МЕТОД АНАЛИЗА МНОГОМЕРНЫХ ВРЕМЕННЫХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ КОРРЕКТИРОВКИ ПРЕДВАРИТЕЛЬНО РАССЧИТАННОЙ ОБРАТНОЙ МАТРИЦЫ: ИССЛЕДОВАНИЕ В СРАВНЕНИИ С ДРУГИМИ МЕТОДАМИ Data Mining

Г.И. Перминов,
к.т.н, доцент кафедры бизнес-аналитики Государственного университета - Высшей школы экономики

В ходе анализа многомерных временных рядов применение традиционных статистических методов определяется соблюдением достаточно строгих предпосылок, позволяющих использовать лежащий в основе этих методов МНК. К ним относятся: отсутствие мультиколлинеарности, гетероскедастичности и автокорреляции. В задачах экономического анализа и многомерного прогнозирования с целью уменьшения числа рассматриваемых переменных и быстрого получения приблизительных закономерностей целесообразно прибегнуть к методам интеллектуального анализа данных. Методы интеллектуального анализа данных позволяют решить проблемы определения структуры математической модели и вырождения обратной матрицы, когда статистические методы не дают должного результата.

1. Проблема выбора структуры математической модели

До настоящего времени основное внимание уделялось вопросам параметрической идентификации, тогда как структурная идентификация системы считалась заданной. Реальные задачи практики часто имеют дело с плохо структурированными данными, когда неизвестна не только сама модель, но и принадлежность её к тому или иному классу: линейная или нелинейная, детерминированная или стохастическая и так далее. Для практики важна проблема структурной идентификации систем. Эта проблема решается в таких методах интеллектуального анализа, как эволюционное и генетическое программирование, построение, обучение и анализ с помощью искусственных нейронных сетей и др.

2. Проблема вырождения обратной матрицы[1]

При возникновении проблемы мультиколлинеарности диагональные элементы матрицы, обратной к матрице системы нормальных уравнений, соответствующие линейно зависимым аргументам, обращаются в бесконечность, что и приводит к возникновению проблемы вырождения обратной матрицы.

Здесь рассматривается метод решения этой проблемы – не рассчитывать заново обратную матрицу при подключении/удалении в модель новых членов, а корректировать обратную матрицу, полученную на предыдущем шаге.

Предлагаемый метод базируется на 3-х теоремах (доказательство теорем 1, 2, 3 здесь опускается):

Теорема 1. При добавлении в модель новой переменной нет необходимости рассчитывать обратную матрицу заново. Можно скорректировать ранее вычисленную обратную матрицу по предложенному правилу.

Пусть $x_1, x_2,\ldots,x_n$ – линейно независимые векторы, x_{n+1} – вектор той же размерности, что и x_i.

Определим матрицы X_n и X_{n+1}; Φ_n и Φ_{n+1}:

$$X_n=(x_1, x_2,\ldots,x_n),\quad X_{n+1}=(X_n, x_{n+1})$$
$$\Phi_n=X_n{}^T X,\quad \Phi_{n+1}=X_{n+1}{}^T X_{n.} \qquad (1)$$

Утверждается, что

$$\Phi_{n+1}^{-1}=\begin{bmatrix} \Phi_n^{-1}+\frac{\lambda\,\lambda^T}{\rho} & -\frac{\lambda^T}{\rho} \\ -\frac{\lambda^T}{\rho} & \frac{1}{\rho} \end{bmatrix} \qquad (2)$$

[1] В разработке алгоритма принимал участие Трубицын Н.Ф.

Здесь $\lambda^T = (\lambda_1, \lambda_2, ..., \lambda_n)$ – коэффициенты линейной комбинации $x_1, x_2, ..., x_n$, аппроксимирующие x_{n+1} по методу наименьших квадратов;

ρ – сумма квадратов погрешностей аппроксимации.

Теорема 2. Если обратная матрица рассчитана, то при удалении переменной из модели нет необходимости определять новую обратную матрицу, достаточно скорректировать имеющуюся матрицу по определённому правилу.

Если $$\Phi_{n+1}^{-1} = \begin{bmatrix} \Phi_n^* & \alpha \\ \alpha^T & C \end{bmatrix}, \text{ то}$$

$$\Phi_n^{-1} = \Phi_n^* - \frac{\alpha\alpha^T}{C} \qquad (3)$$

Φ_n и Φ_{n+1} определены выше в (1) и (2).

Здесь Φn^* – матрица nxn,

a – n – мерный вектор,

C – скаляр.

Теорема 3. Здесь описываются рекуррентные процедуры по включению в модель новых членов и выбрасыванию старых

Пусть

$$\widetilde{x}_i = (x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n),\ \widetilde{\Phi}_i = \widetilde{X}_i^T \widetilde{X}_i .$$

Утверждается, что

$$\widetilde{\Phi}_i^{-1} = \Phi_{n-1}^* - \frac{\alpha^* \alpha^{*T}}{C^*} \qquad (4)$$

Здесь Φ^*_{n-1}, a^*, C^* определяются соотношением

$$\Phi_n^{*-1} = \begin{bmatrix} \Phi_{n-1}^* & \alpha^* \\ \alpha^{*T} & C^* \end{bmatrix}$$

Матрица Φ^*_{n-1} получается из матрицы Φ_{n-1} путём перемещения i-ой строки и i –го столбца в конец.

3. Алгоритм структурно-параметрической идентификации модели, порождающей наблюдаемый процесс

Пусть имеем многомерный процесс с r входами и s выходами.

Цель исследования – поиск механизма, порождающего данный процесс.

Будем представлять процесс следующей феноменологической моделью:

- ✧ выход процесса с r входами и s выходами определяется настоящим и прошлым значением входа процесса

$$u_i(k), ..., u_i(k-n_i),\ i = 1, 2, ..., r$$

- ✧ и прошлыми значениями выходных сигналов

$$z_j(k-1), ..., z_j(k-m_j),$$
$$j=1, 2, .., s,$$

n_i – глубина памяти i-го входа u_i,

m_j – глубина памяти j-го выхода z_j.

Представим информацию об истории процесса в виде:

$$X=\{u_1(k), ..., u_1(k-n_1); ...; u_r(k), ...,$$
$$u_r(k-n_r); z_1(k), .., z_1(k-m_1); ...;$$
$$z_s(k-1), ..., z_s(k-m_s)\},$$

или

$$X=\{x1(k), x2(k), ..., xp(k)\} \qquad (5)$$

Число членов выражения (5) равно

$$p = r + \sum_{i=1}^{r} n_i + \sum_{j=1}^{s} m_j \qquad (6)$$

Пусть нас интересует некоторый a-ый выход $z(k)$.

Представим его в виде нелинейной полиномиальной регрессионной модели:

$$z(k) = a_0 + \sum_{i=1}^{p} a_i x_i(k) + \sum_{i=1}^{p}\sum_{j=i}^{p} a_{ij} x_i(k) x_j(k) +$$
$$+ \sum_{i=1}^{p}\sum_{j=i}^{p}\sum_{l=j}^{p} a_{ijl} x_i(k) x_j(k) x_l(k) + ... +$$
$$+ \sum_{i=1}^{p}\sum_{j=i}^{p} ... \sum_{q=v} a_{ij...vq}(k) ... x_v(k) x_q(k) \qquad (7)$$

q – степень нелинейности.

Заметим, что сложность модели быстро растет с увеличением q. Если p – число прессоров, то число членов модели (6) равно

$$n = \frac{(p+q)!}{p!q!} \qquad (8)$$

Здесь $n!$ = 1*2*3*.....*т.

По нашему мнению следует ограничиться значениями $q <= 4$ и $p < =20$. При этом максимально допустимое число членов будет 10626.

Задача идентификации модели процесса заключается в поиске регрессионных параметров

$$[A_0,\ A_1, A_2, ..., A_{n-1}] = A^T \qquad (9)$$

Здесь $A_0=a_0$, $A_1=a_1$, ..., $A_p = a_p$, $A_{p+1} = a_{11}$, ...

Введем сигнальный вектор – вектор регрессоров

$$V^T(k)=[V_0(k), V_1(k),..., V_{n-1}(k)],$$

где

$V_0(k)=1$
$V_1(k)=u1(k)$
……
$V_{n1+1}(k)=u_1(r-n_1)$
…

$$V_{n-2}(k) = S_s^{q-1} S_{s-1} = S_s^{q-1}(k-m_s) S_{s-1}(k-m_{s-1})$$

$$V_{n-1}(k) = Z_s^q(k-m_s). \quad (10)$$

Теперь модель (7) запишется в виде

$$Z(k)=V^T(k)A \quad (11)$$

Так как модель линейная по параметрам, то элементы вектора A можно определять по методу наименьших квадратов (МНК). Однако непосредственное применение МНК затруднительно по следующим причинам:

1) высокая размерность (много параметров, некоторые из них присутствуют несколько раз с различными лагами, комбинации регрессоров),

2) слабая обусловленность обратной матрицы из-за мультиколлинеарности.

Введение большого количества регрессоров не только усложняет структуру модели, но и обязательно вносит мультиколлинеарность, что приводит к увеличению погрешности. Поэтому определение рациональной структуры модели – важная задача, требующая нетривиальных поисков.

Для структурной идентификации модели результирующей переменной требуется найти простой метод выбора существенных регрессоров. Предлагаемый алгоритм основывается на трех вышеприведённых теоремах.

Идея предлагаемого алгоритма заключается в выборе «перспективных на существенность» членов в массиве исходных данных для включения в модель, при этом после выбора следующего члена делается проверка, нет ли «неперспективных на существенность» членов в модели. Их следует исключить.

Прямой перебор возможных регрессоров в исходном массиве с проверкой перспективности по тем или иным критериям с решением задач МНК на каждом этапе включения члена в модель делает задачу практически неразрешимой. Использование теорем 1–3 позволяет свести прямой пересчёт к поправкам, это даёт возможность избежать многочисленных обращений матриц – самой трудоёмкой операции метода.

Шаг 1.

Задать параметры *r, ni, s, m_j, q* и *N*.

Задать уровень значимости F_a F теста. Например, $F_{0,05} = 3{,}84 + 9{,}9/N$.

Вычислить число максимально возможных членов

$$n = \frac{(p+q)!}{p!q!},$$

где

$$p = r + \sum_{i=1}^{r} n_i + \sum_{j=1}^{s} m_j.$$

Сформировать векторы:

$Z, V_0, V_1,..., V_{n-1}$ (см. выражение 11).

Шаг 2.

Положить $k = 1$. Это означает, что сначала выбирается модель, включающая только один член.

$$R_k=[V_0]$$

$$\Theta_k^{-1} = (R_k^T R_k) = \frac{1}{V_0^T V_0} = \frac{1}{N}$$

$$Q_k = \bar{z} = \frac{1}{N}\sum_{\alpha}^{N} z(\alpha)$$

$$e_k^{(\alpha)} = z(\alpha) - \bar{z}, \quad a=1, 2, ... , N.$$

$$B_k=N\ ln\ S^2{}_k + k\ lnN$$

$$S_k^2 = \frac{1}{N}\sum_{\alpha=1}^{N} e_k^2(\alpha).$$

Шаг 3.

Для оставшихся $n-k$ возможных членов V_i^*, $i = 1, 2, ..., n-k$ вычислим коэффициенты их представления в виде комбинации текущих членов модели

$$\tilde{\alpha}_i = \Theta_k^{-1} R_k^T V_i^*$$

чтобы оценить, что даёт (какую новую информацию) член V_i^*:

$$\tilde{E}_i = V_i^* - R_k\tilde{\alpha}_i \quad \in [\tilde{e}_i(1), \tilde{e}_i(2),..., \tilde{e}_i(N)]^T.$$

Вычислим

$$Q_i = \frac{[\sum_{\alpha=1}^{N} \tilde{e}_i(\alpha) z(\alpha)]^2}{\sum \tilde{e}_i{}^2(\alpha)}, i = 1,..., n-k.$$

Здесь a – индекс суммирования.

Шаг 4.

Выберем член с максимальным значением Q_i, обозначим его $V_{\max}$, соответствующие $\tilde{a}_i$ и $\tilde{E}_i$ обозначим $\tilde{a}_{\max}$ и $\tilde{E}_{\max}$. Сформируем новую матрицу данных: $R_{k+1}=[R_k V_{\max}]$.

Вычислим $C_{\max} = \tilde{E}^T_{\max}\ \tilde{E}_{\max}$. Теперь матрица $\theta^{-1}_{k+1} = (R^T_{k+1}\ R_{k+1})^{-1}$ может быть вычислена, используя приведённые математические факты, весьма просто, без фактического обращения.

$$\Theta_{k+1}^{-1}=\begin{bmatrix}\Theta_k^{-1}+\tilde{\alpha}_{\max}\tilde{\alpha}_{\max}^T & \ldots & -\tilde{\alpha}_{\max}/C_{\max}\\ \ldots. & \ldots & \ldots\\ -\tilde{\alpha}_{\max}^T & \ldots & 1/C_{\max}\end{bmatrix}.$$

Шаг 5.

Вычислим параметры модели, погрешности и информационный критерий F с учетом добавления $V_{\max}$ в модель.

$$\alpha_{k+1}=\Theta_{k+1}^{-1}R_{k+1}^T z$$

$$Ek+1=z\text{-}Rk+1ak+1=def[ek+1(1),\ldots,ek+1(N)]T$$

$$F=\frac{\sum_{\alpha=1}^{N}e_k^2-\sum_{\alpha=1}^{N}e_{k+1}^2(\alpha)}{\sum_{\alpha=1}^{N}e_{k+1}^2(\alpha)}\ (\text{N-k-1})$$

$$S_{k+1}^2=\frac{1}{N}\sum_{\alpha=1}^{N}e_{k+1}^2(\alpha)$$

$$B_{k+1}=N\ lnS^2{}_{k+1}+(k+1)lnN$$

Шаг 6.

Если $B_{k+1} < B_k$ и $F >= F_{0,05}$, то включаем $V_{\max}$ в модель ($k => k+1$) и переходим к шагу 7.

Иначе (если $B_{k+1} > B_k$ или $F < F_a$) новые члены в модель не включаются и процедура построения структуры модели заканчивается.

Шаг 7.

Вычислим для всех k членов F и BIC:

F_i, B_{k-1}, i, $i = 1, \ldots, k$, т.е. критериальные значения, когда отбрасывается i-ый член из k-членной модели. Для этого отбросим i-ый вектор из матрицы R_k и получим матрицу $R_{k-1,i}$ (обозначим через max).

Сдвинем i-ую строку матрицы Φ_k^{-1} и i-ый столбец в конец (вниз). Изменённую матрицу Φ_k^{-1} обозначим как:

$$\Theta_{k,i}^{-1}=\begin{bmatrix}\Theta & \lambda\\ \lambda^T & C\end{bmatrix}.$$

Тогда

$$\Phi^{-1}{}_{k-1,i}=[R^T{}_{k-1,i}R_{k-1,i}]^{-1}=\Theta-\frac{\lambda\,\lambda^T}{C}$$

$$a_{k-1,i}=\Phi^{-1}{}_{k-1,i}\,R^T{}_{k-1,i}\,z$$

$$z=[z(1),\ z(2),\ldots,\ z(N)]^T$$

$$E_{k-1,i}=z\text{-}R_{k-1,i}\,a_{k-1,i}$$

$$F_i=\frac{\sum_{\alpha=1}^{N}e_{k-1,i}^2(\alpha)-\sum_{\alpha=1}^{N}e_k^2(\alpha)}{\sum_{\alpha=1}^{N}e_k^2(\alpha)}\quad(\text{N-k})$$

$$S_{k-1,i}^2=\frac{1}{N}\sum_{\alpha=1}^{N}e_{k-1,i}^2(\alpha)$$

$$B_{k-1,i}=N\ lnS^2{}_{k-1,i}+(k-1)lnN$$

Шаг 8.

Рассмотрим вычисленные на шаге 7 величины. Обозначим наименьшие значения F_i и $B_{k-1,i}$ через $F_{\min}$ и $B_{\min}$ соответственно.

Если $B_{\min} < B_k$, то выбросим соответствующий член, $k <= k\text{-}1$, $R_k <= R_{k,\min}$, $\Phi_k^{-1} <= \Phi_{k,\min}^{-1}$, $a_k <= a_{k,\min}$ и переходим на шаг 7.

Если $F_{\min} < F_a$, отбрасываем соответствующий член и переходим к шагу 7.

Если отброшен последний выбранный член или когда все возможные члены включены в модель, процедура заканчивает построение модели. Во всех других случаях необходимо перейти к шагу 3.

Замечание. В алгоритме шаги 1 и 2 инициируют алгоритм, шаги 3, 4, 5 и 6 используются при выборе члена, шаги 7 и 8 используются при отбрасывании члена.

4. Программная реализация алгоритма и сравнение результатов, полученными различными методами Data Mining

Сравнение результатов, полученных различными методами интеллектуального анализа данных, и с применением предлагаемого алгоритма проводилось на многомерном массиве макроэкономических показателей России с результирующей переменной «Средний индекс РТС».

Описание исходных данных представлено в *табл. 1*:

Приведем некоторые результаты:

4.1. Предлагаемая модель (установлена предельная степень модели – квадратичная) (*рис. 1*)

Квадратичная модель имеет вид:

*RTS=50,4590625+0,00747140*IMQ(t-4)**

*RTS(t-1)-9,72723960*INVFC(t-2)*RTS(t-8)*

В квадратичную модель, помимо самого «Среднего индекса РТС (RTS_M)» с лагом 1 и 8 месяцев, вошли «Индекс производства – добыча полезных ископаемых (IMQ_C)» с лагом 4 месяца и «Инвестиции в основной капитал (INVFC_M)» с лагом 2 месяца.

4.2. Эволюционный алгоритм «Поиск законов (Find Law)» пакета PolyAnalyst

Алгоритм FL предназначен для автоматического нахождения в данных нелинейных зависимостей (вид которых не задаётся пользователем) и представления результатов в виде математических формул, включающих в себя и блоки условий. Алгоритм основан на технологии эволюционного, или генетического, программирования. Поскольку структура и параметры модели эволюционного программирования значительно зависят от расчётного времени, приведём два варианта – расчётное время 0,2 и 0,8 минут.

Таблица 1

Имя	Описание	Комментарий
Time	Дата, которой соответствует исследуемый показатель.	Месяц, год
UNEMPL_M	Количество безработных (на конец месяца) (UNEMPL_M)	млн. чел
EMPLDEC_M	Заявленная потребность в работниках (на конец месяца) (EMPLDEC_M)	тыс. чел.
LESN_SA	Индекс производства Лесное хозяйство и предоставление услуг в этой области, месячный, сглаженный, с сезонной и календарной корректировкой (LESN_SA)	1995.I = 100
LESN	Индекс производства Лесное хозяйство и предоставление услуг в этой области, месячный, исходный ряд (LESN)	1995.I (факт) = 100
FISH_SA	Индекс производства Рыболовство, месячный, сглаженный, с сезонной и календарной корректировкой (FISH_SA)	1995.I = 100
FISH	Индекс производства Рыболовство, месячный, исходный ряд (FISH)	1995.I (факт) = 100
IMQ_C_SA	Индекс производства Добыча полезных ископаемых, месячный, сглаженный, с сезонной и календарной корректировкой (IMQ_C_SA)	1995.I = 100
IMQ_C	Индекс производства Добыча полезных ископаемых, месячный, исходный ряд (IMQ_C)	1995.I (факт) = 100
EPNG_SA	Индекс производства Добыча сырой нефти и природного газа, месячный, сглаженный, с сезонной и календарной корректировкой (EPNG_SA)	1995.I = 100
EPNG	Индекс производства Добыча сырой нефти и природного газа, месячный, исходный ряд (EPNG)	1995.I (факт) = 100
MEEP_SA	Индекс производства Добыча полезных ископаемых, кроме топливно-энергетических, месячный, сглаженный, с сезонной и календарной корректировкой (MEEP_SA)	1995.I = 100
MEEP	Индекс производства Добыча полезных ископаемых, кроме топливно-энергетических, месячный, исходный ряд (MEEP)	1995.I (факт) = 100
IPCDE_SA	Индекс Промышленность (C+D+E) , месячный, сглаженный, с сезонной и календарной корректировкой (IPCDE_SA)	1995.I = 100
IPCDE	Индекс Промышленность (C+D+E) ,месячный, исходный ряд (IPCDE)	1995.I (факт) = 100
RTRD_M_DIRI	Индекс реального оборота розничной торговли (RTRD_M_DIRI)	1994.1 = 100
RTRD_M_DIRI_SA	Индекс реального оборота розничной торговли, с поправкой на сезонность (RTRD_M_DIRI_SA)	1994.1 (факт) = 100
RTRD_M	Оборот розничной торговли в текущих ценах (RTRD_M)	млрд. руб.
WAG_R_M	Реальная зарплата (WAG_R_M)	янв. 93 = 100
WAG_R_M_SA	Реальная зарплата с поправкой на сезонность (WAG_R_M_SA)	янв. 93 (факт) = 100
WAG_C_M	Средняя номинальная заработная плата (WAG_C_M)	рублей в месяц
INVFC_M	Инвестиции в основной капитал (INVFC_M)	млрд. рублей
RDEXRO_M	Официальный курс доллара (RDEXRO_M)	руб/долл.
RDEXRM_M	Курс доллара на ММВБ (RDEXRM_M)	руб/долл.
RTS_M	Средний индекс РТС (RTS_M)	пункты
IB_M	Межбанковская ставка (IB_M)	% годовых
GKO_M	Доходность ГКО (GKO_M)	% годовых
DEP_M	Депозитная ставка (DEP_M)	% годовых
CR_M	Ставка по кредитам (CR_M)	% годовых
RTS_CLASS	Рост или падение среднего индекса РТС	1 – рост 0 – падение

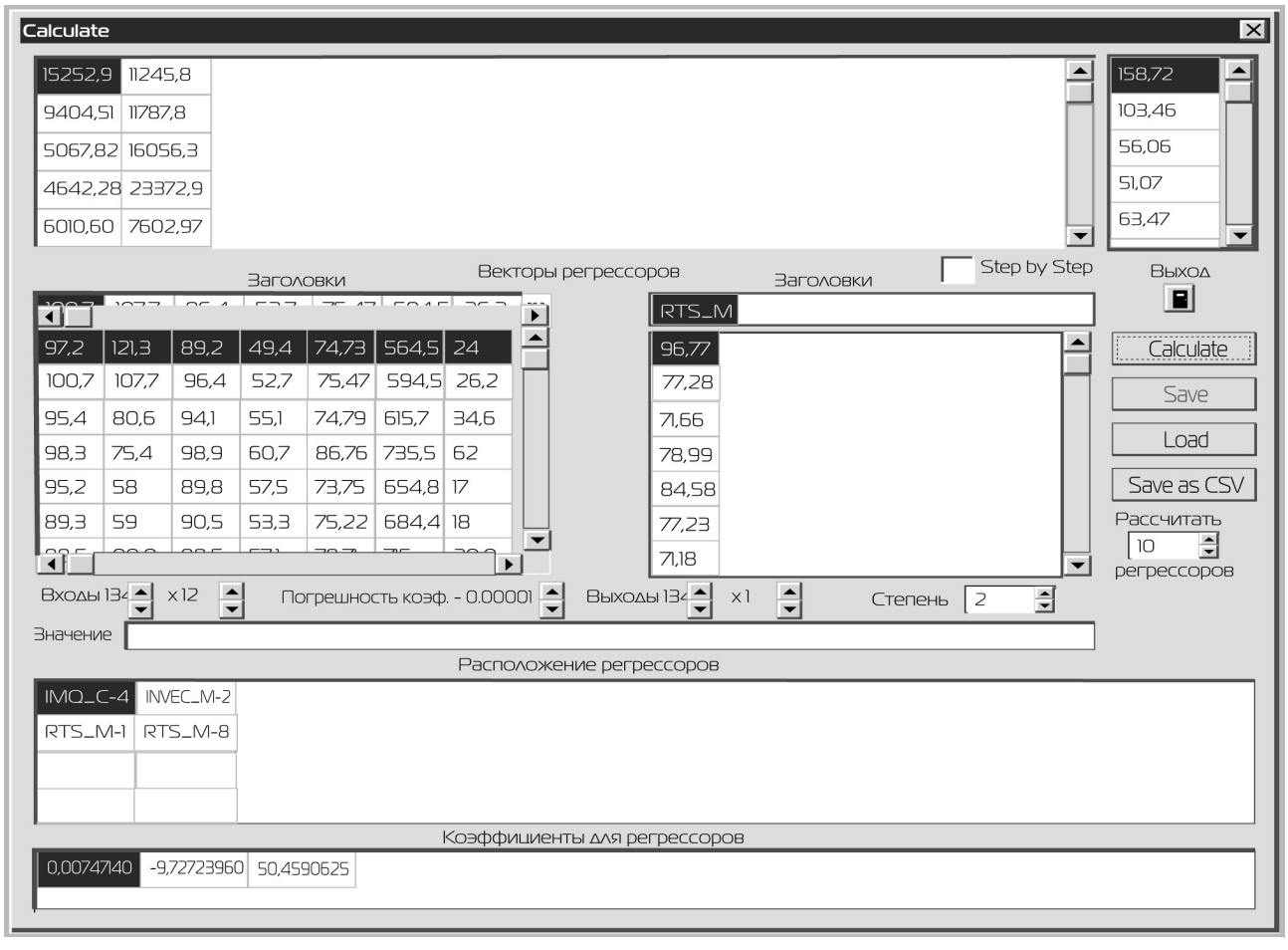

Рис.1. Результаты расчёта по квадратичной модели

4.2.1. Поиск законов 0,2 минуты (FL 1)

Лучшее по значимости правило:

RTS_M = -0.03275 WAG_C_M+0.000231926*RTRD_M_DIRI_SA*RTRD_M_DIRI_SA*RTRD_M_DIRI_SA*

Лучшее по точности правило:

*RTS_M = (-0.0177095 *RTRD_M_DIRI_SA* WAG_C_M*WAG_C_M-19035.2 *WAG_C_M-14411.8 *RTRD_M_DIRI_SA*RTRD_M_DIRI_SA+1.47402e++008)/(RTRD_M_DIRI_SA*WAG_C_M--261.836 * WAG_C_M)*

Результирующие показатели модели:

Критерий	Стандартная ошибка	Стд. отклонение	Значимость	R-sq.
Наибольшая значимость	0.2868	108.4	9.478	0.9178
Наибольшая точность	0.2668	100.9	1.35	0.9288

4.2.2. Поиск законов 0.8 минут (FL 2)

Лучшее по значимости правило:

RTS_M = (0.097235
**RDEXRM_M*RDEXRM_M*DEP_M*
**RTRD_M_DIRI_SA*
**RTRD_M_DIRI_SA-689.648*
**RDEXRM_M*RDEXRM_M*DEP_M-3690.4*
**RDEXRM_M*DEP_M-44640.4*
**RDEXRM_M+2.27516e+006)/(RDEXRM_M*
**RDEXRM_M*DEP_M+6.78154*
**RDEXRM_M*RDEXRM_M+190.493 *GKO_M)*

Лучшее по точности правило:

RTS_M = (0.0998343
**IB_M*RDEXRM_M*RDEXRM_M*DEP_M*
**RTRD_M_DIRI_SA*RTRD_M_DIRI_SA-655.071*
**IB_M*RDEXRM_M*RDEXRM_M*DEP_M-4138.87*
**IB_M*RDEXRM_M*DEP_M-77833*
**IB_M*RDEXRM_M+2.43484e+006*
**IB_M+2.52765e+006)/(IB_M*RDEXRM_M*
**RDEXRM_M*DEP_M+5.09747*
**IB_M*RDEXRM_M*RDEXRM_M+190.493*
**IB_M*GKO_M+5077.62)*

Результирующие показатели модели:

Критерий	Стандартная ошибка	Стд. отклонение	Значимость	R-sq.
Наибольшая значимость	0.146	55.19	2.595	0.9787
Наибольшая точность	0.1375	51.99	not signif.	0.9811

4.3. Нейронная сеть (PolyNet Predictor) пакета PolyAnalyst

Получены следующие результаты:

Индекс значимости:	27.14
Стандартная ошибка	0.225
R-squared:	0.9494
Стандартное отклонение	85.39
Обработано точек:	134
Количество слоев сети	1
Количество узлов сети	3

Найдено правило:

```
(242.094+WAG_C_M*(+WAG_C_M
*(2.65394e-005+-2.30264e-009*WAG_C_M))
+DEP_M*(60.3773+DEP_M
*(-2.08283+0.0158524*DEP_M+0.000962676
*WAG_C_M)+WAG_C_M*(-0.050212+4.85822e
-006*WAG_C_M)))
```

4.4. Линейная регрессия (LR) пакета PolyAnalyst

Реализация этого модуля в системе PolyAnalyst имеет свои особенности – автоматический выбор наиболее значимых независимых переменных и тщательная оценка статистической значимости результатов.

На исследуемом наборе данных методом линейной регрессии найдено следующее правило:

```
RTS_M = -461.338 -6.82264*MEEP_SA
+13.5900*RTRD_M_DIRI_SA
-8.40649*RDEXRM_M
-1.24002*IB_M -1.40264*GKO_M
```

Стандарная ошибка	0.3142
R-squared	0.9013
Станд.откл	119.3
Обработано точек	134
Индекс значимости	60.19

4.5. Модель Data Mining (MS Time Series) пакета Microsoft SQL Server 2005

В результате работы алгоритма получено следующее правило:

```
RTS_M=-139.68-0.26*RTS_M(-2)+1.53
*RTDD_M_DIRI(-4)-
-0.16*RTS_M(-8)+0.23
*RTS_M(-3)+1.06*RTS_M(-1)
```

4.6. Линейная регрессия (Microsoft Linear Regression) пакета Microsoft SQL Server 2005

Алгоритм линейной регрессии на тех же данных дал следующий результат:

```
RTS_M=397.55+4.158*(WAG_R_M-93.43)
-4.45*(IPCDE-97.76)+0.153*(Date-36845.27)-
-2.39*(INVFC_M-129.17)-10.94*(EPNG-112.98)
+18.004*(RTRD_M_DIRI-126.07)
```

4.7. Построение обобщенной модели

Всего было просчитано 20 моделей, включая определение логических правил, искусственная нейросеть пакета «Статистика» и т.д.

Заложенные в каждом методе ИАД различные идеи приведут к большому разнообразию и разбросу результатов. Отбросить «плохие» результаты считаем рискованным, т.к. конкретный метод, рассматривая выборку под своим углом зрения, может увидеть особенности, не улавливаемые другими методами.

Для решения этой проблемы реализуем идею, высказанную Э.Б. Ершовым – можно попытаться найти то общее, что выражается в результатах всех методов (регрессия на главных факторах).

Этот способ объединения частных результатов (прогнозов) состоит в том, чтобы представить комбинированную модель в виде взвешенной суммы частных результатов. Сумма всех весов равна 1, и сами веса находятся в интервале [0,1]. Основная проблема, которая здесь возникает, – это определение весов, поскольку именно они будут характеризовать качество объединённой модели.

Один из возможных методов этого направления объединения прогнозов – использование факторного анализа. В факторном анализе пытаются определить новые переменные, так называемые факторы F_j в значительно меньшем количестве, но наиболее полно воспроизводящие и отражающие исходные переменные X_i. Эти факторы представляют собой линейную комбинацию исходных признаков и находятся из различных условий (чаще всего максимизации суммы квадратов коэффициентов корреляции факторов F_j и признаков X_i). Поэтому новые факторы содержат максимум информации, заключённой в исходных признаках. Идея применения факторного анализа для построения обобщённой модели основана на том, что частные результаты

расчёта, полученные по *i*-му методу прогнозирования $x_{i,t}$ ($i = 1, 2, ..., n$), являются внешним выражением некоторой реально существующей, но непосредственно неизмеримой прогнозной величины. Она и принимается в качестве обобщённого прогноза.

Математически это можно записать так:

$$x_{i,t} = l_i f_t + e_{i,t} \qquad (12)$$

где $x_{i,t}$ – частные прогнозы;

f_t – обобщённый прогноз, обусловливающий корреляционную связь между частными прогнозами;

l_i – нагрузка (вес) обобщенного прогноза ft на частный прогноз xi,t;

$e_{i,t}$ – остаток (характерный показатель), определяющий ту часть прогноза $x_{i,t}$, изменение которой вызвано действием случайных причин.

Выражение, приведённое выше, является моделью факторного анализа с одним генеральным фактором. При этом можно выразить обобщённый прогноз через линейную комбинацию частных прогнозов с весами a_i как регрессию на генеральном факторе.

В случае получения нескольких факторов обобщённый прогноз можно получить через взвешенную сумму регрессий на каждом факторе.

Результаты получения долей каждого метода в обобщённой модели показали примерное равенство рассмотренных методов, как в количественном, так и в качественном анализе.

4.8. Оценка качества модели как вычисление близости к обобщенной модели

Для оценки качества модели был применён метод «ближайший сосед» пакета PolyAnalyst, позволяющий определить степень близости частных прогнозов к обобщённому. Были получены следующие результаты:

Стандартное отклонение: 17.9174
Стандарная ошибка (R sq.): 0.048410 (0.997657)
Индекс значимости: 51.880119

Упорядоченные близости частных рассчитанных моделей к обобщенной приведены в *табл. 2.*

Таблица 2

Метод		Фактор расстояния
Название	**Пояснения**	
Lag_2	Квадратичная модель с корректировкой обратной матрицы	0.00180431
FL2_CLASS	Эволюционный алгоритм со временем расчета 0,8 мин пакета PolyAnalyst	0.00183658
NearNeigh	Ближайший сосед пакета PolyAnalyst	0.00183888
MS_TimeSer_2	MS Time Series пакета Microsoft SQL Server 2005	0.00184421
Neuro_CLASS	Нейросеть пакета Статистика	0.00184485
Lag_1	Линейная модель с корректировкой обратной матрицы	0.00184665
LR_CLASS	Линейная регрессия пакета PolyAnalyst	0.00184958
FL1_CLASS	Эволюционный алгоритм со временем расчета 0,2 мин пакета PolyAnalyst	0.00185628
Neuro	Нейросеть пакета PolyAnalyst	0.00186457
FL2	Эволюционный алгоритм пакета PolyAnalyst	0.00186558
FL1	Эволюционный алгоритм пакета PolyAnalyst	0.00187148
DR1	Дерево решений пакета PolyAnalyst	0.00187591
WizWhy	Построение логических правил пакета WizWhy	0.00188255
DR-Chaid	Дерево решений пакета Clementine	0.00188766
S5	Построение логических правил пакета Clementine	0.00192955
CRT	Построение логических правил пакета Clementine	0.0019315
FD	Нахождение зависимостей Find Dependencies	0.00193986
See 5	Построение логических правил пакета See 5	0.0019784
MS_LR	Линейная регрессия пакета Microsoft SQL Server 2005	0.00201159
MS_TimeSer_1	MS Time Series пакета Microsoft SQL Server 2005	0.00229252

Выводы по сравнению моделей

1. Первый компонент факторного анализа объясняет 96,363 % всей вариации частных моделей, что говорит о тесной корреляции между расчётными данными по всем представленным моделям Data Mining.

2. Вклад частных моделей в обобщённую модель практически одинаков и составляет 0,042–0,045.

3. В частные модели вошли разные показатели (всего вошло 26 показателей). Поэтому уточнённый

расчёт статистическими методами в уменьшенном поле переменных нежелателен.

4. Наиболее близкими к обобщённому прогнозу оказались: квадратичная многомерная модель с лагами и корректировкой обратной матрицы (Lag_2), эволюционные методы пакета PolyAnalyst с поиском структуры модели и «ближайший сосед» (NearNeigh), модель Data Mining MS Time Series пакета MS SQL Server 2005, искусственная нейронная сеть пакета «Статистика» и линейная многомерная модель с лагами и корректировкой обратной матрицы (Lag_1).

5. Время расчёта исходной матрицы размером 30×135 по линейной и квадратичной моделям с лагами и корректировкой обратной матрицы составляет до 1 минуты. Кубическая модель рассчитывается уже десять – двадцать мин., а модель четвёртого порядка –около часа.

6. Предлагаемый алгоритм и его программная реализация делают возможным получать результаты с достаточной точностью с автоматическим нахождением структуры и параметров модели в приемлемое время.

7. Применение в программной реализации критериев прекращения расчётов Акайке и ВИС, упрощающих модель, привело к получению выражения с малым количеством членов, но с достаточно высокой точностью результатов. Так, ниже приведены некоторые модели:

Линейная модель ВВП:

$ВВП_t$=39.118749+1.33908416*Объём промышленного производства$_{t-0}$ +0.11246734*Цена на нефть$_{t-2}$.

Квадратичная модель ВВП:

$ВВП_t$=67.3921875+0.00858213*Индекс цен на строительно-монтажные работы$_{t-0}$* Валовой внутренний продукт$_{t-1}$+ 0,01677165* Официальный курс доллара (на конец периода)$_{t-8}$* Цена нефти$_{t-1}$.

Линейная модель среднедушевых денежных доходов:

Среднедушевые денежные доходы = 205,7412 +1,0093*Валовый внутренний продукт (с лагом 0) – 16,3865* Официальный курс доллара на конец периода (с лагом в 9 кварталов). ■

Литература

1. Akaike H. A new look at the atatistical identification model //IEEE: 1074. –V.19–716–723 p
2. Andrew C. Harvey. Forecasting, Structural Time Series Models and the Kalman Filter. Econometric Theory. 1991,
3. Bollerslev T. Generalized autoregressive conditional heteroscedasticity//Journal of econometrics. –1986, V.31. 307–327 h.
4. Christian Gourieroux and Alain Monfort. Time Series and Dynamic Models// Themes in Modern Econometrics 1996, 425 c.
5. Аболенцев Ю. И., Кильдшиев Г. С. Статистическая адекватность регрессионных моделей и проблема мультиколлинеарности //Экономика и математические методы 1984. Т. XX. Вып. 6.
6. Айвазян С.А. Интеллектуализированные инструментальные системы в статистике и их роль в построении проблемно-ориентированных систем поддержки принятия решений. «Обозрение прикладной и промышленной математики», том 4 (1997), № 2. М.: Научное изд-во ТВП.
7. Гарбер Е. В., Горелик Н.А., Френкель А. А. Развитие адаптивных методов прогнозирования временных рядов// Статистические методы анализа экономической динамики: Ученые записки по статистике. Т. XLVI. М.: Наука, 1983.
8. Ивахненко А.Г., Мюллер Й.А. Самоорганизация прогнозирующих моделей. Киев: Наук. думка, 1985.
9. Канторович Г.Г. Анализ временных рядов. Экономический журнал Высшей школы экономики. Том.6. № 1. № 2. № 3. 2002.
10. Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов. М.: Финансы и статистика, 2003.
11. Савараги Е., Соэда Т., Накамизо Т. Классические методы и оценивание временных рядов. Гл. 2. Современные методы идентификации систем. М.: Мир, 1983.