# Chapter 2 Data Analysis



*There are three data analysis models: CYCLES, IDENTIFY, and ANOVA. These models help you understand data patterns before choosing a forecasting model. CYCLES detects changes in the rate of growth in a time series and identifies turning points in business cycles. IDENTIFY classifies the type of trend in a time series and also detects seasonal patterns. ANOVA computes the proportion of variance in a time series due to trend, seasonality, and noise.*

## 2.1 Business cycles analysis (CYCLES)

The CYCLES model is based on work originally done by forecast analysts at Parker Hannifin Corporation, a manufacturer of metalworking machinery, during the recession of the early 1970s.[1] To illustrate the forecasting problems faced by Parker Hannifin, look at Figure 2-1 which shows new orders for the metalworking machinery industry in millions of dollars from January, 1972, to September, 1974. The economy peaked in November, 1973, and then gradually declined through the official trough month of the recession: February, 1975. New orders for metalworking machinery were essentially flat from March, 1973, through January, 1974. In February and March of 1974, orders increased but fell off again from April through August. Suppose you were forecasting in this industry during the first half of 1974. How do you interpret the behavior of new orders? Will growth pick up again? Should you use a trend in your forecasting model?

The graphs in Figures 2-2 through 2-4 help answer these questions. Figure 2-2 is a graph of index numbers that compare orders for a given month to the same month a year ago. The first point plotted is: [(Jan. '73 orders) / (Jan. '72 orders)] x 100. The second is: [(Feb. '73 orders) / (Feb. '72 orders)] x 100. Multiplying each ratio by 100 forms an index number that is easy to interpret. If the index for a given month is 100, orders for that month are unchanged from a year ago. If the index is greater than 100, orders have grown from a year ago. If the index is less than 100, orders have declined. Such index numbers are called "1/12 pressures." The number 1 means that the index numbers are based on monthly totals and the number 12 means that the totals are separated by 12 months.

The 1/12 pressures declined in an erratic pattern during 1973. All pressures were greater than 100, meaning that orders grew compared to the same month a year ago. But the key point is that growth was slowing because the trend in pressures was negative. In January, 1974, a turning point occurred when the 1/12 pressures bottomed out at near 100. That is, orders in January, 1974, were about the same as the year before. Starting in February, growth picked up again, reaching a pressure level of more than 175 by September.

By including more data in the totals used in the pressure calculations, you can get additional evidence to verify turning points. In general, the more data used in the totals, the more reliable the results. Figure 2-3 shows "3/12 pressures," consecutive three-month totals compared to the same totals a year ago. The first point plotted is:

[(Jan. to Mar. '73 orders) / (Jan. to Mar. '72 orders)] x 100.

The second point plotted is:

[(Feb. to Apr. '73 orders) / (Feb. to Apr. '72 orders)] x 100.

---

[1] *Spyros Makridakis, Steven C. Wheelwright, and Victor E. McGee, Forecasting: Methods and Applications, second edition, New York: Wiley, 1983.*

**Figure 2-1**

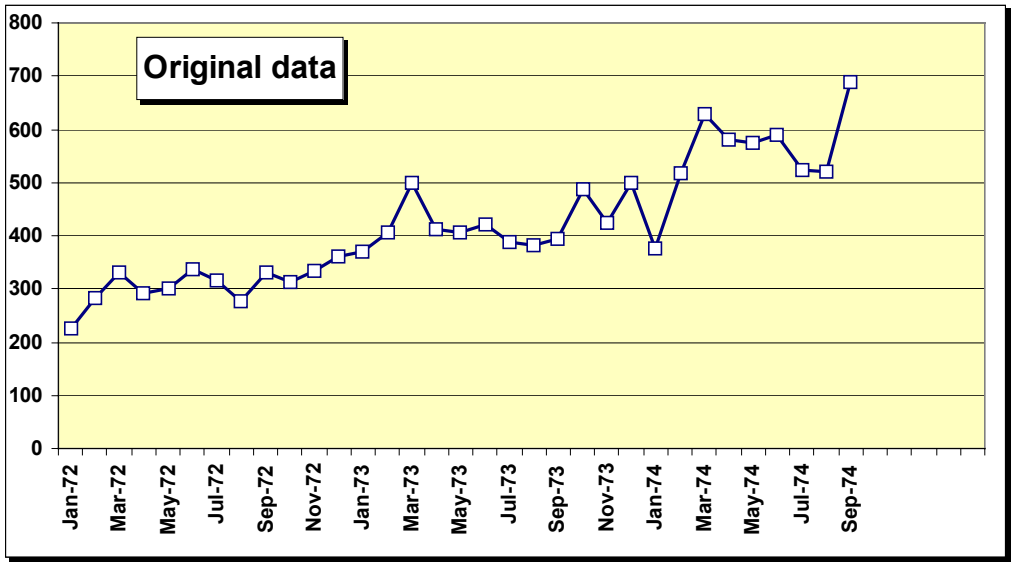



**Figure 2-2**

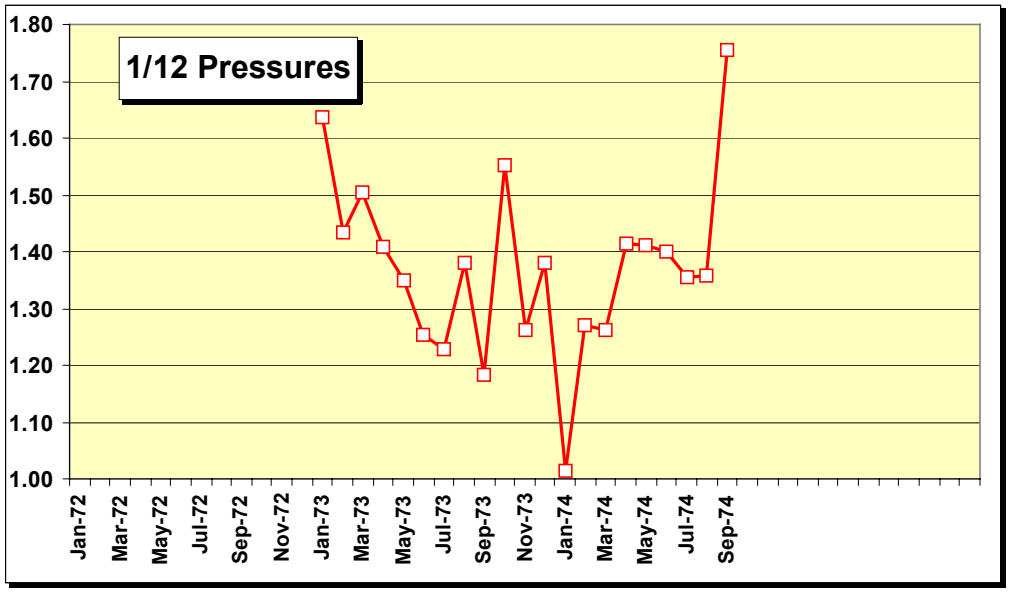


**Figure 2-3**

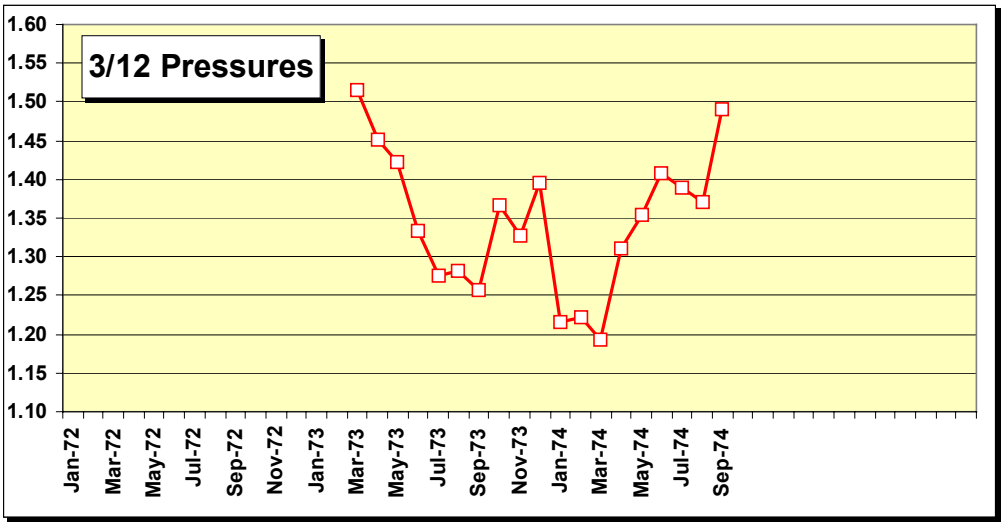
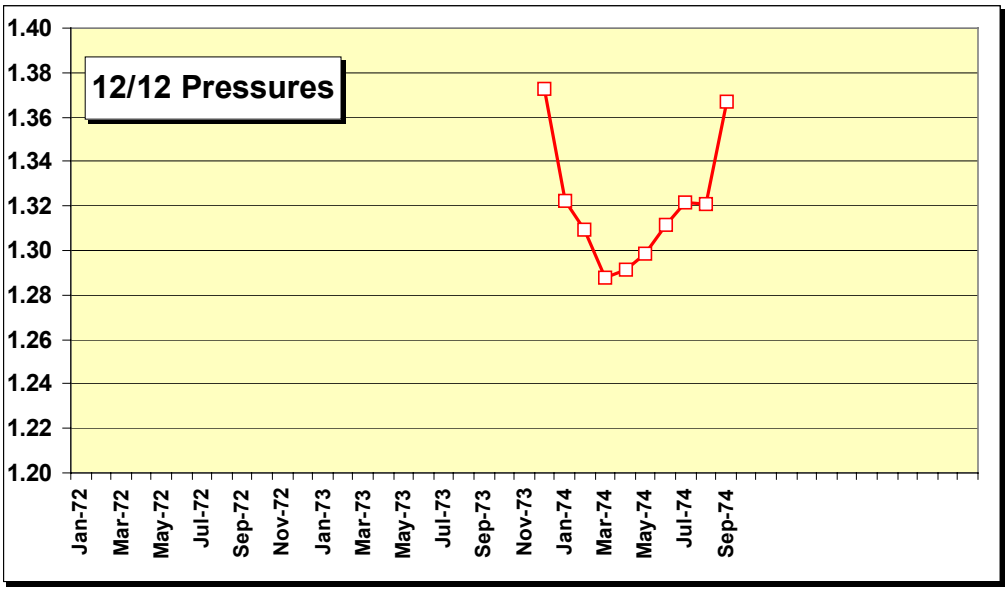



**Figure 2-4**



Note that the 3/12 pressures suggest a turning point in March rather than January, 1974.

Figure 2-4 shows "12/12 pressures," consecutive 12-month totals compared to the same totals a year ago. The first point plotted is:

[(Jan. to Dec. '73 orders) / (Jan. to Dec. '72 orders)] x 100.

The second point plotted is:

[(Feb. '73 to Jan. '74 orders) / (Feb. '72 to Jan. '73 orders)] x 100.

The 12/12 pressures are usually smoother and easier to read than the others. In Figure 2-4, they mark a dramatic turning point in March, 1974, the same month marked by the 3/12 pressures. The implication is that a new period of growth started around April, 1974.

Why did the 1/12 pressures disagree with the others about the timing of the turning point? This outcome simply reflects randomness in the data. So long as the timings of the turning points based on different pressures are in the same ball park, you can have some confidence that things will get better (or get worse if you sight a turning point at the top of a cycle). If the timings of the possible turning points are substantially different, you should be more cautious about the future. Pressures analysis is not foolproof, and randomness may prevent seeing any definite patterns.

Using graphs like Figures 2-2 to 2-4, Parker Hannifin detected the slowdown in growth and cut back on inventories early in the recession. After the turning point, the company began to build inventories again and gained market share from its competitors.

The pressures calculations are shown in Figure 2-5. Data are entered in Columns A and B. Note: data entry sections in all SFM worksheets are shaded in yellow. Column C computes the ratio of each month's data to the same month a year ago, starting at the 13th observation. Column D computes moving 3-month totals, while column E computes ratios of these totals to the same totals a year ago, starting at the 15th observation. Column F computes moving 12-month totals, while column G computes ratios of these totals to the same totals a year ago, starting at the 24th observation.

**Figure 2-5**

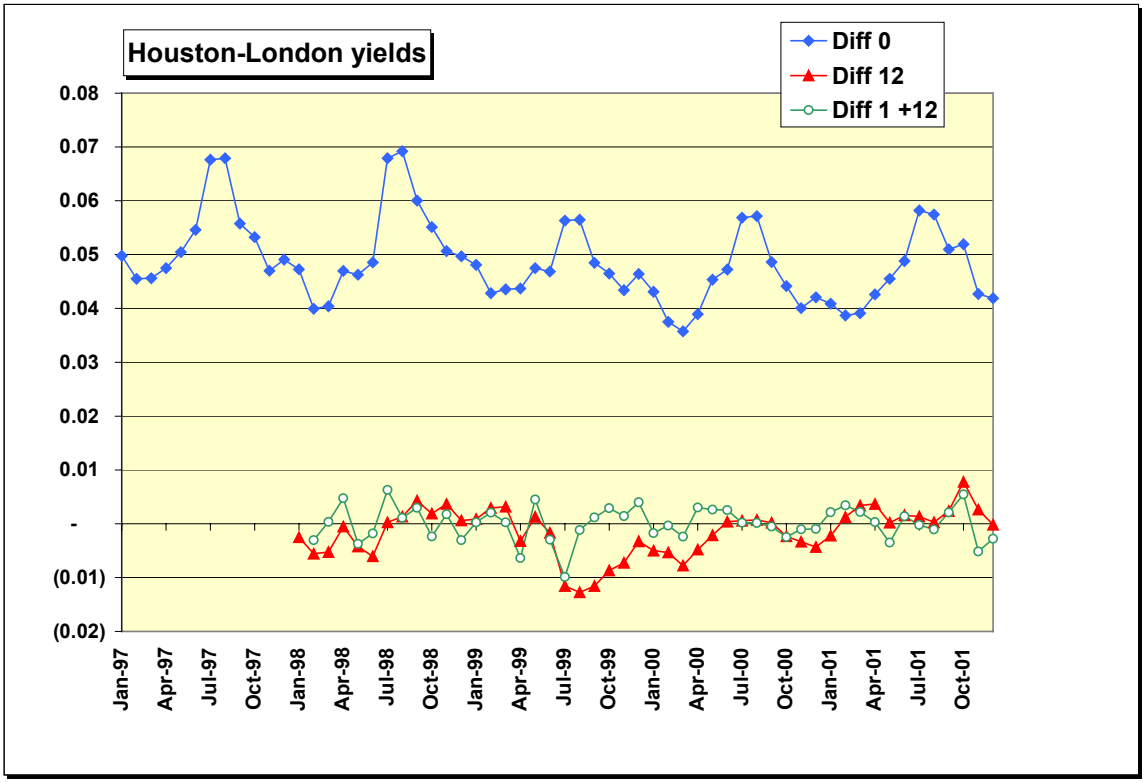
| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | **CYCLES** | | **New orders for metalworking machinery** | | | | |
| 2 | **Pressures analysis** | | | | | | |
| 3 | **Enter month and year.** | | | | | | |
| 4 | **Enter up to 120 observations under "Actual Data"** | | | | | | |
| 5 | | | | | | | |
| 6 | | | | | | | |
| 7 | **Month** | **Actual** | **1/12** | **3-MON.** | **3/12** | **12-MON.** | **12/12** |
| 8 | **& year** | **data** | **RATIO** | **TOTAL** | **RATIO** | **TOTAL** | **RATIO** |
| 9 | **Jan-72** | **226** | | | | | |
| 10 | **Feb-72** | **284** | | | | | |
| 11 | **Mar-72** | **331** | | 841 | | | |
| 12 | **Apr-72** | **292** | | 907 | | | |
| 13 | **May-72** | **301** | | 924 | | | |
| 14 | **Jun-72** | **336** | | 929 | | | |
| 15 | **Jul-72** | **315** | | 952 | | | |
| 16 | **Aug-72** | **277** | | 928 | | | |
| 17 | **Sep-72** | **332** | | 924 | | | |
| 18 | **Oct-72** | **314** | | 923 | | | |
| 19 | **Nov-72** | **335** | | 981 | | | |
| 20 | **Dec-72** | **362** | | 1011 | | 3705 | |
| 21 | **Jan-73** | **370** | 1.64 | 1067 | | 3849 | |
| 22 | **Feb-73** | **407** | 1.43 | 1139 | | 3972 | |
| 23 | **Mar-73** | **498** | 1.50 | 1275 | 1.52 | 4139 | |
| 24 | **Apr-73** | **411** | 1.41 | 1316 | 1.45 | 4258 | |
| 25 | **May-73** | **406** | 1.35 | 1315 | 1.42 | 4363 | |
| 26 | **Jun-73** | **421** | 1.25 | 1238 | 1.33 | 4448 | |
| 27 | **Jul-73** | **387** | 1.23 | 1214 | 1.28 | 4520 | |
| 28 | **Aug-73** | **382** | 1.38 | 1190 | 1.28 | 4625 | |
| 29 | **Sep-73** | **393** | 1.18 | 1162 | 1.26 | 4686 | |
| 30 | **Oct-73** | **487** | 1.55 | 1262 | 1.37 | 4859 | |
| 31 | **Nov-73** | **423** | 1.26 | 1303 | 1.33 | 4947 | |
| 32 | **Dec-73** | **500** | 1.38 | 1410 | 1.39 | 5085 | 1.37 |
| 33 | **Jan-74** | **375** | 1.01 | 1298 | 1.22 | 5090 | 1.32 |
| 34 | **Feb-74** | **517** | 1.27 | 1392 | 1.22 | 5200 | 1.31 |
| 35 | **Mar-74** | **628** | 1.26 | 1520 | 1.19 | 5330 | 1.29 |
| 36 | **Apr-74** | **581** | 1.41 | 1726 | 1.31 | 5500 | 1.29 |
| 37 | **May-74** | **573** | 1.41 | 1782 | 1.36 | 5667 | 1.30 |
| 38 | **Jun-74** | **589** | 1.40 | 1743 | 1.41 | 5835 | 1.31 |
| 39 | **Jul-74** | **524** | 1.35 | 1686 | 1.39 | 5972 | 1.32 |
| 40 | **Aug-74** | **519** | 1.36 | 1632 | 1.37 | 6109 | 1.32 |
| 41 | **Sep-74** | **690** | 1.76 | 1733 | 1.49 | 6406 | 1.37 |

## 2.2 Model identification (IDENTIFY)

The IDENTIFY model, based on research by Gardner and McKenzie,[2] detects seasonal patterns and also classifies the type of trend in a time series. An example of a series with both seasonality and trend is shown in Figure 2-6. The data are monthly yields (average revenue per passenger mile) for discount airline fares between Houston and London.

The first step in the analysis is to compute seasonal differences, that is a series of differences between data for the same month in each year. Seasonal differences, "Diff 12" in the figure, eliminate fluctuations caused by the seasonal cycle and thereby reduce the variance. If seasonality exists, the variance of the seasonal differences must be smaller than the variance of the original data. The next step is to compute a time series of differences between the seasonal differences, shown as "Diff 1 + 12" (for first-order differences between seasonal differences) below. If a trend of any kind exists, the variance of this new series will be smaller than the variance of the seasonal differences.

**Figure 2-6**




[2] *Everette S. Gardner, Jr. and E. McKenzie, "Model Identification in Exponential Smoothing," Journal of the Operational Research Society, Vol. 39, No. 9, pp. 863-867, 1988.*

At this point, you know that a trend exists but you don't know the type of trend. You can narrow the possibilities by computing differences between differences, that is differences between the "Diff 1 + 12" values (not shown in Figure 2-6). If these second-order differences have a smaller variance than the first-order differences, there is a strong trend in the data, either linear or exponential. If the second-order differences do not reduce the variance, you conclude that the trend is moderate, and the most likely model is a damped trend, as in this example. In a damped trend, the amount of growth each period gradually declines over time.

The IDENTIFY worksheet is shown in Figure 2-7 on the next page. The model automatically computes both non-seasonal and seasonal differences. In column C, the difference of order 0 (row 7) indicates that no differencing at all was done and variance is computed for the original data. In column D, the difference of order 1 indicates differences between successive data observations. In Column E, the difference of order 2 indicates differences between the order 1 differences. The variances are converted to indices in row 11, with the minimum variance among the six time series as a base of 100. The type of trend suggested when the minimum variance occurs is shown in row 6. For the airline yields, we conclude that the data are seasonal and the trend is moderate or damped.

When the suggested model includes seasonality, perform a seasonal adjustment using a worksheet from Chapter 3 before choosing a forecasting model. Output from seasonal adjustment can be copied directly to any forecasting model.

**Figure 2-7**

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **IDENTIFY** | | | | | | | | |
| 2 | **Variance of differences** | | | | | | | | |
| 3 | **Houston-London yields** | | | | | | | | |
| 4 | | | | | | | | | |
| 5 | | | **Non-seasonal** | | | **Monthly seasonal** | | | |
| 6 | | **Trend** | **Simple** | **Damped** | **Linear** | **Simple** | **Damped** | **Linear** | |
| 7 | | **Diff order** | **0** | **1** | **2** | **12** | **1+12** | **2+12** | |
| 8 | | **Nbr diffs** | 60 | 59 | 58 | 48 | 47 | 46 | |
| 9 | | **Mean diff** | 0.048696 | -0.000134 | 0.000060 | -0.001571 | 0.000050 | 0.000005 | |
| 10 | | **Variance** | 0.000059 | 0.000031 | 0.000049 | 0.000020 | 0.000010 | 0.000021 | |
| 11 | | **Var. index** | 592 | 311 | 488 | 197 | 100 | 205 | |
| 12 | | | | | | | Minimum | | |
| 13 | | | | | | | | | |
| 14 | | **Month** | **Actual** | | | | | | |
| 15 | | **& year** | **data** | **Diff 0** | **Diff 1** | **Diff 2** | **Diff 12** | **Diff 1 +12** | **Diff 2+12** |
| 16 | 1 | **Jan-97** | **0.04979** | 0.04979 | | | | | |
| 17 | 2 | **Feb-97** | **0.04552** | 0.04552 | (0.00427) | | | | |
| 18 | 3 | **Mar-97** | **0.04563** | 0.04563 | 0.00011 | 0.00438 | | | |
| 19 | 4 | **Apr-97** | **0.04748** | 0.04748 | 0.00185 | 0.00174 | | | |
| 20 | 5 | **May-97** | **0.05048** | 0.05048 | 0.00300 | 0.00115 | | | |
| 21 | 6 | **Jun-97** | **0.05460** | 0.05460 | 0.00412 | 0.00112 | | | |
| 22 | 7 | **Jul-97** | **0.06765** | 0.06765 | 0.01305 | 0.00893 | | | |
| 23 | 8 | **Aug-97** | **0.06791** | 0.06791 | 0.00026 | (0.01279) | | | |
| 24 | 9 | **Sep-97** | **0.05576** | 0.05576 | (0.01215) | (0.01241) | | | |
| 25 | 10 | **Oct-97** | **0.05325** | 0.05325 | (0.00251) | 0.00964 | | | |
| 26 | 11 | **Nov-97** | **0.04700** | 0.04700 | (0.00625) | (0.00374) | | | |
| 27 | 12 | **Dec-97** | **0.04908** | 0.04908 | 0.00208 | 0.00833 | | | |
| 28 | 13 | **Jan-98** | **0.04725** | 0.04725 | (0.00183) | (0.00391) | (0.00254) | | |
| 29 | 14 | **Feb-98** | **0.03994** | 0.03994 | (0.00731) | (0.00548) | (0.00558) | (0.00304) | |
| 30 | 15 | **Mar-98** | **0.04038** | 0.04038 | 0.00044 | 0.00775 | (0.00525) | 0.00033 | 0.00337 |
| 31 | 16 | **Apr-98** | **0.04695** | 0.04695 | 0.00657 | 0.00613 | (0.00053) | 0.00472 | 0.00439 |
| 32 | 17 | **May-98** | **0.04624** | 0.04624 | (0.00071) | (0.00728) | (0.00424) | (0.00371) | (0.00843) |
| 33 | 18 | **Jun-98** | **0.04856** | 0.04856 | 0.00232 | 0.00303 | (0.00604) | (0.00180) | 0.00191 |
| 34 | 19 | **Jul-98** | **0.06789** | 0.06789 | 0.01933 | 0.01701 | 0.00024 | 0.00628 | 0.00808 |
| 35 | 20 | **Aug-98** | **0.06923** | 0.06923 | 0.00134 | (0.01799) | 0.00132 | 0.00108 | (0.00520) |
| 36 | 21 | **Sep-98** | **0.06004** | 0.06004 | (0.00919) | (0.01053) | 0.00428 | 0.00296 | 0.00188 |
| 37 | 22 | **Oct-98** | **0.05514** | 0.05514 | (0.00490) | 0.00429 | 0.00189 | (0.00239) | (0.00535) |
| 38 | 23 | **Nov-98** | **0.05066** | 0.05066 | (0.00448) | 0.00042 | 0.00366 | 0.00177 | 0.00416 |
| 39 | 24 | **Dec-98** | **0.04969** | 0.04969 | (0.00097) | 0.00351 | 0.00061 | (0.00305) | (0.00482) |

## 2.3 Analysis of variance (ANOVA)

Apart from the business cycle, there are three sources of variance in a time series: trend, seasonality, and noise or unpredictable randomness. The ANOVA (analysis of variance) model, developed by AT&T Bell Laboratories,[3] identifies the proportion of total variation due to each source to assist in choosing a forecasting model. Figure 2-8 presents an ANOVA model for the time series of yields discussed in the previous section. The pie chart shows that about 75% of variance is due to seasonality, with 17% due to trend. Only 8% of the variance is due to noise, suggesting that the time series should be relatively easy to forecast.

The computations proceed as follows. The table starting in column E simply repeats the data, one year per column. The row means in column P refer to months. For example, cell P5 is the mean of all January values, cell P6 is the mean of all February values, and so on. If a seasonal pattern exists, there should be some extreme values in the monthly means. The peak month of the seasonal cycle should be much larger than the other months, and the trough or low point of the seasonal cycle should be much smaller. In the airline yields, there is a definite peak in August and a trough in February and March, typical of airline yields to Europe. The column means in row 17 are annual means. If a trend exists, the annual means should grow or decline consistently. In this case, the decline was consistent from 1997-2000 (years 1-4) and appeared to stabilize in 2001.

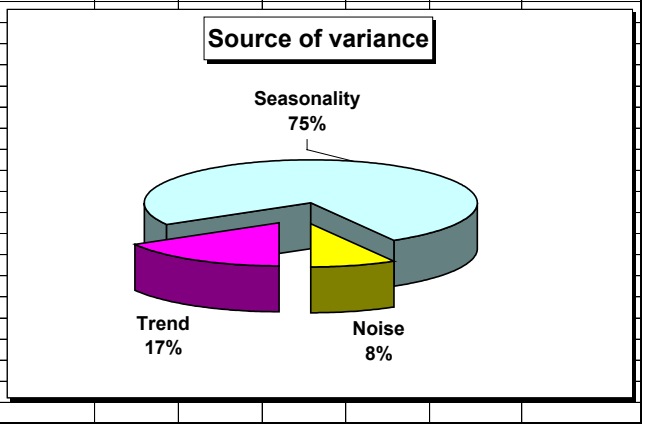
After computing the means, we measure the deviations (differences) between the monthly and annual average values and the "grand mean," the average of all data shown in cell C7. If a seasonal pattern is present, the deviations between the monthly averages and the grand mean should be relatively large. These monthly deviations are computed and squared in column Q and summed in cell C10. By the same reasoning, if a trend exists, the deviations between the yearly averages and the grand mean should be relatively large. Thus yearly deviations are computed and squared in row 17 and summed in cell C9.

The squared deviation values are hard to interpret without some standard for comparison. A logical standard is the total variation in the data, computed in cell C8. Total variation is the sum of the squared deviations between each individual data observation and the grand mean. The individual deviations are computed in column D.

[3] *Hans Levenbach and James P. Cleary, The Beginning Forecaster: The Forecasting Process Through Data Analysis, Belmont, California: Wadsworth, 1981.*

## Figure 2-8

| | A | B | C | D | E | F | G | H | I | J | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **ANOVA** | | | | | | | | | | | |
| 2 | **Analysis of variance (Trend-seasonal-noise decomposition)** | | | | | | | | | | | |
| 3 | **Name** | **Houston-London yields** | | | | | | | | | | |
| 4 | | | | | **Mon\year** | **1** | **2** | **3** | **4** | **5** | **Row mean** | **Row dev^2** |
| 5 | **Nbr. Months** | | 60 | | **Jan** | 0.0498 | 0.0473 | 0.0481 | 0.0431 | 0.0409 | 0.0458 | 0.0000 |
| 6 | **Nbr. Years** | | 5 | | **Feb** | 0.0455 | 0.0399 | 0.0429 | 0.0375 | 0.0387 | 0.0409 | 0.0001 |
| 7 | **Grand mean** | | 0.04870 | | **Mar** | 0.0456 | 0.0404 | 0.0435 | 0.0357 | 0.0391 | 0.0409 | 0.0001 |
| 8 | **Total variance** | | 0.00351 | | **Apr** | 0.0475 | 0.0470 | 0.0437 | 0.0389 | 0.0426 | 0.0439 | 0.0000 |
| 9 | **Trend variance** | | 0.00005 | | **May** | 0.0505 | 0.0462 | 0.0475 | 0.0453 | 0.0455 | 0.0470 | 0.0000 |
| 10 | **Seasonal variance** | | 0.00053 | | **Jun** | 0.0546 | 0.0486 | 0.0469 | 0.0472 | 0.0488 | 0.0492 | 0.0000 |
| 11 | **Source of variance:** | | | | **Jul** | 0.0677 | 0.0679 | 0.0563 | 0.0569 | 0.0582 | 0.0614 | 0.0002 |
| 12 | | **Trend** | 17% | | **Aug** | 0.0679 | 0.0692 | 0.0565 | 0.0572 | 0.0574 | 0.0616 | 0.0002 |
| 13 | | **Seasonality** | 75% | | **Sep** | 0.0558 | 0.0600 | 0.0485 | 0.0486 | 0.0510 | 0.0528 | 0.0000 |
| 14 | | **Noise** | 8% | | **Oct** | 0.0533 | 0.0551 | 0.0465 | 0.0441 | 0.0519 | 0.0502 | 0.0000 |
| 15 | | **Total** | 100% | | **Nov** | 0.0470 | 0.0507 | 0.0434 | 0.0401 | 0.0427 | 0.0448 | 0.0000 |
| 16 | | | | | **Dec** | 0.0491 | 0.0497 | 0.0464 | 0.0421 | 0.0419 | 0.0458 | 0.0000 |
| 17 | | | | | **Col mean** | 0.0528 | 0.0518 | 0.0475 | 0.0447 | 0.0466 | | |
| 18 | ***Enter complete years of data only:*** | | | | **Col dev^2** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | | |
| 19 | | **Month** | **Actual** | **Deviation** | | | | | | | | |
| 20 | | **& year** | **data** | **squared** | | | | | | | | |
| 21 | 1 | **Jan-97** | **0.04979** | 0.000001 | | | | | | | | |
| 22 | 2 | **Feb-97** | **0.04552** | 0.000010 | | | | | | | | |
| 23 | 3 | **Mar-97** | **0.04563** | 0.000009 | | | | | | | | |
| 24 | 4 | **Apr-97** | **0.04748** | 0.000001 | | | | | | | | |
| 25 | 5 | **May-97** | **0.05048** | 0.000003 | | | | | | | | |
| 26 | 6 | **Jun-97** | **0.05460** | 0.000035 | | | | | | | | |
| 27 | 7 | **Jul-97** | **0.06765** | 0.000359 | | | | | | | | |
| 28 | 8 | **Aug-97** | **0.06791** | 0.000369 | | | | | | | | |
| 29 | 9 | **Sep-97** | **0.05576** | 0.000050 | | | | | | | | |
| 30 | 10 | **Oct-97** | **0.05325** | 0.000021 | | | | | | | | |
| 31 | 11 | **Nov-97** | **0.04700** | 0.000003 | | | | | | | | |
| 32 | 12 | **Dec-97** | **0.04908** | 0.000000 | | | | | | | | |
| 33 | 13 | **Jan-98** | **0.04725** | 0.000002 | | | | | | | | |
| 34 | 14 | **Feb-98** | **0.03994** | 0.000077 | | | | | | | | |
| 35 | 15 | **Mar-98** | **0.04038** | 0.000069 | | | | | | | | |
| 36 | 16 | **Apr-98** | **0.04695** | 0.000003 | | | | | | | | |
| 37 | 17 | **May-98** | **0.04624** | 0.000006 | | | | | | | | |
| 38 | 18 | **Jun-98** | **0.04856** | 0.000000 | | | | | | | | |



Now we are ready to compute the proportion of total variance caused by trend and seasonality. The trend proportion is 17% of total variance, computed as follows:

$$\textbf{Trend proportion} = \frac{\textbf{12 x sum of squared yearly deviations}}{\textbf{Sum of squared individual deviations}} \qquad \textbf{(2-1)}$$

The numerator is multiplied by 12 since each yearly deviation is the average of 12 months of data.

The seasonal proportion is 75% of the total, computed as follows:

$$\textbf{Seasonal proportion} = \frac{\textbf{Nbr. of years x sum of squared monthly deviations}}{\textbf{Sum of squared individual deviations}} \qquad \textbf{(2-2)}$$

Finally, the proportion of noise (8%) must be what's left over after trend and seasonality are accounted for:

$$\textbf{Noise proportion} = \textbf{1 - (trend proportion + seasonal proportion)} \qquad \textbf{(2-3)}$$

For the airline yields, IDENTIFY and ANOVA give consistent recommendations. The data are definitely seasonal and we should be cautious in extrapolating any trend or further decline in yields.