможную устойчивость коэффициентов модели $\text{cond}(\mathbf{X}^T \mathbf{X}) = 1$. При невыполнении – структура модели не будет соответствовать «истинной» (правильной) и она будет неустойчивой.
4. При ортогональности эффектов получаем наилучшие условия для случаев перебора (введение большего, чем надо, числа эффектов) и недобора эффектов по отношению к «истинной» структуре модели по критериям несмещенности, состоятельности и статистической эффективности определяемых коэффициентов. При невыполнении – полученная модель не будет соответствовать приведенным критериям.
5. При ортогональности всех эффектов обеспечивается адекватность или близость к ней полученных моделей в реализованных точках плана эксперимента. При невыполнении этого условия модель будет давать ошибочные значения критерия качества моделируемой системы (отклика).
6. Возможность использования плана эксперимента для количественных и качественных факторов. Многофакторные регулярные планы позволяют описывать влияние количественных и качественных факторов.
7. Возможность последовательного планирования для дробного факторного эксперимента. При выполнении условия возможно адекватное описание критерия качества при минимальном числе опытов или, при необходимости, переходе к большему числу опытов с сохранением начального плана или его расширением.
8. Разбиение плана эксперимента, при необходимости, на произвольное число ортогональных блоков. При невыполнении условия дисперсия воспроизводимости может быть значительной и некоторые эффекты модели будут статистически незначимы.
9. Полный факторный эксперимент и многофакторные регулярные планы позволяют выделять различные структуры статистических моделей $\hat{y}_w$ (w – текущий номер определяемой модели, $1 \leq w \leq m$; m – общее число определяемых по плану эксперимента моделей) по одному плану эксперимента. При невыполнении условия необходимо использовать несколько планов, что увеличивает затраты на исследование.

Формализованная структура многофакторной статистической модели задается выражением

$$\prod_{i=1}^k (1 + x_i^{(1)} + x_i^{(2)} + \dots + x_i^{(s_i-1)}) \rightarrow N_{\Pi}, \quad (1)$$

где 1 – значение фиктивного фактора $x_0 \equiv 1$;

$x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(s_i-1)}$ – ортогональные контрасты факторов X_i ;

s_i – число различных уровней фактора X_i ;

k – общее число факторов, $1 \leq i \leq k$;

(1), (2), ..., (s_i–1) – порядок контрастов фактора X_i ;

N_{Π} – число структурных элементов полного факторного эксперимента, равное числу опытов эксперимента.

По теореме Бродского В.З. в любом полном факторном эксперименте все эффекты ортогональны друг к другу [8, с. 26...29]. Расширенная матрица $\mathbf{X}$ главных эффектов и взаимодействий содержит столбец фиктивного фактора $x_0 \equiv 1$, столбцы всех главных эффектов и всех возможных взаимодействий главных эффектов. Если эффекты факторов и взаимодействий факторов выразить в виде системы ортогональных нормированных контрастов, т. е.

$$\begin{aligned} \sum_{u=1}^N x_{iu}^{(p)} &= 0, & \sum_{u=1}^N x_{iu}^{(p)} \times x_{ju}^{(q)} &= 0, \\ \sum_{u=1}^N [x_{iu}^{(p)}]^2 &= N, & \sum_{u=1}^N [x_{iu}^{(p)} \times x_{ju}^{(q)}]^2 &= N, \end{aligned}$$

то матрица дисперсий-ковариаций примет вид:

$$(\mathbf{X}^T \mathbf{X})^{-1} \sigma^2(\varepsilon) = (1/N) \mathbf{E} \sigma^2(\varepsilon),$$

где $x_{iu}^{(p)}$ – значение p -го ортогонального контраста i -го фактора для u -й строки матрицы планирования, $1 \leq u \leq N$, $1 \leq p \leq s_i - 1$;

$x_{ju}^{(q)}$ – значение q -го ортогонального контраста j -го фактора для u -й строки матрицы планирования, $1 \leq q \leq s_j - 1$, $1 \leq i < j \leq k$;

$\mathbf{X}$ – матрица эффектов полного факторного эксперимента;

$\sigma^2(\varepsilon)$ – теоретическое значение дисперсии воспроизводимости результатов опытов;

N – число опытов в плане эксперимента;

$\mathbf{E}$ – единичная матрица.

Анализ требований к многофакторному плану эксперимента и структуре статистической модели показал, что им соответствуют любые полные факторные эксперименты. По теореме Хоттелинга [4, с. 62...63] выполнение требования ортогональности всех эффектов приводит к минимизации дисперсий коэффициентов статистической модели и получению совместно эффективных оценок.

Рассмотрим построение структуры модели для полного факторного эксперимента $2^1 \times 3^1 \times 4^1 // 24$: первый фактор x_1 на двух, второй x_2 на трех, третий x_3 на четырех уровнях. Формализованная структура статистической модели будет следующей:

$$(1 + x_1^{(1)})(1 + x_2^{(1)} + x_2^{(2)})(1 + x_3^{(1)} + x_3^{(2)} + x_3^{(3)}) \rightarrow N_{\Pi} = 24,$$

где $x_1^{(1)}$, $x_2^{(1)}$, $x_3^{(1)}$ – линейные контрасты факторов x_1 , x_2 , x_3 ;

$x_2^{(2)}$, $x_3^{(2)}$ – квадратичные контрасты факторов x_2 , x_3 ;

$x_3^{(3)}$ – кубический контраст фактора x_3 .

Общий вид статистической модели будет следующий:

$$\begin{aligned} \hat{y} = & b_0 x_0 + b_1 x_1^{(1)} + b_2 x_2^{(1)} + b_3 x_2^{(2)} + b_4 x_3^{(1)} + b_5 x_3^{(2)} + b_6 x_3^{(3)} + b_7 x_1^{(1)} x_2^{(1)} + b_8 x_1^{(1)} x_2^{(2)} + \\ & + b_9 x_1^{(1)} x_3^{(1)} + b_{10} x_1^{(1)} x_3^{(2)} + b_{11} x_1^{(1)} x_3^{(3)} + b_{12} x_2^{(1)} x_3^{(1)} + b_{13} x_2^{(1)} x_3^{(2)} + b_{14} x_2^{(1)} x_3^{(3)} + \\ & + b_{15} x_2^{(2)} x_3^{(1)} + b_{16} x_2^{(2)} x_3^{(2)} + b_{17} x_2^{(2)} x_3^{(3)} + b_{18} x_1^{(1)} x_2^{(1)} x_3^{(1)} + b_{19} x_1^{(1)} x_2^{(2)} x_3^{(1)} + \\ & + b_{20} x_1^{(1)} x_2^{(1)} x_3^{(2)} + b_{21} x_1^{(1)} x_2^{(2)} x_3^{(1)} + b_{22} x_1^{(1)} x_2^{(2)} x_3^{(2)} + b_{23} x_1^{(1)} x_2^{(2)} x_3^{(3)}. \end{aligned}$$

Модель содержит семь главных эффектов, одиннадцать двойных взаимодействий и шесть тройных взаимодействий.

Отметим следующие определяющие свойства предложенных многофакторных планов экспериментов и способа выделения структуры статистической модели.

1. В основу подхода положена ортогональность всех эффектов для полного факторного эксперимента. В случае использования дробного факторного эксперимента выбираются планы, которые в максимально возможной степени соответствуют критериям плана полного факторного эксперимента.
2. В статистическую модель, неизвестную априори исследователю, включаются ортогональные и статистически значимые эффекты из множества эффектов для модели полного факторного эксперимента. Если число необходимых экспериментов $N_{\text{д}}$ выбирать с учетом эмпирической формулы

$$N_{\text{д}} \approx (1,5 \dots 2) \sum_{i=1}^k (s_i - 1), \quad (2)$$

где s_i – число уровней i -го фактора; $1 \leq i \leq k$, то практика решения многочисленных задач показала, что адекватность моделей достигается.

3. Структура модели полного факторного эксперимента (1) позволяет выделять по одному и тому же плану эксперимента различные структуры статистических моделей $\hat{y}_w$, каждая из которых будет соответствовать сформулированным основным критериям для структуры статистической модели.

Использование полных факторных экспериментов не всегда возможно из-за значительного числа проводимых экспериментов. Обычно возможное максимальное число факторов 3...4 (при $s_i = 3$ $N_{\text{П}} = 27$ для $k = 3$ и $N_{\text{П}} = 81$ для $k = 4$). Необходимо применять дробные факторные эксперименты, статистические свойства которых близки к свойствам полного факторного эксперимента. Такими планами являются многофакторные регулярные планы и планы на основе ЛП_т равномерно распределенных последовательностей [9, с. 159...169].

В многофакторных регулярных планах все главные эффекты ортогональны друг к другу. Если число необходимых экспериментов $N_{\text{д}}$ выбирать с учетом эмпирической формулы (2), то некоторые из взаимодействий факторов будут ортогональны к главным эффектам или коррелированы с ними сравнительно слабо:

$$|r_{ij}(x_{iu}^{(p)}, (x_{iu}^{(p')} \times x_{ju}^{(q)}))| \leq 0,3 \dots 0,4,$$

где r_{ij} – коэффициент парной корреляции между эффектами; $1 \leq p < p' \leq s_i - 1$; $1 \leq q \leq s_j - 1$.

При использовании дробных факторных экспериментов в регрессионную модель вводятся эффекты (главные и взаимодействия), которые ортогональны или слабо коррелированы со всеми ранее введенными и статистически значимые. Выбор структурных составляющих статистической модели производится из числа структурных элементов модели полного факторного эксперимента (1). Указанное множество элементов необходимо и достаточно для адекватной аппроксимации результатов экспериментов. При выборе структурных составляющих статистической модели используется разработанный алгоритм RASTA3 [9, с. 179...182] и другие алгоритмы.

Множество структурных элементов (1) для модели полного факторного эксперимента образуют систему базисных функций, которые ортогональны друг к другу. Если нормировать эффекты, то они создают ортонормированный базис, который позволяет получить статистические модели наилучшие по критериям адекватности, устойчивости, статистической эффективности и другим. Иная система базисных функций с такими свойствами неизвестна.

Отметим, что введение новых эффектов во множество эффектов (1), отличных от них, нарушает ортогональность базиса, а выведение некоторых из них не позволяет получить, в общем случае, наилучшие статистические модели.

Программное обеспечение решения задач

На кафедре технологии машиностроения Национального технического университета Украины «Киевский политехнический институт» разработано программное средство «Планирование, регрессия и анализ моделей» (ПС ПРИАМ) [9, с. 45...47]. Оно предназначено для выполнения комплекса работ, связанных с генерацией многофакторных планов для оптимизации и моделирования, обработки их результатов, проведения вычислительного эксперимента, визуализации полученных результатов.

ПС ПРИАМ характеризуется следующими функциями.

1. Реализация специально разработанной технологии решения научных и прикладных задач по построению статистических регрессионных моделей и многокритериальной оптимизации.
2. Построение робастных (устойчивых) планов эксперимента.
3. Эффективные алгоритмы определения структуры уравнения регрессии (алгоритмы RASTA3 и др.).
4. Устойчивое оценивание коррелированных факторов в многофакторном уравнении регрессии (алгоритмы RASTA4, RASTA5.1 и др.).
5. Ориентация на массового пользователя. ПС содержит все необходимые средства для решения задач от ее постановки до получения отчета.

Изложенное системное оптимальное планирование эксперимента использовалось при получении статистических регрессионных моделей технических, технологических, измерительных, материаловедческих и других систем [10, с. 102...115]. Его применение возможно при решении конструкторских, технологических, испытательных, эксплуатационных задач, в научно-исследовательской работе, в учебном процессе.

Системное оптимальное планирование эксперимента было проверено при решениях более ста прикладных системных задач и показало хорошие результаты.

В исследовании [11] для трех типов полиэтиленовых труб – тяжелой (Т), средней (С) и средне-легкой (СЛ) были смоделированы и оптимизированы технологические условия сварки: температура нагревательного инструмента (t_n), время оплавления свариваемых торцов ($t_{опл}$), давление осадки труб ($P_{ос}$). Результаты сварки характеризовались шестью критериями качества (разрушающее напряжение при растяжении, высота и ширина грата в зоне сварки, составляющая вязкого разрушения при изломе образца и др.). Использовался многофакторный регулярный план $3^1 \times 4^3 // 32$ [9, с. 333]. С применением ПС ПРИАМ по результатам эксперимента построены статистические модели $\hat{y}_1 \dots \hat{y}_6$ и проведен их анализ. Разработанные по моделям оптимальные условия сварки были внедрены в четырех химических трестах строительного производства. Фактический народнохозяйственный эффект за 1985...1987 г. составил 896 398 руб.

С целью определения оптимальных и предельных условий эксплуатации, обоснованного выбора материалов термонапряженной изоляции (ТНИ) высоковольтного трансформаторно-выпрямительного модуля было проведено исследование влияния 9 простых факторов, сведенным к 4 обобщенным (X_1 – электрофизический, X_2 – геометрический, X_3 – теплоотвода в окружающую среду, X_4 – теплофизический) на 9 критериев качества $y_1 \dots y_9$, характеризующих запас электрической прочности, напряженность электрического поля и температуру диэлектрика в различных точках ТНИ в нестационарном режиме [12]. Использовался равномерный многофакторный регулярный план $3^4 // 27$ (четыре фактора, каждый фактор изменялся на трех уровнях, 27 опытов). Были получены статистические модели для 9 критериев качества ТНИ, по моделям был проведен вычислительный эксперимент и найдены оптимальные значения факторов для различных сочетаний критериев качества.

В работе [13] было проведено повышение точности измерения цифровых весов путем моделирования влияния на показание датчика весов y четырех факторов, (X_1 – гистерезис, X_2 – температура окружающей среды, X_3 – напряжение питания, X_4 – измеряемый вес), из которых первые три создают переменные систематические погрешности, с последующим исключением влияния этих погрешностей с использованием полученной модели. По результатам полного факторного эксперимента $2^1 \times 3^2 \times 6^1 / 108$ построена адекватная, информативная, максимально устойчивая модель $\hat{y}$, с использованием которой точность измерений по критерию среднее абсолютных величин относительных погрешностей аппроксимации $|\bar{e}_{\text{отно}}|$ была повышена в 13,3 раза и по критерию среднеквадратичной погрешности аппроксимации в 11,2 раза.

Изложенная система оптимального планирования эксперимента предполагает, что форма факторного пространства соответствует классической: планированию на кубе, сфере, симплексе. Если форма факторного пространства иная, то необходимо использовать разработанные методы топологического отображения области хорошо обусловленного факторного пространства – прообраза – в область плохо обусловленного пространства – образ [9, с. 183...300], что позволяет устойчиво решать некорректно поставленные задачи.

Выводы и перспективы дальнейших исследований

1. Впервые сформулированы системные требования к плану эксперимента, обеспечивающие, в общем случае, наилучшие системные свойства статистических моделей.
2. Впервые приведен общий вид структуры многофакторной статистической модели, позволяющий получить адекватные, семантические (в информационном смысле), устойчивые модели, структуры которых заранее исследователю не известны.
3. Показано, что сформулированные требования к плану эксперимента и структуре моделей выполняются для полных факторных экспериментов, многофакторных регулярных планов и планов на основе ЛП_т равномерно распределенных последовательностей.

Разработанное системное оптимальное планирование эксперимента обеспечивает выбор плана и получение структуры модели, не известной исследователю заранее, подтвердило эффективность при решении более ста прикладных системных задач по техническим, технологическим, измерительным и др. системам.

Дальнейшее развитие изложенного подхода проводится с использованием топологического метода устойчивого оценивания регрессионных моделей при решении некорректно поставленных задач [14; 15, с. 288...289].

Список литературы

1. Компьютерный анализ и интерпретация эмпирических зависимостей / С.В. Поршневу, Е.В. Овечкина, М.В. Машенко, А.В. Каплан и др. Учебник. – М.: ООО «Бином-Пресс», 2009. – 336 с.
2. Налимов В.В., Голикова Т.И. Логические основания планирования эксперимента. – 2-е изд., перераб. и доп. – М.: Металлургия, 1981. – 152 с.
3. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. – 2-е изд., перераб. и доп. – Л.: Энергоатомиздат. Ленингр. отд-ние, 1991. – 304 с.
4. Себер Дж. Линейный регрессионный анализ / Пер. с англ. В.П. Носко; Под. ред. М.Б. Малютова. – М.: Мир, 1980. – 456 с.
5. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия: В 2-х вып. Вып. 2 / Пер. с англ. Б.Л. Розовского; Под ред. и с предисл. Ю.П. Адлера. – М.: Финансы и статистика, 1982. – 239 с.
6. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследования зависимостей: Справ. изд. / Под. ред. С.А. Айвазяна. – М.: Финансы и статистика, 1985. – 487 с.
7. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере / Под ред. В.Э. Фигурнова. – М.: ИНФРА-М, 1998. – 528 с.

8. Бродский В.З. Введение в факторное планирование эксперимента. – М.: Наука, 1976. – 224 с.
9. Радченко С.Г. Устойчивые методы оценивания статистических моделей: Монография. – К.: ПП «Санспарель», 2005. – 504 с.
10. Радченко С.Г. Анализ экспериментальных данных на основе использования многофакторных статистических математических моделей // Математичні машини і системи. – 2005. – № 3. – С. 102...115.
11. Оптимизация технологических условий сварки полиэтиленовых труб / С.Г. Радченко, Ю.С. Бурбело, Э.В. Котенко, С.Н. Лапач и др. // Пласт. массы. – 1988. – № 9. – С. 29...31.
12. Многофакторная математическая модель термонапряженной электроизоляции / М.Е. Иерусалимов, С.А. Соколовский, С.Г. Радченко и др. // Электричество. – 1991. – № 8. – С. 40...45.
13. Радченко С.Г., Бабич П.Н. Информационная коррекция переменных систематических погрешностей средств измерений и измерительных информационных систем // Радиотехника и информатика. – 1999. – № 3. – С. 82...88.
14. Лаборатория экспериментально-статистических методов исследований – <http://n-t.ru/sp/lesmi/>
15. Радченко С.Г. Методология регрессионного анализа: монография. – К.: «Корнійчук», 2011. – 376 с.

Ранее опубликовано:

Радченко С.Г. Системное оптимальное планирование регрессионного эксперимента / С.Г. Радченко // Заводская лаборатория. Диагностика материалов. – 2012. – Т. 78, № 7. – С. 71...75.

Дата публикации:

28 июля 2013 года

Электронная версия:

© «Наука и техника», n-t.ru