Forecasting Transportation Project Frequency using Multivariate Regression with Elastic Net Regularization

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

Abstract—**Knowledge of the number of upcoming projects and their impact on the company plays a significant role in strategic planning for project-based companies. The current horizon of planning for companies working on public projects are the latest advertised projects for bidding, which in many cases is reported less than a year in advance. This provides a very short-term horizon for strategic project portfolio planning. In this research, a multivariate regression model with elastic net regularization, using economic indices and other environmental factors, is built for Florida Department of Transportation (FDOT) projects to forecast the number of projects they will advertise in the future. The results show that, of the predictors considered, unemployment rate in the construction sector and the Brent oil price are the most significant variables in forecasting FDOT's future project frequency.**

Keywords-Multivariate Regression; Elastic Net Rugularization; Strategic Planning; Project Portfolio Management, Forecasting.

I. INTRODUCTION

Construction companies, as with many other companies working in project-based industries, such as IT, are usually managing multiple projects concurrently while looking for new projects to maintain their business. The task of managing current (ongoing) projects while obtaining projects for continuous business is called Project Portfolio Management (PPM). A crucial part of the management of a portfolio is to make sure that the company resources and on-going projects are optimally balanced to ensure that not only each project meets its objectives but also the whole organization meet its overall goals. Management needs to make sure that they maximize the utilization of their resources by minimizing idle time while not accepting more work than they can complete effectively.

The majority of the literature focuses on internal uncertainties that pertain to PPM. In other words, the most explored aspect of the uncertainties in PPM is the relationships between the projects within the portfolio and the interaction between the current ongoing projects and possible future projects to measure their compatibility in terms of resource demand, and other criteria. However, environmental factors, such as economic conditions and specific industry conditions (for instance, the number of workers in construction) can have a significant impact on the portfolio and company's overall performance. This study aims to integrate the environmental uncertainties and uncertainties regarding the unknown future projects, so that companies can apply this approach in their mid-term to long-term strategic planning. Martinsuo's [1] review of PPM frameworks showed that the uncertainty and continual changes in a company's portfolio has a significant negative correlation with its success. As a result, if users can reduce the extent of the uncertainties in their planning and have a more robust portfolio, it could greatly help their success. In summary, this paper proposes a regression model for forecasting frequency of FDOT's future projects, which helps the user to estimate the number and timing of tendered projects in the future. The novelty of this approach is the consideration of environmental uncertainties in the model and the provision of quantitative insights into the unknown future.

The rest of this paper is organized as follows. Section II describes the impact of uncertainty on PPM and how unknown future projects can impact strategic planning. Section III describes the modeling approach followed in this paper. Section IV addresses the regression model for forecasting the number of projects in the future. Section V presents the conclusions and identifies future directions for the research.

II. UNKNOWN FUTURE PROJECTS AND PORTFOLIO STRATEGIC PLANNING

Planning is vital to the success of any construction entity. In the public sector, governmental agencies try to forecast the needed equity in advance in order to successfully plan the number of their future projects. Historically, governmental agencies have had a short-sighted view towards predicting the future; mainly due to uncertainty in the size of the next year's budget. The process of planning the future is costly, slow and traditionally based on historic-data on past projects. This process is usually projected one year in advance as budgetary issues restrict the ability of governmental agencies to define the scope, number, and types of projects that are needed in later years. In the private sector, the process of defining future projects (in terms of scope, number and types) is better planned compared to the public sector. However, this planning process is still far from ideal.

In project management, the process of targeting goals for multiple projects in a portfolio of a company is referred to as PPM. PPM is defined 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"* [2]. PPM deals with two significant tasks, which are complementary: (1) reinforcing investment decisions by helping companies to select projects that optimize their return on investment and risks associated with them as a whole [3]; and (2) optimizing the allocation of resources across different projects within portfolios in order to meet project goals and minimize risks [4]. The key to effective implementation of PPM within any construction entity is information. The unknown nature of the future is a primary factor that can undermine the success of the PPM process [5].

Uncertainty may influence the success of any organization in any discipline [6][7]. In project management, uncertainty is referred to as the degree of accuracy in determining future work processes, resource variation and work output [8]. The Project Management Institute (PMI) introduces risk management to the broader context of portfolio management. However, PMI does not provide many direct and specific guidelines, recommendations, plans or procedures on how to effectively manage future uncertainties at the portfolio level. Risk management at the portfolio level is restricted to naming only a few risk management techniques. PMI only suggests some vague guidelines on how to detect, monitor and handle uncertainties [5].

At the scientific level, managing uncertainties in projects has usually been handled by analyzing historical project data. Many methods and approaches have been used to collect and analyze historical data to find trends that might help understand how uncertainty impacts the success of projects and/or portfolios. Trippi et al. [9] suggest using Artificial Intelligence (AI) in portfolio management. Henriksen and Traynor [10] developed algorithms to allocate risks and other criteria in project selection and portfolio management. More advanced analysis methods, such as multi-agent modeling [11], multi-objective binary programming [12] and use of Bayesian Networks [13] have also been introduced by researchers to analyze the uncertainty and/or allocated different risks associated with projects at the project and/or portfolio levels. However, incorporating future project forecasts in portfolio planning with consideration of unknown environmental uncertainties remains largely unexplored.

III. MODELING APPROACH

The literature [5][7][14] has looked at forecasting unknown future projects with a univariate modeling approach where the number of future projects are forecasted solely based on the past values of the number of projects. This study builds upon this work by forecasting unknown future projects using multivariate regression in order to incorporate environmental uncertainties in a forecast. The data used in this case study is obtained by text mining FDOT’s historical project letting database. The database covers 12 years (from 2003 to 2015) containing 2816 projects. The features extracted from the database are each project letting date, cost, and duration. Table 1 provides a pool of candidate independent variables including macroeconomics and construction indices compiled from the literature [5][7][14], which were available at the monthly level and did not have any missing values for the explored time frame. Table 1 also provides the abbreviation for each variable and the sources from which they have been obtained. These factors are considered in the regression modeling as the dependent (explanatory) variables.

The integrity and continuity of the data are important as it is a time series. As a result, random cross validation was not appropriate, and a rolling forecast origin approach was adopted for cross-validation, as illustrated in Figure 1. The data were divided into two sections, training and testing. The training period starts with three years and increases by one year in each iteration while the testing period remains steady as the three consecutive years after the training set. In other words, seven models are trained, and the average error is considered as the result of cross-validation.

TABLE I. CANDIDATE INDEPENDENT VARIABLES.

Variable name	Abbreviation of variable	Source
Dow Jones industrial average Vol	DJI	Yahoo Finance
Dow Jones industrial average Closing	DJIC	Yahoo Finance
Money Stock M1	MS1	Federal Reserve System
Money Stock M2	MS2	Federal Reserve System
Federal Fund Rate	FFR	Federal Reserve Systems
Average Prime Rate	APR	Federal Reserve System
Producer Price Index for All Commodities	PPIACO	U.S. Bureau of Labor Statistics
Building Permit	BP	U.S. Bureau of Census
Brent Oil Price	BOP	U.S. Energy Information Administration
Consumer Price Index	CPI	U.S. Bureau of Labor Statistics
Crude Oil Price	COP	U.S. Energy Information Administration
Unemployment Rate	UR	U.S. Bureau of Labor Statistics
Florida Employment	FE	U.S. Bureau of Labor Statistics
Florida Unemployment	FU	U.S. Bureau of Labor Statistics
Florida Unemployment Rate	FUR	U.S. Bureau of Labor Statistics
Florida Number of Employees in Construction	NFEC	U.S. Bureau of Labor Statistics
Number Housing Started	HS	U.S. Bureau of Census
Unemployment Rate Construction	URC	U.S. Bureau of Labor Statistics
Number of Employees in Construction	NEC	U.S. Bureau of Labor Statistics
Number of Job Opening in Construction	JOC	U.S. Bureau of Labor Statistics
Construction Spending	CS	U.S. Census Bureau
Total Highway and Street Spending	THSS	Federal Reserve System

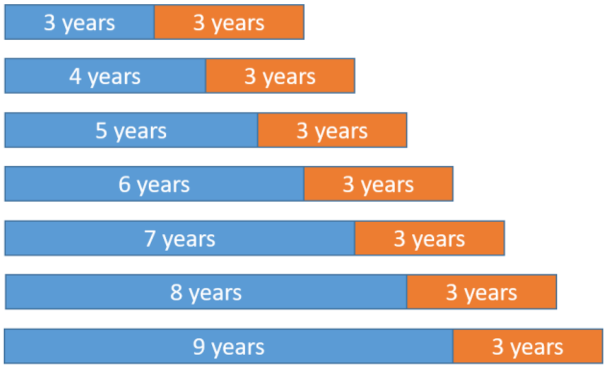

Figure 1. Forecast on a rolling origin cross-validation.

A. *Exploratory data analysis*

To develop the multivariate models, a better understanding of the data characteristics was first necessary, and that information was gained through an exploratory data analysis and the identification of potentially relevant predictors.

The first exploratory analysis consisted of a correlation analysis. Figure 2 provides the correlation plot of the variables. The color indicates the magnitude of the correlation, and the direction of the ellipse illustrates the direction of the relationship. Furthermore, the concentration of the ellipse tells us about the degree of the linear relationship between the variables. Project frequency is represented by "freq" in the last row and column. It appears that none of the exploratory variables had a strong linear relationship with the project frequency.

B. *Feature selection and feature importance*

Feature selection is the process of selecting the most relevant predictors and removing irrelevant variables from the pool of potentially useful predictors. Depending on the model's structure, feature selection can improve a model's accuracy. This process can be carried out by measuring the contribution of each variable to the model's accuracy, and then removing irrelevant and redundant variables while keeping the most useful ones. In some cases, irrelevant features can even reduce a model's accuracy. In general, there are three approaches to feature selection: the filter method, wrapper method, and embedded method.

Embedded methods implement feature selection and model tuning at the same time. In other words, these machine learning algorithms have built-in feature selection elements. Examples of embedded method implementations include LASSO and elastic net. Regularization is a process in which the user intentionally introduces bias into the training, preventing the coefficients from taking large values. This method is especially useful when the number of variables is high. In such a situation, the linear regression is not stable and in which a small change in a few variables results in a large shift in the coefficients. The LASSO approach uses L1 regularization (adding a penalty equal to the magnitude of the coefficient), while ridge regression uses L2 regularization (adding a penalty equal to the square of the magnitude of the coefficient). Elastic net uses a combination of L1 and L2. Ridge regression is effective in reducing a model's variance by minimizing the summation of the square of the residuals. The LASSO method minimizes the summation of the absolute residuals. The LASSO approach produces a sparse model that minimizes the number of coefficients with non-zero values. As a result, this approach has implicit feature selection. The generalized linear method implemented in the next section uses elastic net. This approach incorporates both L1 and L2 regularization and thus has implicit feature selection.

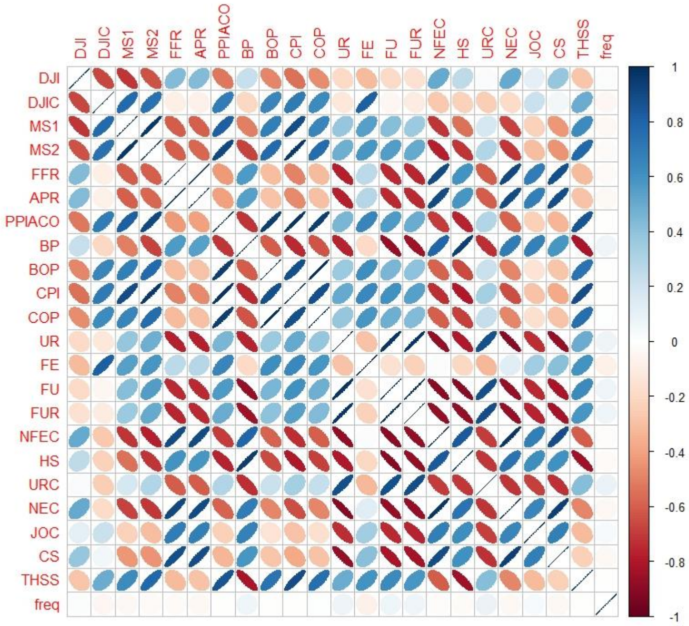

Figure 2. Correlation plot.

Feature reduction methods, such as principal component analysis (PCA), are widely used in studies to reduce the number of independent variables. The output of such methods is a reduced set of new variables extracted from the initial variables while attempting to maintain the same information content. However, using these methods can drastically decrease the ability to interpret the significance of each input, which in itself can be very beneficial. For example, in this study knowing that oil price has a significant impact on the frequency of the projects compared to construction spending can provide valuable insight both for policy makers and contractors. As a result, the authors have chosen not to implement feature reduction methods, such as PCA.

Looking at the correlation between independent variables and the dependent variable, it became evident that a filter method using a correlation analysis was not useful, as all the variables had a nonsignificant relationship with the project frequency. As a result, an elastic net approach is used in the next section.

IV. Regression Model

The general process of model optimization and feature selection consisted of first defining a set of model parameter values to be evaluated. Then, the data was preprocessed in accordance with a 0-1 scale to make sure the high value in some variables are not skewing the model's coefficient and other variables importance. For each parameter set, the cross-validation method discussed earlier served to train and test the model. Finally, the average performance was calculated for each parameter set to identify the optimal values for the parameters.

Ordinary linear regression is based on the underlying assumption that the model for the dependent variable has a normal error distribution. Generalized linear models are a flexible generalization of the ordinary linear regression that allows for other error distributions. In general, they can be applied to a wider variety of problems than can the ordinary linear regression approach. Generalized linear models are defined by three components: a random component, a systematic component, and a link function. The random component recognizes the dependent variable and its corresponding probability distribution. The systematic component recognizes the independent variables and their linear combination, which is called the linear predictor. The link function identifies the connection between the random and systematic components. In other words, it pinpoints how the dependent variable is related to the linear predictor of the independent variables.

Ridge regression uses an L2 penalty to limit the size of the coefficient, while LASSO regression uses an L1 penalty to increase the interpretability of the model. The elastic net uses a mix of L1 and L2 regularization, which makes it superior to the other two methods in most cases. Using a combination of L1 and L2, the elastic net can produce a sparse model with few variables selected from the independent variables. This approach is especially useful when multiple features with high correlations with each other exist.

A generalized linear model was fit to the data using the cross-validation method discussed earlier. Alpha (mixing percentage) and lambda (regularization parameter) were the tuning parameters. Alpha controls the elastic net penalty, where $\alpha=1$ represents lasso regression, and $\alpha=0$ represents ridge regression. Lambda controls the power of the penalty. The L2 penalty shrinks the coefficients of correlated variables, whereas the L1 penalty picks one of the correlated variables and removes the rest. Figure 3 illustrates the results of the generalized linear model (for each set of parameters 7 models according to cross-validation method is trained and the average error is assigned to the set of parameters under study), optimized by minimizing the RMSE with controlling alpha and lambda. The optimized parameters were $\alpha=1$ and $\lambda= 0.56$. The authors also tested λ higher than 0.56 up to 1, however, the coefficients were not well-behaved beyond lambda=0.56.

Figure 4 depicts the LASSO coefficient curves. Each curve represents a variable. The path for each variable demonstrates its coefficient in relation to the L1 value. The coefficient paths more effectively highlight why only two variables were significant in the generalized linear model. When two variables were excluded, all other coefficients became zero at the L1 normalization, and this arrangement yielded the best performance. Figure 5 offers the variable importance for the generalized linear model with all the variables. Only the unemployment rate in construction industry, the Brent oil price, and the unemployment rate (total) had non-zero coefficients. However, the unemployment rate (total) seemed to be relatively insignificant.

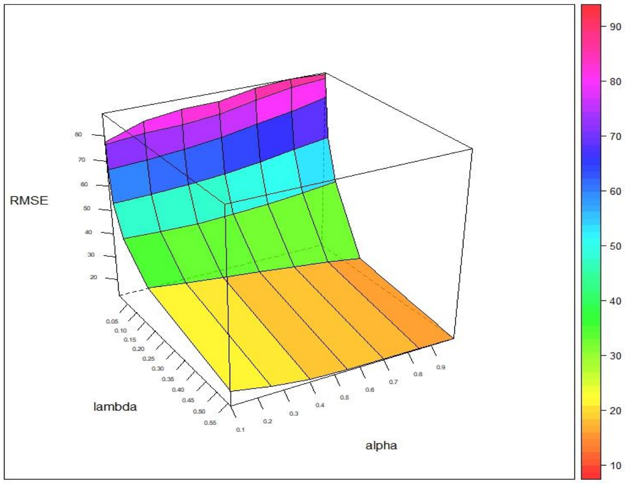

Figure 3. Generalized linear method optimization.

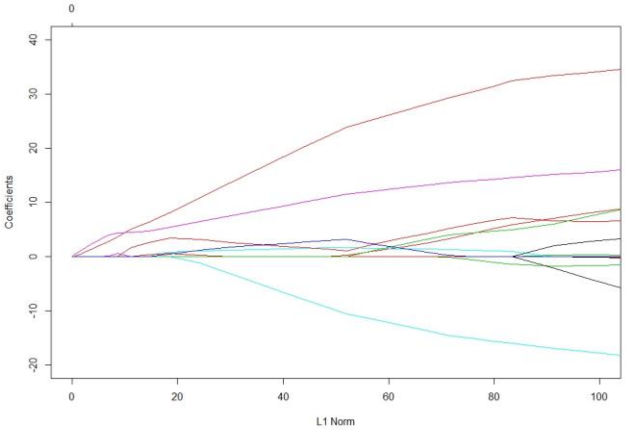

Figure 4. Lasso coefficient curve.

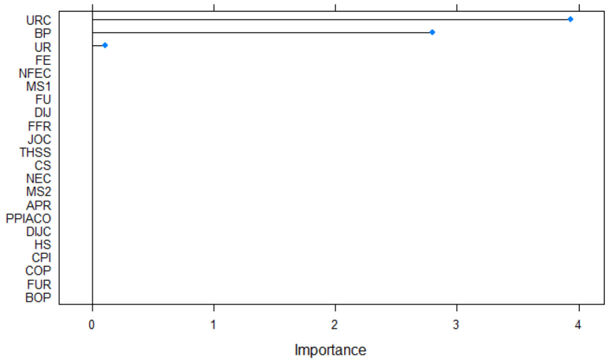

Figure 5. Variable importance of the generalized linear model.

To further prune the generalized linear model, another model with only the unemployment rate in the construction sector and the Brent oil price was trained and tested. Table 2 contains the optimized parameters (coefficients and intercept) for the generalized linear models. The general unemployment rate had a low coefficient and, upon pruning it, the authors saw an improvement in the performance of the model. The most important variable was unemployment rate in construction having the highest coefficient of 4.03.

Table 3 illustrates the performance of the optimized general linear model using a different dataset on the cross-validation sections. It was evident that excluding the unemployment rate improved the model's performance over most of the cross-validation data sections. It is notable that the pruned model performed much better in data section 1 which had the highest error and produced a more evenly distributed error among the different data sections tested. The only variables contributing to the final linear model were the unemployment rate in the construction sector and the Brent oil price.

Table 4 provides a comparison between the regression models proposed in this study and some other univariate models studied previously by the authors [5]. Comparing the error terms shows that the regression model is not outperforming some of the univariate models, such as Autoregressive Moving Average (ARMA). However, it comes close to the best performing example and it provides insight regarding the impact of environmental uncertainties on future project streams and thus could be valuable in long term strategic planning.

It is important to note that the result of this model is the frequency of FDOT's unknown future projects, about which the user would otherwise have no information. Having reliable estimates with known error margins regarding unknown future projects can arguably provide more insight in strategic planning for a company's future compared to the current conjecture-based decision making. It should be noted that the accuracy of the model as long as the model is stable (the error is not systematic but random) is acceptable. The model is forecasting an unknown-unknown variable in the future for which there is no information available regarding their existence. However, users can use the output of this model including the error margin as inputs to their strategic planning.

TABLE II. PARAMETERS OF THE GENERALIZED LINEAR MODELS.

Variables	Coefficients	Coefficients (Pruned by one variable)
URC	3.94	4.03
BP	2.80	2.77
UR	0.11	-----
Intercept	17.14	17.16

TABLE III. PERFORMANCE OF THE GENERALIZED LINEAR MODEL.

Error term	**RMSE**		**MAE**	
Feature set	**All**	**Pruned**	**All**	**Pruned**
1	16.13	9.78	13.24	10.8
2	11.58	11.94	9.64	8.56
3	13.86	13.69	11.6	8.01
4	13.16	13.14	10.82	8.25
5	12.07	10.94	9.55	10
6	11.03	10.27	8.53	8.6
7	10.89	10.87	8.6	11.28
Average	12.67	11.52	10.28	9.36

TABLE IV. PERFORMANCE COMPARISON OF DIFFERENT MODELS.

Model	RMSE	MAE
Regression	11.52	9.36
ARMA(8,8)	10.715	8.45
ARMA(12,12)	11.556	9.23
AR(8)	10.925	8.48
Exponential MA (8)	11.404	9.02

ISBN: 978-1-61208-655-2

The output of this research can provide quantitative insight as a foundation for future planning. It should be noted that this model is not a standalone portfolio management framework, rather it is a supplement to existing models. For example, knowing that there is likely to be a decrease or increase in the number of projects in the future can help a company prepare in terms of consolidating or expanding its resources and assets.

V. Conclusion and Future Work

The importance and impact of upcoming projects on a project portfolio has been established in previous published work. However, little work has been done considering the uncertainties regarding incorporating unknown future projects in long term strategic planning. In this paper, an approach for incorporating environmental uncertainties for forecasting the number of unknown future projects is presented. A multivariate regression model with elastic net regularization was used to forecast FDOT's unknown future projects using economic and construction indices. The results indicate that the approach can reduce the impact of uncertainties on their portfolio and thus enable development of a more robust plan with a better strategic plan. The generalized linear model indicated that the best explanatory variables were the unemployment rate in the construction sector and the Brent oil price. The regression model's performance is no better than other methods tried earlier by the authors, such as a univariate autoregressive moving average model [5] regressing on project frequency's past value. However, this regression model provides insight regarding the impact of environmental uncertainties on future project streams and thus could be valuable in long term strategic planning. The regression model presented in this literature only considers the linear relationship between the variables. Exploring non-linear modeling techniques, such as neural networks for capturing more complicated relationships between the variables would be the next logical step in this research. The model developed in this study is limited to FDOT projects. However, new regression models specific for other databases can be built by following the same steps and adopting appropriate alternative sets of independent variables.

References

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

[2] 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.

[3] H. Markowitz, "Portfolio Selection," J. Finance, vol. 7, no. 1, pp. 77–91, Mar. 1952.

[4] J. S. Pennypacker and L. D. Dye, "Project Portfolio Management and Managing Multiple Projects : Two Sides of the Same Coin ?," in Managing Multiple Projects: Planning, Scheduling, and Allocating Resources for Competitive Advantage, CRC Press, 2002, pp. 1–10.

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

[6] 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.

[7] 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.

[8] J. Dahlgren and J. Soderlund, "Modes and mechanisms of control in Multi-Project Organisations: the R&amp;amp;D case," Int. J. Technol. Manag., vol. 50, no. 1, p. 1-22, 2010.

[9] R. R. Trippi, P. By-Lee, and K. Jae, Artificial intelligence in finance and investing: state-of-the-art technologies for securities selection and portfolio management. McGraw-Hill, Inc., 1995.

[10] A. D. Henriksen and A. J. Traynor, "A practical r&d project-selection scoring tool," IEEE Trans. Eng. Manag., vol. 46, no. 2, pp. 158–170, May 1999.

[11] 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.

[12] A. F. Carazo, et al. "Solving a comprehensive model for multiobjective project portfolio selection," Comput. Oper. Res., vol. 37, no. 4, pp. 630–639, Apr. 2010.

[13] R. Demirer, R. R. Mau, and C. Shenoy, "Bayesian networks: a decision tool to improve portfolio risk analysis," J. Appl. Financ., vol. 16, no. 2, p. 106, 2006.

[14] 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, 2017, no. c, pp. 64–69.