1

Maintenance Forecasting and Capacity Planning

Hesham K. Alfares and Salih O. Duffuaa

1.1 Introduction

Carrying out an effective maintenance operation requires efficient planning of maintenance activities and resources. Since planning is performed in order to prepare for future maintenance tasks, it must be based on good estimates of the future maintenance workload. The maintenance workload consists of two major components: (1) scheduled and planned preventive maintenance, including planned overhauls and shutdowns, and (2) emergency or breakdown failure maintenance. The first component is the deterministic part of the maintenance workload. The second component is the stochastic part that depends on the probabilistic failure pattern, and it is the main cause of uncertainty in maintenance forecasting and capacity planning.

Estimates of the future maintenance workload are obtained by forecasting, which can be simply defined as predicting the future. Clearly, good forecasts of the maintenance workload are needed in order to plan well for maintenance resources. In terms of the time horizon, forecasts are typically classified into three main types: (1) short-term: ranging from days to weeks, (2) intermediate-term: ranging from weeks to months, and (3) long-term: ranging from months to years. Long-term forecasts are usually associated with long-range maintenance capacity planning.

The main objective in capacity planning is to assign fixed maintenance capacity (resources) to meet fluctuating maintenance workload, in order to achieve the best utilization of limited resources. Maintenance capacity planning determines the appropriate level and workload assignment of different maintenance resources in each planning period. Examples of maintenance resources include spare parts, manpower of different skills (craftsmen), tools, instruments, time, and money. For each planning period, capacity planning decisions include the number of employees, the backlog level, overtime workload, and subcontract workload. Proper allocation of the various maintenance resources to meet a probabilistic fluctuating workload is a complex and important practical problem. In order to solve this problem optimally, we have to simultaneously balance the cost and availability of all applicable maintenance resources. A variety of capacity-planning techniques are used for handling this complex problem.

This chapter presents the main concepts and tools of maintenance workload forecasting and capacity planning. Section 1.2 provides a brief introduction to forecasting. Section 1.3 describes qualitative or subjective forecasting techniques. Section 1.4 presents quantitative or objective forecasting models. Section 1.5 covers model evaluation and error analysis. Section 1.6 presents different approaches to maintenance workload forecasting. Section 1.7 outlines the problem of capacity planning in maintenance. Sections 1.8 and 1.9 respectively describe deterministic and stochastic techniques for capacity planning. Finally, section 1.10 gives a brief summary of this chapter.

1.2 Forecasting Basics

Forecasting techniques are generally classified into two main types: qualitative and quantitative. Qualitative (subjective) techniques are naturally used in the absence of historical data (e.g. for new machines or products), and they are based on personal or expert judgment. On the other hand, quantitative (objective) techniques are used with existing numerical data (e.g. for old machines and products), and they are based on mathematical and statistical methods.

Qualitative forecasting techniques include historical analogy, sales force composites, customer surveys, executive opinions, and the Delphi method. Quantitative techniques are classified into two types: (1) growth or time-series models that use only past values of the variable being predicted, and (2) causal or predictor-variable models that use data of other (predictor) variables.

Nahmias (2005) makes the following observations about forecasts: (1) forecasts are usually not exact, (2) a forecast range is better than a single number, (3) aggregate forecasts are more accurate than single-item forecasts, (4) accuracy of forecasts is higher with shorter time horizons, and (5) forecasts should not ignore known and relevant information. To choose a forecasting technique, the main criteria include: (1) objective of the forecast, (2) time horizon for the forecast, and (3) data availability for the given technique. In order to develop a quantitative forecasting model, the steps below should be followed.

1. Define the variable to be predicted, and identify possible cause-effect relationships and associated predictor variables.
2. Collect and validate available data for errors and outliers.
3. Plot the data over time, and look for major patterns including stationarity, trends, and seasonality.
4. Propose several forecasting models, and determine the parameters and forecasts of each model.
5. Use error analysis to test and validate the models and select the best one.
6. Refine the selected model and try to improve its performance.

Quantitative forecasting techniques are classified into time-series and causal models. They aim to identify, from past values, the main patterns that will continue in the future. The most frequent patterns, illustrated in Figure 1, include the following:

1. Stationarity: level or constant demand.
2. Growth or trend: long-term pattern of growth or decline.
3. Seasonality: cyclic pattern repeating itself at fixed intervals.
4. Economic cycles: similar to seasonality, but length and magnitude of cycle may vary.

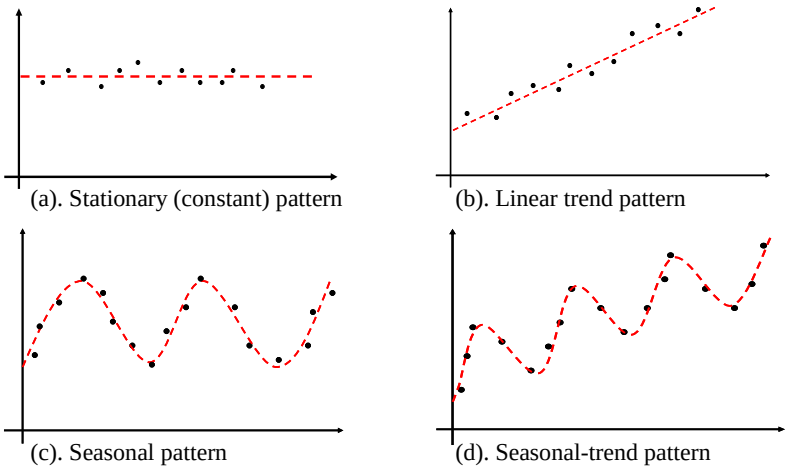

Figure 1.1. Major patterns identified in quantitative forecasting techniques

1.3 Qualitative Forecasting Techniques

Qualitative or subjective forecasting is used in any case where quantitative forecasting techniques are not applicable. Such cases include non-existence, non-availability, non-reliability, and confidentiality of data. Qualitative forecasting is also used when the forecasting horizon is very long, e.g. 20 years or more, such that quantitative forecasting techniques become unreliable. In the absence of numerical data, good qualitative forecasts can still be obtained by systematically soliciting the best subjective estimates of the experts in the given field. For maintenance requirements of new plants or equipment, qualitative forecasting techniques include benchmarking with similar plants and referring to the maintenance instructions provided by the equipment manufacturers. Nahmias (2005) identifies four types of subjective forecasting techniques:

1. Sales force composite: each member of the sales force submits a forecast for items he or she sells, and then the management consolidates.
2. Customer surveys: collect direct customer input; must be carefully designed to find future trends and shifting preferences.
3. Executive opinions: forecasts are provided by management team members from marketing, finance, and production.

4. The Delphi method: a group of experts respond individually to a questionnaire, providing forecasts and justifications. Results are combined, summarized, and returned to experts to revise. The process is repeated until consensus is reached

The most sophisticated technique for qualitative forecasting is the Delphi method, which will be presented in the next section.

1.3.1 The Delphi Method

The Delphi method is a systematic interactive qualitative forecasting technique for obtaining forecasts from a panel of independent experts. The experts are carefully selected and usually consulted using structured questionnaires that are conducted in two or more rounds. At the end of each round, an anonymous summary of the experts' latest forecasts as well as the reasons they provided for their judgments is provided to the experts by a facilitator. The participants are encouraged to revise their earlier answers in light of the replies of other members of the group.

It is believed that during this process the variations in the answers will gradually diminish and that the group will converge towards a consensus. The process is terminated after a pre-defined stopping criterion (e.g. number of rounds, achievement of consensus, and stability of results). According to Rowe and Wright (2001), the mean or median scores of the last round determine the final estimates. The Delphi method was developed in the 1950s by the RAND Corporation in Santa Monica, California. The following steps may be used to implement a Delphi forecasting process:

1. Form the Delphi team to conduct the project.
2. Select the panel of experts.
3. Develop the Delphi questionnaire for the first round.
4. Test and validate the questionnaire for proper design and wording.
5. Send the first survey to the panel.
6. Analyze the first round responses.
7. Prepare the next round questionnaire and possible consensus tests.
8. Send the next round questionnaire to the experts.
9. Analyze responses to the questionnaire (steps 7 through 9 are repeated until the stopping criterion is satisfied).
10. Prepare the report with results, analysis, and recommendations.

The Delphi method is based on the following assumptions: (1) well-informed individuals using their insight and experience can predict the future better than theoretical models, (2) the problem under consideration is very complex, (3) there is no history of sustained communication among participating experts, and (4) exchange of ideas is impossible or impractical.

The strengths of the Delphi method include: (1) it achieves rapid consensus, (2) participants can be anywhere in the world, (3) it can cover a wide range of expertise, and (4) it avoids groupthink. The limitations of the method include: (1) it

neglects cross impact, (2) it does not cope well with paradigm shifts, and (3) its success depends on the quality of the experts.

The Delphi method can be applied in maintenance in several areas, including determining time standards and preventive maintenance time intervals, as well as estimating the remaining useful life of equipment.

1.4 Quantitative Forecasting Techniques

Quantitative or objective forecasting techniques are presented in this section. These models are based on the availability of historical data, and are usually classified into time-series and causal models. A time series is a set of values of the variable being predicted at discrete points in time. Time-series models are considered naïve because they require only past values of the variable being predicted. Causal models assume that other predictor variables exist that can provide a functional relationship to predict the variable being forecasted. For example, the age of a given machine equipment may help in predicting the frequency of failures. The models presented here include methods for stationary, linear, and seasonal data.

1.4.1 Simple Moving Averages

This type of forecasts is used for stationary time series, which is composed of a constant term plus random fluctuation. An example of this could be the load exerted on an electronic component. Mathematically, this can be represented as

$$D_t = \mu + \varepsilon_t \tag{1.1}$$

where

D_t = demand at time period t,
μ = a constant mean of the series,
ε_t = error at time t; a random variable with mean 0 and variance σ^2.

Obviously, our forecast of future demand should be our best estimate of the parameterμ. Let us assume that all N previous observations are assumed to be equally important, i.e. equally weighted. If we use the least-squares method, then we look for the value of μ that minimizes the sum of squared errors (*SSE*).

$$SSE = \sum_{t=1}^{N} (D_t - \mu)^2 \tag{1.2}$$

When we differentiate equation (1.2) with respect to μ and equate the result to zero, we obtain the optimum value of μ as our forecast given by:

$$F_t = \frac{\sum_{i=1}^{N} D_{t-i}}{N} \tag{1.3}$$

where

F_t = forecast for time periods $t, \ldots, \infty$

Since F_t is the average of the last N actual observations (periods $t-1, \ldots, t-N$), it is called a simple N-period moving average, or a moving average of order N.

If simple moving average equation (1.3) is used with a perfectly linear data of the form $D_t = a + bt$, then there will be an error that depends on the slope b and the number of points included in the moving average N. Specifically, the forecast will underestimate or lag behind the actual demand by:

$$\varepsilon_t = D_t - F_t = \frac{(N+1)b}{2} \tag{1.4}$$

Example 1: The breakdown maintenance load in man-hours for the last 5 months is given as

t	1	2	3	4	5
D_t	800	600	900	700	600

Forecast the maintenance load for period 6 using a 3-month moving average.

The forecasted load for month 6 and all future months is:

$$F_6 = \frac{900+700+600}{3} = 733.33$$

1.4.2 Weighted Moving Average

In simple moving average, an equal weight is given to all n data points. Since individual weight is equal to $1/n$, then sum of the weights is $n(1/n) = 1$. Naturally, one would expect that the more recent data points have more forecasting value than older data points. Therefore, the simple moving average method is sometimes modified by including weights that decrease with the age of the data. The forecasting model becomes:

$$F_t = \sum_{i=1}^{N} w_i D_{t-i} \tag{1.5}$$

where

w_i = weight of the ith observation in the N-period moving average

$$\sum_{i=1}^{N} w_i = 1 \tag{1.6}$$

The values of w_t must be non-decreasing with respect to t. These values can be empirically determined based on error analysis, or subjectively estimated based on experience, hence combining qualitative and quantitative forecasting approaches.

Example 2: Using the maintenance load values of example 1, assume that each observation should weigh twice as much as the previous observation. Forecast the load for month 6 using a 3-period weighted moving average:

$w_2 = 2w_1$
$w_3 = 2w_2$
$w_1 + w_2 + w_3 = 1$

Solving the 3×3 system gives:

$w_1 = 1/7$
$w_2 = 2/7$
$w_3 = 4/7$

The forecasted load for month 6 and all future months is:

$$F_6 = \frac{900 + 2(700) + 4(600)}{7} = 671.43$$

1.4.3 Regression Analysis

Regression analysis is used to develop a functional relationship between the independent variable being forecasted and one or more independent predictor variables. In time-series regression models, the only independent variable is time. In causal regression models, other independent predictor variables are present. For example if the cost of maintenance for the current period $m(t)$ is a linear function of the number of operational hours in the same period $h(t)$, then the model is given by

$$m(t) = a + bh(t) + \varepsilon_t \tag{1.7}$$

Equation (1.7) represents a straight-line regression relationship with a single independent predictor variable, namely $h(t)$. The parameters a and b are respectively called the intercept and the slope of this line. Regression analysis is the process of estimating these parameters using the least-squares method. This method finds the best values of a and b that minimize the sum of the squared vertical distances (errors) from the line.

The general straight-line equation showing a linear trend of maintenance work demand D_i over time is

$$D_i = a + bt_i + \varepsilon_i \tag{1.8}$$

where

D_i = demand at time period t_i,
ε_i = error at time period t_i.

Let us assume that n historical data points are available: (t_1, D_1), (t_2, D_2), ..., (t_n, D_n). The least-squares method estimates a and b by minimizing the following sum of squared errors:

$$SSE = \sum_{i=1}^{n} (D_i - a - bt_i)^2 \tag{1.9}$$

Taking partial derivatives with respect to a and b and setting them equal to zero produces a 2×2 system of linear equations, whose solution is given by:

$$b = \frac{n\sum_{i=1}^{n} t_i D_i - \sum_{i=1}^{n} t_i \sum_{i=1}^{n} D_i}{n\sum_{i=1}^{n} {t_i}^2 - \left(\sum_{i=1}^{n} t_i\right)^2} \tag{1.10}$$

$$a = \frac{1}{n}\left(\sum_{i=1}^{n} D_i - b\sum_{i=1}^{n} t_i\right) = \overline{D} - b\bar{t} \tag{1.11}$$

Quite often, the variable being forecasted is a function of several predictor variables. For example, maintenance cost might be a linear function of operating hours, $h(t)$, and the age of the plant, t, which can be expressed as:

$$m(t) = a + bh(t) + ct + \varepsilon_t$$

Least-squares regression methodology can easily accommodate multiple variables and also polynomial or nonlinear functional relationships.

Example 3: Demand for a given spare part is given below for the last four years. Use linear regression to determine the best-fit straight line and to forecast spare part demand in year 5.

Year *t*	1	2	3	4
Spare part demand *D*(*t*)	100	120	150	170

Intermediate calculations for the summations needed in equations (1.10) and (1.11) are shown in Table 1.1 below.

Table 1.1. Data and intermediate calculations for the linear regression example

					Sum
t	1	2	3	4	**10**
***D*(*t*)**	100	120	150	170	**540**
***tD*(*t*)**	100	240	450	680	**1470**
$\boldsymbol{t^2}$	1	4	9	16	**30**

Using equations (1.10) and (1.11), the slope and intercept of the line are estimated as follows:

$$b = \frac{4(1470) - 10(540)}{4(30) - 10^2} \qquad = 24$$

$$a = \frac{1}{4}\left[540 - 24(10)\right] \qquad = 75$$

The equation of the least-squares straight line is $D(t) = 75 + 24t$. Therefore, the forecasted spare part demand in year 5 is:

$$D(5) = 75 + 24(5) = 195 \text{ units}$$

1.4.4 Exponential Smoothing

1.4.4.1 Simple Exponential Smoothing (ES)

Simple exponential smoothing (ES) is similar to weighted moving average (WMA) in assigning higher weights to more recent data, but it differs in two important aspects. First, WMA is a weighted average of only the last N data points, while ES is a weighted average of *all* past data. Second, the weights in WMA are mostly arbitrary, while the weights in ES are well structured. In fact, the weights in ES decrease exponentially with the age of the data. On the other hand, exponential smoothing is very easy to use, and very easy to update by including new data as it becomes available. In addition, we must save the last N observations for WMA, but need to save only the last observation and the last forecast for ES. These characteristics have made exponential smoothing very popular. Basically, the current forecast is a weighted average of the last forecast and the last actual observation. Given the value of smoothing constant α ($0 \le \alpha \le 1$), which is the relative weight of the last observation, the forecast is obtained by:

$$F_t = \alpha D_{t-1} + (1 - \alpha)F_{t-1} \tag{1.12}$$

The greater the value of α, the more weight of the last observation, i.e., the quicker the reaction to changes in data. However, large values of α lead to highly variable, less stable, forecasts. For forecast stability, a value of α between 0.1 and 0.3 is usually recommended for smooth planning. The best value of α can be determined from experience or by trial and error (choosing the value with minimum error). It can be shown that the ES forecast is a weighted average of all past data, where the weights decrease exponentially with the age of the data as expressed by:

$$F_t = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i D_{t-i-1} \tag{1.13}$$

Using equation (1.12), the first forecast F_1 requires the non-existent values of D_0 and F_0. Therefore, an initial value of F_1 must be specified for starting the process.

Usually, F_1 is set equal to the actual demand in the first period D_1, or to the average of the first few observations.

If simple exponential smoothing (1.12) is used with linear data ($D_t = a + bt$), then the error will depend on the slope b and the smoothing constant α. As $t \to \infty$, the forecast will lag behind the actual demand by:

$$\lim_{t\to\infty} \{\varepsilon_t = D_t - F_t\} = \frac{b}{\alpha} \tag{1.14}$$

To make MA(N) and ES(α) consistent, we equate the two lags, ensuring the distribution of forecast errors will be the same, although individual forecasts will not be the same. Equating the exponential smoothing lag of equation (1.14) with the moving average lag of equation (1.4), we obtain the following value for α:

$$\alpha = \frac{2}{N+1} \tag{1.15}$$

Example 4: Given that $\alpha = 0.2$ and $F_1 = D_1$, apply simple exponential smoothing to the data of example 1 to forecast maintenance workload in month 6.

Using equation (1.12), the calculations are shown below. The forecast for month 6 is $F_6 = 736.32$ man-hours.

Table 1.2. Data and intermediate calculations for the simple exponential smoothing example

t	1	2	3	4	5	6
D_t	800	600	900	700	600	
F_t	800	0.2(800) + 0.8(800) = 800	0.2(600) + 0.8(800) = 760	0.2(900) + 0.8(760) = 788	0.2(700) + 0.8(788) = 770.4	0.2(600) + 0.8(770.4) = 736.32

1.4.4.2 Double Exponential Smoothing (Holt's Method)

The simple exponential smoothing equation (1.12) can be used to estimate the parameters for a constant (stationary) model. However, double or triple exponential smoothing approaches can be used to deal with linear, polynomial, and even seasonal forecasting models. Several double exponential smoothing techniques have been developed for forecasting with linear data. One of these is Holt's double exponential smoothing method, which is described below.

Holt's double exponential smoothing method requires two smoothing constants: α and β ($\beta \le \alpha$). Two smoothing equations are applied: one for a_t, the intercept at time t, and another for b_t, the slope at time t:

$$a_t = \alpha D_t + (1-\alpha)(a_{t-1} + b_{t-1}) \tag{1.16}$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \tag{1.17}$$

The initial values b_0 and a_0 are obtained as follows:

$$b_0 = \frac{D_n - D_1}{t_n - t_1} \tag{1.18}$$

$$a_0 = \frac{1}{n}\left(\sum_{i=1}^{n} D_i - b_0 \sum_{i=1}^{n} t_i\right) = \overline{D} - b_0 \bar{t} \tag{1.19}$$

At the end of period t, the forecast for period τ ($\tau > t$) is obtained as follows:

$$F_\tau = a_t + b_t(\tau - t) \tag{1.20}$$

Example 5: Given that $\alpha = \beta = 0.2$, apply Holt's double exponential smoothing method to the data of example 3 in order to forecast spare part demand in year 5.

First, initial conditions are calculated by (1.18) and (1.19)

$$b_0 = \frac{190 - 100}{4 - 1} = 30$$

$$a_0 = \frac{1}{4}[570 - 30(10)] = 67.5$$

Intermediate calculations are shown in the following table.

Table 1.3. Data and calculations for the double exponential smoothing example

t	0	1	2	3	4
D_t		100	120	160	190
a_t	67.5	0.2(100) + 0.8(67.5 + 30) = 98	0.2(120) + 0.8(98 + 30.1) = 126.48	157.005	187.544
b_t	30	0.2(98-67.5) + 0.8(30) = 30.1	0.2(126.48 – 98) + 0.8(30.1) = 29.776	29.926	30.049

The forecasting model at the end of year 4 is: F_t = 187.544 + 30.049(t – 4). Therefore, the forecasted spare part demand in year 5 is given by: F_5 = 187.544 + 30.049(5 – 4) = 217.593.

1.4.5 Seasonal Forecasting

Demand for many products and services follows a seasonal or cyclic pattern, which repeats itself every N periods. Although the term "seasonal" is usually associated with the four seasons of the year, the length of the seasonal cycle N depends on the nature of demand for the particular product or service. For example, demand for electricity has a daily cycle, demand for restaurants has a weekly cycle, while demand for clothes has a yearly cycle. The demand for many products may have

several interacting cyclic patterns. For example, electricity consumption has daily weekly, and yearly seasonal patterns.

Maintenance workload may show seasonal variation due to periodic changes in demand, weather, or operational conditions. If demand for products is seasonal, then greater production rates during the high-season intensify equipment utilization and increase the probability of failure. If demand is not seasonal, high temperatures during summer months may cause overheating and more frequent equipment failures. Plotting the data is important to judge whether or not it has seasonality, trend, or both patterns. Methods are presented below for forecasting with stationary seasonal data and seasonal data that has a trend.

1.4.5.1 Forecasting for Stationary Seasonal Data

The model representing this data is similar to the model presented in equation (1.1), but it allows for seasonal variations

$$D_t = c_t\mu + \varepsilon_t \tag{1.21}$$

where

c_t = seasonal factor (multiplier) for time period t, $1 \le t \le N$,

$\sum_{t=1}^{N} c_t = N$

Given data for at least 2 cycles ($2N$), four simple steps are used to obtain forecasts for each period in the cycle:

1. Calculate the overall average μ.
2. Divide each point by the average μ to obtain seasonal factor estimate.
3. Calculate seasonal factors c_t by averaging all factors for similar periods.
4. Forecast by multiplying μ with the corresponding c_t for the given period.

Example 6: The quarterly totals of maintenance work orders are given below for the last 3 years. Forecast the number of maintenance work orders required per quarter in year 4.

	Quarter 1	**Quarter 2**	**Quarter 3**	**Quarter 4**
Year 1	7,000	3,500	3,000	5,000
Year 2	6,000	4,000	2,500	5,500
Year 3	6,500	4,500	2,000	4,500

Step 1:

Sum of all data = 54,000

Overall average μ = 54,000/12 = 4,500

Step 2 and 3:

Dividing data by 4,500 and averaging columns gives the values in table below.

Table 1.4. Calculations for the stationary seasonal forecasting example

	Quarter 1	**Quarter 2**	**Quarter 3**	**Quarter 4**
Year 1	1.5556	0.7778	0.6667	1.1111
Year 2	1.3333	0.8889	0.5556	1.2222
Year 3	1.4444	1	0.4444	1
Average = c_t	1.4444	0.8889	0.5556	1.1111

Note that sum of the 4 seasonal factors (1.4444 + … + 1.111) is equal to 4, which is the length of the cycle N (4 quarters).

Step 4:

Finally, the forecasted maintenance work orders for each quarter in year 4 are given by:

$F_1 = 1.4444(4{,}500) \cong 6{,}500$ Quarter 1
$F_2 = 0.8889(4{,}500) \cong 4{,}000$ Quarter 2
$F_3 = 0.5556(4{,}500) \cong 2{,}500$ Quarter 3
$F_4 = 1.1111(4{,}500) \cong 5{,}000$ Quarter 4

Of course, we could have obtained the forecasts directly by averaging the original data for each shift. However, going though all 4 steps ensures that the model is completely specified in terms of the mean value μ and seasonal factors $c_1, \ldots, c_N$.

1.4.5.2 Forecasting for Seasonal Data with a Trend

It is possible for a time series to have both seasonal and trend components. For example, the demand for airline travel increases during summer, but it also keeps growing every year. The model representing such data is given by:

$$D_t = c_t(a + bt) + \varepsilon_t \tag{1.22}$$

The usual approach to forecast with seasonal-trend data is to estimate each component by trying to remove the effect of the other one. Thus, several forecasting methods have been developed for this type of data, all of which basically using the same general approach which is to: (1) remove trend to estimate seasonality, (2) remove seasonality to estimate trend, and (3) forecast using both seasonality and trend. Among the simplest of these methods is the cycle average method, whose steps are described below.

1. Divide each cycle by its corresponding cycle average to remove trend.
2. Average the de-trended values for similar periods to determine seasonal factors $c_1, \ldots, c_N$. If $\sum c_t \neq N$, normalize seasonal factors by multiplying them with $N/\sum c_t$.

3. Use any appropriate trend-based method to forecast cycle averages.
4. Forecast by multiplying the trend-based cycle average by appropriate seasonal factor.

Example 7: For a university maintenance department, the number of work orders per academic term is given below for the last 3 years. Forecast the number of maintenance work orders required per term in year 4.

	Term 1	**Term 2**	**Term 3 (summer)**
Year 1	10,000	7,000	5,000
Year 2	12,000	8,000	6,000
Year 3	14,000	9,000	7,000

Unlike the previous example, the above seasonal data has an increasing trend from year to year. Calculations for seasonal factors (steps 1 and 2) are shown in the following tables.

Table 1.5. Calculating cycle averages

Term: t **Year: d**	**1**	**2**	**3**	**Cycle (year) Sum**	**Cycle average: A_d**
1	10,000	7,000	5,000	22,000	7,333.33
2	12,000	8,000	6,000	26,000	8,666.67
3	14,000	9,000	7,000	30,000	10,000

Table 1.6. Calculating seasonal factors by dividing by cycle averages

Term: t **Year: d**	**1**	**2**	**3**
1	1.364	0.955	0.682
2	1.385	0.923	0.692
3	1.400	0.900	0.700
Average = c_t	1.383	0.926	0.691

There is no need to normalize seasonal factors since their sum (1.383 + 0.926 + 0.691) is equal to 3, which is the length of the cycle N (3 terms).

Using regression, calculations for the trend components of cycle averages (step 3) are shown below.

Table 1.7. Calculating seasonal factors

	d	A_d	dA_d	d^2
	1	7,333.33	7,333.33	1
	2	8,666.67	17,333.33	4
	3	10,000	30,000	9
Sum	6	26,000	54,666.67	14

Using equations (1.10) and (1.11), the slope and intercept of the cycle averages are estimated as follows:

$$b = \frac{3(54{,}666.67) - 6(26{,}000)}{3(14) - 6^2} = 1{,}333.33$$

$$a = \frac{1}{3}\left[26{,}000 - 1{,}333.33(6)\right] \cong 6{,}000$$

The forecasting model for period (term) t of cycle (year) d is given by:

$$F_{d,\,t} = c_t[6{,}000 + 1{,}333.33d]$$

Forecasted maintenance work orders required per term in year 4 are calculated as:

$F_{4,\,1} = 1.383[6{,}000 + 1{,}333.33(4)] = 15{,}674$ Term 1
$F_{4,\,2} = 0.926[6{,}000 + 1{,}333.33(4)] = 10{,}495$ Term 2
$F_{4,\,3} = 0.691[6{,}000 + 1{,}333.33(4)] = 7{,}831$ Term 3

1.4.6 Box – Jenkins Time Series Models

Using the data correlation structure, Box-Jenkins models can provide excellent forecasts, but they require extensive data and complex computations, making them unsuitable for manual calculations. Although autocorrelation analysis is used to find best forecasting model for a given data, judgment plays a role, and the model is not flexible to changes in the data. The two basic Box-Jenkins types are the autoregressive (AR) and the moving average (MA) models. In autoregressive (AR) models, the current value of the time series depends on (is correlated with) previous values of same series. An autoregressive model of order p, which is denoted by AR(p), is given by:

$$x_t = a + \phi_1 x_{t-1-2} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t \tag{1.23}$$

where

$a, \phi_1 \ldots \phi_p$ = parameters of fit
ε_t = random error

In moving average (MA) models, the current value of the time series depends on previous errors. In a certain class of problems, the time series x_t can be represented by a linear combination of independent random errors $\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ that are drawn from a probability distribution with mean 0 and variance σ^2_{ε}. Usually the errors are assumed to be normal random variables. A moving average model of order q, denoted by MA(q), is expressed as:

$$x_t = \mu + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \ldots + \psi_q \varepsilon_{t-q} \tag{1.24}$$

where

μ = mean of the series
$\psi_1, \ldots . \psi_q$ = parameters of fit

The two models can be combined to form an autoregressive and moving average model of order p and q, denoted by ARMA(p, q). The order of the model is first determined by autocorrelation analysis, and then the values of the parameters are calculated. The aim is usually to find the model that adequately fits the data with the minimum number of parameters. Box and Jenkins (1970) in their book suggested a general methodology for developing an ARMA(p, q) model. The methodology consists of the three following major steps: (1) a tentative model of the ARMA(p, q) class is identified through autocorrelation analysis of the historical data, (2) the unknown parameter of the model are estimated, and (3) diagnostic checks are performed to establish the adequacy of the model or look for potential improvements.

Frequently, several forecasting models could be used to forecast the future maintenance workload. The forecasting techniques presented in the preceding sections may fit the given data with varying degrees of accuracy. In the following section, error analysis is presented as a tool for evaluating and comparing forecasts.

1.5 Error Analysis

If a single forecasting model is applied, error analysis is used to evaluate its performance and to check how closely it fits the given actual data. If several forecasting models are available for a particular set of data, then error analysis is used to objectively and systematically compare the alternative models in order to choose the best one.

The forecasting error ε_t in time period t is defined as the difference between the actual and the forecasted value for the same period

$$\varepsilon_t = D_t - F_t \tag{1.25}$$

The following error measures are available for checking an individual forecasting model adequacy and comparing among several forecasting models.

1. Sum of the errors (*SOE*)

$$SOE = \sum_{t=1}^{n} (D_t - F_t) \tag{1.26}$$

Usually used as a secondary measure, *SOE* can be deceiving as large positive errors may cancel out with large negative errors. However, this measure is good for checking bias, i.e., tendency of forecast values to consistently overestimate or underestimate actual values. If the forecast is unbiased, *SOE* should be close to zero.

2. Mean Absolute Deviation (*MAD*)

$$MAD = \frac{1}{n}\sum_{t=1}^{n} | D_t - F_t | \tag{1.27}$$

This measure neutralizes the opposite signs of errors by taking their absolute values. If errors are normally distributed, then $1.25 \times MAD$ is approximately equal to the standard deviation of errors, σ.

3. Mean Squared Error (*MSE*)

$$MSE = \frac{1}{n}\sum_{t=1}^{n} (D_t - F_t)^2 \tag{1.28}$$

This measure neutralizes the opposite signs of errors by squaring them. If errors are normally distributed, then *MSE* is approximately equal to the variance of errors σ^2.

4. Mean Absolute Percent Error (*MAPE*)

$$MAPE = \frac{100}{n}\sum_{t=1}^{n} \left| \frac{D_t - F_t}{D_t} \right| \tag{1.29}$$

This measure is an independent yard stick for evaluating the "goodness" of an individual forecast. All the other measures only compare different forecasting models relative to each other.

Example 8: Given the actual forecasted values in the table below, calculate the different error measure for the associated forecasting model.

t	1	2	3	4	5
D_t	7	10	9	11	14
F_t	6	8	10	12	14

Intermediate calculations are shown in the table below.

Table 1.8. Calculating error measures

t	1	2	3	4	5	Sum
ε_t	1	2	– 1	– 1	0	1
$\lvert\varepsilon_t\rvert$	1	2	1	1	0	5
ε_t^2	1	4	1	1	0	7
$100\lvert\varepsilon_t/D_t\rvert$	14.29	20	11.11	9.09	0	54.49

Therefore

SOE = 1

MAD = 5/5 = 1.0

MSE = 7/5 = 1.4
MAPE = 54.49/5 = 10.9%

1.6 Forecasting Maintenance Workload

Different types of maintenance workload require different forecasting approaches. Kelly (2006) categorizes maintenance workload into the following types:

A. First-line maintenance workload: maintenance jobs are started in the same shift in which problems arise and completed in less than 24 hours.

1. *Corrective emergency*. Unplanned and unexpected failures that requires immediate attention for safety or economic reasons. The frequency of occurrence and the volume of work are random variables, but the volume of maintenance work is usually huge.
2. *Corrective deferred minor*. Similar to emergency workload, the frequency and volume of maintenance work are random. However, there is no urge for immediate attention. Therefore, maintenance jobs in this category can be delayed and scheduled when the time and conditions are more convenient.
3. *Preventive routine*. Frequent, short-duration planned maintenance workload, such as inspection, lubrication, and minor part replacement.

B. Second-line maintenance workload: maintenance jobs last less than two days and require one or few maintenance workers.

4. *Corrective deferred major*. Very similar to corrective deferred minor maintenance workload, but requires longer times and greater resources.
5. *Preventive services*. Similar to preventive routine maintenance workload, but the frequency is lower, and the work is usually done offline, usually in the weekend breaks or during scheduled shutdowns.
6. *Corrective reconditioning and fabrication*. Similar to deferred major maintenance workload, but the work is performed away from the plant, by another group of maintenance workers.

C. Third-line maintenance workload: maintenance jobs require maximum demand for resources, long durations and all craft types, at intermediate and long-term intervals.

7. *Preventive major work (overhauls, etc.).* Less frequent, off-line major preventive maintenance that involves overhauling major pieces of equipment or plant sections.
8. *Modifications*. Infrequent, off-line major preventive work that involves process or equipment redesign. This category typically involves the largest capital cost.

Kelly (2006) suggests the following techniques for forecasting the three types of line maintenance workload:

1. *First-line maintenance workload*. A queuing model should be used to represent the size of the first-line maintenance workload. The average maintenance workload is estimated by the average number of man-hours per hour or per day.
2. *Second-line maintenance workload*. The average maintenance workload is estimated by the average number of man-hours per week. This average should be prioritized and updated according to the plant condition.
3. *Third-line maintenance workload*. Long-range (5-year) overhaul and shutdown plans are used to predict maintenance workloads and associated resource requirements.

The above discussion focuses on forecasting maintenance workload for existing plants. For new plants, forecasting the maintenance load is more challenging due to the lack of historical data. In such cases, we must revert to qualitative or subjective forecasting techniques presented in section 1.3.

1.7 Maintenance Capacity Planning

Capacity is the maximum output that can be provided in a specified time period. In other words, capacity is not the absolute volume of work performed or units produced, but the rate of output per time unit. Maintenance capacity planning aims to find the optimum balance between two kinds of capacity: available capacity, and required capacity. Available capacity is mostly constant because it depends on fixed maintenance resources such as maintenance equipment and manpower. On the other hand, required capacity (or maintenance workload) is mostly fluctuating from one period to another according to trend or seasonal patterns.

Effective maintenance capacity planning depends on the availability of the right level of maintenance resources. Resource planning is the process of determining the right level of resources over a long-term planning horizon. Usually, resource planning is done by summing up quarterly or annual maintenance reports and converting them into gross measures of maintenance capacity. Resource planning is a critical strategic function, with serious consequences for errors. If the level of resources is too high, then large sums of capital will be wasted on unused resources. If the level of resources is too little, then lack of effective maintenance resources will reduce the productivity and shorten the life of manufacturing equipment.

Maintenance capacity planning is one function of maintenance capacity management. The other function is maintenance capacity control, in which actual and planned maintenance outputs are compared, and corrective action is taken if necessary. Usually, both available and required capacities are measured in terms of standard work hours. The required capacity for a given period is the sum of standard hours of all work orders, including setup and tooling times. The process of maintenance capacity planning can be briefly described as follows.

1. Estimate (forecast) the total required maintenance capacity (maintenance workload) for each time period.
2. For each time period, determine the available maintenance capacity of each maintenance resource (e.g., employees, contract workers, regular time, and overtime).
3. Determine the level of each maintenance resource to assign to each period in order to satisfy the required maintenance workload.

The main problem in maintenance capacity planning is how to satisfy the required maintenance workload in each period. Typically, in certain time periods, excessive workload or shortage of available resources necessitate the delay of some work orders to later periods. Therefore, maintenance capacity planning has to answer two questions in order to satisfy the demand for any given period: (1) how much of each type of available maintenance capacity (resource) should be used, and (2) when should each type of resource be used. The usual objective of maintenance capacity planning is to minimize the total cost of labor, subcontracting, and delay (backlogging). Other objectives include the maximization of profit, availability, reliability, or customer service.

Capacity planning techniques are generally characterized by 12-month planning horizons, monthly time periods, fluctuating demand, and fixed capacity. Four basic strategies are used to match the fixed capacity with fluctuating monthly demands:

1. Chase strategy: performing the exact amount of maintenance workload required for each month, without advancing or delay.
2. Leveling strategy: the peaks of demand are distributed to periods of lower demand, aiming to have a constant level of monthly maintenance activity.
3. Demand management: the maintenance demand itself is leveled by distributing preventive maintenance equally among all periods.
4. Subcontracting: regular employees perform a constant level of monthly maintenance activity, leaving any excess workload to contractors.

The above capacity planning strategies are considered pure or extreme strategies that usually perform poorly. The best strategy is generally a hybrid strategy, which can be found by several available techniques. Capacity planning techniques are generally classified into two main types: deterministic and stochastic techniques. Deterministic techniques assume that the maintenance workload and all other significant parameters are known constants. Two deterministic techniques will be presented in the following section:

1. Modified transportation tableau method.
2. Mathematical programming.

Stochastic capacity planning techniques assume that the maintenance workload and possibly available capacity and other relevant parameters are random variables. Statistical distribution-fitting techniques are used to identify the probability distributions that best describe these random variables. Since uncertainty always exists, statistical techniques are more representative of real life. However,

statistical models are generally more difficult to construct and solve. The two following stochastic techniques will be presented in section 1.9:

1. Queuing models.
2. Stochastic simulation.

1.8 Deterministic Approaches for Capacity Planning

The modified transportation tableau method and mathematical programming are presented in following subsections.

1.8.1 Modified transportation tableau method

For each maintenance craft, the required capacity is given by the forecasted workload for each period. The available capacity for each period is given by the quantity of available resources of different categories. Each of these categories, such as regular time, overtime, and subcontract, has its own cost. Generally, it is possible to advance some required preventive maintenance work to earlier periods or delay some maintenance work to later periods. However, any advance or delay (backlogging) has an associated cost which is proportional to the volume of shifted work and the length of the time shift. Therefore, the heuristic solution tries to find the least-cost assignment of the required workload in terms of quantity (to different resources) and timing (to different time periods).

The maintenance capacity planning problem is formulated as a transportation model, where the "movement" is not in the space domain, but in the time domain. Maintenance work is "transported" from periods in which the work is performed (sources) to periods where the work is required (destinations). Specifically, each work period is divided into a number of sources that represent the number of maintenance work resources available in the period. Thus, if the planning horizon covers N periods, and if m maintenance resources (e.g., regular, overtime, and subcontract) are available in each period, then the total number of sources in mN. The supply for each source is equal to the capacity of each resource in the given period. The demand for each destination is the required workload for the given period. Notation used in the transportation tableau is defined as:

c_m = cost of maintenance with resource m per man-hour
c_A = cost of advancing (early maintenance) per man-hour per unit time
c_B = cost of backordering (late maintenance) per man-hour per unit time
$Q_{m,t}$ = capacity of maintenance resource m in period t
D_t = maintenance demand (required workload) in period t

The total cost of performing maintenance with resource (m) in month (i) to satisfy demand in month (j) is given by:

$$TC_{i,j,m} = \begin{cases} c_m + (j-i)c_A, & i \le j \\ c_m + (i-j)c_B, & i > j \end{cases} \tag{1.30}$$

The following transportation tableau shows the setup for a three-period planning horizon, with three resources for maintenance work in each period ($m = R$: regular time, $m = O$: overtime, $m = S$: subcontract). Assigning an infinite cost (∞) prohibits assigning any maintenance work to the given (i, j) cell. For example, assigning a cost of (∞) to cells where ($i < j$) would prevent early execution of preventive maintenance work, i.e., execution before the due date.

Table 1.9. Transportation tableau for three-periods and three maintenance resources

Execution Periods	**Resource used**	**Demand Periods**			**Capacity**
		1	**2**	**3**	
1	Regular Time	c_R	$c_R + c_A$	$c_R + 2c_A$	$Q_{R,1}$
	Overtime	c_O	$c_O + c_A$	$c_O + 2c_A$	$Q_{O,1}$
	Subcontract	c_S	$c_S + c_A$	$c_S + 2c_A$	$Q_{S,1}$
2	Regular Time	$c_R + c_B$	c_R	$c_R + c_A$	$Q_{R,2}$
	Overtime	$c_O + c_B$	c_O	$c_O + c_A$	$Q_{O,2}$
	Subcontract	$c_S + c_B$	c_S	$c_S + c_A$	$Q_{S,2}$
3	Regular Time	$c_R + 2c_B$	$c_R + c_B$	c_R	$Q_{R,3}$
	Overtime	$c_O + 2c_B$	$c_O + c_B$	c_O	$Q_{O,3}$
	Subcontract	$c_S + 2c_B$	$c_S + c_B$	c_S	$Q_{S,3}$
Maintenance Demand		D_1	D_2	D_3	

After the modified transportation tableau is constructed, it is solved by the least-cost assignment heuristic. This heuristic assigns as much as possible (the minimum of supply and demand) to the available (unassigned) cell with the least cost. After each assignment, the supply and demand for the given cell are updated, and the process continues until all demands have been assigned. Although this technique does not guarantee an optimum solution, it is an effective heuristic that frequently leads to optimum solutions.

Example 9: The required maintenance workload for the next four months is 400, 60, 300, and 500 man-hours, respectively. The demand can be met by either regular time at a cost of $13 per hour or over-time at a cost of $20 per hour. Regular time and overtime capacities are respectively 400 and 100 hours per month. Early maintenance costs $3 per hour per month, while late maintenance costs $5 per hour per month. Using the modified transportation method, develop the capacity plan to satisfy the required workload.

Table 1.10 shows the modified transportation tableau for this example, with hourly costs at the corners of relevant cells. Using the least-cost assignment heuristic, the capacity plan solution is shown in the table, where the highlighted cells indicate active maintenance assignments. The demands of both month 1 and 3 are entirely met by regular time maintenance in the same month. The demand of month 2 is met by regular time maintenance in month 2 in addition to overtime maintenance in months 2 and 3. Finally, the demand of month 4 is met by both regular time and overtime maintenance in month 4. The total cost (*TC*) of the plan is obtained by multiplying the assigned hours by the corresponding costs.

$$TC = 13(400) + 13(400) + 20(100) + 16(100) + 13(300) + 13(400) + 20(100) = \$25{,}100$$

Table 1.10. Data and solution of Example 9

Execution months	**Resources used**	**Demand months**				**Capacity**
		1	2	3	4	
1	Regular time	13 **400**	18	23	28	400
	Overtime	20	25	30	35	100
2	Regular time	16	13 **400**	18	23	400
	Overtime	23	20 **100**	25	30	100
3	Regular time	19	16 **100**	13 **300**	18	400
	Overtime	26	23	20	25	100
4	Regular time	22	19	16	13 **400**	400
	Overtime	29	26	23	20 **100**	100
Maintenance demand		400	600	300	500	

The transportation tableau method is useful for simple cost functions. More complicated relations and cost structures, e.g., the cost of hiring and firing, require more sophisticated methods such as mathematical programming.

1.8.2 Mathematical programming methods

Taha (2003) provides a thorough discussion of mathematical programming models and solution techniques. Mathematical programming is a class of optimization models and techniques that includes linear, nonlinear, integer, dynamic, and goal programming. In general, a mathematical programming model is composed of decision variables, one or more objective functions, and a set of constraints. The objective function(s) and all constraints are functions of the decision variables and other given parameters. In linear programming (LP), all of these functions are linear functions. The objective function is the target of optimization, such as the maximum profit or minimum cost. The constraints are equations or inequalities representing restrictions or limitations that must be respected, such as limited capacity. The decision variables are values under the control of the decision maker, whose values determine the optimality and feasibility of the solution.

A solution, specified by fixed values of the decision variables, is considered optimal if it gives the best value of the objective function, and is considered feasible if it satisfies all the constraints. Optimum solutions of small models can be the Solver tool in Microsoft Excel. Larger models are solved by specialized optimization software packages such as LINDO and CPLEX. In addition to optimal values of decision variables, LP solutions obtained by these packages include values of slacks and surpluses, dual prices, and sensitivity analysis (ranges of given parameters in which the basic solution remains unchanged).

Many variations of mathematical programming models could be constructed for maintenance capacity planning. Depending on the particular situation, the decision variables, objective function, and constraints must be formulated to match the given needs and limitations. For example, options such as overtime, subcontracting, hiring and firing, and performing early preventive maintenance may or may not be applicable to a given maintenance capacity planning situation. Similarly, each situation calls for a different objective such as minimum cost or maximum safety, reliability, or availability. Examples of different variations of mathematical programming models for maintenance capacity planning are given by Alfares (1999), Duffuaa et al. (1999, pp. 139-144), and Duffuaa (2000).

The mixed integer programming model presented below is only a general-purpose example. Different components of this model could be added, deleted, or modified in order to tailor it to a specific maintenance capacity planning application.

Parameters

c_A (c_B) = cost of advancing (backlogging) each maintenance hour by one month, i.e., cost of early (late) maintenance
c_R(c_O) = cost of regular time (overtime) maintenance per hour
c_S = cost of subcontract maintenance per hour
c_H (c_F) = cost of hiring (firing) one worker
$n_{R,t}$ = number of regular time work hours per worker in month t
$n_{O,t}$ = maximum number of overtime hours per worker in month t
$N_{S,t}$ = number of subcontract work hours available in month t
D_t = demand (forecast) in month t

Decision variables (for each month t)

W_t	= workforce size	R_t	= regular time hours
O_t	= overtime hours	S_t	= subcontracted hours
A_t	= advanced hours	B_t	= backordered hours
H_t	= number hired	F_t	= number fired

Objective function

$$\min TC = \sum_{t=1}^{T} c_R R_t + c_O O_t + c_S S_t + c_A A_t + c_B B_t + c_H H_t + c_F F_t \qquad (1.31)$$

Constraints

$$W_t = W_{t-1} + H_t - F_t, \qquad t = 1, \ldots, T \qquad (1.32)$$

$$A_t - B_t = A_{t-1} - B_{t-1} + R_t + O_t + S_t - D_t, \qquad t = 1, \ldots, T \qquad (1.33)$$

$$R_t = n_{R,t} W_t, \qquad t = 1, \ldots, T \qquad (1.34)$$

$$O_t \leq n_{O,t} W_t, \qquad t = 1, \ldots, T \qquad (1.35)$$

$$S_t \leq N_{S,t}, \qquad t = 1, \ldots, T \qquad (1.36)$$

$$W_t, R_t, O_t, S_t, A_t, B_t, H_t, F_t \geq 0, \quad W_t, H_t, F_t \text{ integer}, \qquad t = 1, \ldots, T \qquad (1.37)$$

The objective function (1.31) aims to minimize the total cost TC of maintenance for all periods in the planning horizon. Constraints (1.32) and (1.33) respectively balance the workforce size and the maintenance workload between adjacent periods. Constraints (1.34) and (1.35) respectively relate regular time and overtime work hours to the number of regular maintenance workers in each period. Constraints (1.36) ensure that the number of assigned subcontract work hours does not exceed the available limit in each period.

Example 10: Maintenance workload for the next five months is 2500, 1500, 1800, 2800 and 2200 man-hours. This workload can be met by employees on regular time at a cost of \$10 per hour, employees on overtime at a cost of \$15 per hour, or subcontractors at a cost of \$18 per hour. The initial workforce size is 10 employees. Each employee works for 150 regular time hours and a maximum of 60 overtime work hours per month. Maximum capacity of subcontract workers is 200 hours per month. Early maintenance costs \$8 per hour per month, while late maintenance costs \$14 per hour per month. For each employee, hiring cost is \$800

and firing cost is \$1000. Assuming zero starting and ending backlog, model and solve this capacity planning problem using mathematical programming.

The integer programming model is given by:

$$\min TC = \sum_{t=1}^{T} 10R_t + 15O_t + 18S_t + 8A_t + 14B_t + 800H_t + 1000F_t$$

subject to

$$W_1 = 10 + H_1 - F_1$$
$$W_t = W_{t-1} + H_t - F_t, \qquad t = 2, \ldots, 5$$
$$A_1 - B_1 = R_1 + O_1 + S_1 - 2500$$
$$A_2 - B_2 = A_1 - B_1 + R_2 + O_2 + S_2 - 1500$$
$$A_3 - B_3 = A_2 - B_2 + R_3 + O_3 + S_3 - 1800$$
$$A_4 - B_4 = A_3 - B_3 + R_4 + O_4 + S_4 - 2800$$
$$0 = A_4 - B_4 + R_5 + O_5 + S_5 - 2200$$
$$R_t = 150W_t, \qquad t = 1, \ldots, 5$$
$$O_t \le 60W_t, \qquad t = 1, \ldots, 5$$
$$S_t \le 200, \qquad t = 1, \ldots, 5$$

The optimum solution of the above model was obtained by the optimization software package LINDO. The minimum total cost *TC* is \$120,920. Decision variables with non-zero values are shown in the table below.

Table 1.11. Integer programming optimal solution of Example 10

Month *t*	1	2	3	4	5
Workforce size W_t	11	11	12	15	15
Regular time hours R_t	1650	1650	1800	2250	2250
Overtime hours O_t	660	0	0	500	0
Subcontract hours S_t	40	0	0	0	0
Backlogged hours B_t	150	0	0	50	0
Hired employees H_t	1	0	1	3	0

1.9 Stochastic Techniques for Capacity Planning

Stochastic models for capacity planning consider various uncertainties ever present in real-life maintenance systems. Uncertainties in maintenance surround both maintenance workload or demand (i.e., timing and severity of equipment failure) and maintenance capacity (i.e., availability and effectiveness of maintenance resources). Usually, uncertainties are represented by probability distributions with specified values of the means and variances. Stochastic models for maintenance capacity planning include queuing models, simulation models, and stochastic programming. Stochastic programming models are mathematical programming models similar to the deterministic models discussed in the previous section, except that some of their elements are probabilistic. Although these models have been used for maintenance capacity planning (e.g., Duffuaa and Al-Sultan, 1999),

they are beyond the scope of this chapter, and thus will not be discussed further. In the remainder of this section, queuing theory models and computer simulation models are presented.

1.9.1 Queuing Models

Queuing models deal with systems in which customers arrive at a service facility, join a queue, wait for service, get service, and finally depart from the facility. Queuing theory is used to determine performance measures of the given system, such as average queue length, average waiting time, and average facility utilization (Taha, 2003). In addition, queuing models can be used for cost optimization, by minimizing the sum of the cost of customer waiting and the cost of providing service. In applying queuing theory to maintenance systems, the maintenance jobs or required maintenance tasks are considered as the customers, and maintenance resources such as manpower and equipment are considered as the servers.

Queuing systems differ from each other in terms of several important characteristics. To clearly define the characteristics of the given queuing situation, a standard notation (Taha, 2003) is used in the following format:

$(a/b/c):(d/e/f)$

where

- a = customer inter-arrival time distribution
- b = service time (or customer departure) distribution
- c = number of parallel servers
- d = queue discipline, i.e., order or priority of serving customers
- e = maximum number of customers allowed in the system (queue plus service)
- f = size of the total potential customer population

Standard symbols are used to represent individual elements of the above notation (symbols a and b). Arrival and service distributions (symbols *a* and *b*) are represented by the symbols *M* (Markovian or Poisson), *D* (deterministic or constant), *E* (Erlang or Gamma), and *G* (general). The queue discipline (symbol *d*) is represented by the symbols: *FCFS* (first come, first served), *LCFS* (last come, first served), *SIRO* (service in random order), and *GD* (general discipline). The symbol *M* corresponds to the exponential or Poisson distributions. If the inter-arrival time is exponential, then the number of arrivals during a specific period is Poisson. These complementary distributions have a significant role in queuing theory because they have the Markovian (or forgetfulness) property, which makes them completely random In order to introduce specific queuing models for maintenance capacity planning, the following notation is defined.

- n = number of customers in the system (queue plus service)
- λ_n = customer arrival rate with n customers in the system
- μ_n = customer departure rate with n customers in the system
- ρ = server utilization = λ_n / μ_n
- p_n = probability of n customers in the system

L_s = expected number of customers in the system
L_q = expected number of customers in the queue
W_s = expected waiting time in the system
W_q = expected waiting time in the queue

Waiting time and the number of customers are directly related by Little's Law, one of the most fundamental formulas in queuing theory:

$$L_s = \lambda_{eff} W_s, \qquad \text{or} \qquad L_q = \lambda_{eff} W_q \tag{1.38}$$

where

λ_{eff} = effective customer arrival rate at the system

Most queuing models are applicable to maintenance capacity planning. Two of these models are presented below, namely the (*M/M/c*):(*GD/∞/∞*) system and the (*M/M/R*) (*GD/k/k*) system.

1.9.1.1 The (M/M/c):(GD/∞/∞) system

This queuing system has and Markovian inter-arrival and service times, *c* parallel servers (repairmen), and general service discipline. Since there are no limits on the number of customers in the system, then $\lambda = \lambda_{eff}$. Defining $\rho = \lambda/\mu$, the steady-state performance measures for this system are given by:

$$L_q = \frac{\rho^{c+1}}{(c-1)!(c-\rho)^2}\, p_0 \tag{1.39}$$

$$L_s = L_q + \rho \tag{1.40}$$

where

$$p_0 = \left\{ \sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!(1-\rho/c)} \right\}^{-1}, \qquad \frac{\rho}{c} < 1 \tag{1.41}$$

The expected number of waiting time in the queue W_q and expected total time in the system W_s are respectively obtained by dividing L_q and L_s by λ.

The above model can be used in maintenance capacity planning to determine the optimum number of servers *c* (maintenance workers). In this case, the objective would be to minimize the total cost *TC* of waiting (i.e., cost of equipment downtime) plus the cost of providing maintenance (i.e., cost of maintenance workers). For example, this objective can be expressed as follows:

$$\min TC(c) = c_M\, c + c_W\, L_s(c) \tag{1.42}$$

where

c_M = cost of maintenance workers per employee
c_W = cost of waiting time in the queue

It should be noted that (1.42) is only a typical example of a relevant objective in maintenance capacity planning. Several alternative objective functions are possible; for instance, c could be replaced by μ, while L_s could be replaced by L_q, W_s, or W_q.

Example 11: A maintenance department repairs a large number of identical machines. Average time between failures is 2 hours and 40 minutes, and average repair time is 5 hours; both are exponentially distributed. The hourly labor cost is \$15 per maintenance employee, while the hourly cost of downtime is \$40 per waiting machine. Use queuing theory to determine the optimum number of maintenance employees.

λ = 1/2.6667 = 0.375
μ = 1/5 = 0.2
ρ = 0.375 /0.2 = 1.875

Since

$\rho/c = 1.875/c < 1$,

then

$c > 1.875$,

or

$c \geq 2$

For $c = 2$, the average number of waiting machines $L_s(2)$ and associated total cost $TC(2)$ are calculated by equations (1.39)-(1.42) as follows:

$$p_0(2) = \left\{ \sum_{n=0}^{2-1} \frac{1.875^n}{n!} + \frac{1.875^2}{2!(1-1.875/2)} \right\}^{-1} = \frac{1}{31} = 0.03226$$

$$L_q(2) = \frac{1.875^{2+1}}{(2-1)!(2-1.875)^2} \left(\frac{1}{31}\right) = \frac{421.875}{31} = 13.60887$$

$L_s(2) = 13.60887 + 1.875 = 15.48387$

$TC(2) = 15(2) + 40(15.48387) = 649.35$

For $c \geq 3$, the average number of waiting machines $L_s(c)$ and associated total cost $TC(c)$ are similarly calculated by equations (1.39) to (1.42). Because $TC(c)$ is convex, we should start with $c = 2$ and increment c by one employee at a time until the total cost $TC(c)$ begins to increase. The calculations are summarized in the

table below, showing that the optimum number of maintenance employees is equal to four.

Table 1.12. Queuing model solution of Example 11

c	$p_0(c)$	$L_s(c)$	$TC(c)$
2	0.03226	15.48387	649.35
3	0.13223	2.52066	145.83
4	0.14924	2.00265	140.11
5	0.15255	1.90328	151.13

1.9.1.2 The (M/M/R):(GD/K/K) system

This queuing system is called the machine repair or machine servicing model. It has Markovian inter-arrival and service times, R parallel servers (repairmen), and a general service discipline. This model represents the situation in a shop with K machines (customers). Therefore, K is both the maximum number of customers in the system and the size of the customer population. For this system, the number of repairmen must not exceed the number of machines, i.e., $R \le K$. Assuming that λ is break-down rate per machine, the steady-state results for this system can be derived as follows:

$$L_s = \sum_{n=0}^{K} n p_n \tag{1.43}$$

$$L_q = \sum_{n=R+1}^{K} (n-R) p_n \tag{1.44}$$

where

$$p_n = \begin{cases} \binom{K}{n} \rho^n p_0, & 0 \le n \le R \\ \binom{K}{n} \frac{n!\,\rho^n}{R!\,R^{n-R}} p_0, & R \le n \le K \end{cases} \tag{1.45}$$

$$p_0 = \left(\sum_{n=0}^{R} \binom{K}{n} \rho^n + \sum_{n=R+1}^{K} \binom{K}{n} \frac{n!\,\rho^n}{R!\,R^{n-R}} \right)^{-1} \tag{1.46}$$

The values of W_q and W_s can be calculated by respectively dividing L_q and L_s by λ_{eff}, which is given by:

$$\lambda_{eff} = \lambda(K - L_s) \tag{1.47}$$

Example 12: A manufacturing facility has 27 identical machines. On average, each machine fails every 4 hours. For each machine failure, average repair time is 30 minutes. Both the time between failures and the time for repair are exponentially distributed. The hourly cost for each repair station is \$18, while the hourly cost of lost production is \$55 per broken machine. Apply queuing theory to determine the optimum number of repairmen for this facility.

λ = 1/4 = 0.25
μ = 1/0.5 = 2
ρ = 0.25/2 = 0.125
K = 27

Starting with R = 1, $p_0(1)$, $p_1(1)$, and $L_s(1)$ are calculated by (1.43), (1.45), and (1.46) as follows:

$$p_0(1) = \left(\sum_{n=0}^{1} \binom{27}{n} 0.125^n + \sum_{n=1+1}^{27} \binom{27}{n} n!\, 0.125^n \right)^{-1} = \frac{1}{13{,}424{,}835.87}$$

$$p_n(1) = \binom{27}{n} \frac{n!0.125^n}{13{,}424{,}835.87}, \qquad n = 1, \ldots, 27$$

$$L_s(1) = \sum_{n=0}^{27} n p_n(1) = 19.00000$$

$$TC(1) = 18(1) + 55(19) = 1063.00$$

Since $TC(R)$ is convex, R is incremented by 1 at a time until $TC(R)$ starts to increase. The calculations are summarized in the table below. The optimum number of repair stations is equal to five.

Table 1.13. Queuing model solution of Example 12

R	$L_s(R)$	$TC(R)$
1	19	1063.00
2	11.07248	644.99
3	5.49428	356.19
4	3.67456	274.10
5	3.18612	265.24
6	3.04971	275.73
7	3.01224	1063.00

1.9.2 Stochastic Simulation

Simulation is a technique in which a computer model is constructed of a real-life system. This model allows us to observe the changing behavior of the system over time and to collect information about the required performance measures. In addition, this technique allows us to perform experiments on the simulation model that would be too expensive, too dangerous, or too time-consuming to perform on the real system. These experiments are performed by running the model under different conditions or assumptions (called scenarios) corresponding to different real-life options. Statistical inference techniques are used to analyze and interpret the results of simulation experiments.

According to Banks et al. (2005), simulation models are classified as static or dynamic, deterministic or stochastic, and discrete or continuous. A static (or Monte Carlo) model represents a system at a single given point in time. A dynamic model represents a system over a whole range of different time periods, showing the changing behavior of the system over time. A deterministic model is completely certain, because has no random variables. A stochastic model includes uncertainty in the form of random variables with specific probability distributions. In discrete simulation models, the system variables change discretely at specific points in time. In continuous simulation models, the system variables may change continuously over time.

Banks et al. (2005) propose the following 12-step procedure for building a sound simulation model.

1. Problem formulation: Develop a clear statement of the problem.
2. Setting of objectives and overall project plan: Specify the question to be answered by simulation, the alternative systems (scenarios) to be considered, and criteria to evaluate those alternatives.
3. Model conceptualization: Construct a simulation model of the real system, as simple as possible while capturing all the essential elements.
4. Data collection: Collect data to run and to validate the simulation model. This step is time consuming and interrelated with model conceptualization.
5. Model translation: Program the model in computer simulation software.
6. Model verification: Debug the program to ensure the model's logical structure is correctly represented in the computer.
7. Model validation: Compare the model to actual system, and calibrate the model make its performance measures as close possible to those of the actual system.
8. Experimental design: Determine length of initialization period, length of simulation runs, and number of replications of each run.
9. Production runs and analysis: Run the model, collect performance measures, and analyze results.
10. Additional runs: Based on analysis, perform more runs if needed.
11. Documentation and reporting: Prepare program documentation and manuals, in addition to report on simulation results and recommendations.
12. Implementation: Apply approved recommendations.

According to Kelly (2007), simulation allows us to consider many complex features of maintenance systems that cannot be easily included otherwise, such as redundant components, stand-by equipment, aging of machines, imperfect repairs, and component repair priorities. Simulation has been effectively used for maintenance capacity planning because it is capable of handling the inherent uncertainty and complexity in maintenance processes. For example, simulation has been used for determining the optimum number and schedule of maintenance workers, the optimum preventive maintenance policy, and the optimum buffer capacity between pairs of successive machines in a production line. Duffuaa et al. (1999, p. 148) consider simulation well suited for maintenance capacity planning, because of the following characteristics of maintenance systems:

- Complex interaction between maintenance functions and other technical and engineering functions.
- High interdependence of different maintenance factors on each other.
- Prevalence of uncertainty in most maintenance processes.

Numerous simulation models have been proposed for different maintenance systems. For example, Sohn and Oh (2004) use simulation to determine the optimal repair capacity at an IT maintenance center. Duffuaa et al. (2001) propose a generic conceptual simulation model that provides a general framework for realistic simulation models of maintenance systems. This model consists of seven modules:

1. Input module: provides all required data for the simulation model.
2. Maintenance load module: generates the maintenance workload.
3. Planning and scheduling module: assigns available resources to maintenance jobs and schedules them to meet workload requirements.
4. Materials and spares module: ensures availability of materials and supply for maintenance jobs.
5. Tools and equipment module: ensures availability of tools and equipment for maintenance jobs.
6. Quality module: ensures the quality of maintenance jobs.
7. Performance measures module: calculates various performance measures of the maintenance system.

Example 13: Alfares (2007) presents a simulation model for days-off scheduling of multi-craft maintenance employees. The maintenance workforce of an oil and gas pipelines department is composed of air conditioning (AC), digital (DG), electrical (EL), machinist (MA), and metal (ME) technicians. Using the workdays/off-days notation, maintenance workers can be assigned to only three days-off schedules: (1) the 5/2 schedule, (2) the 14/7 schedule, and (3) the 7/3-7/4 schedule. The simulation model considers stochastic workload variability, limited manpower availability, and employee work schedules. A simplified flowchart of the simulation model is shown in Figure 2. The model recommended optimum days-off assignments for the multi-craft maintenance workforce. These assignments are expected to reduce the time in the system W_s by an average of 25% for pipeline maintenance work orders.

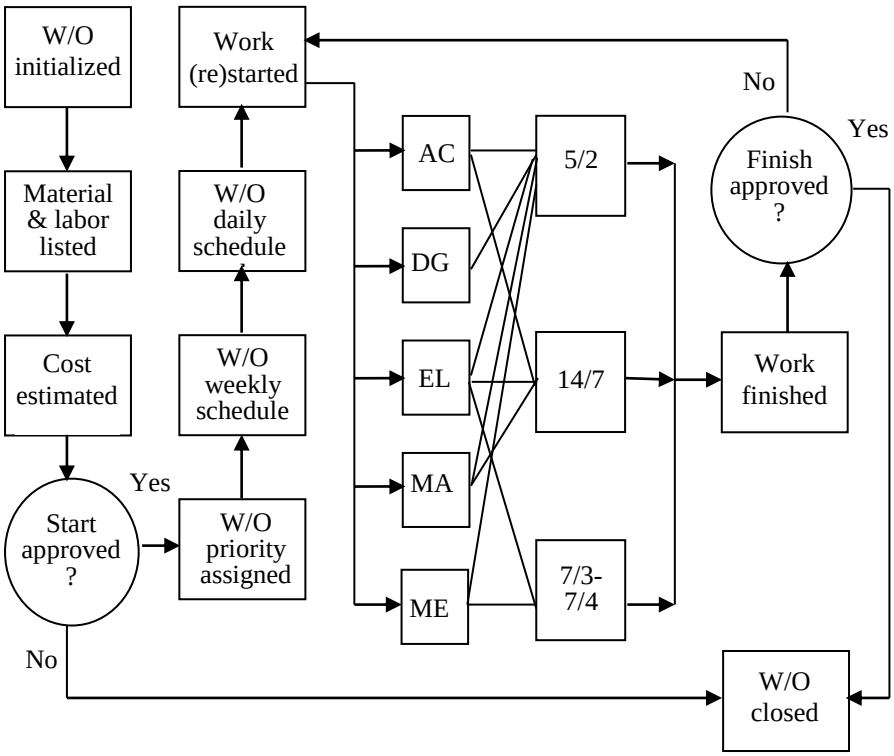

Figure 1.2. Simplified flowchart of the maintenance work order process

1.10 Summary

This chapter presented the basic ideas and procedures in maintenance forecasting and capacity planning. Forecasting has been defined as the prediction of future values, which forms the basis for effective planning. Forecasting techniques are classified into qualitative (subjective) and quantitative (objective). Subjective techniques are used in the absence of reliable numerical data, and include benchmarking, sales force composite, customer surveys, executive opinions, and the Delphi method. Quantitative or objective forecasting techniques are classified into time-series and causal models. Quantitative forecasting techniques presented for stationary, linear, and seasonal data include the moving average, exponential smoothing, least-squares regression, seasonal forecasting and Box-Jenkins ARMA models.

Error analysis was presented as an objective tool to evaluate and compare alternative forecasting models. Different forecasting approaches were presented to deal with the three types of line maintenance workload requirements: first-line, second-line, and third-line maintenance workloads.

Planning has been defined as preparing for the future, and it must be based on forecasting. Maintenance capacity planning aims to best utilize fixed maintenance resources in order to meet the fluctuating maintenance workload. Therefore, it has to determine when and how much of each type of available maintenance resources should be used. Capacity planning techniques are classified into deterministic and stochastic techniques. Deterministic techniques contain parameters that are known constants, and they include: the modified transportation tableau method, and mathematical programming. Stochastic techniques contain parameters that are random variables, and they include: queuing models, and stochastic simulation.

1.11 References

[1] Alfares HK, (1999) Aircraft maintenance workforce scheduling: a case study. Journal of Quality in Maintenance Engineering 5: 78–88

[2] Alfares HK, (2007) A simulation approach for stochastic employee days-off scheduling. International Journal of Modelling and Simulation 27: 9–15

[3] Banks J, Carson JS II, Nelson BL, Nicol DM, (2005) Discrete-Event Simulation, 4th Edition, Prentice Hall, Upper Saddle River, USA.

[4] Box GP, Jenkins GM, (1970) Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco, USA.

[5] Duffuaa SO, (2000) Mathematical Models in Maintenance planning and scheduling, in Ben-Daya M, Duffuaa SO, Raouf A (eds.), Maintenance Modeling and Optimization, Kluwer, Boston, USA, 39–53

[6] Duffuaa SO, Al-Sultan KS, (1999) A stochastic programming model for scheduling maintenance personnel. Applied Mathematical Modelling 23: 385–397

[7] Duffuaa SO, Ben-Daya M, Al-Sultan KS, Andijani AA, (2001) A generic conceptual simulation model for maintenance systems. Journal of Quality in Maintenance Engineering 7: 207–219

[8] Duffuaa SO, Raouf A, Campbell JD, (1999) Planning and Control of Maintenance Systems: Modeling and Analysis, John Wiley & Sons, New York, USA

[9] Kelly AD, (2006) Managing Maintenance Resources, Elsevier

[10] Kelly AD, (2007) The Maintenance Management Framework, Springer, London, 157–184

[11] Nahmias S, (2005) Production and Operations Analysis, Fifth Edition, McGraw-Hill, Singapore.

[12] Rowe, G and Wright, G, (2001) Expert Opinions in Forecasting. Role of the Delphi Technique. In Armstrong (Ed.), Principles of Forecasting: A Handbook of Researchers and Practitioners, Kluwer, Boston, 125–144.

[13] Sohn SY, Oh KY, (2004) Simulation study for the optimal repair capacity of an IT maintenance center based on LRD failure distribution. Computers & Operations Research 31: 745–759

[14] Taha H, (2003) Operations Research: An Introduction, 7th edition, Prentice Hall, Upper Saddle River, USA