Univariate Modeling of the Timings and Costs of Unknown Future Project Streams: A Case Study

Alireza Shojaei, Ian Flood
M.E. Rinker, Sr. School of Construction Management, University of Florida
Gainesville, Florida, USA.
Email: a.shojaei@ufl.edu, flood@ufl.edu

***Abstract*—Providing a practical and comprehensive methodology to facilitate management and coordination of multiple projects in a company's portfolio is a challenging task. Historically, the focus of research has been limited to the selection and prioritization of the set of known projects, current and near future. It is argued that existing portfolio planning models can be improved by adding a stochastic generator of project streams that extends the portfolio and strategic planning horizon to include future unknown projects. The study both identifies the historical factors in the market that are strong predictors of the profile of future project streams and evaluates alternative modeling approaches to the problem. The outputs from the generator are those parameters most critical to a company, namely the occurrence and letting date of a project, its expected duration, and its expected cost. A case study of design-bid-build highway construction projects let by the Florida Department of Transportation (FDOT) is presented for developing, validating and testing the concept of a project stream generator. The results show that FDOT's future projects can be stochastically forecasted by using historical data and autoregressive moving average modeling along with sampling from representative distributions of cost and durations of FDOT's projects.**

Keywords-Project Portfolio Management; Stochastic Forecasting; Time Series Modeling; Strategic Planning; Uncertainty.

I. INTRODUCTION

Previous work has established the need for future projecting portfolio planning, and proposed some basic modeling approaches to address this issue [1]. This paper advances that work by developing and evaluating the proposed methods of forecasting streams of future work that may be added to a future portfolio.

Construction companies are usually involved in multiple projects at any given time. While different projects progress concurrently, they have different goals and objectives. For instance, some projects may have financial objectives while others may be more focused on marketing or strategic networking. Consequently, a key managerial duty is to allocate resources (such as finances, materials, and personnel) between these concurrently ongoing projects and manage their workflow together to maximize the company's performance [2]. The process of coordinating multiple projects as such is a challenging task because each incoming project affects all other ongoing projects in terms of their schedule and progress [3], and without foreseeing these effects, the consequences can be devastating. The goal of this study is to develop a stochastic project stream generator to forecast unknown future projects in order to extend the horizon of strategic planning for construction companies.

The success of a construction company is strongly impacted by its ability to strategically plan for and manage a stream of projects, many of which will overlap in time, and all of which are subject to uncertainty about their occurrence, scope and resource needs. This task can be broadly classified as Project Portfolio Management (PPM). Cooper et al. [4] describe PPM as "…dealing with the coordination and control of multiple projects pursuing the same strategic goals and competing for the same resources, whereby managers prioritize among projects to achieve strategic benefit." Modern portfolio theory was introduced by Markowitz [5] within the context of finance. McFarlan [6] introduced the concept of PPM in an information technology project management context. He suggested using projects as the elements of a portfolio (instead of investments) to better achieve an organization's objectives as well as reduce the overall risk that the organization encounters during execution of those projects.

Providing a practical and comprehensive methodology to facilitate management and coordination of multiple projects in a company's portfolio is a challenging task. There are no appropriate analytical solutions available for dynamic scheduling and resource allocation of project portfolios in real-time [3]. Existing proposed mathematical models (such as those of [7]–[12]) cannot handle the complexity of real world challenges due to a limited consideration of significant uncertainties within their models and a lack of provision for dynamic and real-time analysis. The primary focus of PPM research was initially to improve organizational performance by introducing good practices to choose and prioritize projects and ensure that the right mix of projects was adopted. A recurring theme is the alignment of the projects with the organization's overall strategy. There is also extensive literature on project selection with a mathematical approach [13]–[16]. In this research, it is not proposed that developed models are incorrect. Instead, it is argued they can be advanced by adding a stochastic project generator to extend the portfolio and strategic planning horizon by forecasting the statistical profile of the stream of unknown future projects. The framework discussed in this paper would allow users to take into account unknown future projects in

their portfolio and strategic planning. From this perspective, it is a novel approach, which by the understanding of the authors has not been done before. Using such an extension would allow the user to plan quantitatively for their portfolio of future projects, as opposed to using a conjecture-based approach.

The rest of this paper is organized as follows. Section II provides a review of the shortcomings of existing PPM models and discusses the impact of uncertainties in PPM. Section III describes the project stream generator and the data used for its development and evaluation. Section IV discusses the modeling approach and results. Section V presents the conclusions and identifies future directions for the research.

II. Project Portfolio Management and Uncertainties in Strategic Planning

Selecting projects from available options and planning and scheduling for them collectively have recently received a considerable amount of attention [17]. For construction related organizations, such as investors, developers, and contractors, it is critical to gather and analyze project information to select the best options according to their strategic goals and schedule them within the required timeframe and financial constraints. This is a complex and multifaceted process, which has many contributing factors, such as the market condition, the organization's structure, resource availability and so on [18]. Research on this topic has come from several different points of view, including selection model criteria and scheduling mechanisms [19], yet the primary focus has been choosing the most appropriate projects rather than providing a real-time dynamic model to address the project selection and scheduling issues [3]. Another shortcoming has been to disregard the importance of multiple project scheduling and resource allocation under influential factors and uncertainties, such as the economic situation of the construction industry and companies' organizational changes. Despite the wide range of available modeling approaches, companies still struggle to optimize and manage changes among their projects [19]. One of the reasons for this is that the proposed mathematical models cannot address the complexity of the real world situation [3]. Excluding uncertainties (such as the impact of possible upcoming projects) or changes in the economic and financial situation of the construction industry are some other noteworthy contributing factors to the poor performance of existing models.

The concept of uncertainty is very significant within the field of project portfolio management. This has led to an extensive literature on uncertainty and the ways to manage it. Duncan [20] and Daft [21] demonstrated that changes in the business environment combined with projects with high complexity always result in an increase in uncertainty in parameters, such as the number of projects, their performance, and their adherence to the project plan. Farshchian and Heravi [22] used agent-based modeling to evaluate time and cost uncertainties related to current projects on a project portfolio.

The impact of uncertainty on organizations is well established across many disciplines from psychology to economics [23]. Environmental uncertainties and their relation to organizations are analogous to the state of a person with a shortage of critical information about the environment. Scott [18] provides an example of the definition of environmental uncertainty as variability or the extent of predictability of the environment where work is executed. They also introduce some measures for uncertainty, such as variability of inputs, the number of deviations in the work process, and the number of changes in the main products. In the project management context, uncertainty in a project is defined as the accuracy of predicting the variation of resource consumption, output, and work process. Uncertainty in a project can be seen as a variation from expected performance of the system under investigation.

The Project Management Institute (PMI) standard for portfolio management despite introducing the risk management concept at a portfolio level does not provide much information on how managers should handle uncertainty and risk within their portfolio. They only provide guidelines on categorizing different possible stages and processes plus naming some of the possible techniques available to handle uncertainties. The PMI only suggests monitoring risks and the performance of the project portfolio under the monitoring and control process group. The proposed framework by the PMI also includes monitoring changes in business strategy. This is an important task because when it occurs, it might result in a complete realignment of the portfolio. The mechanisms involved in this realignment are not specified other than restarting the whole PPM process from the beginning. Also, ad-hoc disturbances to the ongoing and approved project portfolios are almost entirely neglected. This oversight is not because the topic lacks interest or that authors assume a stable and predictable environment. Rather, it can probably be explained by the fact that the subject of PPM is relatively young and that the researchers and academics preferred to focus on more pressing issues in this area. For many companies, the environment is unstable, and the high level of uncertainty and unknowns resulting from the dynamic environment lead to some challenges. Upcoming projects significantly affect the performance of a project portfolio [3]. The typical approach when a new project is added to the portfolio is to update the project portfolio's plans and to try to re-optimize everything.

III. Unkown Future Project Stream Generator

This paper presents an approach to statistically represent unknown future projects to extend the portfolio and strategic planning horizon. Forecasting a company's unknown future projects can be based on the company's past and current portfolio data, or it can use historical data from market to forecast all the upcoming projects as project streams and filter those by bidding success models. In an environment, where the supply of the projects is scarce and very competitive, using just the company's past projects to forecast the future unknown projects is potentially less

accurate. Arguably it is more valid to forecast streams of unknown projects (all the available projects in the future) considering the uncertainties in the context and filter those projects by bidding success models to get the final future projects in a company's portfolio. The forecast can statistically generate a single set of outputs or stochastically produce streams of values as output. Considering the uncertainties in the market, the PPM context, and the availability of future projects, stochastic forecasting appears to be the right choice.

This paper reports on the development, validation, and testing of a project stream generator for design-bid-build highway construction projects let by the Florida Department of Transportation (FDOT). The primary data for this study were obtained from FDOT's historical project lettings database covering 14 years (from 2003 to 2017). The last two years (2015 and 2016) data are withheld to be used as a validation set for the final model and not being used in this study to be used after more models are tested for final verification without any kind of bias. Thus, the model training and selection are based on the data from 2003 to 2015, which contains 2,816 design-bid-build project-letting reports. The outputs from the generator are those parameters most critical to a company, namely the occurrence and letting date of a project, its expected duration, and its expected cost. Other factors, such as economic condition can have an impact on the project stream. Table I shows a pool of candidate variables containing 24 potentially relevant predictors including the macroeconomics metrics and construction indices that were compiled from the related sources and literature [24]–[26] that can be used in multivariate modeling.

The data should be split into three sections as a training set, a test and model selection set, and a final validation set for the final model. In this process, different models are trained and tested using the cross-validation method and the best model is validated with the withheld data. The final validation set is the data from 2015 and 2016, and the data from 2003 to 2015 is used for training and testing of different models to find the best performing model and optimize its corresponding values.

The data under study is a time series and so the continuity of the data is important and should not be tampered with by randomly dividing into different sections for validation. As a result, a rolling forecast origin and a rolling window method is used to cross validate the models' performance to avoid overfitting and overestimation of performance. The rolling window method has a fixed window (Figure 1-A), where the training (orange bar) and test (blue bar) sets duration is fixed and rolls through time. In this research, the trainig and testing set were chosen to be three years each and roll one year in each trial. The rolling forecast method on the contrary uses progressive length of data as the training set (as shown in Figure 1-B) in each trial. The initial training set was chosen to be three years and increase one year in each trial while the test set remain three years of consecutive data after the end of the training set for each trial. Using both methods can help better understand the model's performance and give more insights into the characteristics of different time spans of the data.

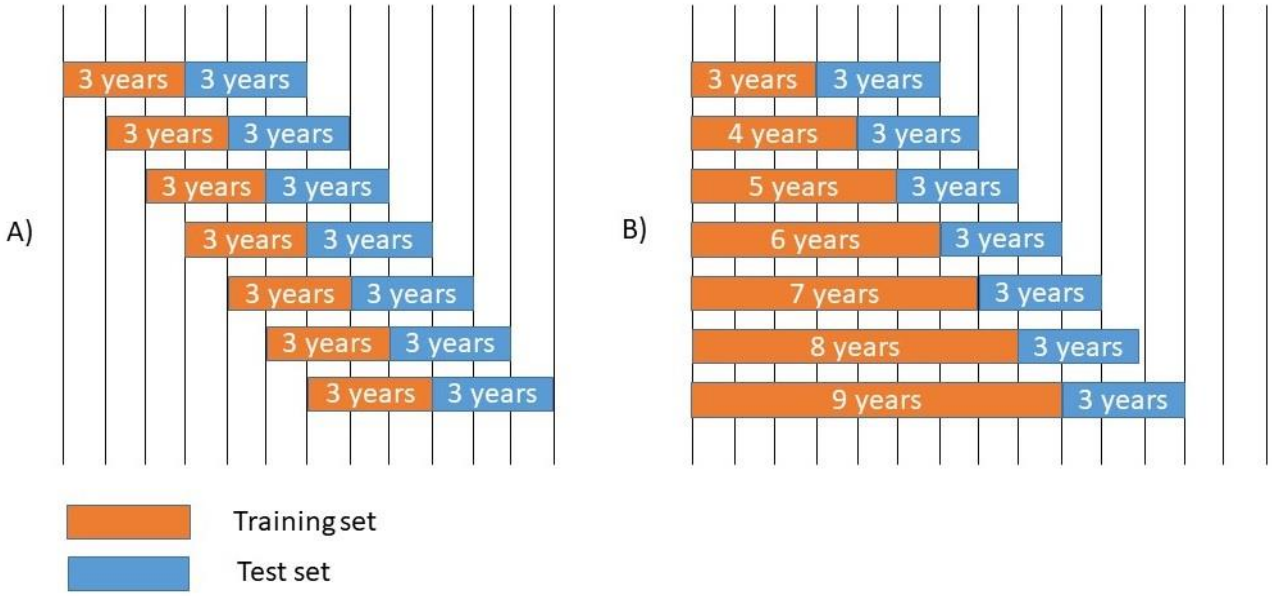

Figure 1. Visual represenation of cross-validation methods used. A) Evaluation based on a fixed window rolling forecast B) Evaluation based on an increasing window rolling forecast

TABLE I. Potentially Relevant Predictors.

CANDIDATE VARIABLES	SOURCE
GROSS DOMESTIC PRODUCTS (GDP)	U.S. Bureau of Economic Analysis
GDP IMPLICIT PRICE DEFLATOR	U.S. Bureau of Economic Analysis
INFLATION RATE	World Bank
CONSUMER PRICE INDEX	U.S. Bureau of Labor Statistics
NATIONAL HIGHWAY COST INDEX (NHCCI)	U.S. Department of Transportation
FDOT'S ANNUAL BUDGET	Florida Department of Transportation
FDOT'S PRODUCT BUDGET	Florida Department of Transportation
FEDERAL FUNDS RATE	Federal Reserve Systems
UNEMPLOYMENT RATE	U.S. Bureau of Labor Statistics
FLORIDA UNEMPLOYMENT RATE	U.S. Bureau of Labor Statistics
NUMBER OF EMPLOYEES IN CONSTRUCTION	U.S. Bureau of Labor Statistics
NUMBER OF EMPLOYEES IN CONSTRUCTION IN FL	U.S. Bureau of Labor Statistics
AVERAGE WEEKLY HOURS	U.S. Bureau of Labor Statistics
PRIME LOAN RATE	Federal Reserve System
BUILDING PERMITS	U.S. Bureau of Census
MONEY SUPPLY	Federal Reserve System
AVERAGE HOURLY EARNINGS	U.S. Bureau of Labor Statistics
EMPLOYMENT COST INDEX (ECI) CIVILIAN	U.S. Bureau of Labor Statistics
DOW JONES INDUSTRIAL AVERAGE	Yahoo Finance
CRUDE OIL PRICE	U.S. Energy Information Administration
BRENT OIL PRICE	U.S. Energy Information Administration
PRODUCER PRICE INDEX	U.S. Bureau of Labor Statistics
HOUSINGS STARTS	U.S. Bureau of Census

The sequence of generating information in the proposed model starts with forecasting the number of projects (project frequency) for the desired time span, using the optimal model based on the training and validation from historical data. This is followed by sampling the project costs from the cost distribution. At each point in time, the number of samples from the distribution is based on the number of projects forecasted in the previous step. Finally, the project durations are sampled from the duration distribution. One important issue is the relationship between these main variables in this process. No logical relationship can be established between duration of the projects and the frequency of the projects. However, the frequency of the projects and their accumulated cost has a high correlation, which can be used in the modeling process. Figure 2 represents the four possible ways that this relationship can be accounted for. One option is using a unidirectional assumption to use cost as an exogenous variable to forecast frequency (number one) along with other variables or vise versa (number two). The third option is to use a recursive model and test for convergence of the values. The last option is to neglect this correlation and assume that it is captured through the individual forecasting of each variable. Another important relationship is the possible correlation between cost and duration, which should be considered in the sampling process from their representing distributions. This could be done by using an empirical copula to build a multivariate probability distribution. As a result, the two variables are assigned simultaneously in each round of sampling with the correlation incorporated in the values.

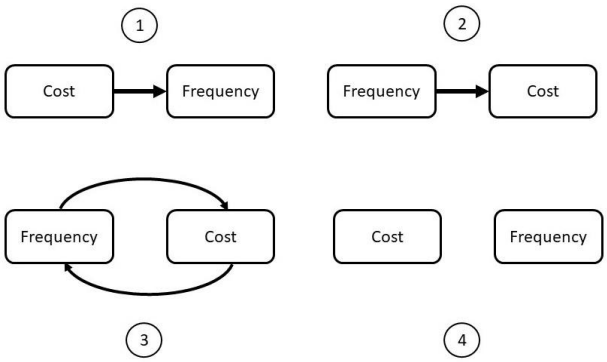

Figure 2. Possible strategies to address the relationship between cost and frequency of the projects. 1) Use cost to forecast frequency 2) Use frequency to forecast cost 3) Recursively use cost and frequency to forecast each other 4) Ignoring the relationship of the cost and frequency in the model

In general, the generator could be implemented as a univariate or multivariate model, and with linear or nonlinear relationships between the inputs and outputs of each model. The complete set of results from the proposed framework can be used as a component in any PPM model to consider unknown future projects in strategic planning.

IV. MODELING APPROACH

Different approaches have been used for time series modeling. Cargnoni et al. [27] used Gaussian models to forecast the number of high-school students in each grade in future school years in the Italian school system. Voyant et al. [28] employed a multilayer perceptron to forecast global solar radiation. Li and Chen [29] used a LASSO (Least Absolute Shrinkage and Selection Operator) based regression to estimate macroeconomic time series, and they demonstrated how this method could be combined with a dynamic factor model to yield a more accurate forecast performance. Exterkate et al. [30] used kernel ridge regression as a multivariate model for economic time-series forecasting by considering the nonlinear relationships among the variables. They found that this method outperformed traditional time-series forecasting techniques based on principal components. Yu and Liong [31] compared the linear ridge regression, ARIMA (Autoregressive Integrated Moving Average), naïve, inverse approach, and support vector machine in forecasting hydrologic time series and concluded that the ridge linear regression outperformed the other models in terms of both performance and time of execution. Choubin et al. [32] compared multiple linear regression, a multilayer perceptron neural network, and an adaptive neuro-fuzzy inference system for forecasting precipitation and concluded that the multilayer perceptron neural network outperformed the other methods. Cao and Tay [33] used a support vector machine for financial time-series forecasting and compared it with a multilayer back-propagation neural network and a regularized radial basis function neural network. They concluded that the support vector machine outperformed the back-propagation neural network and produced a performance similar to that of the regularized radial basis function neural network. The review

of literature shows that dependent on the problem and data, different models perform better. As a result, a set of different models using a systematic approached should be tested to make sure an appropriate model is used for the final forecast.

The scheme used to develop the model is shown in Figure 3. The purpose of this scheme is to look for characteristics of data, to capture them in the model's projections, and then to check to see if the model reproduces them by using the cross-validation tests discussed above. The univariate model, being the simplest, was adopted as a benchmark against which the more complex multivariate models could be compared in terms of forecast accuracy.

The first step is modeling the main variables through univariate modeling methods, such as Autoregressive (AR), Moving Averages (MA), Autoregressive Moving Average (ARMA), and exponential smoothing. More sophisticated approaches such as artificial neural networks can also be implemented considering the availability of the necessary data size to properly the train neural network. After establishing a benchmark, potentially relevant predictors were identified to populate a pool of candidate independent variables based on a literature review and cognitive theories. This introduces the environmental uncertainties to the forecast with the aim of improving the accuracy of the simulation. These variables are not going to have necessarily a causal relationship with the main variables; the only concern here is to be helpful in forecasting the dependent variable.

The next step is exploratory data analysis. It starts with a graphical comparison of the independent and dependent variables, such as scatterplots of pairs of variables. Pearson correlation, unit root (stationary or non-stationary test), Granger causality (helpful for short term forecasting), and cointegration (helpful for long term forecasting) tests are among diagnosis techniques that are relevant.

The last step is to choose a set of multivariate modeling approaches based on the result of the exploratory data analysis and test whether including explanatory variables and models that are more complex can improve the accuracy of the forecast. The range of the models should test for linear and non-linear relationships based on the result of the previous step along with variable selection (pruning), parameter optimization and finding the appropriate lag between variables.

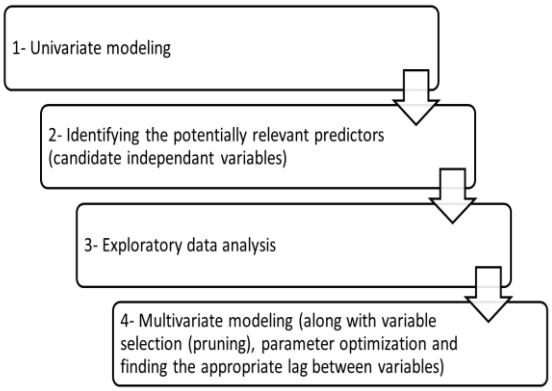

Figure 3. Model development scheme.

Models concerning time series data frequently involve using the value from one or more previous time steps to forecast values at succeeding points in time; in other words, they regress based on past values. In conventional modeling, the assumption is that the independent values are known, and the dependent values are forecast. However, in multivariate time series forecasting, even the independent variables' values in the future are unknown and need to be forecast. As a result, the model contains a system of equations that forecast both independent and dependent variables in the future. This system is recursive when all the causal relationships are unidirectional and non-recursive (simultaneous) when there is reciprocal causation between variables.

Figure 4 shows four of the possible internal structures of the model. Figure 4-A shows the dependencies between the inputs and output in a univariate AR model with a lag of two. In this example, the forecast value at each point in time is based on the two preceding past values. Equation (1) shows the mathematical relationship in such a model, where each value in time is calculated with a linear combination of the past two values plus a constant term (β0) and a white noise term (ϵt).

$$Y_t=\beta_0+\beta_1 Y_{t-1}+\beta_2 Y_{t-2}+\epsilon_t. \quad (1)$$

Figure 4-B shows a recursive multivariate model, where the dependent variable forecast is based on past values of itself and the independent variables. However, each independent variable is only based on its past values. Equation set (2) shows the relationship of such a model with only one lag. It should be noted that in practice there can be many more lags involved in this model and the next two models.

$$Y_t=\beta_{10}+\beta_{11}Y_{t-1}+\beta_{12}\,X_{t-1}+\beta_{13}\,X'_{t-1}+\epsilon_{t,1}. \quad (2)$$
$$X_t=\beta_{20}+\beta_{21}X_{t-1}+\epsilon_{t,2}.$$
$$X'_t=\beta_{30}+\beta_{31}X'_{t-1}+\epsilon_{t,3}.$$

Figure 4-C shows another recursive model, which differs from model 4-B in that the independent variables also act as input to each other.

$$Y_t=\beta_{10}+\beta_{11}Y_{t-1}+\beta_{12}\,X_{t-1}+\beta_{13}\,X'_{t-1}+\epsilon_{t,1}. \quad (3)$$
$$X_t=\beta_{20}+\beta_{21}X_{t-1}+\beta_{22}\,X'_{t-1}+\epsilon_{t,2}.$$
$$X'_t=\beta_{30}+\beta_{31}X_{t-1}+\beta_{32}X'_{t-1}+\epsilon_{t,3}.$$

Figure 4-D shows a sample of a non-recursive (simultaneous) model, where all the variables work as inputs for each other. There is no discrimination between dependent and independent variables in this approach. In this case each variable is a function of its past values and other variables past values.

$$Y_t=\beta_{10}+\beta_{11}Y_{t-1}+\beta_{12}\,X_{t-1}+\beta_{13}\,X'_{t-1}+\epsilon_{t,1}. \quad (4)$$
$$X_t=\beta_{20}+\beta_{21}Y_{t-1}+\beta_{22}X_{t-1}+\beta_{23}X'_{t-1}+\epsilon_{t,2}.$$
$$X'_t=\beta_{30}+\beta_{31}Y_{t-1}+\beta_{32}X_{t-1}+\beta_{33}X'_{t-1}+\epsilon_{t,3}.$$

Figure 5 shows a summary of the possible univariate models considered in this study. At a high level in this classification are two options, neural networks and time series modeling, each with its own set of variations. A long-short term memory is used as the neural network model and AR, MA, ARMA, and different smoothing methods are used as the time series methods.

Figure 6 shows the possible multivariate models that can be used in similar studies. The top categories here are regression, neural networks, time series, and nonlinear autoregressive moving average with exogenous variables (NARMAX), which is a combination of neural networks and time series models.

After training and validating different models, some diagnostic tests should be conducted to check the stability of the best performing model before its implementation. For instance, checking to see if there is an autocorrelation between the residuals of the forecast is an appropriate tool for time series forecasts. Also, checking the way error compounds and undertaking a sensitivity analysis to see how the values of model parameters affect the model's output can give a deeper insight into the performance of the model.

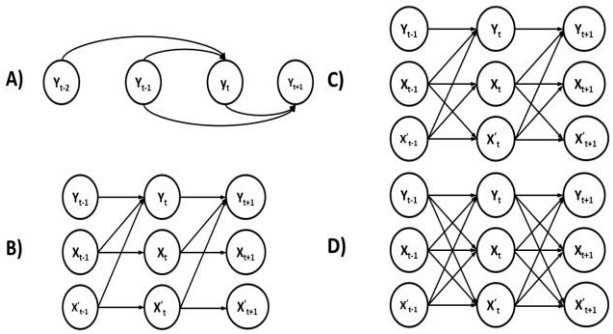

Figure 4. Possible internal structures of the model, illustrating the relationship between the dependent and independent variables. A) Sample illustration of a univariate model B) Sample illustration of recursive multivariate model type 1 C) Sample illustration of recursive multivariate model type 2 D) Sample illustration of recursive multivariate model type 3

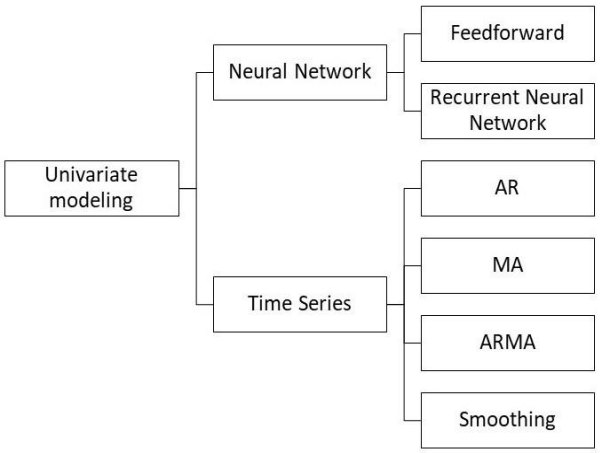

Figure 5. Univariate models surveyed through this study.

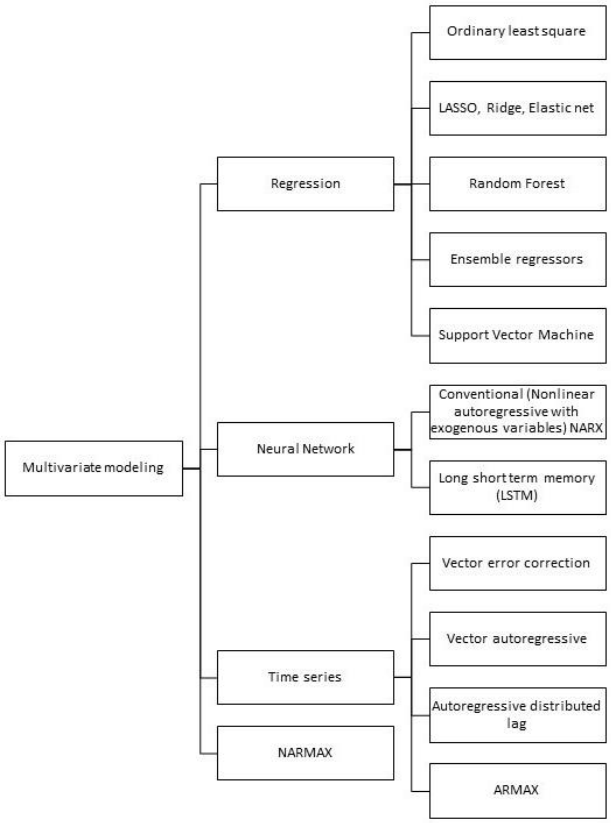

Figure 6. Multivariate models surveyed through this study.

A. *Modeling Project Frequency*

Before modeling the project frequency, it is necessary to conduct some preliminary data analysis to quantify the data's characteristics. Correlogram of autocorrelation and partial autocorrelation reveals that lag 8 and 12 exceeds the significance bounds, which means extending past 8 and 12 values in univariate modeling are the most appropriate options as they demonstrate significant correlations with the original time series under study.

Testing the stationarity of the project frequency is also important. Figure 7 shows the rolling mean and standard deviation of project frequency plotted along with the actual data. It is visually plausible that the data fluctuate around a fixed mean and variance. It can be numerically assessed by using an Augmented Dickey–Fuller test (ADF) to see if the data is stationary. There are three variations of the ADF test, all with the null hypothesis that a unit root is present in a time series sample (series is not stationary). If under any of the three variations the null hypothesis is rejected it can be inferred that the time series is stationary. The ADF test's result (the appropriate lag is chosen based on the Akaike Information Criterion (AIC)) shows that the null hypothesis can be rejected at a 95 percent confidence level. Therefore, the frequency series is considered stationary, meaning it is evolving around a constant mean and variance.

Two approaches can be implemented to forecast project frequency: univariate and multivariate modeling. ARMA and

exponential smoothing are among the most widely used methods to model a univariate time series. ARMA is used to model stationary time series data and is typically represented as ARMA (p,q), where p is the autoregressive order and q is the moving average order. The order of autoregressive and moving average is selected via autocorrelation and partial autocorrelation correlograms. Based on the preliminary data analysis of project frequency an ARMA (p=8, q=8) is the best choice to model the project frequency series. Also, a set of seasonal ARMA models fitted to the data and the best model is selected via AIC. Moreover, other univariate time series methods such as AR, MA, exponential MA, double exponential MA, different variations of ARMA, triple exponential smoothing (Holt-Winters, which takes into account both seasonal changes and trends) method are implemented. This analysis is conducted on the trained and validation data set from 2003 to 2015 using both rolling origin and rolling window methods.

The performance on the test set is the critical measure to compare the performance of the models. Table II presents the summary of the best univariate models and their performance to forecast the project frequency measured by both rolling origin and rolling window cross-validation methods using Root Mean Squared Error (RMSE). Table III presents Mean Absolute Error (MAE) of the same models and cross-validation methods. It should be noted that the results represented here are the average error of the trained models tested on the seven test data sets presented in the cross validation methods of Figure 1. Each test set consist of three different years, so it is safe to assume that there is no over-parametrization or over-fitting represented in the average errors. The results show that almost all the models perform better according to the rolling origin method. This could be due to the fact that in this method the training data is more than what is being used in the rolling window. Thus, the coefficients are calculated more appropriately. Comparing the results of the rolling origin cross-validation method across different models shows that the ARMA (8,8) model outperformed the other models for both the RMSE and MAE measures.

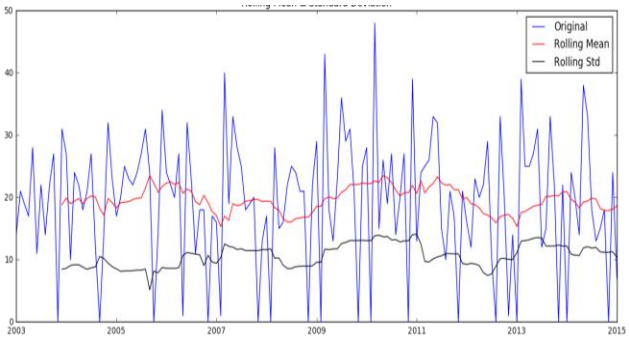

Figure 7. Rolling mean and standard deviation of project frequency plotted against the original data.

TABLE II. SUMMARY OF RMSE OF TIME SERIES MODELS.

Model	Average (rolling origin)	Average (Rolling window)
AR (8)	10.925	10.993
AR (12)	10.934	11.063
MA (8)	11.321	11.343
MA (12)	11.288	11.308
Exponential MA (8)	11.404	11.420
Exponential MA (12)	11.324	11.325
Double Exponential MA (8)	12.050	12.057
Double Exponential MA (12)	11.647	11.683
Auto ARMA	11.057	11.127
ARMA (8,8)	10.715	11.580
ARMA (8,12)	10.830	11.252
ARMA (12,8)	10.870	12.003
ARMA (12,12)	11.556	42.616
Exponential smoothing	11.057	11.138
Holt Winter	10.820	12.814

TABLE III. SUMMARY OF MAE OF TIME SERIES MODELS.

Model	Average (rolling origin)	Average (Rolling window)
AR(8)	8.48	8.551
AR (12)	8.49	8.533
MA (8)	8.85	8.857
MA (12)	8.74	8.765
Exponential MA (8)	9.02	9.040
Exponential MA (12)	8.89	8.894
Double Exponential MA (8)	9.75	9.693
Double Exponential MA (12)	9.3	9.320
Auto ARMA	8.59	8.611
ARMA (8,8)	8.45	9.297
ARMA (8,12)	8.53	9.218
ARMA (12,8)	8.55	9.739
ARMA (12,12)	9.23	31.853
Exponential smoothing	8.59	8.611
Holt Winter	8.7	10.158

The evaluated models so far presented focus only on linear relationships between the inputs. In order to investigate the nonlinear relationship between the inputs a Long-Short Term Memory (LSTM) neural network is used. This method is only implemented with the rolling origin cross-validation method as it requires more data for comprehensive training and the data limitation of the rolling would reduce its performance significantly. Table IV shows the summary of the different LSTM models trained and tested using a grid search to find the optimal number of neurons and lookback number (number of previous values to be considered as input). The results show that the net with two neurons and one lookback performs better than other configurations.

TABLE IV. SUMMARY OF LSTM MODELS PERFORMANCE USING ROLLING ORIGIN METHOD.

lookback	neurons	Average (RMSE)	Average (MAE)
1	1	10.79	8.56
1	2	10.75	8.58
1	3	10.76	8.61
1	4	10.76	8.61
1	5	10.79	8.63
1	10	10.76	8.59
1	20	10.77	8.61
3	1	11.18	8.81
3	2	11.52	9.15
3	3	11.22	9.12
3	4	11.42	9.12
3	5	11.33	9.22
3	10	12.01	9.65
3	20	12.36	9.70
5	1	11.81	9.37
5	2	11.54	9.05
5	3	11.36	9.09
5	4	11.66	9.26
5	5	11.27	9.01
5	10	12.78	10.28
5	20	14.25	11.00
8	1	12.43	10.01
8	2	13.21	10.74
8	3	15.80	12.45
8	4	15.81	12.70
8	5	16.78	13.25
8	10	18.71	14.55
8	20	21.27	15.85
12	1	17.75	13.02
12	2	17.78	13.51
12	3	17.81	13.99
12	4	19.54	15.41
12	5	18.19	14.85
12	10	15.98	13.16
12	20	16.94	13.40

The difference between the performance of the models might seem small. However, considering that it is an average of seven test sections based on the cross validation methods discussed earlier, even small differences are meaningful. Comparing results of the time series model and the LSTM model shows that ARMA (8,8) is the best approach for modeling project frequency. Table V presents a quantitative summary of the training and test set of the last cross validation section for a better understating of how the model compares to the actual data. The mean and median match very well for both the training and test sets. However, the model's variance, standard deviation and range are less than the actual data. Figure 8 provides a more in-depth understanding of the results by a visual illustration of the performance of the ARMA model, illustrating the difference between the actual data and the best performing model. The predicted values are shown in blue, and the actual data are plotted in red. Visual inspection of Figure 8 shows that the model performs better forecasting later values (after 2008) and, likewise, better captures the variance of the actual data in these later years. However, it is evident that the model's variance (blue) is less than the actual data (red) through the whole data set. The gray area represents the prediction intervals for the test data set. The dark grey shows the 80% interval and light grey shows the 95% interval.

Based on the literature [24], [34], [35] including explanatory variables and using multivariate models can yield more accurate results. As a result, following the scheme illustrated in Figure 3 (using multivariate methods to improve project frequency forecast) is part of future work in this study.

B. Modeling Cost and Duration

Cost and duration are the two variables to be sampled from a fitted distribution from past projects. Checking for the correlation between the two variables is essential. A Pearson correlation test shows 0.662 correlation coefficient with 0.00 P-value between the duration and cost at the project level (0.00 P-value shows that the correlation is significant, and it is not due to the chance). This shows a moderately linear relationship between the two variables, and it should be incorporated in the model.

Each member in a set of continuous distributions (consisting of the Inverse Gaussian, Pearson, Fréchet, Normal, Lognormal, Dagum, Fatigue Life, Logistic, Loglogistic, Gamma, Exponential, Triangular, Uniform, Student, and Weibull distributions) has been parametrically fitted to the cross-validation data sections using the maximum likelihood estimation (MLE) method and ranked via AIC. Table VI shows the results from the rolling window cross-validation while Table VII shows the result for the rolling origin cross-validation method. From these tables it can be seen that the distributions that were consistently among the best were the Inverse Gaussian distribution for duration, and the Lognormal distribution for cost.

Figure 9 shows the histogram, and the corresponding fitted distribution for the duration and cost of the projects. An Inverse Gaussian distribution with $\mu = 244.67$ and $\lambda = 273.93$ was found to provide the best fit using AIC for the duration. A lognormal distribution with (mean log) $\mu = 14.413319$ and (standard deviation log) $\sigma = 1.524961$ was found to provide the best fit using AIC for the cost. Through sampling from these distributions, a cost and a duration can be assigned to each forecasted project. As a result, the output of the framework would be the number of projects for each month and a cost and a duration assigned to each project.

TABLE V. QUANTITATIVE SUMMARY OF ARMA MODEL AND ACTUAL DATA

	Training set		Test set	
	ARMA model	Actual data	ARMA model	Actual data
Mean	19.83	19.82	19.92	18.56
Variance	31.33	104.97	38.89	121.57
Std. Dev.	5.60	10.25	6.24	11.03
Median	20.58	21.00	20.96	20.00
Minimum	1.09	0.00	6.87	0.00
Maximum	30.78	48.00	30.75	39.00
Range	29.69	48.00	23.87	39.00

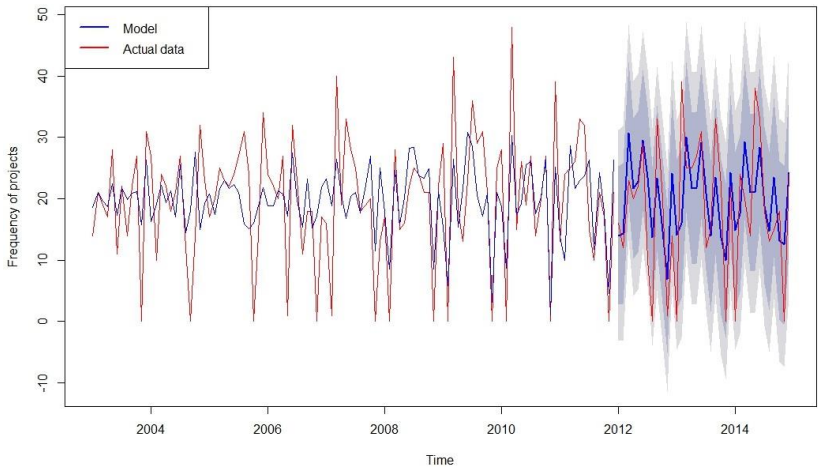

Figure 8. ARIMA (8,0,8) model illustration using 9 years for training and 3 years for testing.

TABLE VI. DISTRIBUTION FIT RESULTS USING ROLLING WINDOW DATA SECTIONS

		2003-2008	2004-2009	2005-2010	2006-2011	2007-2012	2008-2013	2009-2014
Cost	Best distribution	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal
	AIC	45,368.30	46,444.59	47,136.99	46,065.73	45,450.35	45,732.72	45,863.51
Duration	Best distribution	Inverse Gaussian	Inverse Gaussian	Inverse Gaussian	Inverse Gaussian	Inverse Gaussian	Inverse Gaussian	Inverse Gaussian
	AIC	17,724.74	18,115.82	18,322.96	17,963.08	17,821.47	17,933.65	18,022.70

TABLE VII. DISTRIBUTION FIT RESULTS USING ROLLING ORIGIN DATA SECTIONS

		2003-2008	2003-2009	2003-2010	2003-2011	2003-2012	2003-2013	2003-2014
Cost	Best distribution	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal
	AIC	45,368.30	53,914.56	62,259.21	69,807.80	76,323.25	84,308.77	91,330.68
Duration	Best distribution	Inverse Gaussian	Inverse Gaussian	Inverse Gaussian	Inverse Gaussian	Inverse Gaussian	Inverse Gaussian	Inverse Gaussian
	AIC	17,724.74	21,099.32	24,383.72	27,346.81	29,928.68	33,040.76	35,774.54

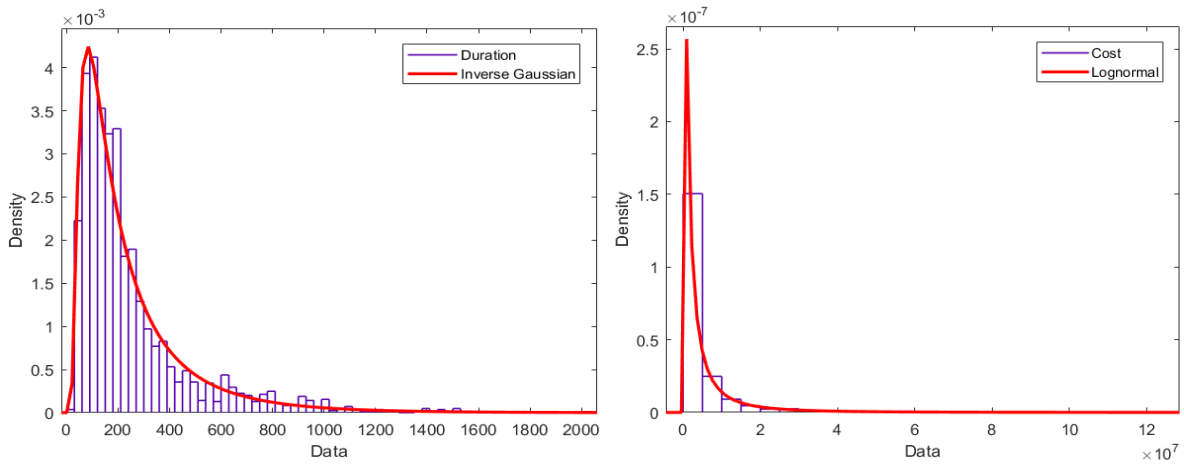

Figure 9. Duration (Left) and Cost (Right) fitted distributions

The performances of the various model components presented in this section indicate the viability of an integrated project stream forecaster that predicts, within a simulation environment, the frequencies of projects and empirical distributions of project duration and cost. Specifically, the generator will produce stochastic streams of unknown future FDOT projects.

V. Conclusion and Future Work.

This paper has proposed an extension to the body of existing project portfolio planning models and discussed a methodology for its development. The proposed model will extend the horizon for portfolio and strategic planning by enabling users to look further into the future and consider unknown (but statistically quantifiable) projects alongside the known and current projects in their planning process.

The proposed model provides an additional component to the current portfolio management models. A general modeling approach with different possible training and validating methods is discussed and results of the research on developing, validating and testing a stream generator to forecast FDOT projects, in terms of time of occurrence, expected duration and expected cost, is presented. It is shown how univariate models can be used to forecast project frequency, and a discussion is provided of the representing distributions for project cost and duration along with their relationship. Results of project frequency univariate modeling showed that ARMA was the best performing model in this case, outperforming the LSTM neural network. This could be explained by the excessive need of such neural networks for large sample datasets. Furthermore, among the tested distributions, the Inverse Gaussian was found to be the most representative of project duration, and the Lognormal distribution was found to be the most representative distribution for project costs.

A set of potentially relevant predictors including the macroeconomics metrics and construction indices have been identified to enable future improvement of the model using multivariate methods. This approach can be applied in different contexts and is not confined to the specific case study discussed in this paper. It is also proposed to expand the scope of the research by adding other characteristics to the project stream generator (such as different project types) and implementing it within various environmental contexts.

The complete framework will allow the user to examine different bidding and project selection strategies to see the impact on a company's portfolio and the future resource demands. Furthermore, it will lead to the selection of a closer to optimal strategy and optimal resource distribution for a user. Finally, taking into account uncertainties in future project streams might decrease the required extent of continuous adjustments to a company's portfolio plan resulting from new projects being added to the portfolio.

References

[1] A. Shojaei and I. Flood, "Extending the Portfolio and Strategic Planning Horizon by Stochastic Forecasting of Unknown Future Projects," in The Seventh International Conference on Advanced Communications and Computation, INFOCOMP 2017, pp. 64–69.

[2] B. S. Blichfeldt and P. Eskerod, "Project portfolio management - There's more to it than what management enacts," Int. J. Proj. Manag., vol. 26, no. 4, pp. 357–365, May 2008.

[3] J. A. Araúzo, J. Pajares, and A. Lopez-Paredes, "Simulating the dynamic scheduling of project portfolios," Simul. Model. Pract. Theory, vol. 18, no. 10, pp. 1428–1441, Nov. 2010.

[4] R. G. Cooper, S. J. Edgett, and E. J. Kleinschmidt, "Portfolio management in new product development: Lessons from the leaders—I," Res. Manag., vol. 40, no. 5, pp. 16–28, 1997.

[5] H. Markowitz, "PORTFOLIO SELECTION*," J. Finance, vol. 7, no. 1, pp. 77–91, Mar. 1952.

[6] W. F. McFarlan, "Portfolio approach to information systems," Harv. Bus. Rev., vol. 59, no. 5, pp. 142–150, 1981.

[7] T. R. Browning and A. A. Yassine, "Resource-constrained multi-project scheduling: Priority rule performance revisited," Int. J. Prod. Econ., vol. 126, no. 2, pp. 212–228, Aug. 2010.

[8] A. F. Carazo, T. Gómez, J. Molina, A. G. Hernández-Díaz, F. M. Guerrero, and R. Caballero, "Solving a comprehensive model for multiobjective project portfolio selection," Comput. Oper. Res., vol. 37, no. 4, pp. 630–639, Apr. 2010.

[9] M. Engwall, "No project is an island: Linking projects to history and context," Res. Policy, vol. 32, no. 5, pp. 789–808, May 2003.

[10] B. Wang and Y. Song, "Reinvestment Strategy-Based Project Portfolio Selection and Scheduling with Time-Dependent Budget Limit Considering Time Value of Capital BT - Proceedings of the 2015 International Conference on Electrical and Information Technologies for Rail Transportat," 2016, pp. 373–381.

[11] B. Canbaz and F. Marle, "Construction of project portfolio considering efficiency, strategic effectiveness, balance and project interdependencies," Int. J. Proj. Organ. Manag., vol. 8, no. 2, p. 103, 2016.

[12] M. Lehnert, A. Linhart, and M. Röglinger, "Value-based process project portfolio management: integrated planning of BPM capability development and process improvement," Bus. Res., vol. 9, no. 2, pp. 377–419, Aug. 2016.

[13] V. Mohagheghi, S. M. Mousavi, B. Vahdani, and M. R. Shahriari, "R&D project evaluation and project portfolio selection by a new interval type-2 fuzzy optimization approach," Neural Comput. Appl., vol. 28, no. 12, pp. 3869–3888, Dec. 2017.

[14] F. Faezy Razi and S. Hooman Shariat, "A hybrid grey based artificial neural network and C&R tree for project portfolio selection," Benchmarking An Int. J., vol. 24, no. 3, pp. 651–665, 2017.

[15] M. Shariatmadari, N. Nahavandi, S. H. Zegordi, and M. H. Sobhiyah, "Integrated resource management for simultaneous project selection and scheduling," Comput. Ind. Eng., vol. 109, pp. 39–47, 2017.

[16] J. M. Hummel, M. D. Oliveira, C. A. B. e Costa, and M. J. IJzerman, "Supporting the Project Portfolio Selection Decision of Research and Development Investments by Means of Multi-Criteria Resource Allocation Modelling," in Multi-Criteria Decision Analysis to Support Healthcare Decisions, Springer, 2017, pp. 89–103.

[17] S.-S. Liu and C.-J. Wang, "Optimizing project selection and scheduling problems with time-dependent resource constraints," Autom. Constr., vol. 20, no. 8, pp. 1110–1119, Dec. 2011.

[18] W. R. Scott, Organizations: Rational, Natural, and Open Systems, 5th Editio. Prentice Hall, 2002.

[19] M. Martinsuo, "Project portfolio management in practice and in context," Int. J. Proj. Manag., vol. 31, no. 6, pp. 794–803, Aug. 2013.

[20] R. B. Duncan, "Characteristics of Organizational Environments and Perceived Environmental Uncertainty," Adm. Sci. Q., pp. 313–327, 1972.

[21] R. L. Daft, "Organziation Theory and Design," in South-Western Cengage Learning, 2009, pp. 138–157.

[22] M. M. Farshchian and G. Heravi, "Probabilistic Assessment of Cost, Time, and Revenue in a Portfolio of Projects Using Stochastic Agent-Based Simulation," J. Constr. Eng. Manag., vol. 144, no. 5, p. 04018028, May 2018.

[23] Y. Petit and B. Hobbs, "Project portfolios in dynamic environments: Sources of uncertainty and sensing mechanisms," Proj. Manag. J., vol. 41, no. 4, pp. 46–58, Sep. 2010.

[24] S. M. Shahandashti and B. Ashuri, "Highway Construction Cost Forecasting Using Vector Error Correction Models," J. Manag. Eng., vol. 32, no. 2, p. 04015040, Mar. 2016.

[25] A. Shojaei and I. Flood, "Stochastic forecasting of project streams for construction project portfolio management," Vis. Eng., vol. 5, no. 1, p. 11, 2017.

[26] A. Shojaei and I. Flood, "Stochastic Forecasting of Unknown Future Project Streams for Strategic Portfolio Planning," in Computing in Civil Engineering 2017, 2017, pp. 280–288.

[27] C. Cargnoni, P. Müller, and M. West, "Bayesian Forecasting of Multinomial Time Series through Conditionally Gaussian Dynamic Models," J. Am. Stat. Assoc., vol. 92, no. 438, pp. 640–647, Jun. 1997.

[28] C. Voyant, G. Notton, C. Darras, A. Fouilloy, and F. Motte, "Uncertainties in global radiation time series forecasting using machine learning: The multilayer perceptron case," Energy, vol. 125, pp. 248–257, Apr. 2017.

[29] J. Li and W. Chen, "Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models," Int. J. Forecast., vol. 30, no. 4, pp. 996–1015, Oct. 2014.

[30] P. Exterkate, P. J. F. Groenen, C. Heij, and D. van Dijk, "Nonlinear forecasting with many predictors using kernel ridge regression," Int. J. Forecast., vol. 32, no. 3, pp. 736–753, Jul. 2016.

[31] X. Yu and S.-Y. Liong, "Forecasting of hydrologic time series with ridge regression in feature space," J. Hydrol., vol. 332, no. 3–4, pp. 290–302, Jan. 2007.

[32] B. Choubin, S. Khalighi-Sigaroodi, A. Malekian, and Ö. Kişi, "Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals," Hydrol. Sci. J., vol. 61, no. 6, pp. 1001–1009, Apr. 2016.

[33] L. J. Cao and F. E. H. Tay, "Support vector machine with adaptive parameters in financial time series forecasting," IEEE Trans. Neural Networks, vol. 14, no. 6, pp. 1506–1518, Nov. 2003.

[34] S. Thomas Ng, S. O. Cheung, R. Martin Skitmore, K. C. Lam, and L. Y. Wong, "Prediction of tender price index directional changes," Constr. Manag. Econ., vol. 18, no. 7, pp. 843–852, Oct. 2000.

[35] J. M. W. Wong and S. T. Ng, "Forecasting construction tender price index in Hong Kong using vector error correction model," Constr. Manag. Econ., vol. 28, no. 12, pp. 1255–1268, 2010.