文章编号：1003-207(2000)zj-0203-09

趋势分析的原理、实现和应用

赵立平[1]，赵阳[2]

（1.中国科学院数学所，北京 100080，
2.国家质量技术监督局信息中心，北京 100088）

摘 要：“趋势分析”依据观测得到的数据资料(时间数列)，进行统计计算，找出内在规律,预测今后的发展趋势。时间数列通常包括四种变化成分:季节变动、长期趋势、周期波动和随机变化。这里主要讲述季节变动的移动平均比率法，其它成分调用统计程序包进行计算。

关键词：趋势分析；预报；预测；时间数列；统计；回归分析；相关分析；数据库；Fortran和Delphi

中图分类号：C931.1 **文献标识码**：A

1 趋势分析的作用

在一个大、中型的管理信息或控制系统中，往往都包含有“趋势分析”模块。趋势分析是“决策支持”子系统的基本组成部分，它依据观测得到的历史数据资料，进行统计计算，找出内在规律,预测今后的发展趋势。例如根据上个月商品每天的销售量预报今后一个星期每天的销售量，以便为决定最佳库存量提供依据。

2 算法设计

2.1 时间数列方法

从观测得到的历史数据资料中，按照相等的时间间隔搜集起来的一组统计数据称为“时间数列”。

时间数列通常包括如下四种变化成分:季节变动、长期趋势、周期波动和随机变化。

“季节变动”是指每年重复出现的四季变化规律。例如夏季肠胃炎发病率高,冬季感冒患者增加。

“长期趋势”是指变量的值在很长的时期内呈现的增加或减少趋势。例如人民生活水平的稳定提高。其规律可以是线性的,也可以是非线性的。

“周期波动”是指周期不是一年的波动规律(一年的波动规律已反映在“季节变动”中)。例如商业的周期性兴衰规律。

作者简介：赵立平，(1938—)，男，汉族，山东莒县人，中国科学院数学研究所计算机科学研究室，副研究员，研究方向：计算机语言编译和数据库技术.

“随机变化”是指不规则的变化。例如个别旅客的来访对某一产品销售量的影响。

2.2 回归分析和相关分析

“回归分析”研究变量之间的关系性质和影响强度，它以历史数据为基础，计算出估计公式的系数，用于预测未知变量的预报值。“相关分析”用于确定变量之间的关连程度。

2.2.1 最小平方法

两个变量之间最简单的关系可以用一条回归直线来反映，其公式为：

$$y=a+bx \tag{1}$$

其中： x 为已知自变量，称为“独立变量”；y 为未知因变量，称为“相关变量”；a 为直线在 y 轴上的截距；b 为直线的斜率，即 x 对 y 的影响系数。

为了确定回归直线,需用“最小平方法”,使得在各个点上回归直线的值与观测值的误差平方和取得最小值。

用它确定回归直线待定系数的公式为:

$$b=\frac{\sum XY-n\overline{XY}}{\sum X^2-n\overline{X}^2} \tag{2}$$

$$a=\overline{Y}-b\overline{X} \tag{3}$$

其中:n 为数据点的个数;X 为独立变量的值;Y 为相关变量的值。

$\overline{X}$ 和 $\overline{Y}$ 为独立变量和相关变量的平均值, 即 $\overline{X}=\frac{\sum X}{n},\overline{Y}=\frac{\sum Y}{n}$

2.2.2 相关分析

用它来衡量回归直线所表示的相关变量变化的精确度。它运用“测定系数”和“相关系数”来表示。测定系数是表示变量之间联系范围和强度的基本方法。

$$\gamma^2=1-\frac{\sum(Y-\hat{Y})^2}{\sum(Y-\overline{Y})^2} \tag{4}$$

其中:Y 为观测数据;$\hat{Y}$ 为回归直线上点的数值;$\overline{Y}$ 为 Y 的平均值。

$$\text{相关系数}\ \gamma=\sqrt{\gamma^2} \tag{5}$$

2.2.3 多元回归分析

在实际应用中,独立变量的个数不只限于一个,例如化工产品的质量取决于温度、压力和流量等因素。增加多个独立变量的数据，则可更精确地确定估计公式。

多元回归公式为：

$$y=a_0+a_1x_1+a_2x_2+a_3x_3+\ldots \tag{6}$$

其中:x_1、x_2和x_3等是完全无关的独立变量，它比通常的拟合公式

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots \tag{7}$$

的含义要广泛得多，后者只是前者的极特殊情况，例如在公式（6）中x_1为 x，x_2则可以是x^2，或者e^x，或者lnx等更复杂的函数关系。

3　功能设计

3.1 时间数列方法的实现

3.1.1 季节变动的移动平均比率法

由于在统计程序包中没有这部分程序，需自行编制，所以作较详细的说明。

除了年中的“季节”变动外，还有年中的“月份”变动和月中的“星期”变动，也属此列。这里根据某一物理量（以下称为“流量”）在某一天之前的 30 天数据，找出其中的“星期”变动规律，列表如下：

表 1.　一个月中各天的流量及计算值

月中日	流量	移动总和	移动平均数	移动平均百分数	调和时间数列
(1)	(2)	(3)	(4)=(3)/7	(5)=(2)/(4)*100	(6)=(2)/SJ(I)
1	22.330	--------	--------	--------	15.434
2	19.488	--------	--------	--------	15.913
3	15.865	--------	--------	--------	16.523
4	19.464	110.863	15.838	122.898	15.853
5	22.652	110.191	15.742	143.899	15.874
6	7.950	109.879	15.697	50.647	15.668
7	3.114	109.974	15.711	19.821	15.103
8	21.658	112.050	16.007	135.302	14.969
::	::	::	::	::	::
::	::	::	::	::	::
23	24.840	137.851	19.693	126.136	20.283
24	19.209	138.236	19.748	97.271	20.006
25	24.216	138.140	19.734	122.710	17.723
26	27.860	137.776	19.683	141.549	19.524
27	9.785	137.896	19.699	49.671	19.285
28	3.894	--------	--------	--------	18.886
29	27.972	--------	--------	--------	19.333
30	24.960	--------	--------	--------	20.381

其中：第二行表示各列的编号及各列之间的计算关系。

第(1)列是一个月中各天的编号，第(2)列是实际测得的流量值。

第(3)列是移动总和，例如前 7 天的流量之和是 110.863，放在第(3)列中前 7 行的中间（即第 4 行上）；第 2 至第 8 天的流量之和是 110.191，放在第(3)列的第 5 行上；余类推。由于上面 3 项和下面 3 项无此数据，所以写上“-------”。

第(4)列由第(3)列除以 7 得到，称为“移动平均数”。

第(5)列由第(2)列除以第(4)列，乘上 100 得到，称为“移动平均百分数”。

将第(5)列的数据按星期横向排列放在下表中：

表 2. 月中日的移动平均百分数

星期 1	星期 2	星期 3	星期 4	星期 5	星期 6	星期日
-------	-------	-------	122.898	143.899	50.647	19.821
135.302	116.791	96.034	129.022	148.148	53.228	20.622
144.711	122.489	95.144	119.163	139.807	50.852	21.173
146.750	126.136	97.271	122.710	141.549	49.671	-------
-------	-------					

为了消除偶然因素的影响，将各列中的最大和最小值删去，剩余的项放入下表中：

表 3. 删去最大和最小值的月中日移动平均百分数

星期 1	星期 2	星期 3	星期 4	星期 5	星期 6	星期日
--------	--------	--------	122.898	143.899	50.647	--------
--------	--------	96.034	--------	--------	--------	20.622
144.711	122.489	--------	--------	--------	50.852	--------
--------	--------	--------	122.710	141.549	--------	--------
--------	-------					

基于上述数据计算出下表各值。

表 4. 计算值

计算值	星期 1	星期 2	星期 3	星期 4	星期 5	星期 6	星期日
修正总和	144.710	122.489	96.034	245.608	285.448	101.499	20.622
修正平均数	144.710	122.489	96.034	122.804	142.724	50.750	20.622
星期指数	155.990	132.037	103.519	132.376	153.849	54.705	22.229

其中：“修正总和”为表 3 各列剩余项之和。“修正平均数”为“修正总和”除以剩余项个数。“修正平均数”的总和等于 700.132，用它去除 700.00 得到“调整系数”0.9998，“修正平均数”乘上“调整系数”就得到最后一行的“星期指数”，分别记作 SJ(1)、SJ(2)、...和 SJ(7),即 SJ(1)= 144.683, SJ(2)=122.466, SJ(3)=96.016, SJ(4)=122.781, SJ(5)=142.697, SJ(6)=50.740, SJ(7)=20.618。

星期指数 SJ(i)就反映出一周中各天的内在比例关系，上述的 SJ(i)值表示星期 1、2、4 和 5 的流量较高，周末流量较低。

表 1 中的第(6)列是由第(2)列的流量分别除以相应的星期指数 SJ(i)得到的，称为“调和时间数列”。

“调和时间数列”表示将时间数列中的星期变动因素剔除后所得到的残差，可用于继续进行后三种因素的统计计算。

3.1.2 长期趋势的回归分析法

这里所说的“回归分析法”正是运用 2.2.3 节的多元回归分析方法，不过公式(6)只取头三项，并且 x_1 成为时间 t，x_2 成为 t^2。也就是要找出调和时间数列与时间 t 存在着的线性和平方关系。

时间编码方法：表 1 中的时间为第 1 天、第 2 天、...、第 30 天,我们将时间表示为：

$t = t_0 + \Delta t * i = 1 + \frac{i}{N} = \frac{N+i}{N}$，其中：N=总天数=30，$\Delta t = \frac{1}{N}$，i=1, 2, . . . , 30。

则 $t = \frac{31}{30}, \frac{32}{30}, \ldots$，2。$t^2 = \left(\frac{31}{30}\right)^2, \left(\frac{32}{30}\right)^2, \ldots$，4。

选择适当参数，调用统计程序包中的模块进行计算。

3.1.3 周期波动和随机变化的统计方法

选择适当参数，调用统计程序包中的模块进行计算。

4 具体实现

4.1 概述

若某系统要求对流量、气温和物价进行三种预测。

对于流量、气温和物价来说，只是数据来源有所区别，分别从相应的数据库中取得历史数据，调用共同的程序进行处理。

4.2 Microsoft Fortran PowerStation 4.0 软件和源程序

Microsoft Fortran PowerStation 4.0 软件是在 Windows 环境下的科学计算语言软件,我们用此软件编写各种功能程序。

源程序 w107.for 包括五个部分：

（1） 实现 3.1.1 节的“季节变动的移动平均比率法”算法，计算“季节变动”成分；

（2） 调用统计程序包“多变量多功能逐步回归分析”算法，计算“长期趋势”成分；

（3） 调用统计程序包“隐含周期的识别与提取”算法，计算“周期波动”成分；

（4） 调用统计程序包“ARMA(K,L)模型的识别、估计与检验”算法，计算“随机变化”成分。

（5） 累计各种成分的预报值，分别乘上 SJ(3) 、SJ(4)、SJ(5)、SJ(6)、SJ(7)、SJ(1)和 SJ(2)，并除以 100.00 而得到总的预报值。

在该软件环境下对源程序 w107.for 进行编译和链接编辑,得到可执行文件 w107.exe。

4.3 对数据的特殊处理

(1) 原始数据的特殊处理

在实测环境下运行该程序模块时，发现采集到的原始数据是非常不规律的，调用统计程序包来处理它们，可能发生异常情况。例如：当原始数据全部为 0.00 时，在“季节

变动”的“移动平均比率法”中计算“移动平均百分数”时分母等于零，从而产生数值计算“溢出”错误，甚至导致系统崩溃；在“长期趋势” 的计算中也可能在计算三角函数的值时出现“溢出”错误。因此需对其做某些特殊处理，例如在判断会出现此种情况时，宁可损失一些计算精度，也不允许出现“溢出”错误。

对于特殊的奇异数据进行了必要的“限幅”处理。

(2) 计算结果的特殊处理

在某些应用中，要求计算结果不能出现负数（例如流量或物价），但在某些情况下（例如数值先大后小，直至为零）则会使预报值为负数，此时可以求绝对值，或将此情况通知系统，由后续模块进行适当处理。

进行这些处理后，该软件真正具有实用性。

5. 应用举例

在 Delphi 4.0 环境下建立项目 day30_7.dpr，其中包括主窗体和子窗体，主窗体见图 1,上面是“运行”按钮，下面是“关闭”按钮。运行该项目时，在主窗体上单击“运行”按钮则出现子窗体，以柱状图形式显示出今后七天的预报流量来，见图 2，在右上角图例框中，右侧为星期几，中间为预报值，左侧为各天柱状图案。

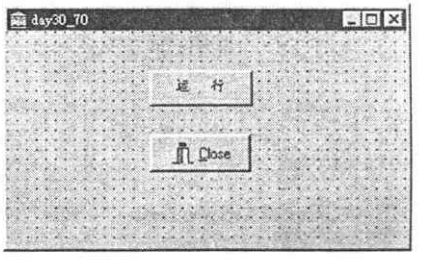

图 1 主窗体

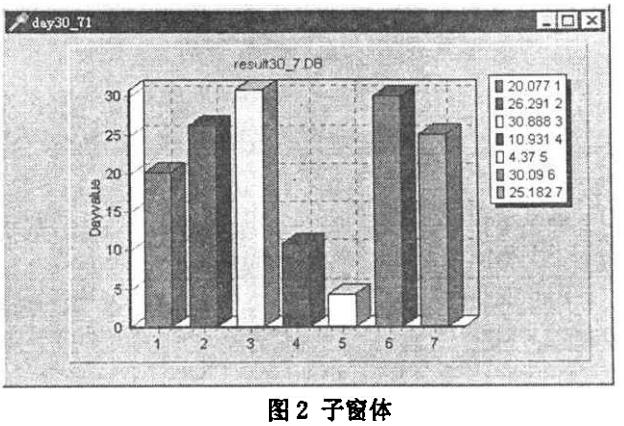

图 2 子窗体

“运行”按钮的单击事件处理程序为：

day30_7.show;

用于显示子窗体。

子窗体的 Create 事件处理程序为：

```
procedure Tday30_71.FormCreate(Sender: TObject);
type
        MyWeekArrayType=array[1..7] of Real;
Var
   I:Integer;
   Week: MyWeekArrayType;
   FiDay:TextFile;
Begin
   ⋮                          {从数据库中选取数据，送到文本文件 abc0126.txt 中}
 winexec('w107.exe',SW_HIDE);                    {执行预报程序}
 Sleep(2000);                                    {等待预报程序执行完毕}
 AssignFile(FiDay,'e:\frt\z4.txt');              {总的预报值文件}
 Reset(FiDay);
 For I:=1 to 7   do
   Read(FiDay,Week[I]);                          {送到数组 week 的第 I 个元素中}
 CloseFile(FiDay);
 Table1.Close;
 Table1.Open;
 Table1.First;
 For I:=1 to 7   do
 begin
  Table1.Edit;
  {将第 I 个预报值送到 Table1 的第 I 个元素中}
  Table1.FieldByName('Dayvalue').AsFloat:= Week[I];
  IF I<7 THEN Table1.FindNext;
 End;
 Table1.Post;                                                {发送到 Table1 中}
End;
```

6　预报结果分析

（1） 源数据文件的内容为：

22.330；19.488；15.865；19.464 22.652；7.950；3.114；21.658；19.176；15.960；21.540；25.536；9.350；3.644；25.452；21.480；16.692；20.964；25.172；9.400；3.990；28.336；24.840；19.209；24.216；27.860；9.785；3.894；27.972；24.9600000。

（2） “星期变动”的数据处理过程详见 3.1.1 节。

（3） “长期趋势” 的回归分析结果：y=4.76+10.29t

预报值为：20.3678100；20.5393300；20.7108600；20.8823900；21.0539200；21.2254500；21.3969800

残差值为：2.118378E-001；5.195661E-001；9.583440E-001；1.161209E-001；-3.385177E-002；.....1.373571E-002；-3.970872E-001；-9.667788E-001；-6.914010E-001；1.849275E-001。

（4）“自动周期”的结果：周期个数=1，周期=10。

（5）“人工周期”的结果：

$$p(t) = -0.28 * 10^{-6} - 0.14\cos\left(\frac{\pi}{5}t\right) + 0.92\sin\left(\frac{\pi}{5}t\right)$$

预报值为：4.284551E-001；8.343102E-001；9.214867E-001；6.566862E-001；1.410537E-001；
-4.284551E-001；-8.343104E-001。

残差值为：-2.166186E-001；-3.147445E-001；3.685743E-002；-5.405649E-001；-1.749051E-001；.....4.421919E-001；4.372238E-001；-4.529166E-002；-3.471333E-002；3.259829E-001。

（6） “随机变化”的结果：

预报值为：1.134149E-001；3.945893E-002；1.372842E-002；4.776352E-003；1.661777E-003；5.781649E-004；2.011582E-004。

（7）“综合结果”产生的总预报值为：20.0765800；26.2911700；30.8883000；10.9313400；4.3703160；30.0904900；25.1824800。

列表如下：

表 5. 计算值

计算值	第 1 天	第 2 天	第 3 天	第 4 天	第 5 天	第 6 天	第 7 天
理论值 Y	20.85	26.69	31.78	11.11	4.44	30.66	25.88
预测值 $\hat{Y}$	20.08	26.29	30.89	10.93	4.37	30.09	25.18
绝对误差 $(Y-\hat{Y})$	0.77	0.40	0.89	-0.14	0.07	0.57	0.70
误差平方 $(Y-\hat{Y})^2$	0.5929	0.1600	0.7921	0.0196	0.0049	0.3249	0.4900
Y 值波动 $(Y-\overline{Y})$	-0.78	5.06	10.15	-10.52	-17.19	9.03	4.25
波动平方 $(Y-\overline{Y})^2$	0.6084	25.6036	103.0225	110.6704	295.4961	81.5409	18.0625

$\overline{Y}$=21.63

根据公式（4）和（5），计算出相关系数

$$\gamma=\sqrt{1-\frac{2.3844}{635.0044}}=\sqrt{1-0.00375}=\sqrt{0.9962}\cong 0.998,$$

接近于理想值 1.00,说明这套统计计算方法相当精确。当然，现场数据不会这么有规律，预测值也就不会这么准确,能达到 0.8 就很好了。

7　其他预报程序

运用 3.1 节的算法，可以类似实现“从过去五年诸季度预测未来四个季度”和“从过去五年诸月预测未来 12 个月”等程序。主要区别在于将“季节”规律变为“季度”规律和“月份”规律。

8　致谢

衷心感谢中国科学院计算数学所魏公毅研究员的大力支持和热情帮助。

参考文献：

[1] 中国科学院计算中心统计组. 概率统计计算[M]. 北京: 科学出版社, 1979.

[2] R.I.Levin 著, 杨美瞳等译. 管理统计[M]. 北京: 电子工业出版社, 1986.

[3] 魏公毅, 张建中. “用 SASD 统计包预报含季节变化数据” [J]. 数理统计与管理, 1989(2).

The principle,implementation and application of trend analysis

ZHAO Li-ping[1], ZHAO Yang[2]

（1. Institute of Mathematics, Academia Sinica, China

2. China stata bureau of quality and technical supervision information center, China）

Abstract: According to observable data document(time number sequence),"trend analysis" does statistic computation finds inherent laws, forecasts further trend. Time number sequence generally involves four changable compositions: Season change, long-term trend, period fluctuation and random change. Here we mainly talk about moving average proportion method of season change, other compositions call statistic program packages.

Key words: Trend analysis; Forecast; Calculation; Time number sequence; Statistic; Regression analysis; Interrelative analysis; Database; Fortran and Delphi