"Anticipating the Unpredictability of Interstate Development Cost File Utilizing Long Momentary Memory

[1]K.Anand, [2]Shruthi V M, [3]M.Swaruparani
[1,2,3]Assistant Professor
[1,2,3] Department of Civil Engineering,
[1,2,3] Ashoka Women's Engineering college

Abstract: For the highway construction business, the highway construction cost index (HCCI) is a composite statistic that shows the overall pricing trend in the industry. The wide range of available indices makes it difficult for governmental agencies to provide precise budget estimates. The index has been predicted several times using quantitative models, however there are still two fundamental issues. Firstly, there are few models that operate well with data that is very variable. Using only stable data to assess a model's predicting capabilities is a waste of time. Having the ability to foresee at diverse periods in time is also critical for a reliable prediction model In the past, Many studies predicted just one index point ahead of time, limiting its applicability in real-world circumstances. LSTM units are used in the encoder and decoder architectures in this work to model and predict the variability of the HCCI. Comparisons were made between the results of a seasonal autoregressive integrated moving average model and data from the Texas Health Care Cost Index. Short-term, medium-term, and long-term forecasts all showed time series models to be useless in predicting the future. Cost engineering and forecasting experts now have the following new insights thanks to this study: Time series models are outperformed in this research by a cost index forecasting approach based on artificial intelligence, especially for volatile cost indexes. It is possible that future researchers might profit from this paper's findings and utilise them as an example. This is the first work in construction management to illustrate how forecasting models function when there is a shape-change in the index.

Authorkeywords: Constructioncostindex;Costprediction;Timeseriesforecasting;Artificialintelligence.

IntroductionandLiteratureReview

When it comes to construction costs, the

Highway Construction Cost Index (HCCI) is an indicator of how much it costs to build roads and highways. Highway construction projects' major line items (materials, labor, and equipment costs) are represented by a unitless price indicator. National and state-level indexes exist in practice. Some states have developed their own state-level index to better reflect local market conditions because the National Highway Cost Index (NHCCI) does not always accurately reflect the market in that area. HCCI can be used in four ways, according to Shrestha et al. (2016), based on a national survey, to measure inflation in the construction industry, gauge the purchasing power of federal and state agencies, and compare market conditions across states or between states that are adjacent to one another. (2) project-specific factors. Users can't keep up with the market's changing trends because of the wide range of prices; this creates a problem for both contractors and owners, such as local or state transportation agencies. There is a lot of research out there trying to find a solution to this problem by using quantitative models to predict and analyse HCCIs.

Time series data analysis and statistical learning methodologies are two of the quantitative tools available for predicting the HCCI's future. There is still a strong preference for time series modelling in this field of research. When used with the ARFIMA model, the ENR construction cost index (CCI) was predicted with an absolute percentage error (MAPE) of 9.5%, according to Moon and colleagues (2018). If you use their methodology, you don't have to distribute the data in a random fashion. Prediction errors of the four univariate time series methods for the ENR CCI were evaluated by Ashuri and Lu (2010a, b). For example, as observed by its developers, the model could not reliably forecast data that varied substantially. Ashuri and Lu (2010a, b) used a similar strategy to anticipate asphalt-cement pricing, and Ilbeigi et al. (2016a, b) followed suit. Rather of using the original index, Ilbeigi et al. (2016b) used a generic autoregressive conditional heteroscedasticity model, which allowed them to connect volatility and residuals rather than ignore it. R2 was greater than 60% for Joukar and Nahmens (2015), which suggests that the model explained 60% of the variance. For multivariate time series analysis, Shahandashti and Ashuri developed methods (2015). A multivariate

time series model built by him while conducting causal study on highway CCI's leading indicators was used to estimate its future (Ashuri et al. 2012). According to the author, the CCI and NHCCI may be predicted using the CPI, PPI, GDP, and money supply, while crude oil prices and average hourly earnings can be used to predict the CCI, using Granger causality analysis. In order to back up his assertion, the predicting accuracy has improved. Building construction costs were studied by Abediniangerabi, et al. (2017) in connection to the Architectural Billing Index (ABI).

ProblemStatement

There are three significant issues with modelling and forecasting the construction index, in addition to the difficulty to anticipate trend change and volatility. Predicting the data required a variety of assumptions, such as those in stationary time series models and in regression approaches, such as the assumption of a Gaussian distribution error. It is impossible to apply the models described here to scenarios when multiple types of data sources are available.

Second, many of the current models were only applicable to low-volatile data, which was a problem for our study. No matter how low a model's error rate, researchers often overlooked two critical questions: whether the dataset (or the issue) necessitated a complex algorithm, and if the method contributed meaningfully to the prediction of the data. The ENR CCI, for example, has 475 data points spanning the period January 1975 to July 2014. A 0.38 percent mean absolute change between successive months is seen in the dataset. On average, the CCI grows or decreases by 0.38 percent as compared to the previous month's number. Two things are revealed as a result of this investigation:

(e.g., fitting a straight line; or using the previous month's value as a forecast for this month's value). Both models are described by the authors as being simple and straightforward. Simple models often have low errors and R2 values around 1, although it's possible that the error measure is overestimated. According to one research, According to MAPE, the ENR CCI had the best out-of-sample prediction error of 1%. At first look, the outcome seems to be good, however the predicted error is substantially larger than the average change ratio (0.38 percent

).Out-of-sample MAPE of 0.18 percent was calculated by the machine learning model, which was an excellent model in terms of accuracy. ” This means that to accurately forecast time series data, you must first examine the variance in your raw data to get a basic idea of which models could work best and what degree of accuracy you can expect from more complicated models.

When making predictions, you should be able to foresee for various time periods. Many earlier studies could only forecast one index point in the future, which limited the usefulness of the findings in practise.

ResearchObjective

Themajorobjectiveofthisresearchistoexplo reasuitablemethodthat can fit highly volatile Cost information. The new system has the potential to be very accurate while yet requiring little time for training. If the new model can predict across a wide range of time periods, the testing must be designed to accommodate a wide range of situations. Time series models such as seasonal ARIMA have been utilised extensively by academics to predict future building cost patterns. ARIMA, for example, was used to estimate the ENR CCI by Ashuri and Lu (2010b). It was discovered in this study that the CCI estimates provided by ENR's own subject matter experts were incorrect It was estimated by using different time series approaches like Holt ES (exponential) and Holt-Winters ES (seasonal) and ARIMA (arithmetic root mean squared. Among the four approaches, ARIMA was determined to produce the best accurate asphalt price estimate. Using the ARIMA approach, Li and Wang (2013) predicted the Tianjin CCI in China using the Tianjin CCI as an example.

ResearchMethodologies

ResearchFramework

Data gathering, model construction, and model validation make up the three key components of the study (Fig. 1). As a measure of error, the model validation use the mean average percent error.

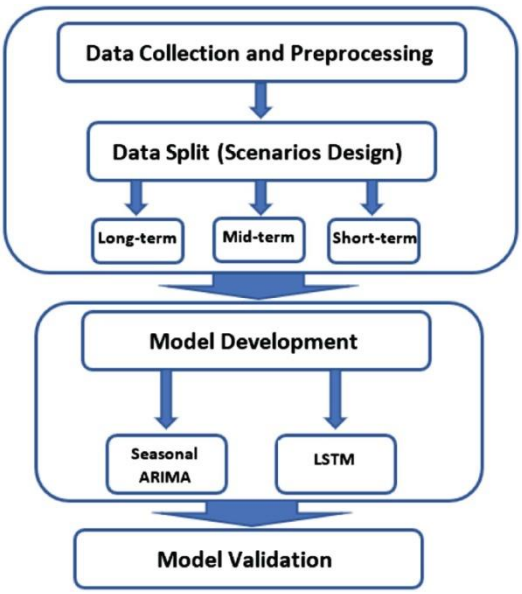

Fig.1.Researchprocedure.

is a complicated model In other words, a bigger dataset should provide greater benefits from the sophisticated model.

Second, this is a classic example of a tough prediction issue characterised by high volatility and frequent trend shifts. The indicator showed a clear shift in trend patterns (Fig. 2). Between 1999 to 2004, there was a period of stability, followed by a two-year period of growth. A three-year era of high volatility preceded the 2008 financial crisis. Five years after a decrease of two years, the index rose significantly, then fell for two years. The prediction model is challenged by this intricate pattern. The change ratio may also be analysed, as stated above. According to the ENR CCI example, the Texas HCCI is far more volatile, with a change ratio of up to 30%. 8.45 percent is the absolute mean change ratio, which is more than 20 times more volatile than the ENR CCI.

LongShort-TermMemory

As a result of its robust structure and rolling-forward forecasting technique across various time horizons, long short-term memory (LSTM) was determined to be an advanced deep learning algorithm that effectively solves these obstacles. Sequence-to-sequence (seq2seq) architecture is LSTM's main selling point.

Although extended short-term memory is a promising and effective prediction tool, only a few researchers have utilised it to solve difficulties in the construction sector at the time of this writing. Long short-term memory (LSTM) and short-term memory (STM) both benefit greatly from LSTM. When compared to other types of recurrent neural networks, this one has an extra-complex unit structure. Scientists Hochreiter and Schmidhuber initially suggested this network in 1997. For neural network training it might be difficult to notice when a gradient disappears, but this structure cleverly addresses that problem.

Work on the LSTM algorithm's implementation was done in Python 3.7. For this application, To get the job done, we used Keras (2.2.0), Numpy (1.15.1), and Scikit-Learn (version 0.20.1). Keras is a high-level neural network package built on top of multiple lower-level frameworks like Tensorflow as an AI model building platform, which was released in 2015. Fast data manipulation and calculation is provided by Tensorflow, which uses a tensor-like computational technique. Using Keras, you can easily call functions and create algorithm structures. If you're more interested in developing tests, implementing the algorithm, and assessing outcomes, Keras is an ideal tool. Keras reduces the amount of time researchers have to spend writing algorithms.

When it comes to numerical computation, Numpy provides a variety of useful matrices and manipulation tools, and it served as the foundation for our study. Data pretreatment was made easier using Scikit-Learn, a popular machine learning software on Github that makes it easy to create real-world applications. Based on Numpy, Scikit-learn is able to do high-performance linear algebra and array operations with ease.

EncoderandDecoderArchitecture(Seq2seq).

The machine learning community was forever altered with the discovery of the LSTM unit, and further study revealed that

various architectures exhibited excellent performance on particular tasks. The seq2seq model, an encoder and decoder architecture, is very helpful in this study. First suggested, NLP architecture called as an LSTM, or "intermediate state," was utilised in the ISP to connect a series of input data to a sequence of output data (also, hidden state and cell state). Since its conception by Google engineers, it has undergone rigorous testing in a number of different languages (Sutskever et al. 2014). It has becoming more common to use deep learning architectures on numerical time series data. On his blog, Wang asserts that the seq2seq structure is particularly good for forecasting time series data (such as climate change). The design can accommodate input and output sequences of varying durations because to its adaptability. Prediction of extreme occurrences and outliers may also be done using Wang's multivariate example of air

pollution levels.(Wang2017).

DataSplitandModelDevelopment

Training and testing sets were created from the dataset. The forecasting method is applied to hyper parameters tuned in the development set, which makes up 10% of the training set. For the following reasons, the training set was selected and consists of data from 1998 to 2008. A smaller training set would lead to overfitting using LSTM, hence the techniques cannot be used. The model was able to reach a high level of accuracy with the smallest amount of data possible, thanks to this sample size. Another factor affecting highway building costs was the 2008 financial crisis, which may be seen in the abnormally high volatility of several indices. From September 2007 to September 2008, Texas' HCCI went from 180 to 240.

It dropped to 170 in September 2009 (29 percent decline) from 33.3 percent in September 2008. An asphalt index index in Kentucky rose from 150 to 380 between January 2007 and July 2008 according to Wangand Liu (2012). (150 percent change). Several studies have shown that estimating the cost of highway building is difficult because of the high degree of unpredictability. The ENR CCI, for example, was the subject of a time series model created by Moon et al. According to Moon et al., the time series model did not perform well in 2008's erratic economic environment. Because the index's value fluctuated so rapidly, the prediction error was rather considerable. The prediction error peaked around the time of the economic crisis in 2008, as observed by Ilbeigi et al. (2017), while using time series analysis to anticipate Georgia's asphalt price index. An effective strategy is needed to generate more accurate projections in the case of significant volatility in the cost index, given the limitations of the current literature. Although it may not be as bad as the 2008 financial crisis, similar occurrences that cause the cost index to fluctuate are common in reality. Research-based forecasting methods should be able to respond to a specific occurrence without the need for human intervention.

The model was trained on the training data. The development set was used to fine-tune hyperparameters, which is important for the LSTM model in particular. It is possible to estimate the parameters of the seasonal ARIMA model ahead of time using a variety of different

ways. This study used k-fold cross validation (CV) to discover the most optimum hyperparameter values for the method. K-fold CV is designed to avoid overfitting by training and testing models on the same dataset. In this study, a 10-fold (k 10) CV was used to construct a reliable forecasting method. Each of the 10 training sets was split into ten equal subsets. A total of ten groups were employed for validation and nine for training. Cross-validation of the procedure was carried out using 10 iterations in which one group was selected at a time.

ResearchResults

Long-TermPrediction

The simplest case is the long-term prediction, which involves just a single round of training and testing. Here is a breakdown of how the dataset was dividedFig.6.

Seasonal ARIMAParameter Selection

Autocorrelation and partial correlation (PACF) statistics are often used to determine the MA and AR terms. In Fig. 7, we can see the results of the ACF and PACF tests.

Only the first lag, the MA order, was related with the ACF figure. First and second delays were correlated, hence the AR order was two using PACF. ' As a result of the split, the seasonality was discovered to be 12. We found that seasonality was not statistically relevant during model training, as predicted from the decomposition picture. The linear regression model. Arima is a linear regression model that makes predictions by using lagged variables and mistakes from earlier runs (Brownlee 2017).ARIMA (p,d,q) is a common abbreviation for ARIMA, which specifies the autoregressive and MA orders, as well as the number of differentiators required to stabilise the data. In the end, we came up with ARIMA (2,2,1). Bayesian information criterion (BIC) is less important than other neighbouring parameter sets, as seen in Figure 8's statistics of in-sample fitting compared to the other coefficients.

LSTMHyperparameterSelection

The following six hyperparameters were taken into account and chosen for this study:

Neurons. The prediction model was built

using a one-layer LSTM algorithm. The LSTM algorithm's neuron structure is more complex than that of a standard neural network. Internally, Because the input data is gradually improved over time, a single neuron may perform the tasks of several LSTM layers (i.e., the out- put of the last layer is used as the input for the next layer). There is a greater risk of overfitting complicated structures if you add more neurons, which will not necessarily enhance predicting accuracy. LSTM was used to solve a difficulty with natural language processing in an experiment by Eckhardt (2018). The accuracy was only improved by 0.2 percent when using a 2-layer LSTM model under ideal circumstances (mea- sured by the predicted number of correct words). It was recommended by Eckhardt that a one-layer LSTM be used to identify the dataset's nonlinear behaviour. The computing time required to train the two-layer LSTM method, according to Eckhardt, is another factor that should be taken into account when deciding the number of layers in the model. For highway construction, training time is an important consideration, as is the model's suitability for use in the field. The number of neurons used in each experiment varied widely. As a result of these Findings

ARIMA and ARFIMA time series models were trained for a better comparisonA number of similarities were found between the LSTM models and the time series models (Figs. 9 and 10). As a result, they were both able to forecast the data's future path but they were unable to discern a diminishing tendency around the year 2016. LSTM outperformed the time series models in spotting the declining trend in 2008 compared to other years. The forecast of LSTM moved in a similar direction to the actual level because of this discrepancy. On the eight-year prediction, the MAPE for the time series models was 42% and 48%, respectively, while the MAPE for Results was 52%.

As can be seen from the time series model, adding a seasonal component to the medium-term forecast improved the out-of-sample accuracy. One explanation is that since the data are volatile, the most cautious forecasting would be to maintain the proper trend and provide an average guess. Adding the seasonality component and more volatile prediction improved accuracy by better capturing variance in the shorter-term forecasts in this experiment.

Figures 12 and 13 show that the most difficult prediction is for the years 2008–2010, due to the quick shift in trend. Even with a bigger inaccuracy than its previous prediction ranges, LSTM was better able to identify this shift. For the forecast from 2008 to 2010, the accuracy of LSTM is much greater than that of time series models. In the second forecasting component, the time series model outperformed the LSTM for the other three prediction periods because they are closer together. Both models included the newly acquired data into their forecasts for the next time period.

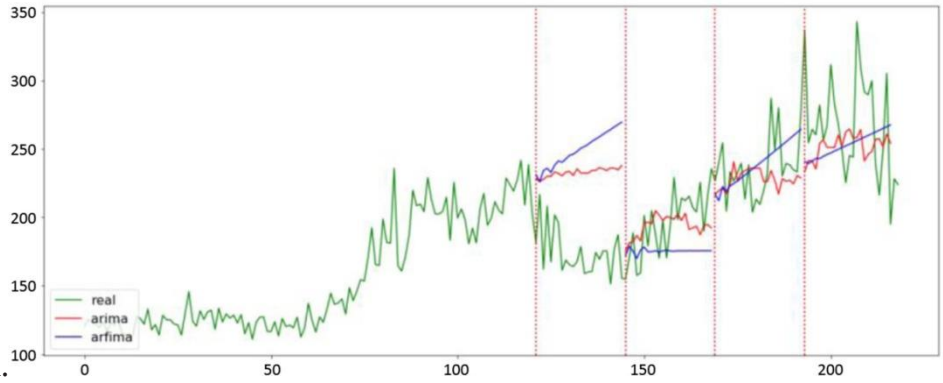

Fig.12.Twoyearsrollingforecastingbyseasonaltimeseriesmodels.

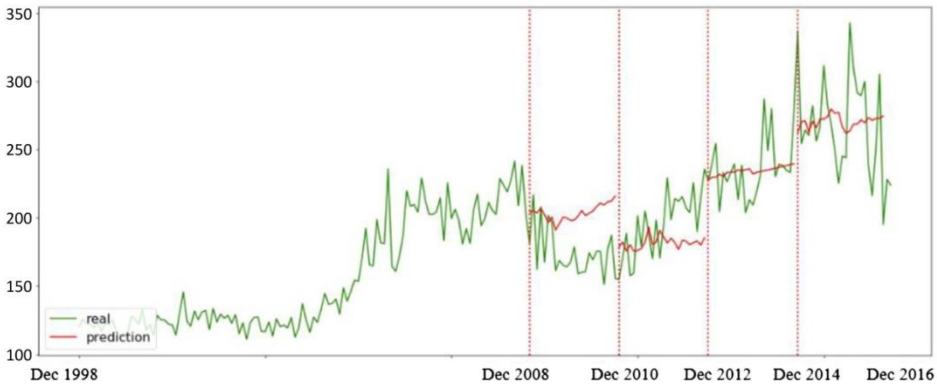

Fig.13.TwoyearsrollingforecastingbyLSTM.

Conclusions

Problems in highway construction cost modelling and index predictions are addressed in this study (which would also apply to other time series data in the construction industry). Additionally, this study mentioned two critical considerations when evaluating a prediction model, one of which is that additional research into a mechanism that can create more accurate predictions is required in the future Consideration #1: The first study of raw data should provide the researcher a rough estimate of how difficult it will be to forecast, and so establish an appropriate expectation for the prediction models. A model's performance may be overestimated if just its error metrics are considered. Another factor is whether or not future information was incorporated in the test procedure for assessing a prediction model's efficacy. As a result, it's possible that the model's efficacy has been exaggerated.

The seq2seq model based on LSTM units was used in this study to predict the HCCI. We utilized ARIMA and ARFIMA as a starting point for evaluating the results. Three different simulations of the time series model and LSTM were run, with forecasts for the short, medium, and long term included in each. While the model's predictive performance was examined, the research also looked at the model's training time to emphasize its potential industrial use. This work introduces an improved artificial intelligence system for cost index forecasting, which delivers more accurate forecasts than current time series models, especially for overly variable cost indices. This study proves that the model development process was done correctly, as well, according to the findings. In the third half of the paper, a novel artificial intelligence technology is employed to predict cost indices. One of the first articles in construction management to demonstrate that forecasting models may be used effectively when a change in index shape is present (which means the index is volatile and hard to predict).

References

1. Ashuri, B., S. M. Shahandashti, and J. Lu. 2012. “Is the information available from historical time series data on economic, energy, and construction mar- ket variables useful to explain variations in ENR construction cost index?” In Proc., 2012 Construction

Research Congress. Reston, VA: ASCE.

2. Bender, E. 2019. “2019 Texas construction industry forecast.” Accessed November 11, 2019. https://www.acppubs.com/articl es/8233-texas
3. -construction-industry-forecast.
4. Brownlee, J. 2017. “How to create an ARIMA model for time series forecasting in Python.” Accessed November 11, 2019. https://machinelearningmastery
5. .com/arima-for-time-series-forecasting-with-python/.
6. Cao, Y., B. Ashuri, and M. Baek. 2018. “Prediction of unit price bids of resurfacing highway projects through ensemble machine learning.” J. Comput. Civ. Eng. 32 (5): 04018043. https://doi.org/10.1061/(ASCE) CP.1943-5487.0000788.
7. Eckhardt, K. 2018. “Choosing the right hyperparameters for a simple LSTM using Keras.” Accessed October 16, 2018. https://towardsdatascience
8. Gransberg, D. D., H. D. Jeong, I. Karaca, and B. Gardner. 2017. Top-down construction cost estimating model using an artificial neural network. No. FHWA/MT-17-007/8227-001. Ames, IA: Iowa State Univ.
9. Guru99. 2019. “What is TensorFlow? Introduction, architecture & exam- ple.” Accessed November 11, 2019. https://www.guru99.com/what-is
10. -tensorflow.html.
11. Huntsman, B., B. Glover, S. Huseynov, T. Wang, and J. Kuzio. 2018. “Highway cost index estimator tool.” Accessed January 15, 2019. https://static.tti.tamu.edu/tti.tam u.edu/documents/PRC-17-73.pdf.