MISSOURI S&T | CURTIS LAWS WILSON LIBRARY
MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

Scholars' Mine

Masters Theses | Student Research & Creative Works

Fall 2010

A comparison of case-based reasoning and regression analysis approaches for cost uncertainty modeling

Karan Banga

Follow this and additional works at: http://scholarsmine.mst.edu/masters_theses

 Part of the Mechanical Engineering Commons

Department:

Recommended Citation

Banga, Karan, "A comparison of case-based reasoning and regression analysis approaches for cost uncertainty modeling" (2010). *Masters Theses.* Paper 4986.

This Thesis - Open Access is brought to you for free and open access by Scholars' Mine. It has been accepted for inclusion in Masters Theses by an authorized administrator of Scholars' Mine. This work is protected by U. S. Copyright Law. Unauthorized use including reproduction for redistribution requires the permission of the copyright holder. For more information, please contact scholarsmine@mst.edu.

A COMPARISON OF CASE-BASED REASONING AND REGRESSION ANALYSIS APPROACHES FOR COST UNCERTAINTY MODELING

by

KARAN BANGA

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN MECHANICAL ENGINEERING

2010

Approved by

Shun Takai, Advisor
Frank W. Liou
Xiaoping Du

PUBLICATION THESIS OPTION

This thesis is presented in publication format and is divided into two separate papers. Pages 1 through 42 have been sent for review in RED 2010, Research in Engineering Design.

Pages 43 through 77 are intended to be sent for publication in RED 2010, Research in Engineering Design.

ABSTRACT

This thesis presents case-based reasoning approach for estimating the cost and modeling cost uncertainty of a new product in the concept selection stage. Case-based reasoning (CBR) is an approach which uses old cases/experiences to understand and solve new problems. The CBR approach consists of creating a knowledge-base (or database) containing past cases (products), defining a new case (concept), retrieving cases similar to the new case, and adjusting the solution of the retrieved cases to the new case. The first paper compares case-based reasoning, in studying the effects of varying design attribute specifications on cost estimation accuracy and cost distribution reliability. Case-based reasoning with cost estimation is compared with three methods: analogy-based cost estimation, case-based reasoning without cost adjustment, and regression analysis. Four automobile concepts with similar performance attribute specifications but varying design attribute specifications are defined and the comparison is made using leave-one-out cross-validation technique to a knowledge-base of 345 automobiles. The second paper further establishes case-based reasoning with cost adjustment by studying the optimum number of design attributes for specifying a concept. The results show that case-based reasoning with cost adjustment performed best for cost estimation accuracy and cost distribution reliability when one design attribute is specified for the concept in addition to performance attributes.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Shun Takai, for his guidance and support, for his insight and perception, for continually encouraging me to broaden my horizons, and for offering me the oppurtunity to pursue a graduate degree while performing the research presented in this thesis. I also wish to thank him for laying the foundation for the work presented in these papers.

Second, I would like to thank Dr. Frank Liou and Dr. Xiaoping Du for serving as committee members and examining the thesis. I would like to thank my friend, Tim Huff for helping me with the programming which has been very crucial in completing my research work. I would also like to thank my friends here in Rolla, who have made the past two years memorable.

The financial assistance provided to me in the form of Graduate Research Assistantship and Graduate Teaching Assistantship through the Intelligent Systems Centre, Interdisciplinary Engineering Department and Department of Mechanical Engineering is gratefully acknowledged.

Finally, I would like to thank my family for their encouragement and patience without which I would not have accomplished this task. I would like to dedicate this thesis to my parents.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

LIST OF TABLES

Table Page

PAPER 1

PAPER 2

PAPER

1. A COMPARISON OF COST ESTIMATION AND COST UNCERTAINTY MODELING APPROACHES

Karan Banga
Department of Mechanical Engineering
Missouri University of Science and Technology
Rolla, MO 65409
kbcgf@mst.edu

Shun Takai (Corresponding Author)
Department of Mechanical Engineering
Missouri University of Science and Technology
Rolla, MO 65409
takais@mst.edu

ABSTRACT

This paper studies case-based reasoning approaches for estimating the cost and modeling cost uncertainty of a new product in the concept selection stage. Case-based reasoning is a procedure to use past cases (experiences) to understand and solve new problems. The case-based reasoning approach consists of creating a knowledge base of past and current products (cases), defining a new product concept, retrieving products similar to the concept, and adjusting costs of the retrieved products to estimate cost and generate cost distribution of the concept. This paper compares case-based reasoning and regression analysis approaches for accuracies of cost estimations and reliabilities of cost distributions. These approaches are compared and effects of defining a concept with a

design attribute, in addition to performance attributes, are studied by applying leave-one-out cross-validation to a knowledge base of automobiles.

KEYWORDS: Case-based reasoning, regression analysis, cost, concept, hierarchical clustering, distribution, leave-one-out cross-validation

1. INTRODUCTION

Product cost is one of the most important factors that determine profitability of a new product. Although accurate cost estimation is essential when selecting a new product concept in the early product development stage, lack of detailed design and assembly process information creates a large degree of uncertainty about product costs and makes accurately cost estimation challenging.

Detailed cost modeling and regression analysis are two widely used methods for estimating a cost of a new product. Cost modeling calculates a product cost by adding part costs, assembly costs, and overhead costs estimated from detailed product information such as bill of materials, design specifications, and assembly process specifications (Ulrich and Eppinger 2004; Otto and Wood 2001; Pahl and Beitz 1996). Because this detailed information is not available in the concept selection stage, cost modeling may not be the optimal approach to estimate the cost of a concept.

Regression analysis (Hamaker 1995; Wyskida 1995) can estimate a product cost using product-level information (i.e., product specifications), and does not necessarily require detailed design and assembly process information. Regression analysis generates a cost estimation relationship (CER), which describes a cost (a dependent variable) as a function of one or more cost-relevant product attributes (independent variables). The cost of a new product is estimated by substituting its product information into the CER. Although regression analysis has a strong theoretical foundation (Neter et al. 1996) and has been widely used in design research (Michalek et al. 2004; Williams et al. 2008; Shiau and Michalek 2009), Braxton and Coleman (2007) identify various challenges in

applying regression analysis in practice. One of these challenges is a poor quality of real world cost data (e.g., missing data and outliers), which can lead to inaccurate cost estimations.

Analogy-based cost estimation is a relatively new method that has been proposed to apply case-based reasoning to estimate cost of software projects (Shepperd and Scofield 1997; Angelis and Stamelos 2000; Mendes et al. 2003; Auer et al. 2006; Jeffery et al. 2000) and more recently to estimate costs of construction projects (Kim et al. 2004; An et al. 2007). Corresponding to a case-based reasoning procedure, which is to use past cases (experiences) to understand and solve a new problem (Kolodner 1993), analogy-based cost estimation creates a knowledge base that contains past projects (cases), defines features of a new project, retrieves up to three past projects that have similar features as the new project, and estimates the cost of the new project from the costs of the retrieved projects (Shepperd and Scofield 1997; Mendes et al. 2003).

When applied to product cost estimation, analogy-based cost estimation can estimate cost of a concept only from product-level specifications without relying on detailed design and assembly process information; however, when applied to cost uncertainty modeling, it may have two limitations. First, it only retrieves up to three projects. The use of a small number of similar projects allows accurate cost estimations; however, three data points may not be sufficient to construct reliable cost distributions (Fox and Safie 1992). Second, most of the analogy-based cost estimation applications do not adjust costs of retrieved projects for the differences between the attribute values of the retrieved projects and those of a new project (Shepperd and Scofield 1997; Mendes et al. 2003); therefore, analogy-based cost estimation may not be fully utilizing information

available in the retrieved projects when modeling cost uncertainty. Jeffery et al. (2000) propose linearly adjusting costs of retrieved projects with respect to a project attribute that has the largest correlation with cost; however, linear adjustment on a single attribute may still fail to take into account all the attribute information available from the retrieved projects.

This paper presents a case-based reasoning approach that utilizes hierarchical clustering to retrieve as many products similar to the concept as possible and adjusts these costs parallel to the regression model obtained from the retrieved products. This paper compares case-based reasoning approaches with analogy-based cost estimation and regression analysis on the basis of accuracy of cost estimation and reliability of cost uncertainty modeling. Jeffery et al. (2000) compare the accuracy of cost estimation between analogy-based cost estimation and ordinary least squares regression using data from company-specific data as well as multi-company data. Although no significant differences are observed between these two techniques when they are applied to the company-specific data, ordinary least squares regression performs significantly better than analogy-based cost estimation when they are applied to the multi-company data. Takai (2009) compares accuracies of cost estimations in a case-based reasoning approach and analogy-based cost estimation with and without a linear adjustment using to a heterogeneous knowledge base (i.e., with missing data). Case-based reasoning provides slightly more accurate cost estimations than analogy-based cost estimations. Because regression analysis may not be able to provide accurate cost estimations when a heterogeneous knowledge base is used, the case-based reasoning approach is not compared against regression analysis. Furthermore, while this study proposes to

represent a concept and products with binary indices (zero or one depending on whether data exist for each product attribute), a similarity measure based on availability of information is not useful when a knowledge base is homogeneous (i.e., when there are no or few missing data). In this study, reliabilities of cost distributions have not been compared.

Case-based reasoning has been used in solving design problems (Bardasz and Zeid 1991; Bardasz and Zeid 1993; Roderman and Tsatsoulis 1993; Maher and Zhang 1993; Shiva Kumar and Krishnamoorthy 1995; Rosenman 2000; Wood and Agogino 1996; Lee and Lee 2002; Al-Shahibi and Zeid 1998). Bardasz and Zeid (1991, 1993) have used it to solve mechanical design problems. Roderman and Tsatsoulis (1993) have created the Pumper Apparatus Novice Design Assistant (PANDA), a case-based design system to assist firefighters who wish to design their pumper engines. Maher and Zhang (1993) have proposed a case-based design process model, CADSYN, to solve new design problems. Cost estimation of a new product in the concept selection stage, however, has not been the scope of these research projects.

This paper illustrates case-based reasoning approaches for a homogeneous knowledge base and compares accuracies of cost estimations and reliabilities of cost distributions against those of analogy-based cost estimation and regression analysis. Furthermore, this paper studies effects of a design specification on accuracies of cost estimations and reliabilities of cost distributions. The remainder of this paper is organized as follows: Section 2 describes a case-based reasoning approach for cost uncertainty modeling; Section 3 illustrates cost estimation and cost uncertainty modeling by case-based reasoning, analogy-based cost estimation, and regression analysis using a

knowledge base of automobiles; Section 4 compares these approaches on the basis of accuracies of cost estimations and reliabilities of cost distributions using leave-one-out cross-validation; Section 5 concludes the paper with discussions for future work.

2. CASE-BASED REASONING APPROACH FOR COST UNCERTAINTY MODELING

Figure 1 schematically illustrates four steps of case-based reasoning approach for cost uncertainty modeling: construction of a knowledge base that contains past and current products (cases), definition of a concept, retrieval of products similar to the concept, and generation of a cost distribution for the concept.

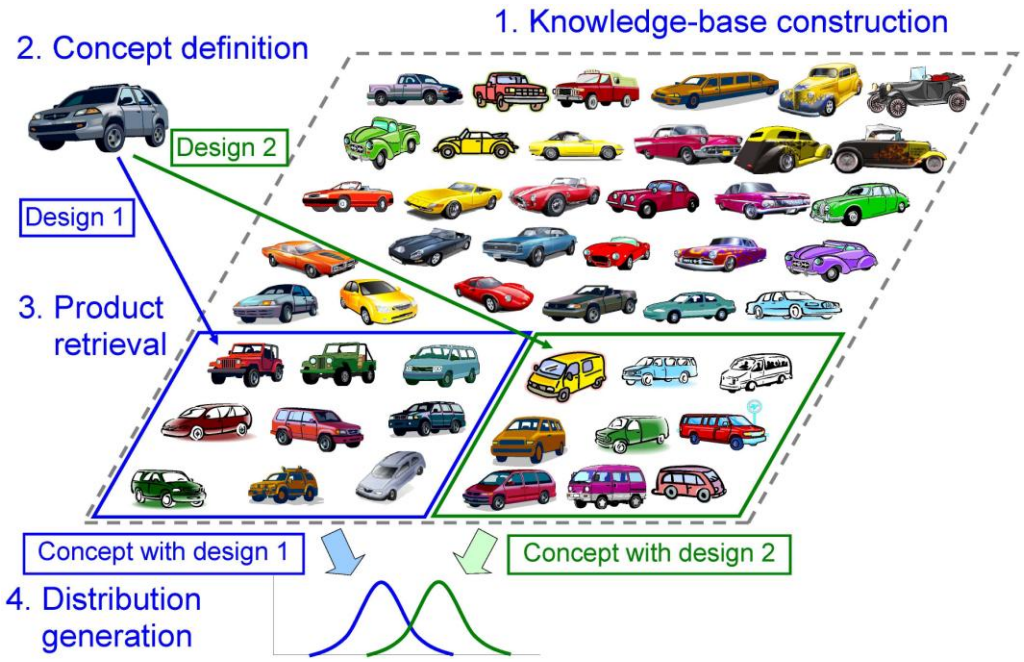


Fig. 1 Case-Based Reasoning Process Flow

2.1 Knowledge Base Construction

The first step is to construct a knowledge base of past and current products. The knowledge base should contain products, together with their attributes and specifications. Product attributes define properties of a product, and product specifications determine specific values of product attributes that a product needs to achieve. In the case of an automobile, fuel efficiency is an attribute for which 25 miles per gallon is a specification. An attribute may be categorical or numerical.

A product attribute may be classified into a performance attribute or a design attribute. Performance attributes describe performance requirements of a product, and they directly affect customers' purchasing decisions. In contrast, design attributes describe design characteristics that enable a product to achieve its performance specifications. For example, "0–60 mph acceleration time" may be defined as a performance attribute and "engine capacity" may be defined as a design attribute.

2.2 Concept Definition

The second step is to define a concept by performance attributes that influence customers' product purchasing decisions. These attributes are identified for example, by first collecting customer needs by interviewing customers and then translating representative needs to corresponding performance attributes. Once performance attributes are identified, performance specifications may be defined.

In addition to performance attributes and specifications, designers may further define a concept by cost-relevant design attributes and specifications. Design specifications enable designers to define a concept in more detail and may provide more

accurate cost estimations and more reliable cost distributions. On the other hand, specifying a concept by too many design attributes may create a risk of biased cost estimations and cost distributions if the design specifications of the final product change from those of the concept. If cost estimations and cost distributions are biased, the initially-selected concept may no longer be an optimum one.

2.3 Product Retrieval

The third step is to retrieve products similar to the concept from the knowledge base using hierarchical clustering. Hierarchical clustering procedure consists of data matrix modification, distance matrix generation, and product retrieval. First, in order to identify products in the knowledge base that are similar to the concept, a data matrix of the initial knowledge base is modified by including the concept in an additional first row. For example, if the knowledge base contains I number of products and J number of attributes, then the data matrix will have I rows for products and J columns for attributes. Initially, the identification number of row of the knowledge base varies from i=1 to I. After including the concept as row i=0, the modified data matrix consists of I+1 rows (i=0 to I) and J columns (j=1 to J).

Second, a distance matrix is generated from the modified data matrix by calculating Euclidian distances between the concept and each product, and between each pair of products. A Euclidian distance, δ between two products, p and p' is defined as

$$\delta(p, p') = \sqrt{\sum_{j=1}^{J} w_j \left(s_{i,j} - s'_{i,j} \right)^2} \qquad (1)$$

where $s_{i,j}$ is the standardized specification of product p (explained later in Equation 3), $s'_{i,j}$ is the standardized specification of product p', and w_j is the weight of attribute j. In this paper, weights of all attributes are set to 1.

Finally, hierarchical clustering is applied to the distance matrix in order to retrieve products similar to the concept. Hierarchical clustering generates upside-down tree-like figures (called dendrograms) based on the distances calculated in the distance matrix. In a dendrogram, the height at which two products, two clusters, or a product and a cluster are grouped together indicates the distance between them. The smaller distances between products/clusters indicate that they are more similar and, therefore, they are grouped together at the lower linkage height in the dendrogram. Hierarchical clustering has been used to group similar cases in the knowledge base before retrieving the most similar case from the group (Reich and Kapeliuk 2004). In this paper, all the products (cases) in the group similar to the concept is retrieved and used to estimate costs and to construct cost distributions.

The three methods that may be used to group similar objects in hierarchical clustering are the single-linkage method, the complete-linkage method, and the average-linkage method. The single-linkage method calculates, element by element, a distance between an element in one cluster and an element in another cluster, and defines a distance of two clusters as the smallest element-by-element distance. On the contrary, the complete-linkage method defines a distance of two clusters as the largest element-by-element distance, and the average-linkage method defines a distance of two clusters as the average element-by-element distance. In this paper, the average-linkage method is

used because of a statistical consistency property that is violated by the other two methods (Kelly and Rice 1990; Hastie et al. 2001).

Figure 2 shows an example dendrogram obtained from applying hierarchical clustering to a knowledge base that consists of a concept (C) and nine products (P1–P9). In Fig. 2, linkage heights at which concept C is grouped with other products or clusters are labeled H1, H2, and H3. Heights H1, H2, and H3 correspond to the linkage height in which the concept is grouped with other products for the first time, for the second time, and for the third time. The differences of these linkage heights are denoted as ΔH; for example, ΔH1 represents the difference in linkage heights between H2 and H1 (ΔH1=H2-H1), and so on. The largest incremental distance (difference of linkage heights) ΔH is used to decide which products are similar to the concept and retrieved from the knowledge base. In this example, because ΔH1 is larger than ΔH2, the dendrogram is cut at the largest distance ΔH1 (for example, at the dashed line in Fig. 2), which indicate two products P2 and P3 are grouped with concept C and retrieved from the knowledge base.

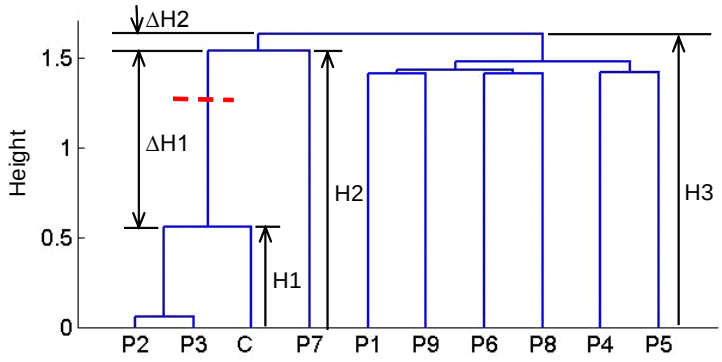


Fig. 2 Example Dendrogram

2.4 Distribution Generation

The final step is to generate cost distributions after adjusting costs of the retrieved products. For cost adjustment, three methods may be used: no adjustment, linear adjustment, and parallel adjustment. In the case of no adjustment, the costs of the retrieved products are used to estimate a cost and generate a cost distribution of a concept. In the case of linear adjustment, one attribute of the retrieved products that is most closely correlated with their costs is identified first. Then, ratios of attribute specifications between the concept and the retrieved products are calculated. Finally, the costs of the retrieved products are adjusted in proportion to these ratios. These adjusted costs are used to estimate the cost and generate the cost distribution of the concept. In the case of parallel adjustment, a regression model is obtained by applying a regression analysis to the retrieved products first, then, the costs of the retrieved products are adjusted parallel to the regression model. These adjusted costs are used to estimate the cost and generate the cost distribution of the concept. The regression model could be a line (in the case of a single cost-relevant attribute) or a surface (in the case of multiple cost-relevant attributes). Figure 3 illustrates these three cost adjustment methods in the case of a single numeric cost-relevant attribute. This paper fits normal distributions to generate cost distributions, but other distributions may also be used.

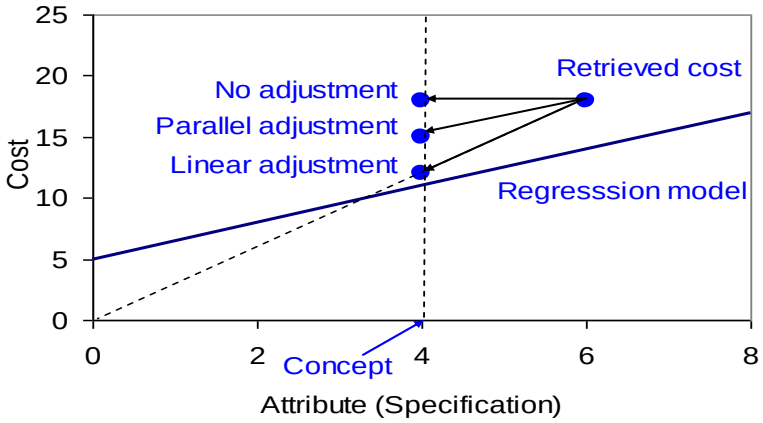


Fig. 3 Cost Adjustment

3. COMPARISON OF COST ESTIMATION AND COST UNCERTAINTY MODELING APPROACHES

Using a knowledge base of automobiles, this section 1) presents case-based reasoning approaches for estimating costs and modeling cost uncertainties of automobile concepts, 2) compares these approaches with analogy-based cost estimation and regression analysis approaches, and 3) studies an effect of a design specification on cost estimation and cost uncertainty modeling when a design attribute is used to define a concept in addition to performance attributes. The reference method is *analogy-based cost estimation (ABCE)*, which retrieves three automobiles most similar to the concept based on the smallest Euclidian distances from the concept. If there are automobiles with same distances from the concept, more than three automobiles may be retrieved. The costs of the retrieved automobiles are not adjusted.

The second method is *case-based reasoning without cost adjustment (CBR)*. In contrast to analogy-based cost estimation, case-based reasoning retrieves as many similar automobiles as possible from the knowledge base by applying a hierarchical clustering with average-linkage method. Automobiles similar to the concept are defined and retrieved from the knowledge base as a result of the largest incremental distance ΔH as discussed in section 2.3. As in analogy-based cost estimation, the costs of the retrieved automobiles are not adjusted.

The third method is *case-based reasoning with cost adjustment (CBR-A)*. In contrast to the case-based reasoning without adjustment, the costs of the retrieved

automobiles are adjusted parallel to a regression model obtained from applying a regression analysis to the retrieved automobiles.

The last method is *regression analysis (RA).* In contrast to case-based reasoning with adjustment, a regression analysis is performed on *all* automobiles in the knowledge base and the costs of all automobiles are adjusted parallel to the regression model.

After the costs of the retrieved automobiles (or of all automobiles in the knowledge base in the case of regression analysis) are adjusted if necessary, a cost of the concept is estimated by averaging these costs and a cost distribution is generated by fitting a normal distribution to these costs.

To study effects of design specifications on a cost of a concept, a concept defined by only performance specifications is considered as a reference concept. Three concepts that are defined with a design specification in addition to the same performance specifications are compared to the reference concept for accuracies of cost estimations and reliabilities of cost distributions.

3.1 Knowledge Base Construction

A knowledge base is constructed by benchmarking automobiles sold in the U.S. from 2003 through 2009 and contains 345 automobiles, with 86 attributes and estimated costs. The automobile attributes and specifications are gathered from a product-evaluation firm's website, and include automobile type (SUV, small car, sedan, minivan, wagon, pickup and sports car), number of cylinders, engine capacity, number of side airbags, acceleration, braking, fuel efficiency, and roadside aid. These attributes are

further classified as either performance or design attributes. There are 51 performance attributes and 35 design attributes.

Because product costs are proprietary information, a cost of an automobile is estimated by subtracting a profit margin from a price similar to the approach used by Williams et al. (2008). In Eq. 2, annual automotive revenue and a total cost are collected from individual automobile company's annual financial reports.

$$\begin{aligned} Cost &= Price \times \left(1 - Average\ profit\ marging\right) \\ &= Price \times \left(1 - \frac{Revenue - Total\ cost}{Revenue}\right) \\ &= Price \times \frac{Total\ cost}{Revenue} \end{aligned} \tag{2}$$

Figure 4 shows a portion of the complete knowledge base used in the analysis. Cost data are used only for estimating a cost and generating a cost distribution of the concept and are not used for calculating distances between the concept and an automobile or between two automobiles in the hierarchical clustering process.

	Automobile Type	Fuel Efficiency (Miles per Gallon)	Engine Capacity (Liters)	Cost ($)
Automobile 1	Convertibles	22	2.7	21,554
Automobile 2	SUV	23	2.4	20,994
Automobile 3	SUV	21	2.4	17,483
Automobile 4	Small Cars	25	2	14,332
Automobile 5	Sedans	29	3	53,100
Automobile 6	Minivans	17	4	37,595
Automobile 7	Wagons	42	1.5	22,575
Automobile 8	Pickups	14	4.7	29,053
Automobile 9	Sporty	20	4.6	27,413

Fig. 4 Portion of the Complete Knowledge Base

3.2 Concept Definition

When defining a concept by performance and design attributes, performance attributes need to be important for customers to make automobile purchasing decisions and design attributes need to be important for designers to estimate automobile costs. Automobile type and fuel efficiency are selected as performance attributes of the concept because they influence customers' purchasing decisions. For example, McCarthy (1996) proposes that vehicle size and type have a direct impact on customers' purchasing decisions. Train and Winston (2007) discuss how Japanese automakers gained an edge over the US manufacturers because of the higher fuel efficiency of their vehicles. Berry et al. (1995, 2004) have proposed that a higher fuel efficiency and vehicle size drive customer demand. Engine capacity is chosen as a design attribute of the concept because it is the most cost-relevant design attribute (design attribute with the smallest p value) identified in a regression analysis that regresses cost against complete set of design attributes in the knowledge base. Vehicle type is a categorical attribute and fuel efficiency and engine capacity are numerical attributes.

Table 1 summarizes a reference concept (Concept 0) and three concepts (Concept 1, 2, and 3) studied in this paper. All concepts are described by two performance attributes (automobile type and fuel efficiency) and their specifications (SUV and 25 miles per gallon). In addition, Concepts 1, 2, and 3 are described by one design attribute (engine capacity) and its specification (2.4, 3.6, and 5.8 liters respectively) in order to study the effects of design specifications. For the purpose of illustration, SUV is arbitrarily chosen as the specification of automobile type and 25 miles per gallon is chosen as the specification close to the maximum fuel efficiency of the SUVs in the knowledge base. Three levels of engine capacity are chosen so that they approximately represent five percentile, median, and 95 percentile of the engine capacity of SUVs in the knowledge base.

Table 1 Concept Definition

		Vehicle Type	Fuel Efficiency (Miles per Gallon)	Engine Capacity (Liters)
Concept 0	(Reference)	SUV	25	-
Concept 1		SUV	25	2.4
Concept 2		SUV	25	3.6
Concept 3		SUV	25	5.8

3.3 Product Retrieval

To calculate distances among the concept and automobiles in the knowledge base, the original knowledge base is coded as shown in Fig. 5, which is demonstrated for

Concept 1. Corresponding to the four concepts in Table 1, there are four knowledge bases; i.e., one knowledge base for each concept (Concept 0, 1, 2, and 3). These knowledge bases differ only for the concept in the first row. In the first attribute, automobile type is broken down into eight categories: convertibles, SUVs, small cars, sedans, minivans, wagons, pickups, and sports cars. One is used if an automobile is of a particular type and zero if otherwise. By using ones and zeros, a Euclidean distance between two automobiles due to automobile type is zero if they are of the same type and 1 if otherwise.

	Type								Fuel Efficiency (Miles per Gallon)	Engine Capacity (Liters)
	Convertibles	SUV	Small Cars	Sedans	Minivans	Wagons	Pickups	Sporty		
Concept 1	0	1	0	0	0	0	0	0	25	2.4
Automobile 1	1	0	0	0	0	0	0	0	22	2.7
Automobile 2	0	1	0	0	0	0	0	0	23	2.4
Automobile 3	0	1	0	0	0	0	0	0	21	2.4
Automobile 4	0	0	1	0	0	0	0	0	25	2.0
Automobile 5	0	0	0	1	0	0	0	0	29	3.0
Automobile 6	0	0	0	0	1	0	0	0	17	4.0
Automobile 7	0	0	0	0	0	1	0	0	42	1.5
Automobile 8	0	0	0	0	0	0	1	0	14	4.7
Automobile 9	0	0	0	0	0	0	0	1	20	4.6

Fig. 5 Portion of the Modified Knowledge Base

All the attributes are then standardized using Eq. 3 so that each attribute has the same degree of influence in the Euclidean distance in Eq. 1.

$$s_{i,\,j} = \frac{x_{i,\,j} - \mu_a}{\sigma_a} \tag{3}$$

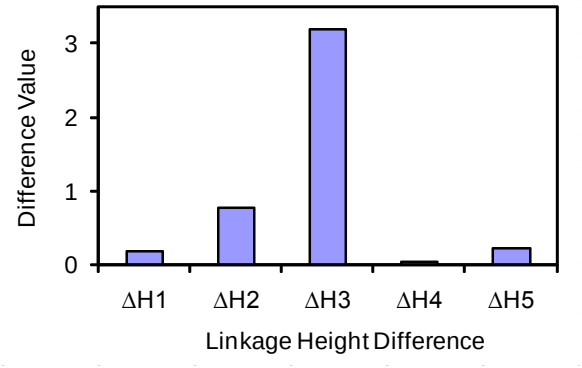
In Eq. 3, the subscript *i* represents a concept and the automobiles in the knowledge base, which varies from 0 to 345 (*i*=0 for the concept). The subscript *j* represents the column of the modified knowledge base in Fig. 5. The value $s_{i,j}$ is the standardized specification, $x_{i,j}$ is the coded specification, and μ_a and σ_a are the average and the standard deviation of specifications of an attribute *a*, where *a* varies from 1 to 3 (*a*=1 for type, *a*=2 for fuel efficiency, and *a*=3 for engine capacity). The mean and standard deviation are calculated only for the automobiles in the knowledge base (i.e., excluding the concept) in order to be consistent when specifications of various concepts are standardized. For the first attribute type (*a*=1), average and standard deviation are calculated for the first eight columns because they all belong to the same attribute; i.e., automobile type. The remaining two numerical attribute averages and standard deviations are calculated across their respective columns. Figure 6 shows the corresponding knowledge base with the standardized values.

For each concept (Concept 0, 1, 2, and 3), distances between a concept and each automobile and between each pair of automobiles are calculated from the standardized knowledge base in Fig. 6; however, for Concept 0, only first two attributes are used (automobile type and fuel efficiency) because the engine capacity is not defined for Concept 0.

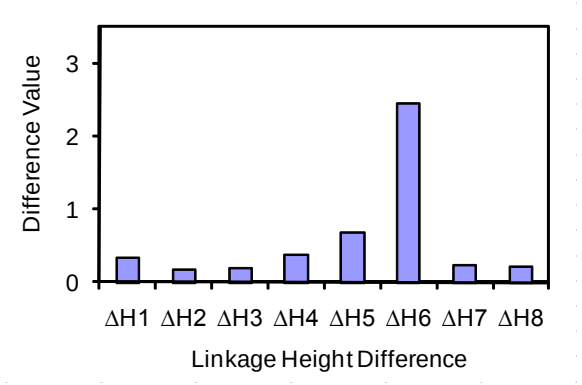
	Type								Fuel Efficiency (Miles per Gallon)	Engine Capacity (Liters)
	Convertibles	SUV	Small Cars	Sedans	Minivans	Wagons	Pickups	Sporty		
Concept 1	-0.4	2.6	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.6	-0.6
Automobile 1	2.6	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.1	-0.4
Automobile 2	-0.4	2.6	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.2	-0.6
Automobile 3	-0.4	2.6	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.1	-0.6
Automobile 4	-0.4	-0.4	2.6	-0.4	-0.4	-0.4	-0.4	-0.4	0.6	-0.9
Automobile 5	-0.4	-0.4	-0.4	2.6	-0.4	-0.4	-0.4	-0.4	1.3	-0.1
Automobile 6	-0.4	-0.4	-0.4	-0.4	2.6	-0.4	-0.4	-0.4	-0.8	0.7
Automobile 7	-0.4	-0.4	-0.4	-0.4	-0.4	2.6	-0.4	-0.4	3.5	-1.3
Automobile 8	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	2.6	-0.4	-1.3	1.3
Automobile 9	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	2.6	-0.3	1.2

Fig. 6 Portion of the Standardized Knowledge Base

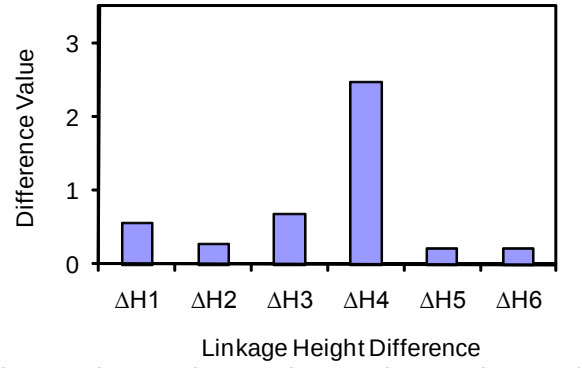
Hierarchical clustering is applied to four distance matrices (one for each concept) and four dendrograms in Fig. 10 in the appendix are generated. The incremental distances ΔH (the differences between every two consecutive linkage heights) are calculated from the dendrograms as illustrated in Fig. 7. As discussed in section 2.3, automobiles similar to the concept are identified based on the largest incremental difference ΔH and retrieved from the knowledge base. For example, in the case of Concept 0, automobiles belonging to H1 through H3 are considered similar because the highest bar ΔH3 (=H4-H3) indicates that the distance is largest between H3 and H4. Similarly, for concepts 1, 2 and 3, automobiles belonging to H1 through H6, H1 through H4, and H1 through H2, respectively, are considered similar.



(a) Concept 0

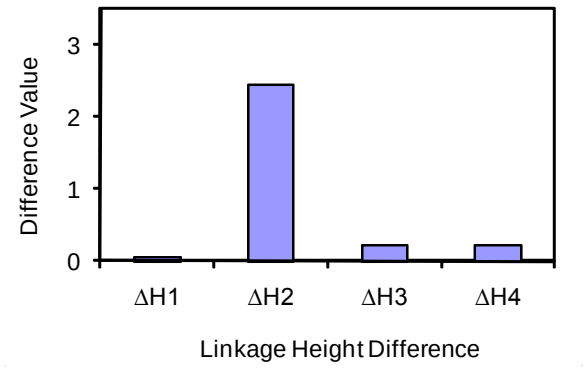


(b) Concept 1



(c) Concept 2

Fig. 7 Automobile Selection Criteria Graphs



(d) Concept 3

Fig. 7 (Continued) Automobile Selection Criteria Graphs

3.4 Distribution Generation

Once automobiles similar to the concept are identified and retrieved from the knowledge base, the next step is to construct a histogram and fit a normal distribution to the costs of the retrieved automobiles (with or without adjustment depending on the method, ABCE, CBR, CBR-A, or RA, outlined at the beginning of section 3). Figure 11 in the appendix summarizes 16 cost distributions obtained from applying four methods (ABCE, CBR, CBR-A, and RA) to four concepts (Concept 0, 1, 2, and 3).

3.5 Observations of Various Approaches

Table 2 summarizes six statistics--the number of retrieved automobiles (n), average (Ave), standard deviation (SD), minimum (Min), maximum (Max), and range (Range) of the costs of the retrieved automobiles (after adjustments if necessary)—in the

four methods (ABCE, CBR, CBR-A, and RA) for four concepts (Concept 0, 1, 2, and 3). Figure 8 plots four statistics: Ave, SD, Min, and Max.

Comparing four methods, standard deviation is very small in ABCE in two cases (Concept 1 and 3), which indicates that *ABCE may generate unreliable (too narrow) distributions* for cost uncertainty modeling.

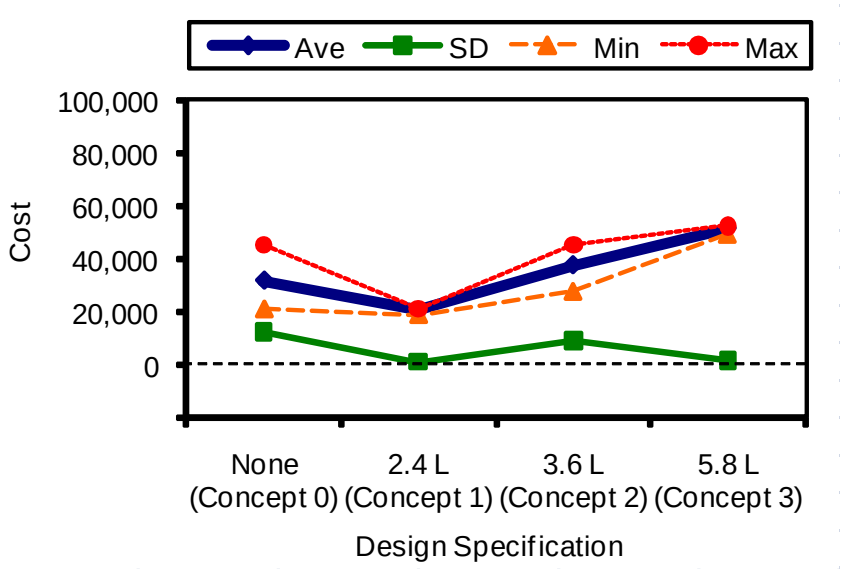
Average costs do not change in CBR because the same sets of automobiles are retrieved for all concepts and no adjustment is performed on the costs of the retrieved automobiles. This indicates that *CBR may provide inaccurate cost estimations*. ABCE does not adjust costs of retrieved automobiles; however, a set of retrieved automobiles and, therefore, statistics of the costs, can be different for each concept.

Except for CBR, average costs increase as an engine capacity increases from 2.4 to 3.6 and to 5.8 liters; thus, *defining the concept by an additional design attribute may provide a more accurate cost estimations*. For RA of Concept 1 (engine capacity 2.4 liter), the minimum cost after an adjustment is negative, which indicates that *distributions obtained in RA may be too wide*.

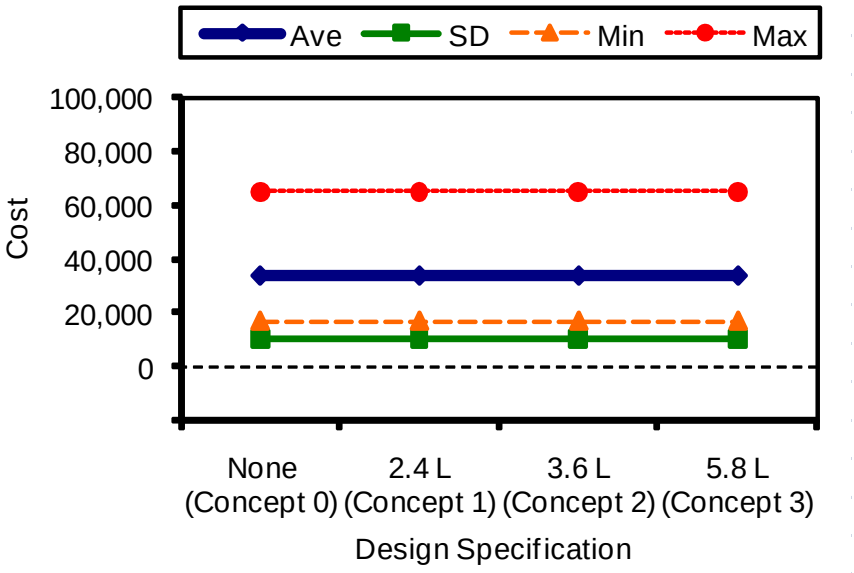
Except for CBR, both ranges and standard deviations are the largest when a design specification is not defined (Concept 0) compared to when a design specification is defined (Concept 1, 2, and 3). *Defining a concept by an additional design attribute (e.g., engine capacity) may provide narrower distributions; however, whether a narrower distribution results in a more reliable cost uncertainty modeling needs further study*.

Table 2 Data for the Cost Distribution Curves

Statistics	Concept	Methods			
		ABCE	**CBR**	**CBR-A**	**RA**
n	Concept 0	4	85	85	345
	Concept 1	4	85	85	345
	Concept 2	3	85	85	345
	Concept 3	3	85	85	345
Ave	Concept 0	31,717	33,482	17,886	22,110
	Concept 1	**20,426**	33,482	**23,593**	**20,699**
	Concept 2	**37,519**	33,482	**33,108**	**32,279**
	Concept 3	**51,330**	33,482	**50,551**	**53,510**
SD	Concept 0	**12,671**	10,471	**8,942**	**11,293**
	Concept 1	**1,179**	10,471	6,935	8,724
	Concept 2	9,077	10,471	6,935	8,724
	Concept 3	**1,885**	10,471	6,935	8,724
Min	Concept 0	20,826	16,816	3,157	2,513
	Concept 1	18,673	16,816	12,934	**-5,438**
	Concept 2	27,508	16,816	22,448	6,143
	Concept 3	49,154	16,816	39,891	27,374
Max	Concept 0	45,212	64,987	45,290	79,751
	Concept 1	21,211	64,987	48,886	65,081
	Concept 2	45,212	64,987	58,400	76,662
	Concept 3	52,419	64,987	75,843	97,893
Range	Concept 0	**24,386**	48,171	**42,133**	**77,238**
	Concept 1	2,538	48,171	35,952	70,519
	Concept 2	17,704	48,171	35,952	70,519
	Concept 3	3,265	48,171	35,952	70,519

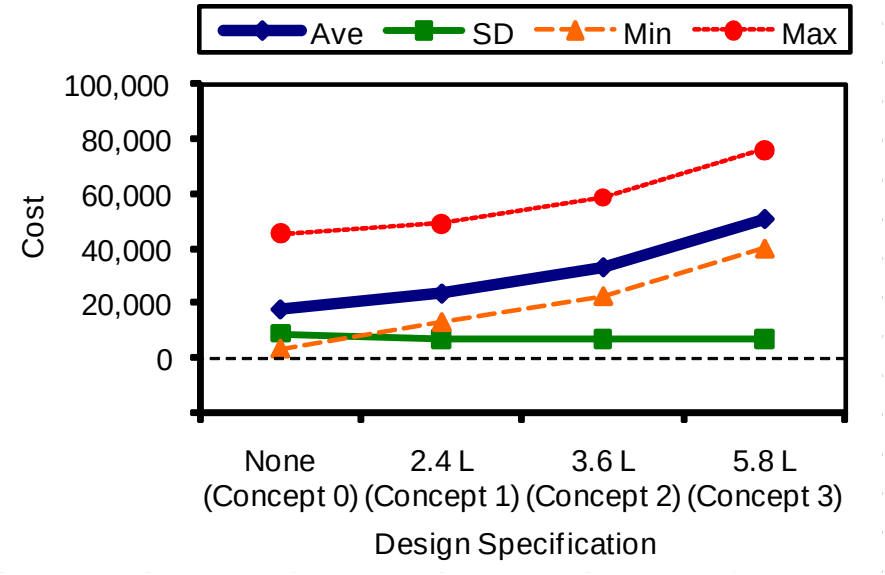


(a) Analogy-Based Cost Estimation (ABCE)

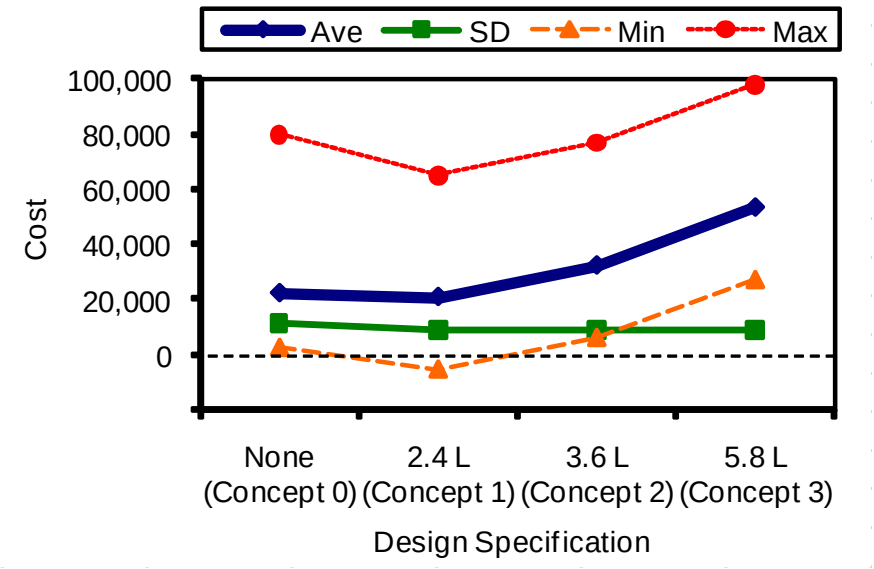


(b) Case-Based Reasoning without Adjustment (CBR)

Fig. 8 Comparison of Various Approaches



(c) Case-Based Reasoning with Adjustment (CBR-A)



(d) Regression Analysis (RA)

Fig. 8 (Continued) Comparison of Various Approaches

4. LEAVE-ONE-OUT CROSS-VALIDATION FOR ACCURACY OF COST ESTIMATION AND RELIABILITY OF COST DISTRIBUTION

To quantitatively verify observations in section 3.5 and evaluate how the four methods (ABCE, CBR, CBR-A, and RA) accurately estimate costs and reliably generate cost distributions of an SUV concept, a leave-one-out cross-validation is performed. In the leave-one-out cross-validation, one of the 85 SUVs is removed from the original knowledge base of 345 automobiles, assuming it is a new concept, and each method (ABCE, CBR, CBR-A, or RA) is applied to the remaining 344 automobiles. To study the effects of design attributes, for each method, leave-one-out cross-validation is performed once with a knowledge base consisting of only with two performance attributes (automobile type and fuel efficiency) and once with a knowledge base consisting of the same two performance attributes and an additional design attribute (engine capacity).

To evaluate an accuracy of a cost estimation, an estimated cost of each concept ($\hat{c}_k$, k=1, …, 85) is compared with the actual cost (c_k). To evaluate a reliability of a cost distribution, a cost distribution is constructed and evaluated by how well this distribution contains the actual cost c_k. This procedure is repeated 85 times (k=1, …, 85), each time assuming a new SUV as a concept.

The accuracy of a cost estimation is compared in terms of "mean magnitude of error" (MME) in Eq. 4 and "mean magnitude of relative error" (MMRE) in Eq. 5. Cost estimation is more accurate if both MME and MMRE are close to zero.

$$MME = \frac{1}{85}\sum_{k=1}^{85}\left|C_k - \hat{C}_k\right| \quad (4)$$

$$MMRE = \frac{1}{85}\sum_{k=1}^{85}\left|C_k - \hat{C}_k\right| \times 100 \quad (5)$$

The reliability of a cost distribution is evaluated in terms of frequency that a 95% data range (a range between 2.5 and 97.5 percentiles) of a normal distribution captures the actual cost c_k. "Reliability of distribution" (R) is defined in Eq. 6, in which I_k=1 if c_k is within the 95% range and I_k=0 if otherwise. Reliability is compared by "magnitude of reliability" (MR) in Eq. 7. Cost distribution is more reliable if R is close to 95% or when MR is close to 0 because, by definition, 95% range should contain only 95% of data. If R is larger than 95%, the distribution is wider than the optimum, and if R is smaller than 95%, the distribution is narrower than the optimum.

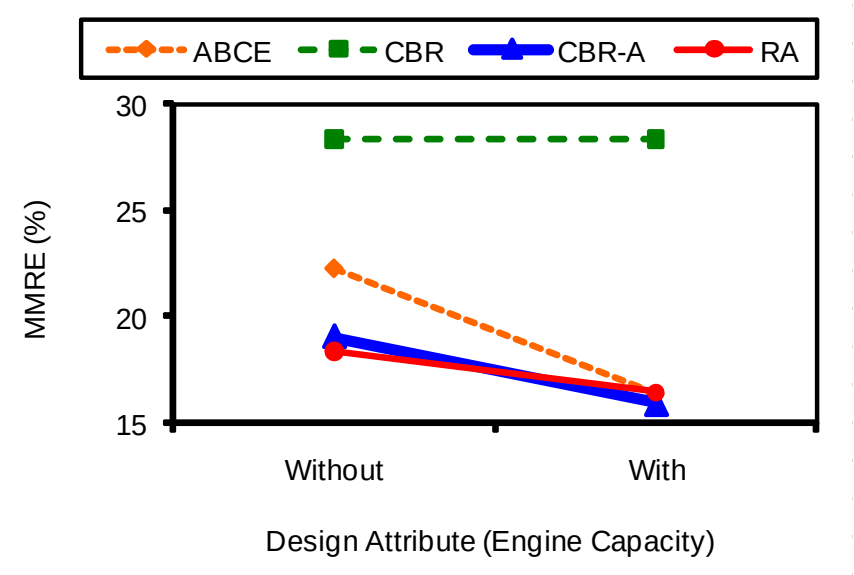
$$R = \frac{1}{85}\sum_{k=1}^{85} I_k \times 100 \quad (6)$$

$$MR = \left|95 - R\right| \quad (7)$$

Table 3 summarizes leave-one-out cross-validation results and Fig. 9 plots two evaluation measures: MMRE to evaluate accuracies of cost estimations and MR to evaluate reliabilities of cost distributions.

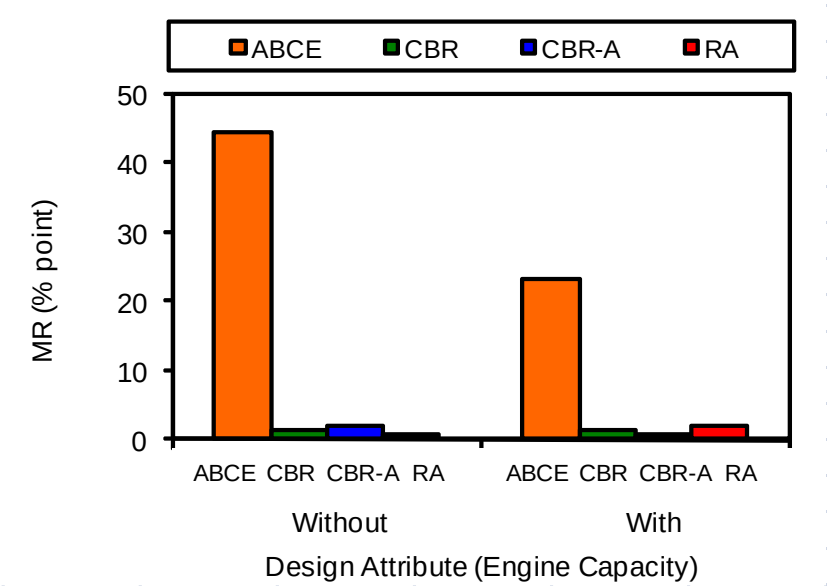
Table 3 Leave-One-Out Cross-Validation Results

Design attribute	Evaluation	Measure	Methods			
			ABCE	CBR	CBR-A	RA
None	Cost estimation	MME ($)	7,800	8,500	6,279	**6,074**
		MMRE (%)	22.2	28.3	19.0	**18.4**
	Cost distribution	R (%)	50.6	96.5	92.9	**94.1**
		MR (% point)	44.4	1.5	2.1	**0.9**
Engine capacity	Cost estimation	MME ($)	5,578	8,500	**5,276**	5,494
		MMRE (%)	16.3	28.3	**15.9**	16.4
	Cost distribution	R (%)	71.8	96.5	**94.1**	92.9
		MR (% point)	23.2	1.5	**0.9**	2.1



(a) Mean Magnitude of Relative Error (MMRE)

Fig. 9 Leave-One-Out Cross-Validation Result



(b) Magnitude of Reliability (MR)

Fig. 9 (Continued) Leave-One-Out Cross-Validation Result

4.1 Discussion of the Leave-One-Out Cross-Validation Results

The results of leave-one-out cross-validation in Table 3 and Fig. 9 indicate that case-based reasoning with adjustment (CBR-A) performs best in both accuracy of cost estimation and reliability of cost distribution when a design attribute (engine capacity) is specified for the concept in addition to performance attributes (automobile type and fuel efficiency). On the other hand, regression analysis (RA) performs best in both accuracy of cost estimation and reliability of cost distribution when only performance attributes are specified for the concept.

Although analogy-based cost estimation (ABCE) provides reasonably accurate cost estimations, cost distributions generated by analogy-based cost estimation are not

reliable; i.e., MRs are large whether or not the design specification (engine capacity) is defined. This indicates that retrieving up to three automobiles similar to a concept (or four if there are automobiles with the same distance) may be too few to construct reliable distributions.

Although case-based reasoning without adjustment (CBR) provides reasonably reliable cost distributions, cost estimation is not accurate; i.e., MMREs are large whether or not the design specification (engine capacity) is defined. This indicates that in addition to retrieving a large number of automobiles similar to a concept, costs of retrieved automobiles need to be adjusted in order to accurately estimate the cost of a concept.

5. CONCLUSION AND FUTURE WORK

This paper studied advantages of case-based reasoning approaches to estimate cost and model cost uncertainty of a new product concept when a knowledge base is homogeneous (i.e., no or few missing data). The comparison of *analogy-based cost estimation (ABCE), case-based reasoning without cost adjustment (CBR), case-based reasoning with cost adjustment (CBR-A),* and *regression analysis (RA)* using the leave-one-out cross-validation indicated that case-based reasoning with adjustment performed best when a design attribute (engine capacity) was specified for the concept in addition to performance attributes (automobile type and fuel efficiency).

Analogy-based cost estimation provided reasonably accurate cost estimations, but it generated unreliable cost distributions. Case-based reasoning without adjustment provided inaccurate cost estimation although it generated reasonably reliable cost distributions.

To further establish case-based reasoning with cost adjustment, optimum product retrieval methods (other clustering and classification methods) and their product retrieval criteria need to be studied together with the optimum number of design attributes for specifying a concept. These case-based reasoning conditions need to be compared with regression analysis for accuracies of cost estimations and reliabilities of cost distributions. These studies are left for future work.

To estimate cost and model cost uncertainty of a product concept that does not exist in the past and the current marketplace (e.g., an innovation), further research is needed to improve the current case-based reasoning approach. This avenue of research

may need to examine functionally similar but physically different products in multiple product categories and determine whether the costs of these products may be used to estimate the cost of the concept. This is another topic for future work.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the University of Missouri Research Board and the Intelligent Systems Center at Missouri University of Science and Technology for supporting this research.

REFERENCES

Al-Shihabi T, Zeid I (1998) A Design-plan-oriented methodology for applying case-based adaptation to engineering design. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 12(5):463-478

An S-H, Kim G-H, Kang K-I (2007) A case-based reasoning cost estimating model using experience by analytic hierarchy process. Building and Environment 42(7):2573-2579

Angelis L, Stamelos I (2000) A simulation tool for efficient analogy based cost estimation. Empirical Software Engineering 5(1):35-68

Auer M, Trendowicz A, Graser B, Haunschmid E, Biffl S (2006) Optimal project feature weights in analogy-based cost estimation: improvement and limitations. IEEE Transactions on Software Engineering 32(2):83-92

Bardasz T, Zeid I (1991) Applying analogical problem solving to mechanical design. Computer Aided Design 23(3):202-212

Bardasz T, Zeid I (1993) DEJAVU: Case-based reasoning for mechanical design. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 7(2):111-124

Berry S, Levinsohn J, Pakes A (2004) Differentiated products demand systems from a combination of micro and macro data: the new car market. Journal of Political Economy 112(1):68-105

Berry S, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. Econometrica 63(4):841-890

Braxton P, Coleman D (2007) Hilbert's problem for cost estimating. Joint ISPA/SCEA International Conference, New Orleans, LA

Fox EP, Safie F (1992) Statistical characterization of life drivers for a probabilistic design analysis. AIAA/SAE/ASME/ASEE 28th Joint Propulsion Conference and Exhibit, Nashville, TN

Hamaker J (1995) Parametric estimating. In: Stewart RD., Wyskida RM, Johannes JD (eds), Cost estimator's reference manual. John Wiley & Sons, New York, NY

Hastie T, Tibshirani R, Friedman, J (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York, NY

Jeffery R, Ruhe M, Wieczorek I (2000) A comparative study of two software development cost modeling techniques using multi-organizational and company-specific data. Information and Software Technology 42(14):1009-1016

Kelly C, Rice, J (1990) Monotone smoothing with application to dose-response curves and the assessment of synergism. Biometrics 46(4):1071-1085

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

Kolodner JL (1993) Case-based reasoning. Morgan Kaufmann Publishers, San Mateo, CA

Lee KH, Lee K-Y (2002) Agent-based collaborative design system and conflict resolution based on case-based reasoning approach. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 16(2):93-102

Maher ML, Zhang DM (1993) CADSYN: a case-based design process model. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 7(2):97-110

McCarthy P (1996) Market price and income elasticities of new vehicle demands. The Review of Economics and Statistics 78(3):543-547

Mendes E, Watson I, Triggs C, Mosley N, Counsell S (2003) A comparative study of cost estimation models for web hypermedia applications. Empirical Software Engineering 8(2):163-196

Michalek JJ, Papalambros PY, Skerlos SJ (2004) A study of fuel efficiency and emission policy impact on optimal vehicle design decisions. Journal of Mechanical Design 128(6):1196-1204

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models. Irwin, Chicago, IL

Otto KN, Wood KL (2001) Product design: techniques in reverse engineering and new product development. Prentice-hall Inc., Upper Saddle River

Pahl G, Beitz W (1996) Engineering design-a systematic approach. Springer Limited, London

Reich Y, Kapeliuk A (2004), Case-based reasoning with subjective influence knowledge. Applied Artificial Intelligence 18(8):735-760

Roderman S, Tsatsoulis C (1993) Panda: a case-based system to aid novice designers. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 7(2):125-133

Rosenman M (2000) Case-based evolutionary design. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 14(1):17-29

Shepperd M, Schofield C (1997) Estimating software project effort using analogies. IEEE Transactions on Software Engineering 23(12):736-743

Shiau C-SN, Michalek JJ (2009) Should designers worry about market systems? Journal of Mechanical Design, 131(1):011011 (9 pages)

Shiva Kumar H, Krishnamoorthy C S (1995) A framework for case-based reasoning in engineering design. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 9(3):161-182

Takai S (2009) A case-based reasoning approach toward developing a belief about the cost of concept. Research in Engineering Design 20:255-264

Train KE, Winston C (2007) Vehicle choice behavior and the declining market share of us automakers. International Economic Review 48(4):1469-1496

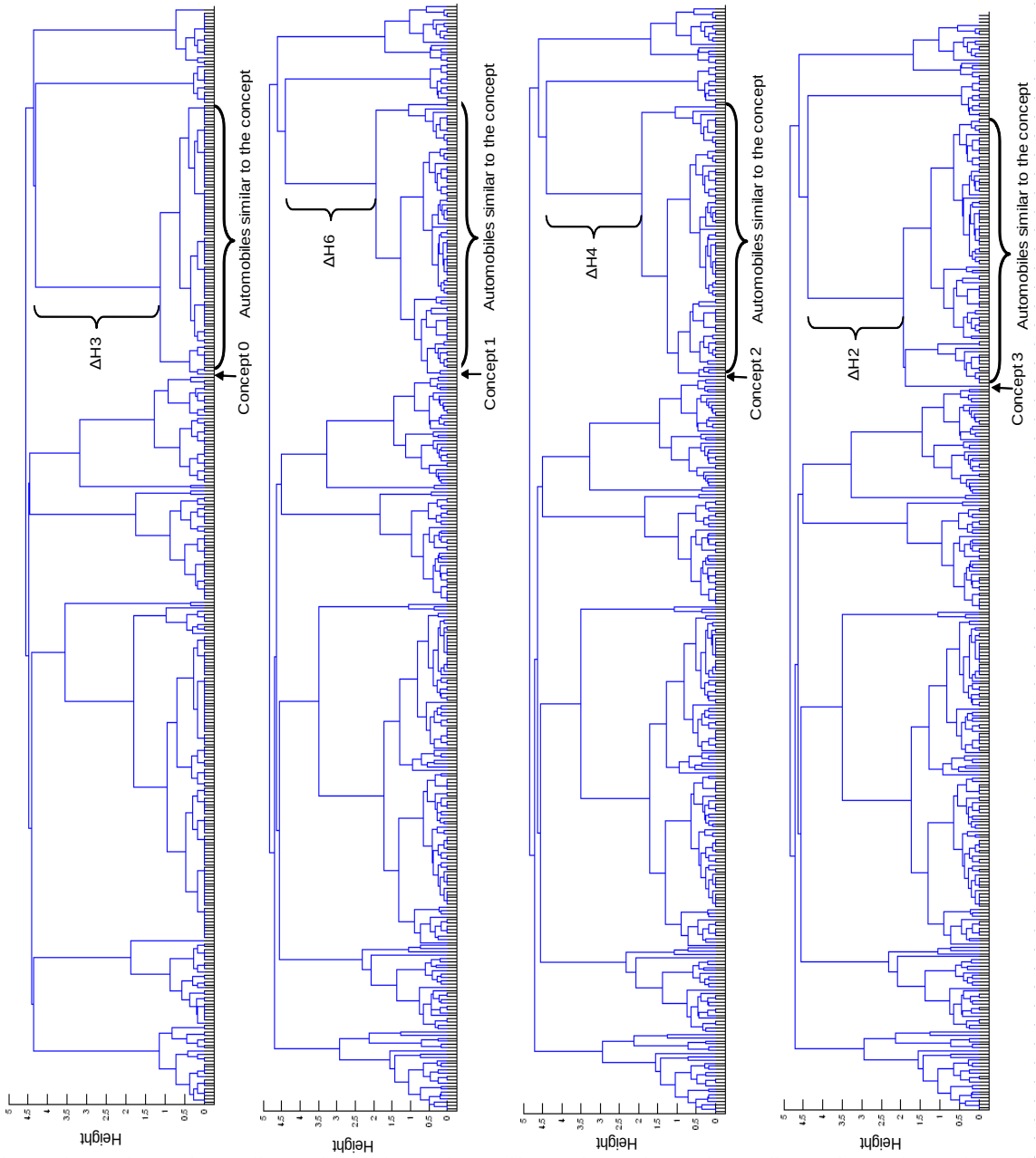
Ulrich KT, Eppinger SD (2004) Product design and development. McGraw-Hill

Williams N, Azarm S, Kannan PK (2008) Engineering product design optimization for retail channel acceptance. Journal of Mechanical Design 130(6):061402 (10 pages)

Wood WH, Agogino AM (1996) Case-based conceptual design information server for concurrent engineering. Computer Aided Design 28(5):361-369

Wyskida RM (1995) Statistical techniques in cost estimation. In: Stewart RD, Wyskida RM, Johannes JD (eds) Cost estimator's reference manual. John Wiley & Sons, New York, NY

APPENDIX



(a) Concept 0 (b) Concept 1 (c) Concept 2 (d) Concept 3

Fig. 10 Dendrograms for Automobile Retrieval

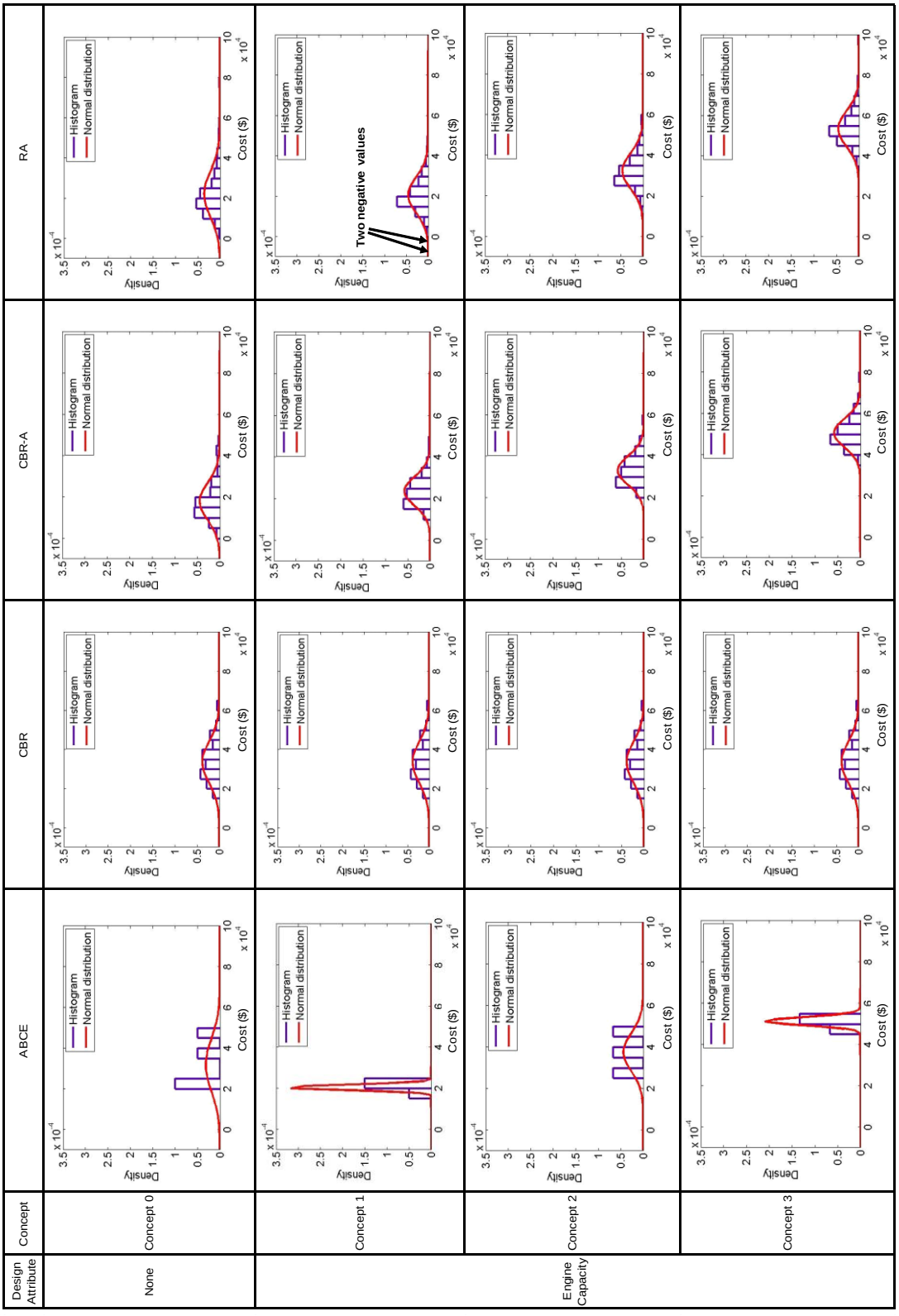


Fig. 11 Cost Distributions

PAPER

2. OPTIMUM NUMBER OF DESIGN ATTRIBUTES FOR COST ESTIMATION AND COST UNCERTAINTY MODELING

Karan Banga
Department of Mechanical Engineering
Missouri University of Science and Technology
Rolla, MO 65409
kbcgf@mst.edu

Shun Takai (Corresponding Author)
Department of Mechanical Engineering
Missouri University of Science and Technology
Rolla, MO 65409
takais@mst.edu

ABSTRACT

Case-based reasoning (CBR) is an approach which uses old cases/experiences to understand and solve new problems. In CBR, a previous case similar to the current case is used to generate a solution for the current case and usually involves adaptation of the generated solution to suit the current case. The CBR approach consists of creating a knowledge-base (or database) containing past cases (products), defining a new case, retrieving cases similar to the new case, and adjusting the solution (cost) of the retrieved cases to the new case. This paper compares CBR approach with regression analysis approach in studying the effects of varying design attribute specifications on cost estimation accuracy and cost distribution reliability. These approaches are compared and

effects of defining a concept with varying design attribute specifications are studied by applying leave-one-out cross-validation to a knowledge-base of automobiles.

KEYWORDS: Cost, concept, case-based reasoning, clustering, histogram, distribution, leave-one-out cross validation

1. INTRODUCTION

Product development involves a sequence of decision making steps that must be taken under uncertainty, including selection of a product concept. Factors influencing the choice of a concept include market size, market share, and cost.

Two popular approaches are available to determine the cost of a product, the cost modeling approach and the regression analysis (RA) approach. The cost modeling approach estimates product cost by adding costs associated with various product attributes and processes. This estimate takes into account part costs, assembly costs, and overhead costs calculated from detailed product information such as the bill of material (BOM) and the design specifications [Ulrich and Eppinger 2004; Otto and Wood 2001; Pahl and Beitz 1996]. This approach requires detailed product design and manufacturing process information that is treated as an uncertainty in the concept selection stage; thus it may not be the optimal method to estimate and model the cost of the final concept.

Different from the cost modeling approach, RA [Hamaker 1995; Wyskida 1995] estimates product cost from product-level information (i.e., product specifications), and does not necessarily require detailed design and manufacturing process information such as BOM, part costs, and assembly costs. RA approximates a cost estimation relationship (CER) in the form of an equation between one dependent variable (cost) and one or more independent variables (attributes influencing cost) [Michalek et al. 2004; Williams et al. 2008; Shiau et al. 2009a]. Once the CER is established, the estimated cost of a concept is calculated by substituting its product information into the CER. Although, RA has a strong theoretical foundation [Neter et al. 1996], a study [Braxton and Coleman 2007]

identified various challenges in applying RA in practice. One of the many challenges is the poor quality of the database (e.g., missing data and outliers) which could lead to inaccurate cost estimates.

Analogy-based cost estimation (ABCE) is a relatively new approach that has been proposed to apply case-based reasoning (CBR) [Kolodner 1993] in cost estimation. Similar to CBR methodology, ABCE consists of creating a database containing past cases, defining a new case (concept), retrieving up to three cases similar to the new case, and adjusting the solution of the retrieved cases to the new case. ABCE does not rely on detailed design and manufacturing information and is thus particularly suitable for estimating the cost of a new product in the concept selection stage.

ABCE has been used to estimate the cost of new software projects [Shepperd and Scofield 1997; Angelis and Stamelos 2000; Mendes et al. 2003; Auer et al. 2006; Jeffery et al. 2000] and to estimate costs of construction projects [Kim et al. 2004; An et al. 2007]. ABCE has also been used in design problems [Bardasz and Zeid 1991; Bardasz and Zeid 1993; Roderman and Tsatsoulis 1993; Maher and Zhang 1993; Shiva Kumar and Krishnamoorthy 1995; Rosenman 2000; Wood and Agogino 1996; lee and Lee 2002; Al-Shahibi and Zeid 1998]. Lately, CBR (more than three similar retrieved cases) has been used to estimate cost of concept products [Takai 2009, Banga and Takai 2010].

It has been a matter of great debate as to which is a better method for cost estimation, CBR or RA. Jeffery et al. 2000 compared the differences in accuracies of cost estimations between ordinary least squares regression and analogy-based estimation using data from multiple companies as well as company-specific data. Although no significant differences were observed between the two techniques as applied to

company-specific data, ordinary least squares regression performed significantly better than analogy-based estimation in the case of multi-company data. Takai 2009 compared the accuracies of cost estimations between CBR approach and analogy-based cost estimation with and without a linear adjustment using a heterogeneous knowledge-base (i.e., with missing data). The results showed that CBR provided slightly more accurate cost estimations than analogy-based cost estimations. Since a heterogeneous knowledge-base was used, it was possible that RA was not able to provide accurate cost estimations and thus, no comparison was made between the two approaches. Banga and Takai 2010 compared CBR approaches (with and without cost adjustment) with ABCE and RA. The analysis was carried out with a homogeneous knowledge-base (i.e., with no missing data) and the comparison was carried out using leave-one-out cross-validation technique. The accuracy of cost estimation and reliability of cost uncertainty modeling using the different methods was then established. The results showed that CBR with cost adjustment (CBR-A) performed better than RA when a design attribute (engine capacity) was specified for the concept in addition to performance attributes (automobile type and fuel efficiency). ABCE provided reasonably accurate cost estimations, but it generated unreliable cost distributions. CBR without adjustment provided inaccurate cost estimation although it generated reasonably reliable cost distributions.

To further establish CBR with cost adjustment (CBR-A), this paper studies the optimum number of design attributes for specifying a concept. A comparison is made between CBR-A and RA approaches in studying the effects of varying design attribute specifications on cost estimation accuracy and cost distribution reliability. The CBR approach uses hierarchical clustering to retrieve from the knowledge-base as many

products as possible that are similar to the new concept. It also uses RA to parallel adjust the cost of the retrieved products and thus constructs a distribution for the cost of a concept.

The remainder of this paper is organized as follows: Section 2 proposes a CBR methodology: constructing a knowledge-base, defining a product concept, retrieving similar products, adjusting the cost to specific attributes, and fitting a distribution to the adjusted costs; Section 3 compares the accuracy between CBR approach and RA using leave-one-out cross validation; Section 4 validates the results obtained in Section 3 using an example. Finally, Section 5 discusses directions for future work.

2. METHODOLOGY: CASE-BASED REASONING

Figure 1 illustrates the four steps to the CBR methodology: construction of a knowledge-base that contains past and current products, definition of a concept, retrieval of products similar to the concept, and generation of a cost distribution for the concept.

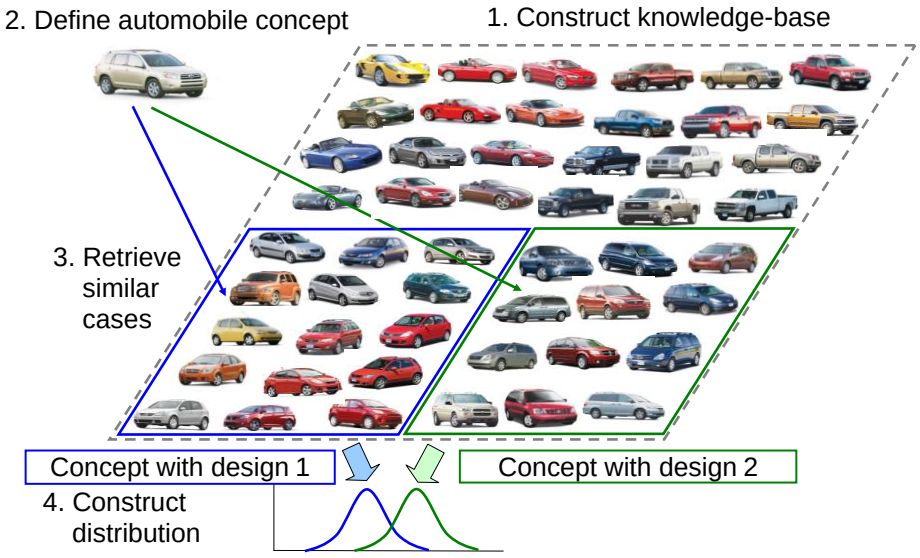


Fig. 1 Case-Based Reasoning Process Flow

2.1 Knowledge-Base Construction

The first step is to construct a knowledge-base of all the past and current products. The knowledge-base includes the products, together with their attributes and specifications. Attributes are the properties defining a product, and specifications are the specific values of those attributes. For example, in the case of an automobile, fuel efficiency is an attribute for which 25 miles/gallon is the specification. The attributes may

be numerical (quantitative) or categorical (qualitative). They may also be classified as performance or design attributes. Performance attributes describe the product functionality which directly influences customers' purchasing decision. In contrast, design attributes describe the design characteristics and manufacturing methodology that makes the functionality possible. For example, fuel efficiency is defined as a performance attribute, and engine capacity is a design attribute. Table 1 shows a portion of the complete knowledge-base used in the analysis.

Table 1 Portion of the Complete Knowledge-Base

	Automobile Type	Fuel Efficiency (miles/gallon)	Engine Capacity (liter)	Cost($)
Porsche Boxster	Convertibles	22	2.7	21,554
Toyota RAV4 4-cyl.	SUV	23	2.4	20,994
Honda CR-V	SUV	21	2.4	17,483
Mitsubishi Lancer ES	Small Cars	25	2.0	14,332
Mercedes-Benz E320	Sedans	29	3.0	53,100
Chrysler Town & Country Limited	Minivans	17	4.0	37,595
Toyota Prius Touring	Wagons	42	1.5	22,575
Dodge Dakota	Pickups	14	4.7	29,053
Ford Mustang V8	Sporty	20	4.6	27,413

The knowledge-base was constructed by benchmarking automobiles sold in the U.S. from 2003 through 2009. The knowledge-base contained 345 automobiles, with 86 attributes. The attributes included automobile type (SUV, small car, sedan, minivan, wagon, pickup and sports car), number of cylinders, engine capacity, number of side airbags, acceleration, braking, fuel efficiency, roadside aid, and many more. The data on

the automobiles and their attributes were gathered primarily from credible sources on the internet, and the data on costs were collected from the annual reports of individual automobile companies available online. The annual automotive revenue and operating income were gathered directly from the annual reports and subsequently, automotive cost to the company was found by subtracting the operating income from the annual revenue.

Automobiles costs were calculated as follows:

$$Cost = pricex\frac{total\ Cost}{total\ revenue} \tag{1}$$

Costs, although recorded in the knowledge-base, were not used in the initial CBR analysis. Costs were used only to construct a distribution for the concept once the automobiles similar to the concept were retrieved.

2.2 Product Concept Definition

The next step in the CBR approach is to define a product concept by attributes and corresponding specifications. These attributes could be identified by first conducting market surveys and hence, identifying customer needs and then converting these needs to corresponding performance attributes and specifications. In addition, the design attributes and specifications may be defined by designers.

2.3 Product Retrieval

The CBR method in this paper relies on hierarchical clustering analysis to retrieve products similar to the concept. Hierarchical clustering permits retrieval from the

knowledge-base of all products similar to the concept. It can be broken down into three steps: (1) Data matrix creation: If the knowledge-base contains *I* number of products and *J* number of attributes, then the data matrix, will have *I* rows and *J* columns. Initially, knowledge-base varies from *i*=1to *I*. When it is used for the purpose of cost estimation, however, an additional row is added for the concept; this row becomes *i*=0. Ones and zeros are used for the categorical attribute (type of automobile) to quantify whether types match or not between two automobiles. Table 2 shows a portion of the modified knowledge-base.

Table 2 Portion of the Modified Knowledge-Base

Row Number		Type								Fuel Efficiency (miles/gallon)	Engine Capacity (liter)
		Convertibles	SUV	Small Cars	Sedans	Minivans	Wagons	Pickups	Sporty		
i=0	Concept	0	1	0	0	0	0	0	0	25	2.4
i=1	Porsche Boxster	1	0	0	0	0	0	0	0	22	2.7
i=2	Toyota RAV4 4-cyl.	0	1	0	0	0	0	0	0	23	2.4
i=3	Honda CR-V	0	1	0	0	0	0	0	0	21	2.4
i=4	Mitsubishi Lancer ES	0	0	1	0	0	0	0	0	25	2.0
i=5	Mercedes-Benz E320	0	0	0	1	0	0	0	0	29	3.0
i=6	Chrysler Town & Country Limited	0	0	0	0	1	0	0	0	17	4.0
i=7	Toyota Prius Touring	0	0	0	0	0	1	0	0	42	1.5
i=8	Dodge Dakota	0	0	0	0	0	0	1	0	14	4.7
i=9	Ford Mustang V8	0	0	0	0	0	0	0	1	20	4.6

(2) Distance matrix creation: The distance matrix is created from the data matrix by calculating the Euclidian between the concept and each product, and between each pair of products. To make sure that the process is not biased toward the units used and so that all the attributes have the same degree of influence for similar automobile retrieval, all the attributes are standardized. Table 3 shows the corresponding portion of the knowledge-base with the standardized values.

Table 3 Portion of the Standardized Knowledge-Base

	Type								Fuel Efficiency (miles/gallon)	Engine Capacity (liter)
	Convertibles	SUV	Small Cars	Sedans	Minivans	Wagons	Pickups	Sporty		
Concept	-0.4	2.6	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.6	-0.6
Porsche Boxster	2.6	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.1	-0.4
Toyota RAV4 4-cyl.	-0.4	2.6	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.2	-0.6
Honda CR-V	-0.4	2.6	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.1	-0.6
Mitsubishi Lancer ES	-0.4	-0.4	2.6	-0.4	-0.4	-0.4	-0.4	-0.4	0.6	-0.9
Mercedes-Benz E320	-0.4	-0.4	-0.4	2.6	-0.4	-0.4	-0.4	-0.4	1.3	-0.1
Chrysler Town & Country Limited	-0.4	-0.4	-0.4	-0.4	2.6	-0.4	-0.4	-0.4	-0.8	0.7
Toyota Prius Touring	-0.4	-0.4	-0.4	-0.4	-0.4	2.6	-0.4	-0.4	3.5	-1.3
Dodge Dakota	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	2.6	-0.4	-1.3	1.3
Ford Mustang V8	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	2.6	-0.3	1.2

The following equation expresses the standardization technique.

$$s_{i,j} = \frac{x_{i,j} - \mu_a}{\sigma_a} \tag{2}$$

where $s_{i,j}$ is the standardized attribute value, $x_{i,j}$ is the original attribute value, and μ_a and σ_a are the average and standard deviation values respectively across an attribute *a*, where *a* varies from 1 to 3. For the first attribute type (*a*=1), the average and standard deviation values were calculated for the first eight columns because they all belong to the same attribute, type of automobile. For the remaining two attributes, the same were calculated across their respective columns. The total number of products in the knowledge-base, including the concept, is denoted by *i*, where *i* varies from 0 to 345 (*i*=0 for the concept).

Once the knowledge-base is standardized, the Euclidian distance, δ between two products, p' and p is calculated as:

$$\delta(p, p') = \sqrt{\sum_{j=1}^{J} w_j \left(s_{i,j} - s_{i',j}\right)^2} \tag{3}$$

where $s_{i,j}$ is the standardized attribute value of product p, $s_{i',j}$ is the standardized attribute value of product p', and w_j is the weight of attribute J. Here, weights for all attributes were set to 1.

(3) Similar product retrieval: Finally, hierarchical clustering is applied to the distance matrix to group products with similar attribute information and thus, obtains products similar to the concept. Hierarchical clustering is used to generate tree figures, also called as dendrograms, based on the distances calculated in the distance matrix. In a dendrogram, the height at which two products, two clusters, or a product and a cluster are grouped together indicates the distance between them. The smaller the distances between products, the more similar the products, and therefore, the lower their group level.

The three methods most commonly used for hierarchical clustering are: the single linkage method, the complete linkage method, and the average linkage method. The single linkage method calculates, element by element, the distances between two clusters and uses the smallest distance as the distance between two clusters. On the contrary, the complete and average linkage methods use the largest and the average distances as the distance between two clusters respectively. In this analysis, average linkage has been used as it has been claimed to have a statistical consistency property which is violated by the other two methods [Kelly and Rice 1990].

Figure 2 shows an example dedrogram for some Concept C. The linkage heights are labeled H1, H2, and H3. Height H1 corresponds to the linkage height of the similar automobiles grouped one level below the concept; H2 corresponds to the linkage height one level above the concept; and H3 corresponds to the linkage height two levels above the concept. The term ΔH1 represents the difference in linkage heights between H2 and H1 (ΔH1=H2-H1), and so on. The largest differences in linkage heights, ΔH is used to determine products similar to the concept and are thus retrieved from the knowledge-base for cost estimation purpose. In figure 2, ΔH1 is larger than ΔH2 and thus, the two

products P2 and P3 are grouped with concept C and are thus retrieved from the knowledge-base.

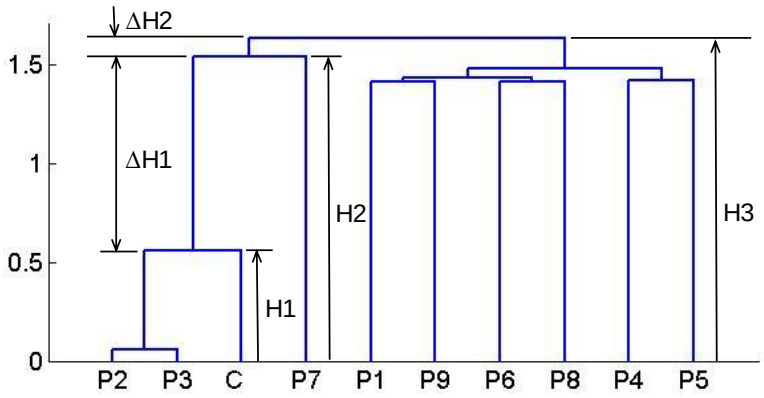


Fig. 2 Example Dendrogram

2.4 Cost Adjustment and Distribution Fitting

The final step in the CBR is to generate cost distributions after adjusting costs of the retrieved products. Three cost adjustment methods may be used to estimate the cost of a concept: no adjustment, linear adjustment, and parallel adjustment. In the past, the CBR applications have used no adjustment. They have calculated point estimates (or averages) from the cost of retrieved products without first adjusting the cost. Linear adjustment first identifies the attribute of the retrieved products that is most closely correlated with the cost, and then calculates a ratio of the attribute specification between the new concept to that of a retrieved product. Finally, it adjusts the cost of the retrieved product in proportion to this ratio. Point estimates are then obtained from these adjusted costs.

The CBR proposed here uses parallel adjustment. In this, a regression model is obtained by applying a regression analysis to the costs of the retrieved products, and then, these retrieved costs are adjusted parallel to the regression model. Finally, these adjusted costs are used to estimate the cost and generate the cost distribution of the concept. The regression model could be a line (in the case of a single cost-relevant attribute) or a surface (in the case of multiple numeric cost-relevant attributes). Figure 3 illustrates these three cost adjustment methodologies in the case of a single numeric cost relevant attribute. In this paper, normal distribution has been used, but other distributions may also be used.

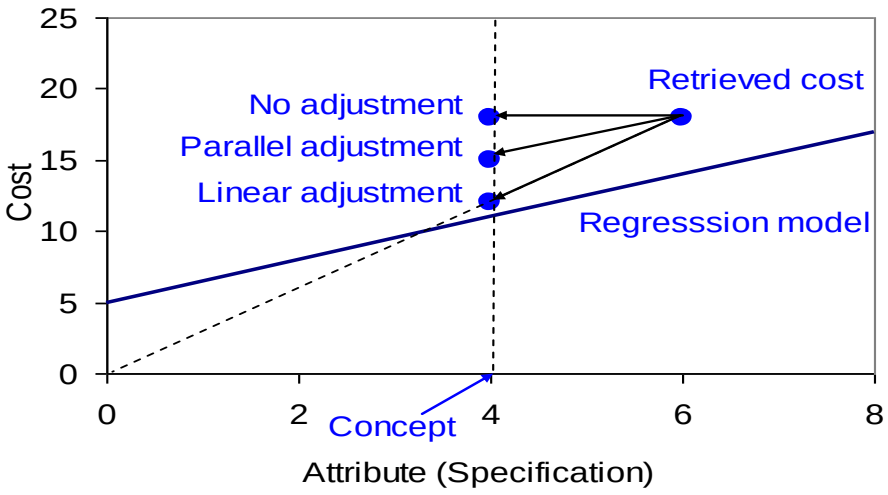


Fig. 3 Cost Adjustment

Once the costs of the retrieved products are adjusted, a distribution must be fitted to the adjusted costs. Generally, histograms of the desired property are constructed first,

and a distribution is then fitted to them. Normal distribution is used most often, but other distributions may also be used. The present analysis uses normal distribution to explain the distribution of the concept cost.

3. COMPARISON OF CBR-A AND RA FOR COST ESTIMATION AND COST UNCERTAINTY MODELING

Using a knowledge-base of automobiles, CBR-A is compared with RA for accuracies of cost estimations and reliabilities of cost distributions for four levels of design specifications: no design attribute, one design attribute, two design attributes and three design attributes. The design attributes chosen for the analysis were engine capacity (liter), accident alert system (available, not available), and type of supercharging (turbo, super, non-turbo and non-super). While the first design attribute is numerical, the remaining two are categorical.

The design attributes were chosen as a result of stepwise regression analysis. The costs were regressed against the complete set of design specifications to find the most significant design attributes. The p values of these attributes were then checked. Finally, the attributes with the lowest three *p* value, in this case engine capacity followed by accident alert system and finally, type of supercharging were identified. The purpose of using a design attribute with the lowest *p* values (i.e., the attribute that is the most significantly associated with cost) was to eliminate subjective judgment in design attribute selection.

However, the performance attributes, namely, automobile type and fuel efficiency were chosen because they had been identified as critical to modeling consumer demand for automobiles in the past. McCarthy 1996 proposed that vehicle size and type have a direct impact on customer decisions. Train and Winston 2007 have discussed how Japanese automakers gained an upper hand over US manufacturers because of the higher

fuel efficiency of their vehicles. Berry et al. 1995, 2004 have proposed that higher fuel efficiency and vehicle size drive customer demand. These studies suggest, therefore, that these two performance attributes have significant impact on customers' automobile purchasing decisions.

The reference method was regression analysis (RA). It was performed on all automobiles in the knowledge-base and the costs of all automobiles were adjusted parallel to the regression model. The second approach used CBR with cost adjustment (CBR-A). CBR analysis was applied to the complete knowledge-base, similar automobiles were retrieved, and finally the costs of the retrieved automobiles were adjusted parallel to a regression model obtained from applying a regression analysis to the retrieved automobiles.

3.1 Leave-One-Out Cross-Validation for Accuracy of Cost Estimation and Reliability of Cost Distribution

To compare cost estimations and cost, a leave-one-out cross-validation method was used. A leave-one-out is a validation technique whereby each data-point is removed from the knowledge-base and the remainder of the data-points are used to predict the desired property (i.e., cost) of the removed data-point. The data-point is then returned to the knowledge-base and the next data-point is removed. The procedure is repeated until all the data-points have been covered.

In this study, an automobile batch consisting of 85 SUVs was used. One of the 85 SUVs was removed from the original knowledge-base consisting of 345 automobiles, assuming it was a new concept, and CBR-A and RA was applied to the remaining 344 automobiles. To study the effects of varying design specifications for each method, leave-

one-out cross-validation was performed on the same knowledge-base four times (four conditions) as shown in table 4. First, it was performed with no design attributes and just the two performance attributes, automobile type and fuel efficiency. Second time, it was performed with the same two performance attributes and an additional design attribute, engine capacity (lowest p value). Third time, it was performed with the same two performance attributes, engine capacity and an additional design attribute, accident alert system (second lowest p value). Finally, it was performed with the same two performance attributes, engine capacity, accident alert system and an additional design attribute, type of supercharging (third lowest p value).

Table 4 Leave-One-Out Cross-Validation Conditions

	Performance Attributes Used	Design Attributes Used
Condition 1	Automobile Type, Fuel Efficiency	None
Condition 2	Automobile Type, Fuel Efficiency	Engine Capacity
Condition 3	Automobile Type, Fuel Efficiency	Engine Capacity, Accident Alert System
Condition 4	Automobile Type, Fuel Efficiency	Engine Capacity, Accident Alert System, Type of Supercharging

To evaluate the accuracy of a cost estimation, estimated cost of each concept $\hat{C}n$ (n=1 through 85) is compared with the actual cost *Cn*. To evaluate the reliability of a cost distribution, cost distribution is constructed and evaluated by how well this distribution

contains the actual cost C*n*. This procedure is repeated 85 times (n=1 through 85), each time assuming a new SUV as the concept.

The accuracy of a cost estimation is compared in terms of "mean magnitude of relative error" (MMRE) in Eq. 5. The closer the MMRE is to zero, the more accurate is the cost estimation.

$$MMRE = \sum_{n=1}^{85} \left(\frac{\left| C_n - \hat{C}_n \right|}{C_n} \right) \frac{100}{85} \tag{4}$$

The reliability of a cost distribution is measured in terms of a frequency that whether the actual cost of the concept, *Cn* falls within a 95% data range (2.5 and 97.5 percentile values) of a normal distribution constructed using the remaining data-points. "Reliability of distribution" (R) is defined in Eq. 5, in which I_n=1 if C*n* is within the 95% range and I_n=0 if otherwise. The distribution is said to be wider than optimum if R is greater than 95% and the distribution is said to be narrower than the optimum if R is lesser than 95%. Also, "magnitude of reliability" (MR) in given Eq. 6 and is used as the main parameter to compare reliability. By definition, the closer the MR is to zero, the more reliable is the cost distribution.

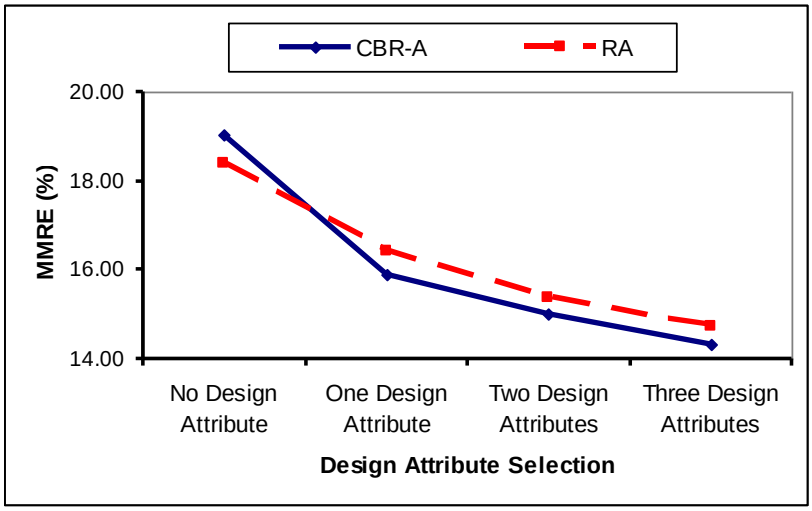
$$R = \frac{1}{85} \sum_{n=1}^{85} I_n \times 100 \tag{5}$$

$$MR = \left| 95 - R \right| \tag{6}$$

Table 5 summarizes the leave-one-out cross-validation results and Fig. 5 shows two plot measures: MMRE to evaluate accuracies of cost estimations and MR to evaluate reliabilities of cost distributions. Also, fig. 5 in appendix shows the dendrograms (for one concept) obtained for CBR-A for the different cases. The similar automobiles retrieved for CBR-A for all the cases were of type SUV. For the first two cases with no design attributes and one design attribute, all the 85 SUVs similar to the concept were grouped together. However, for the remaining two cases with two and three design attributes, the similar SUVs were grouped into two and three clusters respectively. It is interesting to note that the individual reliability for cluster three was the lowest (zero), in case of analysis with three design attributes.

Table 5 Leave-One-Out Cross-Validation Results

Number of Design Attributes	Design Attributes Added	Measure	Methods	
			CBR-A	RA
No Design Attribute	None	MMRE (%)	19.03	18.39
		MR (%)	2.06	0.88
One Design Attribute	Engine Capacity	MMRE (%)	15.90	16.42
		MR (%)	0.88	2.06
Two Design Attributes	Engine Capacity, Accident Alert System	MMRE (%)	15.00	15.36
		MR (%)	2.06	1.47
Three Design Attributes	Engine Capacity, Accident Alert System, Type of Supercharging	MMRE (%)	14.33	14.74
		MR (%)	5.59	1.47



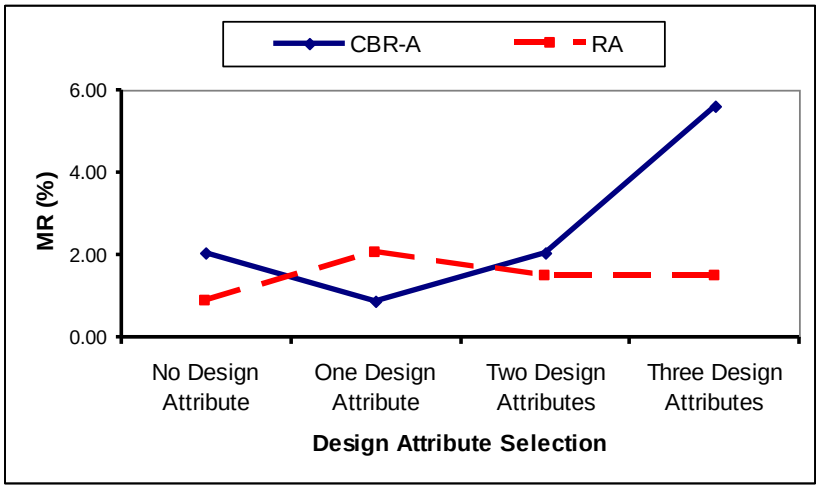


Fig. 4 leave-One-Out Cross-Validation Result

3.2 Discussion of the Leave-One-Out Cross-Validation Results

The results of the leave-one-out cross-validation in Table 3 and Fig. 4 indicate that CBR-A performs best in both accuracy of cost estimation and reliability of cost distribution when one design attribute (engine capacity) is specified for the concept in

addition to performance attributes (automobile type and fuel efficiency). On the other hand, RA performs best in both accuracy of cost estimation and reliability of cost distribution when only performance attributes are specified for the concept.

Furthermore, CBR-A performs better in accuracy of cost estimation when compared to RA with each successive addition of a design attribute.

4. EXAMPLE

To validate the results obtained in Section 3, an automobile was considered from the batch of 85 SUVs as shown in Table 6.

	Automobile Type	Fuel Efficiency (miles per gallon)	Engine Capacity (liter)	Accident alert system	Type of Supercharging
Concept	SUV	18	2.3	None	Turbo

Fig. 5 Concept Definition

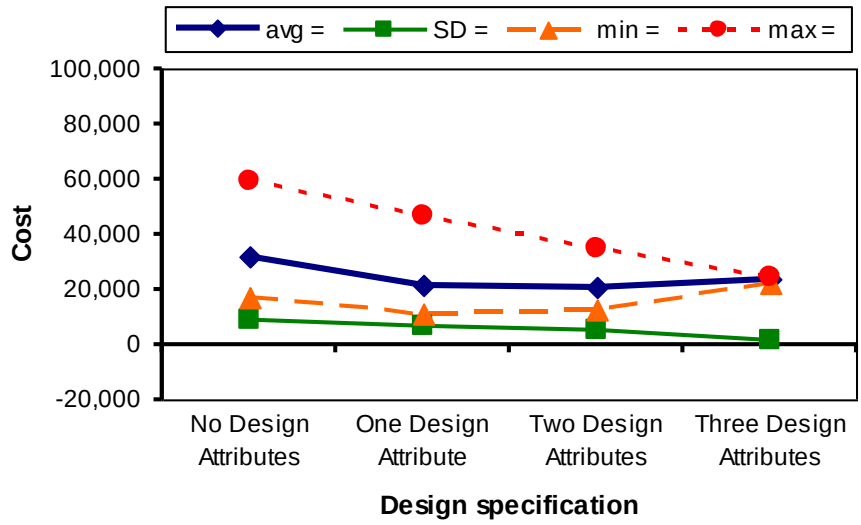
The fuel efficiency and engine capacity for the considered automobile were close to the median values for the complete batch (85 SUVs), therefore it was chosen as an ideal concept.

The next step was to apply CBR-A (explained earlier) to the chosen concept varying the design attribute selection. The hierarchical clustering analysis was applied and products similar to the concept were retrieved. Figure A of the Appendix shows the resulting dendrograms. Once similar automobiles were identified, the next step was to adjust the costs and fit a distribution to those costs. Figure B shows the distribution curves obtained for the two methods for the four design attribute selection conditions. Table 6 summarizes five statistics: the number of retrieved automobiles (n), average (avg), standard deviation (SD), minimum (min), and maximum (max), of the costs of the

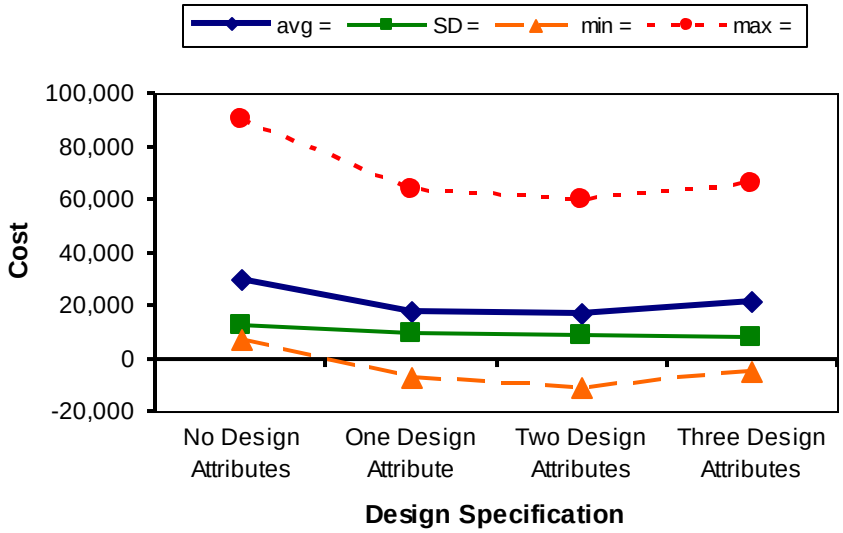
retrieved automobiles (after cost adjustment) in the two methods for four design attribute selection criteria. Figure 6 plots four statistics: avg, SD, min, and max.

Table 6 Data for the Cost Distribution Curves

Number of Design Attributes	Design Attribute Selection	**CBR-A**	**RA**
No Design Attributes	None	n = 85 avg = 31,745 SD = 8,977 min = 16,945 max = 59,101	n = 345 avg = 29,763 SD = 12,115 min = 6,826 max = 89,879
One Design Attribute	Engine Capacity	n = 85 avg = 21,629 SD = 6,958 min = 10,882 max = 46,926	n = 345 avg = 17,372 SD = 9,242 **min = -7,107** max = 63,402
Two Design Attributes	Engine Capacity Accident Alert System	n = 60 avg = 20,692 SD = 5,484 min = 12,687 max = 34,540	n = 345 avg = 16,673 SD = 8,549 **min = -10,880** max = 59,773
Three Design Attributes	Engine Capacity Accident Alert System Type of Supercharging	n = 3 avg = 23,541 SD = 1,224 min = 22,317 max = 24,765	n = 345 avg = 21,673 SD = 8,226 **min = -4,835** max = 65,717



(a) Case-Based Reasoning with Cost Adjustment (CBR-A)



(b) Regression Analysis (RA)

Fig. 6 Comparison of Various Approaches

Comparing the two methods for cost distribution reliability, standard deviation is smaller in CBR-A for all the cases, especially for the last two cases (two and three design attributes), which indicates that CBR-A generates unreliable distributions (too narrow) for cost uncertainty modeling. This is because the number of retrieved automobiles is much smaller compared to the first two cases. This result can also be observed from the cost distribution curves for CBR-A for the last two cases (Figure B in appendix). It is interesting to note that the individual reliability for cluster three was the lowest (zero), in case of analysis with three design attributes.

Comparing the two methods for cost estimation accuracy, we observe negative minimum values in RA for all the design cases except for the first case (Figure B). Since the complete knowledge-base was used for the analysis and some of the data-points could be qualified as potential outliers, RA was unable to cope with it [Braxton and Coleman 2007]. On the other hand, CBR-A retrieved the similar automobiles before adjusting the costs, and thus avoiding the under-adjustment.

5. CONCLUSION AND FUTURE WORK

This paper studied the optimum number of design attributes in defining a concept using case-based reasoning with cost estimation (CBR-A) to estimate cost and model cost uncertainty of a new product concept. A comparison was made between CBR-A and regression analysis (RA) approaches using a homogeneous (i.e., no or few missing data) knowledge-base and a leave-one-out cross-validation technique was used for the comparison. The results showed that CBR-A performed best when one design attribute (engine capacity) was specified for the concept in addition to performance attributes (automobile type and fuel efficiency). Further, it was observed that CBR improved at cost estimation but became worse at cost distribution reliability with each successive addition of a design attribute.

To further establish CBR-A, other product retrieval methods (other clustering and classification methods) and their product retrieval criteria need to be studied together with the optimum number of design attributes for specifying a concept. These case-based reasoning conditions need to be compared with regression analysis for accuracies of cost estimations and reliabilities of cost distributions. These studies are left for future work.

To estimate cost and model cost uncertainty for a new product of a type not yet introduced to the market, further research in the current CBR approach is needed. The next step would be to examine functionally similar but physically different products in multiple product categories and determine whether the cost of these products may be used to estimate cost of the new product. This is another topic for future work.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the University of Missouri Research Board and the Intelligent Systems Center at Missouri University of Science and Technology for supporting this research.

REFERENCES

Al-Shihabi T, Zeid I (1998) A Design-plan-oriented methodology for applying case-based adaptation to engineering design. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 12(5):463-478

An S-H, Kim G-H, Kang K-I (2007) A case-based reasoning cost estimating model using experience by analytic hierarchy process. Building and Environment 42(7):2573-2579

Angelis L, Stamelos I (2000) A simulation tool for efficient analogy based cost estimation. Empirical Software Engineering 5(1):35-68

Auer M, Trendowicz A, Graser B, Haunschmid E, Biffl S (2006) Optimal project feature weights in analogy-based cost estimation: improvement and limitations. IEEE Transactions on Software Engineering 32(2):83-92

Banga K, Takai S (2010) A comparison of cost estimation and cost uncertainty modeling approaches. Research in Engineering Design (under review)

Bardasz T, Zeid I (1991) Applying analogical problem solving to mechanical design. Computer Aided Design 23(3):202-212

Bardasz T, Zeid I (1993) DEJAVU: Case-based reasoning for mechanical design. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 7(2):111-124

Berry S, Levinsohn J, Pakes A (2004) Differentiated products demand systems from a combination of micro and macro data: the new car market. Journal of Political Economy 112(1):68-105

Berry S, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. Econometrica 63(4):841-890

Braxton P, Coleman D (2007) Hilbert's problem for cost estimating. Joint ISPA/SCEA International Conference, New Orleans, LA

Hamaker J (1995) Parametric estimating. In: Stewart RD., Wyskida RM, Johannes JD (eds), Cost estimator's reference manual. John Wiley & Sons, New York, NY

Hastie T, Tibshirani R, Friedman, J (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York, NY

Jeffery R, Ruhe M, Wieczorek I (2000) A comparative study of two software development cost modeling techniques using multi-organizational and company-specific data. Information and Software Technology 42(14):1009-1016

Kelly C, Rice, J (1990) Monotone smoothing with application to dose-response curves and the assessment of synergism. Biometrics 46(4):1071-1085

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

Kolodner JL (1993) Case-based reasoning. Morgan Kaufmann Publishers, San Mateo, CA

Lee KH, Lee K-Y (2002) Agent-based collaborative design system and conflict resolution based on case-based reasoning approach. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 16(2):93-102

Maher ML, Zhang DM (1993) CADSYN: a case-based design process model. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 7(2):97-110

McCarthy P (1996) Market price and income elasticities of new vehicle demands. The Review of Economics and Statistics 78(3):543-547

Mendes E, Watson I, Triggs C, Mosley N, Counsell S (2003) A comparative study of cost estimation models for web hypermedia applications. Empirical Software Engineering 8(2):163-196

Michalek JJ, Papalambros PY, Skerlos SJ (2004) A study of fuel efficiency and emission policy impact on optimal vehicle design decisions. Journal of Mechanical Design 128(6):1196-1204

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models. Irwin, Chicago, IL

Otto KN, Wood KL (2001) Product design: techniques in reverse engineering and new product development. Prentice-hall Inc., Upper Saddle River

Pahl G, Beitz W (1996) Engineering design-a systematic approach. Springer Limited, London

Reich Y, Kapeliuk A (2004), Case-based reasoning with subjective influence knowledge. Applied Artificial Intelligence 18(8):735-760

Roderman S, Tsatsoulis C (1993) Panda: a case-based system to aid novice designers. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 7(2):125-133

Rosenman M (2000) Case-based evolutionary design. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 14(1):17-29

Shepperd M, Schofield C (1997) Estimating software project effort using analogies. IEEE Transactions on Software Engineering 23(12):736-743

Shiau C-SN, Michalek JJ (2009) Should designers worry about market systems? Journal of Mechanical Design, 131(1):011011 (9 pages)

Shiva Kumar H, Krishnamoorthy C S (1995) A framework for case-based reasoning in engineering design. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 9(3):161-182

Takai S (2009) A case-based reasoning approach toward developing a belief about the cost of concept. Research in Engineering Design 20:255-264

Train KE, Winston C (2007) Vehicle choice behavior and the declining market share of us automakers. International Economic Review 48(4):1469-1496

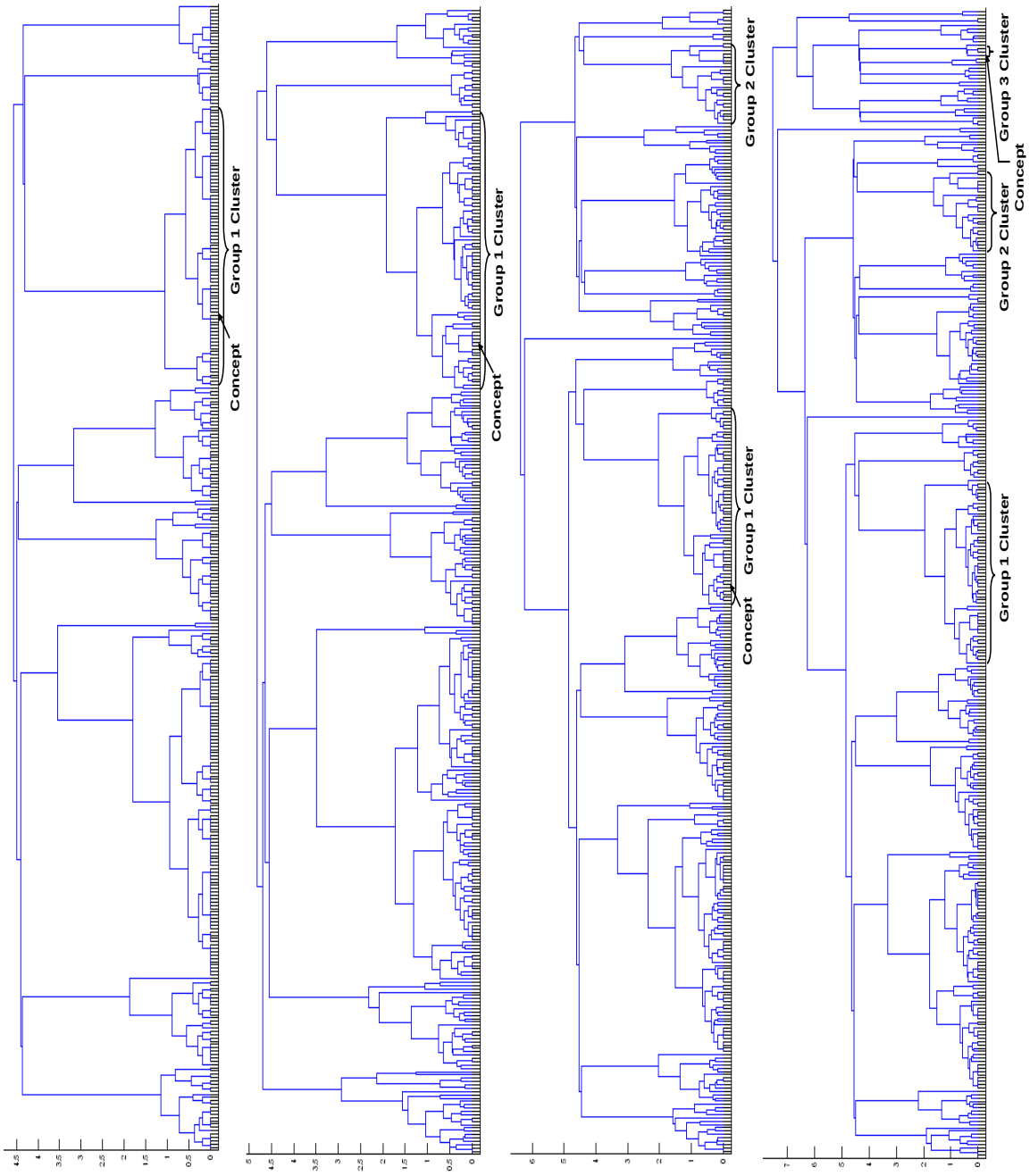
Ulrich KT, Eppinger SD (2004) Product design and development. McGraw-Hill

Williams N, Azarm S, Kannan PK (2008) Engineering product design optimization for retail channel acceptance. Journal of Mechanical Design 130(6):061402 (10 pages)

Wood WH, Agogino AM (1996) Case-based conceptual design information server for concurrent engineering. Computer Aided Design 28(5):361-369

Wyskida RM (1995) Statistical techniques in cost estimation. In: Stewart RD, Wyskida RM, Johannes JD (eds) Cost estimator's reference manual. John Wiley & Sons, New York, NY

APPENDIX



(a) Concept 0 (b) Concept 1 (c) Concept 2 (d) Concept 3

Fig. A Dendrograms for Automobile Retrieval

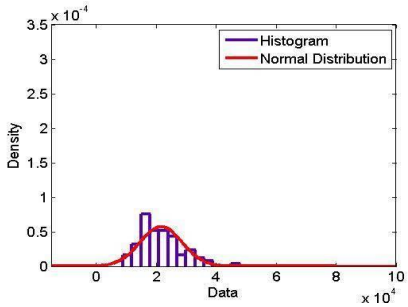
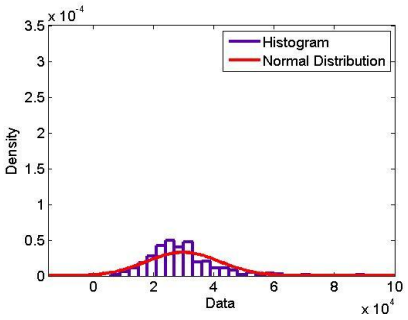
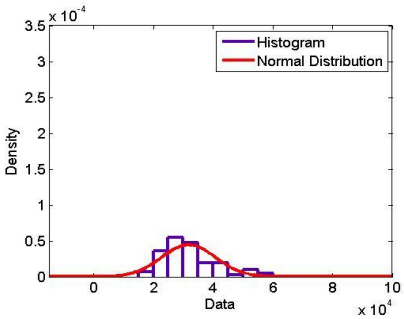
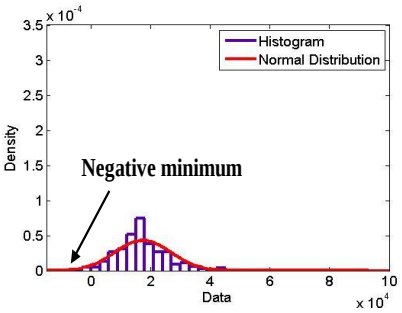
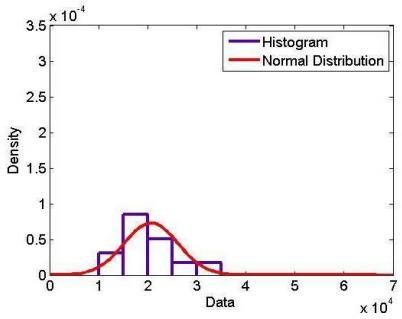
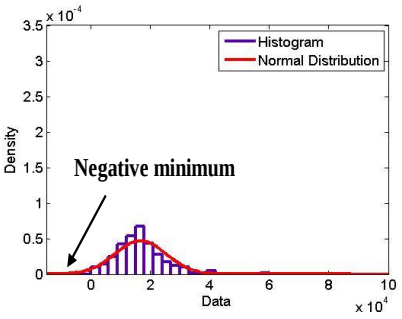
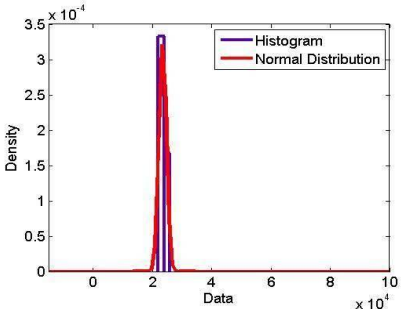
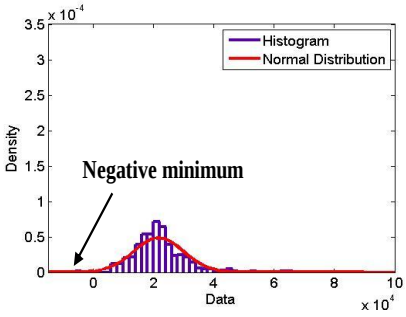
Design Attribute Selection	CBR-A	RA
No Design Attributes		
One Design Attribute		
Two Design Attributes		
Three Design Attributes		

Fig. B Cost Distributions

VITA

Karan Banga was born on July 6, 1985, in New Delhi, India. He received his primary and secondary education in New Delhi, India. He received his Bachelor of Science degree in Mechanical and Automation Engineering in 2007, from the Guru Gobind Singh Indraprastha University, Delhi, India. He has been enrolled in Mechanical Engineering at the Missouri University of Science and Technology, Rolla, since fall 2008. He held a Graduate Research Assistantship under Dr. S. Takai and Graduate Teaching Assistantship under the Department of Mechanical Engineering at the Missouri University of Science and Technology, Rolla. He received his Master's degree in December 2010.