ЭКСПЕРТНЫЕ СИСТЕМЫ И РАСПОЗНАВАНИЕ ОБРАЗОВ
Вычислительные системы
1988 год Выпуск 126

УДК 519.237

ЗАПОЛНЕНИЕ ПРОБЕЛОВ В 3-ВХОДОВЫХ ТАБЛИЦАХ ДАННЫХ ТИПА "ОБЪЕКТ-СВОЙСТВО-ВРЕМЯ"

Н.Г.Загоруйко, Г.В.Ульянов

Введение

Трехвходовые таблицы (3-таблицы), называемые за рубежом "three-way" или "three-mode matrices" - сравнительно новый объект исследования. Как всякий новый тип данных, раньше всех он привлек внимание специалистов по программированию, разведочному анализу данных, методам классификации [1; 2,с.40; 7,с.7]. Позже на подготовленную почву пришли исследователи в области анализа зависимостей. В работе [3] рассматривалась группа алгоритмов семейства ZL для заполнения пробелов в обычных (2-входовых) таблицах данных. В настоящей статье предлагается алгоритм, обрабатывающий пробелы в 3-таблицах ("3-мерный ZL").

§ 1. Структура 3-таблиц

Трехвходовая таблица (3-таблица) $Y = \{y(i,j,t)\}$, где $i = 1,\ldots,m$ - номер *объекта*, $j = 1,\ldots,n$ - номер *свойства*, $t = 1,\ldots,q$ - номер *момента* времени, представляет собой трехмерный массив (*куб данных*), который образуется путем измерения на некотором множестве m объектов некоторого множества n количественных признаков за некоторое множество q моментов времени (рис.1).

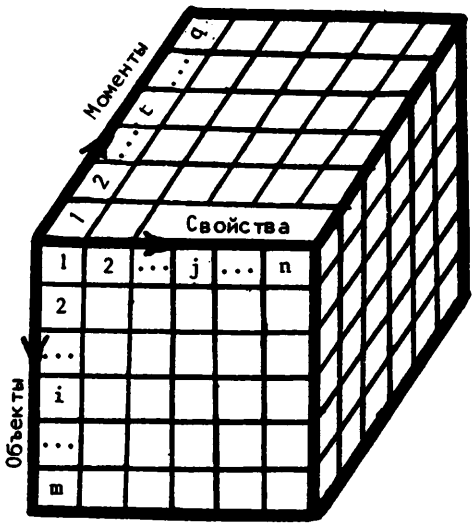

Рис. 1

Если зафиксировать один из индексов 3-таблицы, то получится 2-входовая таблица (*2-таблица*) - обычная матрица. Основными типами 2-таблиц являются : а) таблица "объект-свойство" (ТОС-t) - $Y(*,*,t)$, б) таблица "объект-время" (ТОВ-j) - $Y(*,j,*)$, в) таблица "время-свойство" (ТВС-i) - $Y(i,*,*)$ (рис.2). Кроме основных, есть еще три транспонированных к основным 2-таблицы: ТВО, ТСО, ТСВ. Можно ввести и одномерные элементы 3-таблицы, которые получаются, если в Y зафиксировать два индекса. Тогда (i,t)-строкой 3-таблицы назовем массив $Y(i,*,t)$; (j,t)-столбцом - массив $Y(*,j,t)$; (i,j)-рядом - массив $Y(i,j,*)$ (рис.3).

Каждый элемент $y(i_0,j_0,t_0)$ 3-таблицы Y находится на пересечении трех 2-таблиц - ТОС-t_0, ТОВ-j_0, ТВС-i_0; эти таблицы будем называть *инцидентными* данному элементу; кроме того, тот же элемент лежит на пересечении трех 1-мерных элементов - (i_0,t_0)-строки, (j_0,t_0)-столбца и (i_0,j_0)-ряда, которые мы также назовем *инцидентными* данному элементу.

§ 2. Методы обработки пробелов

В основном режиме каждый пробел (или редактируемый элемент) $y(i_0,j_0,t_0)$ обрабатывается алгоритмом ZL независимо

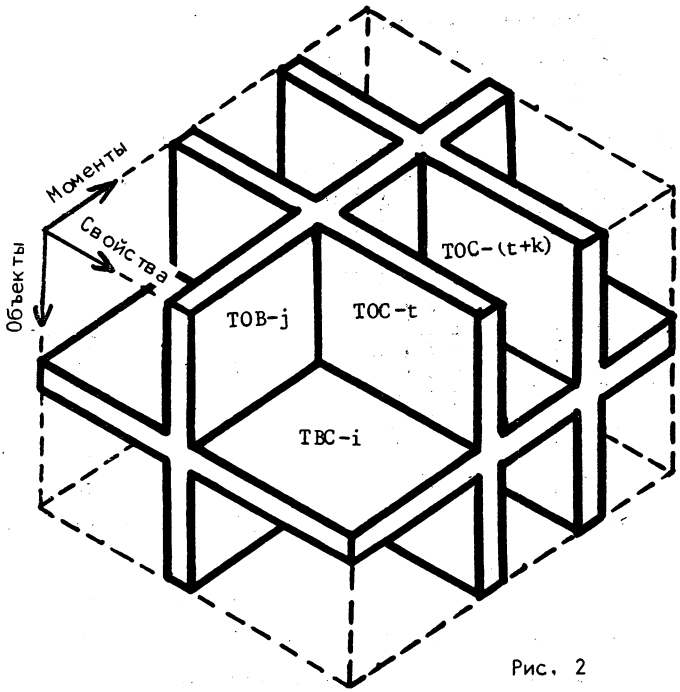
Моменты
Свойства
Объекты
ТОВ-j
ТОС-t
ТОС-(t+k)
ТВС-i

Рис. 2

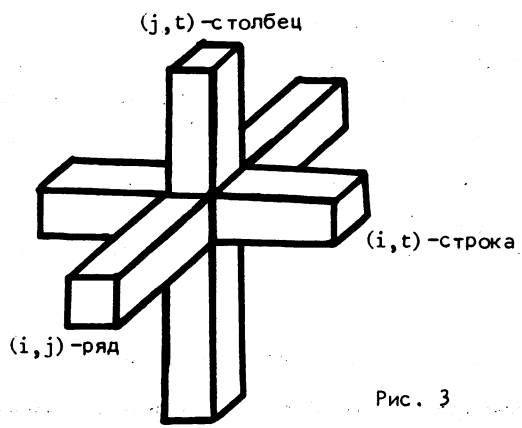
(j,t) -столбец
(i,t) -строка
(i,j) -ряд

Рис. 3

от остальных и результаты обработки одного пробела никак не используются при обработке другого. Поэтому достаточно рассмотреть, как обрабатывает алгоритм какой-нибудь один пробел. Строку, столбец и ряд, инцидентные заполняемому в данный момент пробелу, будем называть *предсказываемыми*. Пробелы в 3-таблице обозначаются уникальным числом PROB, которое должно быть больше каждого элемента таблицы.

Алгоритм ZL сводит задачу обработки пробела в 3-таблице к обработке этого пробела на некотором множестве 2-таблиц, содержащих данный пробел к последующему сравнению полученных прогнозов по некоторому критерию G.

В предыдущей статье авторов [3] описано, каким образом для любого элемента y_0 2-таблицы алгоритм ZL вычисляет прогноз $\hat{y}_0$ и значение критерия G_0 (оценку ожидаемой ошибки прогноза). В некоторых случаях выдается отказ от прогноза пробела по данной 2-таблице. Теперь рассмотрим методы нахождения прогноза для пробела в 3-таблице:

а) метод инцидентных таблиц (3-ТАБ),

б) метод транспонированных таблиц (3-ТАБТ),

в) метод шести таблиц (6-ТАБ),

г) метод одной таблицы с лагами (ТАБ-Л),

д) метод инцидентных таблиц с лагами (3-ТАБ-Л).

§ 3. Метод инцидентных таблиц

В этом методе для каждого пробела $y_0 = y(i_0, j_0, t_0)$ вычисляется прогноз по каждой из трех 2-таблиц, инцидентных данному пробелу (т.е. по ТОС-t_0, ТОВ-j_0, ТВС-i_0), скажем $\hat{y}^0_{TOC}$, $\hat{y}^0_{TOB}$, $\hat{y}^0_{TBC}$, и соответствующее значение критерия G_{TOC}, G_{TOB}, G_{TBC} ($G = PROB$, если имел место отказ от прогноза по данной таблице). Тогда оптимальным считается прогноз $\hat{y}_0$, соответствующий $G = \min \{G_{TOC}, G_{TOB},$

G_{TBC}}. Если же все $G = PROB$ либо $G > REST$, где REST - задаваемая пользователем допустимая величина ошибки, то данный пробел останется вообще без прогноза. Сравнение значений критерия G для разных таблиц правомерно, так как ошибка оценивается в процентах к вычисленным стандартным ошибкам предсказываемых столбцов, соответствующих 2-таблиц.

В каждой 2-таблице согласно 2-мерному алгоритму ZL будет использоваться одна и та же модель: предполагается наличие стохастической зависимости предсказываемого столбца y_{i_0} (отклика) от части остальных столбцов 2-таблицы - $y_{i_1},\ldots,y_{i_p}$ (предикторов): $y_{i_0} = f(y_{i_1},\ldots,y_{i_p}) + \varepsilon_{i_0}$, где ε_{i_0} - случайная ошибка. Какая интерпретация может быть предложена для данной модели в каждой из инцидентных 2-таблиц?

1. В таблице TOC это - стандартная интерпретация. Каждой i-й строке $\tilde{y}_i$ таблицы соответствует *вектор характеристик* (свойств) i-го объекта, измеренных в фиксированный момент времени t_0 ((i,t_0)-столбец): $\tilde{y}_i = \{y(i,1,t_0),\ldots,y(i,j,t_0),\ldots,y(i,n,t_0)\}$, $i = 1,\ldots,m$. Таблица, таким образом, состоит из некоторого множества векторов $\tilde{y}_1,\ldots,\tilde{y}_m$ пространства признаков $y_1,\ldots,y_n$. Предполагается наличие зависимости признака y_{j_0} от части остальных на некотором подмножестве $i \in I$ объектов: $y_{j_0} = f(y_{j_1},\ldots,y_{j_p}) + \varepsilon_{j_0}$. То есть в данной таблице мы имеем дело с наиболее распространенной моделью анализа зависимостей.

ПРИМЕР. Производство шерсти (y_0) в овцеводческих колхозах $i \in I$ зависит от поголовья овец (y_1), количества осадков за лето на пастбищах (y_2), суммы температур за лето (y_3), наличия помещений для овец (тыс.мест, y_4) и т.д.: $y_0 = f(y_1, y_2, y_3, y_4, \ldots)$.

2. В ТОВ i-й строке $\tilde{y}_i$ отвечает одномерный *временной ряд* ((i,j_0)-ряд) значений фиксированного признака j_0 на i-м объекте: $\tilde{y}_i = \{y(i,j_0,1),\ldots,y(i,j_0,t),\ldots,y(i,j_0,q)\}$, $i = 1,\ldots,m$. Таблицу же можно рассматривать как некоторое множество реализаций одномерного случайного процесса. Используемая модель $y_{t_0} = f(y_{t_1},\ldots,y_{t_p}) + \varepsilon_{t_0}$ является обобщением известного процесса авторегрессии порядка p : $y_{t_0} = \sum_{k=1}^{p} \alpha_k y_{t_0-k} + \varepsilon_{t_0}$. То есть предполагается наличие на некотором подмножестве объектов $i \in I$ зависимости между предсказываемым моментом времени и частью остальных моментов.

ПРИМЕР. В колхозах $i \in I$ данного района урожайность пшеницы (y_{86}) в 1986 г. вполне может зависеть от урожайности ее в 1984, 1985, 1988, 1989 гг., если колхозы при прочих равных условиях хозяйствования лежат в одной климатической зоне: $y_{86} = f(y_{84},y_{85},y_{88},y_{89})$.

3. В ТВС t-й строке $\tilde{y}_t$ можно сопоставить *вектор состояния* фиксированного i_0-го объекта в момент времени t ((i_0,t)-столбец): $\tilde{y}_t = \{y(i_0,1,t),\ldots,y(i_0,j,t),\ldots,y(i_0,n,t)\}$, $t = 1,\ldots,q$. Таблица представляет собой многомерный временной ряд $\tilde{y}_1,\ldots,\tilde{y}_m$ состояний объекта i_0 в моменты времени $t_1,\ldots,t_m$. В данном случае предполагается наличие устойчивой во времени (на некотором временном промежутке $t \in T$) зависимости между предсказываемым признаком и частью других признаков: $y_{j_0} = f(y_{j_1},\ldots,y_{j_p}) + \varepsilon_{j_0}$.

ПРИМЕР. На протяжении многих лет $t \in T$ можно проследить довольно устойчивую зависимость количества сдаваемого колхозом i_0 мяса в живом весе (y_0) от количества заготовленных кормов (y_1), емкости мясокомбинатов (y_2) и т.д.: $y_0 = f(y_1,y_2,\ldots)$.

§ 4. Метод шести таблиц

Можно построить некоторые модификации метода инцидентных таблиц, если предположить наличие зависимости между предсказываемой строкой и некоторыми другими строками каждой из инцидентных 2-таблиц: $\tilde{y}_{i_0} = g(\tilde{y}_{i_1}, \ldots, \tilde{y}_{i_p}) + \varepsilon_{i_0}$. О статистических свойствах данных моделей мало что известно. Модель, признающая наличие зависимостей между строками 2-таблицы типа ТОС, рассматривалась в алгоритмах ZET [4,с.58] и в методе многомерной линейной экстраполяции [5,с.49]. Практически данную модификацию можно реализовать, применяя описанный метод к трем 2-таблицам, транспонированным к инцидентным - к ТСО, ТВО и ТСВ. Оптимальным при этом будет прогноз $\hat{y}_0$ с $G = \min\{G_{TCO}, G_{TBO}, G_{TCB}\}$. Этот вариант (*"метод транспонированных таблиц"*) обозначим 3-ТАБТ. Кроме того, можно предложить еще один вариант, в котором оптимальный прогноз отбирался бы среди прогнозов по шести таблицам: по трем инцидентным и трем, транспонированным к ним. Прогноз $\hat{y}_0$ в этом случае соответствует значению $G = \min\{G^3, G^3_T\}$, где G^3 - значение критерия для прогноза по методу 3-ТАБ, а G^3_T - по методу 3-ТАБТ (*"метод шести таблиц"*). Для случая применения к транспонированным таблицам 2-мерного алгоритма ZL трудно найти какие-либо аналоги среди известных моделей.

Следует заметить, что признание наличия зависимостей и по столбцам, и по строкам противоречит требованию некоррелированности наблюдений. На это требование опираются модели регрессии, на которых, в свою очередь, базируется 2-мерный ZL. Однако в таком признании нет необходимости. Процедуру отбора прогнозов можно рассматривать как процесс отбора моделей зависимости, в каждой из которых предполагается наличие зависимости либо по столбцам, либо по строкам, так что метод шести таблиц вполне имеет право на существование.

§ 5. Метод одной таблицы с лагами

В данном методе из одной 2-таблицы $T \in \{TOC, TOB, TBC\}$, задаваемой пользователем, конструируется несколько 2-таблиц V_k, $k \in K$, так называемых *лаговых таблиц*, к которым затем применяется 2-мерный ZL. Здесь $K \subset N$ - некоторое подмножество натуральных чисел, $1 \le k_0+k \le k_T$, $\forall k \in K$, где $k_0 \in \{i_0, j_0, t_0\}$ и $k_T \in \{m, n, q\}$ соответственно для $T \in \{TOC, TOB, TBC\}$. Таблицы V_k (число k называется *лагом*) строятся следующим образом:

$$V_k^{TOC}: \quad v_{ij} = \begin{cases} y(i, j_0, t_0), & j = j_0, \\ y(i, j, t_0+k), & j \ne j_0, \end{cases}$$

$$V_k^{TOB}: \quad v_{it} = \begin{cases} y(i, j_0, t_0), & t = t_0, \\ y(i, j_0+k, t), & t \ne t_0, \end{cases}$$

$$V_k^{TBC}: \quad v_{tjj} = \begin{cases} y(i_0, j_0, t), & j = j_0, \\ y(i_0+k, j, t), & j \ne j_0. \end{cases}$$

Здесь $i = 1, \ldots, m$, $j = 1, \ldots, n$, $t = 1, \ldots, q$. Множество K задается пользователем. В случае TOC удобно определять K в виде некоторого интервала времени $[t_1, t_2]$. Строго говоря, слово "лаг" означает "запаздывание", т. е. имеется в виду задержка во времени эффекта влияния предикторов на отклик; в данном методе "лаг" понимается очень широко: как разность любых двух значений фиксированного индекса, одно из которых относится к отклику, а другое - ко всем предикторам. Будем (в случае TOC) говорить, например, о "лаге вперед" $(K = [0, t_2], t_2 > 0)$, о "лаге назад" $(K = [t_1, 0], t_2 < 0)$ и даже (в случае TOB и TBC) о лаге не по времени, а по индексам, соответствующим объектам и свойствам.

По каждой из таблиц V_k, $k \in K$, образованных из 2-таблицы T, находятся прогноз $\hat{y}_k^0(T)$ и значение критерия G_k^T. Оптимальным считается прогноз при $k = k'$, так что $G = G_{k'}^T = \min_{k \in K} G_k^T$, а $\hat{y}_0 = \hat{y}_{k'}^0(T)$.

Какой смысл имеют эти "лаговые зависимости"?

1. <u>ТОС-(t_0,k)</u>. Предполагается, что на некотором множестве объектов $i \in I$ значения признака y_{j_0} в момент t_0 зависят от значений признаков $y_{j_1}, \ldots, y_{j_p}$ в момент (t_0+k): $y_{j_0}^{t_0} = f(y_{j_1}^{t_0+k}, \ldots, y_{j_p}^{t_0+k})$.

ПРИМЕР. В ряде колхозов $i \in I$ производство овощей в 1988 г. зависит от а) количества внесенных удобрений, б) количества парников, в) цены на овощи (все в 1986 г.). Здесь лаг равен (-2).

2. <u>ТОВ-(j_0,k)</u>. Предполагается, что на некотором множестве объектов $i \in I$ значения признака y_{j_0} в момент t_0 зависят от значений признака y_{j_0+k} в моменты $t_1, \ldots, t_p$:

$$y_{j_0}^{t_0} = f(y_{j_0+k}^{t_1}, \ldots, y_{j_0+k}^{t_p}).$$

ПРИМЕР. В колхозах $i \in I$ производство овощей в 1988 г. зависит от количества внесенных удобрений в 1987, 1985, 1983, 1982 годах.

3. <u>ТВС-(i_0,k)</u>. Утверждается, что в некотором интервале времени $t \in T$ значения признака y_{j_0} для i_0-го объекта зависит от значений признаков $y_{j_1}, \ldots, y_{j_p}$ для объекта (i_0+k): $y_{i_0 j_0} = f(y_{i_0+k, j_1}, \ldots, y_{i_0+k, j_p})$.

ПРИМЕР. В одном районе - несколько колхозов. В первом сеют только гречиху, в остальных - только пшеницу, овес и рожь. Так как погодные условия одни и те же во всех колхозах, то урожайность гречихи в первом колхозе на каком-то временном промежутке, по-видимому, может косвенно зависеть от урожайности пшеницы, овса и ржи в каждом из остальных колхозов. Так что если в данных о гречихе есть пропуски, их можно попробовать восстановить по данным соседних колхозов, хотя описанная зависи - мость и не носит характера причинной связи. Те же примеры можно было бы привести и в предыдущем методе (§ 3), ведь инцидентная таблица - частный случай лаговых таблиц (при $k = 0$).

В *методе инцидентних таблиц с лагами* прогноз вычисляется аналогично, только лаговые таблицы теперь строятся для каждой из инцидентных таблиц $T \in \{TOC, TOB, TBC\}$:

$$G = G^{T'}_{k'} = \min_{T} \left(\min_{k \in K}\{G^{T}_{k}\}\right); \quad y_0 = y^{0}_{k'}(T') \ .$$

§ 6. Экстраполяция

Алгоритм, рассмотренный здесь, не предназначен для экстраполяции. Если 3-таблица состоит из 2-таблиц $TOC-1, \ldots, TOC-q$, а требуется получить прогнозы элементов таблицы $TOC-(q+1)$, то 3-мерный ZL будет бесполезен. То же можно сказать и о 2-мерном ZL по отношению к 2-таблице типа TOB - $(q+1)$-й столбец недоступен. Однако существует хороший прием, позволяющий использовать ZL и при продолжении временных рядов. В работе [4,с.63] для получения прогноза элемента x_{N+1} предлагается ряд $x_1, \ldots, x_N$ располагать сегментами длины k в виде строк некоторой матрицы $Y = \{y_{ij}\}$:

$$Y = \begin{pmatrix} x_1 & \dots & x_{k-1} & x_k \\ x_2 & \dots & x_k & x_{k+1} \\ \dots & \dots & \dots & \dots \\ x_{N-k+1} & \dots & x_{N-1} & x_N \\ x_{N-k+2} & \dots & x_N & x_{N+1} \end{pmatrix},$$

где $x_{N+1} = PROB$. Обозначим через $y_1, \dots, y_k$ столбцы матрицы Y. В этой таблице имеет смысл применять алгоритм ZL для предсказания элемента x_{N+1}, так как многие известные модели временных рядов порождают в данной таблице простые зависимости между столбцами типа $y_k = f(y_1, \dots, y_{k-1})$. Например, если ряд имеет полиномиальный тренд $x_t = \alpha_p t^p + \alpha_{p-1} t^{p-1} + \dots + \alpha_1 t + \alpha_0$, $p \le k-1$, то из теории конечных разностей следует, что между откликом y_k и предикторами будет иметь место линейная зависимость $y_k = \sum_{i=1}^{p} \alpha_i y_{k-i}$, которую можно использовать для прогноза. То же верно и для процесса авторегрессии порядка $p \le k-1$: $x_t = \sum_{k=1}^{p} \alpha_k x_{t-k} + \varepsilon_t$, т.е. $y_k = \sum_{i=1}^{p} \alpha_i y_{k-i} + \varepsilon_t$. При заполнении в таблице Y пробела x_{N+1} непараметрический вариант алгоритма ZL дает оценку, близкую к непараметрической оценке авторегрессии по m_0 ближайшим соседям, предложенную Коломбом [6]. Он приводил условия, при которых эта оценка является состоятельной.

В упомянутых известных моделях p обычно рекомендуется брать небольшим, поэтому и k следует задавать малым, чтобы $N-k-1 \ge k$. Это важно еще и потому, что роль объема выборки играет $(N-k-1)$, а числа оцениваемых параметров - $(p+1)$.

К тому же одно аномальное значение в ряде x_t способно испортить по диагонали сразу k строк матрицы Y.

Если имеется набор временных рядов, объединенных 2-таблицей (типа TOB), то этим приемом ее можно превратить в 3-таблицу, преобразуя в 2-таблицу каждый ряд. Заполнив затем пробелы в полученной 3-таблице 3-мерным алгоритмом ZL, получим прогнозы на один момент времени вперед. Если требуется для 3-таблицы Y получить 2-таблицу прогнозов TOC-(q+1), следует для каждого $1 \le j \le n$ применять данный прием к TOB-j, в результате чего будут вычислены прогнозы для элементов $y(*,j,q+1)$. Все вместе они дадут оценку для TOC-(q+1).

§ 7. Оценки времени

Если алгоритм работает в режиме заполнения, то пробелы вначале следует найти. Процедура поиска требует количества операций порядка $O(mnq)$. При редактировании элементов поиск не требуется, поэтому этот член отсутствует. Сам процесс обработки элементов требует операций порядка

$$O(n_p(l_q t'(m,n) + l_n t'(m,q) + l_m t'(q,n))).$$

Здесь n_p - количество пробелов (редактируемых элементов); l_q, l_n, l_m - количества лаговых таблиц, образованных соответственно из таблиц TOC, TOB, TBC, а $t'(m,n)$ - количество операций, необходимое для обработки одного пробела в 2-таблице $(m \times n)$, см. [3]. В методе инцидентных таблиц $l_q = l_m = l_n = 1$; в методе одной таблицы с лагами два из трех значений l_q, l_m, l_n равны нулю.

§ 8. Эксперименты

Дополнительные возможности 3-мерного алгоритма ZL по сравнению с 2-мерным исследовались в экспериментах с редактированием 16 сельскохозяйственных показателей по 15 республикам

СССР за 12 лет (1970-81 гг.), взятых из статистического ежегодника "Народное хозяйство СССР" за соответствующие годы. Таким образом, в нашем распоряжении оказалась 3-входовая таблица (15x16x12). Для прогноза использовались данные по всем 15 республикам, но редактировались данные только по 13, так как данные по РСФСР и УССР проявили явно выраженную аномальность , при которой по крайней мере по двум 2-таблицам - ТОС и ТОВ - можно было заранее ожидать огромную ошибку прогноза, что и показали предварительные эксперименты. В качестве признаков были взяты 16 показателей:

- производство мяса (тыс.т), масла животного (тыс.т),масла растительного (тыс.т), консервов (млн.банок),сахара (тыс.т), яиц (млн.шт.);

- валовый сбор зерновых (тыс.т), овощей (тыс.т), фруктов и ягод (тыс.т), картофеля (тыс.т);

- поставка минеральных удобрений (тыс.т);

- энергетические мощности на 1 рабочего (л.с.), на 100 га (л.с.);

- расход кормов (млн.т);

- численность сельского населения (тыс.чел.).

Редактированию пятью различными методами подверглась часть (13x16) таблицы ТОС-1979, т.е. данные по 13 республикам за 1979 год (208 элементов). Среди этих методов:

- 2-мерный ZL, примененный к ТОС-79 (ТОС);

- метод инцидентных таблиц (3-ТАБ);

- метод одной таблицы с лагами, таблица ТОС (ТОС-Л);

- то же, таблица ТОВ (ТОВ-Л);

- то же, таблица ТВС (ТВС-Л).

Кроме того, так как 3-мерный ZL базируется на 2-мерном, у этих методов есть разновидности, соответствующие вариантам 2-мерного ZL, а именно: применялись варианты локально-парамет - рические - ZL-СТ (ступенчатый), ZL-ПШ (пошаговый) и непарамет-

рические - ZL-НСПА (с выбором столбцов алгоритмом SPA) и ZL-НСТ (со ступенчатым выбором). Подробнее о вариантах в [3].

Каждый вариант каждого метода вычислял прогнозы $\hat{y}_{ij}$ для всех 208 элементов y_{ij}, после чего для каждого варианта подсчитывались средняя относительная ошибка (R) и средняя медианная ошибка (S):

$$R = \frac{1}{208} \sum_{i=1}^{13} \sum_{j=1}^{16} \frac{|y_{ij} - \hat{y}_{ij}|}{|y_{ij}|};$$

$$S = \frac{1}{208} \sum_{i=1}^{13} \sum_{j=1}^{16} \frac{|y_{ij} - \hat{y}_{ij}|}{AMO_j},$$

где $AMO_j = \underset{i}{\text{med}}|y_{ij} - \underset{i}{\text{med}}\, y_{ij}|, i = 1,\ldots,13$; S используется вместо обычной среднеквадратической ошибки (S_k), которая отличается от S тем, что в знаменателе на месте AMO_j стоит стандартное отклонение, потому что в 2-таблицах работает медианная нормировка столбцов, а не обычная - по дисперсиям: $y'_{ij} = (y_{ij} - \underset{i}{\text{med}}\, y_{ij})/ AMO_j$. Эта нормировка позволяет бороться с сильной неоднородностью данных по республикам. Нужно помнить, что S на неоднородных данных, как правило, значительно больше по величине S_k. Предварительные эксперименты показали, что, например, 2-мерные алгоритмы на ТОС-79 дали S порядка 180-230%, а S_k в это время была порядка 22-30%. Кроме R и S, вычислялись медианы тех и других ошибок - R_m и S_m. Для прогноза использовались подматрицы размером (6x4) для всех 2-таблиц.

Результаты работы каждого варианта приведены в таблице. В каждой клетке таблицы представлены 4 ошибки: в верхней части - R (слева) и S (справа), внизу - R_m и S_m. Из полученных результатов можно сделать следующие выводы.

Таблица

Методы	Варианты методов							
	ZL-СТ		ZL-ПШ		ZL-НСПА		ZL-НСТ	
	R	S	R	S	R	S	R	S
ТОС	134.3 49.1	236.7 81.7	129.7 43.6	210.7 64.9	73.4 44.9	199.3 54.9	110.0 46.3	189.0 71.4
З-ТАБ	16.7 6.1	51.2 13.7	22.9 6.5	49.6 15.7	26.3 12.8	104.0 20.5	22.8 10.9	100.6 19.3
ТОС-Л	114.4 45.6	233.4 73.2	131.6 49.8	232.9 75.2	66.8 41.4	178.6 63.2	69.5 40.7	170.6 61.1
ТОВ-Л	206.7 36.6	290.5 63.3	18.2 6.1	40.1 10.7	46.4 19.1	146.4 25.6	158.5 42.2	281.2 83.2
ТВС-Л	10.0 5.6	93.0 51.5	10.2 5.0	111.6 52.2	10.1 5.3	96.8 53.8	9.7 5.7	92.3 55.4

1. Все 2-мерные алгоритмы на таблице ТОС-79 дали очень плохие результаты: 70-140% (R), 190-240% (S). Это,по-ви - димому, свидетельствует об отсутствии каких-либо четких зави - симостей между признаками. Стоит обратить внимание на то, что ZL-НСПА, как всегда, в неблагоприятной ситуации оказался в наилучшем положении, дав $R = 73.4\%$. Для сравнения по той же таблице был пущен алгоритм ZET [4,с.58], который составил конкуренцию ZL-НСПА, дав близкую $R = 78.4\%$ (в режиме поиска зависимостей по столбцам, а при поиске как по столбцам, так и по строкам $R = 99,9\%$).

2. Поиск лаговых зависимостей (ТОС-Л) по этой же таблице ТОС не спас положения - прогнозы заметно не улучшились, за исключением разве что ZL-НСТ - на 20%(R) и ZL-НСПА - на 8%.

3. Поиск лаговых зависимостей (ТОВ-Л) по таблице ТОВ привел к превосходному результату для ZL-ПШ - достигнут табличный минимум $S = 40.1\%$, $S_m = 10,7\%$ (S уменьшилась в 5 раз против ТОС-79). Так как R при этом не так уж и мала - 18.2%,

то, по-видимому, найдены линейные зависимости. Этот результат достигнут на фоне гигантских ошибок остальных алгоритмов.

4. Первый успех: на лаговых таблицах (ТВС-Л) всем алгоритмам удалось получить достаточно низкие значения R - порядка 10% (табличный минимум R достигнут здесь же - для ZL-НСТ - 9,7%). Таким образом, удалось уменьшить R в 7-13 раз по сравнению с 2-мерными алгоритмами. Большая величина S при этом указывает на то, что хороший прогноз достигнут за счет малой дисперсии столбцов, а не благодаря явно выраженной зависимости. В данной таблице низкая дисперсия столбцов является следствием прекращения ряда с/х показателей в анализируемые годы.

5. Второй успех: метод инцидентных таблиц (3-ТАБ) позволил сократить S в 4 раза - до 49,6% (ZL-ПШ) и до 51,2% (ZL-СТ), что говорит о найденных зависимостях, близких к линейным. Величина R при этом тоже сократилась: в 3-5 раз для ZL-НСПА и ZL-НСТ и в 6-8 раз для ZL-СТ и ZL-ПШ. По сравнению с предварительными экспериментами при обычной нормировке, применение медианной нормировки позволило данному методу сократить R на 2-10%.

В целом можно сказать, что 3-мерный ZL позволил значительно улучшить 2-мерные прогнозы путем использования дополнительных зависимостей из 3-входовой таблицы.

Дополнительно были проведены эксперименты с увеличенными размерами подматрицы - (8x6). При этом 2-мерные алгоритмы, как и следовало ожидать, не только не улучшили прогнозов на ТОС-79, а, как, например, ZET, даже ухудшили (на 8%, R) результаты. Зато метод инцидентных таблиц дал значительные улучшения:

R - 6,5% (ZL-СТ); 9,7% (ZL-ПШ); 4,6% (ZL-НСПА);

S - 23,1% (ZL-СТ); 20% (ZL-ПШ); 25,8% (ZL-НСПА).

К сожалению, время тоже увеличилось и не позволило применить ZL-НСПА к лаговым таблицам с подматрицей (8x6).

Заключение

Эксперименты выявили значительные резервы улучшения прогноза при использовании ретроспективных данных. Развитие 3-мерных методов прогнозирования имеет вполне ощутимые перспективы. Хотелось бы только предостеречь от чрезмерного увлечения лаговыми зависимостями и вообще от злоупотребления сравнением большого количества зависимостей по какому-либо критерию с целью нахождения оптимальной зависимости для прогноза по ней. Найти надежный критерий, т.е. хорошую оценку ожидаемой ошибки, очень трудно. Поэтому разумно при прогнозировании пробелов ограничиваться методом инцидентных таблиц; метод же лаговых таблиц и остальные лучше использовать как инструмент разведочного анализа данных, а не для механического перебора прогнозов с целью отбора оптимального.

Литература

1. COPPI R. Analysis of three-way data matrices based on pairwise relation measures //COMPSTAT-86. Proc. in Comput. Statist. 7th Symp. - Wien, 1986. - P. 122-127.

2. ЖАМБЮ М. Иерархический кластер-анализ и соответствия.- М.: Финансы и статистика,1988.- 342 с.

3. ЗАГОРУЙКО Н.Г., УЛЬЯНОВ Г.В. Локальные методы заполнения пробелов в эмпирических таблицах// Настоящий сборник. - С.75-103.

4. ЗАГОРУЙКО Н.Г., ЁЛКИНА В.Н., ЛБОВ Г.С. Алгоритмы обнаружения эмпирических закономерностей.- Новосибирск:Наука,1985.- 110 с.

5. РАСТРИГИН Л.А., ПОНОМАРЕВ Ю.П. Экстраполяционные методы проектирования и управления.- М.: Машиностроение,1986.-120с.

6. COLLOMB G. Non parametric time series analysis and prediction: uniform almost sure convergence of the window and k-NN autoregression estimates// Statistics.-1985.-Vol.16, N 2.- P.297-308.

7. ДАЙИТБЕГОВ Д.М., КАЛМЫКОВА О.В.,ЧЕРЕПАНОВ А.И. Программное обеспечение статистической обработки данных.- М.: Финансы и статистика,1984.- 192 с.

Поступила в ред.-изд.отд.
26 августа 1988 года