

# Evaluating and Comparing Forecasting Models

Shouyi Wang

shouyi@eden.rutgers.edu

Department of Industrial & Systems Engineering, Rutgers University

Wanpracha Art Chaovalitwongse

wchaoval@rci.rutgers.edu

Department of Industrial & Systems Engineering, Rutgers University

**Abstract:** Forecasting is a very important aspect of any business, and has enormous social, economic, and environmental impacts. Various forecasting models have been developed to help people make right decisions against future uncertainties. However, all forecasting models have distinct advantages and limitations. Selecting appropriate forecasting methods from numerous alternatives is crucial to success. This paper briefly summarizes the state-of-the-art forecasting methods in terms of basic procedure, underlying assumptions, applications and limits. And then the most popular model selection criteria and guidelines are presented.

**Keywords:** forecasting models, forecasting model selection

## 1 Introduction

Strategic planning for the future is extremely important in any business to make definitive action plans. Due to its great importance, the knowledge of forecasting has been growing very rapidly in modern times. Many forecasting models have been developed to empower people in decision-making for various application areas. For example, accurate demand forecasts are essential for manufacturers to determine the optimal production rate by making a tradeoff between stock-outs and high inventory levels. Successful climate forecasting models have been developed to provide early warnings of adverse climatic conditions, such as hurricanes, storms, or frogs [24]. In business activities, forecasting technologies have become indispensable tools in a wide range of managerial decision-making processes, such as finance, banking, investments, employment, mortgages and loans [2].

Forecasts can be made based on either empirical qualitative analysis or mathematical quantitative analysis. Accordingly, forecasting models can be broadly classified as qualitative methods and quantitative methods. The flowchart of a typical forecasting modeling process is shown in Figure 1. And the categorization of forecasting models is shown in Figure 2. Fildes and Lusk [8] argued that no forecaster could be the ‘best’ method from the various forecasting competitions. All forecasting methods have distinct advantages and disadvantages. Therefore, selecting a right forecasting method is of critical importance to all decision makers. This paper briefly reviews the most influential forecasting models, and discusses the evaluating and comparing criterion/algorithms for model selection.

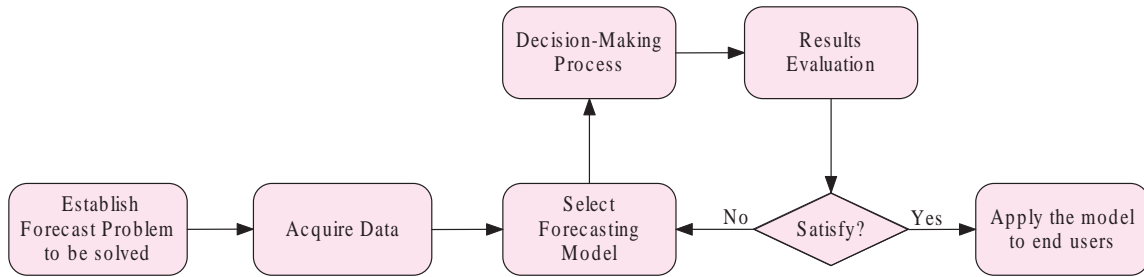


Figure 1: Flowchart of a typical forecasting modeling process.

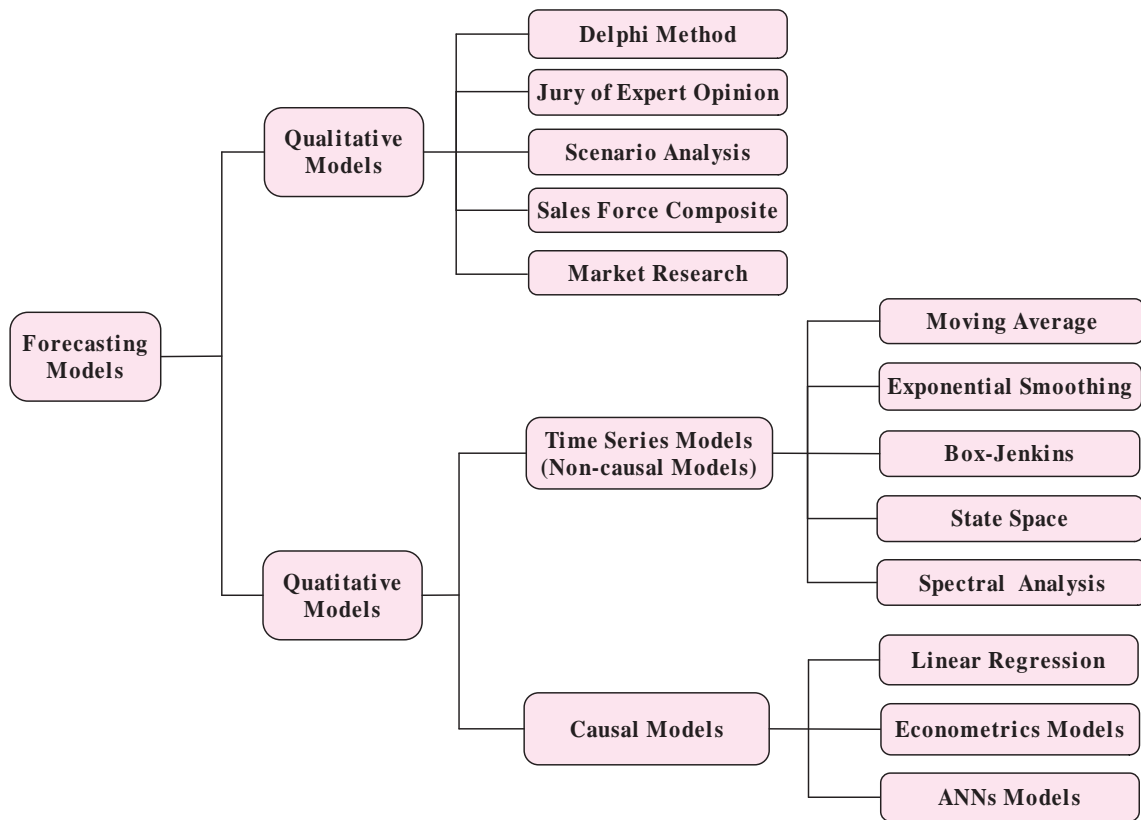


Figure 2: Categorization of Forecasting Models.

## 2 Qualitative Methods

Qualitative forecasting techniques rely primarily on human judgment based on expertise, experience, or intuition. They can be used in a wide range of circumstances where historical data are not available, or circumstances that are changing so rapidly that a mathematical forecasting model based on past data may become irrelevant or questionable. A number of qualitative forecasting methods have been developed to make good use of qualitative judgments from experts. Five of the more popular

qualitative methods are briefly presented in the following.

## **2.1 Delphi Method**

The Delphi method is a group technique which is based on structural surveys of a panel of experts about their perceptions of future events [19]. All of the experts answer a series of survey questions anonymously in several rounds. In the first round, the panel members are asked to write down their intuitive forecasts on the survey papers. Then, all the responses are summarized and fed back to each of the panel members. In the next round, the panel members may modify their original forecasts based on the responses of other panel members. This process usually generates a narrowing of opinions, and can be stopped by a predefined criterion, such as maximum number of rounds, degree of consensus, or stability of results. The final result of the process can be the mean or median results of the final round. The Delphi method provides more accurate forecasts than face-to-face group discussions, which are often highly influenced by those experts who have the best interaction and persuasion skills, or powerful background. With the advances in collaborative tools, such as email, synchronous video conference, web conference, Delphi Method is very easy to implement in modern times. However, one drawback of the Delphi method is that it is usually very time consuming.

## **2.2 Jury of Expert Opinion**

The jury of expert opinion is a method of developing forecasts in which the executives of a company are polled for their best estimate of the future trends, such as likely sales and demands [21]. It is also known as the ‘jury of executive opinion’ method. This method is one of the simplest and widely used forecasting methods in business. A typical outline of the ‘jury of expert opinion’ method is as follows. Firstly, each executive gets familiar with the background information as much as possible. Then in the meeting, all the executives write their estimates on survey papers. The final step is to combine all the options to produce an average and acceptable result among all the executives. The ‘jury of expert opinion’ method can be considered as an informal variant of the Delphi method. The only difference between them is that the ‘jury of expert opinion’ method does not have a mechanism to prevent interactions amongst the meeting participants.

## **2.3 Scenario Analysis**

Scenario analysis is a systematic thinking process to discover how various different factors may function together to create future [7]. Different scenarios can be generated by combining different policy plans with all known facts and anticipated changes about the future. Scenario analysis can often lead to plausible forecasts if the causal relationship between factors and their consequences have been appropriately simulated. Decision-makers can learn from the outcomes of different scenarios without

risking failures and losses in real business. To approach scenario analysis, the major steps are briefly as follows:

1. Determine the overall assumptions and factors on which the scenarios will be built. All assumptions and factors should be carefully made so that they can be conveniently under user's control and inspection. A time scope of the scenarios are also given.
2. Combine all the selected factors and assumptions together to construct a baseline framework. Each setting of the factors corresponds to a scenario. Given different levels of the influential factors, the outcome evolution of each scenario can be quite different.
3. Compare outcomes of the scenarios and pick up the most sensibly fitted scenarios. This step usually requires a considerable amount of debate to determine the selected scenarios.

## **2.4 Sales Force Composite**

Other than management and administration members, sales force can also be an important factor for qualitative forecasting analysis. The most popular method is known as sales force composite, which is based on forecasts of salespeople [5]. The assumption of the 'sales force composite' method is that salespeople are most qualified to predict future market development in their own territories. Since salespeople interact directly with costumers, they are presumably to have good estimate for market changes and trends. One possible disadvantage of this method is that forecasting results of salespeople tend to be optimistic. This is mainly because salespeople may always choose an optimistic prediction if a low estimate could endanger their employment status. Moreover, since sales force in one department may not be aware of impending changes in other areas, the opinions of salespeople in different territories should be combined into consideration in decision-making.

## **2.5 Market Survey**

To forecast future trends, another popular qualitative approach is to ask customers or potential users how they foresee their future consumption of a product or service. It is also called 'user's expectation' method. Since there is no benefit conflict, individual users are supposed to provide forecasts without undue optimism. Market survey has gained its popularity and importance in forecasting the success of new products [16]. However, one problem is that user opinions may be heavily biased by the trend of the moment. For example, most people have pessimistic opinions during periods of recession and have optimistic opinions in periods of growth and prosperity. Another disadvantage of the 'Market Survey' method is that it is often very expensive to obtain a comprehensive survey of costumers expectations.

## 3 Quantitative Methods

Quantitative methods make forecasts based on mathematical models rather than subjective judgment. These methods are the mainstream of forecasting techniques as a result of the great advances in mathematical modeling and computational power in modern times. Quantitative forecasting models have been utilized across a wide spectrum of business and industry. As shown in Figure 2, quantitative methods can be classified as non-causal models and causal models [15]. Non-causal models are also known as time-series models, which make forecasts by extracting systematic patterns (such as trends and seasonality) from historical time series data. Causal models are also known as cause-and-effect models, which investigate how the variable being forecasted is determined by its relevant influential factors. We will take a brief look at the leading quantitative methods in the following.

### 3.1 Time Series Models

Time series models predict values of the variable being forecasted based on historical patterns. Thus the basic underlying assumption of these methods is that future patterns are similar to historical patterns. To extract characteristics of time series patterns, four basic properties of data are often analyzed:

- **Trend.** Given a set of time series data, the term trend refers to a stable tendency of growth or decline exhibited in the data. The trend of a time series can be either linear or nonlinear. Accordingly, linear and nonlinear functions can be utilized to model the trend.
- **Seasonality.** If a pattern always repeats at a fixed interval, it is called a seasonal pattern. Seasonality is a very common characteristic of time series data. For example, air temperature exhibits a strong yearly seasonal pattern.
- **Cycles.** Cyclic patterns are similar to seasonal patterns, except that they repeat at varying intervals. For example, it is common to find nonstationary cycles in financial time series data.
- **Randomness.** Most time series data are assumed to contain both systematic patterns and random noises. The randomness usually makes the pattern difficult to identify. Most time series models include a noise term to take into account the effects of randomness.

Various time series methods have been developed to analyze these properties of time series data. Five of the most popular ones are moving average, exponential smoothing, Box-Jenkins models, state-space models, and spectral methods.

#### 3.1.1 Moving Average (MA)

MA models are simple but popular forecasting methods in time series analysis. A MA model involves taking arithmetic average of  $N$  most recent observations, where

$N$  is a specified number according to the nature of the data to be forecasted. For example, if you are forecasting monthly sales, you might use 12-month MA model, which takes the average sales over the past 12 months. A one-step-ahead MA model of  $N$  periods is given by

$$F_{t+1} = \frac{1}{N} \sum_{i=t-N+1}^t x_i = \frac{x_t + x_{t-1} + \cdots + x_{t-N+1}}{N}, \quad (1)$$

where  $F_{t+1}$  is the forecast for the  $t + 1$  period, and  $x_i$ s,  $i = t - N + 1, \dots, t$  are the observations in the past  $N$  periods. The mean of  $N$  most recent observations is used as the forecast of the next period. The moving averages method is probably the most commonly used technique to smooth out short-term fluctuations and capture characteristics of varying trends in time series data.

### 3.1.2 Exponential Smoothing

Exponential smoothing assigns exponentially decreasing weights as observations getting older. The most commonly used single exponential smoothing is given by

$$F_0 = x_0, \quad (2)$$

$$F_{t+1} = \alpha x_t + (1 - \alpha)F_t = F_t + \alpha(x_t - F_t), \quad (3)$$

where  $\alpha \in [0, 1]$  is the smoothing factor,  $F_{t+1}$  is the new forecast for next period,  $x_t$  is the current observation at period  $t$ , and  $F_t$  is the last forecast made in period  $t - 1$ . In the above formula, one can substitute  $F_t = \alpha x_{t-1} + (1 - \alpha)F_{t-1}$ , and continue so forth to obtain the infinite expansion of  $F_t$  as follows

$$F_t = \sum_{i=0}^{\infty} \alpha(1 - \alpha)^i x_{t-i-1} = \sum_{i=0}^{\infty} \alpha_i x_{t-i-1}, \quad (4)$$

where  $\alpha_i = \alpha(1 - \alpha)^i$ . From this expression, one can see clearly that the weight  $\alpha_i$  decreases exponentially with time. This illustrates why this method is called ‘exponential smoothing’. Single exponential smoothing works best only for stationary time series data. Double exponential smoothing has been developed to handle time series data with linear trends. And triple exponential smoothing has been proposed to deal with both trend and seasonality. A very detailed discussion of exponential smoothing techniques can be found in [9, 10].

### 3.1.3 Box-Jenkins Methods

Box-Jenkins methods, named after the statisticians George Box and Gwilym Jenkins, who applied autoregressive moving average (ARMA) to make forecasts for time series data [4]. An ARMA model can be generally described by

$$x_t = c + \epsilon_t + a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p}, -b_1 \epsilon_{t-1} - b_2 \epsilon_{t-2} - \cdots + b_q \epsilon_{t-q}, \quad (5)$$

$$= c + \sum_{i=1}^p a_i x_i - \sum_{j=1}^q a_j \epsilon_j, \quad (6)$$

where  $x_t$  is the current observation,  $x_{t-1}, \dots, x_{t-p}$  are the observations in the past  $p$  periods, and the  $a_1, \dots, a_p$  are the regression coefficients of the past  $p$  observations. The  $\epsilon_t$  is the current prediction error, the  $\epsilon_{t-1}, \epsilon_{t-2}, \dots, \epsilon_{t-q}$  are the past  $q$  prediction errors, and the  $b_0, b_1, \dots, b_q$  are the associated regression coefficients. ARMA model assumes that the time series to be analyzed is stationary. To handle the nonstationarities such as trend and seasonality, Box and Jenkins proposed a differencing version of ARMA model, which is known as ARIMA model. The ‘I’ stands for ‘Integrated’, since the estimation process is performed on differenced data, and the time series needs to be integrated before making a forecast. More mathematical details of Box-Jenkins model can be referred to Box et al. [4].

### 3.1.4 State Space Models

State space models virtually build up a generalized representation of linear time series models in state space form. For example, one can represent an ARIMA model in state space form. Once a state space model is built, it can be conveniently analyzed by Kalman filter and the associated smoother. Kalman filter is a recursive procedure to compute the optimal estimator of the state vector at time  $t$ , based on the information available at time  $t$ . The parameters of a state space model are usually estimated by maximum likelihood functions. State space method is a sophisticated form of forecasting models. A detailed discussion of various algorithms of state space models can be found in Harvey [14].

### 3.1.5 Spectral Analysis

Spectral analysis represents a group of methods which decompose time series data into a few underlying sine and cosine functions of different frequencies. Compared to ARIMA or Exponential Smoothing techniques, for which seasonal period is known as a priori in the analysis, spectrum analysis is suitable to deal with the seasonal series data for which lengths of cyclic patterns or fluctuations are changing rapidly or difficult to estimate. In spectral analysis, some important recurring cycles of different frequencies in the time series can be discovered. Those patterns may be hidden in random noises and are extremely difficult to find out by other methods. The most common spectrum decomposition process is also referred as Fourier analysis, which can be considered as a linear multiple regression process. The dependent variable is the time series to be studied, and the independent variables are sine and cosine functions of all possible frequencies. In general, a spectral decomposition model is give by

$$x_t = c + \sum_{j=1}^k (a_j \cos(\omega_j t) + b_j \sin(\omega_j t)), \quad 0 < \omega_1 < \dots < \omega_k < \pi \quad (7)$$

where  $\omega_1 < \dots < \omega_k$  are  $k$  possible wave lengths of the cyclic patterns in the time series,  $a_1, a_2, \dots, a_k$ , and  $b_1, b_2, \dots, b_k$  are regression coefficients that represent the degree of corresponding sinusoidal functions are correlated with the data. In other

words, a large sine or cosine coefficient indicates a strong periodicity of the respective frequency in the data. There are various techniques available to perform spectral analysis for time series data. A comprehensive discussion of spectral analysis can be found in Koopmans [17].

## 3.2 Casual Models

Causal models are also known as cause-and-effect models, which establish a causal relationship between the variable being forecasted and all other related variables. The most well known causal models are called regression models, which build up a rigorous mathematical model of causal relationship based on sound statistical techniques. In a regression model, the variable to be forecasted is called dependent or response variable, and the variables that represent the causal factors of the dependent variable are called independent or explanatory variables.

### 3.2.1 Regression Models

Regression models are in principle to investigate the relationship between one dependent variable and its relevant independent predictor variables. In general, a standard regression model takes the form as follows

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n, \quad (8)$$

where  $Y$  is the forecasted value of the dependent variable,  $b_0$  is the intercept, and  $b_1, b_2, \dots, b_n$  are the estimated regression coefficients representing the contribution of the independent predictor variables  $X_1, X_2, \dots, X_n$ , respectively. When linear regression models do not appear to adequately capture the relationships between dependent and predictor variables, nonlinear regression models (such as polynomial regression) can be used. An overview of the popular nonlinear regression models can be found in Seber and Wild [23].

### 3.2.2 Econometrics Models

In many real problems, the cause-and-effect relationship between dependent and independent variables are not straightforward. The estimated model parameters by the standard regression analysis may become inappropriate due to the highly dynamic relationship between the dependent and independent variables. For example, we wish to forecast sales of a product which are related to its price. However, the market price in turn is also affected by sales. Another typical example is the supply and demand model. The interaction of supply and demand jointly determines the equilibrium price and quantity of the product in the market. In such cases, it no longer makes sense to separate dependent and independent variables completely. To handle this problem, a set of simultaneous regression models is necessary to describe dynamics of these systems. The simultaneous regression models are called econometrics models in literature, since they are often applied to analyze the relationships between economic variables that should be jointly determined. One can consider that a single



regression model is a special case of econometrics models. The rigorous mathematical formulations of econometrics can be found in Pindyck and Rubinfeld [18].

### 3.2.3 ANN Models

Artificial neural networks (ANNs) represent another important form of causal models, which have shown powerful capabilities of modeling complex relationships between inputs and outputs. An ANN model consists of a network of neurons connected by arcs with assigned weights. Neurons take some form of basic nonlinear functions. Therefore, an ANN model can be equivalently considered as a nonlinear regression model in mathematics. A typical ANN has three layers, an input layer, a hidden layer, and an output layer. As a causal model, the inputs to an ANN are independent or explanatory variables, and the outputs are dependent or response variables being forecasted. There are various algorithms available to train ANNs, such as Perceptron learning rule and backpropagation [20]. Once the structure and weights of an ANN is determined, it can be employed to perform forecasting. ANNs have been increasingly used in forecast modeling in the past decade. They are suitable for complicated problems which are difficult to be mathematically formulated by regression models or econometric models. In many real applications, ANN methods can often achieve good performance if given enough training data. An overview of the applications of ANNs in forecasting can be found in Zhang et al. [27].

## 4 Comparing and Selecting Forecasting Models

We have discussed the most popular forecasting models above. Table 1 below summarizes the appropriateness of the three major types of forecasting models. To evaluate forecast models, two aspects of terms are often concerned about: accuracy and bias. Accuracy refers to the distance between the forecasts and actual values. And a forecast is biased if the errors in one direction are significantly larger than those in other directions. In general, the basic objective of all forecast models is to maximize accuracy and minimize bias. After fitting several model candidates to a given data set, the next step is to compare and select the best forecasting model. A lot of criteria have been proposed to compare forecasting models, which will be discussed in the following.

### 4.1 Forecast Error Measures

To achieve high prediction accuracy is the primary objective in most forecasting tasks. To evaluate forecasting accuracy, four of the more popular direct error measures are mean squared error (MSE), or its variants such as root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Minimizing these measures is usually the most essential criterion in comparing forecasting models. These measures are most frequently used due to their mathematical convenience. For each of these measures, a smaller value indicates higher prediction

Model Type	Applicable Situations	Horizon	Data Types
Qualitative Models	No Past data, no ideas about causality, data too costly to collect, short-run accuracy no required.	either long-run or short-run	background information, survey
Time Series Models	Historical time series data available, stable patterns exist, short-run accuracy needed.	short-run	time series data
Causal Models	Past data available, causal relationship is clear and stable, explanatory variables are controllable.	short-run	response and explanatory data

Table 1: The applicable situations for different types of forecasting models.

accuracy. Given a set of real data  $y_i, i = 1, 2, \dots, n$ , each of which has an associated forecast value  $\hat{y}_i$ , then these measures are defined as follows

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (9)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}, \quad (10)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (11)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \quad (12)$$

The R-squared ( $R^2$ ), also known as coefficient of determination, is another most commonly used criterion to evaluate a forecast model. The most general form of the  $R^2$  is defined as follows

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{SS_{err}}{SS_{tot}}, \quad (13)$$

where  $\bar{y} = \sum_{i=1}^n y_i/n$ , it is the mean of the observed data. The second term compares the variance of the forecast errors with the total variance of the data. One minus the second term is the proportion of variability in a data set that can be explained by the forecast model.  $R^2$  is used to measure how well a model approximates real data values. The magnitude of  $R^2$  is usually restricted within 0 and 1. An  $R^2$  close to 1.0 indicates that the model perfectly fits the data, while an  $R^2$  close to 0 means that the model cannot explain the data at all.

## 4.2 Information Criterion

A number of model selection criteria have also been developed based on information theory. A well-known criterion is Akaike's information criterion (AIC) developed by Akaike in 1974 [1]. It makes a tradeoff between accuracy and complexity in model construction. In general, this criterion can be defined as:

$$AIC(K) = \log(MSE) + \frac{2K}{n}, \quad (14)$$

where  $K$  is the number of parameters in the model,  $n$  is the number of observations in the data. The MSE has been defined above, and it can be explained as the estimated residual variance in this criterion. In the formula of AIC, the first term indicates model accuracy, and the second term indicates model complexity in terms of the number of parameters. Hence AIC not only rewards prediction accuracy, but also gives a penalty to larger number of model parameters. One major benefit of this penalty is to discourage overfitting. For a set of models, the one with the lowest AIC value is considered as the preferred model. This criterion is particular suitable for comparing a set of nested models. For example, compare an AR( $m$ ) model with an AR( $m+1$ ) for a given set of data. One drawback of AIC is that it is not consistent, since as the number of observations grows, the probability of selecting the correct model does not approach one.

The Bayesian information criterion (BIC), also known as Schwarz criterion, is another well-known criterion to select a set of parametric models with different choices of explanatory parameters [22]. BIC is actually a variant of AIC in a form of:

$$BIC(K) = \log(MSE) + \frac{\log(n)K}{n}. \quad (15)$$

The BIC also makes a tradeoff between accuracy and complexity of a model. For a set of models, the one with the lowest value of BIC is the one to be preferred. It differs from AIC in that the penalty coefficient of  $K$  becomes  $\log(n)/n$  instead of  $2/n$ . The BIC generally penalizes free parameters more strongly than does the AIC. In addition, Hannan and Quinn [13] also proposed an alternative to AIC and BIC called Hannan-Quinn criterion (HQC), which is given by

$$HQC(K) = \log(MSE) + \frac{2K \ln \ln(n)}{n}. \quad (16)$$

Similar to AIC and BIC, the model with the lower value of HQC is preferred. It has been shown that consistency can be obtained by the BIC [22] or HQC [13, 12].

## 4.3 Cross-Validation

Cross-validation is also a commonly used technique to compare different predictive models in practice [11]. Given a set of data, the basic idea of cross-validation is to partition the data set into training and validating subsets, and estimate the predictive accuracy on the validating data by the model obtained from the training data set.

The MSE, RMSE, MAE, and MAPE can be used to measure the expected level of fit of a predictive model. If a model fits the training data set very well but does not fit the validation data, it is called overfitting. A good predictive model is supposed to generate consistent results in both training and validating data sets. To reduce the variability of the model evaluation, multiple trials of cross-validation are usually performed with respect to different partitions to the original data set. The evaluation result of a predictive model is the averaged result over these trials. There are several different approaches to perform multiple steps of cross-validation.

- Repeated Random Partition Validation: this method simply divides the dataset into two subsets randomly each time and repeats the same procedure a number of times. One problem of this method is that the validation subsets may overlap and some observations may never be selected in the validation subsets.
- K-fold cross-validation: the dataset is partitioned randomly into  $K$  subsets. Then the cross-validation is repeated  $K$  times. Each time one of the subsets is reserved as the validation data, and the remaining  $K - 1$  subsets are the training data sets. The validation result is the average over  $K$  results. This approach guarantees that each observation can be used for validation exactly once.
- Leave-one-out cross-validation: this method is actually a special case of K-fold cross-validation, when the number of observations in each subset is one. In other words, only one observation is reserved for validation and the remaining observations are used for training. The procedure is repeated until each observation has been used once for validation.

#### 4.4 Stepwise Model Selection

Stepwise model selection approaches are very useful for automatically selecting a set of nested predictive models, for which there are a large number of potential predictive variables [6]. The selection procedure is generally grounded in some statistical tests and usually takes in the form of partial F-test. Other measures can also be used, such as t-tests,  $R^2$ , AIC, and BIC. Since the basic procedures are similar, only the case of partial F-test is discussed here. To compare two nested models with different number of predictor variables, the partial F-test can be generally formulated as follows:

$$F = \frac{\text{Extra sum of squares/Extra model df}}{\text{SSR of large model/Residual df Large model}}, \quad (17)$$

where SSE denotes the sum of squared residual. The key idea of this test is to check if the ‘extra’ predictors of the large model explain significantly more of the variability compared to the variability that is explained by the predictors that are already in the small model. Based on partial F-test, three approaches are commonly used for model selection:

- Forward selection: starts with the smallest number of possible predictors and adds predictors one by one until a stop criterion is satisfied or the largest model

is reached. The current model satisfying the stop criterion is selected. In particular, suppose that the current model has  $P$  parameters, and we want to test if one of the model with  $P + 1$  parameters is more preferred. If for all models of  $P + 1$  parameters, it satisfies

$$F = \frac{SSE(P + 1) - SSE(P)}{SSE(P + 1)/n - P - 1} < F_{m,P}, \quad (18)$$

where  $F_{m,P}$  is the critical values of F-statistic for a chosen level of significance. Then stop the process and the current model with  $P$  parameters is preferred. Otherwise, select one preferred model of  $P + 1$  parameters, and repeat the test for all models with  $P + 2$  parameters, and so forth.

- Backward selection: starts with the largest number of possible predictors and removes predictors one by one. At each step, it compares all the smaller model candidates with the old larger model and stops the process if

$$F = \frac{SSE(P - 1) - SSE(P)}{SSE(P)/n - P} < F_{m,P}. \quad (19)$$

- Stepwise selection: a modified version of forward-selection which allows the elimination of predictors those become statistically insignificant in the model. At each step of the process, the p-values of all predictors are computed. If the largest of these p-values is greater than a critical value, then the corresponding predictor is eliminated. Other steps are all the same with those of forward selection.

## 4.5 Residual Diagnostics

Residuals represent the portion of the validation data not explained by the model. The graphical residual analysis is commonly used in complementary with the quantitative techniques. A typical residual diagnostics includes plots of residuals versus the predicted values, versus other predictors, and versus time, residual autocorrelation plots, residual histogram, and normal probability plots. In general, the residual analysis can be used to test the following:

- Whiteness test: a good predictive model should have the uncorrelated residuals.
- Independence test: a good model should have residuals uncorrelated with past inputs.
- If there are some extreme influential observations. Identifying and deleting outliers from the training dataset may significantly improve the quality of a model.
- If the residuals exhibit systematic patterns and bias. The residuals of a good model should be approximately dispersed around zero evenly. If systematic patterns are found, the most probably reason is that one or several relevant predictors are missing. A forecast is biased if residuals in one direction are significantly larger than those of the other direction.

## 4.6 Qualitative Considerations

Any forecasting model will be inevitably used by people. Thus qualitative considerations can also be a very important factor in evaluating various forecasting models. The qualitative criteria includes intuitive reasonableness of the model, simplicity of the model, ease of use in practice, ease of explanation and understand, quality of forecast plots and demonstrations, etc. Moreover, if two models have similar predictive performance, the simpler model is usually preferred. One famous principle is called Occam's razor, which suggests that when competing theoretical models are equal in other respects, it is recommended to select the model which introduces the fewest assumptions while still sufficiently answering the question. More details of Occam's razor can be referred to Soklakov [25].

## 5 Conclusion

This paper reviews the most commonly used forecasting methods including both qualitative and quantitative methods. Quantitative methods are based on rigorous mathematical formulations, while qualitative methods are based on subjective judgment. Different forecasting methods have distinct different characteristics and applicable areas. To select appropriate forecasting models, one generally needs to consider the following important factors before model construction:

- What is the objective of forecasting, short-run or long-run?
- What types of data or explanatory variables are available?
- How much accuracy is required?
- Make a tradeoff between the costs and gains of developing a forecasting model.

Once a number of model candidates are selected, one should use one or more evaluation criteria to select the best forecasting model. Although quantitative models are mainstream approaches in forecasting, they are often used in combination with qualitative techniques involving human judgment. Broadly speaking, quantitative methods usually provide tools for decision support, while quantitative and qualitative techniques together are used for decision making in most of the real world cases. Therefore, a comprehensive assessment of forecasting performance usually consists of both quantitative and qualitative analyses to enhance the forecasting rationality in practice. A very good review of combining qualitative methods with quantitative techniques can be found in Webby and O'Connor [26] and Armstrong and Collopy [3].

## References

- [1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [2] A. Anari and J.W. Kolari. *Engineering Statistics Handbook*. Springer, 2009.
- [3] J.S. Armstrong and F. Collopy. *Integration of statistical methods and judgment for time series forecasting: principles from empirical research*, chapter Forecasting with Judgment, pages 269–293. John Wiley & Sons, 1998.
- [4] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time Series Analysis, Forecasting and Control*. Prentice Hall, 1994.
- [5] D.J. Dalrymple. Sales forecasting practices: results from a united states survey. *International Journal of Forecasting*, 3(3-4):379–391, 1987.
- [6] M.A. Efroymson. *Mathematical Methods for Digital Computers*, chapter Multiple regression analysis, pages 191–203. New York: Wiley, 1960.
- [7] L. Fahey and R.M. Randall. *Learning from the Future: Competitive Foresight Scenarios*. John Wiley & Sons, 1997.
- [8] R. Fildes and E.J. Lusk. The choice of a forecasting model. *Omega*, 12(5):427–435, 1984.
- [9] E.S. Gardner. Exponential smoothing: the state of the art. *Journal of Forecasting*, 4(1):1–28, 1985.
- [10] E.S. Gardner. Exponential smoothing: The state of the art—part ii. *International Journal of Forecasting*, 22(4):637–666, 2006.
- [11] S. Geisser. *Predictive Inference*. New York: Chapman and Hall, 1993.
- [12] E.J. Hannan. The estimation of the order of an arma process. *Annals of Statistics*, 8(5):1071–1081, 1980.
- [13] E.J. Hannan and B.G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, B(41):190–195, 1979.
- [14] A.C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1991.
- [15] K. Holden, D.A. Peel, and J. L. Thompson. *Economic Forecasting: An Introduction*, chapter Introduction to Forecasting Methods. Cambridge University Press, 1991.
- [16] K.B. Kahn. An exploratory investigation of new product forecasting practices. *The Journal of Product Innovation Management*, 19:379–391, 2002.

- [17] L.H. Koopmans. *The Spectral Analysis of Time Series*. Academic Press, 1995.
- [18] R. Pindyck and D. Rubinfeld. *Econometric Models and Economic Forecasts*. McGraw-Hill/Irwin, 1997.
- [19] G. Rowe and G. Wright. The delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, 15(4):353–375, 1999.
- [20] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagating errors. *Nature*, 323(9):533–536, 1986.
- [21] L.C. Satorius and N.C. Mohn. Sales forecasting models: a diagnostic approach. Research Monograph 69, College of Business Administration, Georgia State University, Atlanta, GA, 1976.
- [22] G.E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.
- [23] G.A.F. Seber and C.J. Wild. *Nonlinear Regression*. New York: John Wiley & Sons, 1989.
- [24] J. Shukla. Predictability in the midst of chaos: A scientific basis for climate forecasting. *Science*, 282(5389):728–731, 1998.
- [25] A.N. Soklakov. Occam’s razor as a formal basis for a physical theory. *Foundations of Physics Letters*, 15(2):107–135, 2002.
- [26] R. Webby and M. O’Connor. Judgemental and statistical time series forecasting: a review of the literature. *International Journal of Forecasting*, 12(1):91–118, 1996.
- [27] G. Zhang, B.E. Patuwo, and M.Y. Hu. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14(1):35–62, 1998.