Developing a Data Mining Methodology for Oil and Gas Pipelines Cost Prediction

Philip-Mark P. Spanidis [(1)]

[(1)]ASPROFOS Engineering S.A.,

El. Venizelou 284, 17675, Kallithea, Greece, pspani@asprofos.gr

Abstract

Pipelines are means for transportation of oil and gas over long distanced areas. The estimation of the capital and operational expenditures of a new pipeline system is a multidisciplinary task of high complexity and risk. The paper demonstrates the practical problems in estimations of pipelines' cost and suggests the development of a cost prediction methodology applicable in the feasibility study of a pipeline project. The methodology refers to a process developed on the basis of data mining philosophy and dealing with the evaluation, collection and classification of pipeline cost data, structuring of a pipeline Knowledge Base (KB) and performance of a multiple regression analysis for determination of a pipeline cost prediction equation. The methodology constitutes a useful, easy in development and low cost tool for managers and financial analysts, which have to make advisements to decision making bodies for the technical and commercial feasibility of a pipeline project. Discussion on possible improvements and techno-economic views of the methodology along with concluding remarks are also presented.

Keywords: Data Mining, Knowledge Base, Pipeline, Regression.

1. INTRODUCTION

The energy pipelines are safe and environmentally friendly systems of transporting hydrocarbons (crude oil, gas, products, aviation fuels, etc.) over large geographical regions (Dey, 2002; 2006). In the phase of feasibility study, where the concept of a proposed pipeline system is under investigation, the estimation of capital and operational expenditures is a crucial issue. Managers, financial analysts, engineering experts and specialists from various disciplines of science and technology work in synergy to estimate the costs of pipeline route alternatives in order to advise the decision making bodies (energy supply companies, investors, public agencies, steering committees, etc.) to what extent the proposed project is cost effective or not. On the other hand, the limited availability of data significant for a substantial and accuracy cost approach, such as the pipeline size, geo-environmental and regulatory constraints, constructability, accessibility, etc. is a very usual problem with significant influence in the reliability and accuracy of cost estimations.

The objective of this paper is to suggest a methodology for the prediction of energy pipelines cost based on a data mining philosophy and supported by statistical analysis tools. The paper is structured as follows: Section 2 briefs the literature findings on pipeline cost analysis research, Section 3 analyzes the practical problems of cost estimation and addresses research questions, Section 4 describes the suggested methodology and its constituents (i) discovery and collection of pipeline techno-economic data from industry and literature, (ii) structuring of a Knowledge Base (KB) appropriate for organization, storing and use of collected data and (iii) application of the multiple regression analysis for determination of the pipeline cost equation, Section 5 discusses the results of the methodology along with improving proposals and techno-economical limitations and Section 6 concludes the whole study.

2. LITERATURE REVIEW

The investigation of pipelines cost concentrated the attention of researchers since early 1990s, where the rapid liberalization of energy markets in Eurasia and the challenging reformulation of international energy economy revealed new and challenging business opportunities. In parallel, players from countries (mainly) of the former Soviet Union (Azerbaijan, Kazakhstan and Turkmenistan) were steadily entered into the games of international energy competition (Klaasen, 1999). This context, in combination of new finds of oil and gas reserves worldwide has been working as an underline driver for funding and building new pipelines to satisfy the fastest growing international energy demand (CEDIGAZ, 1998; Ellsworth *et al*., 1999; Zhao, 2000; Kandiyoti, 2008).

Zhao (2000) investigated the microeconomic role of international gas transmission lines for the energy markets of the 21st century. The author assessed the share of the cost components in pipeline construction industry and the impact of technological learning and other factors on the costs of gas infrastructures development in Eurasia by drawing upon experience with pipeline development in the US. The author made an approach of indicative cost functions of exponential type $C(x)=aQ^{-b}$, where C, the pipeline unit cost in US$/m^3, Q, the cumulative capacity (mil.m^3), and a, b constant factors. Dey (2002; 2006) demonstrated the significance of pipeline costs as a factor of high impact in multiple criteria evaluations of alternative routes of large scale and capacity oil pipelines proposed in India. Rui *et al*., (2011) provided a reference for the pipeline construction cost by making an analysis of individual pipeline cost components with respect to historical cost data. The

authors analyzed the pipeline costs considering the diameter, length, capacity, year of completion and location of the route based on data collected from 412 pipeline projects constructed between 1992 and 2008. Oliver (2015) analyzed the cost, capacity, mileage and technical data for 254 natural gas pipeline projects constructed in the US over the period 1997–2012 and carried out an empirical analysis of natural gas pipelines expansion costs along with estimates of cost elasticities with respect to pipeline capacity and length. Rui *et al*., (2017) examined the cost overruns by investigating the performance of 200 public oil and gas pipeline projects in the US and documented that the error of cost underestimation is more frequent and greater than that of the overestimation. The authors documented also that the project performance varies in terms of project size, type, region, joint venture information and the year of the final investment decision.

The above review shows that the pipeline cost considerations depend on the nature, scope and target of the performing research. However, for an early financial analysis, the development of a cost equation enabling quick estimates for the capital expenditures of a pipeline project considering parameters like length, diameter, capacity and other, seems to be very practical and to this view, there is a window for further research.

3. PRACTICAL PROBLEMS AND RESEARCH QUESTIONS

Customarily, the cost-effectiveness of a pipeline project is investigated in the pre-investment phase, where several pipeline route alternatives are examined mainly in terms of safety, constructability, expandability, social and environmental acceptance, in order to be investigated if the new project is technically and financially feasible or not. This constitutes a multidisciplinary activity where cost estimators, in cooperation with managers, environmentalists, pipeline route engineers, safety engineers, geologists, process designers and material technology engineers exert intensive effort to make reasonable assumptions allowing substantial estimates of pipeline capital and operational expenditures. These assumptions are based on evaluation of baseline data (geographical and geological maps, satellite imagery and remotely sensed data, ecological surveys, land use maps, etc.) and reporting from field reconnaissance works carried out along the alternative pipeline route corridors. Beyond this work, there are diverse methods followed by cost estimators like brainstorming, proportionate studies based on cost data from past projects, experts' judgement and knowledge aggregation, risk analyses for "digesting" uncertainties due to lack of efficient technical evidence, budgetary estimates, analyses of alternative cost scenarios and/or combination of previous methods (see: Petley, 1997; Dysert, 2003; Whitesides, 2007).

Further to complexity, the aforementioned methods insert practical problems affecting the quality and inserting risks to the whole cost estimation framework. First, pipeline projects are different from each other and the experts engaged may not possess the appropriate background in providing sufficient judgement at all cost aspects. Second, the historical project files stored, usually, in the databases of engineering and consulting companies, may be voluminous, insufficiently updated or poorly formatted and, often, improvements of the database software, that are time and resource consuming, are required for making the stored material (re)usable and appropriate for application. Third, it is usual the effect of cost overestimations or underestimations, especially when the cost estimating philosophy is mainly based on subjective judgements and less on objective quantitative considerations and calculations. By considering the previous, several critical research questions are raised up:

- Is possible data of high reliability for pipeline costs to be collected, validated and used in a form and structure adequate to facilitate and/or improve the work of project cost estimators?
- Could a relational database be properly designed to provide pipeline data sets in combined patterns and multiple forms adequate for diverse cost estimation purposes?
- Is possible a database to be designed to constitute a Knowledge Base (KB) dedicated for pipeline projects and operating as a "learning machine"?
- Can these data be subject to processing with a mathematically/statistically consisted method?
- Is the determination of statistically validated cost prediction equation (combining various data sets fields, e.g. cost vs diameter, cost vs length, capacity vs cost, etc.) feasible?

The above questions are challenging in investigating the possibility for development of a more flexible method enabling efficient collection and storage of actual data for pipeline projects and calculation of project cost(s) using statistical methods suggested in literature as appropriate for reliable predictions of cost and substantial decision making.

4. METHODOLOGY

4.1 Data Mining: an Overview

The methodology suggested is based on Data Mining (DM) philosophy. This is because DM techniques and tools provide a substantial basis for collection, manipulation and exploitation of data in combination with statistical processing methods.

DM is of increasing interest for several scientific fields such as statistics, machine learning, database management and technologies for data visualization (Feelders *et al.*, 2000). DM has been

defined as a process of embodying tasks of data analysis and discovery algorithms in producing enumeration of data patterns and learning models (Fayyad et al., 1996). The objective of DM is the extraction of knowledge, meant as information organized and stored in an advance level of usability and accessibility, from voluminous data stored in files of heterogeneous structure, under different logical models and data manipulation languages (Hand, 1998). Literature shows that there are several DM methods: classification rules, clustering methods, decision trees, Bayesian models, neural networks and statistic regression analysis are the most widespread methods (Berry et al., 1997; Dunham, 2004: Witten et al., 2005). Regression is one of the mostly preferred methods in statistics and operations research that maps data items to a real-valued prediction variable (Hand, 1998; Breiman *et al*., 1984; Gylmour *et al*., 1997), expressed by equations obtaining predictive values of a dependent (or response) variable when other independent (or predictor) variable(s) are also varying accordingly. The regression equation provides the statistical model through which data are being processed using least square computational algorithms (Studenmund, 1992; Koop, 2006) thus giving knowledge to cost estimators. The engineering and consulting companies, that mostly undertake feasibility studies of various pipeline projects, would benefit the most from DM methods insofar information collected from previous projects and validated sources can be properly managed and processed through a regression model that can give mathematically substantial cost prediction results.

4.2 The Data Mining Process

The adopted methodology refers to a typical DM process advised by Feelders *et al*., (2004). **Figure-1** depicts the DM methodology in form of a stepwise algorithmic process, including primary tasks (items 1, 2, 3, 5, 8, 9), decision nodes (items N1, N2 and N3) and a number of secondary tasks required for the execution of interim feedback loops. The methodology comprises six (6) main tasks (steps): (a) *Problem Definition*, (b) *Acquisition of Background Knowledge*, (c) *Selection of Data*, (d) *Pre-processing of Data*, (e) *Analysis and Interpretation* and (f) *Reporting and Use*. The descrption of the main tasks is as follows:

(a) *Problem definition*: critical questions, such as "*what type of project information is appropriate in performing pipeline cost estimations*?" or "*which DM model should be the most convenient for the needs of cost estimation*?" are examined. By making a literature review (O&GJ, 1997; Zhao, 2000) and in collaboration with cost engineers, it was assumed that the greater length and the longer pipeline diameter, the higher cost for engineering,

erection and material procurement. Thus, cost C, being a non-negative function of length L and diameter D, can be expressed by the general form of linear equation F(C):

$$F : \mathbb{R}^2 \rightarrow \mathbb{R};\ D>0,\ L>0,\ F(C)>0;\ \beta_0,\ \beta_1,\ \beta_2 \text{ and } \varepsilon \in \mathbb{R};\ F(C) = \beta_0 + \beta_1\varphi(D) + \beta_2\mu(L) + \varepsilon,$$

where b_0, b_1 and b_2 are constant factors, F(C), φ(D) and μ(L) are functions of C, D and L and ε is the regression error term. $\mathbb{R}$ denotes the set of real numbers;

(b) *Acquisition of Background Knowledge*: the type and extent of required data are specified and the sources of research are identified. Geopolitical corridors, geographical classification, type of transporting product, stakeholders, linear unit cost(s), etc. are some of the background knowledge areas that are useful, not only for cost estimators, but for other specialists involved with the pipeline engineering and management issues;

(c) *Selection of data*: information published in international construction reviews regarding pipelines implemented, or under development, is selected through a cross sectional investigation. In turn, a KB of pipeline projects is designed enabling the storage and organization of this information;

(d) *Pre-processing of data*: homogenization, validation and grouping of cost datasets; data fields containing values of the parameters C, L and D are isolated and regrouped in suitable visualization form (matrix and scatter plots, logarithmic visualization, etc.) enabling prescreening of the data model that seems to be mostly appropriate before the regression analysis is performed;

(e) *Analysis and Interpretation*: the prepared data sets are introduced into the multiple regression statistical software package MINITAB and the cost prediction equation is outlined, along with the calculation report describing the regression output for interpretation and evaluation of the quality of cost equation;

(f) *Reporting and Use*: the testing of regression analysis quality deals with the examination of values of critical statistical parameters regarding, for example, fitting of model to data, variance analysis, multi-collinearity effects, etc. Eventually, the outlined equation is outlined for use, under the limitations of confidence intervals of the statistical analysis.

4.3 Formulating the Knowledge Base

The KB has been designed and implemented as a simple relational database. Each record contains ten (10) data fields: project number, pipeline length (L), overall cost (C), linear cost (C_L), diameter (D), capacity (Q), service mode, project owner/stakeholders, geopolitical classification and source

of information. The *project number* is the primary key of each record and it is consisted of two parts containing four (4) characters each one: an abbreviation showing the geographical classification of every single project (e.g. USCA for US-Canada, EURO for Eurasia, AFRC for Africa and so forth) and the serial number by which any project is entered into the KB. For example, the code of a project classified in the 45th entry for North America pipelines is 'USCA0045'. The second field, *length*, represents the entire pipeline length (in km). The *overall cost* shows the total project cost (millions US$), while the *linear cost* expresses the project cost divided per km of the pipeline route (millions US$/km); *diameter* refers to the nominal pipe size (in inches); *capacity* expresses the pipeline transportation throughput in annual basis (e.g. $mcft^3$, m^3, US gals, bbl, tons etc.). The field *service* specifies the product to be transferred (oil, gas, products, etc.) and the *technology* of installation, refers to the pipeline installation as an onshore or an offshore line. The filed *project owners* refers to consortia owing, planning, constructing and/or operating various pipelines, while the filed geopolitical classification describes the countries crossed by the corridor of a certain pipeline. The data field *source of information* refers to the documentation (e.g. Pipeline & Gas Journal, Oil & Gas Journal, Offshore magazine, etc.) or the web site from which the baseline information is collected. In the KB, 775 major pipeline projects launched, contracted or planned in the period from 1998 up to 2014 have been registered and **Table-1** shows a sample of a records set stored in the KB.

PROJECT NUMBER	LENGTH (Km)	COST (Mil USD)	LINEAR COST (1000 US$/Km)	NOMINAL PIPE SIZE (inches)	CAPACITY	SERVICES	PROJECT OWNER	GEOPOLITICAL CORRIDOR	DATA SOURCE
AFRC0004	47	35	745	14	503 000 cm / hr	Gas Transmission (Offshore)	Nigerian Gas Co	Nigeria-(1)	ILF-Selected References / 05
AFRC0043	1072	3500	3265	30	225000 bpd	Crude Oil	COTCO-Cameroon Oil Transport	Chad-Cameroon	P&GI (9 / 01)
CSAM0022	118,4	150	1267	24		Gas Transmission	BP / Shell	Gulf of Mexico (Na Kika)-US (Destin)	P&GI (7 / 01)
CSAM0075	677	1450	2142	28		Gas Transmission	Transportadora de Gas del Peru	Peru (Camisea-Lima)	P&GJ (8 / 02)
CSAS0024	677	1000	1477	48	8 bcmy	Gas Transmission	Sotuh Caucasus Pipeline (SCP)	Azerbaijan (Baku)-Turkey (Erzerum)	P&GJ (05 / 03)
CSAS0034	1112	1800	1619	52	16 bcmy	Gas Transmission	Trans-Caspian Pipeline Consortium-(2)	Turkmenistan-Caspia Sea-Azerbatijan-Georgia-Turkey	Alexander's Gas & oil Connections, Vol.5 / 07.04.02
EURO0045	47	50	1064	28	490 000 cmhr	Gas Transmission	OMV / TAG WG	Hungary-Austria	ILF-Selected References / 04
EURO0102	215	100	465	16	2,5 Mty	Oil Transmission	ELPET Valkaniki	Greece-FYROM (Thessaloniki-Skopje)	ILF-Selected References / 04
FEST0086	1355	1200	886	30	500 MMcfd	Gas Transmission	Unoacal	India (N. Delhi)-Bangladesh (Bibiyana)	P&GJ (8 / 02)
FEST0091	3900	8500	2179	42	20 bcmy	Gas Transmission	Petrochina (West to East Pipeline)	China (Lunan-Shanhgai)	O&GJ (16.02.04) & China Facts and Figures 2002
MDES0019	248	190	766	26	10 bcmy	Gas Transmission	Egypt-Jordan Goverments	Egypt (Arish Taba)-Jordan (Aqaba)-(1)	Embassy of the Arab Republic of Egypt
MDES0033	720	250	347	22	7,5 Mty	Trunkline, Oil (+ 1PS)	INOC, Kingdom of Jordan	Iraq-Jordan	Alexander's Gas & Oil Connections 2000
SPFC0012	130	568	4369	42	3,85 MMscdf	Gas Transmission	-	Australia (NWS)	O&GJ (01.03.04)
SPFC0032	512	400	781	26		Trunkline, Gas	Chevron Asiatic	Papua (N. Guinea)-Queensland (Australia)	P&GI (9 / 01)
USCA0044	28	57	2036	48		Gas Transmission (+ 1CS)	Union Gas LTD	SW Ontario	P&GI (9 / 01)
USCA0228	608	425	699	36	730 MMcdf	Gas Transmission		US (Cheyenne-Greensburg)	O&GJ (08.03.04)

Table-1: Sample of data records stored in the Knowledge Base (KB)

4.4 Regression Analysis

The regression analysis was performed on a sample of 63 gas pipeline projects. This was a selective approach since the project records stored in the KB were not fully contained actual data of the main parameters of cost, length and diameter. Moreover, data heterogeneities made the use of stored datasets quite problematic. For example: there were projects with two or more different diameters along the same line (e.g. 24-inch/30-inch); capacity was referred in a variety of units (eMtoe, bcmy, Mty, etc.); in some projects the exact mileage or the total cost was missing; in other projects the cost comprised installation of compressors or pump stations while in other similar projects there was not any relevant clarification. Therefore, the statistical research has been focused in a sample of 63 projects including complete data of a three elements vector [Ci, Di, Li]; $1< i < 63$; $i \in \mathbb{N}$, where $\mathbb{N}$: the set of natural numbers.

Before the regression is performed a pre-processing test of Pearson correlation coefficient (R) among the dependent and independent variables was performed based on preliminary scatter plots, in order to inspect visually if possible relationship among data of C versus L and C versus D exists.

The first tests shown that the variable C has been found to be strongly correlated to L and significantly correlated to D. However, the produced distribution scatter plots, due to non-linearity, presented poor visualization defects. After carrying out several trials and taking common logarithms of C and L, the Pearson coefficients shown that the correlation of log(C) vs log(L) and log(C) vs D derives values R=0.786 and R=0.638 respectively.

The final scatter plots depicting the logarithmic values of pre-processed data are shown in **Figures 2 and 3** (sample size: 264 projects). In turn, the vector consisted of the triplet of parameters Log(Ci), Di and Log(Li) inserted to the regression program and the general linear regression model has been formulated in a double-log functional form (Studenmund, 1992):

$$\text{Log}(Cp) = \beta_0 + \beta_1 D + \beta_2 \text{Log}(L)$$

where, Log(C) is the logarithm of predicted cost, C, and β_0, β_1, β_2 the predicted partial regression coefficients. The cost equation derived by the regression process is the following:

$$\text{Log}(C) = - 0{,}448+0{,}018*D+0.935*\text{Log}(L)$$

The regression report is hereunder presented describing the results justifying the quality of the regression process outcome. In particular:

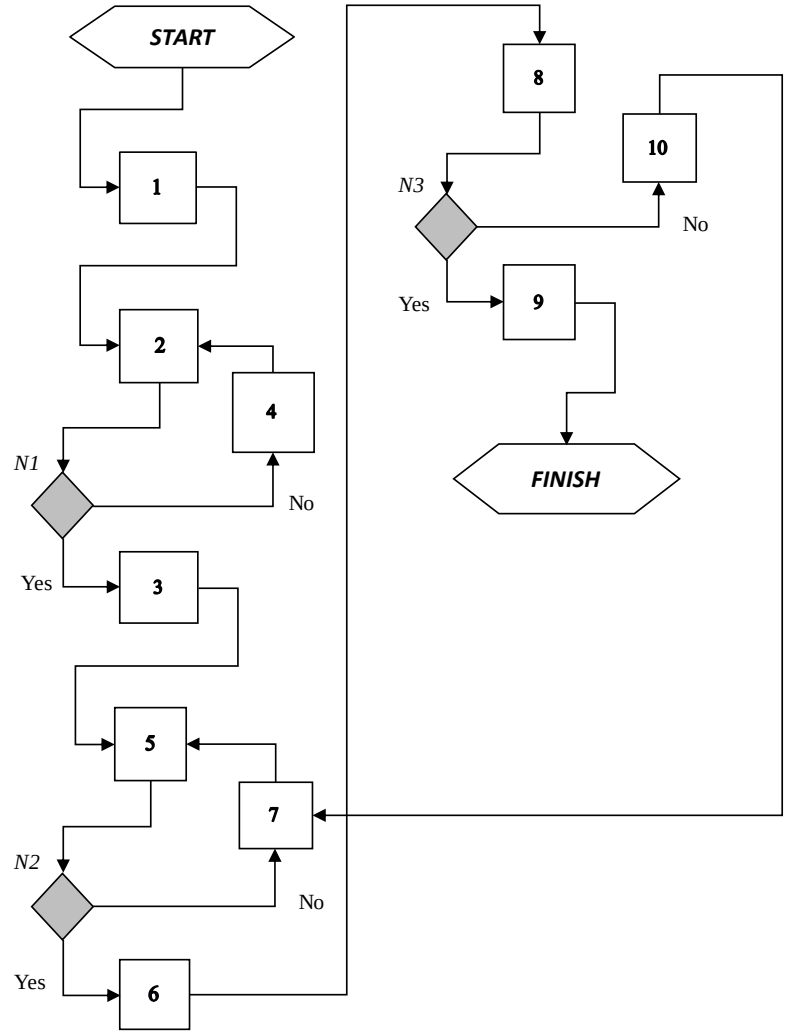
(a) The coefficient of determination R-Sq=85.70% and its adjusted value R-Sq(adj)=85.60% show that the linear model fits the data well and the predictors β_0, β_1 and β_2 can explain sufficiently the variance of the pipeline project cost, when log(L) and D vary accordingly;

(b) In the Analysis of Variance, the value P=0.000 shows that the model is significant at a level of 5% (0.05) and at least one of the constant factors is different from zero (the null hypothesis Ho: $\beta_0=\beta_1=\beta_2=0$ is rejected and the hypothesis Ha: $\beta_0 \neq 0 \vee \beta_1 \neq 0 \vee \beta_2 \neq 0$ is accepted);

(c) The parameter F is much higher than 0: F=1034.42>>P=0.000 that means that at least one of the independent variables D and log(L) has an effect on the dependent variable Log(C);

(d) The Variance Inflation Factors (VIF) for the prediction variables D and Log(L) are equal to 1.2. As $VIF(\beta_1)=VIF(\beta_2)=1.2<5$, the model does not present multi-collinearity effects (Witten et al., 2005; Studenmund, 1992) and thereof, no linear relation exists between D and Log(L) variables;

Figure-1: The Data Mining Methodology



Description of Tasks and Decision Nodes

START

1 Problem Definition (critical and secondary cost parameters)

2 Acquisition of Background Knowledge (from the web)

N1 Is Collected Knowledge sufficient?

3 Data Selection Design/Development of Knowledge Base (KB)

4 Further Selection of Data

5 Pre-Processing of data

N2 Are there any correlation trends (visualization)?

6 Data Refinement

7 Further Investigation of Data relationships

8 Data Analysis and Interpretation (Regression)

N3 Are Regression results acceptable?

9 Reporting and Use

10 Repetition of relationships investigation

FINISH

i = interim task

Ni = decision node

Starting/Terminating node

Workflow

Further to the regression report, the plot of the residuals is also outlined. The plot of residuals (differences between the predicted values Log(Cpi) vs the input values Log(Ci)) and the residuals histogram showing an almost normal distribution of residuals around the zero value.

In conclusion, the produced pipeline cost prediction equation (a) presents an acceptable, in general, fit with the data sets of C, L and D paarmeters, (b) satisfies the statistical acceptability hypotheses for the value of the predictors β_0, β_1 and β_2, and (c) is statistically consistent as the independent variables are not linearly related each other.

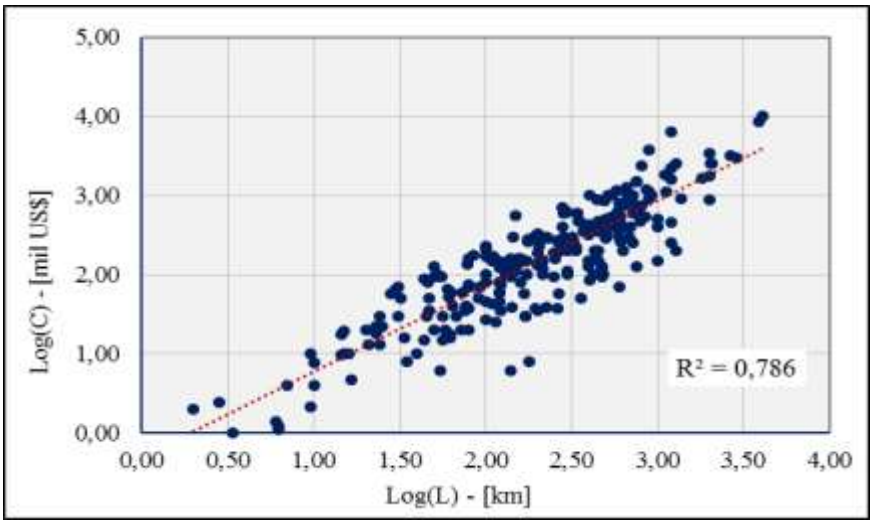


Figure-2: Scatter Plot of Log(C) vs Log(L)

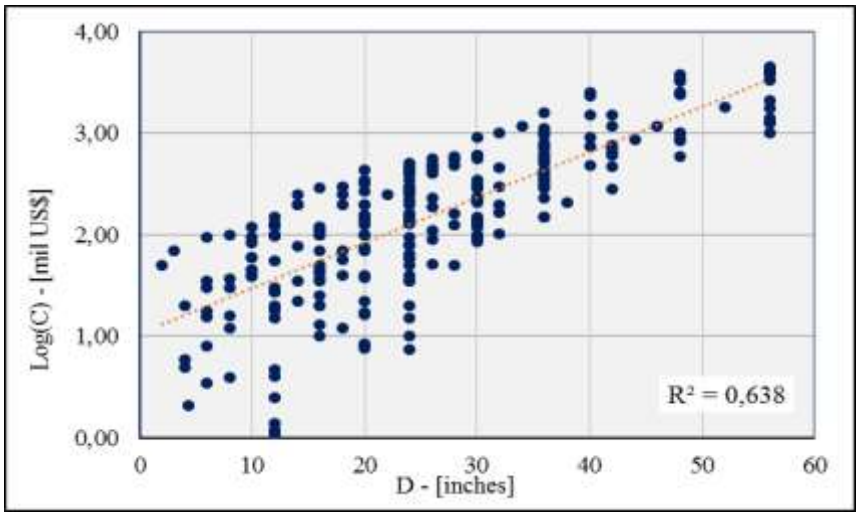


Figure-3: Scatter Plot Log(C) vs D

5. DISCUSSION

The regression equation might be used as a tool for preliminary estimation of pipeline projects cost on supplementary basis and not substituting at all other methods or practices. As indicative, **Table-2** shows the application of regression equation in estimating cost of gas pipelines (in millions of US$) with dimeters varying from 10-inch to 40-inch and lengths from 100 to 1.000 km. Following the methodology proposed herein, a learning context can steadily be established insofar the exploitation of actual pipeline cost data can be processed providing statistically substantial results, useful for project managers, and further, for decision makers. However, the methodology, presents certain strengths and weaknesses. In particular:

Strengths:

(a) Demonstrates that critical techno-economical parameters (like C, D and L and, possibly, others) show considerable trend of correlation and, therefore, are statistically exploitable;

(b) Constitutes a practical solution, especially when historical cost data are of low homogeneity, poorly maintained, less accessible or totally missing;

(c) The produced results, although generic, offer a substantial basis to cost estimators in making reasonable cost scenarios with statistically confident conditions;

(d) It can be easily developed as it requires low cost software and short term training;

Weaknesses:

(a) The parameters used are few, therefore, the pipeline cost prediction is of limited objectivity;

(b) The costs stored in the KB, cover a long time period (1998-2014), therefore, the application of corrective factor for establishing a basis of common reference of time is required;

(c) The results produced by the cost prediction equation have to be further rationalized using risk analysis tools for balancing various uncertainties of the whole method;

(d) The KB design can be furtherly improved enabling more effective manipulation (grouping, classification, clustering etc.) of the recorded datasets;

In a wider view, more parameters, beyond C, D and L, should be taken into account in the cost analyses such as: (i) the “country risk” classified in various levels of significance (low, medium or high) according to the geopolitical intension of the involved countries, (ii) the geomorphology of the pipeline route corridor divided as in smooth, hilly or mountainous terrains, (iii) the annual inflation rates, if available, that affect the engineering, procurement and construction costs of pipelines and (iv) the major geo-environmental constraints along the pipeline routes. The content, the data type and the number of parameters that could be potentially introduced may affect the

accuracy of cost estimation. To improve the performance of cost prediction equation, further testing is required by applying different regression models and choosing the one that fits better to the recorded data. Notwithstanding, experience shows that since the uncertainty of costs in feasibility studies is undoubtedly high, the estimations are moving through tolerances varying from ± 20% up to ± 40% and the result of their effort depends on the reliability of project design data, if available, and the geo-environmental and social factors of regions through which the transmission pipelines are conceptually designed.

The design of the KB is another critical constituent of the proposed methodology. The KB, beyond the regular updating it requires, cannot be merely seen as an instrument for mechanistic accumulation of data from past projects. Instead, KB has to operate as a learning instrument that must be integrated within the corporate system of engineering and consulting companies aiming to enhance the performance of knowledge acquisition and management in this type of organizations. To this regard, the KB design can be furtherly improved enabling more effective manipulation (grouping, classification, clustering etc.) of the recorded datasets. Moreover, the KB should be extended beyond the 10 data field records described herein and supported by a database management system enriched with advanced relational rules enabling the integration of processing capabilities among the attributes of stored datasets. For this reason, the ANSI/SPARC Architecture, which is a standard for a Database Management System (DBMS) development should be adopted as enabling development of an advanced SQL (structured query languages) software that will increase the KB overall performance.

Finally, from the techno-economic point of view, two main categories of corporate costs, due to the DM methodology development can be distinguished: the cost for development of an integrated DM system, C_{DMD}, and the cost of the risk due to low quality cost estimates, C_{CPR}. **Figure-4** depicts a tradeoff diagram of balancing those costs, where the increase of C_{DMD} relates to decrease of C_{CPR} and vice versa. The developers of the DM system (design of the KB, adaptation of statistical tools, updating functions, resources, etc.) have to make an upfront adding value assessment to evaluate to which extent the DM system is necessary and appropriate to be developed keeping an optimum balance between two costs, so that the total cost, $C_{TOTAL} = C_{DMD} + C_{CPR}$ to take the minimum value. This is a critical issue for an engineering and consulting company operation costs and budgetary limitations when planning the development of an integrated DM system.

Length [km]	Diameter [inches]			
	10	20	30	40
100	39,99	60,53	91,62	138,68
200	76,47	115,73	175,17	265,13
300	111,71	169,09	255,92	387,35
400	146,19	221,27	334,91	506,90
500	180,11	272,61	412,61	624,51
600	213,59	323,27	489,30	740,58
700	246,70	373,39	565,15	855,39
800	279,50	423,05	640,31	969,15
900	312,04	472,30	714,85	1081,97
1000	344,35	521,19	788,86	1193,99

Table-2: Pipeline Costs prediction

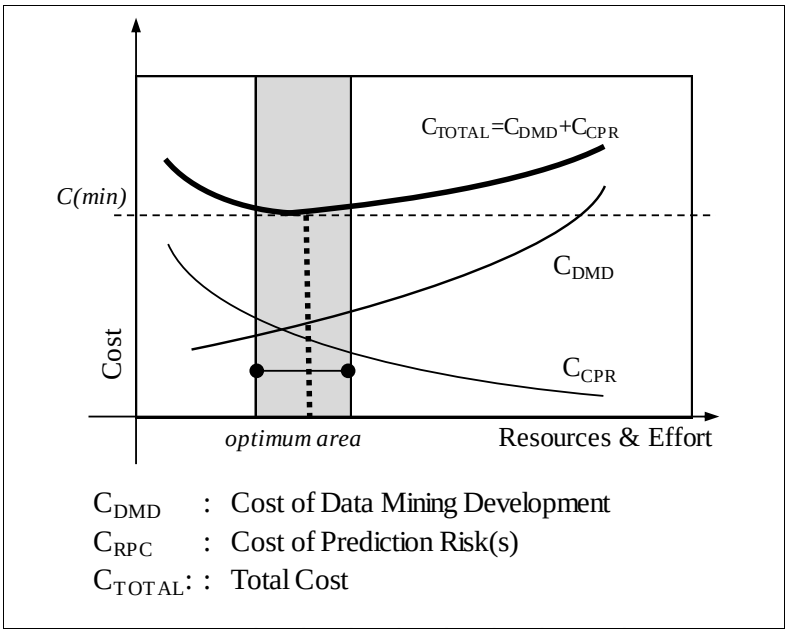


Figure-4: Tradeoff curve of the DM development cost vs Pipeline cost prediction risk

6. CONCLUSION

The proposed methodology suggests adoption of DM philosophy and offers a substantial basis for discovering, collecting and classifying background information from the construction industry of oil and gas pipelines. This information can be stored in a suitably structured KB allowing access and retrieval of data sets containing values of critical techno-economic parameters like C, D and L. In turn, the data sets can be modelled and processed by a multiple regression software for defining a cost prediction equation appropriate for primary predictions of pipeline project costs. The

methodology is an easily developed tool of low cost and sufficient usability, especially in cases where reliable or poor quality pipeline cost data are not available. The analysis of strengths and weaknesses of the methodology shows that there are windows for further research and improvement of its objectivity in terms of statistical reliability and upgrading of the KB design to constitute an integrated knowledge base for utilization of pipeline cost data in more advanced mode for covering needs of cost estimators and requirements of experts from various pipeline engineering and management disciplines.

7. REFERENCES

1. Batzias, F., Spanidis, P. - M., (2008). Bridging Knowledge Gaps in Engineering Companies - The case of Pipeline River Crossings in Greece. *Proceedings 8th Intern. Joint Conf. Knowledge Based Software Engineering (JCKBSE)*, Piraeus, Greece
2. Berry, J. A., Linoff, G., (1997). Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley & Sons, Inc. NY
3. Breiman, L., Friedman, J. H., Olsen R. A., Stone, C. J., (1984). Classification and Regression Trees, Wadsworth
4. CEDIGAZ (1998). Natural Gas in the World: 1998 Survey, Paris
5. Dey, P. K., (2002). An integrated assessment model for Cross Country pipelines, *Environmental Impact Assessment Review*, vol.(22), pp.703-721
6. Dunham, M. H., (2004). Data Mining-Introductory and Advanced Topics, Pearson Education Inc. Prentice Hall
7. Dysert, L.R., (2003). Sharpen your Cost Estimating Skills, *Cost Engineering* vol.45(6), June, pp.22-30
8. Ellsworth, C., Wang, R. (1999). China's Natural Gas Industry awakening, poised growth, *Oil and Gas Journal*, July, pp.23-28
9. Fayyad, U., Piatetsky-Shapiro G., Smyth, P., (1996). From Data Mining to Knowledge discovery in data bases, *AI Magazine*, vol.17(3), pp.37-54
10. Feelders, A., Daniels, H., Holsheimer, M., (2000). Methodological and practical aspects of Data Mining, *Information & Management*, vol.(37), pp.271-281
11. Glymour, C., Madigan, D., Pregibon, D., Smyth, P., (1997). Statistical themes and lessons for data mining, *Data Mining and Knowledge Discovery*, vol.1(1), pp.11–28.
12. Hand, D. J., (1998). Data Mining: Statistics and More? *The American Statistician*, vol.52(2), pp.112-118

13. Kandiyioti, R., (2008). Pipeline-Flowing Oil and Crude Politics, I.B. Tauris & Co Ltd
14. Klaasen, G., Gruber, A., Schrattenholzer, L., (1999). Towards New Energy Infrastructures in Eurasia: a background paper, *IIASA*, IR-99-17
15. Koop, G., (2006). Analysis of Financial Data, John Wiley and Sons, Ltd
16. Nanhay S., Ram S. R., Chauhan, R.K. (2012). Data Mining With Regression Technique, *Journal of Information Systems and Communication*, vol.(3), Issue 1, pp.199-202
17. Oil & Gas Journal, (1997). August 4, pp.37-58
18. Oil & Gas Journal, (1999). June 5, pp.63-66
19. Oliver, M., (2015). Economies of scale and scope in expansion of the US natural gas pipeline networks, *Energy Economics*, vol.(52), part-B, pp.265-276
20. Petley, G.J., (1997). A method for estimating the Capital Cost of Chemical Process Plants, PhD Thesis, Loughborough University
21. Rui Z., Chen, P. M. G., Reynolds, D., Zhoua, X., (2011). Historical Pipeline Construction Cost Analysis, *International Journal of Oil, Gas and Coal Technology*, vol.(3), pp.244-263
22. Rui Z., Peng, F., Ling, K., Chang, H., Chena, G., Zhoua, X., (2017). Investigation into the performance of oil and gas projects, *Journal of Natural Science and Engineering*, vol.(38), pp.12-20
23. Spiegel, M. R., (1975). Probability and Statistics, McGraw-Hill, New York
24. Studenmund, A. H., (1992). Using Econometrics-A Practical Guide/Book and Disk, Harpecollins College Division
25. Whitesides, R.W., (2007). Process Equipment Cost Estmation by Ration and Propoprtion, PDH Course 127, www.PDHonline.org
26. Witten I. H., Frank, E., (2005). DATA MINING - Practical Machine Learning Tools and Techniques, ELSEVIER
27. Yaghini, M., (2010). Data Mining, Part 5. Prediction, 5.7 Regression Analysis (available in the web)
28. Zhao, J. (2000). Diffusion, Costs and Learning in the Development of International Gas Transmission Lines, *IIASA*, IR-00-054