В заключение приведем слова Н. Винера [6, с.76]: «... современный аппарат малых выборок, как только он выходит за рамки простого подсчета своих собственных, специально определенных параметров и превращается в метод положительных статистических выводов для новых случаев, уже не внушает мне доверия, Исключение составляет случай, когда этот аппарат применяется статистиком, который явно знает или хотя бы неявно чувствует основные элементы динамики исследуемой ситуации».

**Таблица 2.**

Критерии адекватности по моделям изменения показателей луговой травы

| Параметр травы | $\Delta_{max}$, % | $\overline{\Delta}$, % | $\bar{y}_\phi$ | $\overline{\Delta}_{кв}$, % | $F_\phi$ | $R^2$ | $R$ |
|---|---|---|---|---|---|---|---|
| $m$, г/м$^2$ | 54.24 | 21.15 | 510.3 | 23.80 | 3.20 | 0.299 | 0.547 |
| $m_c$, г/м$^2$ | 51.31 | 19.22 | 138.6 | 20.62 | 3.82 | **0.461** | **0.679** |
| $m_{в0}$, г/м$^2$ | 59.87 | 22.95 | 364.2 | 26.26 | 3.12 | 0.294 | 0.542 |
| $W$, % | **26.21** | **8.81** | 250.1 | **10.32** | **4.59** | 0.379 | 0.616 |
| $\bar{v}$, г/(м$^2$ ч) | 46.66 | 18.77 | 1.312 | 21.44 | 1.87 | 0.200 | 0.447 |

*Статья опубликована при поддержке гранта 3.2.3/4603 МОН РФ*

СПИСОК ЛИТЕРАТУРЫ:

1. Мазуркин, П.М. Статистическое моделирование. Эвристико-математический подход / П.М. Мазуркин. - Научное издание. - Йошкар-Ола: МарГТУ, 2001. - 100с.

2. Пасхавер, И.С. Общая теория статистики. Для программированного обучения / И.С. Пасхавер, А.Л. Яблочник. М.: Финансы и статистика, 1983. - 432 с.

3. Фёрстер, Э. Методы корреляционного и регрессионного анализа: Руководство для экономистов / Э. Фёрстер, Б. Рёнц. - М.: Экономика и статистика, 1983. - 302 с.

4. Елисеева, И.И. Логика прикладного статистического анализа / И.И. Елисеева, В.О. Рукавишников. - М.: Финансы и статистика, 1982. - 192 с.

5. Мазуркин, П.М. Измерение продуктивности травяного покрова пойменного луга / П.М. Мазуркин, С.И. Михайлова // Современные наукоемкие технологии: материалы заочной электронной конференции. – 2008. - № 7. – С.91-92.

6. Винер, Н. Кибернетика или управление и связь в животном и машине / Н. Винер. - 2-е изд. - М.: Наука, 1983. - 344 с.

## БИОТЕХНИЧЕСКИЙ ЗАКОН И ПОДГОТОВКА ИСХОДНЫХ ДАННЫХ

Мазуркин П.М.
*Марийский государственный технический университет*
*Йошкар-Ола, Россия*

Чаще всего статистическое моделирование выполняется по прошлой количественной информации (качественные значения преобразуются в коды, ранги, отношения), которая зафиксирована в виде текстового описания (эвристическая модель) и таблицы исходных для моделирования данных (табличной модели), а также графиков (чаще всего при однофакторных статистических выборках).

Модель может идентифицироваться и по текущей информации в реальном режиме времени, но при этом процесс идентификации не должен превышать одной трети промежутка времени между получением каждой порции сведений. Однако и здесь на первом сеансе значения параметров искомой модели необходимо вычислить по прошлым данным, то есть необходим анализ некоторой предыстории явления или процесса. Последующие сеансы параметрической идентификации выполняются гораздо быстрее из-за использования готовых после первого сеанса моделей.

На рис. 1 приведены условные примеры аппроксимации (рис. 1а) и параметрической идентификации (рис. 1б).

В первом случае логарифмированием получаем вместо показательного закона $y = a_1 x^{a_2}$ линейную модель $\ln y = \ln a_1 + a_2 \ln x$. Во втором случае точная линеаризация невозможна. Исследователи пытаются практически решить эту задачу статистического моделирования с помощью уравнения $y - a_0 = a_1 x^{a_2}$ путем принятия ориентировочных значений $a_0$. Однако результат такого решения может оказаться некорректным. Тогда снова подбирают значение $a_0$ до тех пор, пока график на рис. 1б не будет глазомерно максимально приближен к множеству экспериментальных данных.

Этот простейший пример неавтоматизированного выбора конструкции модели показывает, что необходим некоторый перебор вариантов значений $a_0$. Если принятое значение $a_0$ удовлетворяет критериям сходимости и адекватности модели к фактическим данным, то он запоминается. Так шаг за шагом выполняется случайный

поиск, в данном примере в неавтоматизированном режиме.

Многие статистические модели не поддаются линеаризации. Многофакторные модели практически всегда невозможно аппроксимировать. Вначале для понимания сущности математических конструктов строятся частные характеристические графики (без масштаба) бинарных отношений типа «фактор $x_i$ - показатель $y$ ».

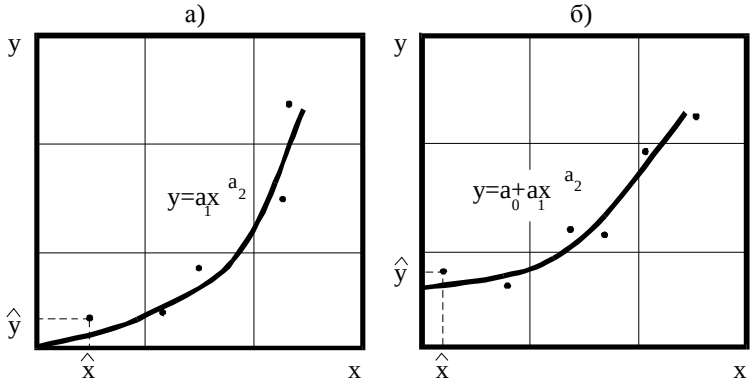


**Рис. 1.** Характеристические графики, построенные по множеству экспериментальных точек: а - модель, приводимая к линейному виду и, соответственно, поддающаяся апроксимации; б - модель, не приводимая (точно) к линейному виду и требующая параметрической идентификации; x, y- координаты экспериментальных точек

На рис. 2 показан условный пример построения составных моделей по какому-то бинарному отношению $x_i \to y$. По схеме на рис. 1а эксперт считает, что изменение $y = f(x_i)$ состоит из двух процессов. Причем он указывает математику-программисту (исходя из эвристик задачи), что оба этих процесса могут быть охарактеризованы показательными законами, то есть общая модель будет $y = b_1 x_i^{b_2} + b_3 x_i^{b_4} + b_5$ (здесь и далее мы произвольные параметры будем перенумеровывать, поэтому в отличие от параметров $a_1 \ldots a_4$ биотехнического закона будем использовать, по возможности, другой символ).

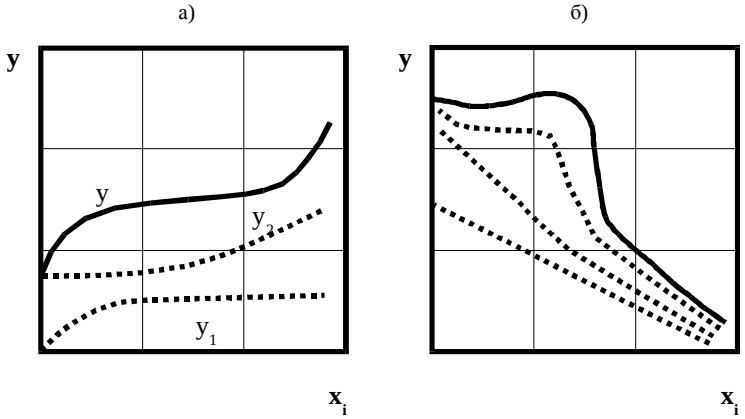


**Рис. 2.** Характеристические графики бинарных отношений xy, приводимые к составным конструкциям (модульным построением при использовании устойчивых законов) регрессионных моделей: а - сумма показательных функций; б - сумма линейной, экспоненциальной и логистической математических функций

Если известны интервалы изменения $x_i$ и ориентировочно (по мысленным представлениям) можно указать на интервалы $[y(x_{i0}), y(x_{i1})]$ изменения показателя (предварительных расчетов не требуется), то возможно указать для ПЭВМ (программы Eureka для малых выборок, ПЭК или CurveExpert-1.3 для матриц данных) значения $b_1 \ldots b_5$. Пусть для нашего примера $b_5 = 2.5$ (главное здесь угадать не значение числа, а только порядок,: если будет введено в ПЭВМ число 2500, то поиск будет затруднен, так как долгий путь машинного поиска предстоит до оконча-

тельного значения параметра, например, $b_5 = 1.8364$).

Исходные значения $b_1$ и $b_3$ угадать труднее, а для интенсивностей можно указать области нахождения числа: $0 < b_2 < 1, b_4 > 1$. Если решается множество однотипных задач, то для второго и последующих примеров принятие исходных значений параметров идентифицируемой модели упрощается, так как значения параметров идентифицируемой модели принимаются по аналогии с первым примером.

Пусть задана матрица данных $\hat{x}_i$, где знак «^» будем принимать для фактических значений. Эта матрица оформляется в виде табл. 1.

**Таблица 1.**

Форма матрицы исходных данных

| | Факторы, участвующие в моделировании | | | | | |
|---|---|---|---|---|---|---|
| № п/п | $\hat{x}_1$ | $\hat{x}_2$ | ... | $\hat{x}_i$ | ... | $\hat{x}_m$ |
| 1<br>...<br>j<br>...<br>n | | | | | | |

Матрица $m \times n$ может быть полностью заполненной. Если имеются пустые клетки, то необходимо учитывать возможность исключения некоторых факторов и групп наблюдений в некоторых математических конструктах.

Далее строятся структурные модели, например, типа:

$$\begin{cases} x_2 = f(x_1, x_3, x_7); \\ x_6 = f(x_2, x_4, x_5, x_3); \\ x_7 = f(x_1, x_5, x_8, x_9, x_{10}). \end{cases} \quad (1)$$

Такие структурные модели только указывают на зависимость одних факторов от других. Эксперт-специалист это обязан выполнить. Причем основным условием конструирования является структурная избыточность. Лучше, если конструкция каждой из составляющих математической модели будет избыточной, до полной формы. Так же желательно, если бинарные отношения будут записаны в усложненной форме, например, вместо формулы $y = ax$ следует использовать $y = ax^b$ или даже $y = ax^b \exp(-cx)$ и т.п.

В системе структурных уравнений (1) левые части становятся показателями, то есть $x_2 \to y_1$, $x_6 \to y_2$, $x_7 \to y_3$ и т.д. Так выполняется разделение факторов на объясняющие переменные $x_i$ и показатели $y_k$. При факторном анализе структурные модели типа (1) не строятся, так как как будут известны модели всех бинарных отношений между отдельными факторами.

Мы ранее указывали, что множество $y_k$ можно свернуть в обобщенный критерий (или принять несколько общих критериев) оптимизации. Эта работа при идентификации не выполняется, поэтому в данной книге не рассматривается.

При однофакторном моделировании табл. 1 превращается в двухстолбцовую таблицу со столбцами $\hat{x}$ и $\hat{y}$.

**Свойства исходных данных.** Для работы по методике МЭРА не требуется выполнять корреляционный и дисперсионный анализы. Причем общеизвестно, что существующие методы статистического моделирования исходят именно из допущения о нормальном законе распределения исходных данных.

На рис. 3 приведены практически возможные случаи распределения наблюдений в однофакторном эксперименте. Отсутствие влияния $x \to y$ будет по схемам на рис. 3а,в описываться моделью типа $y = a_1 x^0 = a_1$. При нормальном законе распределения (рис. 3а) получим $a_1 = \bar{\hat{y}} = n^{-1} \sum \hat{y}_i$, где $\bar{\hat{y}}$ - среднее арифметическое фактических значений, $n$ - общее число наблюдений.

На рис. 3б показано дискретное изменение $x$, когда при каждом значении $\hat{x}$ образуется статистическая частная выборка $\hat{y}$, которая равномерно распределена одинаково для значений $\hat{x}$. В итоге образуется линия регрессии по значениям $y = f(x)$. Эта линия равновероятно отстоит в

пределах доверительных границ $\hat{y}_1$ и $\hat{y}_2$. Очевидно, что такое распределение возможно аппроксимировать. Однако, как показали наши примеры, идентификация многофакторных моделей и здесь эффективнее.

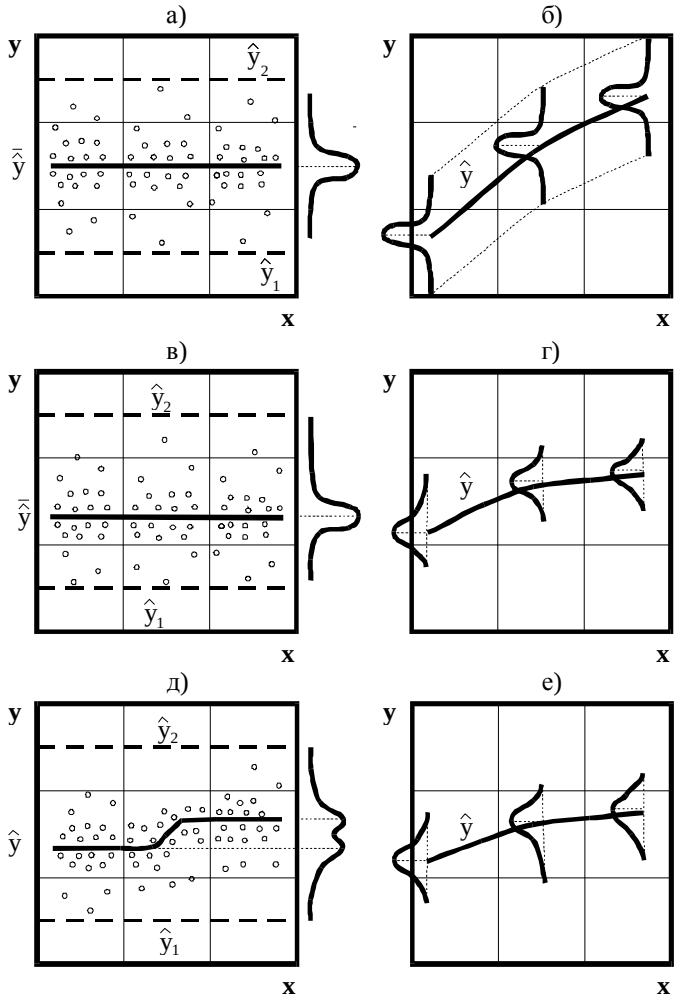


**Рис. 3.** Возможные случаи распределения повторностей наблюдения : а - случайные изменения x, y и нормальное распределение y; б - равномерно нормальные распределения y при дискретнозаданном изменении x (обычно планированием эксперимента); в - асимметрия нормального распределения y; г – равномерно асимметричное распределение выборок y при дискретных x; д - появление эксцесса у нормального распределения; е - случайные изменения асимметрии распределения

С отклонением законов распределения от нормального погрешность аппроксимации возрастает. По схеме на рис. 3в происходит значительная асимметрия исходных данных. Линия регрессии (рис. 3г) фактически проходит по «сгущенным» множествам экспериментальных точек, а аппроксимированная линия идет по среднеарифметическим значениям и поэтому отклоняется от сгущенностей наблюдений. Чем больше асимметрия, то тем существеннее разница между линией моды $\tilde{\hat{y}}$ и линией среднеарифметической $\bar{\hat{y}}$.

Появление эксцесса (рис. 3д) может произойти из-за каких-то структурных сдвигов (например, включилось во времени влияние не учтенного фактора) или из-за резкого скачка погрешностей измерения. Вот почему рекомендуется эксперименты проводить быстро, не давая времени повлиять на ход процесса самого эксперимента. Однако управление временем эксперимента чаще всего возможно выполнить только в технических исследованиях.

Эксцесс появляется также от неучтенного порогового эффекта нелинейного скачкообразного влияния фактора (переход в новое качество, например от закона Гука к упруго-пластической деформации, от стабильности экономики к кризису и др.). В условиях производства это может быть изменение самоорганизации персонала и др.

Способом идентификации объяснимые скачки (например, работа в праздничные дни и др.) вполне можно учесть и включить в виде отдельных математических конструктов.

На рис. 3е показано изменение линии регрессии при дискретных замерах и различных законах распределения исходных данных по отдельным выборкам. Методика МЭРА позволяет получить регрессионную модель, проходящую по вершинам различных типов распределений. Это означает, что вид частных законов распределения выборок не влияет на результат параметрической идентификации.

В лесном деле часты случаи со взаимно связными факторами, когда необходимы взаимно обратимые математические функции типа $y = f(x)$ или $x = \varphi(y)$. Например, высота и диаметр дерева взаимосвязаны, а сама функциональная связность (прямая и обратная) зависит от параметров местообитания этого дерева.

На рис. 4 схематически показано, что в статистических выборках $\hat{x} \leftrightarrow \hat{y}$ появляются так называемые зоны устойчивости исходных данных. При идентификации методами случайного поиска к ним стремится линия регрессии.

На рис. 4а показаны взаимные нормальные распределения неэкспериментальных данных, то есть данных, не зависящих от воли исследователя. Зона устойчивости крестообразной формы получается в виде двух прямых $y = \bar{\hat{y}}$ и $x = \bar{\hat{x}}$. При этом центр устойчивости $C(\bar{\hat{x}}, \bar{\hat{y}})$ превращается в точку.

По схеме на рис. 4б нет четкого проявления какого-то закона распределения. Предельными случаями становятся пуассоновское случайное распределение или регулярное (посадка деревьев в плантациях) размещение неэкспериментальных и экспериментальных точек [16]. Центра устойчивости здесь нет.

А зона случайного изменения $f(\hat{x}, \hat{y})$ охватывает всю область точек. В этом случае моделирование становится бессмысленным процессом, так как можно провести бесчисленное множество кривых, от которых точки будут равноотстоящими по двум одинаковым частям множества исходных данных.

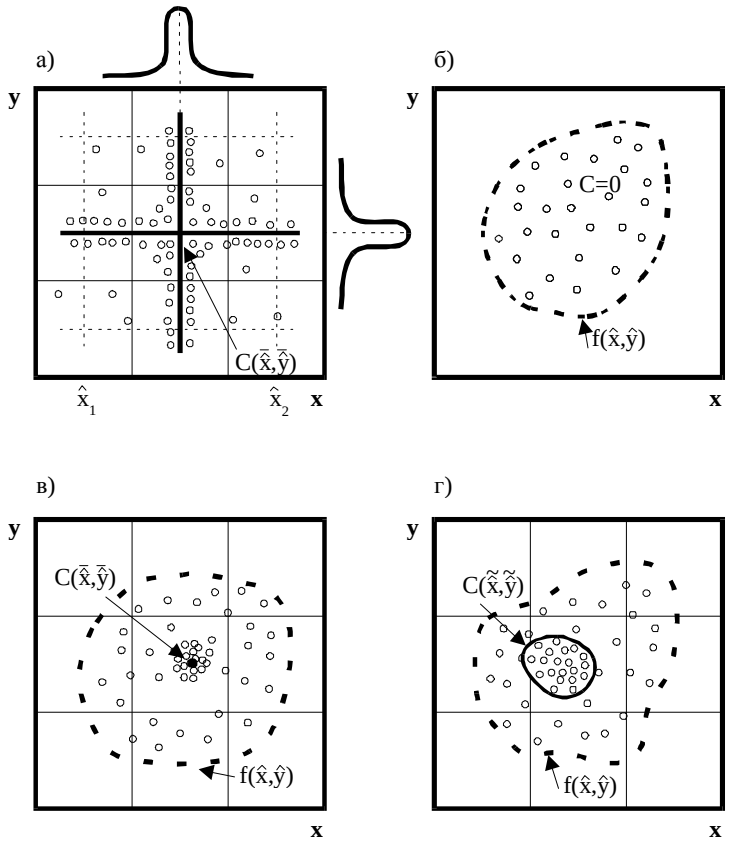


**Рис. 4.** Схемы, показывающие появление зон устойчивости исходных данных при случайных зависимостях x и y: а - нормальное распределение приводит к устойчивости в ориентациях xy и yx (вид "креста"); б - неустойчивая зона во всей области f(x,y); в - устойчивая сходимость зоны в центр C(x,y); г - сходимость зоны устойчивости по статистической информации в область линии моды C(x,y)

Если нет эвристической модели, а наблюдения выполнены без содержательного обоснования, то на практике чаще всего это происходит по многим причинам: а) неверно подобраны интервалы изменения $\hat{x}$ и $\hat{y}$; б) нет увязки между эвристикой и математикой; в) слишком малы интервалы изменения $\hat{x}$ и $\hat{y}$ и т.п.

На рис. 4в показан идеальный случай, когда зона устойчивости исходных данных сводится в центральную точку $C(\bar{\hat{x}}, \bar{\hat{y}})$. Эта точка является генеральной средней арифметической величиной. Очевидно, что точка С может образоваться и при других законах распределения, а также при их различных сочетаниях.

Процесс параметрической идентификации очень быстро сходится к устойчивым значениям параметров модели. Причем небольшие изменения (оператором ПЭВМ) значений параметров модели не влияют, так как все же эти параметры модели сходятся к одному набору чисел.

В реальных явлениях и процессах этого не происходит. Поэтому, как показано на рис. 4г, появляется влияние эксцесса. В области точек $f(\hat{x}, \hat{y})$ появляется зона устойчивости $f(\tilde{\hat{x}}, \tilde{\hat{y}})$ по модам или медианам. Для множества факторов это будет какое-то замкнутое пространство, внутри которого линия регрессии может колебаться из-за сочетаний различных значений параметров модели. Сходимость параметрической идентификации протекает дольше и исследователю иногда приходится выбирать какое-то сочетание значений по модели (1) по каким-то эвристическим соображениям. Такой случай стохастичности параметров модели появляется редко, да и то с увеличением количества параметров модели. Вычислительными экспериментами было установлено, что при числе факторов более 15 ($m > 15$) и числе переменных модели более 25 моделирование становится неустойчивым, то есть в этом случае трудно предсказуемым процессом.

Для преодоления этого явления и повышения устойчивости исходных данных необходимо моделировать комплекс формул, поочередно идентифицируя по матрице исходных данных каждую модель (2) в отдельности.

*Статья опубликована при поддержке гранта 3.2.3/4603 МОН РФ*

***Проблемы экологического мониторинга***

***Технические науки***

## БИОТЕХНИЧЕСКИЙ ЗАКОН И ЧИСЛЕННОСТЬ НАБЛЮДЕНИЙ

Мазуркин П.М.
*Марийский государственный технический университет*
*Йошкар-Ола, Россия*

Полнота количественной части исходной информации влияет на результаты параметрической идентификации. Вначале необходима сортировка исходной информации для отсеивания тех наблюдений, которые заведомо «испорчены» влиянием «чужих» факторов. Пусть такая процедура проведена, хотя она нами не рекомендуется из-за того, что любой член статистической выборки имеет право на существование. Лучше всего отбросить резко отклоняющуюся от других точку после моделирования идентификацией предложенного нами биотехнического закона [1].

Однако всегда необходимо знать, сколько же наблюдений необходимо зарегистрировать, то есть определить число наблюдений, которое значимо будет меньше для процедур идентификации устойчивых законов распределения. Поэтому методология идентификации биотехнического закона примерно в 2-5 раз экономит время и издержки на проведение последующих измерений по осознанным методикам экспериментов.

Если число наблюдений хорошо предсказуемо в технических однофакторных и даже многофакторных исследованиях и в планируемых экспериментах, то многофакторные исследования эргатических (с участием человека) или природохозяйственных (с участием природных объектов) систем, характеризующихся мультисвязностью факторов, не имеют пока обоснованных методических рекомендаций по численности необходимых наблюдений.

Принято среди ученых аксиоматически, что выборочное наблюдение, объем которого не превышает 20 единиц, следует считать малой выборкой [2, с.298]. Если есть некоторая спасительная граница, то естественно, экономисты и инженеры часто не выходят за пределы малых выборок.

Например, по рекомендациям [3, с.189] для получения достоверной многофакторной регрессионной модели рекомендуется общее число наблюдений $n$ принять из условия

$$n \geq 20N + N, \qquad (1)$$