Forecasting Pipeline Construction Costs Using Recurrent Neural Networks

Sooin Kim[1]; Bahram Abediniangerabi[2]; and Mohsen Shahandashti, M.ASCE[3]

[1]Doctoral Student, Dept. of Civil Engineering, Univ. of Texas at Arlington, Arlington, TX (corresponding author). Email: sooin.kim@uta.edu
[2]Postdoctoral Research Fellow, Dept. of Civil Engineering, Univ. of Texas at Arlington, Arlington, TX. Email: bahram.abediniangerabi@mavs.uta.edu
[3]Assistant Professor, Dept. of Civil Engineering, Univ. of Texas at Arlington, Arlington, TX. Email: mohsen@uta.edu

ABSTRACT

Pipe material and labor costs comprise about 70% of the total pipeline construction cost. Pipe and labor costs experience volatile fluctuations over time, which mostly cause cost overruns in lengthy and large-scale pipeline projects. The accurate forecasting of pipe and labor costs is critical for cost estimators to prepare accurate bids and manage the cost contingencies. The objective of this research is to develop recurrent neural networks (RNNs) to forecast pipeline construction costs. Pipe material (reinforced concrete pipe, corrugated steel pipe) and labor (common labor, skilled labor) costs from January 1995 to December 2018 were collected from *Engineering News-Record* (ENR) to develop the RNNs. The out-of-sample forecasting accuracies of the RNNs were validated using the ENR pipe and labor cost observations of 12 months in 2019. The results show that the RNNs consistently outperform the seasonal autoregressive integrated moving average (SARIMA) models, which are the most accurate univariate time series model in forecasting pipe and labor cost fluctuations. This research contributes to the pipeline construction community by assisting cost engineers and project managers in enhancing bidding, budgeting, and cost estimating for pipeline projects.

INTRODUCTION

Construction costs experience significant fluctuations over time and result in cost overruns in many construction projects (Shahandashti 2014). Cost fluctuations in construction materials are ranked as the first significant cause of cost overruns in large construction projects (Abdul Rahman et al. 2013). The longer and larger construction projects are more susceptible to cost fluctuations leading to cost overruns (Touran and Lopez 2006).

Pipeline construction projects, which mostly require large-scale and long-term processes, are more prone to experience volatile cost fluctuations (Khodahemmati and Shahandashti 2020). Pipe and labor costs account for 71 percent of a total pipeline construction project cost (Rui et al. 2011). Pipe costs incur an average cost overrun of 5 percent, and labor costs incur an average cost overrun of 22 percent, which are greater than the other cost overrun rates, such as equipment costs (Rui et al. 2012). Therefore, it is critical to accurately forecast future costs of pipe material and labor for successful bidding and budgeting in pipeline construction projects.

Quantitative methods have been implemented to forecast pipeline construction cost fluctuations. Rui et al. (2011) estimated pipeline construction costs, using regression models. Since regression models can misinterpret a spurious correlation between two independent variables, Kim et al. (2020) forecasted pipeline construction costs using univariate time series models. They found

that the seasonal autoregressive integrated moving average (SARIMA) models can most accurately predict the pipe and labor costs among the univariate time series models.

Time series models are one of the most prevalent quantitative methods for forecasting construction costs. Hwang et al. (2012) forecasted construction material costs with ARIMA models. Ashuri and Lu (2010a) concluded that a SARIMA model provides the most accurate forecasts for the ENR construction cost index (CCI). Multivariate time series models, such as vector error correction (VEC) models, have been used to forecast the fluctuations in the national highway construction cost index (NHCCI) and ENR CCI (Shahandashti and Ashuri 2016; Shahandashti and Ashuri 2013). Choi et al. (2020) predicted city-level construction cost index using ARIMA and VEC models. Time series methods have also approximated the fluctuations in construction spending and construction investments (Abediniangerabi et al. 2017, Ahmadi and Shahandashti 2017).

Despite the accurate forecasting abilities of time series models for construction costs, linear time series models, including ARIMA and SARIMA, often fail to forecast volatile fluctuations (Ashuri and Lu 2010b). Supervised machine learning algorithms can provide more accurate forecasts by capturing volatile fluctuations in the historical data, utilizing nonlinear activation functions (Bontempi and Flauder 2015). Shiha et al. (2020) forecasted construction material prices in Egypt with artificial neural networks. Cao et al. (2015) utilized neural networks to forecast Taiwan construction cost index.

The forecasting accuracies of time series models and neural networks have been compared in the literature. Lam and Oshodi (2016) found that neural networks outperform ARIMA in forecasting construction outputs of Hong Kong. Oshodi et al. (2017) reported that neural networks could more accurately forecast the tender price index than ARIMA. Cao and Ashuri (2020) concluded that RNN is more accurate in forecasting highway construction cost index than ARIMA. Even though forecasting volatile fluctuations in pipeline construction costs is necessary for accurate bidding, budgeting, and cost adjusting in pipeline projects, RNNs have not been used to forecast the pipeline construction cost fluctuations.

The objective of this research is to develop recurrent neural networks (RNNs) to forecast pipeline construction costs. This research contributes to improving forecasting accuracies of pipeline construction cost by developing RNNs for forecasting pipe material and labor costs with higher accuracies.

RESEARCH METHODS

Figure 1 presents the research methodology to develop and validate RNNs. The first step is data collection. The monthly data of construction pipe material and labor costs published by ENR are collected from January 1995 to December 2019. The second step is forecasting model development. Statistical procedures are conducted to develop RNNs. The final step is the model validation. The forecasting accuracies of RNNs are evaluated based on three typical error measures: mean absolute percentage errors (MAPE), root mean squared errors (RMSE) and mean absolute errors (MAE).

Data Collection

Pipe and labor cost time series are published monthly by ENR. The ENR material and labor costs are the average costs of the line items in twenty U.S. cities. The ENR costs are widely used

as an average input price for contractors and cost engineers to prepare cost estimates, bids, and budgets in capital projects (Ashuri et al. 2012).

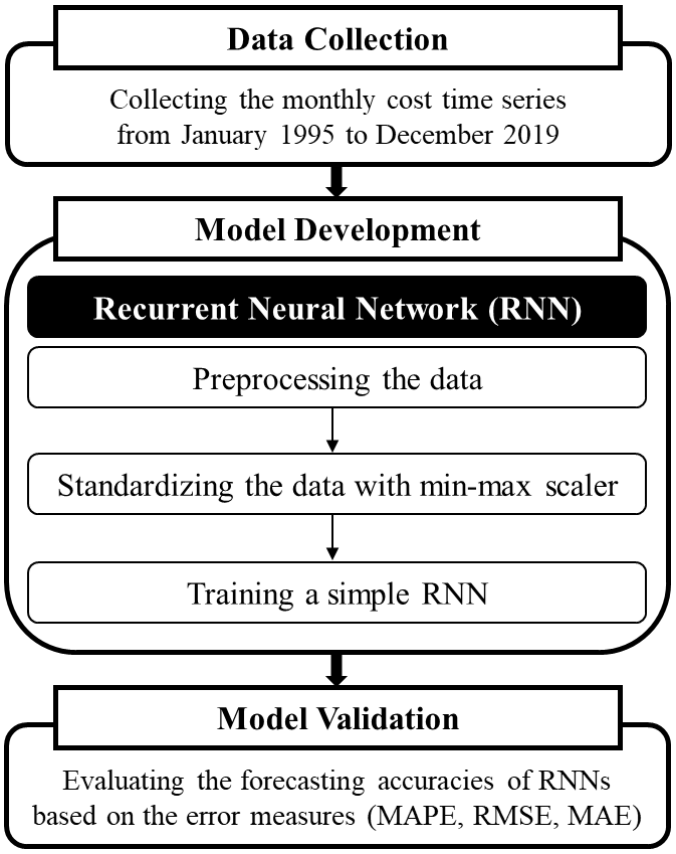

Figure 1. Flowchart of the statistical procedures to develop and validate RNNs

The ENR reinforced concrete pipe costs, corrugated steel pipe costs, common labor costs, and skilled labor costs are collected from January 1995 to December 2019. The collected datasets are split into training and testing datasets. The monthly time series from 1995 to 2018 are utilized as a training set to develop RNNs. Twelve monthly time series in 2019 are utilized as a testing set for one-year out-of-sample forecasting to evaluate the one-year short-term forecasting accuracies of the developed RNNs.

RNN Model Development

A recurrent neural network (RNN) is a feed-back neural network handling a variable-length sequence input with recurrent hidden layers (Chung et al. 2014). An RNN contains special hidden layers composed of recurrently connected neurons (Bandara et al. 2020). These recurrently connected neurons help the network to process the sequential information from inputs and learn long-term dependencies by allowing a memory of previous outputs to persist in the network's internal state (Goodfellow et al. 2016). Since the recurrent connections between neurons in hidden layers reflect a temporal sequence, the RNN can explain the temporal dynamic behavior of time series (Ndikumana et al. 2018). The RNN for reinforced concrete pipe (RCP) costs is expressed by Equation (1).

$$a_t = g_1(W_{aa}a_{t-1} + W_{ax}RCP_t + b_a)$$
$$RCP_{t+1} = g_2(W_{ay}a_t + b_y) \quad (1)$$

where t is the time step; a_t is the activation function at time t; RCP_t is the reinforced concrete pipe cost at time t; W_{aa}, W_{ax}, W_{ay}, b_a, and b_y are coefficients; g_1 and g_2 are activation functions.

The RNN for corrugated steel pipe (CSP) costs is represented by Equation (2).

$$a_t = g_1(W_{aa}a_{t-1} + W_{ax}CSP_t + b_a)$$
$$CSP_{t+1} = g_2(W_{ay}a_t + b_y) \quad (2)$$

where t is the time step; a_t is the activation function at time t; CSP_t is the corrugated steel pipe cost at time t; W_{aa}, W_{ax}, W_{ay}, b_a, and b_y are coefficients; g_1 and g_2 are activation functions.

Preprocessing the data

An RNN is a supervised machine learning model that requires input and target values for model development. The collected ENR cost time series are lagged one-step to generate the input value at the time *t-1* and a target value at time *t*. Then, the one-step lagged datasets are split into training and testing sets with preserving the order of observations. Two hundred twenty-eight observations (96 percent of the observations) are used for training the RNN, and twelve observations in 2019 (4 percent of the observations) are separated for testing.

Standardizing the data with min-max scaler

Since an RNN is sensitive to the scale of input values, standardizing the data to the range of 0 to 1 with a min-max scaler is needed before training an RNN. Standardizing the data with a min-max scaler can improve the speed of model training (Jayalakshmi and Santhakumaran 2011).

Training a simple RNN

A simple RNN is developed and trained by using the training dataset from January 1995 to December 2018 to forecast the cost time series from January 2019 to December 2019. Sequential models with three dense layers are developed for each pipe and labor cost time series. Neurons in a dense layer receive inputs from the neurons in the previous dense layer. The number of neurons in dense layers can be arbitrary and experimentally selected (Brezak et al., 2012). The epoch parameter for the number of iterations is set to 100 to fit the model. The rectified linear unit (ReLu) is used as the activation function. Adam is used as the optimization algorithm to adapt the learning rate for each weight of the recurrent neural network.

Model Validation

The forecasting accuracies of RNNs are compared with the forecasting accuracies of SARIMA models based on the error measures: MAPE, RMSE, and MAE. These error measures are calculated by Equations (3), (4), and (5).

$$MAPE\ (\%) = \left(\frac{1}{N}\sum_{t=1}^{N}\frac{|\widehat{Y}_t - Y_t|}{Y_t}\right) \times 100 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\widehat{Y}_t - Y_t)^2} \quad (4)$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|\widehat{Y}_t - Y_t\right| \quad (5)$$

where $\widehat{Y}_t$ is the forecasted cost by forecasting model at time t, Y_t is the actual cost observation at time t, and N is the total number of forecasted values.

MAPE, RMSE, and MAE measure the forecast errors by comparing the forecasted values with the actual observations. MAPE, RMSE, and MAE have been widely used to evaluate the forecasting accuracies of the models for construction costs (Zhang et al. 2018; Shahandashti and Ashuri 2013).

SARIMA models most accurately forecasted pipe material and labor costs among the univariate time series models (Kim et al. 2020). The SARIMA $(p, d, q)(P, D, Q)_S$ model for forecasting reinforced concrete pipe costs is represented by Equation (6).

$$(1-B)^d(1-B^S)^D RCP_t = \frac{\theta(B)\Theta(B^S)}{\emptyset(B)\Phi(B^S)}Z_t + \mu \quad (6)$$

where B is the backshift operator; d is the differencing order; D is the seasonal differencing order; S is the period of seasonality; μ is the mean of time series; $\phi(B)$ is the autoregressive (AR) operator for non-seasonal components (i.e., $\phi(B) = 1 - \phi_1(B^1) - \cdots - \phi_p(B^p)$); $\Phi(B)$ is the AR operator for seasonal components (i.e., $\Phi(B) = 1 - \Phi_1(B^1) - \cdots - \Phi_P(B^P)$); $\theta(B)$ is the moving average (MA) operator for non-seasonal components (i.e., $\theta(B) = 1 + \theta_1(B^1) + \cdots + \theta_q(B^q)$); $\Theta(B)$ is the MA operator for seasonal components (i.e., $\Theta(B) = 1 + \Theta_1(B^1) + \cdots + \Theta_q(B^q)$); Z_t is the white noise.

EMPIRICAL RESULTS

Forecasting Reinforced Concrete Pipe Costs

A sequential RNN with three dense layers was developed for forecasting reinforced concrete pipe costs. The number of neurons of the RNN was set to ten in the first dense layer, seven in the second dense layer, and one in the last dense layer. Table 1 shows that the RNN has lower forecasting errors for the reinforced concrete pipe costs than the SARIMA model. Figure 2 illustrates the out-of-sample forecasts of the RNN and the SARIMA model and the observed reinforced concrete pipe costs for twelve months in 2019. Figure 2 clearly shows that the RNN more accurately forecasts the volatile fluctuations of the reinforced concrete pipe cost time series than the SARIMA model.

Table 1. Out-of-sample forecasting errors for the reinforced concrete pipe costs

Model	MAPE (%)	RMSE	MAE
RNN	0.05	0.03	0.03
SARIMA	0.4	0.36	0.24

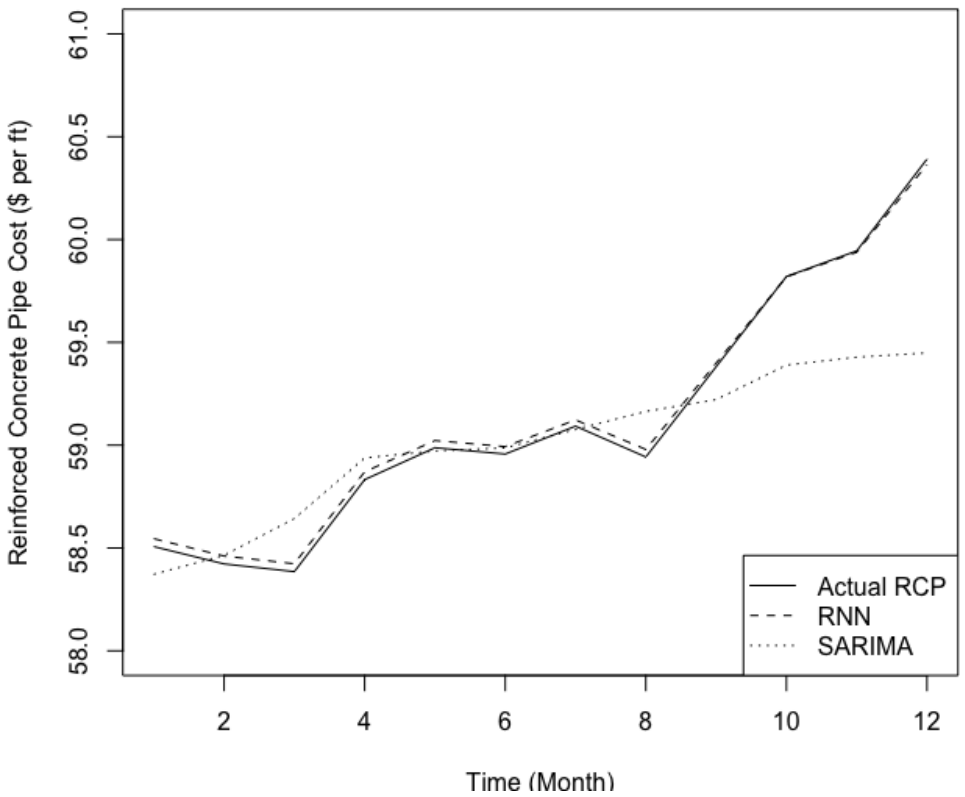

Figure 2. Forecasts by the RNN and the SARIMA model and the observed reinforced concrete pipe costs

Forecasting Corrugated Steel Pipe Costs

An RNN with three dense layers was developed for forecasting corrugated steel pipe costs. The number of neurons of the RNN was set to twenty-four in the first dense layer, twelve in the second dense layer, and one in the last dense layer. Table 2 shows that the RNN has lower forecasting errors for the corrugated steel pipe costs than the SARIMA model. Figure 3 compares the out-of-sample forecasts of the RNN and the SARIMA model and the observed corrugated steel pipe costs for twelve months in 2019. Figure 3 demonstrates that the RNN can more accurately forecast the discrete jumps of the corrugated steel pipe costs than the SARIMA model.

Table 2. Out-of-sample forecasting errors for corrugated steel pipe costs

Model	MAPE (%)	RMSE	MAE
RNN	0.05	0.02	0.02
SARIMA	0.26	0.11	0.11

Forecasting Common Labor Costs

An RNN with three dense layers was developed for forecasting common labor costs. The number of neurons of the RNN was set to twelve in the first dense layer, seven in the second dense layer, and one in the last dense layer. Table 3 shows that the RNN has lower forecasting errors for

the common labor costs than the SARIMA model. Figure 4 plots the out-of-sample forecasts of the RNN and the SARIMA model and the observed common labor costs for twelve months in 2019. Figure 4 demonstrates that the RNN more accurately forecasts the fluctuations as well as the trend of common labor costs than the SARIMA model.

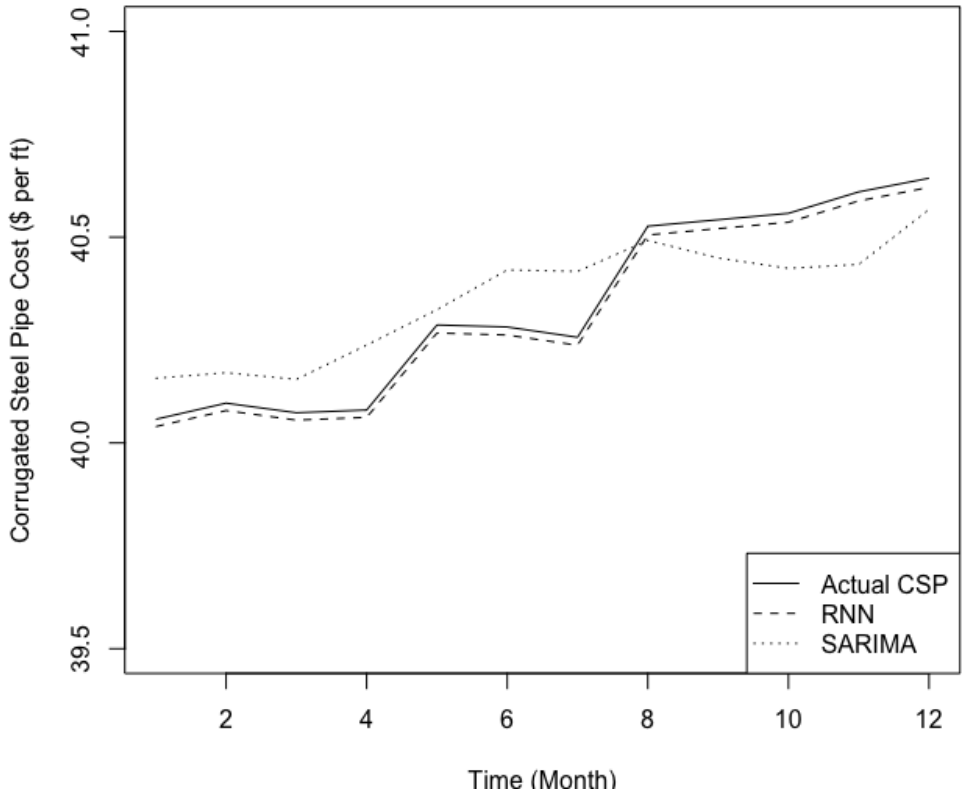

Figure 3. Forecasts by the RNN and the SARIMA model and the observed corrugated steel pipe costs

Table 3. Out-of-sample forecasting errors for common labor costs

Model	MAPE (%)	RMSE	MAE
RNN	0.03	7.94	7.16
SARIMA	0.48	129.83	114.32

Forecasting Skilled Labor Costs

An RNN with three dense layers was developed for forecasting skilled labor costs. The number of neurons of the RNN was set to twenty-four in the first dense layer, twelve in the second dense layer, and one in the last dense layer. Table 4 shows that the RNN outperforms the SARIMA model in forecasting skilled labor costs for twelve months in 2019. Figure 5 illustrates the out-of-sample forecasts of the RNN and the SARIMA model and the observed skilled labor costs for twelve months in 2019. Figure 5 shows that RNN more accurately forecasts the volatile fluctuations as well as the trend of skilled labor costs than the SARIMA model.

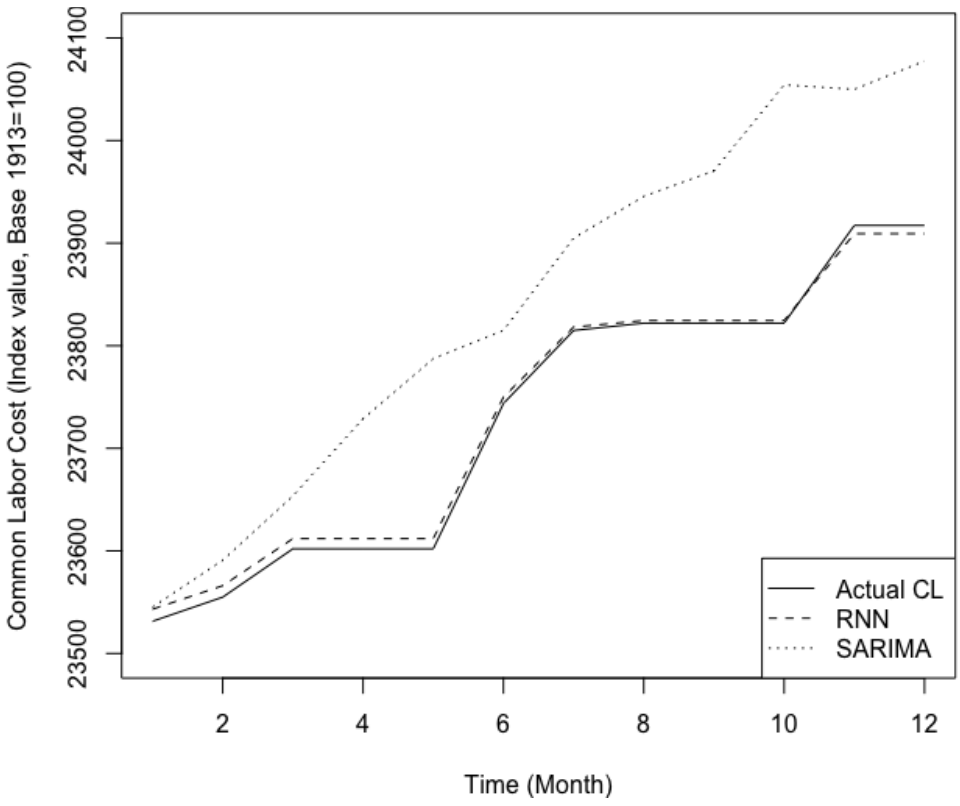

Figure 4. Forecasts by the RNN and the SARIMA model and the observed common labor costs

Table 4. Out-of-sample forecasting errors for skilled labor costs

Model	MAPE (%)	RMSE	MAE
RNN	0.09	9.02	9.02
SARIMA	0.28	34.9	29.1

Based on the out-of-sample forecasting accuracies, RNNs consistently outperform SARIMA models in forecasting pipe material and labor costs for twelve months in 2019. RNNs could more accurately predict the volatile fluctuations and discrete jumps in the pipe material and labor costs than SARIMA models.

CONCLUSION

ENR reports monthly costs of pipe and construction labor in the United States. These cost time series are subject to volatilities over time. These volatilities are problematic for cost estimation and can result in cost overruns in lengthy and large-scale pipeline projects. These volatilities can enlarge the differences between the budgeted cost and the actual cost of a project, especially if the pipeline project requires large amounts of construction resources over time. For example, the forecasting errors of ten cents per foot of the corrugated steel pipe costs can incur over 106 thousand dollars of cost overruns in the twenty-mile pipeline project. Therefore, it is crucial to accurately forecast the cost fluctuations to prevent bid failure, cost overruns, or profit loss in

pipeline projects. It is certainly significant to improve accuracies of forecasting pipe material and labor costs, which account for 71 percent of the total cost in the U.S. pipeline projects on average. This research developed recurrent neural networks to forecast pipe material and labor costs and evaluated the forecasting accuracies of the recurrent neural networks based on mean absolute percentage errors (MAPEs), root-mean-squared errors (RMSEs), and mean absolute errors (MAEs).

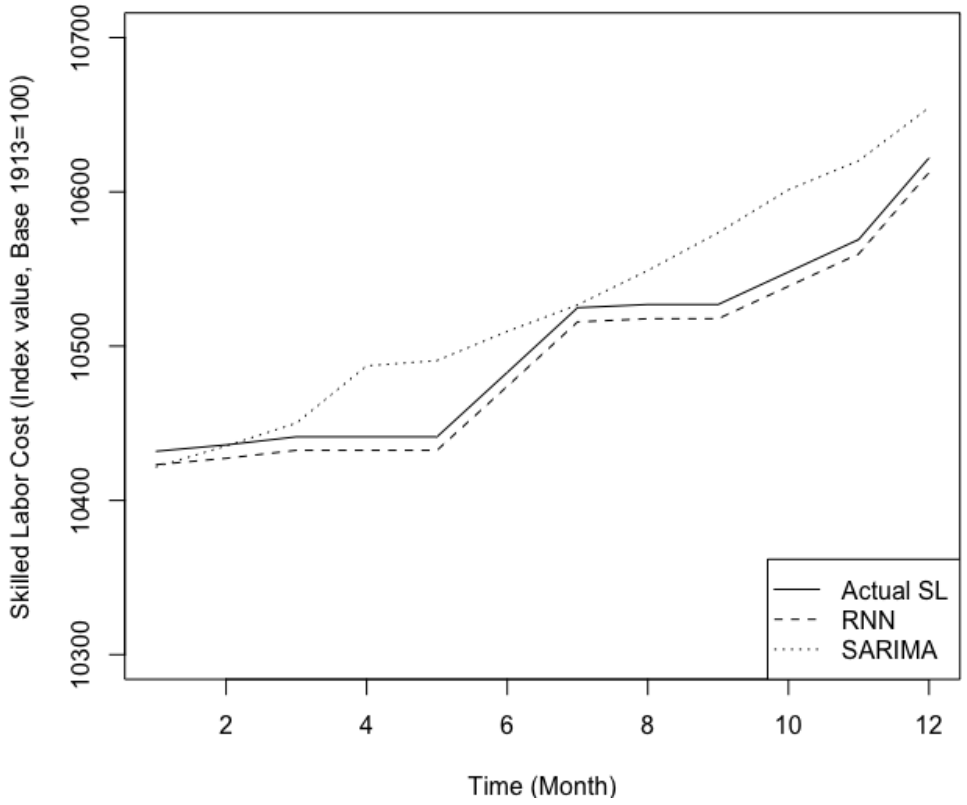

Figure 5. Forecasts by the RNN and the SARIMA model and the observed skilled labor costs

The empirical results of the research showed that recurrent neural networks more accurately forecast the future values of pipe material and labor cost time series than seasonal autoregressive integrated moving average (SARIMA) models. Recurrent neural networks can approximate the volatile fluctuations and discrete jumps in pipe and labor cost time series with higher accuracy, utilizing nonlinear activation functions. Nonlinear activation functions provide recurrent neural networks with higher flexibility to forecast nonlinear patterns in the time series than the seasonal autoregressive integrated moving average (SARIMA) models, which are linear time series models.

The findings of this research contribute to the state of knowledge by developing recurrent neural networks, which can more accurately forecast volatile fluctuations in pipe material and labor costs than seasonal autoregressive integrated moving average (SARIMA) models. Recurrent neural networks can be used for cost estimation and adjustments to prepare more accurate bids and budgets in pipeline projects as well as other construction projects. It is expected that the findings of this research assist cost engineers and project managers to enhance the accuracy of cost estimation in bidding, budgeting, and cost adjusting for future changes in lengthy and large-scale pipeline projects.

ACKNOWLEDGMENT

This research is based upon work supported by the National Science Foundation under Grant No. 1926792.

REFERENCES

Abdul Rahman, I., Memon, A. H., and Abdul Karim, A. T. (2013). Significant factors causing cost overruns in large construction projects in Malaysia. *Journal of Applied Science*, 13(2), 286-293.

Abediniangerabi, B., Shahandashti, S. M., Ahmadi, N., and Ashuri, B. (2017). Empirical investigation of temporal association between architecture billings index and construction spending using time-series methods. *Journal of Construction Engineering and Management,* 143(10), 04017080.

Ahmadi, N., and Shahandashti, M. (2017). Comparative empirical analysis of temporal relationships between construction investment and economic growth in the United States. *Construction Economics and Building*, 17(3), 85-108.

Allouche, E. N., Ariaratnam, S. T., & Lueke, J. S. (2000). Horizontal directional drilling: Profile of an emerging industry. *Journal of Construction Engineering and Management*, 126(1), 68-76.

Ashuri, B., and Lu, J. (2010a). Time series analysis of ENR construction cost index. *Journal of Construction Engineering and Management*, 136(11), 1227-1237.

Ashuri, B., and Lu, J. (2010b). *Exponential smoothing time series models for forecasting ENR construction cost index*. Economics of the Sustainable Built Environment (ESBE) Lab, School of Building Construction, Georgia Institute of Technology.

Ashuri, B., Shahandashti, S. M., and Lu, J. (2012). "Empirical tests for identifying leading indicators of ENR construction cost index." *Construction Management and Economics*, 30(11), 917-927.

Bandara, K., Bergmeir, C., and Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications,* 140, 112896.

Bontempi, G., and Flauder, M. "From dependency to causality: a machine learning approach." *J. Mach. Learn. Res.* 16, no. 1 (2015): 2437-2457.

Brezak, D., Bacek, T., Majetic, D., Kasac, J., and Novakovic, B. (2012, March). A comparison of feed-forward and recurrent neural networks in time series forecasting. In *2012 IEEE Conference on Computational Intelligence for Financial Engineering and Economics (CIFEr)* (pp. 1-6). IEEE.

Cao, M. T., Cheng, M. Y., and Wu, Y. W. (2015). Hybrid computational model for forecasting Taiwan construction cost index. *Journal of Construction Engineering and Management*, 141(4), 04014089.

Cao, Y., and Ashuri, B. (2020). Predicting the Volatility of Highway Construction Cost Index Using Long Short-Term Memory. *Journal of Management in Engineering*, 36(4), 04020020.

Choi, C. Y., Ryu, K. R., and Shahandashti, M. (2020). Predicting City-Level Construction Cost Index Using Linear Forecasting Models. *Journal of Construction Engineering and Management*, 147(2), 04020158.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.

Hwang, S., Park, M., Lee, H. S., and Kim, H. (2012). "An automated time-series cost forecasting system for construction materials." *J. Constr. Eng. Manage.*,, 1259–1269.

Jayalakshmi, T., and Santhakumaran, A. (2011). Statistical normalization and back propagation for classification. *International Journal of Computer Theory and Engineering*, 3(1), 1793-8201.

Khodahemmati, N., and Shahandashti, M. (2020). Diagnosis and Quantification of Postdisaster Construction Material Cost Fluctuations. *Natural Hazards Review*, 21(3), 04020019.

Kim, S., Abediniangerabi, B., and Shahandashti, M. (2020). Forecasting Pipeline Construction Costs Using Time Series Methods. In *Pipelines 2020* (pp. 198-209). Reston, VA: American Society of Civil Engineers.

Lam, K. C., and Oshodi, O. S. (2016). Forecasting construction output: a comparison of artificial neural network and Box-Jenkins model. *Engineering, Construction and Architectural Management*.

Ndikumana, E., Ho Tong Minh, D., Baghdadi, N., Courault, D., and Hossard, L. (2018). Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France. *Remote Sensing*, 10(8), 1217.

Oshodi, O. S., Ejohwomu, O. A., Famakin, I. O., and Cortez, P. (2017). Comparing univariate techniques for tender price index forecasting: Box-Jenkins and neural network model. *Construction Economics and Building*, 17(3), 109-123.

Rui, Z., Metz, P. A., Reynolds, D. B., Chen, G., and Zhou, X. (2011). "Historical pipeline construction cost analysis." *International Journal of Oil, Gas and Coal Technology*, 4(3).

Rui, Z., Metz, P. A., and Chen, G. (2012). "An analysis of inaccuracy in pipeline construction cost estimation." *International Journal of Oil, Gas and Coal Technology*, 5(1).

Shahandashti, S. M., and Ashuri, B. (2016). "Highway construction cost forecasting using vector error correction models." *J. Manage. Eng*. 32 (2): 04015040.

Shahandashti, S. M. (2014). *Analysis of construction cost variations using macroeconomic, energy and construction market variables* (Doctoral dissertation, Georgia Institute of Technology).

Shahandashti, S. M., and Ashuri, B. (2013). Forecasting engineering news-record construction cost index using multivariate time series models. *Journal of Construction Engineering and Management*, 139(9), 1237-1243.

Shiha, A., Dorra, E. M., and Nassar, K. (2020). Neural Networks Model for Prediction of Construction Material Prices in Egypt Using Macroeconomic Indicators. *Journal of Construction Engineering and Management*, 146(3), 04020010.

Touran, A., and Lopez, R. (2006). Modeling cost escalation in large infrastructure projects. *Journal of construction engineering and management*, 132(8), 853-860.

Zhang, R., Ashuri, B., Shyr, Y., and Deng, Y. (2018). Forecasting Construction Cost Index based on visibility graph: A network approach. *Physica A: Statistical Mechanics and Its Applications*, 493, 239-252.