Preliminary Estimation of Natural Gas Pipeline Construction Costs with Regression Analysis

C. Çakmak and M. Erdal
Gazi University, Department of Civil Engineering, Ankara, Turkey
coskuncakmak@gazi.edu.tr, merdal@gazi.edu.tr

Abstract

Natural gas pipeline construction (NGPC) projects are inherently includes high changefulness like any other construction works. Considering that variability, it can be stated that to make an accurate cost estimation is difficult. The main objective of this paper is to achieve a coherent mathematical equation to be able to make a preliminary cost estimation of NGPC costs. To this end, a number of pipeline cost data, which is in US dollar currency, was obtained for the period 1986-2010. With this data, a simple regression equation was tried to be revealed by working through statsmodels, one of the libraries used for the Python programming language. It was concluded that a significant part of the NGPC cost could be explained with the diameter and length variables. In this respect, this paper has the potential to assist the institutions responsible for the execution of the NGPC processes of the Turkey or the contractors considering investment or at the decision-making stage. Moreover, the paper presents a first attempt at both predicting of NGPC costs in Turkey and executing that operation by using Python programming language.

Keywords: *natural gas pipeline, preliminary cost estimation, python.*

Introduction

An accomplished cost estimation is one of the key factors of project success in construction sector (Kim et al., 2004). It plays a vital role for each part of a construction project. It is crucial for owners to be able to create a systematic investment program in the meantime, while it is lifesaving for contractors for calculation of profitability and forecasting of cash flow. It should be noted that a real-like cost estimation would also set the designer's limits. Especially, preliminary cost estimation (PCE) techniques with limited project information are very helpful for partners of construction works like owners, designers, contractors and so on (Arage & Dharwadkar, 2017). The earlier successful cost estimation before beginning of the project brings early decisions among those partners. Thus, an environment arises in which each partner of construction works can foresee the future and people can make more solid decisions. Just as an accurate PCE may increase the profitability of the contractor, an improper one may affect the profitability of contractors negatively and lead indispensably project failure (Cheng & Hoang, 2014). It is a well-known fact that a high level of detail does not exist at PCE stage. That is why we need an amount of data of similar previous projects to make estimation. When we take the competitive environment of construction sector into consideration, it is very clear to see the necessity of a fast and efficient estimation method for

the people who are employee in cost and planning department of construction companies (Uğur, 2007).

Natural gas pipeline constructions (NGPCs), like other construction works, have pretty much uncertainty in terms of project cost because of its nature. To be able to make an accurate cost estimation is thereby considerable not easy. In the literature, there are a small number of research scrutinizing costs of NGPC projects. Those studies, almost all of which were studied in United States, indicate that there are several cost components in NGPC like labor, material (pipes), miscellaneous and right of way (ROW) (Parker, 2004; Rui et al., 2011a, 2011b, 2012; Thaduri, 2012; Ulvestad & Overland, 2012; Zhao, 2000). In these studies, details of which will be shared in literature part, while some researchers try to obtain an equation that gives total estimated cost with those four cost components, others contributed the descriptive statistics of the total costs of pipeline projects in the past to the literature. In Turkey, unfortunately, there is no a database where the costs of the NGPCs are kept by dividing them into components unlike in the United States. Therefore, it is not possible to study with aforementioned cost components and to obtain an equation, which is result of those. Under these circumstances, this research has focused on constructional independent variables that are inherent for NGPCs instead of general cost components like labor, material, miscellaneous and ROW. Details of constructional independent variables will be presented in later sections. The research suggests an early estimation technique of NGPC costs via an equation created with statsmodels library of Python. Since the maximum possible number of data is reached during the data acquisition phase, it was necessary to comment that the research on a whole count, not a sample, for our study. This is the most important limit of the research and that is why further statistical evaluations were not executed such as normal distribution of residuals.

Literature

There are many studies covering cost estimation techniques in construction sector via various methods (Erdal, 2021; Sueri & Erdal, 2022; Ugur et al., 2019). However, few of them focus on the NGPCs. Lack of the information in detail about NGPCs or arrangement difficulty of existing complex database may be the main reason of that deficiency. In this part, a few studies addressing only cost estimations in NGPCs or other pipeline systems have been explained. In addition to these, although not directly aimed at estimating natural gas pipeline construction costs, a few related studies are also included in this section.

Zhao (2000) reported the distribution and cost variation of natural gas pipelines in Europe, Asia and United States. Researcher stated in her study that material and labor costs cover 80% of the NGPC costs for onshore while it is 84% for offshore (Zhao, 2000). Parker (2004) investigated the cost differences in hydrogen transmission lines based on the costs of natural gas, fuel oil and oil pipelines. Author concluded that NGPC costs consist of 26% material costs, 45% labor, 7% miscellaneous expenses (research, inspection, insurances, permits, etc.) and 22% ROW (Parker, 2004). Rui et al. (2011a) developed five different regression models and tried to estimate the costs of 412 onshore natural gas pipelines built between 1992 and 2008 in different regions in the United States. In the study, the authors determined that large-scale natural gas pipeline projects have more affordable unit costs (Rui et al., 2011a). In another study in same year by Rui et al. (2011b), researchers stated that learning the job by repetition reduces the labor costs over time (Rui et al., 2011b). In his study, Thaduri (2012) collected data on 180 natural gas pipelines and 136 pressure stations completed within the borders of the United States and Canada between 2000 and 2008 and presented descriptive

statistical data. Because of the evaluations, it was stated that the cost items with the highest share in the costs of natural gas pipelines were labor with 40% and material costs with 31% (Thaduri, 2012). Ulvestad and Overland (2012) examined the cost analysis of two different natural gas transmission options, pipeline and liquefied natural gas (LNG). Authors concluded that pipeline transport is a better option for larger natural gas volumes and shorter distances while it is LNG for smaller volumes and longer distances (Ulvestad & Overland, 2012). Bai et al. (2013) estimated the annual cost for the pipelines required for CO_2 supply, with the support of the existing literature in their studies. In the study, where two different routes with similar characteristics were selected in two different regions of China, they estimated the annual costs to be 52.52 RMB/t on average (Bai et al., 2013). Matthews et al. (2015) evaluated infrastructure projects within the scope of the concept of "social cost". Especially in infrastructure construction or rehabilitation projects where excavation is applied, calculation methods are proposed for each social cost category because of the inability to use the road, noise, pollination due to excavation, etc. Researchers determined that the social cost could be calculated between 1% and 9% compared to the project cost (Matthews et al., 2015). Zhang et al. (2016) conducted a research on interregional natural gas supply and infrastructure expansion in China by 2035, using an optimization model. In the study, in which four different scenarios were evaluated, it was stated that natural gas wholesale price and import cost are the most important points for China. In addition, it was evaluated what kind of natural gas supply changes/problems these scenarios will create in the country (Zhang et al., 2016). Kaiser (2021) examined the costs of onshore and offshore oil and gas pipelines in the United States. The author also mentioned pipeline requirements, cost estimation procedures and normalization processes in his study, where he presented the similarities and differences in the construction and demolition processes of onshore and offshore pipelines (Kaiser, 2021).

Preparation of Dataset

Data of available approximately 100 NGPCs in Turkey were reached by a not easy way. The information of all projects made between 1986 and 2010 was examined manually and dataset was finalized. After the examinations, in first stage the independent variable candidates that can be used in our regression model were pipe diameter (inch), length of pipeline (km), excavation amount (m^3), ideal depth (m), line valve (number), take-off valve (number), pigging station (number) and project duration (day). In examined papers, information of some projects were missing. Therefore, number of data decreased to 57 because of lack of information that can be used as a variable like project duration, cost etc. As an example, information of first five projects of 57 natural gas pipeline projects were showed in Table 1 which was created with codding *df.head()* (df is name of the dataset) in *Pandas* library which is one of the most popular modules in Python.

After finalizing the dataset some descriptive statistics of variables and correlation between variables were calculated with codding *df.describe().T* and *df.corr()*. Descriptive statistics including important explanatory statistical information such as mean and standard deviation are presented in Table 2. Table showing correlation values between variables and formatted as a heat map is presented in Figure 1.

Table 1. First five data item of the dataset.

Number	Diameter	Length	Excavation	Depth	Line	Take-off	Pigging	Project	Cost

	(inch)	(km)	(m^3)	(m)	valve (number)	valve (number)	station (number)	duration (day)	($)
0	16	1.57	1992.31	1.41	0	0	0	343	1478615
1	40	110.96	385488.94	2.32	3	0	1	747	47364121
2	12	12.64	12375.07	1.22	0	2	2	14	860314
3	40	102.47	355973.83	2.32	3	0	1	771	44619968
4	16	25.00	31644.00	1.41	0	2	1	46	2640478

Table 2. Descriptive statistics of 57 NGPC costs.

Parameters	**Mean**	**Standard Deviation**	**Min**	**25%**	**50%**	**75%**	**Max**
Diameter (inch)	21.68	13.51	6.00	10.00	16.00	36.00	48.00
Length (km)	77.89	68.97	1.57	32.22	59.62	100.15	320.00
Excavation (m^2)	185415.93	280057.34	1992.31	31255.85	101466.29	204683.31	1311421.46
Depth (m)	1.59	0.53	0.97	1.07	1.41	2.11	2.52
Line valve (number)	2.33	2.42	0.00	0.00	2.00	3.00	10.00
Take-off valve (number)	2.16	4.58	0.00	0.00	1.00	3.00	34.00
Pigging station (number)	1.14	0.61	0.00	1.00	1.00	1.00	3.00
Project duration (day)	456.00	325.77	14.00	205.00	435.00	662.00	1205.00
Cost ($)	24988742	36524306	860314	2966139	12061949	27678047	174505208

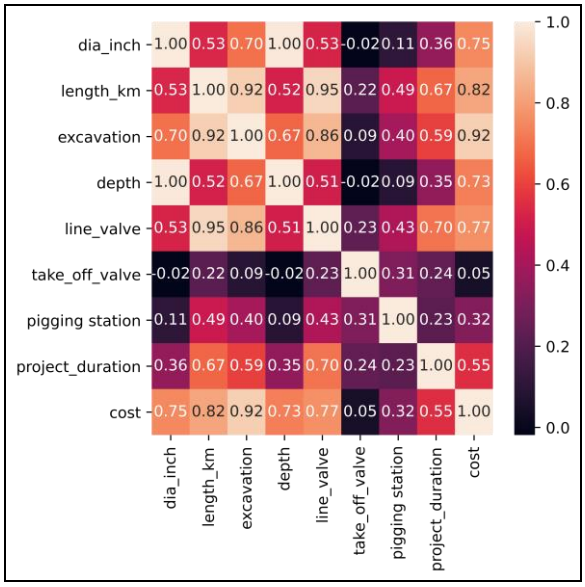


Figure 1: Heat-map of correlation between variables.

Data Preprocessing

To increase the quality of data is the first step of executing of any model because it can lead fallacious results (Pyle et al., 1999). At first blush, it is easy to see that dataset includes cost data taking places in a wide range by looking at Table 2. Even in Table 1, which shows only first five data of all dataset, there is a considerable amount of variability. Data set with an average of about $25M has a standard deviation of about $36M by the way. However, a correlation problem of independent variables was determined. In Figure 1 one can easily say that excavation and depth parameters look like a result of a mathematical equation due to the fact that correlation higher than 0.9 value. Same comment is available for line valve variable

by the way because of its high correlation value. That means we do not need those variables to be able to explain the variability of cost data as the diameter and length parameters are enough. It is most probable that excavation and depth parameters are calculated values by length and depth respectively. Therefore, it is a better way to eliminate these parameters from dataset. On the other hand, we should be awake to differences between 75% quartile cost value and maximum cost value. The gap between those two values makes one think about existing of outlier values. At this stage, it was decided to define and eliminate outlier values by creating a boxplot of cost data. The boxplot created with *seaborn library* in Python in which 25% quartiles can be seen is given in Figure 2.

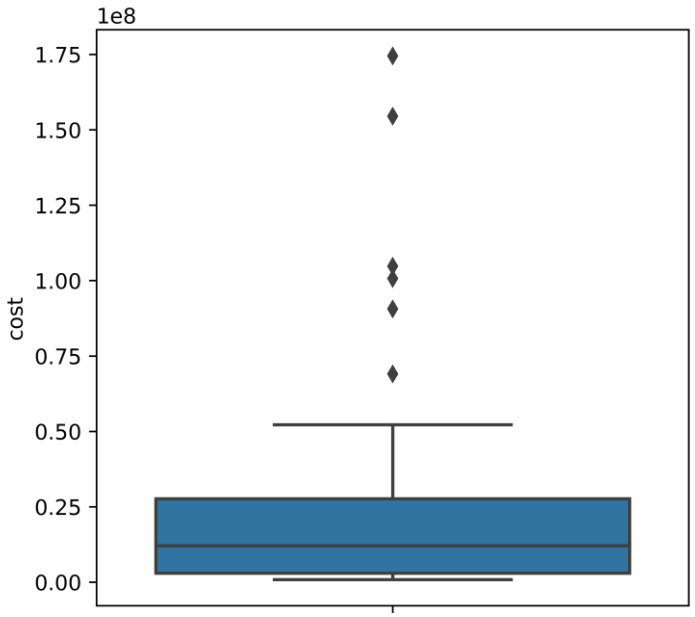


Figure 2: Boxplot of cost data.

According to plot, cost values higher than approximately $60M are outlier values and should be subtracted from dataset because it is thought that those values would affect negatively the final equation. After all these arrangements number of cases decreased to 51 and 3 independent variables were taken out of the dataset. After these processes, descriptive statistics and correlation matrix of variables were created and shown in Table 3 and Figure 3.

Table 3. Descriptive statistics of 51 NGPC costs.

parameter	**mean**	**standard deviation**	**min**	**25%**	**50%**	**75%**	**max**
Diameter (inch)	19.22	11.96	6.00	10.00	14.00	24.00	40.00
Length (km)	59.89	36.06	1.57	30.98	58.28	74.89	164.50
Take-off valve (number)	2.06	4.82	0.00	0.00	1.00	2.50	34.00
Pigging station (number)	1.06	0.54	0.00	1.00	1.00	1.00	2.00
Project duration (day)	399.12	285.81	14.00	174.00	358.00	585.50	1205.00
Cost ($)	14317338	15111751	860314	2799282	8350746	22154375	52219433

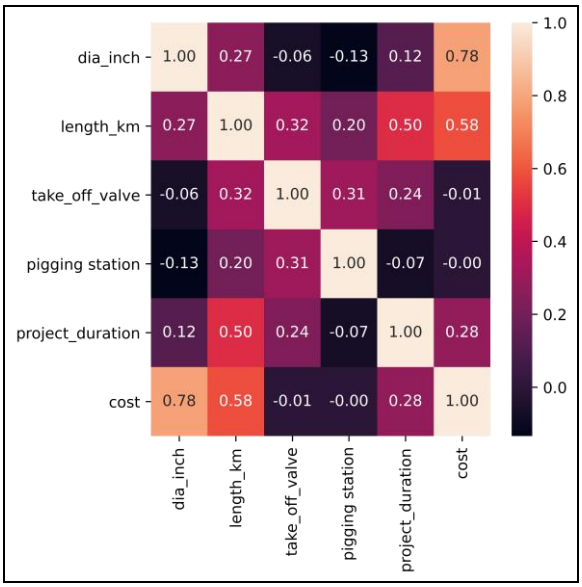


Figure 3: Heat-map of correlation between variables after parameter elimination.

Regression Analysis by Statsmodels Library

In this part of the research, a regression analysis were executed by using statsmodels library. Statsmodels library is one of the useful modules of Python for statistical models, hypothesis tests and data exploration (Seabold & Perktold, 2010). It is beneficial in terms of information it gives and the time it saves. Actually, exiguousness of data was a supreme challenge for this kind of an analysis. In this research, to have just about 50 data to be able to use in creating regression model is an example of that challenge. The more data actually the more learning and the more deep analysis. We conducted the operations under these conditions with statsmodels library by using least squares methods. After data preprocessing, dataset finalized and regression model was executed by statsmodels library and its functions. Codes showing how to create model with statsmodels are given in Figure 4.

```
import statsmodels.api as sm
import statsmodels.formula.api as smf

df.drop(df[df.cost > 60000000].index, inplace = True)
df1 = pd.DataFrame(df)
df1.drop(columns=["depth", "excavation", "line_valve"], inplace=True)
df2 = pd.DataFrame(df1)

X = df2.drop(columns =["cost"])
y= df2["cost"]
```

```
X = sm.add_constant(X)
lm = sm.OLS(y, X)
model = lm.fit()
model.summary()
```

Figure 4: Codes for creating model by statsmodels.

A regression model using ordinary least squares method was created by means of statsmodels. Results of this code and details of the final model, including important metrics such as R-squared, adjusted R-squared and Prob (F-statistic), are given in Table 4 and Table 5.

Table 4. Ordinary least squares regression results.

Dep. Variable:	cost	R-squared:	**0.777**
Model:	OLS	Adj. R-squared:	**0.752**
Method:	Least Squares	F-statistic:	31.39
No. Observations:	51	Prob (F-statistic):	1.30e-13
Df Residuals:	45	Log-Likelihood:	-876.66
Df Model:	5	AIC:	1765.
Covariance Type:	nonrobust	BIC:	1777.

Table 5. Ordinary least squares regression results.

parameters	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.347e+07	3.43e+06	-3.922	**0.000**	-2.04e+07	-6.55e+06
dia_inch	8.419e+05	9.48e+04	8.884	**0.000**	6.51e+05	1.03e+06
length_km	1.781e+05	3.76e+04	4.741	**0.000**	1.02e+05	2.54e+05
take_off_valve	-3.833e+05	2.46e+05	-1.557	0.126	-8.79e+05	1.12e+05
pigging_station	1.156e+06	2.16e+06	0.536	0.595	-3.19e+06	5.5e+06
project_duration	1266.9667	4433.852	0.286	0.776	-7663.269	1.02e+04

In first trial, independent variables except diameter and length had p-value higher than 0.05. It statistically means that those variables are inadequate to be able to explain the variability of target cost value. At this stage, an equation can be formed with the values in Table 5 and the model can be terminated, or the independent variables with high p-values can be removed from the model and a trial can be made again. On the purpose of presenting some real statistical result, those variables were removed from the dataset. Therefore, only diameter and length variables were left for the regression model. Results of this final trial were presented in Table 6 and Table 7.

Table 6. Ordinary least squares regression results.

Dep. Variable:	cost	R-squared:	**0.765**
Model:	OLS	Adj. R-squared:	**0.755**
Method:	Least Squares	F-statistic:	78.15
No. Observations:	51	Prob (F-statistic):	8.00e-16
Df Residuals:	48	Log-Likelihood:	-878.01
Df Model:	2	AIC:	1762.
Covariance Type:	nonrobust	BIC:	1768.

Table 7. Ordinary least squares regression results.

parameters	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.223e+07	2.41e+06	-5.071	0.000	-1.71e+07	-7.38e+06
dia_inch	8.549e+05	9.17e+04	9.324	0.000	6.71e+05	1.04e+06
length_km	1.689e+05	3.04e+04	5.555	0.000	1.08e+05	2.3e+05

As it can be seen from the Table 6 that created model has a $R^2 = 0.765$ value. It may be accepted as a not bad result specific to engineering problems. In the model, two independent variables have p-value less than 0.05, which means it is an appropriate choice to be able to explain the variability in dependent variable with those parameters. At this stage, as a PCE option that Eq. 1 emerged:

Cost ($) = -12226990.71 + 854865.97 × “Pipe Diameter” + 168936.70 × “Pipeline Length” (1)

where cost is in US dollars, pipe diameter in inch and length in km. To simplify the equation “D” can be used instead of Pipe Diameter and “L” for Pipeline Length. The revised Eq. 2 is given below.

Cost ($) = -12226990.71 + 854865.97 × D + 168936.70 × L (2)

The constant of the model was located in equation unpleasingly. Moreover, it is clear to see that some misestimations may come true because of such a huge constant value. Even if no one can say negative constants are awkward for regression analyses, it is normal to boggle because of having a constant lowering the result about $12M. It will be beneficial to look at the predicted costs and real costs. Some helpful statistics were presented in Table 8 to measure success of the final model. Explanations about details of the table are below the table.

Table 8. Real cost – predicted cost relationship.

Cost ($) = -12226990.71 + 854865.97 × D + 168936.70 × L		
Predictions	**Number of projects**	**Average real project costs of these projects ($)**
Negative cost results	6	1751854.27
Underestimation	25	11455554.97
Overestimation	26	17069053.52
Error range under $1M	12	6388981.00
Error range between $1M and $2M	5	16424855.22
Error range between $2M and $3M	7	9122743.09
Error range between $3M and $4M	8	3383372.35
Error range between $4M and $5M	2	24305072.25

The model made negative cost estimation 6 times as projected after creating the equation. Average real project costs of those six ones was about $1.75M and none of them higher than $2.4M. The model made 25 underestimation while it is 26 for overestimation. For 12 ones of 51 projects, model can be presented as successful because of estimation with an error under $1M. In addition, it estimated the cost with an error range between $1M and $2M for five projects while the average real project cost of which was nearly $16.5M. On the other hand, to have an error range between $3M and $4M while the average of real costs was just $3.83M

looks like a failure for PCE. Nevertheless, the model can be seen as partially successful in view of limitations and challenges about getting and arranging the data besides missing and lack of the useful data.

Conclusion

This research has been conducted to produce an equation for a PCE method of NGPC costs. Much as lack and acquisition of information tended to demotivate, the studies carried out in United States encouraged the authors about workableness of that subject. To prepare the dataset manually and missing of a number of values of cases were the main challenges of the research. After all this process, an equation was obtained that can reveal the NGPC cost estimates while the projects are still in the preparation stage, with a not bad coefficient of determination which is $R^2 = 0.765$. This study covered only a basic regression model. The proposed method can be used in the construction sector considering the statistics presented in last paragraph before conclusion part. In our future research, we intend to concentrate on increasing the number of the data and running analysis by machine learning techniques. In addition, in such large works in terms of cost, it would be desirable to conduct a study that will reveal how much of a mistake would be considered tolerable by the investor in a preliminary estimation that can be made with limited information.

References

Arage, S. S., & Dharwadkar, N. V. (2017). Cost estimation of civil construction projects using machine learning paradigm. In *Proc., International Conference on I-SMAC: IoT in Social, Mobile, Analytics and Cloud*.

Bai, B., Li, X., & Yuan, Y. (2013). A new cost estimate methodology for onshore pipeline transport of CO_2 in China selection and/or peer-review under responsibility of GHGT. *Energy Procedia*, *37*, 7633-7638.

Cheng, M. Y., & Hoang, N. D. (2014). Interval estimation of construction cost at completion using least squares support vector machine. *Journal of Civil Engineering and Management, 20*(2), 223-236.

Erdal, H. (2021). Prediction of pipeline projects construction costs utilizing machine learning techniques. In *Proc., International Marmara Sciences Congress* (pp. 218-223).

Kaiser, M. J. (2021). A review of onshore and offshore pipeline construction and decommissioning cost in the USA - part 1: specifications, cost estimation and onshore construction. *International Journal of Oil, Gas and Coal Technology, 27*(3), 247-285.

Kim, G. H., An, S. H., & Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment, 39*(10), 1235-1242.

Matthews, J. C., Allouche, E. N., & Sterling, R. L. (2015). Social cost impact assessment of pipeline infrastructure projects. *Environmental Impact Assessment Review, 50*, 196-202.

Parker, N. C. (2004). “Using Natural Gas Transmission Pipeline Costs to Estimate Hydrogen Pipeline Costs”, Institute of Transportation Studies, UC Davis.

Pyle, D., Cerra, D. D., & Kaufmann, M. (1999). Data preparation for data mining.

Rui, Z., Metz, P. A., & Chen, G. (2012). An analysis of inaccuracy in pipeline construction cost estimation. *International Journal of Oil, Gas and Coal Technology*, *5*(1), 29-46.

Rui, Z., Metz, P. A., Reynolds, D. B., Chen, G., & Zhou, X. (2011a). Regression models estimate pipeline construction costs. *Oil and Gas Journal*, *109*(14), 120-127.

Rui, Z., Metz, P. A., Reynolds, D. B., Chen, G., & Zhou, X. (2011b). Historical pipeline construction cost analysis. *International Journal of Oil, Gas and Coal Technology*, *4*(3), 244-263.

Seabold, S., & Perktold, J. (2010). Statsmodels: econometric and statistical modeling with Python. In *Proc., 9th Python in Science Conference.*

Sueri, M., & Erdal, M. (2022). Early estimation of sewerage line costs with regression analysis. *Gazi University Journal of Science*, *35*(3), 822-832.

Thaduri, R. K. (2012). Oil and gas pipeline construction cost analysis and developing regression models for cost estimation.

Uğur, L. (2007). Analysis of Construction Costs with Artificial Neural Networks. *PhD Thesis, Gazi University*.

Ugur, L. O., Kanit, R., Erdal, H., Namli, E., Erdal, H. I., Baykan, U. N., & Erdal, M. (2019). Enhanced predictive models for construction costs: a case study of Turkish mass housing sector. *Computational Economics*, *53*(4), 1403-1419.

Ulvestad, M., & Overland, I. (2012). Natural gas and CO2 price variation: impact on the relative cost-efficiency of LNG and pipelines. *International Journal of Environmental Studies*, *69*(3), 407-426.

Zhang, Q., Li, Z., Wang, G., & Li, H. (2016). Study on the impacts of natural gas supply cost on gas flow and infrastructure deployment in China. *Applied Energy*, *162*, 1385-1398.

Zhao, J. (2000). “Diffusion, Costs and Learning in the Development of International Gas Transmission Lines”, Interim Report, IR-00-054.