# Automated Time-Series Cost Forecasting System for Construction Materials

Sungjoo Hwang[1]; Moonseo Park, M.ASCE[2]; Hyun-Soo Lee, M.ASCE[3]; and Hyunsoo Kim[4]

**Abstract:** As large-scale building projects increase in frequency, their construction costs become a matter of great concern, especially because of their lengthy construction periods. In particular, recent volatile fluctuations of construction material prices have fueled problems like cost forecasting. Many researchers try to accurately estimate cost escalations, but price forecasting for numerous construction materials requires a simplified and automated process. The research in this paper develops an automated time-series material cost forecasting (ATMF) system including both autoselected procedures for determining a best-fitting model and an autoextracting module for forecasting values using the Box-Jenkins approach. If the modeling process is simplified and iterative arbitrary decisions for the modeler eliminated, each future prices of a large number of materials can be forecast differently. Thus, the ATMF system can be utilized for predicting future trends in construction material costs. Further, an out-of-sample forecast applying several material price data confirms that this system can be effectively applied to material cost estimation at a more detailed level in object-based cost planning. The proposed system can thus help decision makers in the construction industry deal with changes in economic conditions and design by estimating cost escalations caused by volatile factors such as inflation. **DOI: [10.1061/(ASCE)CO.1943-7862.0000536](10.1061/(ASCE)CO.1943-7862.0000536).** © *2012 American Society of Civil Engineers.*

**CE Database subject headings:** Auto-regressive moving-average models; Construction costs; Predictions; Time series analysis; Construction materials.

**Author keywords:** Autoregressive moving-average models; Construction costs; Pricing; Predictions; Time-series analysis.

## Introduction

Large-scale construction projects have recently increased in number for residential, commercial, and government facilities worldwide. Examples of this trend are the numerous high-rise buildings (over 100 stories) being constructed or planned as urban landmarks. In these circumstances, construction costs for large-scale building projects have been a matter of great concern because of their significant cost implications and frequent design changes during their lengthy construction periods. Over the long time span from project startup to completion, many factors affect the final project cost (Shane et al. 2009). These volatile factors, such as resource prices, can lead to under- or overestimations of the total project cost, as resource prices vary because of changes in demand, market conditions, and macroeconomic conditions (Williams 1994). For example, structural steel prices tripled from 2001 to 2008 in Korea, according to recent global economic reports [Construction Association of Korea (CAK) 2008]. Therefore,

additional material costs caused by inflation can decrease general contractors' profitability over the course of 5 or more years for most large-scale projects, which are generally awarded with a lump-sum contract. As material costs represent approximately one-fourth of the total project cost, an accurate prediction of raw material prices is important for an accurate prediction of the total cost.

To solve these problems, many researchers have attempted to accurately estimate cost escalations in construction projects with a focus on total costs or construction cost index using estimating techniques such as time-series analysis, artificial neural networks, or other probabilistic methods. Time-series analysis, in particular, has been widely used because it represents the time-lagged relationship of correlated observations among both single and interrelated multiple series (Hwang and Liu 2010). Despite such strengths, however, two main problems are associated with applying this method to forecasting material cost escalations. First, numerous materials are required for construction projects. Because prices of various materials increase or decrease at different rates, total material cost escalation should be the sum of each material's price escalation. Second, the complicated and iterative procedures of most forecasting methods require significant time and effort both to determine suitable models and to forecast price escalations for each material required. For these reasons, a simplified and automated forecasting system is desirable for material cost estimation on large-scale construction projects.

This research develops an automated time-series system to easily determine a suitable forecasting model for a corresponding material. This system includes both autoselected procedures for determining a best-fitting forecasting model and an autoextracting module for forecast values.

According to estimators in construction companies, the material price trends to be forecast have the following characteristics: (1) no seasonal variations can be found in most material price time-series data; (2) material prices tend to remain steady even if they increase

[1]Ph.D. Student, Dept. of Architecture and Architectural Engineering, Seoul National Univ., 599 Gwanak-ro, Seoul, Korea. E-mail: nkkt14@snu.ac.kr

[2]Professor, Dept. of Architecture and Architectural Engineering, Seoul National Univ., 599 Gwanak-ro, Seoul, Korea (corresponding author). E-mail: mspark@snu.ac.kr

[3]Professor, Dept. of Architecture and Architectural Engineering, Seoul National Univ., 599 Gwanak-ro, Seoul, Korea. E-mail: hyunslee@snu.ac.kr

[4]Ph.D. Student, Dept. of Architecture and Architectural Engineering, Seoul National Univ., 599 Gwanak-ro, Seoul, Korea. E-mail: verserk13@naver.com

JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT © ASCE / NOVEMBER 2012 / **1259**

J. Constr. Eng. Manage., 2012, 138(11): 1259-1269

once in a recessionary period; (3) because numerous factors affect material prices, extensive data collection is required to make a single prediction (Williams 1994). To check the nonseasonality of material price data, this research analyzes the plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF) with regard to three main kinds of construction material: high-tensile rebar (SD400, D51, 15.9 kg/m), H-type steel beam (H700, B300, 185 kg/m), and ready-mixed concrete (30 Mpa, 15 cm, 40 cm). Fig. 1 shows the ACF and PACF of seasonal-transformed price data of the three materials. When there is seasonality in data sets, significant autocorrelation can be found at all high time lags except for the first or second time lag in the PACF. According to these plots, which have no significant seasonal autocorrelation, the nonseasonality of the material price data can be identified.

With a focus on nonseasonality and trends, this research applies the autoregressive and moving average (ARIMA) modeling method suggested by Box and Jenkins (1994), one of the univariate time-series modeling methods. Although various economic variables are believed to affect trends in material prices, multivariate approaches are not always feasible for quantifying the effect of various factors or representing relationships mathematically. In addition, future values of some factors are not available but must be estimated (Hwang 2010). On the other hand, the univariate time-series approach requires an examination of the relationship based solely on past behavior (Wong et al. 2005). Despite the simplicity of this analysis, the ARIMA method often produces the most accurate forecasting models for any set of data and employs a more systematic approach to building, analyzing, and forecasting time-series models by considering the trends and cyclic and seasonal elements of past data (Lu and AbouRizk 2009). When there are both non-stationary behavior and seasonality in time-series data, seasonal

ARIMA (SARIMA) can be applied by conducting additional seasonal differencing to remove seasonality. With material price data that have nonseasonality, however, this process for seasonality analysis can be omitted. The ARIMA model can thus be effectively applied to develop a simplified and automated forecasting system specialized for a large number of construction materials.

The paper begins with a study of existing cost-forecasting methods and time-series forecasting. With a focus on possibilities and limitations, an automated time-series forecasting system for construction materials is developed. Using an out-of-sample forecast on several structural construction materials, the proposed system is then evaluated. Finally, application methods of the system are outlined with conclusions.

## Literature Review

### Research on Cost Escalations

Many studies have focused on the rapidly changing construction material market and have attempted to address cost escalation factors to make cost planning more feasible. The main issues here are identifying escalation factors and estimating project costs accurately and simply. Ranasinghe (1996), for example, presented a simplified model for total project cost by considering the effects of inflation. Akpan and Igwe (2001) developed a suitable model for the evaluation of cost overruns due to inflation, governmental policies, and delay during project execution. Trost and Oberlender (2003) presented a mathematical model for evaluating the accuracy of early estimates using factor analysis and multivariate regression analysis. Touran (2003) proposed a probabilistic model according to Poisson processes for the calculation of project cost contingency. Sonmez (2008) developed an integrated approach to conceptual cost estimation including the advantages of parametric and probabilistic estimating techniques such as regression models. Finally, Shane et al. (2009) categorized individual cost increase factors to assess future total project cost.

Although these models are useful for addressing cost escalation factors and preliminary estimation in the early design phase, some restrictions exist for time-varying variables and for reflecting different time lags between influence factors. Because much time-related data are dependent or have an autocorrelation in reality (Lu and AbouRizk 2009), one way to overcome these limitations is to apply time-related approaches to predicting trends in material prices.

### Time-Series Models for Cost Estimating

In an attempt to solve time-related problems in the aforementioned methods, time-series approaches, which determine future trends of a certain variable on the basis of past values of itself and other escalation factors, have been applied to cost estimation in construction projects. For example, Fellows (1991) provided reliable forecasts of tender prices, building costs, and the effects of inflation on building projects using a time-series model. Williams (1994) illustrated the difficulties of applying neural networks to predict changes in construction cost indexes. Akintoye et al. (1998) identified leading indicators of construction cost oscillations in the United Kingdom using a time-series approach. Ng et al. (2004) outlined the procedures for integrating regression analysis and time-series models to develop a tender price index for Hong Kong construction projects. Wong et al. (2005) applied a time-series projection approach in the labor market of the Hong Kong construction industry using the Box-Jenkins model. Hwang (2009) presented a method for predicting changes in construction costs due to economic conditions in the market using a dynamic regression
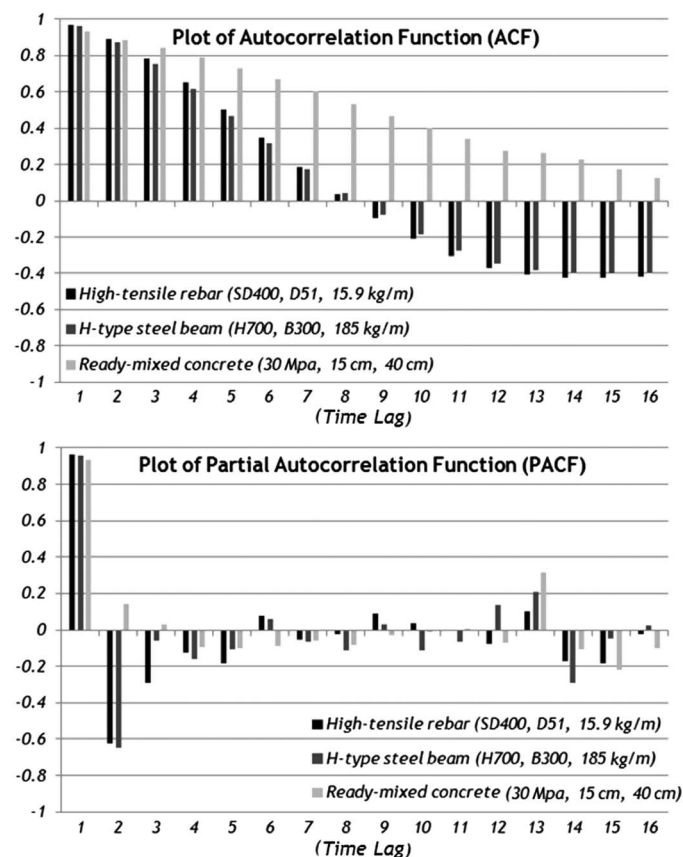


**Fig. 1.** Seasonality analysis of material price data

model. Finally, Ashuri and Lu (2010) developed a time-series model that determines trends based on past values and corresponding errors and provides a more accurate prediction of a construction cost index.

These time-series models provide systematic and time-related approaches to forecasting trends. That is, it is possible to make useful projections based on historical patterns (Wong et al. 2005). However, any research that predicts weighted aggregate indexes of prices or total cost will be limited by the necessary inclusion of detailed information required by construction managers, particularly in recent object-based cost-planning practices, even though it is useful for owners and bidders. Because existing time-series forecasting approaches also involve complicated iterative procedures and arbitrary decisions, a modeler must spend considerable time and effort on forecasting different future trends in numerous types of construction material. Also, as new data are acquired during lengthy project durations, the model must be rebuilt, requiring additional expenditures of time and effort. Lu and AbouRizk (2009) thus tried to develop an integrated, robust, and automated Box-Jenkins modeling tool that could be applied to the capital planning of infrastructure systems.

To make the research in this paper applicable to detailed and updatable material cost estimating, an automated forecasting system is developed on the basis of both the ARIMA modeling process and a simplified forecasting procedure proposed by Lu and Abou Rizk. The suggested ATMF system improves data input and output processes and the process of determining suitable models compared to previous research. Data usability is also supplemented so that the system can be applied to both current data-based management tools (mostly performed in Microsoft Excel-based platforms) and object-based parametric estimating tools (such as building information modeling). Particularly because construction management is performed mostly via data-model-based computing software, data usability and compatibility should be a very important consideration. Using Microsoft Excel 2007 and its Visual Basic for Applications (VBA), this system can be further simplified and made more suitable for related cost-estimating software.

## Time-Series Analysis

### Time-Series Models and the Box-Jenkins Approach

Time-series data sets are a sequence of data points, typically measured at successive times and spaced at uniform time intervals. A time-series model determines trends based on past values and corresponding errors. Since the model requires only the historical data of the forecast variable, it is widely used to develop predictive models (Ashuri and Lu 2010). Among these models, the Box-Jenkins approach—a procedure suggested by Box and Jenkins (1994) for carrying out stochastic time-series modeling—is useful for univariate time-series forecasting. The Box-Jenkins approach is also called the autoregressive integrated moving average (ARIMA) model, which considers the underlying trends, cyclic and seasonal elements, and the particular repetitive continuing patterns exhibited by past data (Ng et al. 2004). The ARIMA model takes into consideration any particular repetitive or continuing pattern exhibited in the historical trend (Fan et al. 2010). The autoregressive (AR) model estimates the stochastic process underlying a time series where its values exhibit a nonzero autocorrelation, while the moving average (MA) model estimates the process where the current value is related to the random errors from previous time periods (Ng et al. 2004). In other words, the AR model is a relational equation between the current value and its past values, and the MA

model represents a relational equation between the current value and its past forecast errors. The ARIMA model, including both the AR and MA models, calculates current value on the basis of both past values and past forecast errors.

### Existing Analysis Procedures of ARIMA Models

The Box-Jenkins approach has three stages: (1) tentative model identification, (2) parameter estimation, and (3) diagnostic checking. Because the ARIMA model is applied to analyze stationary time-series data, first it must be determined whether or not the data are stationary, in terms of both mean and variance (Bowerman et al. 2005). In the tentative model identification stage, therefore, stationary time-series data are judged and nonstationary data are transformed into stationary data by applying normal differencing or logarithmic transformation. Fig. 2 is an example of data transformation using normal differencing. The nonstationary raw price data of high-tensile rebar (SD400, D51, 15.9 kg/m) can be transformed using first- or second-order differencing. Because transformed data show stationarity with regard to their mean, they can be applied to the following modeling stage. Then, to determine tentative ARIMA $(p, d, q)$ models, the order of the AR model $(p)$, the order of the MA model $(q)$, and the order of differencing $(d)$ are determined by analyzing the ACF plot and PACF plot.

In the parameter estimation stage, fitted ARIMA models are identified based on goodness-of-fit (GOF) statistics, and model parameters are estimated. To select the most suitable model, the following GOF criteria can be used: smaller values of root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE), Akaike's information criteria (AIC), and normalized Bayesian information criterion (BIC). These indicate a higher GOF for predicting values; the same is indicated by larger values of normalized R-squared. BIC rules are used to select tentative parameters for $p$ and $q$.

The most suitable model is finally determined by the residual test and an overfitting examination in the diagnostic checking stage. The residual ACF is used to examine the residual series of the tentative models. Residual means the difference between the actual (observed) and fitted values. This means that the closer to zero and the more random the residual, the better the fit (Khosrowshahi and Alani 2003). The Ljung-Box chi-squared test examines the autocorrelation of residuals.

### Required Improvement on Application to Material Cost Forecasting

The ARIMA model described previously analyzes trends in historical data, estimates appropriate values for parameters, checks the model's suitability, and then repeats this process until a best-fitting model is determined (Lu and AbouRizk 2009). In other words, the traditional analysis procedure of the ARIMA model involves a complicated and iterative process with arbitrary decisions made by the modeler, particularly in judging stationarity, in determining parameters of the model, and in selecting a suitable model based on GOF criteria. To solve these problems, Lu and AbouRizk (2009) developed a simplified and integrated analysis tool with the aid of computer technology like MATLAB. Although this research improved existing procedures, problems remained, and so the tool was not applicable to material cost forecasting due to numerous types of material to be forecast and a requirement for detailed information in current object-based estimating practices. Lu and AbouRizk's system was also developed for the capital planning of infrastructure systems for the purpose of combining the forecast method and simulation techniques, as well as enhancing data exchange with other software within the simulation procedure.
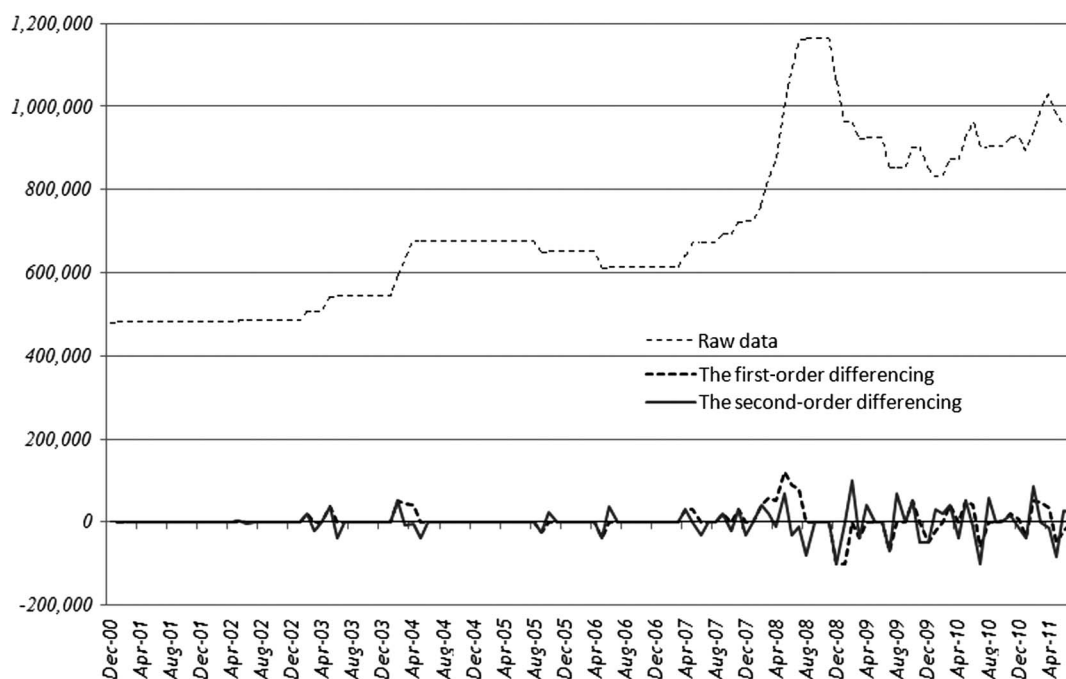
**Fig. 2.** Example of data transformation: Rebar (SD400, D51, 15.9 kg/m) (KPI 2011)

Despite its strength, the model requires a direct data input process, several user choices and decisions within the modeling phases, and an output data transmission process. While the system can effectively be applied in the case of analyzing macrolevel variables such as price index, its application might be time consuming given detail-level variables such as numerous types of material. Also, because forecast price data should be matched with their input time and quantity within the object- and spreadsheet-based database, a spreadsheet-based forecasting system can eliminate the data transmission process.

On the basis of the advantages and disadvantages of previous work, the research in this paper analyzes the required improvement for material cost estimation, as described in Table 1. The proposed ATMF system is developed in an MS Excel environment in order to be compatible with existing estimating software. Excel-based calculation methods, an optimal solution algorithm, a GOF score, and an improved data autoinput method are also suggested to minimize the time, effort, and user choices required to forecast the prices of numerous types of material.

**Table 1.** Improvement of This Research over Existing Methods

| Category | Existing methods | Lu and AbouRizk's system (2009) | ATMF system |
|---|---|---|---|
| Analysis tool | e.g., SPSS, SAS | MATLAB | Microsoft Excel and VBA |
| Data input | Direct input of raw data | Direct input of raw data | Select material names (data input from material unit price database) and forecast duration and time |
| Determination of tentative model | Graphical judgment and unit root test | Input tentative 32 models except logarithmic transformation and seasonal differencing ($p$, $d$, $q = 0$ to 3) | Input all tentative 48 models ($p$, $d$, $q = 0$ to 2, logarithmic transformation) |
| Logarithmic transformation | Graphical judgment and unit root test | Performed by computer program | Included in tentative models |
| Parameter estimation | Calculating by, e.g., SPSS and SAS using various estimation methods | Calculating by programs using maximum likelihood estimation | Calculating by Excel using unconditional least-squares method in optimal solution algorithm |
| Goodness-of-fit test | Modelers' choice among various goodness-of-fit criteria | RMSE | Goodness-of-fit score considering RMSE, MAE AIC, BIC, and R-square |
| Output | Best-fitting model determined by modeler and forecast value | Best-fitting model and forecast value | Best-fitting three model and forecast values |
| Arbitrary decision | Analysis of ACF and PACF, determination of differencing and logarithmic transformation, determination of tentative model, goodness-of-fit test | Determination of differencing and logarithmic transformation | None |
| User choice and run | Numerous choices and runs | Several choices and runs | 3 choices (material name, forecast time, and duration) and 1 run |

## Automated Time-Series Material Cost Forecasting System

This research develops an automated time-series forecasting system for construction materials. The system focuses on usability for construction practitioners, compatibility with existing estimation software, and simplicity in analysis procedures. This system involves a simplified and automated input module using a price database and autoinput of raw time-series data, a parameter estimation module for all tentative models using an optimal solution algorithm, a GOF calculation module using a GOF score, and a diagnostic checking module.

### Model Development

Fig. 3 illustrates the modeling procedures of this paper's ATMF system as compared to an existing method. As shown in this procedure, all tentative models—including the parameters of the order of the AR model ($p$), the order of the MA model ($q$), and the order of differencing ($d$) from 0 to 2, as well as the existence of logarithmic transformation—are incorporated into the system. In other words, the number of combinations of $p$ (0 to 2) and $q$ (0 to 2) is 8 (i.e., combination of $p = 0$ and $q = 0$ is excluded), and each combination can be transformed into three types of the order of differencing (0 to 2) and two types of logarithmic transformation (exist or not). The total number of tentative models is thus calculated to be 48. Goodness-of-fit and test statistics corresponding to each model allow for examinations of suitability and are automatically calculated; this eliminates arbitrary decisions that otherwise would have been made by the modeler.

The ATMF system framework, as described in Fig. 4, shows the detailed modeling procedures. The system is composed of four modules: time-series data input, parameter estimation, GOF calculation, and diagnostic checking.

- Time-series data input module

    To input raw data of material prices to be forecast in this paper's system, a user selects required material names and forecast time and duration, whereas in the previous system raw data had to be typed in by hand. To implement the module, this research firstly material unit prices database. Construction materials are categorized by their type and prices of each material according to date are assigned. Material unit price data from December 2000, including structural steels and ready-mixed concrete, are provided by the Korea Price Information (KPI) Corporation. When a user selects required material names, corresponding row time-series data of prices are extracted from a material price database and then automatically entered into the system. In addition, row time-series data are automatically transformed into five types of data by normal differencing and natural logarithmic transformation: data transformed by first-order differencing, second-order differencing, logarithmic transformation, both first-order differencing and logarithmic transformation, and both second-order differencing and logarithmic transformation. Finally, transformed data are input into the system to be utilized for parameter estimation and diagnostic checking for all 48 tentative models.

- Parameter estimation module

    The parameter estimation module estimates parameters of all tentative ARIMA models and then tests the significance of the parameters. While the traditional method requires
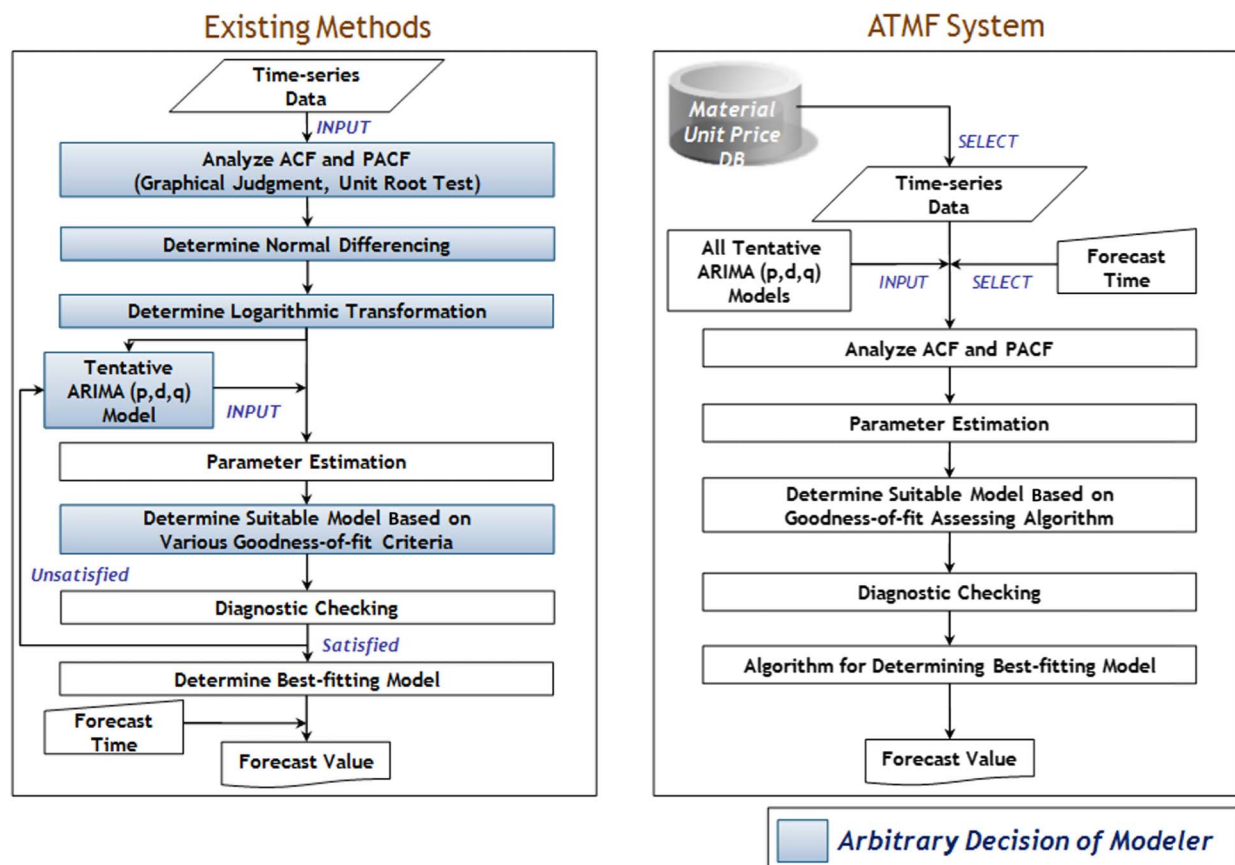


**Fig. 3.** Model determination procedures of ATMF system compared to existing methods
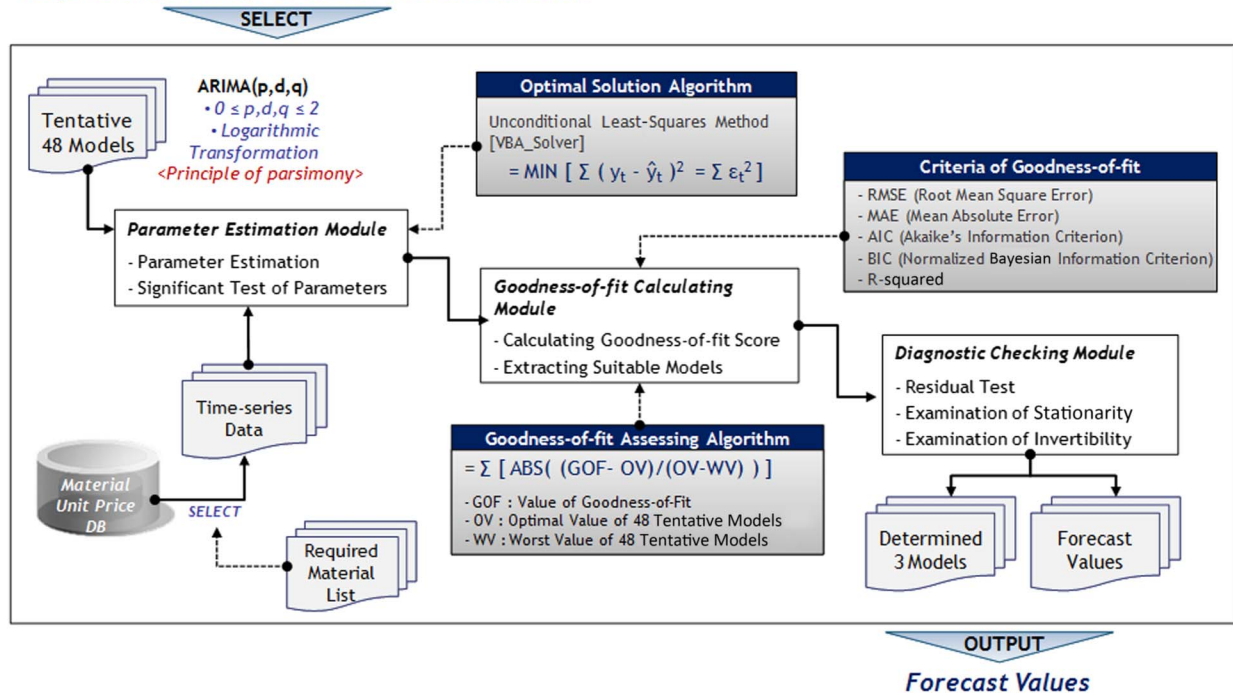
**Fig. 4.** ATMF system framework

significant time and effort to estimate the parameters of even a single model, this new system performs parameter estimation of all tentative models simultaneously, with only one automatic run, by using optimal solution algorithms. The unconditional least-squared method is utilized in this algorithm with the aid of VBA. Using this method, the system automatically calculates optimum values of parameters of the order of $p$ ($\varphi_1$ and $\varphi_2$), the order of $q$ ($\theta_1$ and $\theta_2$) from 0 to 2, and a constant term included in an estimation equation that minimizes the sum of squared residuals (SSR). The whole ATMF system, including the computing codes of the optimal solution algorithm for parameter estimation, can be found on the author's research website (http://blog.naver.com/nkkt14).

- Goodness-of-fit calculating module

The GOF calculating module determines the suitability of all tentative models by using five GOF criteria: root mean square error (RMSE), mean absolute error (MAE), Akaike's information criteria (AIC), normalized Bayesian information criterion (BIC), and R-squared values. These are calculated by the following equations, respectively:

$$\text{RMSE} = \sqrt{\frac{1}{N-k}\sum_{1}^{N}(y_t - \hat{y}_t)^2} \tag{1}$$

$$\text{MAE} = \frac{1}{N}\sum_{1}^{N}|y_t - \hat{y}_t| \tag{2}$$

$$\text{AIC} = \ln(\text{RMSE}^2) + \frac{2(p+q)}{N} \tag{3}$$

$$\text{BIC} = \ln(\text{RMSE}^2) + k\frac{\ln N}{N} \tag{4}$$

$$R^2 = 1 - \frac{\Sigma_1^N(y_t - \hat{y}_t)^2}{\Sigma_1^N(y_t - \bar{y}_t)^2} \tag{5}$$

where $N$ = total number of data points; $y_t$ = actual material price; $\hat{y}_t$ = forecasted material price; $\bar{y}_t$ = mean of actual material prices; and $k$ = total number of estimated parameters. Because each criterion has a different standard of judgment, the greatest GOF of a single criterion cannot guarantee the best predictability of the model. This research therefore suggests a GOF score composed by combining the five criteria. Goodness-of-fit scores are thus calculated using the following equation:

$$\text{GOF Score} = \Sigma[\text{ABS}((\text{GOF} - \text{OV})/(\text{OV} - \text{WV}))] \tag{6}$$

where GOF = value of GOF of corresponding model; OV = optimal value of 48 tentative models; and WV = worst value of 48 tentative models.

- Diagnostic checking module

The diagnostic checking module performs a residual test with a focus on residuals' correlations using the Ljung-Box chi-squared test. The larger the Ljung-Box Q-statistic ($Q^*$) is, the smaller the probability of a type 1 error and the larger the autocorrelation of the residual. $Q^*$, which follows the chi-square distribution, is calculated by the following equations (Bowerman et al. 2005):

$$Q^* = N(N+2)\sum_{k=1}^{K}\frac{\gamma_k^2(\hat{\varepsilon})}{N-k}, \quad \gamma_k = \frac{\Sigma_{t=1}^{N-k}(\hat{\varepsilon}_t - \bar{\varepsilon})(\hat{\varepsilon}_{t+2} - \bar{\varepsilon})}{\Sigma_{t=1}^{N}(\hat{\varepsilon}_t - \bar{\varepsilon})^2} \tag{7}$$

where $N$ = total number of data points; $\hat{\varepsilon}_t$ = forecasted residual; $\bar{\varepsilon}$ = mean of residuals; and $k$ $t$ = total number of estimated parameters. When the probability of $Q^*$ is larger than the level of significance (generally 0.05), the null hypothesis that correlation coefficients of white noises are all zero

[i.e., $H_0 : \rho_1(\varepsilon) = \rho_2(\varepsilon) = \cdots = \rho_k(\varepsilon) = 0$] is adopted. This is because the error term, $e(t)$, should be a white noise process to check whether residual series of tentative models follow the white noise assumption (Ashuri and Lu 2010). The stationarity and invertibility of the parameters are also examined. The stationary condition of the AR(1) process is $|\varphi_1| < 1$, and those of the AR(2) process are $|\varphi_2| < 1$, $\varphi_2 + \varphi_1 < 1$, and $\varphi_2 - \varphi_1 < 1$. The invertibility condition of MA(1) is also $|\theta_1| < 1$, and those of the MA(2) process are $|\theta_2| < 1$, $\theta_2 + \theta_1 < 1$, and $\theta_2 - \theta_1 < 1$ (Bowerman et al. 2005). By performing the residual test, stationarity test, and invertibility test, any overfitting of the model is examined, and inappropriate models are excluded from their tentative counterparts. Among filtered appropriate models, the three best-fitting ones, which have high goodness-of-fit scores, are finally selected for the forecasting model. The forecast values that correspond to each best-fitting model are automatically calculated using the following equation:

$$\hat{Y}_{t+1} = \text{IF}[\text{LT} = \text{''No''}, \text{IF}\{d = 0, \hat{y}_{t+1}, \text{IF}(d = 1, Y_t + \hat{y}_{t+1}, 2Y_t$$
$$- Y_{t-1} + \hat{y}_{t+1})\}, \text{EXP}[\text{IF}\{d = 0, \hat{y}_{t+1}, \text{IF}(d = 1, \text{Ln}(Y_t)$$
$$+ \hat{y}_{t+1}, 2\text{Ln}(Y_t) - \text{Ln}(Y_{t-1}) + \hat{y}_{t+1})\}]] \qquad (8)$$

$$\hat{y}_{t+1} = \hat{\varphi}_1(y_t - \mu) + \hat{\varphi}_2(y_{t-1} - \mu) - \hat{\theta}_1 \varepsilon_t - \hat{\theta}_2 \varepsilon_{t-1} + \mu \qquad (9)$$

where $\hat{Y}_t$ = forecasted values of row data; $\hat{y}_t$ = forecasted values of transformed data $\mu$ a constant term; and LT = logarithmic transformation.

### System Templates

Fig. 5 represents the main templates of the ATMF system. To predict future trends of unit prices of required materials, a user selects corresponding material names, forecast time, and forecast duration in Fig. 5(a). The time-series data of selected materials are automatically input into the left part of Fig. 5(b). When the user clicks the RUN button, three best-fitting models and their GOF scores are determined in Fig. 5(c). Future values predicted by each model are shown in the right part of Fig. 5(b).

The report screen (Fig. 6) presents the modeling process. If the estimated parameter values of AR(1), AR(2), MA(1), MA(2), and a constant term exist, then they are described in Fig. 6(a). Their significant test result [Fig. 6(a)], goodness-of-fit values [Fig. 6(b)], and test results for residual, stationarity, and invertibility [Fig. 6(c)] all correspond to the 48 tentative models also presented.

## System Evaluation

This research conducts a comparative analysis of the prediction capability of three selected models using an out-of-sample forecast. At this stage, 124 data sets of material prices from December 2000 to March 2011 are employed. To perform an out-of-sample forecast, only 112 data sets from December 2000 to March 2010 are included in the modeling phase. The last 12 data sets from April 2010 to March 2011 are used only to analyze the predictive power of the models and to evaluate the system._Three main kinds of construction material—high-tensile rebar (SD400, D51, 15.9 kg/m), H-type steel beam (H700, B300, 185 kg/m), and ready-mixed concrete (30 Mpa, 15 cm, 40 cm)—are selected to evaluate the system. The forecast values and their absolute errors are determined by three best-fitting models (Table 2). The mean absolute percentage
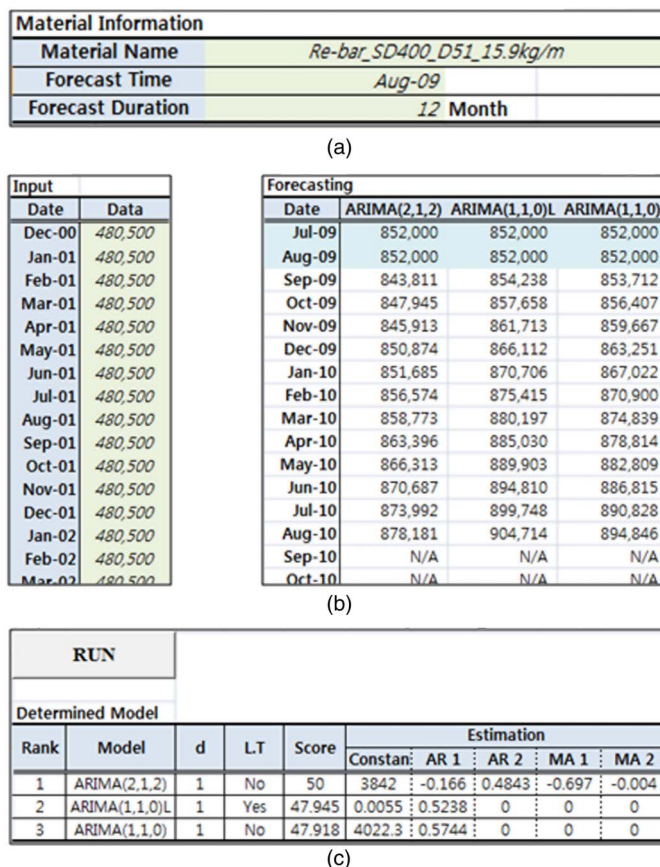


**Fig. 5.** System templates for data input and output: (a) material information; (b) time-series and forecasting data; (c) results of run

error (MAPE) values of the three best-fitting models, with respect to the three materials, are also presented to compare the predictability of the three models. The MAPE is calculated using the following equation:

$$\text{MAPE} = \frac{100}{N} \sum_{1}^{N} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \qquad (10)$$

The generally accepted MAPE of a robust forecast model is 10% (Fan et al. 2010). In this table, the MAPE of the forecast values ranges from 0.97 to 3.50%, and the overall MAPE ranges from 0.52 to 1.85%, which means the model offers significant prediction capability. Also, the higher the goodness-of-fit score a model has, the better prediction capability it has with the smaller MAPE values. However, when there are only slight differences between the GOF scores among determined models, the models' prediction capabilities need to be checked. For example, the ARIMA (1,1,0)L model with rank 2 (GOF score of 45.88) for the material RMC 30_15_40 has a better prediction capability (MAPE value of 0.97) than the ARIMA (0,1,1)L model with rank 1 (GOF score of 45.88 and MAPE value of 0.99). This is the only case where the goodness-of-fit scores among determined models are much the same. For this reason, this system provides the top three scoring models as well as their forecast values, each GOF value, and test results for residual, stationarity, and invertibility. By comprehensively considering all these results as well as conducting predictability test, users can make a final decision according to their preferences (e.g., whether to focus on the predictability or reliability of the determining procedures).

| Model | D | L.T. | Constants | | | AR 1 | AR 2 | MA 1 | MA 2 |
|---|---|---|---|---|---|---|---|---|---|
| | | | Value | t | Result | Value | Value | Value | Value |
| ARIMA(1,1,0) | 1 | No | 4022.27 | 1.34703 | X | 0.57441 | | | |
| ARIMA(2,1,0) | 1 | No | 3931.93 | 1.34703 | X | 0.52514 | 0.08598 | | |
| ARIMA(0,1,1) | 1 | No | 3958.17 | 1.34703 | X | | | -0.4438 | |
| ARIMA(0,1,2) | 1 | No | 3812.09 | 1.34703 | X | | | -0.5538 | -0.2527 |
| ARIMA(1,1,1) | 1 | No | 3968.36 | 1.34703 | X | 0.66757 | | 0.139 | |
| ARIMA(2,1,1) | 1 | No | 3823.41 | 1.34703 | X | -0.1635 | 0.48625 | -0.6943 | |
| ARIMA(1,1,2) | 1 | No | 3825.43 | 1.34703 | X | 0.65266 | | 0.12813 | -0.0223 |
| ARIMA(2,1,2) | 1 | No | 3842.04 | 1.34703 | X | -0.1665 | 0.48425 | -0.6974 | -0.0045 |

(a)

| Model | RMSE | | MAE | | AIC | | Normalized BIC | | R-Square | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Value | Score | Value | Score | Value | Score | Value | Score | Value | Score |
| ARIMA(1,1,0) | 22142.0 | 9.8 | 11487 | 9.87 | 20.03 | 9.07 | 20.0998 | 9.204 | 0.98526 | 9.97 |
| ARIMA(2,1,0) | 22177.2 | 9.8 | 11450 | 9.87 | 20.052 | 9.03 | 20.1476 | 9.122 | 0.98536 | 9.97 |
| ARIMA(0,1,1) | 23522.5 | 9.75 | 12831 | 9.82 | 20.151 | 8.87 | 20.2207 | 8.996 | 0.98336 | 9.97 |
| ARIMA(0,1,2) | 22633.6 | 9.78 | 11968 | 9.85 | 20.093 | 8.96 | 20.1884 | 9.052 | 0.98475 | 9.97 |
| ARIMA(1,1,1) | 22179.7 | 9.8 | 11460 | 9.87 | 20.052 | 9.03 | 20.1478 | 9.121 | 0.98535 | 9.97 |
| ARIMA(2,1,1) | 22221.0 | 9.8 | 11631 | 9.86 | 20.075 | 8.99 | 20.1962 | 9.038 | 0.98544 | 9.98 |
| ARIMA(1,1,2) | 22286.1 | 9.79 | 11420 | 9.87 | 20.081 | 8.98 | 20.2021 | 9.028 | 0.98536 | 9.97 |
| ARIMA(2,1,2) | 16422.8 | 10 | 7923.7 | 10 | 19.49 | 10 | 19.6361 | 10 | 0.99213 | 10 |

(b)

| Model | Residual Test | | | | Stationary test | Invertible test | Total |
|---|---|---|---|---|---|---|---|
| | Q | DoF | Prob. | Result | | | |
| ARIMA(1,1,0) | 18.09035 | 24 | 0.798603 | O | O | O | 47.9 |
| ARIMA(2,1,0) | 15.98033 | 23 | 0.85611 | O | O | O | 47.8 |
| ARIMA(0,1,1) | 36.70379 | 24 | 0.046828 | X | O | O | – |
| ARIMA(0,1,2) | 22.02205 | 23 | 0.518921 | O | O | O | 47.6 |
| ARIMA(1,1,1) | 2.085606 | 23 | 1 | O | O | O | 47.8 |
| ARIMA(2,1,1) | 15.38539 | 22 | 0.845175 | O | O | O | 47.7 |
| ARIMA(1,1,2) | 15.92705 | 22 | 0.81949 | O | O | O | 47.7 |
| ARIMA(2,1,2) | 15.39684 | 21 | 0.802526 | O | O | O | 50.0 |

(c)

**Fig. 6.** System templates for reports of determining procedures: (a) description of estimated parameter values and constant term with significant test reults; (b) goodness-of-fit values; (c) residual, stationarity, and invertibility test results

Table 3 shows the results of comparative analysis between the time-series model determined by the ATMF system and the linear regression model in which time period of the data sets is considered the independent variable. The lower MAPE values of the time-series model show that in terms of predictability the ATMF model outperforms other simple models such as linear regression.

Despite the simplified and automated procedures of this system, its considerable estimating accuracy means it can be effectively applied to forecasting future prices of construction materials. Further, the system has the potential to accurately estimate material costs in a rapidly changing economic environment.

## Application

The proposed ATMF system can forecast future prices of all materials required in construction projects by automated and simplified procedures; thus, it can be easily applied to material cost estimating in object-based estimating environments at a detailed level. Fig. 7 shows the material-cost-forecasting framework after applying trends of material price fluctuations forecast by an ATMF system. For material cost estimation, future values of material prices predicted

by the ATMF system can be useful if combined with other object properties related to scheduling and quantity estimation.

In the object-based model, each object has a unique ID, classified according to the area of the object's location and installation. The ID also includes various properties encompassing types and sizes of the installed materials (length, area, and volume). This information is imported to the quantity-estimating module connected to the scheduling module. Information of activity start times and material stocked times extracted from this model are added to objects as new properties. With respect to the predicted values of material unit prices, these prices' properties are extracted from the suggested time-series system and then added to objects corresponding to activity time periods when each material is installed. Finally, each object includes properties about material types, quantity information, stocked time, and future prices corresponding to the materials' stocked times.

By using extracted properties including material type, quantity information, stocked time, and future prices, the quantities and times of inputs are classified according to corresponding activity start time. Thus monthly (and weekly or quarterly) material quantity inputs can be obtained from the quantity take-off system.

In the cost-estimating module, each monthly quantity is multiplied by forecasted material prices of equivalent periods, yielding

**Table 2.** Forecast Values Determined by Selected Models

| Material | Rebar_SD400_D51 | | | Steel H-Beam_700_300 | | | RMC_30_15_40 | | |
|---|---|---|---|---|---|---|---|---|---|
| Rank | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| Forecast model | ARIMA (2,1,2) | ARIMA (1,1,0)L | ARIMA (1,1,0) | ARIMA (2,1,2) | ARIMA (1,1,0)L | ARIMA (1,1,0) | ARIMA (0,1,1)L | ARIMA (1,1,0)L | ARIMA (1,1,1)L |
| GOF score | 50 | 47.95 | 47.92 | 49.23 | 49.23 | 49.15 | 45.88 | 45.88 | 45.82 |
| **Forecast value (Won) [absolute error (%)]** | | | | | | | | | |
| Apr 2010 | 894,997 (2.64) | 895,363 (2.68) | 895,698 (2.72) | 992,174 (0.22) | 992,542 (0.26) | 991,566 (0.16) | 63,875 (0.15) | 63,858 (0.12) | 63,895 (0.18) |
| May 2010 | 911,507 (1.14) | 909,939 (1.31) | 910,468 (1.25) | 995,440 (4.28) | 996,661 (4.17) | 994,192 (4.40) | 63,972 (0.30) | 63,953 (0.27) | 63,994 (0.34) |
| Jun 2010 | 917,389 (4.64) | 919,993 (4.37) | 920,349 (4.33) | 999,400 (4.82) | 1,001,768 (4.59) | 997,537 (5.00) | 64,070 (0.45) | 64,052 (0.43) | 64,091 (0.49) |
| Jul 2010 | 924,307 (2.47) | 927,752 (2.86) | 927,554 (2.83) | 1,003,693 (0.37) | 1,007,502 (0.75) | 1,001,368 (0.14) | 64,167 (0.61) | 64,151 (0.58) | 64,187 (0.64) |
| Aug 2010 | 927,854 (2.87) | 934,364 (3.59) | 933,292 (3.47) | 1,008,178 (2.88) | 1,013,640 (3.43) | 1,005,529 (2.60) | 64,265 (0.76) | 64,251 (0.74) | 64,284 (0.79) |
| Sep 2010 | 932,628 (3.40) | 940,410 (4.26) | 938,227 (4.02) | 1,012,761 (3.34) | 1,020,045 (4.09) | 1,009,912 (3.05) | 64,363 (0.91) | 64,352 (0.90) | 64,381 (0.94) |
| Oct 2010 | 936,086 (3.55) | 946,186 (4.67) | 942,722 (4.28) | 1,017,398 (0.73) | 1,026,633 (1.65) | 1,014,447 (0.44) | 64,461 (1.07) | 64,452 (1.05) | 64,477 (1.09) |
| Nov 2010 | 940,262 (1.76) | 951,841 (3.01) | 946,977 (2.49) | 1,022,063 (1.72) | 1,033,352 (0.64) | 1,019,084 (2.01) | 64,559 (1.22) | 64,553 (1.21) | 64,574 (1.25) |
| Dec 2010 | 943,859 (1.60) | 957,452 (3.06) | 951,100 (2.38) | 1,026,744 (0.32) | 1,040,168 (0.99) | 1,023,791 (0.60) | 64,657 (1.38) | 64,653 (1.37) | 64,671 (1.40) |
| Jan 2011 | 947,823 (6.02) | 963,056 (7.72) | 955,150 (6.84) | 1,031,433 (2.12) | 1,047,064 (3.67) | 1,028,544 (1.84) | 64,755 (1.53) | 64,754 (1.53) | 64,768 (1.55) |
| Feb 2011 | 951,518 (0.80) | 968,673 (2.61) | 959,161 (1.61) | 1,036,127 (5.81) | 1,054,026 (4.18) | 1,033,329 (6.06) | 64,854 (1.68) | 64,855 (1.69) | 64,866 (1.70) |
| Mar 2011 | 955,394 (3.79) | 974,313 (1.88) | 963,151 (3.01) | 1,040,823 (9.49) | 1,061,048 (7.73) | 1,038,136 (9.73) | 64,953 (1.84) | 64,957 (1.84) | 64,963 (1.86) |
| MAPE (%) | 2.88 | 3.5 | 3.27 | 3.01 | 3.01 | 3.01 | 0.99 | 0.97 | 1.02 |

**Table 3.** Comparative Analysis of Predictive Power with Simple Models

| Material | Rebar_SD400_D51 | | | Steel H-Beam_700_300 | | | RMC_30_15_40 | | |
|---|---|---|---|---|---|---|---|---|---|
| Rank | 1 | 2 | — | 1 | 2 | — | 1 | 2 | — |
| Forecast model | ARIMA (2,1,2) | ARIMA (1,1,0) L | Linear regression $y = \beta t + \alpha$ $\beta = 4720.0$ $\alpha = 404585$ | ARIMA (2,1,2) | ARIMA (1,1,0) L | Linear regression $y = \beta t + \alpha$ $\beta = 6158.9$ $\alpha = 363127$ | ARIMA (0,1,1) L | ARIMA (1,1,0) L | Linear regression $y = \beta t + \alpha$ $\beta = 38.31$ $\alpha = 57487$ |
| GOF score | 50 | 47.95 | — | 49.23 | 49.23 | — | 45.88 | 45.88 | — |
| **MAPE (%)** | | | | | | | | | |
| Total data (Dec 2000–Mar 2011) | 1.22 | 1.73 | 8.75 | 1.47 | 1.83 | 9.78 | 0.52 | 0.54 | 3.56 |
| Forecast data (Apr 2010–Mar 2011) | 2.88 | 3.50 | 5.07 | 3.01 | 3.01 | 5.94 | 0.99 | 0.97 | 2.81 |

JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT © ASCE / NOVEMBER 2012 / **1267**
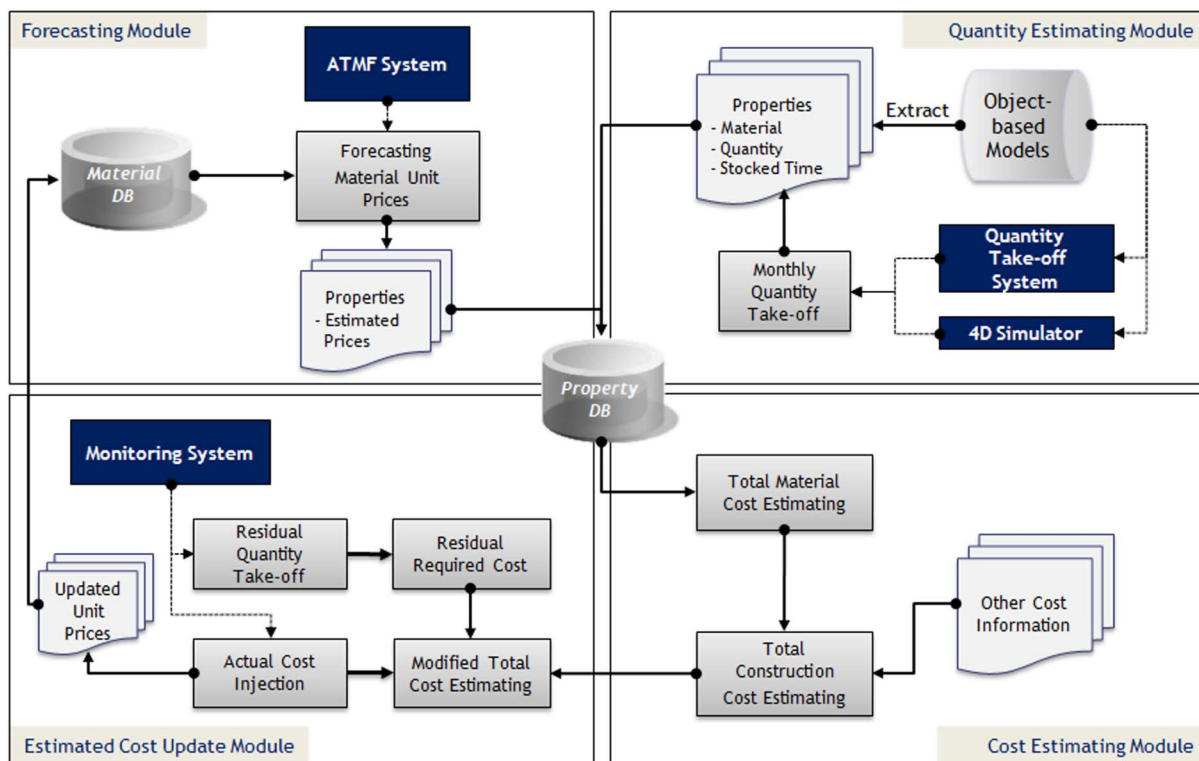
J. Constr. Eng. Manage., 2012, 138(11): 1259-1269

**Fig. 7.** Material cost estimation using ATMF system

total material costs. Total construction costs can also be estimated by applying other cost information such as labor, subcontract, and equipment costs.

On the other hand, new time-series data sets will be updated as the construction progresses. Updating allows for forecasting more accurate estimated values of material prices to modify budgeted costs for residual quantities and work schedules. The cost update module can thus more accurately estimate modified total construction costs by combining the budgeted cost of residual work and the actual cost of work performed. As described previously, the ATMF system can play a salient part in an object-oriented cost-estimating system by offering detailed information on respective material price escalation in a rapidly changing economic environment.

## Conclusions

The research in this paper developed an automated time-series forecasting system for construction materials, including both auto-selected procedures for determining a best-fitting forecasting model and an autoextracting module for forecast values, using Box-Jenkins methodologies and an Excel-based statistical approach. The ATMF system can automatically determine the optimal forecasting model for corresponding materials and forecast trends of unit prices. This system was evaluated through an out-of-sample forecast with respect to time-series price data of three structural construction materials: rebar, steel beam, and ready-mixed concrete. The autodetermination procedures involved in this system can be used to predict future trends in prices of a large number of construction materials by simplifying the forecast modeling process and by eliminating the iterative arbitrary decisions made by the modeler. Construction practitioners can thus minimize the time and effort spent on material cost estimation.

Although this system was evaluated using only material price data in Korea, which do not have seasonality, it can be applied to any geographical region as long as there is no seasonality in the region's material price data. Although it is hard to identify seasonal fluctuations in most regions, further supplementing the seasonality analysis of the ATMF system could enhance applicability worldwide.

With the recent volatile economic conditions, advance preparation for material price escalation can help increase the accuracy of cost estimates by being applied to both current data-based management platforms and object-based cost planning at a detailed level. Therefore, the proposed system can help in dealing with changes in economic conditions and designs by estimating future material prices.

This research attempted to simplify and automate the forecast process for required construction materials. Nevertheless, the developed system can produce large forecast errors if discontinuities occur within the projection time periods. Minor statistical shortcomings of the system, such as a unit root test used for statistically determining stationarity, are acknowledged and should be addressed in future research. Lastly, the developed system can make trends reliable only in the short run. In future research, several attempts should be made to reduce prediction errors by developing additional functions in this system and, further, by building multivariate time-series and intervention models that consider various influence factors and sudden changes in market conditions.

## Acknowledgments

## References

Akintoye, A., Bowen, P., and Hardcastle, C. (1998). "Macro-economic leading indicators of construction contract prices." *Constr. Manage. Econ.*, 16(2),159–175.

Akpan, E. O., and Igwe, O. (2001). "Methodology for determining price variation in project execution." *J. Constr. Eng. Manage.*, 127(5), 367–373.

Ashuri, B., and Lu, J. (2010). "Forecasting ENR construction cost index: A time series analysis approach." *Proc., Construction Research Congress 2010*, ASCE, Reston, VA, 1345–1355.

Bowerman, B. L., O'Connell, R. T., and Koehler, A. B. (2005). *Forecasting, time series, and regression: An applied approach*, 4th Ed., Duxbury Press, Pacific Grove, CA, 401–590.

Box, G., and Jenkins, G. M. (1994). *Time series analysis: Forecasting and control*, 3rd Ed., Prentice Hall, Upper Saddle River, NJ.

Construction Association of Korea (CAK). (2008). "Statistics in production cost of completed construction projects." ⟨http://www.cak.or.kr⟩ (Dec. 21, 2010).

Fan, R., Ng, T., and Wong, J. (2010). "Reliability of the Box-Jenkins model for forecasting construction demand covering times of economic austerity." *Constr. Manage. Econ.*, 28(3), 241–254.

Fellows, R. F. (1991). "Escalation management: Forecasting the effects of inflation on building projects." *Constr. Manage. Econ.*, 9(2), 187–204.

Hwang, S. (2009). "Dynamic regression models for prediction of construction costs." *J. Constr. Eng. Manage.*, 135(5), 360–367.

Hwang, S. (2010). "Cross-validation of short-term productivity forecasting methodologies." *J. Constr. Eng. Manage.*, 136(9), 1037–1046.

Hwang, S., and Liu, L. Y. (2010). "Contemporaneous time series and forecasting methodologies for predicting short-term productivity." *J. Constr. Eng. Manage.*, 136(9), 1047–1055.

Khosrowshahi, F., and Alani, A. M. (2003). "A model for smoothing time-series data in construction." *Constr. Manage. Econ.*, 21(5), 483–494.

Korea Price Information (KPI). (2011). "Construction material prices services." ⟨http://www.kpi.or.kr⟩ (May 31, 2011) (in Korean).

Lu, T., and AbouRizk, S. M. (2009). "Automated Box-Jenkins forecasting modelling." *Autom. Constr.*, 18(5), 547–558.

Ng, S. T., Cheung, S. O., Skitmore, M., and Wong, T. C. (2004). "An integrated regression analysis and time series model for construction tender price index forecasting." *Constr. Manage. Econ.*, 22(5), 483–493.

Ranasinghe, M. (1996). "Total project cost: A simplified model for decision makers." *Constr. Manage. Econ.*, 14(6), 497–505.

Shane, J. S., Molenaar, K. R., Anderson, S., and Schexnayder, C. (2009). "Construction project cost escalation factors." *J. Manage. Eng.*, 25(4), 221–229.

Sonmez, R. (2008). "Parametric range estimating of building costs using regression models and bootstrap." *J. Constr. Eng. Manage.*, 134(12), 1011–1016.

Touran, A. (2003). "Probabilistic model for cost contingency." *J. Constr. Eng. Manage.*, 129(3), 280–284.

Trost, S., and Oberlender, G. (2003). "Predicting accuracy of early cost estimates using factor analysis and multivariate regression." *J. Constr. Eng. Manage.*, 129(2), 198–204.

Williams, T. P. (1994). "Predicting changes in construction cost indexes using neural networks." *J. Constr. Eng. Manage.*, 120(2), 306–320.

Wong, J., Chan, A., and Chiang, Y. (2005). "Time series forecasts of construction labour market in Hong Kong: The Box-Jenkins approach." *Constr. Manage. Econ.*, 23(9), 979–991.

JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT © ASCE / NOVEMBER 2012 / **1269**

J. Constr. Eng. Manage., 2012, 138(11): 1259-1269