Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 55 (2016) 41 – 46

www.elsevier.com/locate/procedia

5th CIRP Global Web Conference Research and Innovation for Future Production

An approach for selecting cost estimation techniques for innovative high value manufacturing products

Oliver Schwabe[a*], Essam Shehab[a] and John Erkoyuncu[a]

[a]*Cranfield University, Building 50, Cranfield, Bedford MK43 0AL, U.K.*

* Corresponding author. Tel.: +49 (0) 170 9053671;. *E-mail address:* o.schwabe@cranfield.ac.uk

Abstract

This paper presents an approach for determining the most appropriate technique for cost estimation of innovative high value manufacturing products depending on the amount of prior data available. Case study data from the United States Scheduled Annual Summary Reports for the Joint Strike Fighter (1997-2010) is used to exemplify how, depending on the attributes of a priori data certain techniques for cost estimation are more suitable than others. The data attribute focused on is the computational complexity involved in identifying whether or not there are patterns suited for propagation. Computational complexity is calculated based upon established mathematical principles for pattern recognition which argue that at least 42 data sets are required for the application of standard regression analysis techniques. The paper proposes that below this threshold a generic dependency model and starting conditions should be used and iteratively adapted to the context. In the special case of having less than four datasets available it is suggested that no contemporary cost estimating techniques other than analogy or expert opinion are currently applicable and alternate techniques must be explored if more quantitative results are desired. By applying the mathematical principles of complexity groups the paper argues that when less than four consecutive datasets are available the principles of topological data analysis should be applied. The preconditions being that the cost variance of at least three cost variance types for one to three time discrete continuous intervals is available so that it can be quantified based upon its geometrical attributes, visualised as an n-dimensional point cloud and then evaluated based upon the symmetrical properties of the evolving shape. Further work is suggested to validate the provided decision-trees in cost estimation practice.

Peer-review under responsibility of the scientific committee of the 5th CIRP Global Web Conference Research and Innovation for Future Production

Keywords: Computational complexity; Cost estimate uncertainty; Method selection

1. Introduction

This paper presents an approach for determining the most appropriate technique for cost estimation of innovative high value manufacturing products depending on the amount of a priori data available [1]. High value manufacturing products are understood as such products which are the results of "…the application of leading edge technical knowledge and expertise…" and result in "…the creation of products, production processes, and associated services which have strong potential to bring sustainable growth and high economic value…" [2]. Exemplary high value manufacturing products are aerospace engines and airframes, sea vessels and defence ground vehicles. Innovation is declared to exist when no verified and accurate cost models are available. Under the associated conditions of deep uncertainty and small data it is especially the plethora of varying plausible future scenarios that calls for a deeper understanding of available methods. Uncertainty is defined as unintended cost variance with a probability of 100% and an unknown impact [3]. The presented approach provides guidelines for when the following cost estimation techniques can be applied with confidence: analogy or expert opinion, topological estimation, parametric estimation and standard statistical regression. The cost estimation technique based upon topological estimation is novel in the field of cost estimation and deemed suitable for

Peer-review under responsibility of the scientific committee of the 5th CIRP Global Web Conference Research and Innovation for Future Production
doi:10.1016/j.procir.2016.08.017

filling the gap between zero and four discrete time elements of prior data [4]. The guidelines are based upon the complexity of a priori data as defined by Kolmogorov [5] and applied to an evolving bit string describing the propagation direction of cost variance metrics deemed relevant by the estimator (i.e. the direction of cost variance propagation).

Section 2 describes how the principles of Kolmogorov complexity are used to determine the time windows for differing forecast techniques and Section 3 presents the case study data. Section 4 presents the fundamental selection guidelines and Section 5 presents the process for determining which forecast technique is most applicable when. Section 6 discusses the concept of innovativeness and Section 7 shares a series of common estimation situations and describes how the presented approach leads to specific forecast technique recommendations. Section 8 shares a conclusion and provides recommendations for future research.

2. Complexity for determining time windows

It is suggested that Kolmogorov complexity [5] is a suitable indicator for determining when a specific forecasting technique is applicable or not. The metric of Kolmogorov complexity signifies the degree of compression a binary string can be subject to whereby compression is understood as the process of converting the sequence of bits into the description of the pattern represented by that bit sequence. The bit sequence is hence transformed into a program that can generate exactly that bit sequence. The program consists of a descriptor language which explains how a sequence of instructions is applied by a Turing Machine in order to generate the bit string.

The data of interest is the prior data related to the financial cost variance of the high value manufacturing product, specifically across at least three dimensions of cost variance such as cost changes due to variance in requirements, schedule or units ordered. This data needs to cover iterative discrete time intervals prior to the point in time where the cost estimate is being performed.

The first boundary suggested by Kolmogorov complexity is that the data from at least 42 discrete time intervals is required before pattern recognition approaches can be applied for forecasting purposes. This includes the application of standard regression techniques [5].

The second boundary suggested by Kolmogorov complexity is that depending on the length of the bit string the actual complexity score of individual bit strings can be grouped into groups of identical complexity. Bit strings of length one or two have the same complexity group. It is first the bit strings with a length of three elements that demonstrate this behavior. It is then with the fourth element that a first determination of stability can be made. The authors therefore suggest that while at time interval zero no techniques other than analogy or expert opinion are feasible, starting with the fourth time intervals parametric models (that depend on an understanding of cost estimating relationships) are applicable. From one to three elements a gap exists that the authors propose to fill with the technique of topological estimation.

3. Case study data

The data selected for exemplifying the selection technique is drawn from the United States Scheduled Annual Summary Reports [6] for the Joint Strike Fighter in the time period of 1997 to 2010 as illustrated by Table 1.

Table 1. Case study data

Base-line	Year	Quan-tity	Sche-dule	Engin-eering	Estim-ating	Oth-er	Support
1994	1997	0	0	0	140	0	0
1994	1998	0	0	1121	105	0	0
1994	1999	0	0	1121	105	0	0
1994	2000	0	0	1121	105	0	0
2002	2001	0	19	5452	7438	0	0
2002	2002	16249	0	2458	2330	0	2595
2002	2003	16249	8024	4370	17153	0	2735
2002	2004	16249	8139	7940	11998	0	5092
2002	2005	16249	8208	8279	17838	0	8054
2002	2006	16249	8797	9687	18849	0	11218
2002	2007	16249	8797	9687	30738	0	58
2002	2008	16249	8797	9687	30738	0	58
2002	2009	16119	8797	9687	52380	0	6753
2002	2010	16119	8797	9687	77984	0	13151

The “baseline” is the year in which the technical baseline estimate was created. The “year” is the discrete time period for which accounting data was available. The factors “quantity” through “support” are the reasons for cost variance assessed in the accounting period and the numbers entered represent the total absolute financial variance in USD$ million as compared to the baseline estimate. This variance may represent cost increases or reductions whereby the focus of the research is absolute variance from target.

As illustrated in Table 2 the case study data is now analysed to determine whether the cost variance for an accounting time period is higher (“1”), lower (“0”) or equal (“0”) to the previous time period. If the year of the baseline estimate changes a “1” is also assigned.

Table 2. Case study data

Year	Quan-tity	Sche-dule	Engin-eering	Estim-ating	Other	Sup-port	Total
1997	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1998	0	0	0	0	0	0	1
1999	0	0	0	0	0	0	0
2000	0	0	0	0	0	0	0
2001	1	1	1	1	1	1	1
2002	1	0	0	0	0	1	1
2003	0	1	1	1	0	1	1
2004	0	1	1	0	0	1	1
2005	0	1	1	1	0	1	1
2006	0	1	0	1	0	1	1
2007	0	0	0	1	0	0	1
2008	0	0	0	0	0	0	0
2009	0	0	0	1	0	1	1
2010	0	0	0	1	0	1	1

The string for any cost variance factor can now be examined for relevant changes in complexity groups. The complexity group is based on using a three bit sliding window along an exemplary bit string. For exemplary purposes the string for total change is used therefore "1001111111011" as illustrated in Table 3. The complexity score is calculated using http://www.complexitycalculator.com and categorized into complexity groups [5, p. 30, table 5]. Complexity group 1 corresponds to a complexity score of 5.40 for a three bit string. Complexity group 2 corresponds to a complexity score of 5.45 and complexity group 3 to a complexity score of 5.51.

Table 3. Case study data exemplary bit string analysis

Three bit string	Complexity	Complexity group
100	5.45	2
001	5.45	2
011	5.45	2
111	5.40	1
111	5.40	1
111	5.40	1
111	5.40	1
110	5.45	2
101	5.51	3
011	5.45	2

These changes in complexity groups are illustrated in Figure 1. Each change in complexity group indicates a potential need for (re-) selecting a cost estimation technique.

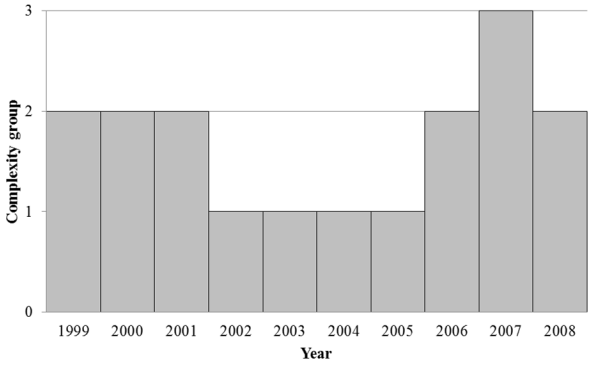

Fig. 1. Cost estimate (re-) selection points

4. Selection guideline

The previous section has discussed how complexity groups can be identified with the minimum prior information of three consecutive bits as examined using a three bit sliding window approach. The moment the complexity group changes the minimum a priori data counter must be reset to "0" and the recommended estimation technique is re-evaluated as indicated in Table 4. In essence the moment a complexity group repeats itself geometric relationships are deemed identifiable and topological estimation can be applied; analogies / expert opinions no longer need to be relied on. The moment a complexity group repeats itself four times the estimator can move to parametric estimation techniques.

Table 4. Selection guideline

Minimum a priori data counter	Estimation technique	Reasoning
0	Analogies / expert opinion	No data available
1-3	Topological	Geometric relationships can be identifed
4-41	Parametric	Arithmetic relationships can be identified
> 42	Standard regression	Central Limit Theorem applicable

The selection guidelines are based on the available minimum a priori data as related to a bit string describing the direction of cost variance propagation (or other metric such as geometrical symmetry). Geometric relationships are understood as the interdependency of topological boundaries when the shape of point clouds is used as an organising principle for data while arithmetic relationships are such that can be described through parametric approaches such as dependency models [11].

When variance data is not available for any time interval then only analogies or expert opinion can be used for forecasting cost estimate uncertainty propagation.

If variance data for one to three time increments is available then topological estimation can be applied since it draws upon the redundant information provided in the point cloud shape as indicators for propagation. The approach should be applied as an enhancement to a structured analogy or expert estimating process.

When variance data for four increments is available then parametric estimation can begin to be applied and the confidence in the generated results will grow iteratively until variance data for more than 42 time increments is available and standard statistical regression techniques become most effective.

Important to note however is that assuming a normal accounting period of one year then it can be safely assumed that at no time will standard regression techniques in fact become applicable unless this is compensated for by production volumes of identical products. The question of production volumes is hereby closely related to ramp up rates in a manufacturing facility so that it can be assumed that once at least 4 units have been produced (independent of time taken for this) this may suffice for the application of parametric techniques. In this respect the production of a single unit may suffice for the application of topological estimation techniques. The authors assume that the whole product life cycle phases of concept and development are the shortest and may at best need three or less years to complete. The phases of production, utilization and support are expected to last (in parallel) perhaps 30-40 years and the retirement phase between four and ten years [13].

At the beginning of each phase the conditions for estimating and forecasting usually change in respect to methods, stakeholders and timelines so that it can safely be assumed that the minimum a priori data counter is reset. Estimating and forecasting must therefore occur initially based on analogy or expert opinion to then be replaced by topological estimation until the end of time increment three

and then followed by parametric estimation approaches. If only small volumes of products are manufactured (therefore less than 42) then standard regression techniques may not become applicable at any point in time.

5. Selection process

The process for selecting the most appropriate estimation technique is based on the following factors:

- n = the number of historical time intervals with a common complexity group for which data is available
- d = the number of consistent cost variance dimensions n is available for
- CERs means “Cost Estimating Relationships”
- ∫ means the integral version of the dependency model describing the entirety of CERs

Figure 2 illustrates the relevant decision making process.

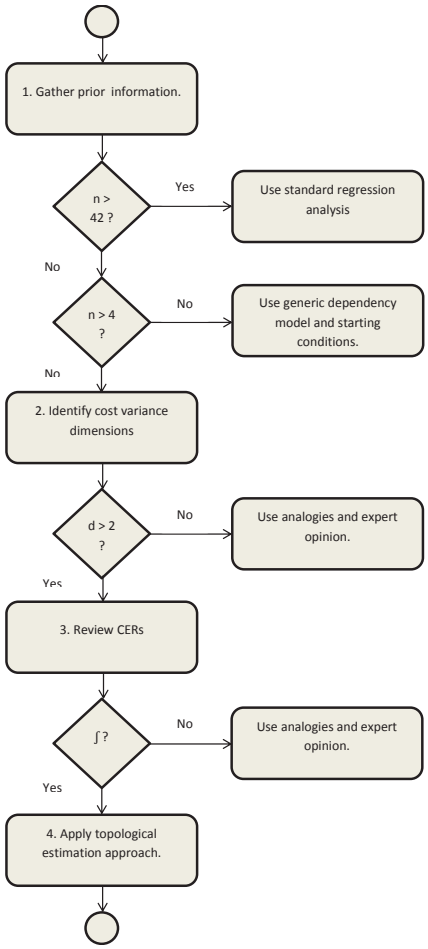

Fig. 2. Selection process

The approach begins with the cost estimator gathering all prior cost variance information available for historical discrete time intervals, i.e. the cost variance propagation in the time periods before the time of the estimate and determining its complexity group:

- If the number of discrete time intervals (t) for which data is available (n) is greater than 42 then standard regression analysis approaches can be applied since the Kolmogorov complexity of the underlying data strings is sufficient to allow for lossless compression [5].
- If the number of discrete time intervals for which data of a common complexity group is available (n) is greater than four but less than 42 then a cost estimate can be made using the generic dependency model and generic starting conditions presented.
- If the number of discrete time intervals for which data is available (n) is less than four then the cost estimator proceeds to determine for how many dimensions (d) cost variance data is available.
- If more than two dimensions are available (i.e. cost variance due to quantity, schedule and engineering, changes) then topological data analysis can be considered applicable since the minimum number of dimensions for such has been arrived at (therefore three) otherwise the cost estimator must depend on using analogies and expert opinion for proceeding with the cost estimate [4,7,8,9,10,11,12, 13].

The cost estimator can then proceed to review the relationships between the cost variance factors in order to determine whether a context specific cost estimating relationship (CER) can be defined and transformed into a dependency model. This review can occur through data analysis and / or expert opinion as deemed relevant. Cost estimation relationship models are used to correlate key independent and dependent variables for calculating the technical baseline estimate (including inherent uncertainty). The presented approach focuses on those variables related to the variance of key performance indicators for the whole product life cycle itself. On the simplest level these are cost variances related to schedule, cost and requirements. If a dependency model can be defined the cost estimator must then determine whether this relationship can be described using the integral form (∫) [14]. While continuous change is best modelled using differential equations which support infitessimal accumulation processes discrete event simulations such as the one represented by this stock-flow diagram are better modelled through integral equations as suggested by Forrester [14] and Ossimitz and Mrotzek [15] in their opinions that these explain the relationship between a model and the real world more effectively. If this is not the case then the cost estimator must rely on expert opinion for the cost estimate. If the integral form can be described then the presented method can be used for quantifying and forecasting cost estimate uncertainty.

Applying the selection technique to the case study data results in the selection of techniques as shown in Table 5.

Table 5. Case study data selection techniques

Year	Estimation technique
1999	Analogies and expert opinion
2000	Topological
2001	Topological
2002	Topological
2003	Analogy and expert opinion
2004	Topological
2005	Topological
2006	Topological
2007	Analogy and expert opinion
2008	Analogy and expert opinion

Important to note is that changes in the technique will always occur one time period after the complexity group has changed since only then is the cost variance data for assessing the complexity group available.

6. Assessing innovativeness

Although the authors introduce a definition of innovativeness based on the availability of a verified and accurate cost model it is important to provide the estimator with greater guidance on applying this in practice. For purpose of exploration a high value manufacturing product can be considered to consist of multiple components that each consists of multiple sub-components as illustrated in Figure 3.

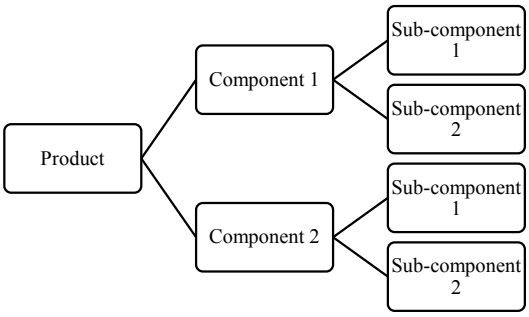

Fig. 3. Component breakdown

While individual components or sub-components may have verified and accurate cost models the object of analysis is the product as a whole or the system the product is embedded in. Therefore the moment any (sub-component) does not have a verified and accurate cost model the product (or system it is embedded in) must be considered as innovative and the cost selection technique presented applied.

7. Application to common estimation scenarios

The presented selection process is not based on specific whole product life cycle phases or technology readiness level as is common in practice. The presented selection process is based on the availability of a priori data for discrete time periods of common complexity groups. When exploring relevant scenarios for the estimator which trigger the application of the method certain generic points in the whole product life cycle can be identified. It is at least at these points where the presented method should find application. Generic points include:

- Change in whole product life cycle phase
- Major milestones applied during a project
- After re-baselining of an estimate
- Significant changes in schedule or requirements
- Change in responsible cost estimator
- Change in key assumptions

These generic points all relate to significant changes in the open complex system [16] represented by the whole product life cycle and hence by default will lead to significant changes in perspectives and valuations of the involved stakeholder groups (which also change accordingly).

8. Conclusions and recommendations for further research

The relevance of the presented approach for industry practice may be significant because it potentially disqualifies the use of standard regression based techniques (i.e. Monte Carlo simulation) in most cost estimation scenarios. When estimating the cost for innovative high value manufacturing products the attribute of innovativeness defers any initial cost estimate technique automatically to analogies or expert opinions. Only after the cost variance data from the first time increment is available can structured methods such as the topological approach be applied whereby with the fourth element parametric approaches appear to become feasible.

From a practitioner perspective this means that the more innovative a high value manufacturing product the less feasible a cost estimate uncertainty forecast for the whole product life cycle can be. It is only as the number of units produced grows that cost estimates can be validated and hence the innovativeness drops. The less prior data is available the more the estimator must hence be given permission to doubt the accuracy of estimates thus moving from single point estimates, to ranges and probability fields [17, 18].

One common mitigation strategy in respect to the financial uncertainty is to reduce the innovativeness to a degree that at least parametric models can be applied with relative confidence. The authors argue that the need for such a strategy could be reduced by introducing the concept of topological estimation which requires only one set of cost variance data in order to generate a first forecast if only for a short time window.

A further common mitigation strategy is to introduce a research and development phase that precedes the concept phase for a specific innovative high value manufacturing product. The research and development phase is budgeted for separately from a specific product and is intended to reduce the innovativeness of the solution per se before being confronted with the need for a commercially binding estimate.

In parallel by introducing a focus on factors affecting the variance of cost (and their interdependency) a path is opened for optimizing the accountability and tracking of such among the relevant stakeholders which in itself helps to stabilize the context the estimate is being made in as well.

Comparative validation of the presented approach using a variety of case studies is part of ongoing research activities.

Acknowledgements

The authors wish to thank the members of the LinkedIn Group "Cost Risk and Uncertainty" [19] for their continued support and the anonymous referees for their constructive comments and helpful suggestions for improving this paper.

Key terms and definitions

This article is based on a number of important terms whose definition is provided in Table 6 for the sake of clarity.

Table 6. Key terms and definitions.

Term	Definition
Baseline estimate	The agreed cost of producing a unit or delivering agreed support services. This cost consists of costed technical line items (often called the technical baseline estimate) and a risk contingency.
Complexity	As defined by Kolmolgorov this metric quantifies the length of the shortest computer program that reproduces a specific binary string.
Complexity group	The Kolmolgorov complexity shared by different binary strings of equal length.
Cost uncertainty	Manifested and unintended future cost variance with a probability of 100% and an unknown quantity.
Cost variance	Deviations from the baseline estimate.
Deep uncertainty	A decision-making situation where Knightian uncertainty, conflicting divergent paradigms and emergent decision making are relevant.
Forecast	Predictions of future baseline estimate changes.
High value manufacturing product	Products, production processes, and associated services which have the potential to create sustainable growth and high economic value.
Minimum a priori data	Historical cost variance known in advance of estimation which suffices for cost estimating.
Innovativeness	A product attribute which exists when no verified cost estimates are available.
Open complex system	A group of dependent variables that form a purposeful whole, interacts with its environment and exhibits unpredictable behavior.
Pattern	Recurring behaviour of data as it propagates.
Prior information	The probability distribution function applied to a data set before the identification of relevant evidence.
Scenario	A future use case for a product or service for which a business model has been created.
Small data	Data sets which are significantly smaller than those encountered in daily practice and arise from a context of few measurement points, little prior experience, little to no known history, low quality and conditions of deep uncertainty.
Stability	The consistency of the complexity group over time.

References

[1] Scales, J.A., Tenorio, L. (2001) "Prior information and uncertainty in inverse problems", Geophysics, Vol. 66, No. 2, pp. 389–397

[2] Technology Strategy Board (2012) "High Value Manufacturing Strategy"

[3] Schwabe, O., Shehab, E., Erkoyuncu, J.A. (2015a) "Uncertainty Quantification Metrics for Whole Product Life Cycle Cost Estimates in Aerospace Innovation", Journal Progress in Aerospace Sciences, pp. 1-24, DOI: 10.1016/j.paerosci.2015.06.002

[4] Schwabe, O., Shehab, E., Erkojuncu, J. (2015b) "Geometric Quantification of Cost Uncertainty Propagation: A Case Study", Procedia CIRP, Vol. 37, 2015, pp.158-163, CIRPe 2015 - Understanding the life cycle implications of manufacturing

[5] Soler-Toscano, F., Zenil, H., Delahaye, J-P., Gauvrit, N. (2014) "Calculating Kolmogorov Complexity from the Frequency Output Distributions of Small Turing Machines", Preprint submitted to the Journal of Theoretical Computer Science, March 2015

[6] United States Department of Defence (2015) "Selected Acquisition Reports (SAR) Summary Tables"

[7] Edelsbrunner, H., Letscher, D., Zomorodian, A. (2001) "Topological persistence and simplification", Discrete and Computational Geometry, Vol. 28, pp. 511-533

[8] Zomorodian, A., Carlsson, G. (2005) "Computing persistent homology", Discrete and Computational Geometry, Vol. 33, pp. 249-274

[9] Bubenik, P., Kim, P.T. (2007) "A statistical approach to persistent homology", Homology Homotopy and Applications, Vol. 9, pp. 337-362

[10] Ghrist, R. (2008) "Barcodes: The Persistent Topology of Data", Bulletin (New Series) of the American Mathematical Society, Vol. 45, pp. 61-75

[11] Bubenik, P., Carlsson, G., Kim, P. (2009) "Data Analysis using Computational Topology and Geometric Statistics", Banff International Research Station (BIRS), Research workshop, March 8-13, 2009

[12] Schwabe, O., Shehab, E., Erkoyuncu, J. (2016) "A Framework for Early Life Cycle Visualisation, Quantification and Forecasting of Cost Uncertainty in the Aerospace Industry", Journal Progress in Aerospace Sciences

[13] Carlsson, G. (April 2009) "Topology and data", Bulletin (New Series) of the American Mathematical Society, Vol. 46, pp. 255-308

[14] Forrester, J. W. (1968) "Principles of Systems", Text and Workbook. Cambridge, Mass., MIT Press

[15] Ossimitz, G, Mrotzek, M. (2008) "The Basics of System Dynamics: Discrete vs. Continuous Modelling of Time", International System Dynamics Conference, Athens, Greece

[16] Mitleton-Kelly, E. (2003) "Complex Systems and evolutionary perspectives on organisations: The application of complexity theory to organisations", Elsevier ISBN: 0-08-043957-8

[17] Berberian, S.K. (1961) "Introduction to Hilbert Space", American Mathematical Society, AMS Chelsea Publishing, ISBN-10: 0821819127

[18] Robbin, T. (2015) "Topology and the Visualization of Space", Symmetry, Vol. 7, pp. 32-39

[19] Cost Risk and Uncertainty (2016) LinkedIn Group.