ГЛАВА 6
СГЛАЖИВАНИЕ ВРЕМЕННЫ́Х РЯДОВ

6.1. ЛИНЕЙНЫЕ ФИЛЬТРЫ

Одним из распространенных способов выявления тренда является сглаживание временно́го ряда. Его суть сводится к замене фактических значений ряда расчетными, полученными после удаления высокочастотных колебаний. Поэтому сглаживание ряда часто называют фильтрованием, а преобразование ряда (оператор), с помощью которого осуществляется фильтрование, — фильтром.

Наиболее часто на практике используются линейные фильтры. Общая формула линейного фильтра такова

$$Y(t) = \sum_{r=-l}^{k} a_r X(t+r), \tag{6.1}$$

где $Y(t)$ — сглаженное (отфильтрованное) значение временного ряда в момент времени t;

a_r — вес, приписываемый значению исходного ряда, находящемуся на расстоянии r от рассматриваемого момента времени t.

Фильтр (6.1) учитывает k значений (уровней) ряда после момента времени t и l уровней до него. Число $k + l + 1$ значений исходного ряда, одновременно участвующих в сглаживании, называется шириной интервала сглаживания. Если $k = l$, то сглаживание центрированное. Сглаженный ряд короче исходного ряда на $k + 1$ значение. В зависимости от выбора ширины интервала сглаживания, величины весов a_r применяются различные методы сглаживания, самым простым из которых является метод простой скользящей средней.

6.2. МЕТОД ПРОСТОЙ СКОЛЬЗЯЩЕЙ СРЕДНЕЙ

Если $\sum a_r = 1$ и a_r = const, то фильтр (6.1) означает вычисление средней арифметической, которую называют скользящей средней.

Для удобства сопоставления сглаженного и исходного рядов ширину интервала сглаживания чаще выбирают нечетным числом $m = 2k + 1$. Тогда $a_r = 1/m$ и из (6.1) получаем

$$Y(t) = \frac{1}{2k+1} \sum_{r=-k}^{k} X(t+r). \tag{6.2}$$

Такое сглаживание будет симметричным ($a_{-r} = a_r$) и центрированным. Чаще всего для сглаживания берут $m = 3; 5; 7$. Если дисперсия уровней исходного ряда постоянная и равна σ^2, а сами члены ряда $X(t_i)$ независимы между собой, то дисперсия сглаженного ряда $Y(t)$ равна σ^2/m. Таким образом, при увеличении колеблемости исходного ряда $X(t)$ (при наличии большой дисперсии σ^2) для уменьшения амплитуды колебаний у сглаженного ряда $Y(t)$ необходимо увеличивать ширину интервала сглаживания m либо проводить процедуру сглаживания повторно. Кстати, по степени уменьшения дисперсии у повторно сглаженных рядов можно судить о степени зависимости между собой членов исходного ряда, т.е. о наличии «долговременной памяти» у исходного ряда.

Если ряд имеет периодические колебания с продолжительностью цикла меньше m, то они полностью исчезают при сглаживании с помощью скользящей средней с шириной интервала сглаживания m.

Расчет $Y(t)$ при $m > 3$ можно упростить, применяя рекуррентную формулу

$$Y(t) = Y(t-1) + \frac{X(t+k) - X(t+k-1)}{2k+1}, \quad t = k+2; k+3; \ldots, n-k. \qquad (6.3)$$

При расчетах по формуле (6.3) первое сглаженное значение $Y(k+1)$ вычисляется по формуле (6.2)

$$Y(k+1) = (X(1) + X(2) + \ldots + X(m))/m, \; m = 2k+1.$$

Недостатком метода простой скользящей средней является равное участие в сглаживании значений ряда, отстоящих от момента сглаживания t на разном расстоянии. В результате этого могут быть потеряны важные для анализа свойства ряда. Рассмотрим более «тонкие» методы сглаживания.

6.3. МЕТОДЫ ВЗВЕШЕННЫХ СКОЛЬЗЯЩИХ СРЕДНИХ

Суть методов взвешенных скользящих средних заключается в том, что значениям исходного ряда приписывается вес ar, зависящий от расстояния до середины интервала сглаживания, т. е. от $|r|$. Тогда $a - r = ar$ и сглаживание по этим методам является центрированным и симметричным. Для определения весов прибегают к различным подходам.

Рассмотрим первый подход. Пусть весами являются члены разложения бинома $(0{,}5 + 0{,}5)^{2k}$, $m = 2k + 1$. Тогда

$$a_r = C_{2k}^{k-r}(0{,}5)^{k-r}(0{,}5)^{2k-(k-r)},$$
$$r = 0, 1, ..., k. \qquad (6.4)$$
$$a_r = C_{2k}^{k-r}(0{,}5)^{2k},$$

где C_{2k}^{k-r} — число сочетаний из $2k$ элементов по $k - r$ элементов.

При этом $a_{-r} = C_{2k}^{k-(-r)} = C_{2k}^{k+r}$. По свойству сочетаний имеем

$$C_{2k}^{k-r} = C_{2k}^{2k-(k-r)} = C_{2k}^{k+r} = \frac{(2k)!}{(2k-k-r)!(k+r)!} = \frac{(2k)!}{(k-r)!(k+r)!}.$$

Получаем:
при $m = 3$ ($k = 1$) $a_{-1} = 1/4$, $a_0 = 1/2$, $a_1 = 1/4$;
при $m = 5$ ($k = 2$) $a_{-2} = 1/16$, $a_{-1} = 1/4$, $a_0 = 3/8$, $a_1 = 1/4$, $a_2 = 1/16$;
при $m = 7$ ($k = 3$) $a_{-3} = 1/64$, $a_{-2} = 3/32$, $a_{-1} = 15/64$, $a_0 = 5/16$, $a_1 = 15/64$, $a_2 = 3/32$, $a_3 = 1/64$.

Второй подход заключается в подборе полинома регрессии к данным, содержащимся в интервале сглаживания. При этом свободный член a полинома регрессии выбирается равным расчетному значению ряда $Y(t)$.

Рассмотрим случай, когда уравнение регрессии квадратичное, т.е. сглаживание происходит на основе уравнения параболы. В этом случае для каждого набора m последовательных членов исходного ряда составляется система $m = 2k + 1$ уравнений для расчета по методу наименьших квадратов:

$$Z(t + i) = a + bi + ci^2; I = -k; -k + 1; -1; 0; 1; ...; k - 1; k, \qquad (6.5)$$

где $Z(t + i)$ — расчетные значения квадратичного уравнения регрессии.

Сглаженное значение ряда $Y(t)$ выбирается по формуле

$$Y(t) = Z(t + 0) = a. \qquad (6.6)$$

Запишем систему нормальных уравнений для определения коэффициентов параболы a, b, c

$$\left.\begin{cases} am + b\sum\limits_{i=-k}^{k} i + c\sum\limits_{i=-k}^{k} i^2 = \sum\limits_{i=-k}^{k} X(t+i) \\ a\sum\limits_{i=-k}^{k} i + b\sum\limits_{i=-k}^{k} i^2 + c\sum\limits_{i=-k}^{k} i^3 = \sum\limits_{i=-k}^{k} iX(t+i) \\ a\sum\limits_{i=-k}^{k} i^2 + b\sum\limits_{i=-k}^{k} i^3 + c\sum\limits_{i=-k}^{k} i^4 = \sum\limits_{i=-k}^{k} i^2X(t+i) \end{cases}\right\}. \qquad (6.7)$$

Так как $\sum_{i=-k}^{k} i = \sum_{i=-k}^{k} i^3 = 0$, получаем из системы (6.7)

$$\left\{\begin{array}{l} am + c\sum_{i=-k}^{k} i^2 = \sum_{i=-k}^{k} X(t+i) \\ b\sum_{i=-k}^{k} i^2 = \sum_{i=-k}^{k} iX(t+i) \\ a\sum_{i=-k}^{k} i^2 + c\sum_{i=-k}^{k} i^4 = \sum_{i=-k}^{k} i^2 X(t+i) \end{array}\right\}. \quad (6.8)$$

Исключив c из первого и третьего уравнений системы (6.8), получаем формулу для расчета коэффициента a

$$Y(t) = a = \frac{\sum_{i=-k}^{k} i^4 \sum_{i=-k}^{k} X(t+i) - \sum_{i=-k}^{k} i^2 \sum_{i=-k}^{k} \left(i^2 X(t+i)\right)}{m\sum_{i=-k}^{k} i^4 - \sum_{i=-k}^{k} i^2}. \quad (6.9)$$

При $m = 5$, $k = 2$ имеем

$$a = -(3/35)X(t-2) + (128/35)X(t-1) + (17/35)X(t) + + (12/35)X(t+1) - (3/35)X(t+2), \quad (6.10)$$

т.е. $a_r = -3/35;\ 12/35;\ 17/35;\ 12/35;\ 3/35$.

Аналогично для $m = 7$, $k = 3$ и для $m = 9$, $k = 4$ соответственно получаем

$$a_r = -2/21;\ 3/21;\ 6/21;\ 7/21;\ 6/21;\ 3/21;\ -2/21, \quad (6.11)$$

$$a_r = -21/231;\ 14/231;\ 39/231;\ 54/231;\ 39/231;\ 14/231;\ -21/231. \quad (6.12)$$

Можно легко проверить, что если в качестве сглаживающего многочлена взять прямую, то коэффициенты $a_r = 1/m$, т.е. совпадут с коэффициентами сглаживания с помощью метода простой скользящей средней.

Предположим, дисперсия уровней исходного ряда постоянная и равна σ^2, а сами члены ряда $X(t_i)$ независимы между собой. В этом случае дисперсия сглаженного по квадратичному полиному ряда $Y(t)$ равна

$$\sigma_y^2 = \frac{3(3k^2 + 3k - 1)}{(2k-1)(2k+1)(2k+3)}\sigma^2.$$

При $m=5, k=2\ \sigma_y^2=\frac{17}{35}\sigma^2$, а при $m=7, k=3\ \sigma_y^2=\frac{1}{3}\sigma^2$, т.е. тенденция к уменьшению дисперсии с ростом m сохраняется.

При сглаживании с помощью скользящей средней нет возможности получить сглаженные значения для k первых и k последних членов ряда $X(t)$. В случае сглаживания с помощью полинома регрессии для крайних членов исходного ряда могут быть получены сглаженные значения — значения полинома регрессии в этих точках. Но для этого надо оценить не только свободный член a полинома, но и остальные коэффициенты полинома регрессии (коэффициенты b, c в квадратичном случае). Из системы (6.8) получаем

$$b=\frac{\sum_{i=-k}^{k} iX(t+i)}{\sum_{i=-k}^{k} i^2},\quad c=\frac{\sum_{i=-k}^{k} X(t+i)\ -am}{\sum_{i=-k}^{k} i^2}. \tag{6.13}$$

Для $m = 5$, $k = 2$ из (6.10), (6.13) для начальных значений ряда имеем

$$a = Z(0) = Y(3) = -(3/35)X(1) + (12/35)X(2) + (17/35)X(3) + \\ + (12/35)X(4) - (3/35)X(5), \tag{6.14}$$

$$b=\frac{1}{10}\big(-2X(1)-X(2)+X(4)+2X(5)\big),$$
$$c=\frac{1}{10}\big(X(1)+X(2)+X(3)+X(4)+X(5)-5a\big). \tag{6.15}$$

$$Y(1) = Z(-2) = a + b(-2) + c(-2)^2 = a - 2b + 4c, \tag{6.16}$$

$$Y(2) = Z(-1) = a + b(-1) + c(-1)^2 = a - b + c. \tag{6.17}$$

Для последних пяти членов ряда аналогично получаем (n — объем выборки):

$$a = Z(0) = Y(n-2) = -(3/35)X(n-4) + (12/35)X(n-3) + \\ + (17/35)X(n-2) + (12/35)X(n-1) - (3/35)X(n), \tag{6.18}$$

$$b=\frac{1}{10}\big(-2X(n-4)-X(n-3)+X(n-1)+2X(n)\big),$$
$$c=\frac{1}{10}\big(X(n-4)+X(n-3)+X(n-2)+X(n-1)+X(n)-5a\big). \tag{6.19}$$

$$Y(n-1) = Z(1) = a + b(1) + c(1)^2 = a + b + c, \qquad (6.20)$$

$$Y(n) = Z(2) = a + b(2) + c(2)^2 = a + 2b + 4c. \qquad (6.21)$$

6.4. ПРОСТОЕ ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ

Выше рассмотрены симметричные фильтры для сглаживания временны́х рядов. Вместе с ними широкое распространение при прогнозировании временны́х рядов получили асимметричные фильтры. Простейший из них — скользящая средняя, которая заменяет последнее значение ряда внутри интервала сглаживания,

$$Y(t) = \frac{1}{m}\sum_{r=0}^{m} X(t-r), \quad t = m+1, m+2, ..., n. \qquad (6.22)$$

Из формулы (6.22) следует простая рекуррентная формула

$$Y(t) = Y(t-1) + \frac{1}{m}\big(X(t) - X(t-m)\big), \quad t = m+1, m+2, ..., n. \qquad (6.23)$$

Первое слагаемое в формуле (6.23) указывает на то, что процесс обладает инерцией развития. Второе слагаемое отражает последние изменения в процессе, причем вес, с которым учитывается новое значение, зависит от ширины интервала сглаживания и равен $1/m$.

В предложенном методе (см. (6.22)) прошлые значения временно́го ряда выбирались с постоянным весом. Однако естественно прогнозировать будущие значения ряда с учетом устаревания информации. Один из приемов сглаживания ряда, позволяющих учитывать устаревание данных, является придание веса α текущему наблюдению и веса $1 - \alpha$ предыдущему сглаженному значению ряда

$$Y(t) = \alpha X(t) + (1-\alpha)Y(t-1), \; 0 < \alpha < 1. \qquad (6.24)$$

Если формулу (6.24) последовательно применить к $Y(t-1)$, $Y(t-2)$, ..., $Y(2)$, то получим

$$Y(t) = \sum_{r=0}^{t-1} \alpha(1-\alpha)^r X(t-r). \qquad (6.25)$$

Так как $0 < \alpha < 1$, веса при $X(t-k)$ быстро уменьшаются с ростом k и влияние $X(t-k)$ на $Y(t)$ падает. Аналогично (6.23) можно переписать формулу (6.24)

$$Y(t) = Y(t-1) + \alpha(X(t) - Y(t-1)), \; 0 < \alpha < 1. \qquad (6.26)$$

Из формулы (6.26) видно, что α является коэффициентом поправки на «новизну» в сглаженном временном ряде $Y(t)$. Отметим, наконец, что за $Y(1)$ обычно принимают $X(1)$ и процедуру сглаживания начинают со второго члена ряда.

Для простой экспоненциальной средней, в предположении о независимости членов исходного ряда и постоянности дисперсии σ^2, можно вывести формулу для дисперсии сглаженного ряда

$$D(Y(t)) = D(\alpha X(t) + (1-\alpha)Y(t-1)) = D(\alpha X(t)) + \\ + D((1-\alpha)Y(t-1)) = \alpha^2 D(X(t)) + (1-\alpha)^2 D(Y(t-1)) = \\ = \alpha^2\sigma^2 + (1-\alpha)^2 D(Y(t-1)).$$

Так как $\delta^2 = D(Y(t)) = D(Y(t-1))$, то $\delta^2 = \alpha^2\sigma^2 + (1-\alpha)^2\delta^2$. Отсюда $\delta^2 = \dfrac{\alpha}{2-\alpha}\sigma^2$. Следовательно, дисперсия сглаженного ряда меньше дисперсии исходного ряда.

Теперь рассмотрим вопрос о выборе параметра α. Во многих руководствах рекомендуется выбирать α из промежутка [0,1; 0,3].

Однако в работе С. Макридакиса[11] рассмотрены примеры, когда лучшие результаты сглаживания были получены при $\alpha > 0{,}3$. Поэтому в последних статистических пакетах (например, в пакете STATISTICA) предлагается находить оптимальное значение параметра α методом перебора значений с заранее выбранным шагом h.

Метод простого экспоненциального сглаживания получил популярность главным образом благодаря его привлекательности как инструмента прогноза и простоте реализации. Так, С. Макридакис[12] показал, что простое показательное сглаживание было лучшим выбором для прогноза «один период вперед» из числа 24 других методов. Таким образом, независимо от теоретической модели процесса, лежащей в основе наблюдаемого временнóго ряда, простое показательное сглаживание часто дает весьма точные прогнозы.

6.5. ЭЛЕМЕНТЫ ДИАЛОГА В МОДУЛЕ «АНАЛИЗ ВРЕМЕННЫ́Х РЯДОВ» ПП STATISTICA. ПРОГНОЗИРОВАНИЕ

В переключателе модулей пакета STATISTICA выберем модуль «Анализ временны́х рядов», высветив его и щелкнув ОК либо дважды

[11] *Makridakis S., Andersen A., Carbone R., Fildes R., Hibon M., Lewandowski R., Newton J., Parzen R., & Winkler R.* The accuracy of extrapolation (time series) methods: Results of a forecasting competition // Journal of Forecasting. 1982. N 1.

[12] *Makridakis S.G.* Empirical evidence versus personal experience // Journal of Forecasting. 1983. N 2.

щелкнув на его имени. Появится стартовая панель выбранного модуля (рис. 6.1).

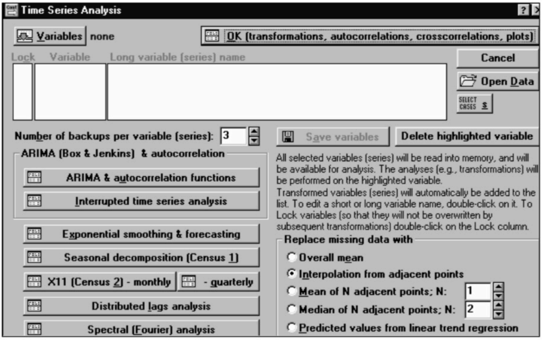

Рис. 6.1

Нажатие кнопки 

(ОТКРЫТЬ ДАННЫЕ) позволяет открыть файл данных. Из открытого файла данных выбираем нужный столбец-переменную последовательным нажатием кнопок 

и ОК, имена выбранных столбцов-переменных появляются в светлом информационном окне.

Нажимаем на панель 

для обработки временнóго ряда методами экспоненциального сглаживания и прогнозирования. Часть этих методов была описана выше. Данные методы позволяют сгладить ряд, выделить из него случайную составляющую-шум и прогнозировать будущие значения.

Открывается панель СГЛАЖИВАНИЕ ВРЕМЕННЫ́Х РЯДОВ (рис. 6.2).

При обработке ряда в модели сглаживания могут быть учтены различные виды трендов и сезонность. Могут также использоваться модели с аддитивным и мультипликативным шумом. Сначала строят график исходного временного ряда. Для этого достаточно перейти в окно таблицы данных, подсветить нужную переменную и нажатием левой кнопки мыши перейти к панели опций, включающих и построение графика (рис. 6.3).

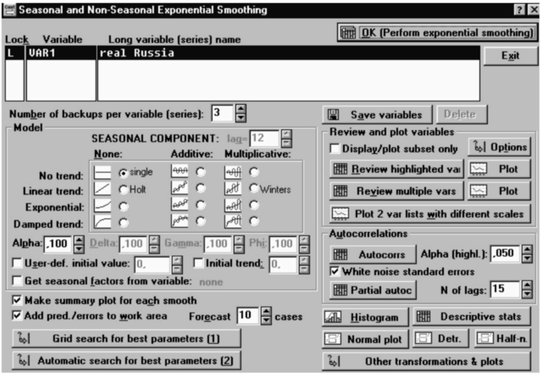

Рис. 6.2

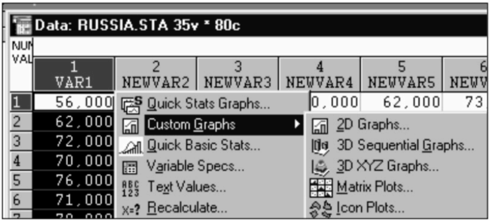

Рис. 6.3

Далее по виду полученного графика исходной переменной (рис. 6.4) подбираем подходящую модель сглаживания, руководствуясь пиктограммами на панели (рис. 6.5).

Исходя из графика (см. рис. 6.4) можно предположить, что в нашем случае подходящей является модель с демпфированным (подавляемым) трендом без сезонной компоненты либо более точная модель с демпфированным трендом и мультипликативной сезонной компонентой. Отметим курсором модель с демпфированным трендом внизу первого слева столбца пиктограмм. Затем начнем сглаживание, нажав на панель 

. В результате получа-

Рис. 6.4

Рис. 6.5

ем таблицу исходных данных, расчетных значений и остатков, фрагмент которой представлен на рис. 6.6.

TIME SERIES	Damped trend, no season ; Alpha=,100 Gamma=,100 Phi=,100 VAR1 : real Russia		
Case	VAR1	Smoothed Series	Resids
24	146,0000	106,3450	39,6550
25	125,0000	110,3560	14,6440
26	168,0000	111,8396	56,1604
27	146,0000	117,5137	28,4863
28	101,0000	120,3966	−19,3966
29	165,0000	118,4410	46,5590
30	139,0000	123,1419	15,8581
31	177,0000	124,7480	52,2520
32	121,0000	130,0275	−9,0275
33	139,0000	129,1212	9,8788
34	106,0000	130,1186	−24,1186

Рис. 6.6

Чтобы получить результаты работы на графике, откроем снова с помощью опции ОКНО панель СГЛАЖИВАНИЕ ВРЕМЕННЫХ РЯДОВ (рис. 6.7).

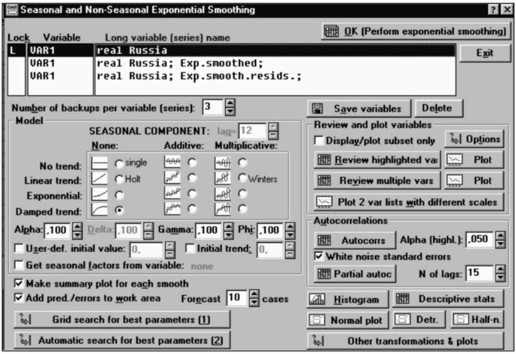

Рис. 6.7

Далее нажимаем последовательно кнопки 

и 

и получаем совместный график исходных значений, сглаженных расчетных значений и остатков, т.е. отклонений расчетных значений от исходных (рис. 6.8).

Рис. 6.8

Как видно из графика (см. рис. 6.8), величины остатков весьма значительны и качество приближения нельзя назвать хорошим. Поэтому применим для сглаживания ряда модель с демпфированным трендом и сезонной компонентой. Для этого в окне СГЛАЖИВАНИЕ ВРЕМЕННЫХ РЯДОВ (см. рис. 6.7) отмечаем курсором данный метод и выбираем длину сезонной компоненты равной трем SEASONAL COMPONENT: lag= 3. Затем в изложенном выше порядке проводим обработку данных и получаем новый график (рис. 6.9).

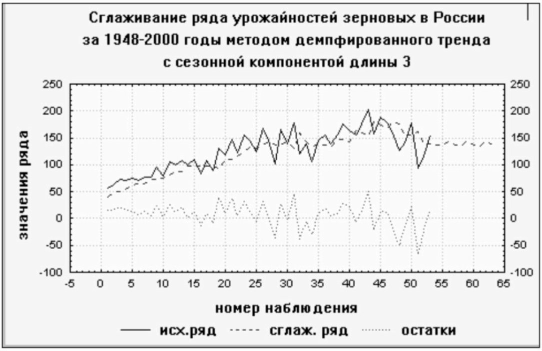

Рис. 6.9

Нетрудно заметить, что качество сглаживания явно улучшилось.

Контрольные вопросы и задания

1. Что собой представляет линейный фильтр для временного ряда?
2. Как строится простая скользящая средняя временного ряда?
3. В чем заключается метод взвешенных скользящих средних?
4. Опишите процесс простого экспоненциального сглаживания временного ряда.