NN3 Forecasting Competition Methodology

Edward Thomas Lewicke

Abstract—**Forecasts were produced using equation-based models. The equation consists of a trend plus a series of sinusoidal error terms. Substitution of future time values into the equation produces an estimate of the dependent variable.**

I. INTRODUCTION

PARABOLIC SYSTEMS PARACASTER J forecasting and modeling software was used to produce the forecasts submitted as entries in the NN3 forecasting competition. This software fits a trend line, straight line, curved line or flat line to the data. The differences between the data and the trend line can usually be described as a sinusoidal line. ParaCaster J calculates the frequency and starting point of the sine curve. The sine term is added to the previously determined trend and sine terms. The coefficients of the all the equation terms are found using conventional least squares regression

II. TREND LINE

The selection of the trend line shape is not as easy as looking at the approximate shape of the graph of the data. If the data must stay within an upper and lower limit as in the case of an unemployment rate, a flat trend line can be selected. Data with an exponential growth rate can be described using a curved trend line. In most other cases, a straight trend line can be used.

Things become a little more complicated because the cycle selection part of the program can identify a long cycle that will describe the curvature in many data sets. A flat or straight trend can be used to describe data sets with a curved graph. At present, the best way to select a trend shape is to review the model, the model's error and the model's shape.

III. CYCLICAL COMPONENTS

The difference between the trend and the data can be described as a series of sinusoidal error terms. Each cyclical term requires the estimation of three constants, the coefficient, the point at which the cycle starts and the length of the cycle. The cycle start point is the point where the cycle crosses the x-axis in a positive direction. This point corresponds to the 0 angle point of the sine curve.

The first approximation of the cycle can be determined by estimating the length of most significant lobe of the error line above or below the x-axis. If the lobe is above the x-axis the start point is the beginning of the lobe. If the lobe is below the x-axis, the 0 angle of the cycle is the end of the lobe. The cycle length is twice the length of the lobe. The estimated cycle length and cycle start point are refined using iterative trials to find the combination of start point and cycle length that best fits the error. This new cyclical term describing the incremental error is added to the previously determined equation. The coefficient is determined using least squares regression.

This process is repeated until the error is reduced to the point that additional error terms do not make an appreciable increase in the accuracy of the model. At this point the model may contain a number of cycles that are not statistically significant.

IV. MODEL REFINEMENT

Short cycles that add 'roughness' to the graph of the model are removed one at a time, as are cycles with a 't' statistic of less than 1.96. A maximum number of terms in the equation of the model can be specified and the less significant model terms are removed one at a time until the specified number of terms is reached. When terms are removed, the term with the lowest 't' statistic is removed, the regression coefficients are recalculated and the 't' statistic is recalculated. This process is repeated until the required number of terms is reached or the lowest 't' statistic is above the specified minimum.

V. FINAL MODEL

The equation of the final model will have the form:

$$y = a + bx + cd^x + e_1 \sin\left(\frac{x - s_1}{w_1} 2\pi\right) + \ldots e_n \sin\left(\frac{x - s_n}{w_n} 2\pi\right)$$

a, b, c, and e are regression coefficients
d is 1 plus the growth rate
s is the cycle start point
w is the cycle length
x is a point in time (the independent variable)
y is the dependent variable

The curved trend contains the three first terms. The straight trend contains the first two terms but not the growth rate term. The flat trend contains only the first term and the sine terms. A sample equation, in this case the equation for data set 56, is:

$$y = 138359 + 4.577x - 166528 \times 1.00161^{x} + 475.431 \sin\left(\frac{x - 29999.57}{369.839} 2\pi\right) + 104.084 \sin\left(\frac{x - 30156.63}{788.018} 2\pi\right) + 120.664 \sin\left(\frac{x - 30115.05}{181.526} 2\pi\right)$$

Points in time, the independent variable and cycle start point, are in days after December 31, 1899 and the cycle length is in days.

VI. Conclusion

The method of modeling data presented here is applicable to time series data as well as physical dimensions such as topography. The use of an equation allows extrapolation of the data to determine future values for forecasts. Using an equation eliminates any problems due to outliers, as the equation is not unduly influenced by a single observation. The equation can be used to interpolate between known observations to estimate the values of missing data.

The ability to limit the minimum cycle length permits the model to be smoothed. The trend line shape options allow the model to accurately describe various graphical shapes of data.