Cost Modeling Based on Support Vector Regression for Complex Products During the Early Design Phases

Guorong Huang

Dissertation Submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
in
Industrial and Systems Engineering

Michael P. Deisenroth, Chair
F. Frank Chen
Kimberly P. Ellis
Roberta S. Russell

August 9, 2007
Blacksburg, Virginia

Keywords: Cost Estimation, Support Vector Regression, Cost Driver Identification, Cost Driver Selection, Semiparametric Approach, Sensitivity Analysis, Tabu Search

Cost Modeling Based on Support Vector Regression for Complex Products During the Early Design Phases

Guorong Huang

(ABSTRACT)

The purpose of a cost model is to provide designers and decision-makers with accurate cost information to assess and compare multiple alternatives for obtaining the optimal solution and controlling cost. The cost models developed in the design phases are the most important and the most difficult to develop. Therefore it is necessary to identify appropriate cost drivers and employ appropriate modeling techniques to accurately estimate cost for directing designers. The objective of this study is to provide higher predictive accuracy of cost estimation for directing designer in the early design phases of complex products.

After a generic cost estimation model is presented and the existing methods for identification of cost drivers and different cost modeling techniques are reviewed, the dissertation first proposes new methodologies to identify and select the cost drivers: Causal-Associated (CA) method and Tabu-Stepwise selection approach. The CA method increases understanding and explanation of the cost analysis and helps avoid missing some cost drivers. The Tabu-Stepwise selection approach is used to select significant cost drivers and eliminate irrelevant cost drivers under nonlinear situation. A case study is created to illustrate their procedure and benefits. The test data show they can improve predictive capacity.

Second, this dissertation introduces Tabu-SVR, a nonparametric approach based on support vector regression (SVR) for cost estimation for complex products in the early design phases. Tabu-SVR determines the parameters of SVR via a tabu search algorithm improved by the author. For verification and validation of performance on Tabu-SVR, the five common basic cost characteristics are summarized: accumulation, linear function,

power function, step function, and exponential function. Based on these five characteristics and the Flight Optimization Systems (FLOPS) cost module (engine part), seven test data sets are generated to test Tabu-SVR and are used to compare it with other traditional methods (parametric modeling, neural networking and case-based reasoning). The results show Tabu-SVR significantly improves the performance compared to SVR based on empirical study. The radial basis function (RBF) kernel, which is much more robust, often has better performance over linear and polynomial kernel functions. Compared with other traditional cost estimating approaches, Tabu-SVR with RBF kernel function has strong predicable capability and is able to capture nonlinearities and discontinuities along with interactions among cost drivers.

The third part of this dissertation focuses on semiparametric cost estimating approaches. Extensive studies are conducted on three semiparametric algorithms based on SVR. Three data sets are produced by combining the aforementioned five common basic cost characteristics. The experiments show Semiparametric Algorithm 1 is the best approach under most situations. It has better cost estimating accuracy over the pure nonparametric approach and the pure parametric approach. The model complexity influences the estimating accuracy for Semiparametric Algorithm 2 and Algorithm 3. If the inexact function forms are used as the parametric component of semiparametric algorithm, they often do not bring any improvement of cost estimating accuracy over the pure nonparametric approach and even worsen the performance.

The last part of this dissertation introduces two existing methods for sensitivity analysis to improve the explanation capability of the cost estimating approach based on SVR. These methods are able to show the contribution of cost drivers, to determine the effect of cost drivers, to establish the profiles of cost drivers, and to conduct monotonic analysis. They finally can help designers make trade-off study and answer "what-if" questions.

Keywords: Cost Estimation, Support Vector Regression, Cost Driver Identification, Cost Driver Selection, Semiparametric Approach, Sensitivity Analysis, Tabu Search

Acknowledgements

I would like to express my deepest appreciation to my advisor Dr. Michael P. Deisenroth for his help, support, invaluable guidance and advice through these years. Without his influence, my research would never have been gone thus far. The inspiring discussions and close interactions with him always reinvigorated my hope once again and made me learn how to steadfastly move to the next step. He always presented me challenging problems and gave me the freedom to tackle them in my own ways. His extraordinary ability to guide student in the right direction has always benefited me. Whenever I had trouble with my research, he were always there helping me.

I also want to give my sincere thanks to my committee members, Dr. F. Frank Chen, Dr. Kimberly P. Ellis, and Dr. Roberta S. Russell. I am grateful for their kindness, encouragement along with their insightful comments, constructive suggestions to my research.

Finally, I would like to express a full heart of thankfulness to my wife Yang Xu. Her support, encouragement and love have enabled me to complete this dissertation. I would like to dedicate this dissertation to my parents for their unconditional love and support for my past, present and future journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Background

The primary focus of this research is on accurately estimating cost to assist designers and decision makers to control cost during the early design phases for complex products. The cost model developed during the design phases is used in critical design decisions which shape the cost of the entire project. Ten to fifteen percent [1] of the total cost spent during the design phase commits eighty percent of the total cost in the life cycle. The cost models developed in the design phases are the most important and the most difficult to develop. Even though a rather small proportion of the total life cycle cost is spent during the design phases of a product, it is important that there are effective measures to control cost during the design phases to minimize total life cycle cost. Experience has shown that the greatest potential for cost reduction is in the early design phases. Cost estimation is necessary to provide opportunities to allow designers, planners and decision-makers to consider better alternatives for cost control.

However, during the early design phases available information is inadequate because the product is not fully defined. This is especially true for complex products; they have complex designs and complex manufacturing processes and thus cost estimation is not an easy task. The relationships between costs and cost drivers are complex. They often include nonlinear properties and may be very hard to define in some cases. Therefore, it is necessary, especially for complex products, to employ appropriate models and techniques to accurately estimate the cost for directing the designers.

1.2 Motivation

During the design phases, designers and decision-makers often need to know accurate cost information to assess and compare multiple alternatives to obtain the optimal solution. They need to identify cost reduction opportunities and tradeoffs to meet targets (requirement, performance, and schedule). They also need to evaluate cost reduction ideas and alternatives affecting system performance factors for their impact and compare

the results with the original “baseline” design. A cost estimating model must be reasonably, accurate, robust, fast and capable of operating on data of the detail typically available in the related phase, to support conduct the cost trade-off studies for designers and decision makers.

Accuracy is crucial to a cost model for a complex product. It is the most basic need for estimating cost. In the early design phases, inaccuracy results in overestimates or underestimates of cost. An overestimate may cause a designer to abandon an appropriate design in favor of a cheaper design with worse performance. When costs are underestimated in the early design phases, initial plans for materials, scheduling, tooling, processing etc., are not attainable. There would be a redesign, replanning, reproduction and possibly the addition of personnel and equipment at the later phases. These eventually increase costs more than originally budgeted. The accuracy of cost estimates is therefore essential for designers to control cost. The most realistic estimate with greatest accuracy would incur the most economical cost of a complex product.

Accuracy improves with more available information. Generally, with the evolving of design, more and more information can be provided to make estimation more accurate. Creese and Moore [2] provided a discussion on the degree of accuracy in the different stages as shown in Figure 1-1. During the conceptual design phase, available information is inadequate and cost estimation must rely primarily on the use of known essential product functions and hence accuracy ranges from -30% to +50% of the real cost. As design progresses, more design information becomes available and cost estimates can be made based on available historical cost data. The accuracy of cost estimates in this stage ranges from -15 to +30%. During the detail design phase, when most information about the product is known, the degree of cost estimating accuracy should be within - 5 to +15%.

Generally, with more information, better accuracy is obtained. In reality, the development of a cost estimating model is a process of identifying and employing information (cost drivers, historical data and apriori knowledge) to estimate costs. Early and even some current cost models, miss some important cost drivers which reflect key engineering design parameters and key management decisions [3], even when the historical data are available. This results in a bias for cost estimation and lack of

credibility for designers and decision-makers. Hence, identifying and selecting cost drivers is an important task for cost estimation.

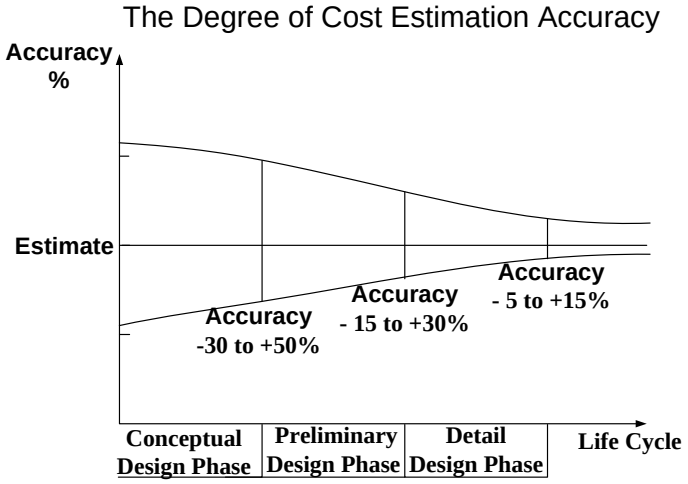


Figure 1-1 The Degree of Cost Estimation Accuracy

Besides identification of cost drivers, the cost modeling techniques that employ the cost drivers greatly affect the degree of estimating accuracy. Different cost estimating approaches will produce different degrees of accuracy even under the same situation. This is especially true for complex products. The relationship between cost and cost drivers often includes nonlinear and discontinuous properties, which are difficult to define. Even when the same cost drivers and the same set of historical data are used, the appropriate choice of cost modeling approaches can significantly improve the estimating accuracy. Finding a feasible and appropriate approach to improve the estimating accuracy is another main task of cost estimation.

In summary, the motivations of this study are addressed as:

- The cost estimating model would be continuously and concurrently applicable for design use, which could help designers achieve good trade-off decisions. It can provide designers with a “what if” capability to easily test the impact of design alternatives on complex product costs.
- The appropriate cost model provides designers with accurate cost. Accuracy is the basic need of a cost model. To improve the estimating accuracy for a complex product, it is necessary to:

- Fully employ all available significant cost drivers as possible; and
- Choose the feasible and appropriate cost estimating approach.

1.3 Objectives

The main research issue is how to provide higher predictive accuracy of cost estimation for directing designers at the early design phases of complex products.

It will address these questions:

1. How are available significant cost drivers identified and selected?
2. Is an approach based on support vector regression (SVR) applicable in cost estimation? Is it better than other traditional cost estimating approaches such as parametric method, neural networks, and case-based reasoning?
3. Can the nonparametric approach based on support vector regression be combined with parametric approaches to improve cost estimates?
4. Can a cost estimating model based on support vector regression be used to provide a guide for designer to realize the impact of design decisions on costs?

Given these research questions, the primary objective of this research is to provide new methodologies to obtain higher predictive accuracy of cost estimation and guide designers at the early design phases of complex products. It can be broken down into sub-objectives that include:

- Presenting and summarizing current cost estimating approaches and model techniques;
- Proposing two new methodologies, Causal-Associated (CA) approach and Tabu-Stepwise selection approach, to identify cost drivers and then select the most significant cost drivers by eliminating the irrelevant and redundant cost drivers using Tabu-Stepwise selection approach;
- Exploring a new technique Tabu-SVR, a nonparametric approach based on support vector regression (SVR) combining with a tabu search algorithm, which is applied in cost estimating model for the higher predictive accuracy during the early design phases of complex products; this objective involves:

  - Choosing an appropriate kernel function and tuning parameters for support vector regression to accurately estimate cost,
  - Summarizing the common basic cost characteristics to produce test cases for experiments,
  - Conducting experiments to compare different kinds of approaches under the different scenarios: parametric estimating approach; neural network approach; case-based reasoning approach and the approaches based on support vector regression,
- Investigating three semiparametric algorithms based on SVR for cost estimation and comparing them with pure parametric method and pure nonparametric approach based on SVR;
- Introducing two existing methods to conduct a sensitivity analysis based on the nonparametric approach based on SVR for directing designers.

1.4 Scope

This research will focus on cost estimation for complex products at the early design phases. The methodology proposed in this study is applicable to most products throughout the phases of the entire life cycle. But for complex products at the early design phases, the methodology is especially useful when the cost structure is not clear or when relationships between costs and cost drivers are unknown and include nonlinear properties. This cost modeling methodology will result in higher accuracy of cost estimation and help designers conduct cost tradeoff studies at the time of concurrent design.

Most examples and existing cost models in this study come from the area of aerospace cost estimation. The methodology for cost estimation in the area of aerospace can undoubtedly be used in broader domains such as automobile, semiconductor, etc.

1.5 Organization

This dissertation is organized into eight chapters. Chapter 1 introduces motivations, objectives and scope of this study. Chapter 2 first presents a generic cost estimation model. Based on this generic framework and classification of cost modeling techniques, it reviews the existing methods about identification of cost drivers and different kinds of cost modeling techniques. Then Chapter 2 gives the reviews of tabu search and support vector regression to support the next chapters. In Chapter 3, the framework of this study is first presented and the research methodology for the whole study is briefly introduced. Chapter 4 proposes the Causal-Associated method for identifying cost drivers and the Tabu-Stepwise selection technique for selecting cost drivers. A case study of an AC motor is presented to show the feasibility and benefits of proposed methods. In Chapter 5, Tabu-SVR, the nonparametric cost estimating approach based on SVR is then given. Based on summarized basic common cost characteristics, and the Flight Optimization Systems (FLOPS) cost module (engine part) [4-6], the test cases (data sets) are produced. Choosing the appropriate kernel and tuning corresponding parameters via a tabu search algorithm are studied. Comparison between SVR and other three conventional approaches is presented. Chapter 6 introduces three semiparametric algorithms based on SVR under different type and amount of known information for cost estimation. After that, two methods of sensitivity analysis based on SVR are introduced in Chapter 7. Finally the dissertation is concluded in Chapter 8. The direction of future research is also presented in this chapter.

Chapter 2 Literature Review

2.1 Generic Cost Estimation Model

A generic cost estimation model can be written as (2-1):

$$C = f(x;\beta) \tag{2-1}$$

where C is the desired cost.

The first variable, x, is a vector of real numbers with dimension d, $x \in R^d$. It can be written as $x = \{x_1, x_2, x_3, \cdots, x_d\}$. Each x_i $(1 \le i \le d)$ is a cost driver which is assumed to be related to and influence the cost (C). The set of cost drivers should include all of the inputs that significantly impact the cost (C).

The second variable β is a vector of parameters. For different models which use different estimating strategies and/or approaches, β has different meanings. It often depends on the form of $f(\bullet;\bullet)$ and the nature of input space. For example, in a linear model, the parameters reflect the model structure and the nature of historical data. The parameters are more fundamental to the model than the input variable set.

The function, $f(\bullet;\bullet)$, expresses the relationship between x, β and C. It includes two aspects: a structural relationship and a functional relationship.

In the cost estimating area, the structural relationship is often called the cost breakdown structure (CBS). It can help a designer understand the detail cost information associated with the product. Additionally, it can simplify the problem. Generally, when and how $f(\bullet;\bullet)$ is decomposed relies on knowledge about the product. For example, $f(\bullet;\bullet)$ can be broken down according to product structure and/or the phases of the product life cycle.

In Figure 2-1, cost (C) is broken down into three components: C_1, C_2, and C_3. Their relationships can also be expressed as Equations (2-2). For instance, the acquisition cost of engines (C) in an aircraft is composed of three components: 1) the cost of research, development, test and evaluation (RDT&E) (C_1); 2) the cost of production (C_2); and 3) the other cost (C_3) such as administration, collaboration, etc.

C_1 is composed of C_{11}, C_{12} and C_{13}. In the first component of engines cost, C_{11} is the cost of research, C_{12} is the cost of development, and C_{13} is the cost of test and evaluation.

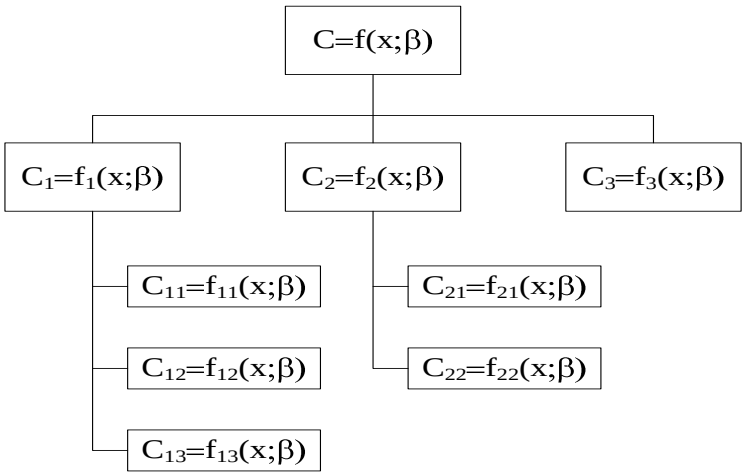


Figure 2-1 An Example of Cost Breakdown Structure (CBS)

C_2 is made of C_{21} and C_{22}. For the engines in an aircraft, the cost of production can be the sum of direct production cost (C_{21}) and indirect production cost (C_{22}).

$$\begin{aligned} C &= C_1 + C_2 + C_3 = (C_{11} + C_{12} + C_{13}) + (C_{21} + C_{22}) + C_3 \\ C &= f(x;\beta) \\ &= f_1(x;\beta) + f_2(x;\beta) + f_3(x;\beta) \\ &= (f_{11}(x;\beta) + f_{12}(x;\beta) + f_{13}(x;\beta)) + (f_{21}(x;\beta) + f_{22}(x;\beta)) + f_3(x;\beta) \end{aligned} \quad (2\text{-}2)$$

The input space or parameter space of subcomponents is a subset of the input space or parameter space of the product. In the above example, the input space of C_{13} is a subset of input space of C.

A functional relationship $f(\bullet;\bullet)$ is a mapping relation between the input space and the output (C) (see Figure 2-2). It is generally derived from historical data, experience, and physical properties via statistical techniques, empirical studies, laws of physics, etc. It can be expressed by a cost estimating approach.

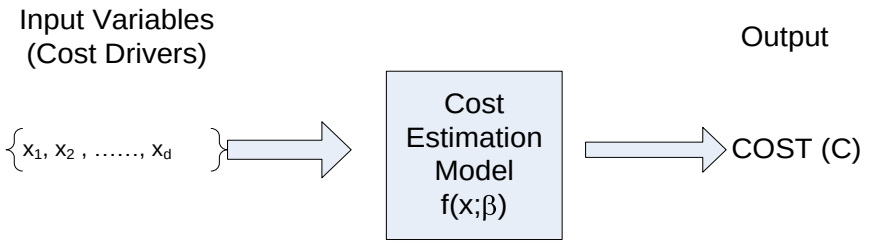


Figure 2-2 Generic Cost Function Relation *f(x; β)*

There are many different cost estimating techniques identified in literatures. They can be organized into two broad classifications:

- Based on functional relationship:
    1) Expert judgment (see Section 2.3.1);
    2) Parametric modeling (see Section 2.3.2);
    3) Neural network modeling (see Section 2.3.3);
    4) Case-based reasoning (see Section 2.3.4);
- Based on type of inputs and structural relationship:
    1) Feature-based estimating (see Section 2.4.2);
    2) Activity-based costing (see Section 2.4.3);
    3) Process-based costing (see Section 2.4.4);
    4) Simulation (see Section 2.4.5).

The functional relationships form the building blocks of a cost model, such as $f_{12}(x; \beta)$, or $f_{22}(x; \beta)$ in Figure 2-1. If there is not enough knowledge to form a cost breakdown structure, the high level functional relationship $f(x; \beta)$ would be investigated as a basic element. Functional relationships can be established by a few different cost modeling techniques: expert judgment, parametric method, neural network approach, and case-based reasoning approach. These cost estimating techniques can be applied during the entire life cycle of a product (see Figure 2-3).

With increased information and knowledge about a product, there are additional cost estimating techniques that can be utilized to take advantage of the additional information and knowledge. The cost estimating techniques, which are based on the type of inputs and the structural relationship, are often applied in the later phases of the product life

cycle (see Figure 2-3). These techniques include feature-based estimating, activity-based costing, process-based costing, and simulation.

In this research, after the Causal-Associated method and Tabu-Stepwise algorithm are presented to identify and select cost drivers, new cost estimating approaches based on SVR, including nonparametric and semiparametric, are proposed and studied. The nonparametric approach will be compared with other formalized conventional cost estimating models (parametric modeling, neural network modeling, and case-based reasoning). Finally, there are two methods presented to determine the impact of cost drivers on output (C) and provide designers with guidance for sensitivity analysis.

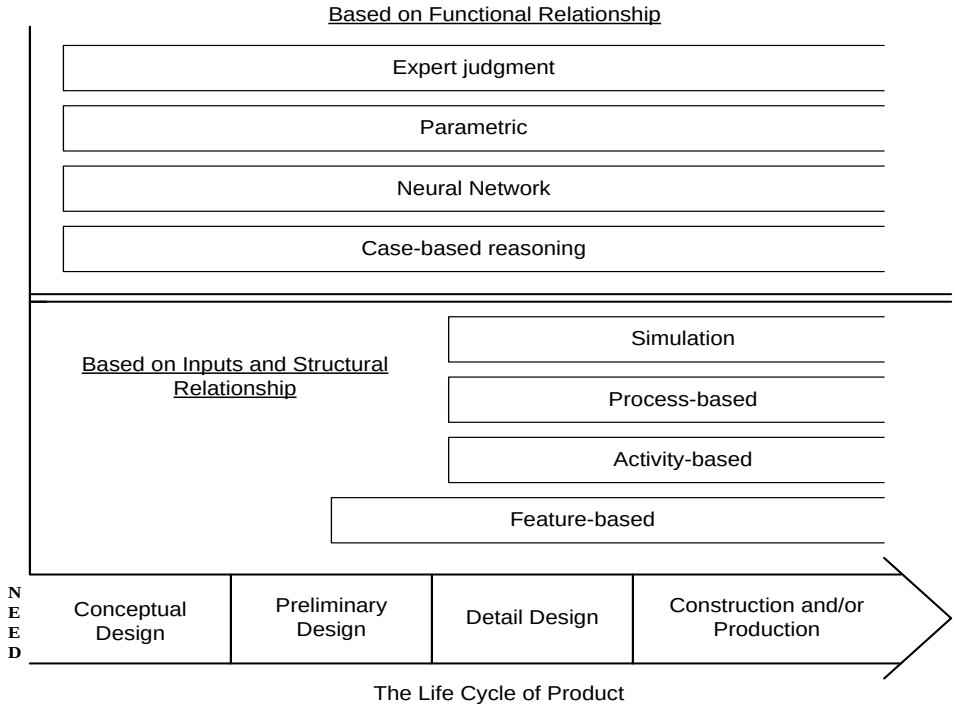


Figure 2-3 The Cost Estimating Techniques and the Life Cycle of a Product

2.2 Identification and Selection of Input Variables (Cost Drivers)

2.2.1 Input Variables (Cost Drivers)

As indicated earlier, the input variables x_1, x_2, ... , x_d are called cost drivers and are assumed to significantly impact and/or relate to the final cost (C). Analyzing their effects on the final cost can help determine which design properties deserve the most attention

during design phases of a product. Cost drivers can be associated with or include such things as: recourses, design attributes, product features, product structures information, performance, reliability, maintainability, production processes, production plans, management information, general operations such as activities performed in the life cycle, etc. As a design evolves, more detail is known about the product and more detailed cost models can be constructed which can introduce more new cost drivers.

Harwick [7] indicated that most cost drivers in space transportation economics can be grouped in different categories: management (design team, production team), technical (size, stages/structure, motors/rocket propulsion, electronics), cultural (parallel organizations, certification requirements, schedule), and market (new design, prototypes, production quantity, design repeat, year of technology, mechanic and electronic).

Complexity, an important index in cost modeling, is a typical type of cost drivers. Bashir and Thomson [8] proposed five complexity measure criteria: intuition, sensitivity, consistency, generality, and simplicity. Harwick [7] categorized the economic properties in different areas: diseconomies and economies of scale (size), impact of schedule (rate of development or rate of production), production learning curve factor, impact of economic externalities at the system level. These complexity measure and economic properties are widely used in most cost models [9-13].

2.2.2 Identification and Selection of Cost Drivers

The identification and selection of cost drivers is very important for the performance of a cost model. Appropriate and complete cost drivers are a prerequisite for accurate cost estimates.

Early cost models only consider part of the available information (cost drivers) such as product performance and technical indicators. The result is that their degree of estimating accuracy is not high. In 1998, Marx, Mavris, et al. [14] stated that existing aircraft cost models were based only on product design variables. For improving accuracy and model fidelity, the cost model must represent the use of advanced materials and processes. The aircraft designers can thus determine life-cycle cost implications of production change. Prince [3] presented two reasons that NASA’s management did not

believe the cost estimate: the cost models “do not reflect key engineering design parameters; and they do not reflect key management decisions”. Therefore, identification and selection of cost drivers is crucial to cost estimating.

Most current models [15-17] first identify candidate cost drivers via many sources: personal and/or other experience, and published information. The final set of cost drivers is selected by expert survey and/or statistical analysis (see Figure 2-4).

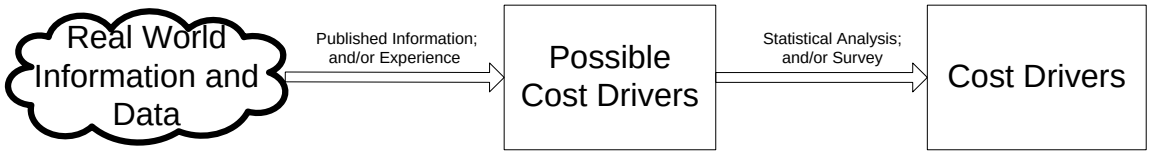


Figure 2-4 The Existing Methods to Identify and Select Cost Drivers

The parametric estimating handbook [16] published by the Department of Defense presents a typical process for identifying possible cost drivers. First, using brainstorming techniques, several alternatives for potential cost drivers are identified. Several experts are then surveyed to obtain their feedback on the merits of each potential cost driver. Finally the best cost driver candidates are selected for further analysis.

Seo, Park, et al. [15, 18] adopted another typical process to identify the cost drivers. First they formed a set of candidate product attributes based upon the literatures and the experience of experts. Second they grouped and reviewed these candidate cost drivers. Third, they refined these candidate drivers using first-order relationships based on the data. After bivariate correlations were computed, those cost drivers were selected if correlation tests to 95% statistical significance.

2.3 Cost Estimating Approaches Based on Functional Relationship

The following sections describe cost estimating approaches based on functional relationship. These approaches include expert judgment, parametric modeling, neural network modeling, and case-based reasoning.

2.3.1 Expert Judgment

Although designers and decision-makers feel more comfortable with the use of algorithmic and formalized methods, expert judgment for cost estimate is one of most widely used approaches in the whole life cycle of a product. Hughes [19] indicated that Heemstra and Kusters [20] stated that over half of the production estimates are based on intuition while approximate 16% of the estimates employ a formalized estimation methodology.

Expert judgment is based on the experience and knowledge of experts in the field. By nature, a cost estimate is a cost prediction of experts. Based on their experience and understanding of the complex product, experts achieve a cost estimate of the product under design. The apparent weakness of this method is that an estimate cannot be better than the experts' opinion. Additionally, the apparently subjective and unstructured nature of expert opinion makes it appear particularly vulnerable.

The advantages and disadvantages of this technique are summarized [19, 21] as following:

- Advantages:
  - Quick to produce, and easily incorporate knowledge of past experiences;
  - Requires little resource in terms of time and cost.
- Disadvantages:
  - No better than the experts;
  - Subjective: different experts with the same starting information will provide different cost estimates, use of expert judgment is not consistent and an unstructured process;
  - Prone to bias: personal experience, political aims, resources, time pressure, memory recall;
  - Estimate reuse and modification is difficult;
  - Difficult to quantify and validate the estimates;
  - Estimate depends on level of experience.

For expert judgment, perhaps the most formal and rigorous method for capturing expert opinion is the Delphi technique [22]. This method has been used as an effective way of achieving consensus. It can lower individual biases and improve the estimate.

This method will force the experts involved in the cost estimation to voice their opinions anonymously through an intermediary. After analyzing the opinion gathering in each round and then feeding back to participants in subsequence rounds, the estimating cost will converse on a consensus to be good estimator of the true cost.

2.3.2 Parametric Modeling

Parametric estimating [16] is an approach that employs historical cost data and statistical techniques to establish cost estimating relationships (CERs) between cost and cost drivers during design, production, operation and support, and retirement phase for predicting future cost. CERs are mathematical expressions or formulas that are used to estimate the cost of an item or activity as a function of one or more relevant cost drivers.

Dean [23] gave a commonly used CER as (2-3):

$$c = e^{\beta_0} \prod_{i=1}^{r} e^{\beta_i x_i} \prod_{j=r+1}^{s} x_j^{\beta_j} \tag{2-3}$$

where x_i and x_j are cost drivers, β_0, β_i, and β_j are parameters, $r <= s$, and $s <= d$, d is the dimension of input space.

The linear form of Equation (2-3) is as (2-4):

$$\ln c = \beta_0 + \sum_{i=1}^{r} \beta_i x_i + \sum_{j=r+1}^{s} \beta_j \ln x_j \tag{2-4}$$

In this linear form (2-4) the coefficients β_0, β_i, and β_j can be obtained via least squares regression. There are four steps to establish a CER [16, 23]:

1) Selection of cost drivers (see Section 2.2);
2) Appropriate structure of the formula;
3) Computation of parameter β by the statistical technique, generally a multiple linear regression;
4) Examination and validation of the estimates of the CERs.

Currently, most formalized cost models are parametric models, which are mainly composed of a set of CERs and a structural relationship. Total system cost model are more complex than individual CERs because they incorporate the functional relationship and structural relationship which consist of “many equations, ground rules, assumptions,

logic, and variables that describe and define the particular situation being studied and estimated" [16].

In the area of aerospace and aircraft, most cost models are established based on parametric methods, such as the cost module of Flight Optimization Systems (FLOPS) [4-6], DAPCA-III[24], TRANSCOST Model [13], PRICE H [9, 11, 25, 26], SEER-H [10], NASA/Air Force Cost Model (NAFCOM) [12], Unmanned Space Vehicle Cost Model (USCM), and Small Satellite Cost Model (SSCM) [27, 28]. For example, the cost module of FLOPS adapted by Johnson [4, 5] is a typical life cycle cost model that is a parametric model. McCullers and NASA Langley Research Center [6] has continued to improve the module. The inputs to this module include four types of data:

1) Cost calculation data (variables related to calculating cost that do not change as the optimization proceeds);
2) Mission performance data (design mach number; maximum dynamic pressure; cruise velocity; block fuel (fraction of aircraft fuel capacity; block time));
3) Cost technology parameters (cost technology parameters to account for the cost associated with advanced technologies);
4) Configuration and data from other modules (weight; number of engines per aircraft; maximum thrust per engine; number of seats; total number of crew; maximum total fuel capacity; cargo weight; wing area).

The outputs of the cost module are: airframe RDT&E cost; airframe production cost; engine RDT&E cost; engine production cost; manufacturing cost; manufacturing cost with spares; manufacturers profit; total acquisition cost (price); direct operating cost; indirect operating cost; total life cycle cost.

Depending on pertinent CERs and inputs, costs are computed. For instance, when computing airframe production cost, the CERs of three components (wing, tail and body) are as follows:

- *WING:* $C_{WING} = 1730*WTS(1)^{0.766}Q218*FMWING*FMCOMP$
- *TAIL:* $C_{TAIL} = 1820*(WTS(2) + WTS(3))^{0.766}Q218*FMTAIL*FMCOMP$
- *BODY:* $C_{BODY} = 2060*WTS(4)^{0.766}\ Q218*FMBODY*FMCOMP$

where *WTS(1)*, *WTS(2)*, *WTS(3)*, and *WTS(4)* are the weights of the related components. *FMWING*, *FMCOMP*, *FMBODY*, and *FMTAIL* are cost technology parameters

representing complexity (their default value are 1). *Q218* reflects the price index.

2.3.3 Analogy Models (Cased-Based Reasoning (CBR))

The analogical modeling method (case-based reasoning method) seeks an estimate solution based on the past cases. By analogy it identifies a similar product or component and adjusts its costs for differences between it and the target system or entity (see Figure 2-5).

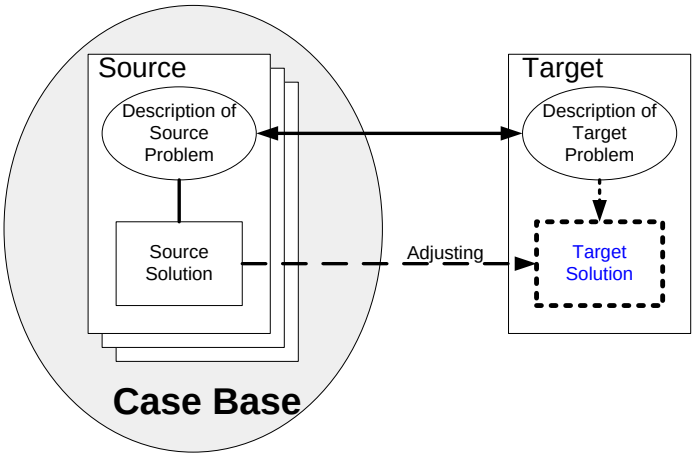


Figure 2-5 Case-Based Reasoning (CBR) [29]

The analogous method tends to be good for new products or innovative designs. It attempts to evaluate the cost of a product from similar products. The method is similar to the cognitive reasoning process of an individual: "the recognition of the problem, the recall of similar experiences and their solutions, the choice and the adaptation of one of the solutions (source case) to the new problem (target case), the evaluation of the new situation and the learning of the solved problem" [29]. Based on the creation of a link between the source problem specifications and the target problem specifications, the analogical method transposes the solution of the source to the target and adapts the known new solution to the target problem (See Figure 2-5).

The method can be thought of as a four step process. The first step is the description of problem (indexation). It includes structural description and contents description. This

step would focus on contents in a case, which are cost drivers in CBR. The second step implements similarity measures and retrieves similar cases from case base. Then through adaptation procedures, the solution is obtained in the third step. Finally, the fourth step determines whether the solved case needs to be put in the case base. Details of the process are given below.

2.3.3.1 Description of Problem (Indexation) with Cost Drivers of CBR

For the CBR method, a problem is solved by retrieving a previous case suitable for solving the new problem. Hence the structure and content of CBR's collection of cases is very important. It is necessary that the case searching and matching processes must be both effective and reasonably time efficient. Furthermore the new case is needed to integrate into the case base. Therefore, the description of problem in CBR including: 1) contents in a case, 2) an appropriate structure for describing case contents, and 3) the organization and index of cases for effective retrieval and reuse.

From the analysis of the product design process and its decision making processes, Wenstink, ten Brinke, et al. [30] presented four product characteristics that should be considered for the committed product costs (the costs that are fixed during the product design process): geometry (shape, dimensions, accuracy, etc); material (material costs may occupy about 50% of the total product costs); production process(es); and production plans.

2.3.3.2 Similarity Measures and Retrieved Process

Starting with a problem description (sometimes partial), the similarity measures and retrieved process end when a best matching case has been acquired. The target description features are matched against the description features of cases in the case base and a measure of similarity is computed. The retrieved cases would be ranked according to their similarity to the target, and a best matching case would be found. There are two major case retrieval approaches [31]:

- The distance-based or computational approach calculates the distance between cases. The most similar case is determined by the evaluation of a similarity measure.
- The indexing or representational approach searches the similar case by indexing structures which connect the cases. The similar case is coded into the structure of the case base itself.

Note that the above two approaches might be combined. The first approach is most widely used. Commonly, CBR systems use the inverse of weighted normalized Euclidian distance (Equation 2-5) as the similarity measure.

$$SIM(X, Y) = 1 - DIST(X, Y) = 1 - \sqrt{\sum_i w_i^2 dist^2(x_i, y_i)} \qquad (2\text{-}5)$$

where X is the attribute vector of a source case and Y is the attribute vector of the target case, w_i is normalized importance of i^{th} attribute. The normalized distance, $dist(x_i,y_i)$, is defined as $dist(x_i,y_i) = |x_i-y_i|/|max_i-min_i|$. Besides the above conventional approaches, Liao, Zhang, et al. [31] showed that combining the fuzzy set theory with the conventional CBR system greatly enhances measure's capability. Herrmann, Balasubramanian, et al. [32] proposed a special design similarity measure using the artificial neural networks (ANN) to help decision-makers.

2.3.3.3 Adaptation Procedures to Get the Solution

Once a matching case is retrieved, the adaptation process adjusts for prominent differences between the retrieved case and the target problem by applying formulas or rules to modify the solution of the retrieved case for those differences. Daengdej, Lukose, et al. [33] proposed a method of adapting the solution applying the statistical methods called closeness factor. As was the case used in the similarity measure, ANN is also used in the process of adaptation. To some extend, by adaptation, the CBR method can incorporate parametric modeling and ANN in determining similarity measures and adaptation procedures. Once the similar cases are selected, the final cost of the problem would depend on the available data of these cases. It is possible to use a parametric model on these similar case data or train ANN using them to compute the cost of the target.

2.3.3.4 Self-Learning

One of the advantages of CBR is its powerful learning capacity. When the CBR method solves a new problem, it can retain the solution of new problem in the case database. As more problems are solved, the CBR method can be applied in a larger variety of situations and the estimates increase in accuracy. In short, newly solved problems can be learned by storing their specifications and solutions together as new cases in the case base.

Rehman and Guenov [34] proposed a methodology to estimate the manufacturing cost at the design phases, which incorporates the use of case-based and rule-based reasoning. They employed case-based reasoning to retrieve a similar product model completely described, and applied rules to derive the process plan for cost estimate. In a series of papers written by ten Brinke, Lutters, Weustink, et al. [30, 35, 36], a generic framework for cost estimating was developed as the basis for the control of the production costs. The framework takes design, process planning and production planning aspects into account. The authors also proposed a variant-based (case-based) cost estimation method based on the product information structure related to the manufacturing engineering reference model. Based upon the CAD information exported from the CAD system in STEP format, El-Mehalawi [37] developed a cost estimation model for Net-Shape Manufacturing (NSM) using a case-based reasoning approach.

For software cost estimation, much research has been done on the use of CBR [38-41]. Shepperd and Schofield [39] presented an approach to estimating software project effort based upon the use of analogies (case-based reasoning). After characterizing projects in terms of these projects' features, such as the number of interfaces, the development method, and the size of the functional requirements document, the case base for the source projects was established. One of the most similar projects was then selected and adjusted to predict the project effort.

2.3.4 Neural Networks

Artificial neural networks (NNs) simulate biological neurons using computers [42] to model a system with an unknown input-output relation. Artificial NNs are trained through modifying the parameters to minimize a loss function via the stochastic gradient decent method. For example, a back-propagation neural network is a common neural network architecture, which is composed of an input layer, an output layer, and some hidden layers between the input and output layers. Each layer has a number of processing unit (neuron). A neuron simply computes the sum of their weighted inputs, subtracts its threshold from the sum, and passes the results through its transfer function. This can be expressed mathematically as Equation (2-6):

$$y_i = f_i(\sum_{j=1}^{n} w_{ij} x_j - s_i) \qquad (2\text{-}6)$$

where y_i represents the output of neuron, w_{ij} represents the weight associated with the input j, s_i represents the threshold value of the neuron, and f_i represents the transfer function.

Artificial Neural Networks (NNs) are purely data driven models. Funahashi [43] and Hornik, Stinchcombe, et al. [44] have proven that multilayer feedforward networks, with as few as one hidden layer, are indeed capable of universal approximation. However, they did not address the issue of how many neurons are needed to attain a given degree of approximation.

Using a back-propagation neural network Zhang and Fuh [45] proposed a feature-based prototype system to estimate the costs of packaging products only based on design information. However Zhang and Fuh indicated, determining the number of hidden layers and the number of neurons in each hidden layer is a trial and error process, which can be time consuming. Neural network training requires experience and relies on accurate historical cost data. Using artificial neural networks, Seo, Park, et al. [15, 18] proposed an approximate method --- learning life cycle cost (LCC) for providing preliminary life cycle cost. With regard to accuracy and stability, the model resulted in better prediction than the statistical regression model did.

Based on pilot cost data from a manufacturing company and artificially created simulative data, Bode [46, 47] compared cost estimation performance between

conventional methods, i.e. linear and nonlinear parametric regression and neural networks. He indicated that neural networks achieve lower deviations in their cost estimations. Bode concluded that neural networks can detect hidden relationships among training data and they seem most appropriate for cost estimation in the conceptual phase of routine design and configuration design tasks. However, they are inappropriate for cost estimation of radical innovations. Smith and Mason [48] examined the performance, stability and ease of cost modeling using regression versus neural networks. Their paper indicates that neural networks have advantages when dealing with data where there is little apriori knowledge of CER function form for regression. Furthermore, the artificial neural network is a "black box" CER and does not provide any explanation for users.

2.4 Cost Estimating Approaches Based on Inputs and Structural Relationship

In Section 2.3, the approaches based on functional relationship were classified into: expert judgment, parametric method, case-based reasoning approach, and neural network approach. They are building blocks for the approaches based on inputs and structural relationship which are discussed in this section. The performance of these cost estimating approaches presented later depends on the identification of inputs, structure and the approaches based on functional relationship. This section focuses on cost breakdown structure (CBS) and four approaches based on inputs and structural relationship: feature-based modeling, process-based approach, activity-based costing estimating and simulation.

2.4.1 Cost Breakdown Structure (CBS)

Breaking a complex problem into a set of subproblems is a common strategy for solving a complex and/or new problem. It can make the complex problem easy to understand and solve. With enough information, especially in the later design phases, employing a reasonable cost breakdown structure (CBS) can simplify the degree of problem and help designers and decision-makers trace the cost detail information. A CBS

partitions a complex product into smaller components to improve the accuracy of cost estimates and provides a mechanism for collecting and organizing actual costs, cost control for design and decision making [1].

The cost breakdown structure (CBS) constitutes a functional breakdown of costs. For the life cycle cost of a product, the entire life cycle could be considered and identified in a CBS. This includes research and development cost, production and construction cost, operation and system support cost, and retirement and disposal cost. In life cycle cost modeling, the CBS includes all costs related to customer, contractor, supplier, and consumer (user) activities over entire life cycle [1].

The Life Cycle Cost (LCC) module of the FLOPS [4-6] is a typical parametric model (see Section 2.3.2) with a functional breakdown structure. The module is composed of elements to calculate RDT&E (Research, Development, Testing and Evaluation) cost production cost, DOC (Direct operation cost), IOC (Indirect operating cost) (see Figure 2-6).

The cost breakdown structure of a computer model for estimating Development and Procurement Costs of Airframe (DAPCA-III)[24] is showed in Figure 2-7. It is composed of two components – the development and the production costs. The development cost is composed of total engineering for flight-test aircraft, total tooling for flight-test aircraft, nonrecurring manufacturing labor, recurring manufacturing labor for flight-test, quality control, nonrecurring manufacturing materials, recurring manufacturing materials for flight-test, and flight test. The production cost includes total engineering for production aircraft, total tooling for production aircraft, recurring manufacturing labor for production aircraft, recurring manufacturing materials for production aircraft, and quality control for production aircraft.

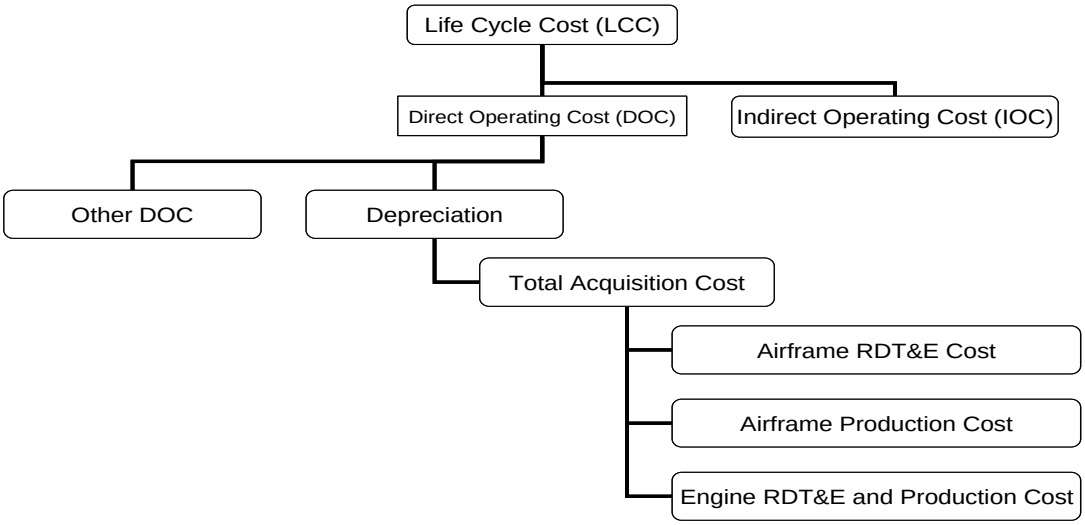


Figure 2-6 The Cost Breakdown Structure of the Cost Module in FLOPS

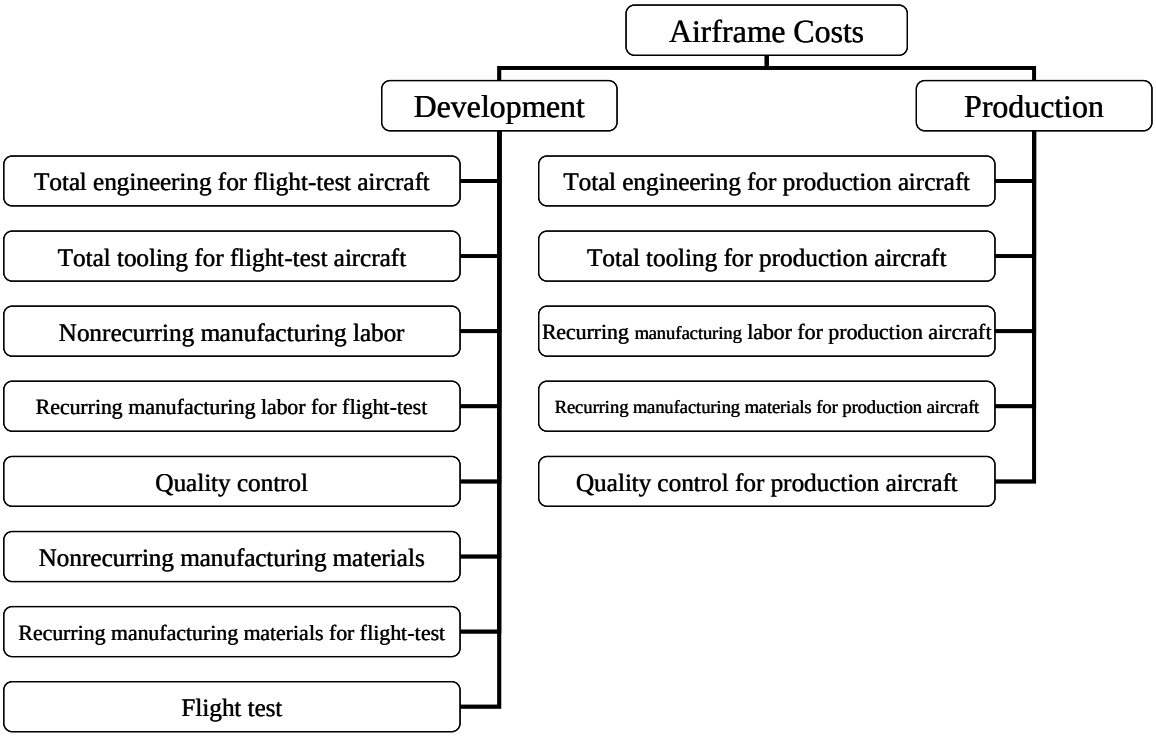


Figure 2-7 The Cost Breakdown Structure of DAPCA – III

2.4.2 Feature-Based Modeling

Shah [49] states that "features are elements used in generating, analyzing, or evaluating designs." From the manufacturing view, they represent shapes and technological attributes associated manufacturing operations and tools. Because there exists correlations between design features and cost, feature-based cost modeling uses these features as cost drivers. Feng, Kusiak, et al. [50] focused on the cost evaluation of machining form features. They indicated that the machining cost of a part depends not only on the type of form features, but also on the relationship among the features which has a significant impact on the machining cost imposed by changeovers and setups. Due to the type of information required, these models could only be used in the detail design stage. Jung [51] classified the features into four major categories and then further classified them based on machining operation. According to these detailed features, Jung developed a cost estimating system at the early design stage. Though he indicated the system could be applied at the early design stage, detailed design information is not typically available at the stage.

With the growth of CAD/CAM technology and 3D modeling, the feature-based approach has become very popular in part design. In their paper, Ouyang and Lin [52] stated that manufacturing cost was determined by shape complexity, product precision and tooling process. They estimated the manufacturing cost of a design according to the shapes and precision of its features through the integration with commercial CAD. But Ou-Yang and Lin [52] simplified machining processes. They did not consider the tool, the cutting speed, and the feed rate, which would impact the surface roughness.

In the feature-based approach, as in the parametric and analogy estimation approaches, the neural network approach could be combined with features-based modeling for cost estimation. Using a back-propagation neural network Zhang and Fub [45] proposed a feature-based cost prototype system to estimate the cost of packaging products.

Because the cost drivers of feature-based approach are design features. It makes it easy and convenient to let designer know how design features influence committed cost directly. However, there are limitations for feature-based cost modeling:

1) There is no widely accepted consensus on the definition of features. Even different CAD/CAM systems have different definition for a same feature;

2) The relation between features and costs is not easily defined;
3) Generally, it can be applied in the preliminary phase or later. The reason is that the feature-based approach needs more knowledge about product structure and features.

2.4.3 Process-Based Approach

The process-based approach is mainly employed in cost estimation for manufacturing a product. Haffner [53] stated that there exists an inherent relation between product design, processes, and product costs. The process-based approach maps material types, process technologies, design changes, and productions conditions to the part cost by establishing a relation between product design, material choice, process selection, and processing costs. The cost equations are established based on the physics of the underlying production process. The basic laws of physics for a variety of processes often provide the scaling between part design and the processing time.

The fundamental tenet of process-based cost modeling is a first order cost model [53]. It was observed that many manufacturing operations (humans and machines) can be represented as dynamic systems with first order velocity response to a step input. This behavior is amenable to a physical model by the following equation (2-7):

$$v = v_0(1 - e^{-t/\tau}) \tag{2-7}$$

where v_0 is the steady-state process velocity, τ is the dynamic time constant, and t is the process time.

From the above equation, the process time can be obtained and the cost can be calculated by using the following cost form (2-8) for the corresponding process [54]:

*Cost = (Manufacturing Process Step Time) * (Manufacturing Time, Resource Cost Relationship)* (2-8)

COSTADE [54, 55] is a process-based cost model which addresses fabrication, design and analysis costs. The fabrication costs depend on the time for the process step when the resources are consumed. Their cost equations were built through selection of critical design parameters (surface area, length, width, quantity, etc.), and a cost equation functional form, which best represent the physics of the problem for a range. This

physical insight can be used to define the critical design variables and their functional relationship to cost.

The process-based approach can predict the cost for a complex product, and new product which adopts new technology without directly applicable production data. The major disadvantage of process-based approach is that its development is often time consuming and expensive. In addition, it requires some engineering knowledge of the processes and the evaluated parts. Therefore, the process-based approach can often be only applied at the detailed design phase, at which there is a detail manufacturing plan for the product.

2.4.4 Activity-Based Costing (ABC) Estimating

In activity-based cost estimating, all cost drivers are associated with the activities required to produce the product. The design, manufacturing, usage and recycle /disposal for a product can be divided into all kinds of defined activities that are mutually exclusive. Based on historical, observed, or estimated data, the cost per unit of the activity's output is calculated. The estimated cost for a new product can be obtained according to the product's consumption of these activities.

Park and Kim [56] presented a set of activity-based cost drivers in their economic evaluation model for advanced manufacturing systems (see Table 2-1):

Table 2-1 Some Cost Drivers of Activity-based Estimating Approach [56]

Activity Type	Details
Costs as Needed	Direct material cost; Direct labor cost.
Activity Costs	Processing activity cost (utilities, equipment depreciation, insurance and property taxes, maintenance and repair, and floor space cost); Tooling activity cost; Quality control activity costs; Setup activity cost; Material handling activity cost; Inventory handling activity cost; Purchase order activity cost; Software-Related activity cost.
Nonactivity Costs	Unused activity costs, waiting time cost, inventory holding cost, and idle time cost.

Ong [57] developed an activity-based cost estimating system to help designers estimate the manufacturing cost of a printed circuit board assembly at the early concept stage of design. The author declares the model could be employed at the conceptual design phase, however the data needed would most probably available at and after the preliminary design stage.

Velcu, Ben-Arieh and Qian [58, 59] indicated that the advantages of ABC include the following: ABC is a relatively accurate method to estimate costs; it can track details of indirect cost-to-cost objectives, provide product cost information, and monitor cost behavior. However, they indicated that the shortcomings of ABC include the following: the ABC method requires greater effort and expense in obtaining the information required for the analysis; it is time-consuming analytical processes; they cannot always clearly identify the causal relationship between activities and products; ABC cannot self-learn like CBS and NNs; any change in business or manufacturing processes results in the reconstruction of the ABC model.

Based on the aforementioned properties, the ABC method can be generally employed at the detailed design phase when there is enough accounting information for the analysis. But it is not good for all products and all phases in the entire lifecycle, especially during the early design phases.

2.4.5 Simulation

Simulation is a powerful visualization method used to analyze system performance and thus improve the qualitative understanding of how cost is incurred in products. Moreover, simulation can model the stochastic nature inherent in a system. It thus provides a good tool for the analysis in nondeterministic situations, and more specifically for the study of risk and uncertainty in cost estimating area. Combining with ABC, the simulation-based model was proposed by Ozbayrak, Akgun, et al. [60] to estimate the product costs in an advanced manufacturing organization. Steele and Cope [61] proposed a methodology to estimate operational costs of reusable launch vehicle, which uses activity-based simulation as the platform to analyze the operations. Asiedu, Besant, et al.

[62] proposed a simulation modeling approach via kernel estimation techniques and thus applied this approach to a bidding problem.

Simulations can be used to study the dynamics of cost behavior. Forrester pioneered the work on systems dynamics and referred to his research as a simulation methodology [63-66]. In his paper, Forrester [66] said "system dynamics uses concepts drawn from the field of feedback control to organize available information into computer simulation models." Sterman [67] introduce the thinking tools of system dynamics which are mainly composed of causal loop diagrams and stock and flows. A causal loop diagram, the basic building block of system dynamics, denotes the cause and effect relationship (using an 'arrow') between two variables via a causal linkage. Stocks and flows and feedback are the two cornerstones of system dynamics. Stocks are accumulations due to differences in the inflow and outflow rates of a process. Stocks characterize the state of the system and provide information to base decisions or actions upon.

Abdel-Hamid and Madnick [68, 69] developed a dynamics model that estimates the time distribution of effort, schedule and residual defect rates using inputs (cost drivers) such as staffing rates, experience-mix, training rates, personnel turnover, and defect introduction rates for the software development. This model can continually reestimate effort and cost and compare targets against actual expenditure at each major milestone. It also incorporates managerial decision-making dynamics into continuous estimation models. Abdel-Hamid and Madnick [68, 69] stated that modeling technique could not only increase the fidelity of such models, but could also enable management to search for and test alternative interventions on a continuous basis. Monga, Damle and Scott [70-72] used system dynamics model to study the cost of the integration, the development, the operations, maintenance and disposal of new technologies for ship systems. In these system dynamics models, the trend and sensitivity of costs associated with some input variables are easily study. However, formulating the dynamics model is not an easy task. It needs strong expert knowledge and data in the process of establishing the model.

Therefore, simulation can be employed with enough information. In general, it does a good job to estimating cost and analyzing the cost behavior and distribution at the detail phase. However, as ABC and process-based cost modeling, it is not good for estimating costs during the early design phases.

2.5 Supporting Methodologies

This section provides literatures review about tabu search and support vector regression (SVR). They are supporting methodologies for this study.

2.5.1 Tabu Search

Tabu search, initially proposed by Glover [73, 74], is a mathematical optimization method, which has been widely used for combinatorial optimization. Tabu search is a meta heuristic that uses a memory function to avoid being trapped at a local minimum. To explore regions left unexplored by the local search procedure and then escape local optimality, its short term memory structures prevent search cycles. To perform an exhaustive search in the entire space by generating solutions that are not seen before or to analyze in depth a subset of promising solutions, its long term memory helps implement diversification and intensification mechanisms.

2.5.2 Support Vector Regression (SVR)

The foundations of Support Vector Machines (SVM) have been developed by Vapnik [75]. The SVM can be applied to both classification and regression problems. The SVM for regression is called Support Vector Regression (SVR) and applied in regression analysis. SVM has its solid mathematical foundation based on statistical learning theory (Vapnik-Chervonenkis (VC) theory) [75-81]. A major goal of VC theory is to characterize the generalization error instead of the error on specific data sets, which enable SVM to generalize well to unseen data. Unlike conventional regression techniques, the SVR attempts to minimize the upper bound on the generalization error based on the principle of structural risk minimization rather than minimizing the training error. This approach is expected to perform better than the empirical risk minimization principle employed in the conventional approaches. Moreover, the SVR is a convex optimization, which ensures that the local minimization is the unique minimization. Support vector machine has three key features:

1) Better generalization capability;

2) Global optimal solution using optimization theory;

3) Kernel functions for nonlinearity.

The following section introduces the structural risk minimization used by SVM compared to empirical risk minimization principle employed in the conventional approaches. Then Section 2.5.2.2 give a brief description of SVR. A more detailed description of SVR refers to [75-81]. The attractive features and limitation of SVR are summarized in Section 2.5.2.3.

2.5.2.1 Structure Risk Minimization versus Empirical Risk Minimization

The generalization error (structure risk) is a key concept in the SVM [75-81]. The goal is to estimate unknown real-valued function in the relationship (2-9):

$$y = f(x) \tag{2-9}$$

with a training data set $\{(x_i, y_i)\}_{i=1}^{l}$. The training data are independent, identically distributed (i.i.d) samples generated according to some (unknown) joint probability density function *p(x,y)*. An estimation procedure selects the best model *f(x)* from a set of possible models by minimizing (unknown) prediction risk, generalization risk (or error) (2-10):

$$R[f] = \int L(f(x), y)\, p(x, y)\, dxdy \tag{2-10}$$

The loss function can be defined as (2-11):

$$L(f(x), y)_\varepsilon = |y_i - f(x_i)|_\varepsilon \tag{2-11}$$

The following error defined on the training data set is usually called the training error or empirical risk (2-12):

$$R_{emp}^{\varepsilon} = \frac{1}{l}\sum_{i=1}^{l} L(f(x), y)_\varepsilon = \frac{1}{l}\sum_{i=1}^{l} |y_i - f(x_i)|_\varepsilon \tag{2-12}$$

Unfortunately, *p(x,y)* is unknown so that *R[f]* is difficult to formulate. However, Vapnik gave an upper bound (Equation 2-13) to this generalization error.

$$R[f] \le R_{emp}^{\varepsilon} + \sqrt{\frac{h(\ln(2l/h)+1) - \ln(\eta/4)}{l}} \tag{2-13}$$

Vapnik showed that for i.i.d. data and $l>h$, the bound holds with probability $1-\eta$, where the second term is confidence term which depends on the VC dimension *h* that

characterizes the capacity of the set of functions and it is a combinatorial measure for the model complexity.

For a linear function, it can be expressed as (2-14):

$$f(x) = \langle w, x \rangle + b \tag{2-14}$$

In the support vector regression, the risk function [80] of (2-14) is as (2-15):

$$\frac{1}{2}\|w\|^2 + C \bullet R_{emp}^{\varepsilon} \tag{2-15}$$

the first term measures the model complexity, the second term measures the training error or empirical risk. The goal is to minimize the generalization error which can be achieved by obtaining a small training error R_{emp}^{ε} while keeping the capacity of the set of functions (model complexity) as small as possible.

2.5.2.2 Support Vector Regression

Given a training data set $\{(x_i, y_i)\}_{i=1}^{l}$, where $x \in R^d$ is the input space. The SVR developed by Vapnik [75] relies on estimating a linear regression function (Equation 2-14):

$$f(x) = \langle w, x \rangle + b$$

where *w* and *b* are the slope and offset of the regression line. As above mentioned, the regression function is calculated by minimizing the objective function (it is also called the primal objective function) and it is subjected to the corresponding constraints (2-16):

$$\begin{aligned} \min \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{l} \left(\xi_i + \xi_i^*\right) \\ S.t. \quad & \begin{cases} y_i - w^T x_i - b \le \varepsilon + \xi_i \\ -(y_i - w^T x_i - b) \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \end{aligned} \tag{2-16}$$

where $\frac{1}{2} w^T w$ is the term characterizing the model complexity (smoothness of $f(\mathbf{x})$) and $C \sum_{i=1}^{l} \left(\xi_i + \xi_i^*\right)$ is the loss function determining how the distance between $f(\mathbf{x}_i)$ and the target values y_i should be penalized. The slack variables ξ_i and ξ_i^* are introduced for the

situation that the target value exceeds more than ε, see Figure 2-16. The constant $C>0$ determines the trade-off between the flatness of f (model complexity) and the amount to which deviations larger than ε are tolerated. The commonly used ε-insensitive loss function was introduced by Vapnik. This ε-insensitive loss function $|\xi|_\varepsilon$ is defined by (2-17):

$$|\xi|_\varepsilon = \begin{cases} 0 & if\ |\xi| \le \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{2-17}$$

In fact, this particular constraint defines a tube with radius ε around the hypothetical regression function in such a way that if a data point is positioned in this tube the loss function equals 0, while if a data point lies outside the tube, the loss is proportional to the magnitude of the Euclidean difference between the data point and the radius ε of the tube. The points lying outside the ε tube are named support vectors (SVs), because they will be used to estimate regression function. This implies that all other data points are in fact not important for inclusion into the model and can be removed after the SVR model has been constructed. Hence, usually (much) less training points do constitute the regression model.

By introducing a dual set of variables (Lagrange multipliers), the Lagrange function is defined as Equation (2-18).

$$\begin{aligned} L = \frac{1}{2} w^T w + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i - y_i + w^T x_i + b) \\ - \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* + y_i - w^T x_i + b) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*) \end{aligned} \tag{2-18}$$

Karush-Kuhn-Tucker (KKT) theorem states, a solution to the primal problem must satisfy the following (2-19):

$$\begin{aligned} &\partial_b L = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0 \\ &\partial_w L = w - \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i = 0 \\ &\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \\ &\alpha_i^{(*)}, \eta_i^{(*)} \ge 0 \\ &\alpha_i^{(*)} \quad refer \quad to \quad \alpha_i \quad and \quad \alpha_i^* \end{aligned} \tag{2-19}$$

Substitute all into L yields the dual problem (2-20):

$$\begin{aligned} \max \quad & -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_i,x_j\rangle-\varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*) \\ S.t. \quad & \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \quad and \quad \alpha_i,\alpha_i^*\in[0,C] \end{aligned} \tag{2-20}$$

So $w=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)x_i$, thus $y(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\cdot\langle x_i,x\rangle+b$

To solve a nonlinear regression or functional approximation problem, the SVR constructs a linear regression hyper plane in a high-dimensional feature space, which is nonlinear in the original input space via the mapping function: $\Phi: x\in X\mapsto\Phi(x)\in F$. However, it is suffices to know $k(x,x_i)=\langle\Phi(x),\Phi(x_i)\rangle$ rather than Φ for support vector regression. $K(x, x_i)$ is called the kernel function. It has been shown that a suitable kernel function makes it possible to map a non-linear input space to a high-dimensional feature space where linear regression can be carried out. Several kernel functions have been proposed in literatures, but the particular choice of a kernel to map the non-linear input space into a linear feature space depends highly on the nature of the data representing the problem at hand. The four widely used kernel functions are shown below:

- Linear (2-21):

$$K(x,x_i)=\langle x,x_i\rangle \tag{2-21}$$

- Polynomial (2-22):

$$K(x,x_i)=(\gamma\langle x,x_i\rangle+\beta)^d \tag{2-22}$$

- Radial basis function (2-23):

$$K(x,x_i)=\exp\left(-\gamma\|x-x_i\|^2\right) \tag{2-23}$$

- Hyperbolic tangent kernel (2-24):

$$K(x,x_i)=\tanh(\beta+\gamma\langle x,x_i\rangle) \tag{2-24}$$

Theoretically, the kernel function must satisfy Mercer theorem. (However, the last kernel function --- tangent kernel does not satisfy the Mercer theorem but has been successfully used in practice (for details see [76, 77])).

After mapping using the kernel, the regression formulas are as follows:

Linear (2-25):

$$y(x)=\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)\cdot\langle x_i,x\rangle+b \tag{2-25}$$

Nonlinear (2-26):

$$y(x)=\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)\cdot\langle\Phi(x_i),\Phi(x)\rangle+b \tag{2-26}$$

General (2-27):

$$y(x)=\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)\cdot K(x_i,x)+b \tag{2-27}$$

2.5.2.3 The Attractive Features and Limitations of SVR

Support vector regression (SVR) has many attractive features:

- It has the ability to model non-linear relationships;
- It has the ability to select only the necessary data points (support vector) to solve the regression function, which results in a sparse solution;
- The regression function is related to a quadratic problem (QP) which has a unique global solution in general;
- VC theory characterizes properties the generalization error which enable SVR to generalize well to unseen data;
- The SVR technique can be used when there are few samples than variables, which is also called small n large p problems [82].

But, SVR has some limitations:

- SVR raises a quadratic optimization problem of the same size as the training data set. There is a computationally demanding optimization problem.
- Currently, there is not a structure method to choose kernel function.

- There are a number of free parameters that need to be defined by the user. Since the generalization performance of the SVR models depends on a proper setting of these parameters, this is still a hard problem for applying SVR.

The problem of optimal parameter selection is further complicated by the fact that the SVR model complexity and its generalization performance depends on all of these parameters together (interaction of parameters). This means that a separate optimization of each parameter is not sufficient to find the optimal regression model. For this reason, usually a very time-consuming grid search optimization method is invoked to find the optimal SVR parameter settings, or some formulas based on the empirical study [83] are used to determine the appropriate parameters set. Chapter 5 will further discuss the choice of kernel functions and parameters.

2.6 Summary

First this chapter presented a generic cost model. It consists of three components: output (cost C), the relationship $f(x;\beta)$, and an input space x. Based on the generic model, the classification of cost estimating approaches is given as: approaches based on functional relationship and approaches based on inputs and structural relationship.

The cost drivers are fundamental to a cost model. The cost drivers are factors which have significant effect on final costs. Section 2.2.1 gave some cost drives published in literatures. Two traditional methods of identifying cost drivers in the literatures were then introduced.

This chapter has then presented an overview of a variety of estimation techniques. This discussion was organized into the approaches based on functional relationships and based on inputs and structural relationships.

Based on functional relationship, the approaches were classified into: expert judgment, parametric method, neural network approach and case-based reasoning approach. The significant drawback of expert judgment is its subjective nature, which makes the designer and decision-maker uncomfortable in using it. Parametric method is the most widely used formalized modeling method for cost estimating. But apriori knowledge of the functional form is needed. Also it is very difficult to deal with

nonlinearity. Case-based reasoning method is a good estimating approach which overcomes some drawbacks of parametric method. It does not need functional forms apriori and can deal with nonlinearity. However, it is hard to define a similarity measure and adjusting methods for case-based reasoning. It thus cannot guarantee the accuracy of estimating cost. Artificial neural network approach is a much more accurate technique. Artificial NNs are able to capture the nonlinearities, discontinuities, interactions among the cost drivers, and have capability of learning and adaptivity. But neural network approach has some weakness: it is a "black box"; it lacks explanation capabilities and does not provide an environment for directing user; producing near optimal neural network models is still a challenge task; and there are over fitting problems when there are lots of historical data.

The approaches based on the inputs and structural relationship (feature-based cost approach, activity-based cost estimating approach, process-based cost approach and simulation) are often applied in the preliminary or later phase and respectively do a good job under a certain situation and scenario. In the different situation and scenarios, different approaches have respective advantages and disadvantages. But at the early design stage these four methods may not work because the information about the structure is usually incomplete or uncertain.

The approaches based on functional relationship are building blocks for the approaches based on structural relationship and inputs. The performance of these four approaches based on inputs and structural relationship depends on the identification of inputs, structure and the approaches based on functional relationship. If there is not a good approach based on functional relationship, there would be no accurate estimation. The approach based on functional relationship is corner stones for cost estimation.

Cost estimation has always been difficult at the early stage of product development when only a few conceptual attributes of the product in question are known or for complex product. The relationship between these attributes and cost is very hard to obtain. And the discontinuity and nonlinearity often may exist in these relationships. From the above summary, the neural network approach has advantages over other estimating approaches when there is little apriori knowledge and nonlinearity and/or discontinuity to the CER and multicollinearity existing among the cost drivers. However the neural

network approach has over fitting problems. Also there is not a structure way to produce near optimal neural network architecture, training methods and stopping criteria. Moreover, it is a "black box" CER and does not provide any explanation for users.

Section 2.5 introduces the two methods, tabu search and support vector regression based on statistical learning theory. Tabu search is a memory-based stochastic optimization strategy. The SVR maps the data into a high-dimensional feature space via a kernel function and then performs linear regression in this space. Therefore SVR could model nonlinear relationships and have better generalization capability with a global solution than conventional methods. SVR should provide a better performance comparing to these conventional cost modeling approaches.

Therefore, this study will focus on a new way to identifying and selecting cost drivers and new cost estimating approaches based on SVR, which can be applied in the entire life cycle, especially at the early design phases. This approach will overcome the “black box” problem to be able to provide guide to designers.

Chapter 3 Research Methodology Framework

3.1 Introduction

A cost estimating model must be accurate and capable of operating on data of the detail typically available in the related phase, to support cost trade-off studies for designers and decision makers. The main purpose of this research is to provide new methodologies to obtain higher predictive accuracy of cost estimation and guide designers at the early design phases of complex products. However, accurately estimating cost is not an easy task at the early stage of complex product development when only a few conceptual attributes of the product are known. The relationship between these attributes and cost is very hard to obtain. Furthermore, discontinuity and nonlinearity often exist between them.

New methodologies for the generic cost estimation model (presented in Section 2.1) are discussed and studied in this study. First these methodologies can be used to identify the cost drivers from causal and associated aspect and then select the significant cost drivers. Second they will estimate the cost based on support vector regression (SVR) using pure nonparametric approach and semiparametric approach when existing nonlinear and discontinuous properties during the early design phases. After that, they will direct designer and decision-maker based on sensitivity analysis supported by SVR.

3.2 Research Methodology Framework

The cost drivers, the relationship $f(x; \beta)$, and the desired cost estimate are three basic elements for the generic cost estimation model. The appropriate cost driver set and appropriate relationship influence the final desired cost estimate. The identification and selection of cost drivers is important in a cost estimating model. Under certain situation, especially when not enough information is available for cost estimation, the cost driver set used may actually influence what relationship will serve as the model. In this study a new method, the Causal-Associated approach, is introduced to identify the cost drivers. A Tabu-Stepwise algorithm is then proposed to select appropriate cost drivers set.

To improve the predicable accuracy of cost estimating, it is important to choose an appropriate approach to estimate the cost. In the estimation of the cost of the complex product, if the parametric form of underlying function $f(x; \beta)$ is known apriori, a parametric cost estimating approach should be used. There are many references found in the literature (see Section 2.3.2) which discuss the parametric approach. However, in the initial design phases, for a complex product there may be inadequate knowledge of the relationships that exist, or there may exists nonlinearity and discontinuity in the relationship. Hence it is very hard to define a cost model. This study proposed a pure nonparametric cost estimating approach based on SVR and semiparametric cost approaches based SVR to estimate costs. If a product is very complex and/or the estimating process occurs at the very early stage of design and the functional form of the model is unknown, a pure nonparametric cost estimating approach should be used. For the situation in between, there is knowledge about the partial parametric form but this form is not adequate throughout the entire inputs, the parametric approach would not be appropriate because the resulting fit would be misleading (biased) at points where the data deviates from the specified model. However, it is not wise to ignore the knowledge and only use the pure nonparametric approach. Semiparametric approach, combining the parametric technique with the nonparametric technique, would be a good way for cost estimation in this situation. Therefore, in this study, new cost estimating nonparametric and semiparametric approaches, based on support vector regression (SVR), are presented to deal with the situation when there is not apriori knowledge of the functional form of the cost model or there exists some incomplete apriori knowledge about the model. These new approaches can improve the accuracy of cost estimation over conventional methods.

A cost estimation model not only needs to estimate cost as accurate as possible. At the design phases, but also the objective of a cost model is to help designer achieve good cost trade-off decisions. The cost estimating approach thus can provide the cost transparency (the impact of design alternatives on complex product costs) to designers. Two methods of sensitivity analysis are introduced to help designers evaluate the contribution of input variables for the final cost and provide explanation to the above nonparametric cost model based on SVR.

The proposed cost modeling approach in this research is composed of four parts: 1) identifying cost drivers via Causal-Associated (CA) method and eliminating the insignificant cost drivers using Tabu-stepwise method (Chapter 4); 2) estimating cost via the nonparametric approach based on support vector regression (Chapter 5); 3) estimating cost using semiparametric approach based on support vector regression (SVR) (Chapter 6); and 4) indicating the effect of cost drivers on cost for cost modeling based on SVR via sensitivity analysis (Chapter 7). The framework of the proposed cost estimating approach is as the Figure 3-1.

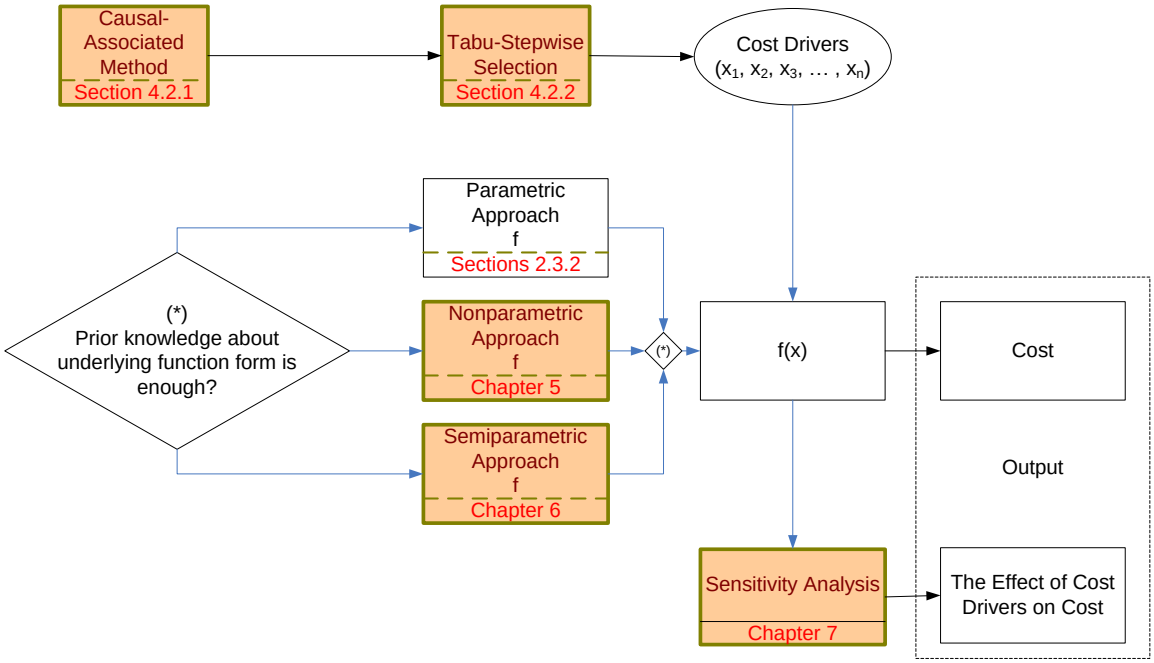


Figure 3-1 The Framework of the Proposed Cost Estimating Approach

The new method, Causal-Associated (CA), in this research is proposed to identify the cost drivers, which is different with traditional methods for identifying cost drivers. The acceptable and available cost drivers at the current stage, are found via this method. The CA method includes five components: cost breakdown structure (CBS), root cost drivers, associated cost drivers, relationships, and assumptions. Generally, there are five steps for identifying the cost drivers: 1. Decomposition; 2. Listing Root Cost Drivers; 3. Analysis; 4. Substitution; and 5. Gathering. The CA method not only reduces the chance of missing

data, but also provides a way to analyze the assumptions and preconditions of cost estimation. The detailed CA would be discussed in Chapter 4.

After identification, using a Tabu-Stepwise variable selection technique, the unrelated or insignificant cost drivers will be eliminated to reduce the variance in the model output and the cost of collecting the data. The Tabu-Stepwise algorithm based on Tabu-SVR selects the significant cost drivers to form a candidate set for cost estimates. This algorithm searches better candidate subset of cost drivers via 5-fold cross validation and employs RMSE as its criterion. The Tabu-Stepwise method is a stepwise search method and employs tabu-list in the searching process. The tabu list would record a number of history steps and prohibit repeated calculation at the future steps. The initial subsets would choose the results of Mallow's Cp, Adjusted R-square methods, or start from the first variable. The detailed Tabu-Stepwise would be discussed in Chapter 4.

The nonparametric approach based on SVR, Tabu-SVR, estimates cost combining support vector regression with the tabu search algorithm mentioned previously in Section 2.5. For a cost estimating nonparametric approach based on SVR, there are three steps to get final cost: 1. Data Preprocessing; 2. Choosing the kernel and parameters; 3. Training the SVR and computing the final cost. The parameters are determined using the tabu search algorithm via the cross-validation procedure. The performance criterion to choose the parameters is Root Mean Square Error (RMSE). There are three types of kernel (linear kernel, polynomial kernel, and radial basis function (RBF) kernel) to be investigated in this study. More information about the Tabu-SVR would be found in Chapter 5.

At times there may be limited some knowledge about the parametric form but full information about this form is not known, the semiparametric approach would be a good way for cost estimation in this situation. The semiparametric approach is able to combine a parametric component based on the researcher's knowledge of the underlying model with a nonparametric component designed to capture any structure in the data that the parametric fit fails to explain. According to different combining strategies on the nonparametric component and the parametric component, three semiparametric algorithms based on SVR are discussed in Chapter 6.

The development of a cost model for complex product design is not only done to provide accurate cost estimation but also to explain those complex and often non-linear relationships. In Chapter 7, there are two existing methods introduced for sensitivity analysis based on SVR. In the first method, a certain number of points with equal interval are produced in the range of the studied cost driver. In the second method, the cost estimating approach (nonparametric approach and semiparametric approach) based on SVR adjusts the input values of one variable while keeping all the others constant to approximate a gradient.

In this study, for verifying and validating the cost estimating approaches, five common basic cost characteristics are summarized in Chapter 5. They are: accumulation; linear function; power function; step function; and exponential function. These five common basic cost characteristics are often combined to represent the cost characteristics of a complex product. Based on those fundamental cost characteristics and general rules for combining terms along with FLOPS cost module, test cases (data sets) are produced to verify and validate the nonparametric and semiparametric cost estimating approaches.

Chapter 4 Identification and Selection of Cost Drivers

4.1 Introduction

Cost drivers are any factors which cause a change in the cost of work performed in the lifecycle of a product. The identification and selection of cost drivers is fundamental to a cost estimating model. The cost driver set that is used can determine what relationship (cost estimating approach) is applied in the cost model. Appropriate selection of the cost driver sets can make the estimation more accurate (smaller bias or unbiased) and reduce the variance of the estimate.

This research proposes a new method, Causal-Associated (CA) approach (see Figure 4-1), to identify the cost drivers, which is different with traditional methods for identifying cost drivers. The CA method utilizes the cost breakdown structure to list the root cost drivers and then find associated cost drivers to substitute for root cost drivers that are not available or too expensive to obtain. The more complete acceptable and available cost drivers at the current stage are identified via the CA method. After identification, using the Tabu-Stepwise selection technique, the unrelated or insignificant cost drivers will be eliminated. This can reduce the variance in the model output and the cost of collecting the data. The procedure of identification and selection of cost driver is shown in Figure 4-1. A case study is then conducted to show how the CA method and the Tabu-Stepwise technique identify and select cost drivers.

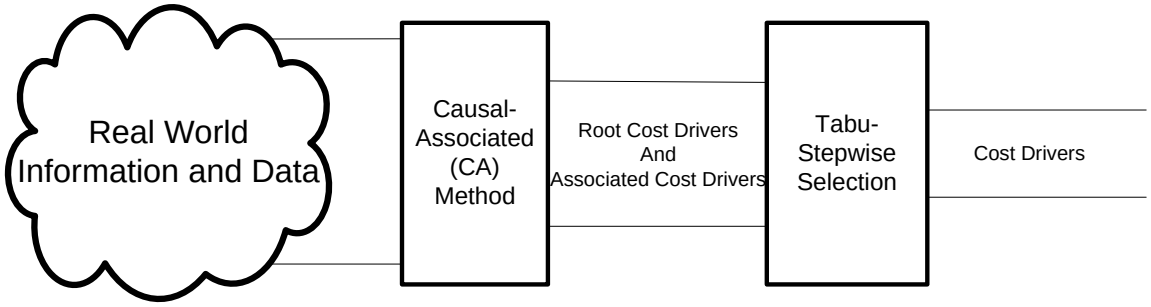


Figure 4-1 The Procedure of Identification and Selection of Cost Drivers

4.2 Methodology of Identifying and Selecting Cost Drivers

4.2.1 Causal-Associated (CA) Method

Correlation is not causation. A statistically significant link or high correlation between two variables does not imply that one causes the other because this could be coincidence or the result of another unmeasured variable related to the two variables. Correlation is often used as measure of the linear relationship between variables, and therefore it is inappropriate when the relationship is strongly nonlinear. From the statistical aspect [84], the associated relation means: *X* and *Y* are associated if and only if $\exists x1 \neq x2$ then $P(Y \mid X = x1) \neq P(Y \mid X = x2)$. The causal relation means: *X* is a cause of *Y* if and only if $\exists x1 \neq x2$ then $P(Y \mid X \text{ set}= x1) \neq P(Y \mid X \text{ set}= x2)$. This indicates the associated relation would function in some points of the set but the causal relation would function for all of the points of the set.

Causal relationship analysis [85] would bring more complete and correct understanding and explanation to the cost analysis. It therefore results in an improved predictive capacity. Completeness helps show what drives the cost and it formulates guiding principles and useful rules. Correctness provides greater insight and detail to cost analysis.

At the early design stage, not all root cost drivers discussed in Section 4.2.2.1 are measurable and affordable. While a root cost driver may not be available, there could exist some associated cost drivers defined in Section 4.2.2.1 to represent these unavailable or unacceptable root cost drivers. In this situation, care must be taken in the assumptions or preconditions when associated cost variables represent root cost drivers. They would be assumptions or preconditions of the final cost model. The following section provides a new method, Causal-Associated (CA) method, to list all available and acceptable cost variables including root cost drivers and associated cost drivers. All of these cost variables would be candidate cost drivers for the model. At the same time, the assumptions and/or preconditions associated with associated cost drivers would be identified in the process.

4.2.1.1 The Framework of Causal-Associated (CA) Method

CA method includes five components: cost breakdown structure (CBS), root cost drivers, associated cost drivers, relationships, and assumptions:

1. Cost Breakdown Structure (CBS): The CBS is a conceptual model for understanding and analyzing the root cost drivers. The desired cost output is decomposed into cost components. There are three ways to decompose a cost:
    a) Time phase method --- which depends on temporal sequence or phases in the whole process. The cost component of an aircraft can be divided into four cost components: the cost in the design phases, the cost in the production phase, the cost in the operation phase, and the cost in the disposal phase.
    b) Physical structure method --- which depends on physical constituents of the product. For example, the cost of an aircraft is composed of the cost of airframe, the cost of engine and the assembling cost. The cost of the airframe is sum of the costs of the following components: wing, tail, body, gear, nacelle, propulsion system, flight control, hydraulics, electrical, pneumatics, air condition, anti-icing, auxiliary power, furnishing and equipment, instrument, avionics, and assembling cost.
    c) Mixed method --- which combines time phases with physical structure. For example, the cost of an aircraft at the design and production phase is composed of five cost components: the cost for airframe in the design phases, the cost for airframe at production phase, the cost for engine at design phases, the cost for engine at production phase, the other miscellaneous cost.
2. Root Cost Drivers (see Figure 4-2)

    Root cost drivers are causal variables to cost. While there exist many different ways to categorize root cost drivers, one way will be considered as follows:

    - Materials
        - The Type of Material
        - The Volume of Material
    - Time/Count:

- Product Design Properties
- Production Process
- Technology
- Management
- Schedule
- Quantity of production
- Facility and Equipments

- Environment Variables
  - The Unit Cost of Material
  - The Unit Cost of Labor
  - The Unit Cost of Facility and Equipment Consumption
  - The Unit Cost of Energy
- Economic Factors
  - Price Index
  - Quantity
  - Learning Factor

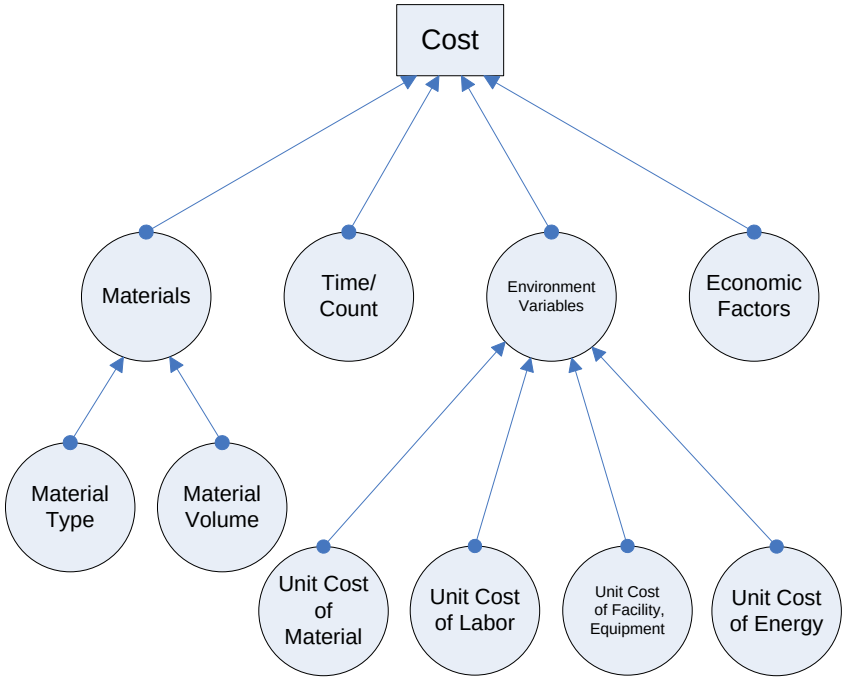


Figure 4-2 Root Cost Drivers

3. Associated Cost Drivers

   Associated cost drivers are the input variables that would have a significant effect on the final cost but are not truly root cost drivers. Associated cost drivers generally have a high correlation with cost, which can include performance, reliability, maintainability, and general operations. Correlation is not causation. For example, most aerospace cost models use weight as a cost driver. Weight is strongly correlated with the cost but generally it is not considered as a root cost driver for all cost components. Under some assumptions and preconditions, the associated cost drivers can be used to estimate the cost.

4. Relationships

   Causal-Associated method includes three relationships: accumulation, causality and associated relationship. The accumulating relationship exists between a cost component and all cost components in its sublevel (see Figure 2-1). Causal relationships exist between a cost component and its root cost drivers, and between root cost drivers in different level (see Figure 4-2). Associated relationship exists between a root cost driver and its associated cost drivers.

5. Conditions and Assumptions

   Because correlation is not causation, correlation generally happens under some conditions or assumptions. When associated cost drivers are found and associated relationships are established, there must be some assumptions and preconditions. These assumptions and preconditions would be indicated as the assumptions and preconditions of the cost estimating model while using these associated cost drivers.

4.2.1.2 The Procedure of Causal-Associated (CA) Method

Generally, the identification of cost drivers has five steps for identifying the cost drivers (see Figure 4-3):

1. Decomposition: The cost components are decomposed until they cannot be broken down more in a meaningful way. The cost of a product is conceptually broken into a number of cost components. This helps simplify the problem and

determine the cost drivers. Generally a Cost Breakdown Structure (CBS) is constructed according to time phases and/or physical properties. The CBS is an accumulation process.

2. Listing Root Cost Drivers: At the lowest lever of CBS, all important root cost drivers are listed from four perspectives: material; time/count; environment variables; and economic factors. Each cost component at the lowest level must be linked to some root drivers which are the causes of cost of the corresponding components. All root cost drivers of each component are listed as completely as possible.
3. Analysis: The availability and acceptability of the root cost drivers are analyzed. An analysis is conducted to determine if the root cost drivers are available and acceptable.
4. Substitution: If the root cost drivers are unavailable or unacceptable, the associated cost drivers are found to substitute them under corresponding assumptions. If these associated cost drivers are not available and acceptable, the causal and associated analysis will continue until all cost drivers are acceptable and available under some assumptions.
5. Gathering: All available and acceptable root cost drivers and associated cost drivers are gathered; and all redundant drivers to form the possible set of cost driver are eliminated. At the final step, all available and acceptable root cost drivers and associated cost drivers, which are gathered and output, are candidate cost drivers. They will be employed to estimate costs of a product.

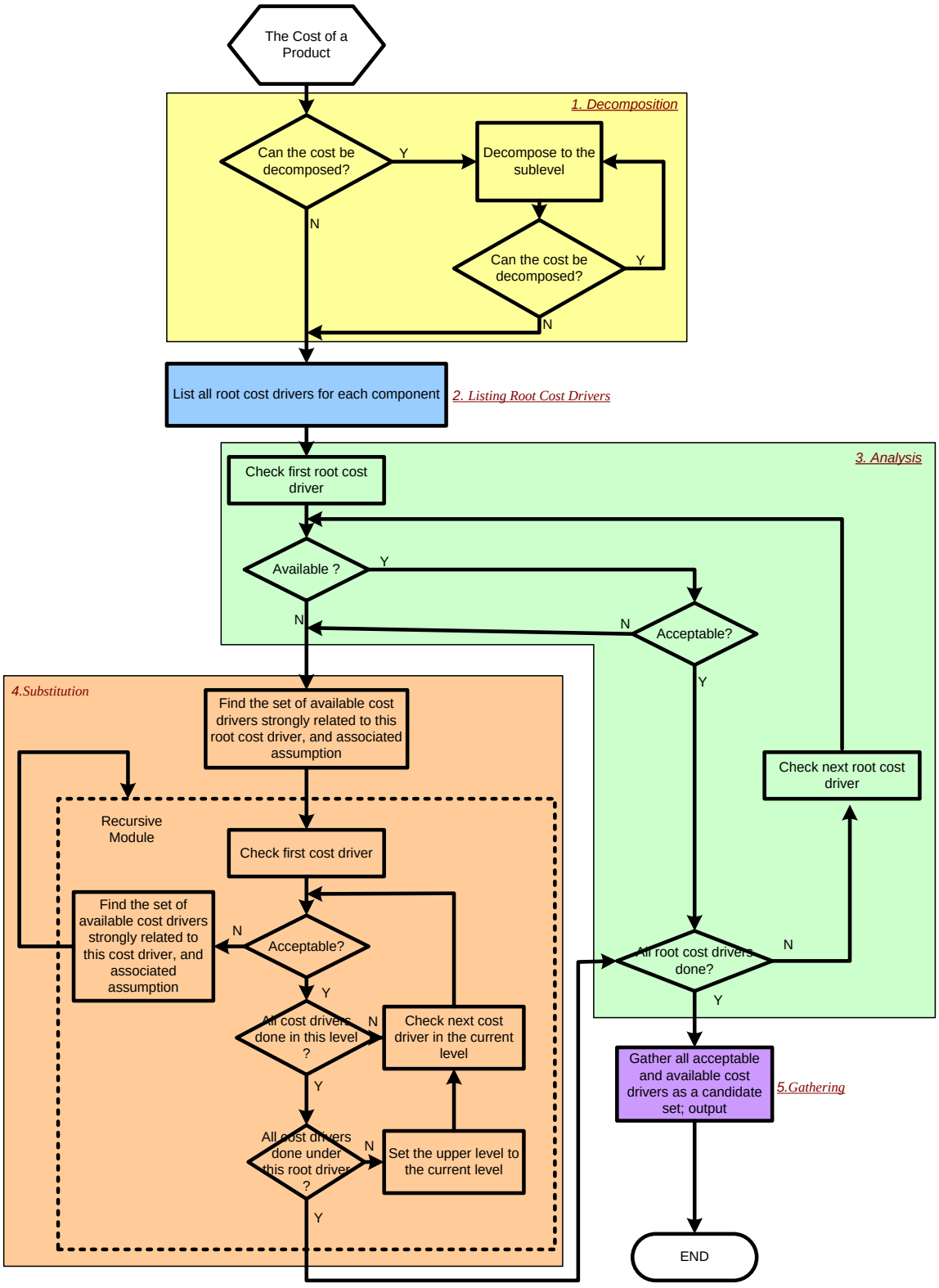


Figure 4-3 The Procedure of Causal-Associated (CA) Method

4.2.1.3 Comparisons with Traditional Methods of Identification of Cost Drivers

The Causal-Associated method assists the designer in considering all factors and in avoiding missing some cost drivers. This can help reduce the bias and improve the degree of estimating accuracy. When using associated cost drivers to represent some root cost drivers, the assumptions and preconditions are easily identified.

This Causal-Associated method is different with the traditional method (see Figure 4-4). The traditional methods identify potential cost drivers from references found in the literature and experience (see Section 2.2.2). After these potential cost drivers are grouped and reviewed, the candidate cost drivers are then determined using statistical analysis based on the data or expert survey. This method cannot guarantee the completeness and correctness of cost drivers. It just selects cost drivers from the existing literatures and experience. This method potentially misses some input information and cannot provide assumptions and preconditions for cost estimation as the Causal-Associated method.

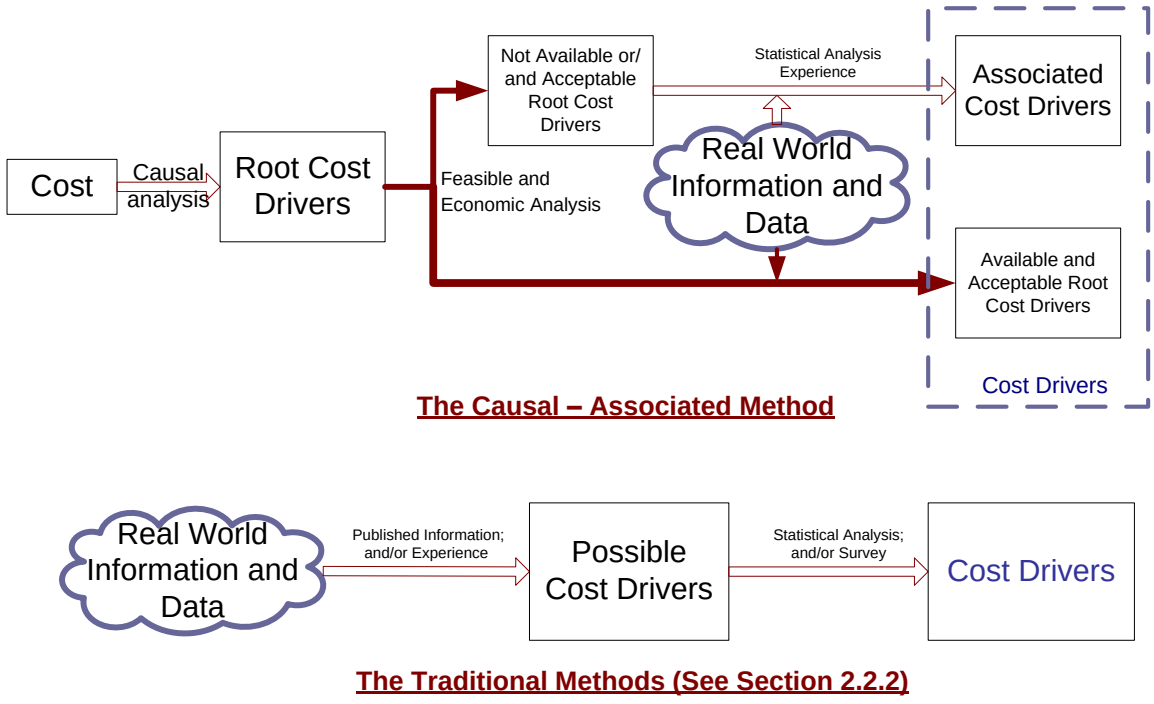


Figure 4-4 Comparison between Causal-Associated Method and Traditional Methods

A comparison between the Causal-Associated method and the traditional methods is illustrated in Figure 4-5. This example is assumed to happen at an early design phase.

The root cost drivers are x_1^*, x_2^*, x_3^*, x_4^*. The root cost driver x_3^* is correlated with x_1, x_2, x_3 under some assumptions. The root cost driver x_4^* is correlated with x_4, x_5 under some assumptions.

The root cost drivers, x_3^*, x_4^*, are not available currently.

Possible cost drivers for the product based on experience and other published resources are included in the set A. However all product information related to the cost are included in the set B. The real world data and information about the product is in the set R. Their relationship is $A \subseteq B \subset R$.

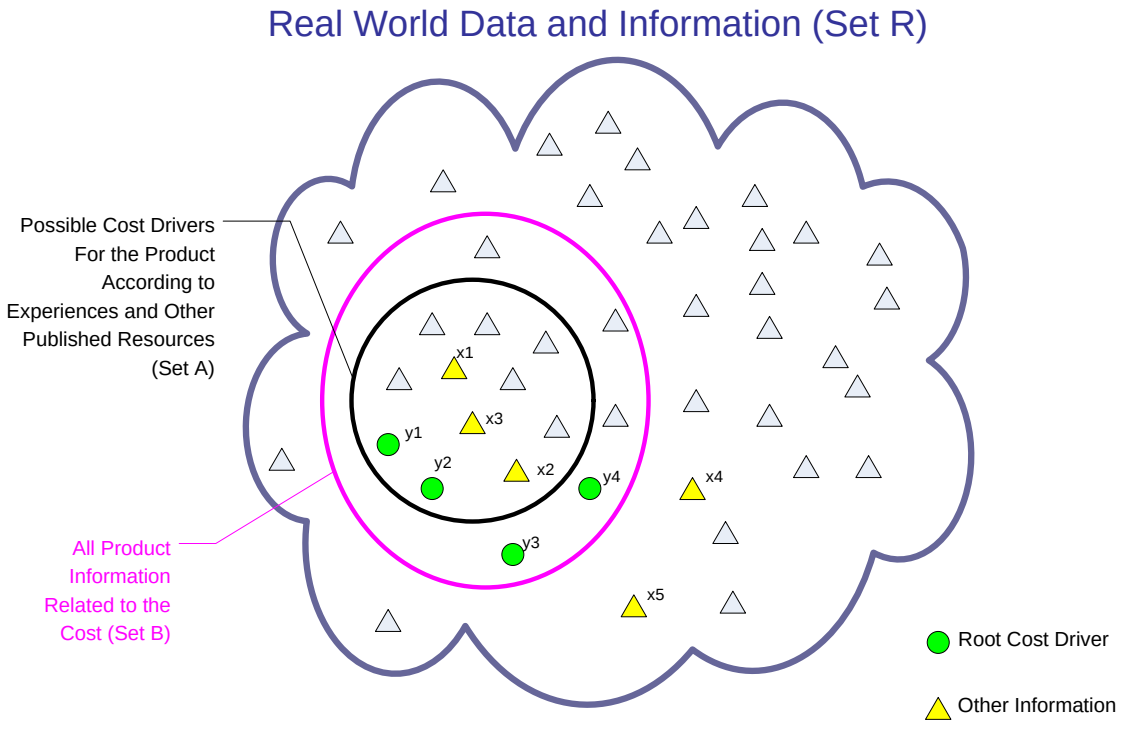


Figure 4-5 A Comparison Example for Identifying Cost Drivers

- Traditional Methods

  From the literature and experience, possible cost drivers are listed based on Set A. After grouping, reviewing, statistical analysis or expert survey, x_1, x_2, x_3, x_1^*, and x_2^* are the final cost drivers as the result of traditional methods.

- The Causal–Associated Method

  From the causal analysis, the root cost drivers will be x_1^*, x_2^*, x_3^*, and x_4^*. Because x_3^* and x_4^* are not available, associated cost drivers would be identified to represent them and appropriate assumptions and conditions could be defined.

  The Causal-Associated method can find the variables x_4 and x_5 to represent x_4^* which cannot be identified in the traditional methods. Additionally, the Causal-Associated method determines the assumptions for x_1, x_2, and x_3 representing x_3^*. These assumptions would be the preconditions of the cost estimating model. The traditional methods ignore these assumptions. Finally, x_1, x_2, x_3, x_4, x_5, x_1^*, and x_2^* are the candidate cost drivers under these assumptions.

In summary, the CA method not only reduces the chance of missing data, but also provides a way to analyze the assumptions and preconditions of cost estimation. It assists the designer in considering all factors and in avoiding missing some cost drivers. This can help reduce the bias and improve the degree of estimating accuracy. When replacing some root cost drivers with associated cost drivers, the assumptions and preconditions are easily identified.

4.2.2 Tabu-Stepwise Selection Based on Tabu-SVR

4.2.2.1 Introduction of Tabu-Stepwise

The variable selection methods have extensively been studied in linear models. Generally, the first type of approach to variable selection is a sequential approach. It includes three possible methods: forward selection method, backward elimination method, and stepwise selection method. The forward selection method adds one variable at a time to a model until the addition of another variable does not significantly improve the performance criterion. The backward elimination method begins with the model with all variables and drops one variable at a time until eliminating a variable significantly worse the performance criterion. The stepwise selection method is the most popular. It is a

combination of both forward selection method and backward elimination method. Its procedure begins by adding one variable at a time to a model but each time a new variable is added, all previously entered variables are re-evaluated and possibly dropped. The stepwise procedure ends by adding and dropping variables until the "best" subset is found. Another type of approaches to variable selection would examine all possible models from the total list of future variables. This approach includes R-square, MSE, Adjusted R-square, Mallow's Cp, etc.

The biggest drawback of first type of approach is that the methods cannot guarantee that they will find the best solution. The major shortcoming of the second type of approach is that the methods must examine all possible models (the number of models = $2^{the\ number\ of\ variables} - 1$). For example, if only fifteen variables were considered for the model, the number of possible models would be *32,767*. This can result in computation times that are not acceptable.

Additionally, for a complex product in the early design phases, it is known that there are nonlinear relationships between cost drivers and cost and generally there is not enough information about function form and cost relationships. The above variable selection methods, based on a linear model, are not adequate for the cost driver selection for complex products during the early design phases.

In a word, the purpose of the method of cost driver selection is to find the preferred solution without consuming excessive computational resources for complex products during the early design phases.

Therefore, an improved stepwise method, the Tabu-Stepwise selection method based on Tabu-SVR, is proposed to deal with the problem of cost driver selection for complex products during the early design phases. The basis of this selection method, the cost model based on Tabu-SVR, is a nonparametric model based on support vector regression, which is presented and discussed in Chapter 5. The Tabu-Stepwise algorithm employs Tabu-SVR to find the appropriate parameters via 5-fold cross validation, and use a stepwise search and a tabu list in the searching process to reduce the calculation time. Additionally, it can start from initial subsets which are the results of the Mallow's Cp and Adjusted R-square methods, or the first variable.

4.2.2.2 The Procedure of Tabu-Stepwise

The Tabu-Stepwise is based on Tabu-SVR (see Chapter 5) to find a best subset of cost drivers. The performance criterion (CV-MSE) is Mean Square Error via 5-fold Cross Validation. It is a stepwise search method and employs a tabu list in the searching process. The tabu list would record a number of history steps and reduce the chance of repeated calculation for the future search.

The flow chart of Tabu-Stepwise is as Figure 4-6. The procedure of Tabu-Stepwise is as follows:

Step 1: Construct the initial subset (The final results of Mallow's Cp, Adjusted R-square, or the first variable were chosen as the initial subset in this study);

Step 2: Calculate the CV-MSE of initial subset using Tabu-SVR;

Step 3: Initialize the tabu list, loop variables i, j;

Step 4: Add in the i^{th} variable if i^{th} is not in the model; form the new subset, and then calculate the performance of this new subset using Tabu-SVR and record the new subset in the tabu list if it is not in the previous tabu list; compare the MSE and then determine the candidate subset;

Step 5: Drop the j^{th} variable if j^{th} is in the model; form the new subset, and then calculate it using Tabu-SVR and record it in the tabu list if it is not in the previous tabu list; compare the CV-MSE and then determine the candidate subset;

Step 6: Return Step 5 if j is less than the number of variables; return step 4 if i is less than the number of variables;

Step 7: Output the final subset with the smallest MSE.

In summary, the Tabu-Stepwise method can select the "best" cost driver set for a complex product in the early design phases. It takes advantage of the benefit of stepwise selection and tabu list, which not only have better performance than the forward method and back elimination method but also can save plenty of computation resource.

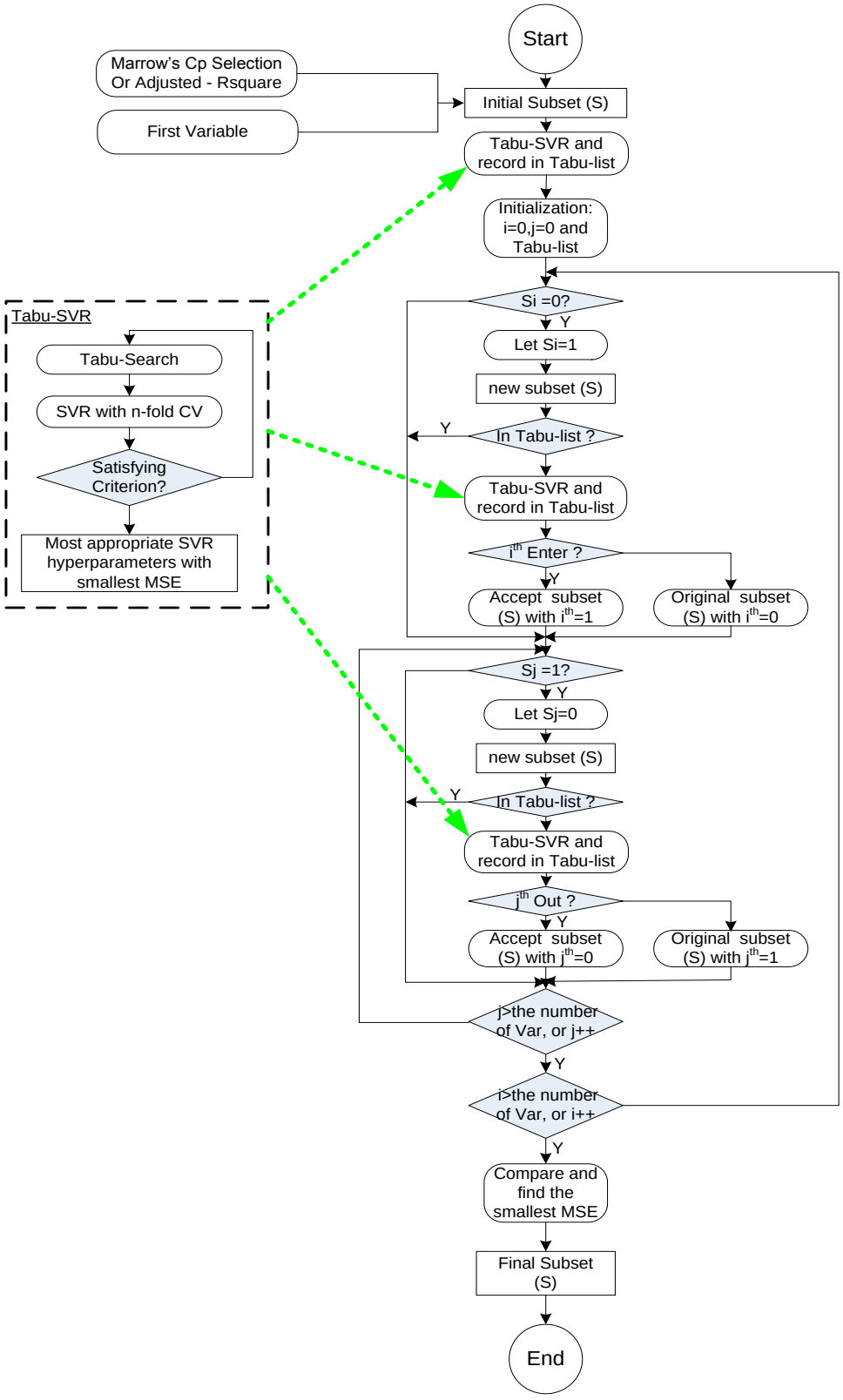


Figure 4-6 Flow Chart of Tabu-Stepwise Selection Method

4.3 Case Study

4.3.1 Overview of Case Study

The purpose of this case study is to demonstrate the feasibility of the Causal-Associated method to identify cost drivers and to illustrate the use of the Tabu-Stepwise based on Tabu-SVR selection methods to select significant cost drivers and eliminate irrelevant cost drivers. It is intended to show the value of the methods in being able to identify all cost drivers, provide a way to analyze the assumptions and preconditions of cost estimation, and then avoid adding extra noise, deteriorating the accuracy of the model, and clouding meaningful relationships which exist between important variables.

An electric motor (AC) is used as the object of this case study for the identification of cost drivers. Electric motors are often a component of a more complex system. According to the design and manufacturing process of an AC motor and following the Causal-Associated method described before, the cost are broken down, the root cost drivers are listed and analyzed. For those unacceptable and unavailable root cost drivers, the corresponding associated cost drivers and assumptions would be obtained. Finally all available and acceptable cost drivers would become the set of possible cost drivers.

To eliminate irrelevant variables and improve the accuracy of cost estimation, the Tabu-Stepwise based on Tabu-SVR selection method is used to select significant cost drivers set from all possible cost drivers. Different initial sets were used in Tabu-Stepwise, which are the results of Adjusted R-square using SAS, the results of Mallow's Cp using SAS, and the first variable. Because Tabu-SVR is employed in this search, the nonlinear properties existing between cost and cost drivers and among cost drivers are not ignored.

The final selected cost drivers by Tabu-Stepwise selection method would form the candidate cost drivers set. They would be used for cost estimation based on support vector regression discussed in Chapter 5 and Chapter 6.

4.3.2 Description of Case Study Background

4.3.2.1 Choice of Product

When a complex product is decomposed by its physical structure, the product cost is often the sum of the cost components as illustrated in Figure 4-7. In this case study, this cost component chosen for research is the cost of an AC motor (460v, three-phase, 100 hp, 1800 Speed). For many complex products, an electric motor is one of their important components. It often requires being designed and then manufactured for different complex products and satisfying special needs of some customers. Therefore, the process for identifying and selecting cost drivers for an AC motor is typical and needed in analyzing components of a complex product.

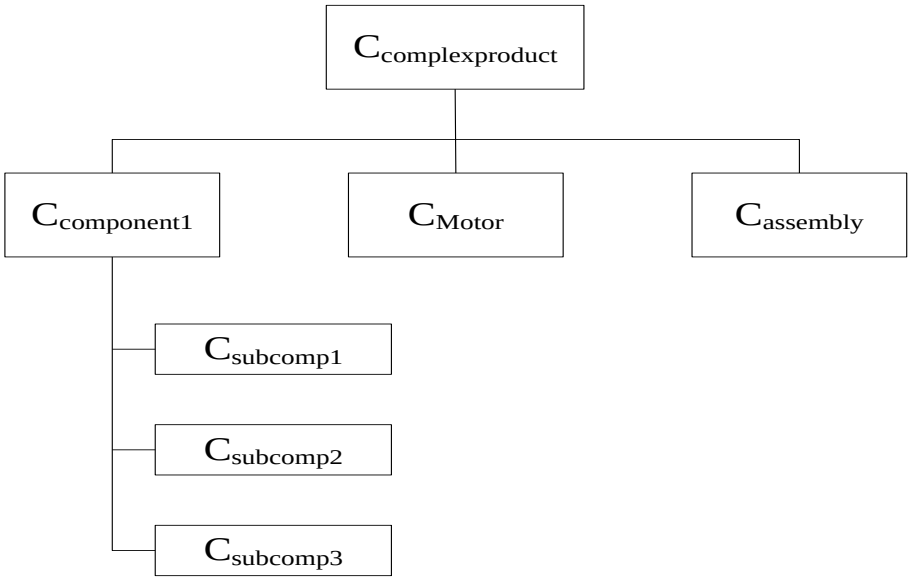


Figure 4-7 A Complex Product of Cost Breakdown Structure

4.3.2.2 Structure and Components of AC Induction Motor

An induction motor is comprised of the following basic components:

- Stator --- It consists of a number of coils of wire wrapped on laminated iron cores. The windings utilize electrical power to produce a rotating magnetic field in the rotor-stator gap, and thereby transfer drive power to the rotor.

- Rotor --- It consists of a cylindrical arrangement of copper or aluminum conducting bars attached to two end rings at either end of the bars. The magnetic field of the rotor couples with the rotating magnetic fields of the stator to produce mechanical torque and drive the load.
- Frame --- There are two parts: housing and end frame. The motor frame supports the rotor and stator during operations, and provides enclosure of the motor environment. The frame attaches to the foundation, supplying support and reaction to driver torques.
- Other Miscellaneous Parts --- Bearings support the rotating shaft within the stationary motor housing. Lead wire and terminations are used for the connections between the power supply and motor.

4.3.2.3 Design, Materials and Manufacturing Process of AC Induction Motor

- **Design of AC Induction Motor**

  There are several CAD packages and programs available to assist motor designers and researches. They make motor design and analysis much easier and faster. But for original design and special requirements on motor, the designer always needs to perform the actual procedure. The crucial stages of the design procedure are: specification elaboration; design considerations; dimensioning procedure; performance calculation; initial evaluation; design formation and layout; final evaluation; finalized design layout; and technical drawings and documents [86].

- **Materials and Manufacturing Process of AC Induction Motor**

  Materials for manufacturing of an AC induction motor and its parts include iron and aluminum castings, steel tubing and shafting, copper wire, steel laminations, purchased bearings, epoxy coating, varnish, adhesive, cleaning chemicals, etc. The four separate parts of the product (wound stator core, rotor core with shaft, frame and other miscellaneous parts) are manufactured or purchased separately and then assembled into complete motor units as Figure 4-8. The following sections briefly introduce these four separate parts separately and their materials.

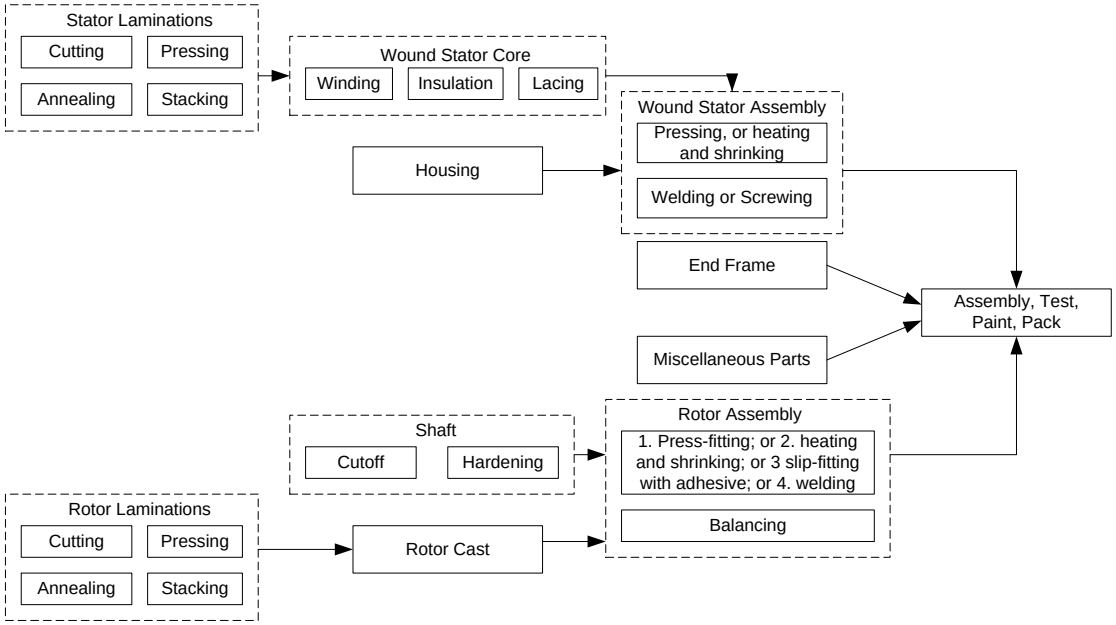


Figure 4-8 The Manufacturing Process of an AC Motor ([87])

- **Stator Core:**

  Typical stators are comprised of steel laminates with uniform slotting around the inner diameter (ID) (see Figure 4-9). The laminations are stacked into a core. Copper coils, insulated for the appropriate voltage level are wound into the slots (a wound stator) to deliver the supply power in the correct spatial and phase orientation.

  1. Stator Laminations (Figure 4-9)

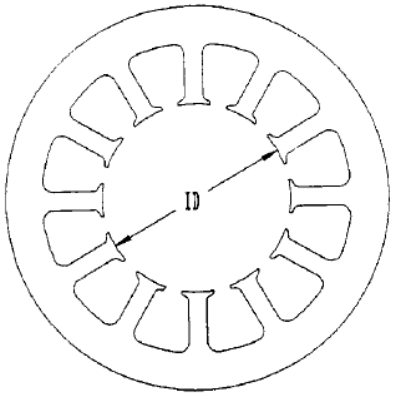


Figure 4-9 Stator Lamination

Different types of steel have different magnetic properties. The appropriate steel will depend on motor design. After pressing and annealing, laminates are cut or pressed from the sheet. Stator laminations are then stacked together to form a stator core via welding, bonding, or cleating.

2. Insulators

   The stator winding must be insulated from the stator core. Two processes are often used: epoxy coating and slot liners.

3. Coils

   Copper coils insulated for the appropriate voltage level are wound and injected into the slots (a wound stator). Then Lacing is used for tightly holding all of the magnet wires in the end turn through the resin bonding process.

- **Rotor:**

The rotor is comprised of a stack of thin insulated laminations shrunk, pressed or welded onto the shaft, with stack plates at each end to maintain a high interlaminar pressure. Conducting bars, usually made of copper alloy, run down the length of the rotor body either parallel to the axis of rotation, or with a uniform skew.

1. Shaft

   High strength steel with good fatigue characteristics is typically required. Most motor manufacturers use SAE 1045 in either cold-rolled or hot-rolled steel (CRS or HRS), or sulfurized SAE 1117, SAE 1137, SAE 1144, hot-rolled SAE 1035, and cold-rolled SAE 1018([87]). The CNC Swiss turning machines is mostly used to complete the shaft. Then the shaft is hardened.

2. Rotor laminations

   The materials and manufacturing process are mostly the same as stator laminators. An insulation coating, applied to prevent induction of axial current, keeps rotor body losses to a minimum.

3. Rotor bars

   Aluminum, copper, and copper alloys with low electrical resistance are most often used.

4. Connection ring and retaining ring.

   Comparable to the rotor bars, the connection ring requires low electrical resistance to minimize losses. Aluminum, copper and higher strength copper alloys can be used. Retaining ring typically comprises high-strength alloy steel with good fatigue characteristics.

- **Frame:**

The housing or frame is used to cover the stator, provide heat transfer and protection, provide a location for mounting the end frames, and serve as an attachment for other components, such as outlet boxes and lifting hooks. The end frames are used to contain the shaft bearings, support the rotor assembly, and act as a heat transfer device.

1. Housing

   The housings are made of cast iron; in rolled, wrapped, and tube steel; or in both cast and extruded aluminum tube. Different materials have different manufacturing process. For cast iron completed on either manual machines or CNC machining centers, the processes are: machine and drill the mounting feet; bore the inner diameter (ID); turn the end frame registers; drill and tap for the end frame attachment; and mill for the outlet box attachment. For rolled steel, after a stamping press this piece is formed around a mandrel, welded, machine-faced to length; a stamped mounting base is welded to the housing. For wrapped steel, the manufacturing processes are the same as for a rolled housing except that the stator core is used as the mandrel. For tube steel, the process is: cut to length, machine end frame diameter, and weld mounting feet. For aluminum castings, they are machined like cast iron with the same type of equipments. For aluminum tubing, the

material is cut to length and the mounting feet are then welded or screwed to the housing.

2. End Frames

   Like housings, end frames come in cast-iron, steel, zinc, or aluminum castings. For cast-iron castings, a computer numerically controlled (CNC) machine prepares the bearing bore and end frame diameter, and a manual drill is used to prepare the holes for the housing attachment. For the steel material, it is processed through a stamping press. For zinc or aluminum end frames, they are usually cast in a horizontal die caster.

- **Miscellaneous Parts:**

1. Bearings

   Bearing systems are used to support the rotor and shaft within the stationary motor housing and reduce the friction between the shaft and the end frames.

2. Lead wire and terminations (Studs, screws and terminals).

   Lead wire and terminations are used for motor to connect the motor to a power source.

This study will focus on the cost of an AC motor associated with design and manufacturing phases while estimates happen during the early design phases. With limited historical cost information and knowledge of these phases, the cost of AC motor can be break down to the basic cost object unit. For this basic cost object, the identification of cost drivers would be conducted according to break down structure and the detailed procedures during these phases.

For an AC motor, depending on temporal sequence, there are two phases: design and manufacturing phases. Then for the design phase of an AC motor, there are crucial stages: specification elaboration; design considerations; dimensioning procedure; performance calculation; initial evaluation; design formation and layout; final evaluation; finalized design layout; and technical drawings and documents. For the manufacturing phase of an

AC motor, there are four components to be considered in this research: stator; rotor; frame; other miscellaneous parts. For different parts, there are different processes which also vary with their materials and manufacturer.

4.3.3 Identification of Cost Drivers

4.3.3.1 The Procedure of Identification of Cost Drivers

Generally, the identification of cost drivers for an AC motor consists of five following steps as discussed earlier.

1. Decomposition: the cost components are decomposed until they are acceptable;
2. Listing Root Cost Drivers: at the lowest lever of CBS, all important root cost drivers are listed from four perspectives: material; time/count; environment variables; and economic factors;
3. Analysis: the availableness and acceptableness are analyzed for the above root cost drivers;
4. Substitution: If the root cost driver is unavailable or unacceptable, the associated cost drivers are found to substitute that root cost driver under corresponding assumptions;
5. Gathering: all available and acceptable causal and associated cost drivers are gathered; all redundant drivers to form the possible set of cost driver are eliminated.

4.3.3.2 Decomposition

As indicated earlier, there are two cost components that are considered in this case study: the cost of AC motor design and the cost of production of the AC motor.

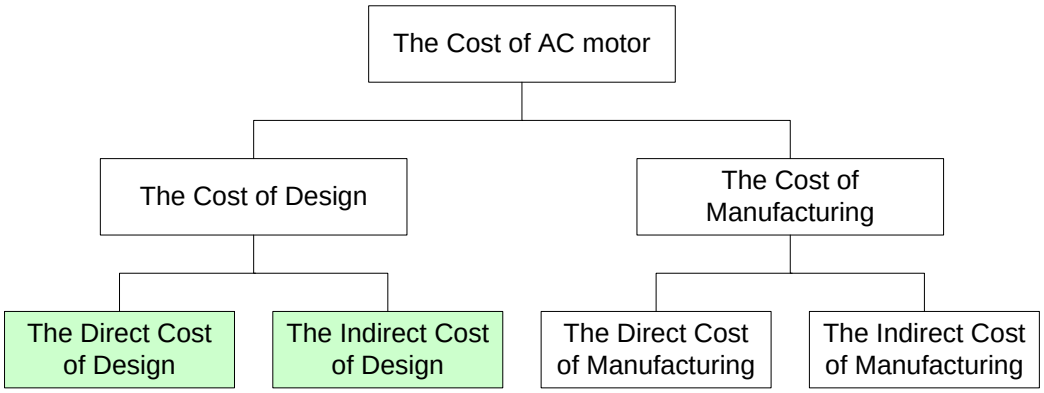


Figure 4-10 Decomposition of the Cost of AC Motor

- **The Cost of AC Electric Motor Design**

For simplicity, it is assumed that there will not be any prototype motor to manufacture. If this were not true, the process of analysis would be same as the manufacturing phase. Although the design process could be divided into multiple stages, in this case study it is considered as a single direct cost component plus one indirect cost component (see Figure 4-10). This is reasonable in that design is a labor intensive iterative process. The direct cost component includes direct labor cost, materials cost, and equipment charges associated with design. Indirect cost component incorporates the administration cost, related indirect material cost, and equipment charges. Those do not directly impact the design but are a definitely component of the final design cost.

- **The Cost of AC Motor Manufacturing**

The cost of the AC motor manufacturing has two cost components: direct cost of manufacturing and indirect cost of manufacturing. The direct costs are costs directly attributable to the manufacturing of an AC motor. Indirect costs are costs not directly allocated to an AC motor associated with manufacturing such as depreciation or supervisory expenses.

1. The direct cost of manufacturing

The direct cost of manufacturing is decomposed according to manufacturing processes of the corresponding AC motor. This study

focuses on the processes discussed before (see Figure 4-8). If the cost breakdown structure follows the basic motor components, the direct cost of manufacturing consists of five components: stator costs, end frames cost, miscellaneous parts costs, rotor costs, and assembly costs.

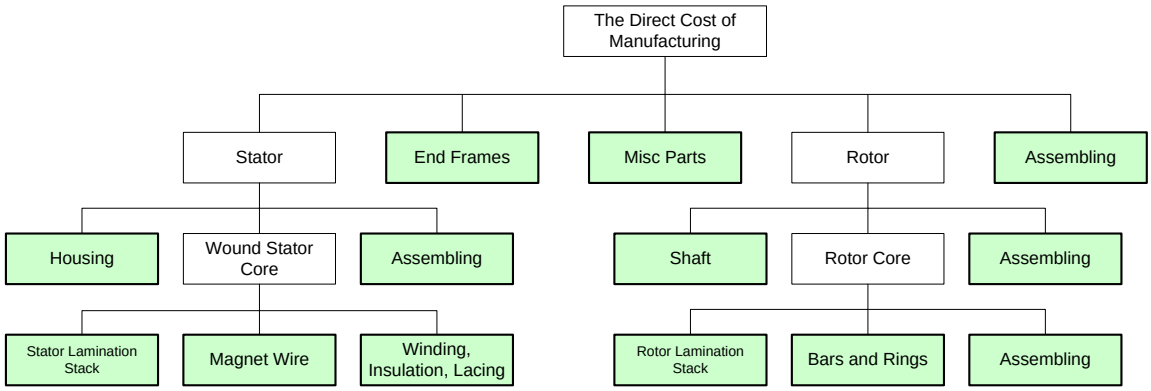


Figure 4-11 Decomposition of the Direct Cost Manufacturing (AC Motor)

For the decomposable cost components, they are broken down to subcomponents and the analysis continues until decomposition of the lower levels is no longer needed. The cost of stator includes three subcomponents: housing cost, wound stator core costs, and assembly costs. The cost of wound stator core is divided into stator lamination stack cost, magnet wire cost, and the cost of component for winding, insulation and lacing. The cost of rotor also includes three components: shaft cost, rotor core cost, and the cost of the component for assembly. The rotor core comprises of three subcomponents: rotor lamination stack cost, the cost of rotor bars and the rings for connection and retaining, and the cost of assembly. All shade rectangles in Figure 4-11 are indecomposable including end frames; miscellaneous parts; final assembly; housing; stator assembly; stator lamination stack; magnet wire; the component for winding, insulation, and lacing; shaft; rotor assembly; rotor lamination stack; bars and rings; rotor core assembly. For these indecomposable components, the root cost drivers are listed.

2. The indirect cost of manufacturing

   The indirect cost of manufacturing often includes two subcomponents: production overhead and corporation overhead. In this study, for simplicity, the cost associated with supply and demand is not considered, although the identifying processes of their cost drivers are the same as the processes described here.

4.3.3.3 Listing Root Cost Drivers

For indecomposable cost component, the root cost drivers are considered based on four basic types: materials, time/count, environment variables, and economic factors (see Figure 4-12) and listed under these categories. This section respectively lists the root cost drivers for all indecomposable cost components in last section.

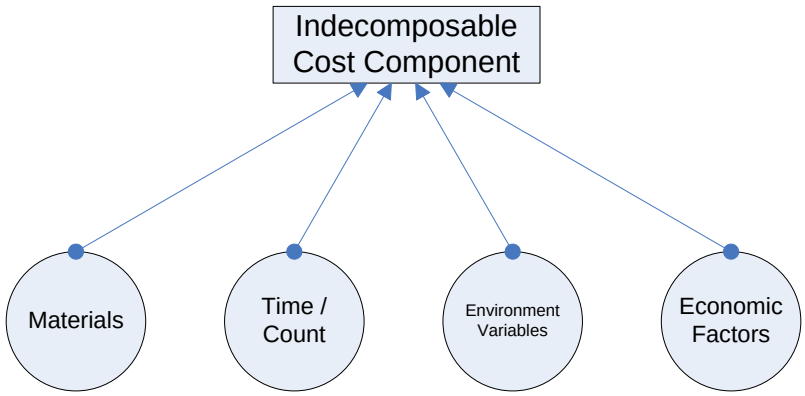


Figure 4-12 The Studied Indecomposable Cost Component

According to Figure 4-10, after decomposing the cost of design, two following cost components are not decomposed, which are listed root cost driver based on the above four basic types:

1. Direct Cost Component for AC Motor Design

   Design is a labor intensive process. Materials typically are accounted for a very small part of the total design cost, which can be neglectable. Moreover, modern AC motor designer often employs CAD or other software to aid design, which would incur the cost. So for this component, design labor time is a major root cost driver

which mostly impacts the cost. Also equipment time being used is another root cost driver. Correspondingly, environment variables compose of two drivers: design labor cost per unit time and equipment cost per unit time. Economic factors are current price index.

2. Indirect Cost Component for AC Motor Design

Indirect cost component also is a labor intensive process. It is mostly considered the cost associated with design. Therefore materials could be considered neglectable. With respective to time/count, there are two root cost drivers: the number of initialization for design (which is same as setup while a machine starts, or like fix cost), administration labor time for design. Correspondingly, environment variables compose of two drivers: initialization unit cost and administration labor cost per unit time for design. Economic factors are the current price index.

According to Figure 4-11, after decomposing the direct cost component of manufacturing, thirteen cost components are not able to be decomposed further, whose root cost driver are listed based on the above four basic types:

1. Cost Component for End Frames

Generally, end frames are made from cast-iron, steel, zinc, or aluminum castings. Here, the AC induction motor in this study is assumed to be a 100 hp motor and cast-iron would be used for the end frames. The manufacturing processes was discussed before: a computer numerically controlled (CNC) machine prepares the bearing bore and end frame diameter, and a manual drill is used to prepare the holes for the housing attachment. Therefore, with respect to materials, the root cost driver is cast-iron per unit weight. Manufacturing processes involve machine time, the number of setups, and labor time. Correspondingly, environment variables compose of two drivers: unit cost for cast-iron, labor cost per unit time, machine cost per unit and setup cost per unit. Economic factors could include the current price index.

2. Cost Component for Miscellaneous Parts

For simplicity, two miscellaneous parts, bearings and terminators, are considered. The bearings are considered as purchased parts. Terminators with lead wire also do not involve production. From the perspective of material, the number of bearings with particular specification, the length of lead wire, the number of studs and screws are root cost drivers. No root cost drivers associated with time/count are in this component. Environment variables include the unit cost for the bearing with particular specification, the unit cost for lead wire, and the unit cost for studs and screws. Economic factors are same as above.

3. Cost Component for Final Assembly

The final assembly of the AC motor includes assembling, testing, painting and packing processes. The weight of painting material and its unit cost are root cost drivers. The costs incurred in this component are time based. Labor time for final assembly and its labor cost per unit time are very important root cost drivers here.

4. Cost Component for Housing

Housings come in cast-iron; in rolled, wrapped, and tube steel; or in both cast and extruded tube aluminum. Since the AC induction motor in this study is assumed to be 100 hp, cast-iron is used for housing. The manufacturing operations are completed on either manual machines or a computer numerically controlled (CNC) machine. Therefore, with respect to materials, one root cost driver is the weight of cast-iron. Manufacturing processes involve machine time, the number of setups, and labor time. Correspondingly, environment variables compose of two drivers: unit cost for cast-iron, labor cost per unit, machine cost per unit and setup cost per unit. Economic factors are the current price index.

5. Cost Component for Stator Lamination Stack

The stator lamination stack is made of stator lamination via welding, bonding, or cleating. Stator laminators are made of specific types of steel sheets via cutting, pressing, and annealing. With respect to materials, the weight of particular steel sheets is a root cost drivers (other associated materials are neglectable for simplicity).

Labor time is needed to weld, bond, or cleat for stacking laminations. Environment variables include unit cost for particular steel sheets, labor cost unit time. Economic factors are same as before.

6. Cost Component for Magnet Wire

The component only is related to material cost. The length or gauge of the magnet wire and unit length of magnet wire is its root cost drivers.

7. Cost Component for Winding, Insulation, and Lacing

Some winding and lacing operations are completed on machines. The weight of insulation material (epoxy powder) and its unit cost are root cost drivers. Varnish for bonding stator also is an important material in this part. The weight of varnish and its unit cost are root cost drivers. From the perspective of time, labor time, machine time, the number of machine setups are major causes of cost. Environment variables are labor cost per unit time, machine cost per unit time and setup unit cost.

8. Cost Component for Stator Assembly

The housing needs to be cleaned first. The wound stator is pressed into the housing by a hydraulic machine. The cleaning materials are root cost drivers during this process. They include zinc phosphate/caustic sludge, spent solvent and acetone-wetted cloth. The cost is also incurred by time. Labor time and machine time for stator assembly, the weight of cleaning material, the corresponding cleaning material unit cost, labor cost per unit time and the corresponding machine cost per unit time are very important root cost drivers here.

9. Cost Component for Shaft

The shaft is made of cold-rolled steel. The manufacturing operations are completed on special CNC Swiss turning machines. Therefore, with respect to materials, the root cost driver is the weight of rolled steel. The root cost drivers associated with time/count and include machine time, the number of setups, and labor time. Correspondingly, environment variables compose of following drivers: unit cost

for rolled steel, labor cost per unit, machine cost per unit and setup cost per unit. Economic factors are the current price index.

10. Cost Component for Rotor Lamination Stack

The rotor lamination stack is made of rotor laminations via welding, bonding, or cleating. Rotor laminations are made of specific types of steel sheets via cutting, pressing, annealing. With respect to materials, the weight of particular steel sheets is a root cost drivers (other associated materials are neglectable for simplicity). Labor time is needed to weld, bond, or cleat for stacking laminations. Environment variables include unit cost for particular steel sheets, labor cost unit time. Economic factors are same as before.

11. Cost Component for Bars and Rings

The cost component of bars and rings are only related to material cost. The root cost drivers include the weight of aluminum (alloy) with specific specification for connection bars and the weight of steel with specific specification for retaining rings. Their corresponding environment variables are unit cost of aluminum and unit cost of steel.

12. Cost Component for Rotor Core Assembly

The material cost associated with rotor core assembly could be neglectable. Assembling rotor bars, connecting rings, retaining rings, and rotor lamination stack together incurs labor cost. The root cost drivers include labor time and labor cost per unit time.

13. Cost Component for Rotor Assembly

The shaft is inserted into rotor core. The labor and machine cost is the only cost considered. Labor time and machine time for rotor assembly, the corresponding labor cost per unit time and the corresponding machine cost per unit time are very important root cost drivers.

The indirect cost component of manufacturing can be broken down into two following cost components, whose root cost drivers are listed based on the four basic types:

1. Cost Component for Production Overhead

   Cost component for production overhead also is often based on labor. It is mostly considered as the cost associated with production management. Therefore materials could be considered neglectable. With respect to the basic type of root cost driver (time/count), there are two root cost drivers: administration labor time for production. Correspondingly, environment variable is administration labor cost per unit time for production. Economic factors are the current price index.

2. Cost Component for Corporation Overhead

   Corporation overhead also is a labor intensive process. It is mostly considered as the cost associated with overhead in the range of corporation associated the motor production. Therefore materials could be considered neglectable. With respect to the basic root cost driver (time/count), administration labor time for motor production at the corporation level is a root cost driver. Correspondingly, environment variable is administration labor cost per unit time for production at the corporation level.

In the prediction of AC motor, the costs are influenced by their manufacturing scale (quantity) and learning factors. For simplicity, in this study, these factors are included in the economic factors. After listing all important root cost drivers for each bottom component, the final list of root cost drivers sees Table 4-1 through Table 4-4.

Table 4-1 The Cost Drivers Associated with Material

The Type of Root Cost Driver	Phases	Property	Root Cost Driver	Unit	Mark of Availability	Associated Cost Drivers
Material	Design	Direct	(Neglectable)			
		Indirect	(Neglectable)			
	Production	Direct	Cast Iron (for Housing)	Weight	Available	
			Magnet Wire	Length	Not Available	Voltage, Frequency, Temperature, Power (output), Effiency, Speed, Winding density Factor
			Integral epoxy powder	Weight	Not Available	Voltage, Temperature, Power (output), Efficiency, Stator Laminations
			Steel Sheet	Weight	Not Available	Voltage, Frequency, Temperature, Power (output), Effiency, Speed, Winding density Factor
			Cold-Rolled Steel (CRS)	Weight	Available	
			Cast Iron (for End Frame)	Weight	Available	
			Bearing	number	Not Available	Speed, operation, torque, operating conditions, vibration, temperature, shock-impact loads
			Aluminum (alloy) for Bars and Connection Ring (Rotor)	Weight	Not Available	Voltage, Frequency, Temperature, Power (output), Effiency, Speed, Winding density Factor
			Steel (alloy) for Retaining Ring	Weight	Not Available	Voltage, Frequency, Temperature, Power (output), Effiency, Speed, Winding density Factor
			Lead Wire	Length	Available	
			Studs	number	Available	
			Screws	number	Available	
			Terminals	number	Available	
			Varnish	Weight	Available	
			Adhesives	Weight	Available	
			Zinc phosphate/caustic sludge(for Cleaning)	Weight	Available	
			Spent solvent (for Cleaning)	Weight	Available	
			Acetone-wetted cloth (for Cleaning)	area	Available	
			Paint solids ((including plastic sheets, filters, and precipitated paint from the paint booth water curtain))	Weight	Available	
			Paint liquids	Weight	Available	
		Indirect	(Neglectable)			

Table 4-2 The Cost Drivers Associated with Time/Count

The Type of Root Cost Driver	Phases	Property	Root Cost Driver	Unit	Mark of Availability	Associated Cost Drivers
Time/ Count	Design	Direct	Design Labor Time	Time	Not Availabe	Complexity (Voltage, Frequency, Temperature, Power (output), Effiency, Speed, Winding density Factor, Specific requirements), Design Institute (Experience, Tools)
			Equipment Time	Time	Not Available	Design Time--- Direct
		Indirect	Administration labor time	Time	Not Available	Design Time--- Complexity
			The Number of Initialization for Design	Number	Avaiable	
	Production	Direct	CNC Machining Time (Housing)	Time	Not Available	Specific requirements (Geometry), Quantity
			The Number of CNC Setups (Housing)	Number	Available	
			Labor Time (Housing)	Time	Not Available	CNC machining time and setup times (Housing)
			CNC Machining Time (End Frame)	Time	Not Available	Specific requirements (Geometry), Quantity
			The Number of CNC Setup(End Frame)	number	Available	The number of setups
			Labor Time (End Frame)	Time	Not Available	CNC machining time and setup times (End Frames)
			Wound Stator Assembly Labor Time	Time	Not Available	Housing Material, the size, the electrical efficiency
			Laminations Machining Time	Time	Not Available	Voltage, Frequency, Temperature, Power (output), Effiency, Speed, Winding density Factor
			Laminations Labor Time	Time	Not Available	Laminations Machining Time
			Labor Time for Stator Laminations Stack	Time	Avaiable	
			Machining Time for Stator Winding, Insulation and Lacing	Time	Not Avaiable	Voltage, Frequency, Temperature, Power (output), Effiency, Speed, Winding density Factor
			Labor Time for Stator Winding, Insulation and Lacing	Time	Not Available	Voltage, Frequency, Temperature, Power (output), Effiency, Speed, Winding density Factor
			The Number of Setups for Stator	Number	Available	
			CNC Swiss Turning Machining time (Shaft)	Time	Not Available	Power, Speed, Voltage, Specific requirements (Geometry),
			The Number of CNC Swiss Turning Setups(Shaft)	number	Available	
			Labor Time for Machining Shaft	Time	Not Available	CNC Swiss Turning Machining Time
			Rotor Assembly Labor Time	Time	Not Available	Speed, operation, torque, operating conditions, vibration, temperature, shock-impact loads
			Rotor Assembly Machine Time	Time	Not Available	Speed, operation, torque, operating conditions, vibration, temperature, shock-impact loads
			Varnish Impregnation Process Time	Time	Available	
			Assembly Labor Time	Time	Not Available	Complexity (Voltage, Frequency, Temperature, Power (output), Effiency, Speed, , Specific requirements),
		Indirect	Administration Labor Time (Production)	Time	Not Available	Labor Time, Machine Time and Materials in Production
			Administration Labor Time (Corporation)	Time	Not Available	Production Scale, Complexity, Manufacturer Efficiency

Table 4-3 The Cost Drivers Associated with Environment

The Type of Root Cost Driver	Property	Root Cost Driver	Mark of Availability
Environments	Design	Design Labor Cost per Unit Time (Direct)	Unvailable*
		Labor Cost per Unit Time (Indirect)	Available
		Equipment Cost per Unit Time	Available
		Administration Labor Cost Per Unit Time	Available
		Initializing Unit Cost (Design)	Available
	Production	Cost per Pound (cast iron) for Housing	Available
		Cost per Pound (cast iron) for End Frame	Available
		Cost per unit length (Magnet Wire)	Available
		Cost per Pound (Integral epoxy powder)	Available
		Cost per Pound (Steel Sheet) for Laminations	Available
		Cost per Pound (Steel) for Retaining Rings	Available
		Cost per Pound (Aluminum) for Rotor Bars and Connection Rings	Available
		Cost per Pound (CRS) for Shafts	Available
		Cost per unit bearing (Particular specification)	Available
		Cost per unit Length (Leading Wire)	Available
		Unit Cost for studs (Particular specification)	Available
		Unit Cost for screw (Particular specification)	Available
		Unit Cost for terminates (Particular specification)	Available
		Cost per Pound (Varnish)	Available
		Cost per Pound (Adhesive)	Available
		Cost per Pound (Zinc phosphate)	Available
		Cost per Pound (Spent solvent)	Available
		Cost per Square Feet (Acetone-wetted cloth)	Available
		Cost per Pound (Paint liquids)	Available
		Labor Cost per Unit Time (CNC machine)	Available
		Machine Cost per Unit Time (CNC machine)	Available
		Cost per Setup (CNC machine)	Available
		Labor Cost per Unit Time (Wound)	Available
		Cost per Unit Time (Machine for Wound, Lacing)	Available
		Cost per Setup (Machine for Wound, Lacing)	Available
		Cost per Unit Time (CNC Swiss Turning)	Available
		Cost per Unit Time (Machine for Rotor Assembly)	Available
		Cost per Setup (CNC Swiss Turning)	Available
		Labor Cost per Unit Time(CNC Swiss Turning)	Available
		Labor Cost per Unit Time (General)	Available
		Labor Cost per Unit Time(Production)	Available
		Labor Cost per Unit Time (Corporation)	Available

Note:

* Its associated cost drivers are labor unit cost, design labor unit cost for other product, and labor unit cost for other product.

Table 4-4 The Cost Drivers Associated with Economic Factors

The Type of Root Cost Driver	**Root Cost Driver**	**Mark of Availability**
Economic Factors	Price Index	Available
	Quantity	Available
	Learning Curve Factor	Available

4.3.3.4 Analysis of Availableness and Acceptableness

More and more cost information is provided with the evolving of a design. During the early design phases, much cost information is not available. While cost is estimated at this phase, unavailable information associated with costs cannot be used as cost drivers to estimate cost. Therefore, after listing root cost drivers, the availability of root cost drivers must first be checked. Secondly, when cost information is available but very expensive to obtain compared to the cost of the motor, this type of cost information also cannot be practically used. Lack of these appropriate cost information could lead to cost estimation inaccuracy. When cost drivers are unavailable or unacceptable, associated cost drivers can be substituted for the missing root cost drivers to improve the accuracy.

This case study examines the cost of an AC motor in the early design phases. The performance parameters and customer requirements of the AC motor are known. Some materials, general design processes and manufacturing processes for the particular AC motor also are known. But special and customized requirements will result in change of some of the processes. For example, the machining time and labor hours for some processes cannot be known before completing detailed design. Additionally, some types of material specification cannot be determined until the detailed design phase. The above mentioned information is important to cost estimation but they are not available early in the design process. They must be marked for finding a substitution. Another type of root cost drivers must also be marked for substitution because it is available but very expensive to obtain. For instance, some manufacturing process information can be purchased from other organization at a very high price. This study only considers availability of root cost drivers.

After analyzing, the marked root cost drivers are as indicated in Table 4-1 through Table 4-4.

4.3.3.5 Substitution (Associated Cost Drivers)

As mentioned in the last section, there are unavailable and unacceptable root cost drivers, some of which significantly impact the cost estimate of AC motor. Associated cost drivers must be found at current design phase by making assumptions and

preconditions. As the design evolves, the root cost driver known at that future phase would be employed instead of their substitutions (associated cost drivers). Cost estimation would thus be much more accurate.

All marked in last analysis step (see Table 4-1 through Table 4-4) would be found associated cost drivers. Generally, there are two types of substitution: direct and indirect. A direct substitution means that a root cost driver has a direct relationship with associated cost driver(s). The root cost driver can be derived from the associated cost driver(s) under some conditions. For instance, the length of magnet wire cannot be determined in the early design phases. But the length of magnet wire is related to voltage, frequency, temperature, power (output), efficiency, winding density factor. Using empirical formula or physics properties, the length of magnet wire can be obtained (Figure 4-13).

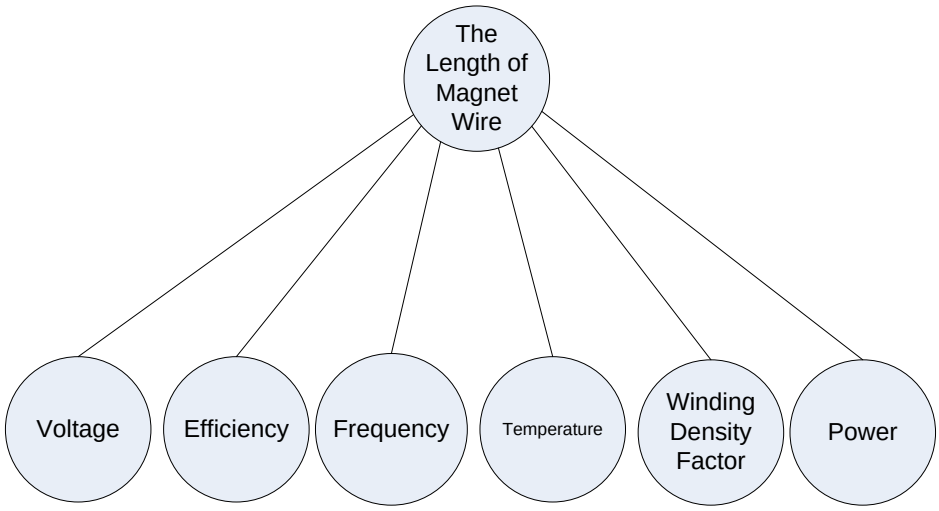


Figure 4-13 Associated Cost Drivers for the Length of Magnet Wire

Other substitutions are based on indirect relationships. For example, it is assumed that the labor unit cost is known for motor production but design labor unit cost is unknown for motor design. From experience, the labor unit cost for producing some products has stable relationship with the design labor unit cost for designing those products if both are done in the same area (city or country). Therefore, if the labor unit cost and design labor unit cost for other product are known at the same area, the design labor unit cost for the AC motor can be derived (Figure 4-14). The assumption or condition for this indirect

substitution is relationships in the same area. If this assumption or condition is not satisfied, the substitution would be not appropriate.

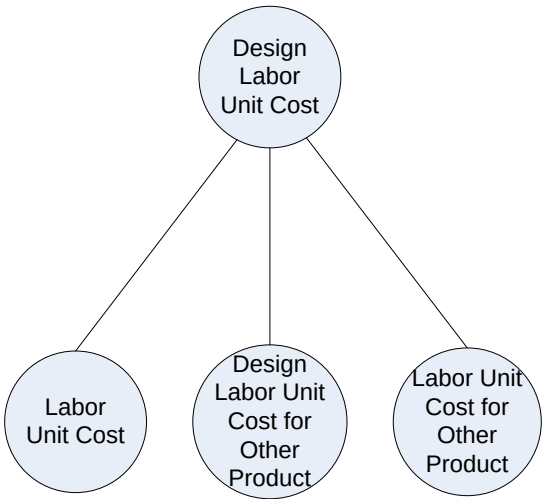


Figure 4-14 Associated Cost Drivers of the Design Labor Unit Cost for AC Motor

The associated cost drivers for all marked root cost driver are derived via those two types of relationship: direct or indirect. Please see Table 4-1 through Table 4-4.

4.3.3.6 Gathering for Future Possible Cost Drivers to Model

All available and acceptable root cost drivers and associated cost drivers are gathered as Table 4-5. They are used for cost model to estimate the cost.

4.3.3.7 Effect of Causal-Associated Method

A Causal-Associated method helps consider all factors and avoid missing some cost drivers. This can help reduce the bias and improve the degree of estimating accuracy. When using associated cost drivers to represent some root cost drivers, the assumptions and preconditions will be easily obtained.

This Causal-Associated method is different with the traditional method. The traditional methods identify potential cost drivers from references found in the literature and experience. After these potential cost drivers are grouped and reviewed, the candidate cost drivers are then determined using statistical analysis based on the data or expert

survey. This method cannot guarantee the completeness and correctness of cost drivers. It selects cost drivers from the existing literatures and experience. And it only considers available information and cannot provide assumptions and preconditions for cost estimation as the Causal-Associated method. The CA method not only reduces the chance of missing data, but it also provides a way to analyze the assumptions and preconditions of cost estimation.

Table 4-5 Final Cost Drivers of an AC motor

Availale Root Cost Drivers		
the weight of cast iron (for Housing)	the weight of Integral epoxy powder	the weight of Cold-Rolled Steel (CRS)
the weight of Cast Iron (for End Frame)	the number of bearing	the length of leading wire
the number of studs	the number of screws	the number of terminals
the weight of varnish	the weight of coolant	the weight of Zinc phosphate
the weight of Spent solvent	the area of Acetone-wetted cloth	the weight of paint solid
the paint of liquid	the number of setups of CNC (Housing)	the number of setups of CNC (End Frame)
Stator Laminations Annealing time	Rator Laminations Annealing time	CNC Swiss turning Machining time (Shaft)
Shaft Hardening Time	Varnish Impregnation Process Time	Labor Cost per Unit time (Direct)
Labor Cost per Unit time (Support)	Machine Cost per Unit Length	Unit Cost for design (other)
Cost per pound (cast iron) for Housing	Cost per pound (cast iron) for End Frame	Cost per pound (Integral epoxy powder)
Cost per pound (Steel Sheet) for Laminations	Cost per pound (CRS) for Shafts	Cost per unit Length (Leading Wire)
Unit Cost for studs (particular specification)	Unit Cost for screw (particular specification)	Unit Cost for terminates (particular specification)
Cost per pound (varnish)	Cost per pound (adhesive)	Cost per pound (coolant)
Cost per pound (Zinc phosphate)	Cost per pound (Spent solvent)	Cost per square feet (Acetone-wetted cloth)
Cost per pound (Paint liquids)	Labor Cost per Unit Length (CNC machine)	Cost per Unit Length (CNC machine)
Cost per Setup (CNC machine)	Labor Cost per Unit Time (Wound)	Cost per Unit Time (Machine for Wound Lacing)
Unit Cost per Laminations	Cost per Unit Time (Annealing)	Cost per Unit Time (CNC Swiss Turning)
Labor Cost per Unit Time (CNC Swiss Turning)	Labor Cost per Unit Time (General)	Labor Cost per Unit Time (Support)
Price Index	Quantity	Learning Factors
Available Associated Cost Drivers		
Voltage	Frequency	Temperature
Power (output)	Effiency	Speed
Winding density Factor	Operating conditions	Torque
Vibration	Temperature	Shock-impact Loads
Design Institute (Experience Tools)	Management Efficiency of Manufacturer	Specific Requirements
Design Labor Unit Cost for Other Product	Labor Unit Cost for Other Product	

4.3.4 Selection of Cost Drivers

4.3.4.1 Description of the Problem

For simplifying the problem and illustrating the method of selection of cost drivers, the wound stator core of the AC motor was taken as the subject of the case study developed in this research. The wound stator core is assumed there are 15 available and acceptable possible cost drivers at the current stage: Horse Power (HP); Speed (SPEED); Efficiency (EFF); Quantity (QUAN); Machine Cost per Unit (MCU); Setup Cost per Unit(SCU); Labor Cost per Unit (LCU); Design Labor Cost per Unit (LCUD); the Number of Setups (STs); Design Time (DT); Material 1 Cost (MC1); Material 2 Cost (MC2); Material 3 Cost (MC3); Material 1 Weight (MW1); Material 2 Weight (MW2). Secondly, for verifying Tabu-Stepwise cost driver selection based on SVR, five noise variables are added into the motor: V1, V2, V3, V4, and V5.

Some of the data for this study come from the public resources. The others were artificially generated according to motor cost properties based on manufacturing processes.

4.3.4.2 Method for Selection of Cost Drivers for Wound Stator Core

The methods of variable selection have been extensively studied in linear models. For complex products during the early design phases, these selection methods would not have good performance if there were nonlinear relationships between cost drivers and cost or there were no knowledge about function form.

Support Vector Regression (SVR) for cost estimates is a good method to deal with the above situation. The cost model based on Tabu-SVR is presented and discussed in Chapter 5. Based on Tabu-SVR, this study proposed a hybrid approach, Tabu-Stepwise, to select cost drivers. The Tabu-Stepwise algorithm employs Tabu-SVR to find the appropriate parameters via 5-fold cross validation, and use a stepwise search along with a tabu list in the searching process to find the better subset with less calculation time. The tabu list would record a number of history steps and reduce the chance of repeated calculation. This reduces computation time to a third of its original value.

4.3.4.3 Results and Analysis

Using SAS, based on Adjusted R-Square selection, twelve cost drivers were selected from the original set of cost driver. They are listed as follows (see Table 4-6):

– HP; QUAN; MCU; SCU; LCU; LCUD; STS; DT; MC1; MC3; MW1; MW2

Using SAS, based on Cp selection, ten cost drivers were selected as Table 4-6.

– HP; SPEED; SCU; LCU; LCUD; STS; DT; MC1; MW1; MW2

Table 4-6 The Results of Adjusted R-square and Cp

	HP	Speed	EFF	Quan	MCU	SCU	LCU	LCUD	STs	DT	MC1	MC2	MC3	MW1	MW2
Adjust-R	√			√	√	√	√	√	√	√	√		√	√	√
Cp	√	√				√	√	√	√	√	√			√	√

The proposed search methodology starts the search for the best set of cost drivers using the initial set of cost drivers provided by the best result of Adjusted R-square and Cp method along with the first variable(see Appendix B). The Table 4-7 lists the partial results based on SVR, when the starting point was the result of Adjusted R-square. The Table 4-8 lists the partial results based on SVR, when the starting point was the result of Cp. A third starting point was considered where the first variable identified in both methods was considered (see Appendix B). Different starting points have different results. The best subset of cost drivers to estimate cost is the result with the smallest Mean Square Error via 5- fold Cross Validation (CV-MSE).

Table 4-7 The Partial Searching Results Based on SVR (Starting Point: the Result of Adjusted R-square)

Index*	HP	Speed	EFF	Quan	MCU	SCU	LCU	LCUD	STs	DT	MC1	MC2	MC3	MW1	MW2	γ**	C**	ε**	CV-MSE***
30713	√			√	√	√	√	√	√	√	√		√	√	√	0.052811	7.77E+08	567.720	1.175E+09
30712				√	√	√	√	√	√	√	√		√	√	√	0.023797	6.38E+08	1644.441	1.375E+09
30715	√	√		√	√	√	√	√	√	√	√		√	√	√	0.010689	6.22E+08	425.652	1.553E+09
30711	√	√	√		√	√	√	√	√	√	√		√	√	√	0.022399	2.51E+08	6553.866	1.491E+09
30703	√	√	√	√		√	√	√	√	√	√		√	√	√	0.021285	4.00E+08	1396.906	1.196E+09
30687	√	√	√	√	√		√	√	√	√	√		√	√	√	0.012904	9.96E+08	4678.882	1.245E+09
30655	√	√	√	√	√	√		√	√	√	√		√	√	√	0.020879	6.00E+08	446.637	1.331E+09
30591	√	√	√	√	√	√	√		√	√	√		√	√	√	0.005587	4.96E+08	260.615	1.392E+09
30463	√	√	√	√	√	√	√	√		√	√		√	√	√	0.008514	2.24E+08	3904.206	1.507E+09
30207	√	√	√	√	√	√	√	√	√		√		√	√	√	0.006466	5.89E+08	54.954	1.584E+09
29695	√	√	√	√	√	√	√	√	√	√			√	√	√	0.018772	2.74E+08	73.349	1.494E+09
30719	√	√	√	√	√	√	√	√	√	√	√		√	√	√	0.018574	5.06E+08	23.575	1.343E+09
32767	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	0.016341	7.02E+08	1397.497	1.218E+09
28671	√	√	√	√	√	√	√	√	√	√	√	√		√	√	***0.024208***	***3.35E+08***	***38.647***	***7.613E+08*** ←

Note:
* index means a unique integer representing the set of cost drivers.
** γ, C, ε are the SVR parameters. Their value were found by Tabu-SVR via 5-fold cross validation.
***CV-MSE is the performance criterion using Mean Square Error (MSE) via 5-fold Cross Validation (CV).

Table 4-8 The Partial Searching Results Based on SVR (Starting Point: the Result of Cp)

Index*	HP	Speed	EFF	Quan	MCU	SCU	LCU	LCUD	STs	DT	MC1	MC2	MC3	MW1	MW2	γ**	C**	ε**	CV-MSE***
26595	√	√				√	√	√	√	√	√			√	√	***0.044072***	***9.78E+08***	***716.035***	***5.800E+08*** ←
26594		√				√	√	√	√	√	√			√	√	0.030207	6.63E+08	270.934	6.935E+08
26593	√					√	√	√	√	√	√			√	√	0.052748	9.57E+08	98.415	7.566E+08
26599	√	√	√			√	√	√	√	√	√			√	√	0.029658	9.38E+08	42.097	7.639E+08
26607	√	√	√	√		√	√	√	√	√	√			√	√	0.050608	9.61E+08	250.935	7.325E+08
26591	√	√	√	√	√		√	√	√	√	√			√	√	0.011578	7.04E+08	1562.584	8.546E+08
26559	√	√	√	√	√	√		√	√	√	√			√	√	0.011213	7.60E+08	93.500	8.645E+08
26495	√	√	√	√	√	√	√		√	√	√			√	√	0.012671	8.03E+08	304.639	9.839E+08
26367	√	√	√	√	√	√	√	√		√	√			√	√	0.030614	9.16E+08	22.182	9.122E+08
26111	√	√	√	√	√	√	√	√	√		√			√	√	0.038852	8.57E+08	107.224	1.046E+09
25599	√	√	√	√	√	√	√	√	√	√				√	√	0.036888	5.88E+08	154.114	6.522E+08
26623	√	√	√	√	√	√	√	√	√	√	√			√	√	0.024645	5.96E+08	420.363	7.360E+08
32767	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	0.016341	7.02E+08	1397.497	1.218E+09
28671	√	√	√	√	√	√	√	√	√	√	√	√		√	√	0.014945	5.97E+08	1026.168	7.303E+08

Note:
* index; ** γ, C, ε; ***CV-MSE have same meanings as Table 4-7.

As the results indicated in Table 4-9, the partial searching results based on SVR are better than above results in this case.

Table 4-9 The Partial Searching Results Based on SVR (Starting point: First Variable)

Index*	HP	Speed	EFF	Quan	MCU	SCU	LCU	LCUD	STs	DT	MC1	MC2	MC3	MW1	MW2	γ**	C**	ε**	CV-MSE***
1	√															4.3879513	7.06E+08	10132.145	2.003E+09
7	√	√	√													***4.7479872***	***3.39E+07***	***2459.6004***	***2.507E+08***

Note:
* index; ** γ, C, ε; ***CV-MSE have same meanings as Table 4-7.

The final cost driver subset is:

- Horse Power (HP); Speed (SPEED); Efficiency (EFF);

The eliminated cost drivers:

- Quantity (QUAN); Machine Cost per Unit (MCU); Material 2 Cost (MC2); Material 3 Cost (MC3); Setup Cost per Unit(SCU); Labor Cost per Unit (LCU); Design Labor Cost per Unit (LCUD); the Number of Setups (STs); Design Time (DT); Material 1 Cost (MC1); Material 1 Weight (MW1); Material 2 Weight (MW2);

This subset has the smallest MSE, which means it has the smallest predicting error.

For verifying the Tabu-Stepwise selection, five variables were added into the subset to create noise. The same three starting points were used: the result of Adjusted R-square; the result of Cp; and the first variable (see Appendix B). The result of Adjusted R-square introduced one noise variable (V5). The Tabu-Stepwise did not eliminate it. However, the result set of cost drivers after the search is better over the original set as shown in Table 4-10. The Cp method is better as it does not select any noise variables. Tabu-stepwise reduces the CV-MSE results via reselecting the cost drivers (Table 4-11). In Table 4-12, the best result is still same as the no-noise variables discussed before. As before, starting with the only one variable eliminates all noise variables and gets the best subset of cost drivers with smallest CV-MSE.

Therefore, the best subset including five noise variables is:

- Horse Power (HP); Speed (SPEED); Efficiency (EFF);

Table 4-10 The Partial Searching Results Based on SVR (Starting Point: the Result of Adjusted R-square with 5 Noise Variables)

Index*	HP	Speed	EFF	Quan	MCU	SCU	LCU	LCUD	STs	DT	MC1	MC2	MC3	MW1	MW2	V1	V2	V3	V4	V5	γ**	C**	ε**	CV-MSE***
554985	√			√		√	√	√	√	√	√		√	√	√					√	0.018856	6.25E+08	140.159	1.428E+09
554984				√		√	√	√	√	√	√		√	√	√					√	0.039944	5.88E+08	113.409	1.241E+09
554983	√	√	√			√	√	√	√	√	√		√	√	√					√	0.026018	5.74E+08	523.340	1.314E+09
554975	√	√	√	√	√		√	√	√	√	√		√	√	√					√	0.014438	6.49E+08	4203.534	1.224E+09
552959	√	√	√	√	√	√	√	√	√	√	√	√		√	√					√	***0.019056***	***4.29E+08***	***617.508***	***1.178E+09*** ←

Note:
* index; ** γ, C, ε; ***CV-MSE have same meanings as Table 4-7.

Table 4-11 The Partial Searching Results Based on SVR (Starting Point: the Result of Cp with 5 Noise Variables)

Index*	HP	Speed	EFF	Quan	MCU	SCU	LCU	LCUD	STs	DT	MC1	MC2	MC3	MW1	MW2	V1	V2	V3	V4	V5	γ**	C**	ε**	CV-MSE***
26521	√			√	√			√	√	√	√			√	√						0.036593	7.48E+08	233.552	6.619E+08
26520				√	√			√	√	√	√			√	√						0.033536	9.91E+08	1823.965	6.419E+08
26511	√	√	√	√				√	√	√	√			√	√						***0.021391***	***7.04E+08***	***766.690***	***6.379E+08*** ←
25599	√	√	√	√	√	√	√	√	√	√				√	√						0.039823	6.37E+08	558.234	7.161E+08

Note:
* index; ** γ, C, ε; ***CV-MSE have same meanings as Table 4-7.

Table 4-12 The Partial Searching Results Based on SVR (Starting Point: First Variable with 5 Noise Variables)

Index*	HP	Speed	EFF	Quan	MCU	SCU	LCU	LCUD	STs	DT	MC1	MC2	MC3	MW1	MW2	V1	V2	V3	V4	V5	γ**	C**	ε**	CV-MSE***
1	√																				4.387951	7.06E+08	10132.145	2.003E+09
7	√	√	√																		***4.747987***	***3.39E+07***	***2459.600***	***2.507E+08*** ←

Note:
* index; ** γ, C, ε; ***CV-MSE have same meanings as Table 4-7.

4.3.4.4 Effects of Selection of Cost Drivers

Tabu-stepwise selecting method based on Tabu-SVR improves the accuracy of the cost estimating prediction by eliminating irrelevant variables, and it reduces expenditure of the collection, storage, and computation load in the process of cost estimation. Also the Tabu-Stepwise selection method based on Tabu-SVR can identify the nonlinear

correlations among cost drivers and find the nonlinear relationships between cost drivers and cost. It makes selection of cost drivers feasible and effective under nonlinear conditions.

4.3.5 Summary of Case Study

This case study illustrates the feasibility and the procedure of Causal-Associated method for identifying the cost drivers. From this case study, it is seen that the Causal-Associated method helps reduce the chance of missing some cost drivers. This can reduce the bias and improve the degree of estimating accuracy. When using associated cost drivers to represent some root cost drivers, the assumptions and preconditions can be identified. The Tabu-Stepwise selecting method, based on Tabu-SVR, was used to select the cost drivers of wound stator core. The test data shows it improves the accuracy of the cost estimating prediction by eliminating irrelevant variables, and it reduces expenditure of the collection, storage, and computation load in the process of cost estimation.

Chapter 5 Cost Estimating Nonparametric Approach Based on Support Vector Regression

Support Vector Machine (SVM) is widely used in signal processing, pattern analysis, data classification, facial expression classification, text analysis. Additionally it has been applied to several financial applications. No literature on the application of SVM for cost modeling was found. This research focuses on the application of SVR for cost estimation in the early design phases of complex products and comparison among SVR and other traditional cost estimating methods.

The objective of this chapter is to present the cost estimating nonparametric approach based on support vector regression (SVR) and to test the applicability of support vector regression for cost estimation during design stage of a product. The performance of SVR is compared with conventional methods: linear regression, neural networks, case-based reasoning. This chapter firstly presents how to apply support vector regression in cost estimating area. Next how the test data (data sets) are created under the simulated and pilot scenarios is explained. Thirdly the application of support vector regression in cost estimation is investigated to see the influence of the selection of appropriate parameters and the choice of kernel function. Finally, the performance of Tabu-SVR, when kernel respectively is the polynomial and radial basis function, is compared with those of other traditional cost estimating methods: regression, case-based reasoning, and neural network.

5.1 The Cost Estimating Nonparametric Approach Based on SVR

The nonparametric approach based on SVR estimates cost using Equation (5-1). The nonlinear input space is mapped to the linear feature space with high dimensions by $\psi(x)$. Vapnik [75] proposed the structure risk minimization principle to minimize the upper bound of the generalization error. The coefficient *w* and *b* of Equation (5-1) are estimated by minimizing the regularized risk function under the constraints (see Problem 5-2). The Primal Problem (5-2) can be transformed to its Dual Problem (5-3) with its constraints. The Dual Problem (5-3) is a quadratic optimization problem, which has a global optimal solution. After solving this quadratic optimization problem, the final nonparametric

function can be expressed by Equation (5-4). The α_i and α_i^* are the solutions of optimization problem. Parameter *b* is a by-product of the optimization process [76]. The kernel function is defined as $k(x,x_i)=\langle\psi(x),\psi(x_i)\rangle$. For additional information on SVR, see Section 2.5.2.

$$f(x)=\langle w,\psi(x)\rangle+b \tag{5-1}$$

$$\begin{aligned}&\min \quad \frac{1}{2}w^T w+C\sum_{i=1}^{l}\left(\xi_i+\xi_i^*\right)\\ &S.t. \quad \begin{cases} y_i-w^T\psi(x_i)-b\le\varepsilon+\zeta_i \\ -(y_i-w^T\psi(x_i)-b)\le\varepsilon+\zeta_i^* \\ \xi_i,\xi_i^*\ge 0\end{cases}\end{aligned} \tag{5-2}$$

$$\begin{aligned}&\max \quad -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k(x_i,x_j)-\varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*)\\ &S.t. \quad \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \quad and \quad \alpha_i,\alpha_i^*\in[0,C]\end{aligned} \tag{5-3}$$

$$f(x)=\sum_{i=1}^{N}\left(\alpha_i-\alpha_i^*\right)\cdot k(x_i,x)+b \tag{5-4}$$

The root mean square error (RMSE) of prediction and accuracy degree are commonly used as generalization criterion between measured and predicted values. Also they can be used as performance criterion of the models. The RMSE of prediction value is defined by Equation (5-5).

$$RMSE=\sqrt{\frac{\sum_{i=1}^{l}(\hat{y}_i-y_i)^2}{l}} \tag{5-5}$$

where $\hat{y}_i$ and y_i denote the predicted value and the measured value and *l* is the number of points.

And accuracy degree is equal to (1 - Mean Absolute Relative Error (MARE)), where MARE can be defined as (5-6):

$$MARE=\frac{1}{l}\sum_{i=1}^{l}\frac{|\hat{y}_i-y_i|}{y_i} \tag{5-6}$$

For a cost estimating nonparametric approach based on SVR, there are four steps to get the final cost:

1. Data Preprocessing (Section 5.1.1)
2. Choosing the kernel and parameters (Section 5.1.2)
3. Training the SVR (Section 5.1.3)
4. Computing the final cost using the SVR model and the input (Section 5.1.3).

These steps are discussed in detail in the next three sections.

5.1.1 Data Preprocessing

The original data are scaled into the range of (0, 1). The goal of linear scaling is to independently normalize each cost drivers to the specified range. It avoids the larger value input variables dominate smaller values inputs and avoids numerical difficulties during the calculation. This hence reduces prediction errors.

The whole data set is divided into two parts: a training data set and a test data set. The training data is firstly used to choose a kernel and parameters via cross-validation (the idea is to split the data into two parts, to train on one part and then to test the accuracy of the predictor on the rest of data). The training data can then be used to determine α_i ; α_i^* and b in Equation (5-4) via solving the optimization Problem (5-3). The test data is for the verification purposes.

5.1.2 Choosing Kernel and Corresponding Parameters via Tabu-Search

In the SVR algorithm, the kernel function, its parameters, and two SVR training parameters (C, and ε) for ε-insensitive loss function play a key role in the SVR performance. Parameter C is the trade off between model complexity (flatness) and the degree of deviations allowed in the optimization formulation. Parameter ε controls the width of the ε-insensitive zone, which affects the number of support vectors used to construct the regression function. The kernel function represents the mapping instrument that is necessary to transform the non-linear input space to a high-dimensional feature space where linear regression is possible. The mapping depends on the intrinsic

topological structure of the data and application-domain knowledge. It implies that the kernel type and all parameters need be optimally chosen to get the best performance. However, there are no structural methods for determining efficiently the selection of kernel and all parameters. Moreover, because C and ε-values affect the model in a different way and kernel parameters and C are dependent, the kernel function and all parameters cannot be chosen separately. Three different kernel functions are often found in the literature associated mapping process. Therefore, the following method (Figure 5-1) is used to choose the appropriate kernel and all parameters to get the solution.

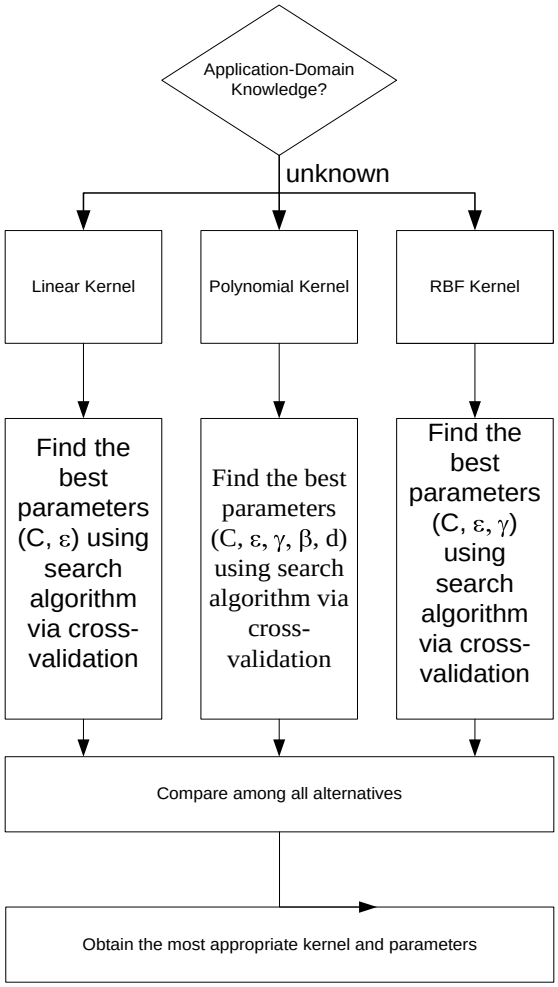


Figure 5-1 The Method to Choose Kernel and Parameters

If the intrinsic topological structure of the data and application-domain knowledge is known, the particular kernel would be chosen. The parameters are determined based on a

tabu search algorithm. The performance obtained by the cross-validation procedure is criterion to choose the parameters.

Otherwise, there are generally three types of kernel (linear kernel, polynomial kernel, and radial basis function (RBF) kernel). The better parameters of each kernel can be chosen through the cross-validation procedure. Then after comparison with performance of all kinds of kernel with the chosen parameters, the most appropriate kernel and corresponding parameters are obtained. This study focused on which kernel in these three kernel functions had better performance in the cost estimating area.

The following procedure is for choosing parameters based on RBF kernel. The procedure for other kernel functions is same.

For RBF kernel, γ, C, ε play an important role on the generalization performance of SVR. The proposed search algorithm combines the empirical study [83] and a tabu search algorithm. The starting point is based on the training data. Cherkassky and Ma [83] proposed that C values should be based on the training data without resulting in re-sampling using the following estimation:

$$C = \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|) \tag{5-7}$$

where $\bar{y}$ and σ_y are the mean and standard deviation of the y values of the training data. Using the idea of Central Limit Theorem, they proposed that ε be given by:

$$\varepsilon = 3p \times [\ln l \times l^{-1}]^{-\frac{1}{2}} \tag{5-8}$$

where p is the standard deviation of the input noise and l is the number of training samples. Since the value of p is not known beforehand, the following equation can be use to estimate p using the idea of k-nearest-neighbor method:

$$\hat{p} = \sqrt{\frac{l^{1/5}k}{l^{1/5}k - 1} \frac{1}{l} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}\ ,\ 2 \le k \le 6 \tag{5-9}$$

where l is the number of training samples, k is the low-bias/high variance estimators, and $\hat{y}$ is the predicted value of y by fitting a linear regression to the training data to estimate the noise variance.

And γ=0.5∗1/σ^2 is set to σ^d ~ (0.1-0.5) where all d cost drivers are pre-scaled to [0, 1]. The degree of dimensions is given by d.

The kernel function and all parameters would be put into Problem (5-3). The Equation (5-4) is used to predict the value of y_i corresponding to the sample x_i. The predictive capability of a set of given parameters (C_1, ε_1, γ_1) (the starting point) is evaluated using the RMSE defined as Equation (5-5).

Tabu search is a memory-based stochastic optimization algorithm [74, 88, 89], modified for SVR as follows:

Step 1: Define the starting point (C_1,ε_1,γ_1) based on empirical study (Equations 5-7, 5-8, 5-9) as current point; initialize the best point (C^*,ε^*,γ^*), the tabu list, the ranges of C, ε and γ; setup the number of neighbor points, the parameters for aspiration criterion, diversification, intensification, and termination criterion.

Step 2: Generate set of neighbor points based on current point according to neighbor generating policy; compute RMSE of the neighbor points via 10-fold cross validation.

Step 3: Choose the best neighbor (C', ε', γ'), which is not in the tabu list or does not satisfied aspiration criterion.

Step 4: Replace the best point (C^*,ε^*,γ^*) with (C', ε', γ') if RMSE of (C', ε', γ') is less than RMSE of (C^*,ε^*,γ^*).

Step 5: Go to Step 2, if termination criterion is not satisfied, define (C', ε', γ') as current point and update tabu list; or go to Step 6.

Step 6: Terminate and output the best point (C^*,ε^* ,γ^*).

For the tabu search algorithm, the procedure of tabu search is almost the same. However the policy of defining neighborhood, the structure of tabu list, the strategies of intensification and diversification, and the criteria of aspiration and termination determine the performance of tabu search. These six elements of tabu search are defined as follows:

- Neighborhood Definition:

  The neighborhood structure is very important to the efficiency of TABU SEARCH algorithm. Three sets of neighborhood points (see Figure 5-2) are

separately produced according to corresponding generating policy. This policy can help explore most of the space and finally converges to the global optimum.

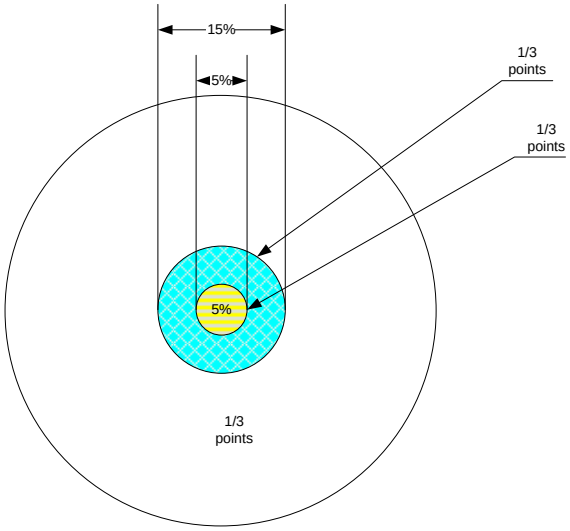


Figure 5-2 The Neighborhood Definition

First Neighborhood:

$$Y = X + R \cdot 0.025 \cdot h \cdot \alpha \cdot \beta / \eta \tag{5-10}$$

where *X* is the current value, *R* is a random parameter out of the interval of [-1, 1], *h* is the range of current variable, α is defined as:

$$\alpha = 0.5(1 + \sin(\frac{i \cdot \theta \cdot \pi}{N/3})) \tag{5-11}$$

where *i* is the index of the neighbor points, *N* is the total number of neighbor points generated at each iteration; θ is used to control the oscillation period of α [88]. α makes the generated randomized first-neighbors much closer center on the current solution. The sine function explores the closer promising area better. β is defined as:

$$\beta = (10^{-\frac{n}{M}})^{iter} \tag{5-12}$$

where M is the total number of iterations, n measures the complexity of the optimization problem, which is set between 1 and 2, *iter* is the current number of iterations. β is shrink factor, which shrinks as the number of iteration increases. Another shrink variable is defined as (5-13):

$$\eta = 10^{\phi} \tag{5-13}$$

where ϕ is a frequency index. If the search visits one area frequently, the possibility that this area would include a promising solution would be much higher than other areas. Using η can more accurately explore this area. ϕ, the frequency index is associated with the hit frequency.

The purpose of three parameters α, β, η is to implement intensification strategies via generating continuous neighborhood points and to provide more precise final solutions.

Second Neighborhood:

$$Y = X + R \cdot 0.075 \cdot h \cdot \alpha \cdot \beta / \eta \tag{5-14}$$

R, h, α, β, η, are same as previous definition.

Third Neighborhood:

$$Y = Current\_LowerBound + R \cdot h \tag{5-15}$$

R is a random parameter out of the interval of [0, 1], h is the range of current variable.

- Tabu List:

A tabu list illustrated in Figure 5-3 stores solutions that have recently been selected, which are used to escape from being recycled. The tabu list is based on two ideas: recency-based and frequency-based. The recently-based tabu list stores the most recently visited points and prohibits revisiting unpromising points for a specified number of iterations. The frequency-based tabu list is used to record history and hit frequency of the search area.

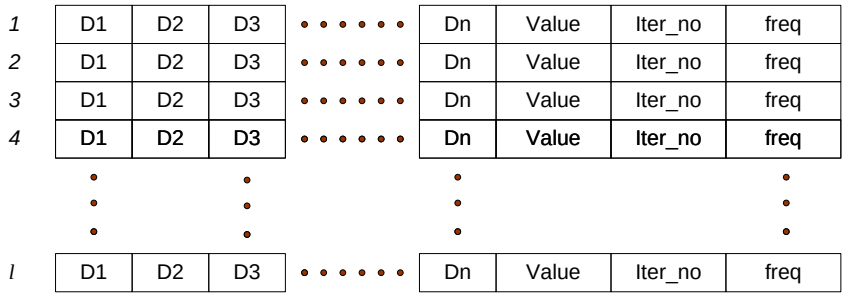


Figure 5-3 The Structure of Tabu List

❖ Intensification:

Intensification strategies help tabu search explore specific areas more thoroughly. This study employs two strategies to intensify the search:

- Generating policy of three neighborhood points.
- Frequency-based shrink variable η. shrink factor β, which dynamically decreases the search space for each variable.

❖ Aspiration Criterion

For a broadly diversified search, the aspiration criterion allows tabu search to override the tabu property when the equation (5-16) satisfies certain conditions [88].

$$f(k) = \frac{1}{1+e^{-\sigma(k-k_{center}\cdot M)}} \tag{5-16}$$

where k is the current iteration number and k_{center} determines at which point, $f(k) = 0.5$. A uniform random number P ($0 \le P \le 1$), is generated at each iteration. If P is less than or equal to $f(k)$, the tabu property is overridden; or the best non-tabu point in the neighbor space is used as a new starting point.

❖ Diversification Strategies:

For ensuring all areas of the search space have been adequately explored, there are following three strategies to diversify the search:

- Escape strategy based on tabu-list frequency,
- Aspiration criterion, and
- Generating policy of three neighborhood points

❖ Termination Criterion:

The stopping conditions for the three processes are defined as follows:

- The program will stop after a given number of iterations without any improvement on the value of the objective function as (5-17):

$$\left| \frac{f_k(x) - f_{k-\Gamma}(x)}{f_{k-\Gamma}(x)} \right| < \delta \qquad (5\text{-}17)$$

where *k* is the current number of iterations, Γ is a given number of iterations, δ is a pre-defined value of the objective function.

- The search procedure will stop after a pre-defined maximum number of iterations.

5.1.3 Training the SVR and Computing the Final Cost

After obtaining the kernel and all parameters, the following (Equation 5-3) can be solved to get an optimal solutions (α_i ; α_i^* and *b*) using the training data set.

$$\max \quad -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*)$$

$$S.t. \quad \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \quad and \quad \alpha_i, \alpha_i^* \in [0, C]$$

Using these α_i, α_i^* and *b* and corresponding historical data points along with the kernel function, the cost of any inputs can be calculated via Equation 5-4 as follows:

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) \cdot k(x_i, x) + b$$

5.2 Test Cases

Cost data of a real product in a company or organization are often confidential. It is hard to find the real cost data of a complex product to verify and validate the proposed model in this study. Even though there are small data sets in literature, the amount of these small cost data are too small to test performance of the proposed cost estimating approach. Therefore, this study provided two types of test data, simulated test data and pilot test data, to verify and validate the proposed cost estimating approach. The simulated test data come from the summarized common basic cost characteristics. The pilot test data are produced by the cost module (engine part) of Flight Optimization Systems (FLOPS). Section 5.2.1 and Section 5.2.2 will discuss these types of test data further.

5.2.1 Simulated Data Sets Based on Common Basic Cost Characteristics

This section first summarizes common basic cost characteristics based on literatures. These common basic cost characteristics are then expressed by mathematical functions. Combining the mathematical functions and following the general rules, the formulas are constructed to produce test data for the future verification and validation.

5.2.1.1 Common Basic Cost Characteristics

Complex products have complex designs and complex manufacturing processes and thus cost estimation is not an easy task. The relationships between cost and its cost drivers are complex and often include nonlinear relationships and may be very hard to define in some cases. Here, common basic cost characteristics are summarized as following to form the test cases (data sets). These test cases would be used to test cost modeling techniques.

As mentioned in previous chapters, the cost can be broken down into cost components illustrated in Figure 5-4. Therefore, for complex products, the first basic characteristic of cost is accumulation. Cost of a complex product can always be expressed

as the sum of all cost components (Equation 5-18) via appropriate cost breakdown structure.

$$C = C_1 + C_2 + \cdots\cdots + Cn \tag{5-18}$$

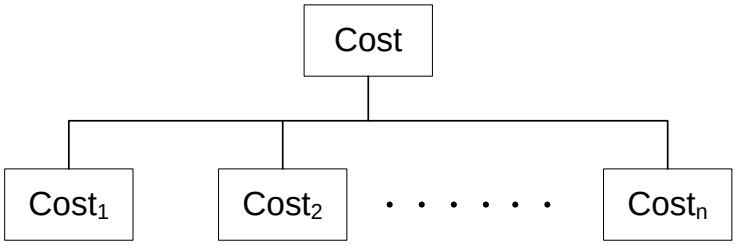


Figure 5-4 Cost Breakdown Structure

The second basic characteristic of cost can be expressed by linear function (5-19):

$$C = ax + b \tag{5-19}$$

This is very common characteristic in cost estimating area. For example, the labor cost (C) can approximately equal variable cost plus fixed cost under some condition. Here, fixed cost is, the intercept b and a can be expressed by unit cost per time, the cost driver x is the time length.

The third basic characteristic of cost can be expressed by power function (5-20):

$$C = a_1 x^{a_2} \tag{5-20}$$

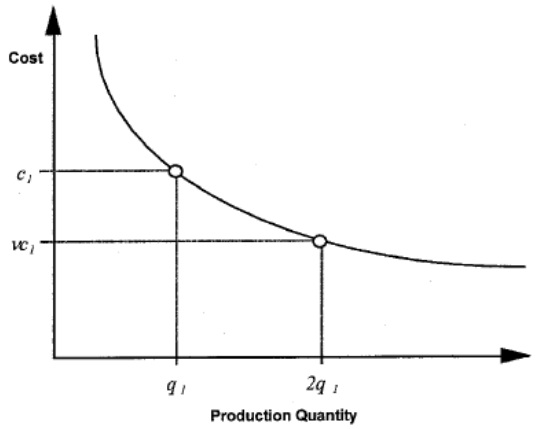


Figure 5-5 The Power Relation ([90])

The power function can normalize cost for learning curve theory or structural size. Under the assumptions of learning, the unit cost of production decreases as the quantity increase, and the rate of decrease typically decreases as the number of units increases. For example (Figure 5-5), C is cost of x^{th} unit, a_1 is the cost of the first unit produced, a_2 is a parameter measuring the rate labor hours are reduced as cumulative output increases, and quantity x is cost driver. In the area of aerospace, the weight-size cost also can be expressed as the power function (5-3). The cost driver x is the weight or size, a_1 is the cost of the theoretical first pound cost, a_2 is a parameter measuring the amount costs decrease with respect to the weight (size).

The fourth basic characteristic of cost can be expressed by step functions (5-21):

$$C = a_1 f(x) \tag{5-21}$$

where $a_i \in R$, $f(x)=1$ if $x \in [a_2, a_3)$ and 0 otherwise, $i=1,2,3$. Generally, the step function is combined with other functions to express cost. For example, the cost in semiconductor industry has typical associated cost curve in Figure 5-6. The cost driver here is one measurement of product performance. The contribution of technology breakthrough would make one manufacturing cost curve with corresponding performance shift to another curve. This situation can be expressed by a step function with other characteristics.

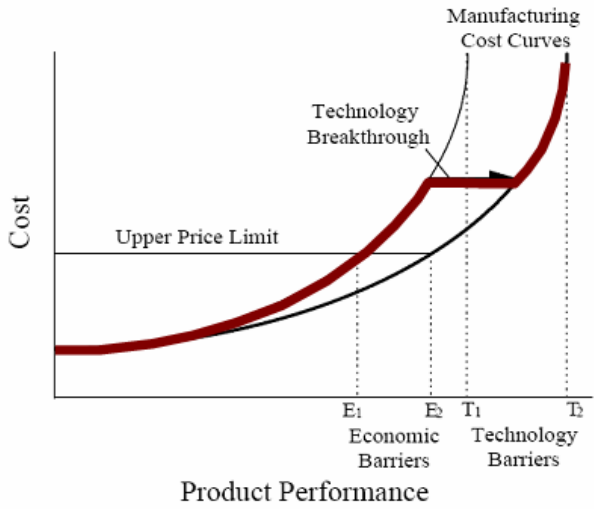


Figure 5-6 The Cost Expressed by the Combination of Step and Other Functions [91]

The fifth basic characteristic of cost is an exponential function (5-22), which does an excellent job of normalizing cost for temporal effects such as inflation and technology escalation. The exponential functions often combine with other function to express cost.

$$C = a_1 e^{a_2 x} \tag{5-22}$$

In summary, there are five above basic cost characteristics often combined to use in cost modeling: accumulation; linear function; power function; step function; and exponential function. The cost of complex product can be expressed by these five characteristics. For example, Figure 5-7 shows typical cost curve on time, which at least includes linear, power, step function. Based on these five common basic cost characteristics and general combining rules, the six simulated formula are constructed to produce test data in Section 5.2.1.2.

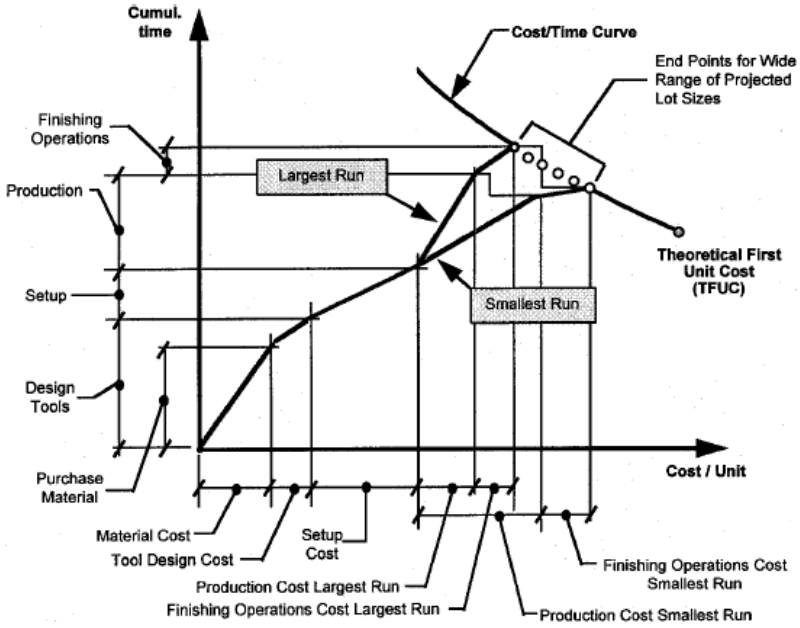


Figure 5-7 Typical Cost Curve on Time [90]

5.2.1.2 Six Formulas for Producing Simulated Test Cases (Data Sets)

Following the common cost modeling mathematical expression and based on above common basic cost characteristics, the following six formulas are constructed for producing test cases (data sets). The formulas have five to six cost drivers to represent

multiple inputs. Step function and exponential function could not express cost only by themselves.

Linear Function with five cost drivers is as (5-23):

$$C = 500 + 250x_1 + 150x_2 + 300x_3 + 100x_4 + 200x_5 \tag{5-23}$$

The range of x_i is 0 to 1. A noise component is also added in Equation (5-23) (coefficient of variation, c.v. = σ/μ=0.05).

The Linear-Step Function (5-24) with five cost drivers is as (5-24):

$$C = 500 + 250x_1 + 150x_2 + 300x_3 + 100x_4 + 200x_5 * f(x_5)$$
$$f(x_5) = \begin{cases} 7 & x_5 < 0.7 \\ 1.5 & x_5 \geq 0.7 \end{cases} \tag{5-24}$$

The range of x_i is 0 to 1. A noise component is also added in each cost drivers (coefficient of variation, *c.v.* = σ/μ=*0.03*).

The Power Function with five cost drivers is as (5-25):

$$C = 120(x_1^{0.05})(x_2^{0.5})(x_3^{5})(x_4^{2})(x_5^{-1}) \tag{5-25}$$

The range of x_i is 0 to 1. A noise component is also added in each cost drivers (coefficient of variation, *c.v.* = σ/μ=*0.03*).

The Linear-Power Function with five cost drivers is as (5-26):

$$C = 20 + 40x_1 + 10x_2 + 30x_3^{0.5} + 15(x_4^{0.8})(x_5^{1.5}) \tag{5-26}$$

The range of x_i is 0 to 1. A noise component is also added in each cost drivers (coefficient of variation, *c.v.* = σ/μ=*0.03*).

The Power-Step Function with five cost drivers is as (5-27):

$$C = 120(x_1^{0.05})(x_2^{0.5})(x_3^{5})(x_4^{0.5f(x_5)})$$
$$f(x_5) = \begin{cases} 0.8 & x_5 \geq 0.8 \\ 0.5 & 0.8 > x_5 > 0.2 \\ 0.2 & 0.2 \geq x_5 \end{cases} \tag{5-27}$$

The range of x_i is 0 to 1. A noise component is also added in each cost drivers (coefficient of variation, *c.v.* = σ/μ=*0.03*).

The Linear-Power-Step Function with six cost drivers is as (5-28):

$$C = 20 + 40x_1 + 10x_2 + 30x_3^{0.5} + 35(x_4^{0.8})(x_5^{1.5})e^{f(x_6)}$$
$$f(x_6) = \begin{cases} 0.6 & x_6 \geq 0.66 \\ 0.4 & 0.66 > x_6 > 0.33 \\ 0.2 & 0.33 \geq x_6 \end{cases} \quad (5\text{-}28)$$

The range of x_i was 0 to 1. A noise component is also added in each cost drivers (coefficient of variation, *c.v.* = σ/μ=*0.03*).

In a summary, the above formulas are listed in Table 5-1. They would produce test cases to examine the performance of the cost estimating techniques.

Table 5-1 Six Formulas to Produce Test Cases (Data Sets)

No	Type	Formula
1	Linear	$C = 500 + 250x_1 + 150x_2 + 300x_3 + 100x_4 + 200x_5$
2	Linear-Step	$C = 500 + 250x_1 + 150x_2 + 300x_3 + 100x_4 + 200x_5 * f(x_5)$ $f(x_5) = \begin{cases} 7 & x_5 < 0.7 \\ 1.5 & x_5 \geq 0.7 \end{cases}$
3	Power	$C = 120(x_1^{0.05})(x_2^{0.5})(x_3^{5})(x_4^{2})(x_5^{-1})$
4	Linear-Power	$C = 20 + 40x_1 + 10x_2 + 30x_3^{0.5} + 15(x_4^{0.8})(x_5^{1.5})$
5	Power-Step	$C = 120(x_1^{0.05})(x_2^{0.5})(x_3^{5})(x_4^{0.5 f(x_5)})$ $f(x_5) = \begin{cases} 0.8 & x_5 \geq 0.8 \\ 0.5 & 0.8 > x_5 > 0.2 \\ 0.2 & 0.2 \geq x_5 \end{cases}$
6	Linear-Power-Step-Exponential	$C = 20 + 40x_1 + 10x_2 + 30x_3^{0.5} + 35(x_4^{0.8})(x_5^{1.5})e^{f(x_6)}$ $f(x_6) = \begin{cases} 0.6 & x_6 \geq 0.66 \\ 0.4 & 0.66 > x_6 > 0.33 \\ 0.2 & 0.33 \geq x_6 \end{cases}$

5.2.2 Pilot Data Set from a Real Detailed Cost Model

Detailed cost models or some commercial parametric cost model have incorporated knowledge about cost breakdown structure (CBS) and functional relationship. These cost models can be used as a generator of data to produce pilot data. These data can be applied to test the performance of other tested cost models when these tested models do not have enough knowledge about CBS and/or functional relationship. For the tested cost model in this section, given the subset of cost drivers, it could be assumed that there is no knowledge about CBS and the functional form. The pilot data by the commercial parametric cost model are employed to choose kernel function and hyper parameters of SVR, to test the robustness of SVR, and to compare the performance between the proposal cost estimating method based on SVR and other existing methods.

An available cost model was chosen, which is associated with the cost for aircraft engines [5, 6, 92]. The model for the cost of subsonic engine research, development, test, evaluation (RDT&E) and production is a function of the maximum thrust of the engine at sea-level static conditions, weight, specific fuel consumption at sea-level static, turbine inlet temperature, and a pressure term.

Five input variables were chosen for cost drivers. They are *WTS_25*, *NENG*, *THRMAX*, *SMACH*, *QMAX*. *WTS_25* is total weight of engines. It is assumed between *10,000 lbs* and *70,000lbs*. *NENG* is the number of engines per aircraft. In this example, *NENG* is set *2* or *4*. *THRMAX* is the maximum thrust per engine ranging from *20,000 lbs* to *90,000 lbs*. *SMACH* is the maximum Mach number at best altitude and ranges between 0.7 Mach and 1 Mach. *QMAX* is the maximum dynamic pressure ranging from *200 lb/ft*^{2} to *600 lb/ft*^{2}. All data based on these five variables were randomly and uniformly produced in their range and input to the FLOPS cost module (engine part) [6] while other input variables were held constant default value (See Figure 5-8). The FLOPS cost module [6] was originally developed using FORTRAN. For this study, the FLOPS cost module was converted to EXCEL, which was then verified using the data from Johnson's dissertation [5].

Then all variables are scaled to [0, 1] using following formula (5-29) as:

The transformed value = (current value – minimum)/(maximum-minimum) (5-29)

The above linear scaling independently normalizes each input variable to the specified range [0, 1]. It would avoid larger value input variables overwhelm smaller value inputs and avoid numerical difficulties during the calculation to hence help reduce prediction errors [93].

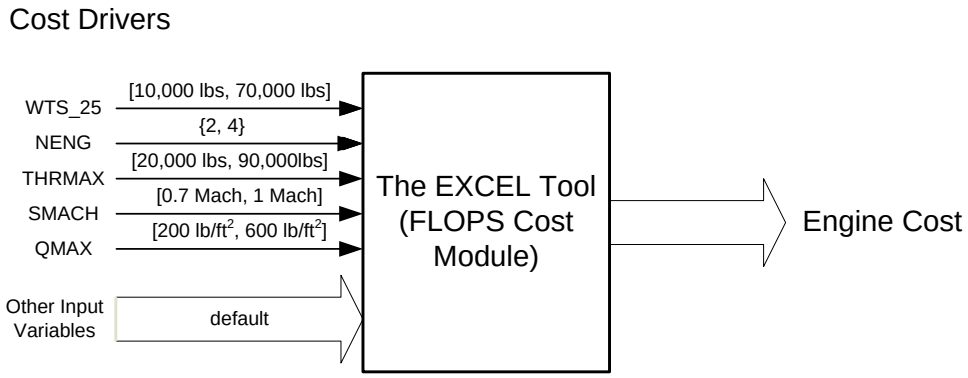


Figure 5-8 The Pilot Data Produced by an EXCEL Tool (FLOPS COST Module)

5.2.3 The Method to Produce Data Set

The six formulas in Section 5.2.1 randomly produced 120 points. The EXCEL tool developed for simulating FLOPS cost module also randomly generated 120 points in the ranges of five cost drivers. These data points were first scaled using Equation 5-29. The first 60 points were used for training data. Two groups (each 30 points) were used as test data sets. The training data set and one test data set are used in the study of choice of hyper parameters and kernels for SVR and comparison of performance. The second test data set and the training data along with first test data were used to study the robustness of SVR to the sample data.

5.3 Experiments

5.3.1 Implementation of Methods for Experiments

In experiments, it is assumed that there is no apriori knowledge about the cost breakdown structure and the functional relationship. Support vector regression, parametric method, neural network and case-based reasoning are respectively used to

build cost models and make a comparison. The root mean square error (RMSE, see Equation 5-5) and accuracy degree (1-MARE, see Equation 5-6) are the performance criterion of the models.

5.3.1.1 Support Vector Regression and Support Vector Regression with Tabu Search Algorithm

For support vector regression, the details are presented in previous chapter. The parameters of SVR were calculated according to empirical study [83]. This method of choosing parameters is called SVR for comparison purposes. The parameters were obtained via tabu search algorithm developed in C++. This method of choosing parameters is called Tabu-SVR. The experiments of support vector regression were performed through the program developed by the author using C++ and CPLEX (Figure 5-9, also see Appendix A).

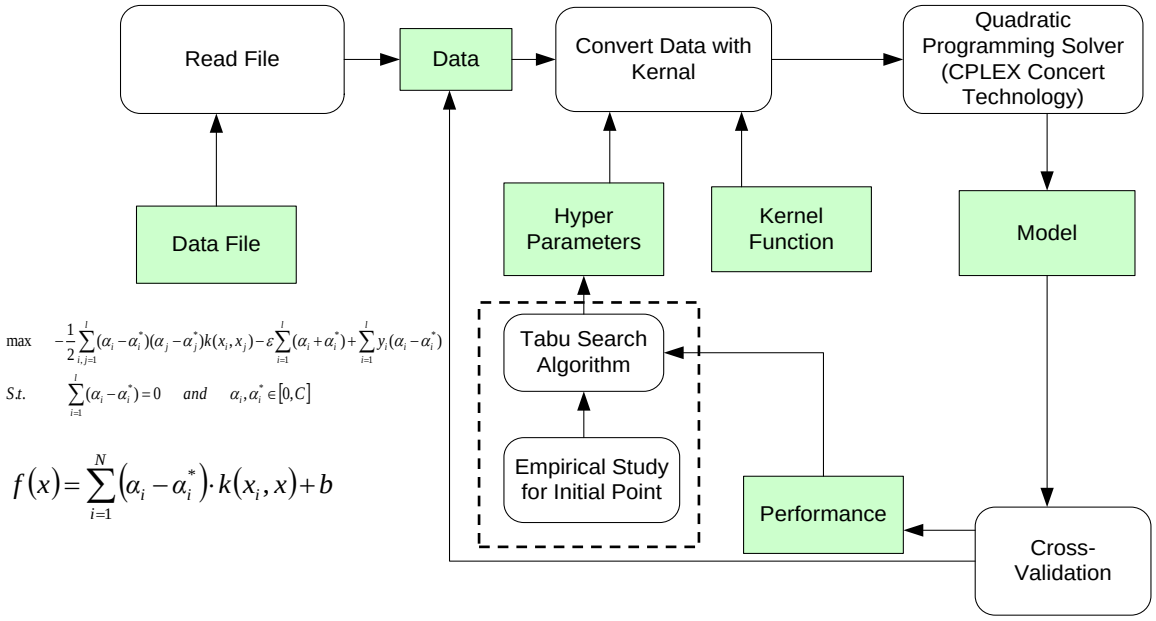


Figure 5-9 The Software Framework of Tabu-SVR

5.3.1.2 Parametric Method (Linear and Log-Linear)

In applying the parametric method, the following cost estimating relationship (CER)(5-30) of these five independent variables is used to estimate cost,

$$C = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \qquad (5\text{-}30)$$

In the general CER given by Dean[23], another common CER can be obtained by the logarithm transformation in both sides (5-31):

$$\ln(C) = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \beta_3 \ln x_3 + \beta_4 \ln x_4 + \beta_5 \ln x_5 \qquad (5\text{-}31)$$

5.3.1.3 Neural Networks

Two types of neural network were tested. The first type of neural networks is feed-forward back propagation (NN1) (see Figure 5-10). It has two layers. Five neurons are in the first layer and one neuron is in the second layer (output layer). The transfer function in the first layer is hyperbolic tangent sigmoid. The transfer function in the second function is linear.

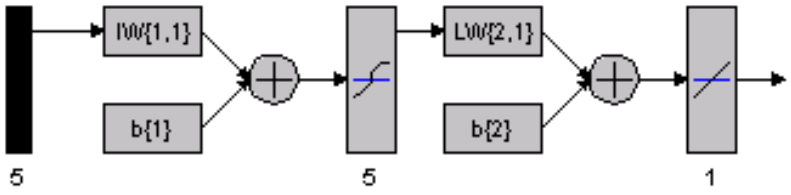


Figure 5-10 Feed-Forward Backpropagation Neural Networks (NN1)[1]

The second type of neural network (NN2) is radial basis networks which consist of two layers: a hidden radial basis layer and an output linear layer of neurons (see Figure 5-11). The network has the number of neurons same as the number of historical data points, which can produce a network with zero error on training vectors. Hence thirty hidden neurons are in the first layer and one neuron is in the second layer.

[1] For Data Sets which have five cost drivers and are associated with a cost function, such as Linear, Power, Power-Step, Linear-Power, Linear-Step, but not with Linear-Power-Step cost function

The experiments of these two types of neural networks (NN1 and NN2) are performed using Matlab 7.1.

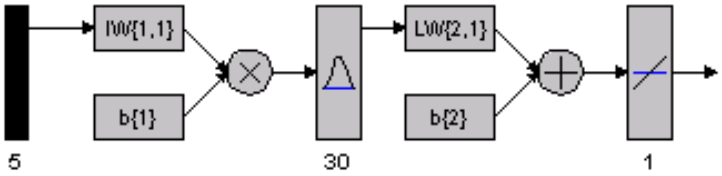


Figure 5-11 Radial Basis Neural Networks (NN2)[1]

5.3.1.4 Case-Based Reasoning

For case-based reasoning (see Section 2.3.3), based on five cost drivers, the degree of similarity was calculated according to the following equation (5-32):

$$SIM(X, Y) = 1 - DIST(X, Y) = 1 - \sqrt{\sum_i w_i^2 dist^2(x_i, y_i)} \qquad (5\text{-}32)$$

where weight (w_i) is up to coefficient of linear model. Two adaptation procedures are then applied to get the solution. In the first procedure (CBR1), the output value (y) of one training point, whose value of similarity measurement is the largest, is chosen as the predicted value for the test point. In the second procedure (CBR2), the corresponding predicted value is determined as the average value of two most similar cases.

5.3.2 Result Analysis

5.3.2.1 Appropriate Parameters for SVR

Proper parameters setting can improve the cost estimating predicable accuracy. This section shows that the cost model can more accurately predict unknown data via identifying optimal choice of hyper parameters. From Chapter 2.5.2, the SVR performance depends on kernel, its parameters and model parameters (C, and ε) for ε-

[1] For Data Sets which have five cost drivers and are associated with a cost function, such as Linear, Power, Power-Step, Linear-Power, Linear-Step, but not with Linear-Power-Step cost function

insensitive loss function. Generally, they affect performance together. The selection of optimum values for these training parameters (C and ε) and kernel's parameters is an active area of research. However there is not a structure method to select optimum values of these parameters. In this study the tabu search algorithm would be used to choose appropriate hyper parameters to improve the performance. All hyper parameters were obtained by empirical study or by the tabu search algorithm. The hyper parameters of SVR and Tabu-SVR under all data sets are showed in Table 5-2. The performance results via the tabu search algorithm were compared with those via empirical study under different scenarios in Table 5-3. Then the Wilcoxon signed ranks tests were conducted in Table 5-4, to examine whether the selection of hyper parameters via tabu search algorithm significantly improved the performance of SVR.

Table 5-3 shows the RMSE and accuracy results of Tabu-SVR and SVR. Except for the data set associated with a power cost function, the performance of the other data sets was improved by using Tabu-SVR. For data sets associated with a linear cost function, RMSE of linear kernel was reduced from around 132 to around 35 and accuracy was increased from around 88% to around 97%. For the data set associated with a linear-power cost function, RMSE of linear kernel was decreased from 10.71 to 3.53 and accuracy was improved from 86.12% to 95.20%; RMSE of polynomial kernel was decreased from 12.94 to 1.21 and accuracy was improved from 86.12% to 95.20%; RMSE of RBF kernel was reduced from 12.93 to 1.32 and accuracy was increased from 83.32% to 98.50%.

From the perspective of accuracy, it was more improvement for the data set associated with a linear-power-step cost function, RMSE of linear kernel was dropped from 28.80 to 10.50 and the accuracy rose from 78.36% to 92.91%; RMSE of RBF kernel was reduced from 28.30 to 3.19 and accuracy was improved from 78.33% to 97.65%; though RMSE and accuracy of polynomial kernel was not changed. The largest improvement was on the data set associated with a linear-step cost function and data set from a real cost detailed model (aircraft Engine). For the data set associated with a linear-step cost function, RMSE of all kernels was decreased from 650 to 248 (linear), 177.99(polynomial), and 222.90(RBF); accuracy was improved from 54% to above 82%. For the data set associated with the aircraft engine, RMSE was decreased from 752,054 to

338,816 (linear kernel), from 770,889 to 124,678 (polynomial kernel), and from 770,848 to 109,972; accuracy was improved from 59% to 86% (linear kernel), from 60% to 96% (polynomial and RBF kernel). For the data set associated with a power-step cost function, except linear and polynomial kernel, RMSE and accuracy of RBF kernel improved a lot, respectively from 12.63 to 3.02 and from 7.96% to 56.20%. Moreover, for the data set associated with a power cost function, RMSE and accuracy of all kernels were not improved.

Since values are not always known to be normal distributed, the Wilcoxon signed test is to be preferred over the Student t-test. Here Wilcoxon test for choice of hyper parameters (H_0: the accuracy of SVR and Tabu-SVR is equal) was conducted. Table 5-3 shows that RMSE and accuracy were improved under all data sets except the data set produced by power function. The *p-values* in Table 5-4 suggested: except for the data set associated with a power cost function under all kernels, a power-step cost function under polynomial kernel, and a linear-power-step cost function under polynomial kernel, all improvements of the tabu search algorithm for choice of hyper parameters are significant.

In summary, the tabu search algorithm for choosing parameters for SVR (Tabu-SVR) can significantly improve the performance over SVR with empirical study for most data sets.

Table 5-2 The Hyper Parameters of SVR with Empirical Study and Tabu-SVR for the Data Sets

		Kernel		
Data Sets		Linear (C; ε)	Polynomial (C; ε; γ; β; d)	RBF (C; ε; γ)
Linear	SVR	(1418; 355)	(1418; 355.9; 0.000488; 0; 3)	(1418; 355.9; 0.000488)
	Tabu-SVR	(671361482; 1.61)	(8207921; 1.74; 8.33; 63.93; 1)	(199600833; 46.30; 0.1493)
Linear-Step	SVR	(2071; 2663)	(2071; 2663; 0.000488; 0; 3)	(2071; 2663; 0.000488)
	Tabu-SVR	(2196; 96.52)	(84355032; 57.10; 56.47; 84.24; 2)	(421917211; 0.2923; 3.38)
Power	SVR	(499.5; 1587)	(499.5; 1588; 0.000488; 0; 3)	(499.5; 1588; 0.000488)
	Tabu-SVR	(819105583; 4.44)	(31306882; 851863; 72.73; 42.81; 7)	(255195575; 97091; 10.79)
Linear-Power	SVR	(108.3; 17.93)	(108.3; 17.93; 0.000488; 0; 3)	(108.3; 17.93; 0.000488)
	Tabu-SVR	(10.07; 0.05118)	(80459428; 0.02011; 13.51; 44.35; 2)	(474749259; 0.7482; 0.07294)
Power-Step	SVR	(39.39; 77; 93)	(39.39; 77; 93; 0.000488; 0; 3)	(39.39; 77; 93; 0.000488)
	Tabu-SVR	(121535774; 1e-007)	(87210032; 2854029; 71.00; 95.24; 7)	(480356155; 0.001414; 0.9006)
Linear-Power-Step	SVR	(191.9; 73.72)	(191.9; 73.72; 0.000488; 0; 3)	(191.9; 73.72; 0.000488)
	Tabu-SVR	(31.50; 0.120025)	(35917982; 4602666; 96.33; 73.55; 7)	(779005577; 0.5482; 0.04027)
Aircraft Engine	SVR	(2918482; 2823387)	(2918482; 2823387; 0.000488; 0; 3)	(2918482; 2823387; 0.000488)
	Tabu-SVR	(609051117; 549.54)	(29305139; 218.1; 3.408; 6.635; 4)	(507358263; 35.74; 0.2802)

Table 5-3 The Choice of Hyper Parameters

		Kernel							
		Linear			Polynomial			RBF	
Data Sets	Measurement	SVR	Tabu-SVR		SVR	Tabu-SVR		SVR	Tabu-SVR
Linear	RMSE	132.04	36.06		132.22	28.58		132.22	47.33
	Accuracy	88.68%	97.42%		88.70%	98.05%		88.70%	96.24%
Linear-Step	RMSE	650.21	248.53		651.30	177.99		651.30	222.90
	Accuracy	54.48%	82.78%		54.39%	89.21%		54.39%	87.24%
Power	RMSE	29.89	61.78		28.14	24.75		28.14	31.61
	Accuracy	9.13%	4.08%		9.54%	10.60%		9.54%	8.33%
Linear-Power	RMSE	10.71	3.53		12.94	1.21		12.93	1.32
	Accuracy	86.12%	95.30%		83.31%	98.55%		83.32%	98.50%
Power-Step	RMSE	12.60	7.79		12.63	13.31		12.63	3.02
	Accuracy	7.78%	18.02%		7.96%	11.77%		7.96%	56.20%
Linear-Power-Step	RMSE	28.08	10.50		28.30	28.30		28.30	3.19
	Accuracy	78.36%	92.91%		78.33%	78.33%		78.33%	97.65%
Aircraft Engine	RMSE	752054.73	338815.80		770889.35	124678.03		770848.57	109972.68
	Accuracy	58.86%	86.37%		60.49%	95.85%		60.49%	96.32%

Table 5-4 Wilcoxon Signed Rank Test for Choice of Hyper Parameters

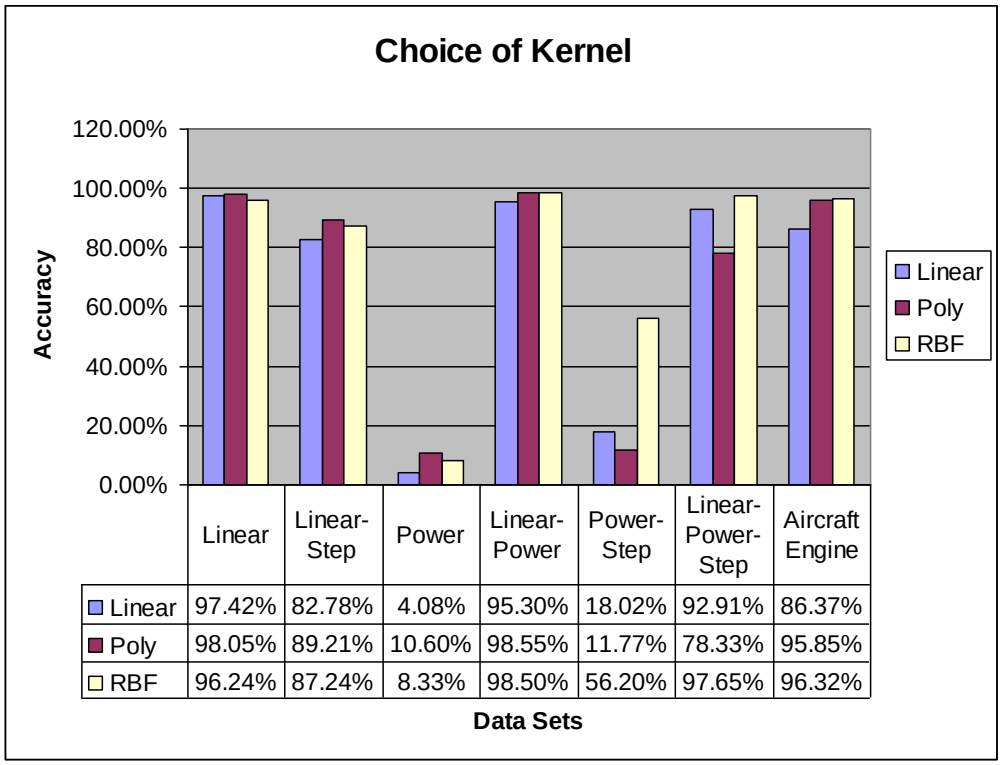
p-value (Wilcoxon Signed Rank Test)	Linear Kernel	Polynomial Kernel	RBF Kernel
	SVR vs. Tabu-SVR	SVR vs. Tabu-SVR	SVR vs. Tabu-SVR
Linear	<.0001	<.0001	<.0001
Linear-Step	<.0001	<.0001	<.0001
Power	0.2188*	1.0000*	0.6875*
Linear-Power	<.0001	<.0001	<.0001
Power-Step	0.0098	0.8311*	<.0001
Linear-Power-Step	<.0001	1.0000*	<.0001
Aircraft Engine	<.0001	<.0001	<.0001

* denote α>0.1 (insignificant difference)

5.3.2.2 Appropriate Kernels of SVR for Cost Estimates

As mentioned previously, SVR performance depends heavily on the kernel function. However the kernel should reflect the intrinsic topological structure of the data and application-domain knowledge. Which kernel is much more appropriate for cost estimates is the major focus in this section.

The results of this study indicate that one kernel function performs better than another two under most conditions. From Table 5-3 and Figure 5-12, the polynomial kernel outperformed linear kernel and RBF kernel for the data sets associated with a linear cost function, a linear-step cost function, a power cost function, and a linear-power cost function. But the performance of RBF is very close to polynomial kernel. For example, for the data set associated with a linear-power cost function, the RMSE of polynomial and RBF kernels are respectively 1.32 and 1.21 comparing to RMSE (3.53) of linear kernel. Their accuracy shows same trend: 98.55% (polynomial), 98.50% (RBF) and 95.30% (linear). For the data sets associated a linear cost function, a linear-step cost function, a power cost function, a linear-power cost function, and aircraft engine, the difference of accuracy between polynomial and RBF kernel is less than 2%. However, for the data sets associated with a power-step cost function and a linear-power-step cost function, the performance of RBF is much better than those of polynomial and linear kernel. Even the linear kernel has much better accuracy and RMSE than polynomial kernel does.



	Linear	Linear-Step	Power	Linear-Power	Power-Step	Linear-Power-Step	Aircraft Engine
Linear	97.42%	82.78%	4.08%	95.30%	18.02%	92.91%	86.37%
Poly	98.05%	89.21%	10.60%	98.55%	11.77%	78.33%	95.85%
RBF	96.24%	87.24%	8.33%	98.50%	56.20%	97.65%	96.32%

Figure 5-12 Choice of Linear Kernel, Polynomial Kernel and RBF Kernel

Wilcoxon signed rank test statistically compared the performance of accuracy for different data sets among linear, polynomial and RBF kernels. As shown in Table 5-5, although three kernels have good performance (their accuracy greater than 95%), the polynomial and linear kernels significantly outperformed RBF kernel. However the performance of polynomial and linear is not significantly different. Table 5-6 shows the polynomial kernels have much better performance than linear kernel. And the performance of polynomial kernels is not significantly different with that of RBF kernel. For data set associated with a power cost function (Table 5-7), there is not significant difference among three kernels. Moreover, for data set associated with a linear-power cost function, Table 5-8 shows the accuracy of RBF and polynomial kernels are much better than that of linear kernel, and the performance of polynomial kernels is not significantly different with that of RBF kernel. Table 5-9 shows for the data set associated with a power-step cost function, the performance of RBF kernel is significantly better than those of linear kernel and polynomial kernel. For the data set

associated with a linear-power-step cost function, polynomial kernel has worst performance, RBF kernel is best and the linear kernel is in the middle (Table 5-10). They are all significantly different. Table 5-11 shows the RBF and polynomial kernel have better accuracy than linear and they are not significantly different.

In conclusion, RBF is best kernel under all kinds of above data sets except for the data set associated with a linear cost function. However, its accuracy is 96.24% for the data set associated with a linear cost function, which is acceptable. The polynomial kernel often has good performance. But for data sets associated with a power-step cost function and a linear-power-step cost function, its performance is not good and even worse than that of linear kernel.

Table 5-5 Wilcoxon Signed Rank Test for Kernels under Data Set (Linear)

Data Set (Linear)

RMSE	Accuracy	p-value(Wilcoxon Signed Rank Test)	Linear	Polynomial	RBF
36.06	97.42%	Linear		0.1074*	0.0106
28.58	98.05%	Polynomial			<.0001
47.33	96.24%	RBF			

* denote α>0.1 (insignificant difference)

Table 5-6 Wilcoxon Signed Rank Test for Kernels under Data Set (Linear-Step)

Data Set (Linear-Step)

RMSE	Accuracy	p-value(Wilcoxon Signed Rank Test)	Linear	Polynomial	RBF
248.53	82.78%	Linear		0.0106	0.094*
177.99	89.21%	Polynomial			0.1755*
222.90	87.24%	RBF			

* denote α>0.1 (insignificant difference)

Table 5-7 Wilcoxon Signed Rank Test for Kernels under Data Set (Power)

Data Set (Power)

RMSE	Accuracy	p-value(Wilcoxon Signed Rank Test)	Linear	Polynomial	RBF
61.78	4.08%	Linear		0.2188*	0.25*
24.75	10.60%	Polynomial			0.5625*
31.61	8.33%	RBF			

* denote α>0.1 (insignificant difference)

Table 5-8 Wilcoxon Signed Rank Test for Kernels under Data Set (Linear-Power)

Data Set (Linear-Power)

RMSE	Accuracy	p-value(Wilcoxon Signed Rank Test)	Linear	Polynomial	RBF
3.53	95.30%	Linear		<.0001	<.0001
1.21	98.55%	Polynomial			0.5001*
1.32	98.50%	RBF			

* denote α>0.1 (insignificant difference)

Table 5-9 Wilcoxon Signed Rank Test for Kernels under Data Set (Power-Step)

Data Set (Power-Step)

RMSE	Accuracy	p-value(Wilcoxon Signed Rank Test)	Linear	Polynomial	RBF
7.79	18.02%	Linear		0.5195*	0.0001
13.31	11.77%	Polynomial			<.0001
3.02	56.20%	RBF			

* denote α>0.1 (insignificant difference)

Table 5-10 Wilcoxon Signed Rank Test for Kernels under Data Set (Linear-Power-Step)

Data Set (Linear-Power-Step)

RMSE	Accuracy	p-value(Wilcoxon Signed Rank Test)	Linear	Polynomial	RBF
10.50	92.91%	Linear		<.0001	<.0001
28.30	78.33%	Polynomial			<.0001
3.19	97.65%	RBF			

Table 5-11 Wilcoxon Signed Rank Test for Kernels under Data Set (Aircraft Engine)

Data Set (Aircraft Engine)

RMSE	Accuracy	p-value(Wilcoxon Signed Rank Test)	Linear	Polynomial	RBF
338,816	86.37%	Linear		<.0001	<.0001
124,678	95.85%	Polynomial			0.5131*
109,973	96.32%	RBF			

* denote α>0.1 (insignificant difference)

5.3.2.3 Data Sensitivity Test

After choosing appropriate kernel with corresponding parameters, the SVR model can be easily built to estimate estimates. However the sensitivity of this model to data is very

important. If the model was not robust to data, the performance of this model would not be reliable. This section will study the sensitivity of Tabu-SVR to data.

Table 5-12 shows RMSE and accuracy under different kernels for two test data sets respectively from each formula. RMSE and accuracy seems very different for the data sets associated with a linear-step cost function (RBF kernel), power (linear kernel, polynomial kernel, RBF kernel), aircraft engine.

Wilcoxon test (Table 5-13) statistically compared the performance (accuracy) for two data sets associated same cost functions under different kernel functions. For seven data sets, the performance under RBF kernel for two test data sets is not significantly different. The accuracy of polynomial kernel for two test data sets associated with a linear-power cost function is significantly different at the level α=0.1. That means polynomial kernel is sensitive to the sample sets when the data have linear-power property. The accuracy under linear kernel for two test sets from linear and aircraft engine is significantly different at the level α=0.1. Thus linear kernel is sensitive to the sample sets when the data have linear property.

The RBF kernel is most robust. For all test data sets, it has not significant difference in performance. The polynomial kernel is sensitive to the data set associated with a linear-power cost function. The linear kernel was worst, which has significant difference when data sets produced by the linear cost function and the real cost detail model (aircraft engine).

Table 5-12 Data Sensitivity Test

		Kernel							
		Linear (Tabu-SVR)			Poly (Tabu-SVR)			RBF (Tabu-SVR)	
	Measurement	Data Set1	Data Set2		Data Set 1	Data Set2		Data Set1	Data Set2
Linear	RMSE	36.06	40.50		28.58	28.65		47.33	36.49
	Accuracy	97.42%	96.40%		98.05%	97.63%		96.24%	97.26%
Linear-Step	RMSE	248.53	245.43		177.99	133.28		222.90	166.93
	Accuracy	82.78%	84.46%		89.21%	91.57%		87.24%	90.19%
Power	RMSE	61.78	99.15		24.75	100.19		31.61	94.66
	Accuracy	4.08%	1.21%		10.60%	10.84%		8.33%	6.33%
Linear-Power	RMSE	3.53	3.58		1.21	0.77		1.32	0.86
	Accuracy	95.30%	95.56%		98.55%	99.12%		98.50%	98.97%
Power-Step	RMSE	7.79	10.18		13.31	16.23		3.02	4.08
	Accuracy	18.02%	21.19%		11.77%	12.49%		56.20%	60.16%
Linear-Power-Step	RMSE	10.50	9.50		28.30	16.23		3.19	3.11
	Accuracy	92.91%	91.37%		78.33%	77.57%		97.65%	97.35%
Aircraft Engine	RMSE	338815.80	182883.25		124678.03	61007.67		109972.68	56185.22
	Accuracy	86.37%	90.22%		95.85%	97.11%		96.32%	97.45%

Table 5-13 Wilcoxon Signed Rank Test for Data Sensitivity

p-value (Wilcoxon Signed Rank Test)	Linear	Polynomial	RBF
	Set1 vs. Set2	Set1 vs. Set2	Set1 vs. Set2
Linear	**0.0168***	0.3069	0.0782
Linear-Step	0.3911	0.3167	0.4141
Power	0.5000	0.5469	0.6250
Linear-Power	0.9920	**0.0451***	0.4258
Power-Step	0.2661	0.6788	0.4954
Linear-Power-Step	0.0894	0.9122	0.9122
Aircraft Engine	**0.0476***	0.4025	0.1964

* denote α <0.1 (two side test).

5.3.2.4 Comparison with Traditional Methods

For complex products during the early design phases, the functional form of parametric methods (the linear or log-linear) must be determined based on experience. Moreover, nonparametric methods such as SVR, case-based reasoning, and neural networking, do not suffer from this disadvantage. With the exception of a case of an exact fit, the performance of the parametric methods is worse than that of nonparametric methods.

From Table 5-14 and Table 5-15, for data sets associated with a linear cost function, undoubtedly, parametric method (linear) and Tabu-SVR with polynomial kernel outperformed other methods. Their RMSE (28.58 of Tabu-SVR (poly) and 29.13 of linear) are best. The performance of Tabu-SVR (RBF) is very close to them. Its accuracy is 96.24%, which is much better than those of other methods. This is also true for data set associated with a power cost function and a power-step cost function, the parametric method with the form (log-linear) has the better performance. For the data set associated with a power cost function, none of the nonparametric methods and parametric methods (linear) has good performance. Moreover, from the perspective of RMSE, Tabu-SVR (poly) and Tabu-SVR (SVR) have much better performance than the parametric method (linear), case-based reasoning 2, and neural networking 1 and 2. For the data set associated with a power-step cost function, Tabu-SVR (RBF) also has better performance than Tabu-SVR (polynomial kernel), parametric method (linear), case-based reasoning 1 and 2, neural networking 1. Its performance is very close to that of neural networking 2.

For other data sets associated with a linear-step cost function, or a linear-power cost function, Tabu-SVR (RBF) and Tabu-SVR (polynomial kernel) have much better performance than other methods. For the data set associated with a linear-step cost function, the performance of Tabu-SVR (polynomial kernel) is significantly better than those of parametric methods (linear, log-linear), case-based reasoning (1 and 2), and neural networking (1 and 2). For the data set associated with a linear-power cost function, the performance of Tabu-SVR (polynomial and RBF kernels) is significantly better than that of parametric methods (linear, log-linear), case-based reasoning (1 and 2), and neural networking (1 and 2). For the data sets associated with a linear-power-step cost function and the real cost detailed model, the performance of Tabu-SVR (RBF kernel) is significantly better than of parametric methods (linear, log-linear), case-based reasoning (1 and 2), and neural networking 1. For the data set associated with a linear-power-step cost function, the RMSE (3.19) of Tabu-SVR (RBF kernel) is far less than RMSEs of other methods (see Table 5-14). The accuracy (96.32%) of Tabu-SVR (RBF) for the data set associated with the aircraft engine is much higher than accuracy of other methods: of parametric methods (linear, log-linear), case-based reasoning (1 and 2), and neural networking (1 and 2).

Thus, Tabu-SVR has better performance over other methods under most situations. It seems to be quite robust to the complexity of the hidden relationship among cost drivers and cost except for data sets associated with a power cost function and power-step cost function. Even for these two data sets, their performance is not worse than those of most other methods. The exception to this is that the log-linear method does model power relationship well as should be expected.

Table 5-14 Comparison with Traditional Methods

Data Sets	Measurement	Tabu-SVR(Poly)	Tabu-SVR(RBF)	Log-Linear	Linear	CBR1	CBR2	NN1	NN2
Linear	RMSE	**28.58**	47.33	60.32	29.13	67.91	61.46	72.29	415.76
	Accuracy	**98.05%**	96.24%	94.72%	97.92%	93.98%	94.67%	94.17%	68.14%
Linear-Step	RMSE	**177.99**	222.90	248.995792	271.61	376.35	284.1266	271.19	361.21
	Accuracy	**89.21%**	87.24%	83.61%	81.31%	75.31%	0.80	83.73%	77.72%
Power	RMSE	24.75	31.61	**10.40**	72.04	22.49	109.65	185.32	132.80
	Accuracy	<50%	<50%	**80.04%**	<50%	<50%	<50%	<50%	<50%
Linear-Power	RMSE	**1.21**	1.32	5.40	2.57	5.33	4.18	2.70	3.16
	Accuracy	**98.55%**	98.50%	93.73%	96.88%	93.79%	94.73%	96.95%	96.72%
Power-Step	RMSE	13.31	3.02	**2.26**	7.32	5.12	4.83	4.48	2.83
	Accuracy	<50%	56.20%	**84.34%**	<50%	51.02%	54.23%	53.52%	53.81%
Linear-Power-Step	RMSE	28.30	**3.19**	13.23	6.68	14.22	13.28	25.91	5.67
	Accuracy	78.33%	**97.65%**	92.25%	95.89%	90.84%	91.23%	81.09%	96.76%
Aircraft Engine	RMSE	124678.03	**109972.68**	873792.97	316000.95	248326.98	275598.93	426500.34	192653.92
	Accuracy	95.85%	**96.32%**	73.65%	82.09%	90.17%	90.22%	79.90%	87.03%

Table 5-15 Wilcoxon Signed Rank Test for Comparison with Traditional Methods

Data Sets	p-value (Wilcoxon Signed Rank Test)	Log-Linear	Linear	CBR1	CBR2	NN1	NN2
Linear	**Tabu-SVR(Poly)**	<.0001	0.1122*	<.0001	<.0001	<.0001	<.0001
	Tabu-SVR(RBF)	0.1624*	0.0001	0.0044	0.0344	0.0216	<.0001
Linear-Step	**Tabu-SVR(Poly)**	0.0074	0.0011	0.0008	0.0003	0.0274	0.0001
	Tabu-SVR(RBF)	0.1384*	0.0476	0.0038	0.0148	0.1688*	0.0005
Power	Tabu-SVR(Poly)	<.0001	0.6406*	0.0005	0.0027	0.8926*	0.0313
	Tabu-SVR(RBF)	<.0001	0.8125*	0.0005	0.0015	0.8457*	0.2500*
Linear-Power	**Tabu-SVR(Poly)**	<.0001	0.0004	<.0001	<.0001	0.0003	0.0025
	Tabu-SVR(RBF)	<.0001	0.0055	<.0001	<.0001	0.0038	0.0168
Power-Step	Tabu-SVR(Poly)	<.0001	0.0179	<.0001	<.0001	<.0001	<.0001
	Tabu-SVR(RBF)	0.0203	0.0232	0.5634*	0.8073*	0.6475*	0.1876*
Linear-Power-Step	Tabu-SVR(Poly)	<.0001	<.0001	<.0001	<.0001	0.1122*	<.0001
	Tabu-SVR(RBF)	0.0004	0.0055	<.0001	<.0001	<.0001	0.0711*
Aircraft Engine	**Tabu-SVR(Poly)**	<.0001	<.0001	0.0011	0.0023	<.0001	0.0005
	Tabu-SVR(RBF)	<.0001	<.0001	0.0002	0.0006	<.0001	0.0003

* denote α>0.1 (insignificant difference)

5.4 Conclusions

In experiments, according to cost estimating basic common characteristics, six formulas and an EXCEL tool equivalent to FLOPS cost module were used to create data sets for testing. Using these data, a study was made in choosing parameters and the kernel function of SVR for cost estimation. Cost models based on support vector regression, parametric modeling, neural network and case-based reasoning were constructed. Their performance was then compared. The root mean square error (RMSE) and accuracy degree of prediction were used as performance criterion.

From the results, Tabu-SVR significantly improved the performance of the cost models based on SVR which choose appropriate parameters via empirical study. The RBF and polynomial kernel showed better performance over a linear kernel under most data sets. Moreover, the RBF kernel was much more robust to data of the problem.

When function forms are known, the nonparametric methods are not necessary and do not perform well. For example, when it is known that the data set would be produced by a linear function or the data set would be produced by a power function and a power-step function, the parametric method (linear) or the parametric method (log-linear) would be a good choice. However, when an apriori CER (functional form) is unknown, the nonparametric methods, such as support vector regression, case-based reasoning, and neural networking, have better performance.

Tabu-SVR cost modeling yielded good performance over other cost modeling techniques. The Tabu-SVR was able to capture these nonlinearities and discontinuities, along with interactions among cost drivers. Tabu-SVR hence had strong predicable capability. When the function form cannot be determined because of inadequate information, the Tabu-SVR can be used effectively. Therefore, the cost model based on SVR has a great potential to accurately estimate cost for complex product during the early design phases.

The largest benefits of the Tabu-SVR are the facts that a global solution exists and is found with appropriate parameters and kernel function in contrast to neural networks which have to be trained with randomly chosen initial weight setting. Furthermore, due to specific optimization procedure it is assured that overtraining is. A drawback of the Tabu-

SVR is that the searching time for appropriate parameters could be much longer than parametric methods, case-based reasoning, and neural network. Finding appropriate parameters of Tabu-SVR might spend hours of computation resource. However, cost estimating area does not generally need real-time estimates. Spending hours of training and model building is still acceptable for cost estimates.

Chapter 6 Cost Estimating Semiparametric Approach Based on Support Vector Regression

6.1 Introduction

At some points in times, while there may be limited knowledge about the parametric form of the cost relationships, the form is not adequately known throughout the entire range of data. At such times, the parametric approach would not be appropriate because the resulting fit would be misleading (biased) at points where the data deviates from the specified model. However, it is not wise to ignore the knowledge and only use a nonparametric approach.

A semiparametric approach is presented for such situation that combines a parametric approach with a nonparametric approach. The semiparametric approach is able to handle different amounts of model misspecification by combining a parametric regression fit, which is based on the researcher's knowledge of the underlying model, with a nonparametric regression fit, which is designed to capture any structure in the data that the parametric fit fails to explain. It can provide noticeable improvements over the two approaches when used individually. Therefore, the semiparametric approach has two components as (6-1): the parametric component $f_{param}(\beta_{param}; x)$ and the nonparametric component $f_{nonpar}(\beta_{nonpar}; x)$.

$$f(x) = f_{param}(\beta_{param}; x) + f_{nonpar}(\beta_{nonpar}; x) \qquad (6\text{-}1)$$

where β_{param} and β_{nonpar} are their parameters and x is vector of cost drivers.

There are limited studies in literatures [94, 95] on semiparametric approaches based on support vector regression. Smola, Frieb, et al.[94] extended the support vector regression to a semiparametric model. However, they did not consider multiple inputs and the data sets that included the properties of power and step functions. Pai and Lin[95] proposed a hybrid ARIMA and support vector regression to forecast stock price. The parametric component in their hybrid model only employed the first-order terms of inputs. Therefore there are limitations when applying these semiparametric approaches in cost estimation. Moreover, the application of a semiparametric approach for cost estimation has not been found in other literatures.

The main task of this chapter is to effectively and efficiently use semiparametric approach in the area of cost modeling. This chapter first introduces three semiparametric algorithms. Second three test data sets are produced using basic common cost characteristics summarized in last chapter. Extensive studies for these three semiparametric algorithms (A1, A2, and A3) on multiple inputs were then performed under different situations. Comparisons were made among these three algorithms (A1, A2, and A3), a pure nonparametric approach Tabu-SVR, and parametric approaches for cost estimation.

6.2 Semiparametric Approaches Based on SVR

For the semiparametric approach based on SVR, there are three algorithms considered in this chapter. Algorithm 1 (A1) is the most common idea in the semiparametric area. Algorithm 2 (A2) was presented by Pai and Lin[95]. Algorithm 3 (A3) was proposed by Smola, Frieb, et al.[94]. They are introduced respectively as follows.

Algorithm 1 (A1): The first algorithm uses residuals from the parametric model to train the nonparametric portion. There are three steps:

First, the parametric part (6-2) performs a regression using the training data for all β_{param}^{j}.

$$f_{param}(\beta_{param}; x) = \sum_{j=1}^{n} \beta_{param}^{j} \phi_j(x) = y \qquad (6\text{-}2)$$

Second, after β_{param}^{j} in Equation (6-2) is obtained, the parametric portion of semiparametric model, $f_{param}(\beta_{param}; x)$, is determined. The residual of the parametric portion becomes the output of nonparametric portion (6-3).

$$f_{nonpar}(\beta_{nonpar}; x) = e = y - \sum_{j=1}^{n} \beta_{param}^{j} \phi_j(x) = \langle w, \psi(x) \rangle \qquad (6\text{-}3)$$

For the nonparametric portion $f_{nonpar}(\beta_{nonpar}; x)$, all input *x* and output *e* are from the train data and residual. For this investigation, the RBF kernel was chosen. The parameter, β_{nonpar}, can then be found using the tabu search algorithm presented in Chapter 5.

Third, the final cost is sum of the parametric portion and the nonparametric portion.

Algorithm 2 (A2): A second algorithm was presented by Pai and Lin [95]. It is same as the algorithm 1 except that it considers the parametric portion when choosing the parameters of SVR.

First, the parametric part (6-2) performs a regression using the training data for all β_{param}^{j}. This step is the same as the first step in Algorithm 1.

After β_{param}^{j} is obtained, the parametric portion of semiparametric model is determined. The residual (6-3) of the parametric portion becomes as the output of the nonparametric portion. This is also same as Algorithm 1.

However, for the nonparametric portion, the parameters (β_{nonpar} such as *C,* ε*,* kernel parameters) are chosen to make the final performance best when considering the parametric portion and nonparametric portion. This is different with Algorithm 1 (see Table 6-1). Algorithm 1 chooses the parameters considering only the nonparametric portion.

After choosing the parameters β_{nonpar} using the tabu search algorithm, $f_{nonpar}(\beta_{nonpar}; x)$ is constructed as the same way in Chapter 5.

The final cost is sum of the parametric portion and the nonparametric portion.

Algorithm 3 (A3): Smola, Frieb, et al.[94] denoted that algorithm 1 generally would not lead to finding the minimum generalization error. Hence, the semiparametric procedure should involve simultaneously fitting a parametric and a nonparametric model. Smola, Frieb, et al. [94] proposed a feasible way to fit the parametric part and nonparametric part at the same time.

Using a general parametric part $\sum_{i=1}^{n} \beta_{param}^{i} \phi_i(x)$ replaces the b in Equation 5-1 to get Equation (6-4):

$$f(x) = f_{nonpar}(\beta_{nonpar}; x) + f_{param}(\beta_{param}; x) = \langle w, \psi(x) \rangle + \sum_{i=1}^{n} \beta_{param}^{i} \phi_i(x) \qquad (6\text{-}4)$$

where $\langle w, \psi(x) \rangle$ is the nonparametric part $f_{nonpar}(\beta_{nonpar}; x)$. The parameter β_{nonpar} here includes model parameters such as *C*, ε and kernel parameter.

Smola, Frieb, et al.[94] made an extension for SVR when considering the parametric parts. The Primal Problem (2-16) and the Dual Problem (2-20) were changed as follows Equations (6-5) and (6-6):

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ S.t. \quad & \begin{cases} y_i - \langle w, \psi(x) \rangle - \sum_{j=1}^{n} \beta_{param}^{j} \phi_j(x_i) \le \varepsilon + \xi_i \\ -y_i + \langle w, \psi(x) \rangle + \sum_{j=1}^{n} \beta_{param}^{j} \phi_j(x_i) \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \end{aligned} \tag{6-5}$$

$$\begin{aligned} \max \quad & -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \\ S.t. \quad & \begin{cases} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\phi_j(x_i) = 0 \qquad for\,all \quad 1 \le j \le n \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases} \end{aligned} \tag{6-6}$$

The RBF was chosen as its kernel here. The parameters β_{nonpar} can be found using tabu search algorithm described in Chapter 5. The optimization problem (α_i ; α_i^* and β_{param}^{j}) is solved when appropriate β_{nonpar} are determined. This is different with Algorithm 1 and Algorithm 2 (see Table 6-1).

The semiparametric cost estimating function can be found as (6-7):

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_i, x) + \sum_{j=1}^{n}\beta_{param}^{j}\phi_j(x) \tag{6-7}$$

The cost can be calculated via Equation 6-7 with known optimal α_i, α_i^*, β_{param}^{j} and an appropriate kernel function.

Table 6-1 Three Semiparametric Algorithms Based on SVR

	Algorithm 1 (A1)	Algorithm 2 (A2)	Algorithm (A3)
Step 1	Obtain β_{param} according to training data; Construct $f_{param}(\beta_{param};x)$	Obtain β_{param} according to training data; Construct $f_{param}(\beta_{param};x)$	Obtain β_{param} , β_{nonpar} according to training data simultaneously; Construct $f_{param}(\beta_{param};x)$ and $f_{nonpar}(\beta_{nonpar};x)$
Step 2	Obtain β_{nonpar} according to training data (residual and input *x*); Construct $f_{nonpar}(\beta_{nonpar};x)$	Obtain β_{nonpar} according to training data (residual and input *x*) while considering $f_{param}(\beta_{param};x)$; Construct $f_{nonpar}(\beta_{nonpar};x)$	Sum of $f_{param}(\beta_{param};x)$ and $f_{nonpar}(\beta_{nonpar};x)$
Step 3	Sum of $f_{param}(\beta_{param};x)$ and $f_{nonpar}(\beta_{nonpar};x)$	Sum of $f_{param}(\beta_{param};x)$ and $f_{nonpar}(\beta_{nonpar};x)$	

This chapter would make comparisons between the above three algorithms (see Table 6-1), a pure parametric model, and a pure nonparametric model via the experiments under different scenarios. The advantages and drawbacks under different scenarios are discussed later.

6.3 Experiments

6.3.1 Data

In the last chapter (Section 5.2.1), five common basic cost characteristics were summarized: accumulation, linear function, power function, step function, and exponential function. They are often combined in cost modeling area to express the cost relationship of a complex product. Based on these five common basic cost characteristics and general combining rules, the test cases (data sets) for semiparametric cost estimating approaches were produced. In the last chapter, experiments have shown that nonparametric cost estimating approach based on SVR had very good performance under most scenarios except those included power properties. Therefore, for semiparametric

cost approaches based on SVR, only the data including strong power properties were produced to test the performance.

Combing with five common basic cost characteristics, three equations were created in Table 6-2 to use in generating test data sets under the desire to include power and step functions. Each formula has five cost drivers with nonlinear features: strong power properties and step function. The formula of test case 1 does not include any interaction among x_1, x_2 and x_3. The formulas of test case 2 and 3, there exists interaction among x_2, x_4 and x_5. Moreover, the formula of test case 3 has larger parameters for its step function and greater power for its power property comparing to the formula of test case 2.

The data of each cost driver x_i ($i=1,2,3,4,5$) were uniformly and randomly produced in the range between 0 and 1. Noise was then added in each cost drivers (coefficient of variation, $c.v. = \sigma/\mu$=0.03). The corresponding cost was produced using the formulas of test cases as Table 6-2.

Table 6-2 Three Formulas to Produce Test Cases (Data Sets)

Name	Formula
Test Case 1 Producer	$C = 20+100x_1+150x_2 \cdot f_1(x_2)+500x_3^{-0.5}+250(x_4^{0.8})(x_5^{2})e^{f_2(x_4)}$ $f_1(x_2)=\begin{cases}2.5 & x_2 \geq 0.6\\ 0.5 & x_2 < 0.6\end{cases}$ $f_2(x_4)=\begin{cases}1.05 & x_4 \geq 0.6\\ 1.15 & x_4 < 0.6\end{cases}$
Test Case 2 Producer	$C = 20+100x_1+150x_2 \cdot f_1(x_2)+500x_3^{-0.3}+250(x_4^{0.8})(x_5^{1.3})e^{f_2(x_2)}+100x_5^{1.5}$ $f_1(x_2)=\begin{cases}1.5 & x_2 \geq 0.6\\ 0.5 & x_2 < 0.6\end{cases}$ $f_2(x_2)=\begin{cases}1.05 & x_2 \geq 0.6\\ 1.15 & x_2 < 0.6\end{cases}$
Test Case 3 Producer	$C = 20+100x_1+150x_2 \cdot f_1(x_2)+500x_3^{-0.5}+250(x_4^{0.8})(x_5^{-0.3})e^{f_2(x_2)}+200x_5^{3}$ $f_1(x_2)=\begin{cases}2.5 & x_2 \geq 0.6\\ 0.5 & x_2 < 0.6\end{cases}$ $f_2(x_2)=\begin{cases}1.05 & x_2 \geq 0.6\\ 1.15 & x_2 < 0.6\end{cases}$

6.3.2 Methods

The semiparametric approaches include two components: the parametric part and the nonparametric part as noted in Equation (6-1). The nonparametric parts are implemented using the Tabu-SVR presented in Chapter 5. The parametric part is generally based on knowledge of the underlying model. However, the knowledge of underlying model might be inexact, partial, or exact. The performance of three semiparametric algorithms, the pure nonparametric approach and pure parametric approach, under different degree of exactness and different amount of known knowledge of underlying model, is the focus of this chapter.

According to three test cases in Section 6.3.1, there are following situations considering in this study. Based on these situations, the corresponding parametric components of semiparametric algorithms or the parametric approach are defined in Table 6-3 through Table 6-6.

- "L" assumes all function forms of cost drivers are first-order terms.
- Known partial or exact function form of one cost driver
- Known partial or exact function forms of multiple cost drivers
- Inexact function form(s) of cost driver(s). This means that the function form is unknown. The first-order term of cost driver(s) is as a substitute of its function form.

1. "L" assumes all function forms of the cost drivers are first-order terms.

When there are unknown function forms of cost drivers, the first-order term is often used as the function form of cost driver in the parametric approach. Therefore, when the function forms of all cost drivers are the first-order terms, the relationships of parametric approach and the parametric portion of the semiparametric approach are listed in Table 6-3.

Table 6-3 The Relationship When the Function Forms of All Cost Driver Are First-Order

Test Case	Known Cost Drivers	Abbr*	Parametric Function Form**
Test Case 1	$!x_1, !x_2, !x_3, !x_4, !x_5$	Param-L	$f(\beta, x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$
		A1-L, A2-L, A3-L	$f_{param}(\beta_{param}; x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$
Test Case 2	$!x_1, !x_2, !x_3, !x_4, !x_5$	Param-L	$f(\beta, x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$
		A1-L, A2-L, A3-L	$f_{param}(\beta_{param}; x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$
Test Case 3	$!x_1, !x_2, !x_3, !x_4, !x_5$	Param-L	$f(\beta, x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$
		A1-L, A2-L, A3-L	$f_{param}(\beta_{param}; x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$

Note:
*Abbr: The "Param-L"denotes parametric approach with all first-order terms of cost drivers.
.The "A1-L", "A2-L", or "A3-L" respectively denotes that in the parametric component of semiparametric algorithms, the function forms of all cost drivers are the first-order term.
** Parametric Function Form means:
1. for the parametric approach, the form is its relationship.
2. for the semiparametric algorithms, the form is the relationship of its parametric component
! means "unknown but instead of the first-order term". The function form of its cost driver is expressed as the first-order term.

2. Known partial or exact function form of one cost driver

Under test case 1, the exact function forms of x_2 or x_3 are known. Under test case 2 and test case 3, the exact function form of x_3 is known. The partial function form of x_2 *or* x_5 is known because the cost drivers x_2 or x_5 involve the interaction term which are assumed unknown. Therefore, the relationship of parametric approach and parametric component of semiparametric algorithms can be expressed as Table 6-4.

3. Known partial or exact function forms of multiple cost drivers

Under test case 1, the exact function form of x_2 and x_3 might be known together. Under test case 2 and test case 3, both the partial function form of x_2 and the exact function form of x_3 might be known; or all three of the partial function form of x_2, the exact function form of x_3, and the partial function form of x5 might be known. Moreover, the interaction term is assumed unknown. Therefore, the relationship of parametric approach and parametric component of semiparametric algorithms can be expressed as Table 6-5.

Table 6-4 The Relationship with Known Partial or Exact Function Form of One Cost Driver

Test Case	Known Cost Drivers	Abbr*	Parametric Function Form**
Test Case 1	x_2	Param-2	$f(\beta,x)=\beta_0+\beta_1x_1+\beta_2x_2\cdot f_1(x_2)+\beta_3x_3+\beta_4x_4+\beta_5x_5$ $f_1(x_2)=\begin{cases}2.5 & x_2\geq 0.6\\ 0.5 & x_2<0.6\end{cases}$
		A1-2 or A2-2 or A3-2	$f_{param}(\beta_{param};x)=\beta_2x_2\cdot f_1(x_2)$ $f_1(x_2)=\begin{cases}2.5 & x_2\geq 0.6\\ 0.5 & x_2<0.6\end{cases}$
	x_3	Param-3	$f(\beta,x)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3^{-0.5}+\beta_4x_4+\beta_5x_5$
		A1-3 or A2-3 or A3-3	$f_{param}(\beta_{param};x)=\beta_3x_3^{-0.5}$
Test Case 2	^x_2	Param-2	$f(\beta,x)=\beta_0+\beta_1x_1+\beta_2x_2\cdot f_1(x_2)+\beta_3x_3+\beta_4x_4+\beta_5x_5$ $f_1(x_2)=\begin{cases}1.5 & x_2\geq 0.6\\ 0.5 & x_2<0.6\end{cases}$
		A1-2 or A2-2 or A3-2	$f_{param}(\beta_{param};x)=\beta_2x_2\cdot f_1(x_2)$ $f_1(x_2)=\begin{cases}1.5 & x_2\geq 0.6\\ 0.5 & x_2<0.6\end{cases}$
	x_3	Param-3	$f(\beta,x)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3^{-0.3}+\beta_4x_4+\beta_5x_5$
		A1-3 or A2-3 or A3-3	$f_{param}(\beta_{param};x)=\beta_3x_3^{-0..3}$
	^x_5	Param-5	$f(\beta,x)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5^{1.5}$
		A1-5 or A2-5 or A3-5	$f_{param}(\beta_{param};x)=\beta_5x_5^{1.5}$
Test Case 3	^x_2	Param-2	$f(\beta,x)=\beta_0+\beta_1x_1+\beta_2x_2\cdot f_1(x_2)+\beta_3x_3+\beta_4x_4+\beta_5x_5$ $f_1(x_2)=\begin{cases}2.5 & x_2\geq 0.6\\ 0.5 & x_2<0.6\end{cases}$
		A1-2 or A2-2 or A3-2	$f_{param}(\beta_{param};x)=\beta_2x_2\cdot f_1(x_2)$ $f_1(x_2)=\begin{cases}2.5 & x_2\geq 0.6\\ 0.5 & x_2<0.6\end{cases}$
	x_3	Param-3	$f(\beta,x)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3^{-0.5}+\beta_4x_4+\beta_5x_5$
		A1-3 or A2-3 or A3-3	$f_{param}(\beta_{param};x)=\beta_3x_3^{-0..5}$
	^x_5	Param-5	$f(\beta,x)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5^{3}$
		A1-5 or A2-5 or A3-5	$f_{param}(\beta_{param};x)=\beta_5x_5^{3}$

Note:

*Abbr: Param means the parametric approach. The A1, A2, and A3 respectively denote Semiparametric Algorithm 1, Algorithm 2, and Algorithm 3.

** Parametric Function Form means:

1. for the parametric approach, the form is its relationship.
2. for the semiparametric algorithms, the form is the relationship of its parametric component

^ means "partial known"

Table 6-5 The Relationship with Known Partial or Exact Function Form of Multiple Cost Drivers

Test Case	Known Cost Drivers	Abbr*	Parametric Function Form**
Test Case 1	x_2, x_3	Param-23	$f(\beta, x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdot f_1(x_2) + \beta_3 x_3^{-0.5} + \beta_4 x_4 + \beta_5 x_5$ $f_1(x_2) = \begin{cases} 2.5 & x_2 \geq 0.6 \\ 0.5 & x_2 < 0.6 \end{cases}$
		A1-23 or A2-23 or A3-23	$f_{param}(\beta_{param}; x) = \beta_2 x_2 \cdot f_1(x_2) + \beta_3 x_3^{-0.5}$ $f_1(x_2) = \begin{cases} 2.5 & x_2 \geq 0.6 \\ 0.5 & x_2 < 0.6 \end{cases}$
Test Case 2	^x_2, x_3	Param-23	$f(\beta, x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdot f_1(x_2) + \beta_3 x_3^{-0.3} + \beta_4 x_4 + \beta_5 x_5$ $f_1(x_2) = \begin{cases} 1.5 & x_2 \geq 0.6 \\ 0.5 & x_2 < 0.6 \end{cases}$
		A1-23 or A2-23 or A3-23	$f_{param}(\beta_{param}; x) = \beta_2 x_2 \cdot f_1(x_2) + \beta_3 x_3^{-0..3}$ $f_1(x_2) = \begin{cases} 1.5 & x_2 \geq 0.6 \\ 0.5 & x_2 < 0.6 \end{cases}$
	^x_2, x_3, ^x_5	Param-235	$f(\beta, x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdot f_1(x_2) + \beta_3 x_3^{-0.3} + \beta_4 x_4 + \beta_5 x_5^{1.5}$ $f_1(x_2) = \begin{cases} 1.5 & x_2 \geq 0.6 \\ 0.5 & x_2 < 0.6 \end{cases}$
		A1-235 or A2-235 or A3-235	$f_{param}(\beta_{param}; x) = \beta_2 x_2 \cdot f_1(x_2) + \beta_3 x_3^{-0..3} + \beta_5 x_5^{1.5}$ $f_1(x_2) = \begin{cases} 1.5 & x_2 \geq 0.6 \\ 0.5 & x_2 < 0.6 \end{cases}$
Test Case 3	^x_2, x_3	Param-23	$f(\beta, x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdot f_1(x_2) + \beta_3 x_3^{-0..5} + \beta_4 x_4 + \beta_5 x_5$ $f_1(x_2) = \begin{cases} 2.5 & x_2 \geq 0.6 \\ 0.5 & x_2 < 0.6 \end{cases}$
		A1-23 or A2-23 or A3-23	$f_{param}(\beta_{param}; x) = \beta_2 x_2 \cdot f_1(x_2) + \beta_3 x_3^{-0..5}$ $f_1(x_2) = \begin{cases} 2.5 & x_2 \geq 0.6 \\ 0.5 & x_2 < 0.6 \end{cases}$
	^x_2, x_3, ^x_5	Param-235	$f(\beta, x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdot f_1(x_2) + \beta_3 x_3^{-0..5} + \beta_4 x_4 + \beta_5 x_5^3$ $f_1(x_2) = \begin{cases} 2.5 & x_2 \geq 0.6 \\ 0.5 & x_2 < 0.6 \end{cases}$
		A1-235 or A2-235 or A3-235	$f_{param}(\beta_{param}; x) = \beta_2 x_2 \cdot f_1(x_2) + \beta_3 x_3^{-0.5} + \beta_5 x_5^3$ $f_1(x_2) = \begin{cases} 2.5 & x_2 \geq 0.6 \\ 0.5 & x_2 < 0.6 \end{cases}$

Note:
*Abbr: Param means the parametric approach. The A1, A2, and A3 respectively denote Semiparametric Algorithm 1, Algorithm 2, and Algorithm 3.
** Parametric Function Form means:

1. for the parametric approach, the form is its relationship.
2. for the semiparametric algorithms, the form is the relationship of its parametric component

^ means "partial known"

4. Inexact or unknown function form(s) of cost drivers

This part further studies the influence of inexact knowledge of function form on estimating accuracy based on Semiparametric Algorithm 3. If there are not known exact

function forms, the corresponding first-order term (s) of cost driver is (are) as the parametric component of semiparametric algorithms. These function forms of parametric components are listed in Table 6-6.

Table 6-6 The Relationship with Inexact Function Forms of Cost Driver(s)

Test Case	Known Cost Drivers	Abbr*	Parametric Function Form**
Test Case 1	$!x_2$	A3-02	$f_{param}(\beta_{param}; x) = \beta_2 x_2$
	$!x_3$	A3-03	$f_{param}(\beta_{param}; x) = \beta_3 x_3$
	$!x_2$, $! x_3$	A3-0203	$f_{param}(\beta_{param}; x) = \beta_2 x_2 + \beta_3 x_3$
Test Case 2	$!x_2$,	A3-02	$f_{param}(\beta_{param}; x) = \beta_2 x_2$
	$!x_3$	A3-03	$f_{param}(\beta_{param}; x) = \beta_3 x_3$
	$!x_5$	A3-05	$f_{param}(\beta_{param}; x) = \beta_5 x_5$
	$!x_2$,$! x_3$,	A3-0203	$f_{param}(\beta_{param}; x) = \beta_2 x_2 + \beta_3 x_3$
	$!x_2$,$! x_3$,$!x_5$	A3-020305	$f_{param}(\beta_{param}; x) = \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5$
Test Case 3	$!x_2$,	A3-02	$f_{param}(\beta_{param}; x) = \beta_2 x_2$
	$!x_3$	A3-03	$f_{param}(\beta_{param}; x) = \beta_3 x_3$
	$!x_5$	A3-05	$f_{param}(\beta_{param}; x) = \beta_5 x_5$
	$!x_2$,$! x_3$,	A3-0203	$f_{param}(\beta_{param}; x) = \beta_2 x_2 + \beta_3 x_3$
	$!x_2$,$! x_3$,$!x_5$	A3-020305	$f_{param}(\beta_{param}; x) = \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5$

Note:
*Abbr: The "A3" denotes Semiparametric Algorithm 3.
** Parametric Function Form means: the form is the relationship of its parametric component
! means "unknown". The function form of its cost driver is expressed as the first-order term.

6.3.3 Results and Discussion

6.3.3.1 Comparison between Semiparametric Algorithms Based on SVR and Parametric Approach

According to the amount and type of known information, the different parametric components (see Table 6-3 through Table 6-4) and the nonparametric component were combined to construct different semiparametric models. Their performance was listed and compared under three scenarios: test case 1, 2, and 3. Table 6-7, Table 6-8 and Table 6-9 respectively show comparison between parametric approach and three semiparametric algorithms (A1, A2, and A3) under these test cases.

In Figure 6-1, Figure 6-2, and Figure 6-3, first column in each series represents the performance of the pure parametric approach. The last column in each series represents the performance of a pure nonparametric approach based on SVR. Between them, second

column represents Semiparametric Algorithm (A1) under the different amount and type of know information.

From Table 6-7 through Table 6-9 and Figure 6-1 through Figure 6-3, it can be seen that the second column (semiparametric A1) in each series is always higher than or almost equal to the corresponding first column (the pure parametric approach). This means that the performance of A1 is always better than the performance of the pure parametric approach with same amount and types of known information. From the Table 6-7 through Table 6-9, the *p-values* of Wilcoxon signed rank test also show under some situations the semiparametric A1 is significantly better than corresponding pure parametric approach. Under test case 1, if the function forms for "x_2 and x_3" are known, the semiparametric A1 has highest accuracy. It is significantly better than the corresponding pure parametric approach. Under test case 2, semiparametric A1 models with know function forms about "x_5" or "x_2, x_3 and x_5" are respectively and significantly better than their corresponding parametric approach with known function forms on "x_5" or "x_2, x_3 and x_5".When assuming function forms of all cost drivers are first-order terms, A1-L has significant improvement over Param-L.

Moreover, the performance of semiparametric A2 is not good. This is discussed in Section 6.3.3.5. The performance of semiparametric A3 with known function form on "x_3" is very good. Even under the test case 3, it is best in the Table 6-9 and Figure 6-3, which have significant improvement oven the corresponding Param-3. However, for semiparametric A3, the known function forms on other cost drivers did not bring any benefits for estimating accuracy under test case 3. This is discussed in Section 6.3.3.5.

Table 6-7 Comparisons (Accuracy and Wilcoxon Signed Rank Test) under Test Case 1

	Accuracy	0.7181	0.7257	0.7166	0.9321	*0.9757*	0.6702	0.7164	0.2266	0.3052	0.4420	0.7399	0.9493	0.6989	0.7285	0.7268	0.6621
Accuracy	p-value (Wilcoxon Signed Rank Test)	**Nonpar**	**A1-L**	**A1-2**	**A1-3**	**A1-23**	**A2-L**	**A2-2**	**A2-3**	**A2-23**	**A3-L**	**A3-2**	**A3-3**	**A3-23**	**A3-02**	**A3-03**	**A3-0203**
0.7257	**Param-L**	0.7112	0.4590	0.3655	***0.0001***	***0.0000***	***0.0128***	0.3655	***0.0000***	***0.0000***	***0.0000***	0.8050	***0.0000***	0.5237	1.0000	0.9836	0.1746
0.7166	**Param-2**	0.8531	0.3547	0.0613	***0.0000***	***0.0000***	0.0957	0.4466	***0.0000***	***0.0000***	***0.0000***	0.6216	***0.0000***	0.6509	0.8531	0.8855	0.2022
0.9321	**Param-3**	***0.0001***	***0.0001***	***0.0000***	0.8130	***0.0008***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	0.0878	***0.0000***	***0.0001***	***0.0000***	***0.0000***
0.9476	**Param-23**	***0.0000***	***0.0000***	***0.0000***	***0.0237***	***0.0018***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	0.6959	***0.0000***	***0.0000***	***0.0000***	***0.0000***
0.7181	**Nonpar**		0.7112	0.8531	***0.0001***	***0.0000***	0.3235	0.8370	***0.0000***	***0.0000***	***0.0009***	0.6658	***0.0000***	0.6216	0.2757	0.9672	0.2096

Table 6-8 Comparisons (Accuracy and Wilcoxon Signed Rank Test) under Test Case 2

	Accuracy	0.9117	0.9030	0.8822	0.9592	0.8944	0.9604	*0.9607*	0.8560	0.8618	0.2204	0.8556	0.2574	0.3270	0.3411	0.8626	0.9359	0.7673	0.7231	0.1745	0.8428	0.8606	0.8764	0.8159	0.7371
Accuracy	p-value (Wilcoxon Signed Rank Test)	**Nonpar**	**A1-L**	**A1-2**	**A1-3**	**A1-5**	**A1-23**	**A1-235**	**A2-L**	**A2-2**	**A2-3**	**A2-5**	**A2-23**	**A2-235**	**A3-L**	**A3-2**	**A3-3**	**A3-5**	**A3-23**	**A3-235**	**A3-02**	**A3-03**	**A3-05**	**A3-0203**	**A3-020305**
0.8549	Param-L	***0.0005***	***0.0031***	0.0292	***0.0000***	***0.0037***	***0.0000***	***0.0000***	0.2096	0.5787	***0.0000***	1.0000	***0.0000***	***0.0000***	***0.0000***	0.7577	***0.0003***	***0.0066***	***0.0055***	***0.0000***	0.5372	0.5928	0.3038	0.3135	***0.0058***
0.8613	Param-2	***0.0055***	***0.0075***	0.1558	***0.0001***	***0.0058***	***0.0000***	***0.0000***	0.8531	0.0999	***0.0000***	0.4590	***0.0000***	***0.0000***	***0.0000***	0.9672	***0.0010***	***0.0066***	***0.0055***	***0.0000***	0.3038	0.9018	0.5509	0.2172	***0.0009***
0.9445	Param-3	0.0840	***0.0161***	***0.0004***	0.0378	***0.0237***	***0.0190***	***0.0152***	***0.0000***	***0.0003***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	0.2329	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0005***	***0.0002***	***0.0001***	***0.0000***
0.8541	Param-5	***0.0003***	***0.0023***	0.0585	***0.0000***	***0.0011***	***0.0000***	***0.0000***	0.7112	0.3991	***0.0000***	0.1086	***0.0000***	***0.0000***	***0.0000***	0.7266	***0.0000***	***0.0037***	***0.0045***	***0.0000***	0.4716	0.7421	0.2096	0.3337	***0.0037***
0.9485	Param-23	0.0359	***0.0055***	***0.0002***	0.1229	***0.0017***	0.1332	0.1332	***0.0000***	***0.0001***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	0.1746	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0005***	***0.0000***	***0.0001***	***0.0000***
0.9374	Param-235	0.3135	***0.0613***	***0.0029***	***0.0033***	0.0672	***0.0010***	***0.0004***	***0.0001***	***0.0008***	***0.0000***	***0.0001***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	0.8693	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0011***	***0.0006***	***0.0004***	***0.0000***
0.9117	Nonpar		0.5928	***0.0144***	***0.0003***	0.0878	***0.0001***	***0.0001***	***0.0006***	***0.0058***	***0.0000***	***0.0004***	***0.0000***	***0.0000***	***0.0000***	***0.0033***	0.3038	***0.0000***	***0.0001***	***0.0000***	***0.0012***	***0.0085***	0.1086	***0.0040***	***0.0001***

Table 6-9 Comparisons (Accuracy and Wilcoxon Signed Rank Test) under Test Case 3

	Accuracy	0.7804	0.7746	0.7653	0.9338	0.7697	0.9499	0.9392	0.7649	0.7653	0.6202	0.7700	0.6759	0.5764	0.7269	0.7668	*0.9578*	0.7247	0.7970	0.6955	0.7773	0.7493	0.7382	0.7849	0.7207
Accuracy	p-value (Wilcoxon Signed Rank Test)	**Nonpar**	**A1-L**	**A1-2**	**A1-3**	**A1-5**	**A1-23**	**A1-235**	**A2--L**	**A2-2**	**A2-3**	**A2-5**	**A2-23**	**A2-235**	**A3-L**	**A3-2**	**A3-3**	**A3-5**	**A3-23**	**A3-235**	**A3-02**	**A3-03**	**A3-05**	**A3-0203**	**A3-020305**
0.7746	**Param-L**	0.4590	0.9672	0.6959	***0.0002***	0.0585	***0.0000***	***0.0001***	0.6071	0.4590	***0.0003***	0.0417	***0.0090***	***0.0001***	0.2022	0.7892	***0.0000***	0.2757	0.5647	0.0735	0.6509	0.5372	0.8210	0.3991	0.2172
0.7653	**Param-2**	0.3991	0.6959	0.4107	***0.0001***	0.9836	***0.0000***	***0.0001***	0.8531	0.8210	***0.0007***	0.9508	***0.0136***	***0.0002***	0.3038	0.8050	***0.0000***	0.4843	0.4843	0.1132	0.6071	0.5372	0.9508	0.3991	0.3038
0.9338	**Param-3**	***0.0007***	***0.0002***	***0.0001***	0.6216	***0.0002***	0.1229	0.3038	***0.0002***	***0.0000***	***0.0000***	***0.0002***	***0.0000***	***0.0000***	***0.0000***	***0.0001***	***0.0114***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0001***	***0.0001***	***0.0001***
0.7697	**Param-5**	0.3765	0.0585	1.0000	***0.0002***	0.9672	***0.0000***	***0.0002***	0.8210	0.7734	***0.0005***	0.9508	***0.0136***	***0.0001***	0.2410	0.7266	***0.0000***	0.3877	0.3991	0.0999	0.5787	0.7266	0.9508	0.3655	0.2249
0.9343	**Param-23**	***0.0005***	***0.0002***	***0.0001***	0.8370	***0.0001***	0.0917	0.4716	***0.0002***	***0.0000***	***0.0000***	***0.0001***	***0.0000***	***0.0000***	***0.0000***	***0.0001***	***0.0152***	***0.0000***	***0.0000***	***0.0000***	***0.0001***	***0.0000***	***0.0001***	***0.0001***	***0.0000***
0.9392	**Param-235**	***0.0005***	***0.0001***	***0.0001***	0.3038	***0.0002***	0.2942	0.7577	***0.0002***	***0.0001***	***0.0000***	***0.0001***	***0.0000***	***0.0000***	***0.0000***	***0.0001***	0.2579	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0001***	***0.0001***	***0.0000***
0.7804	**Nonpar**		0.4590	0.3991	***0.0007***	0.3765	***0.0003***	***0.0005***	0.6071	0.4716	***0.0015***	0.3547	***0.0071***	***0.0008***	0.1812	0.3991	***0.0001***	***0.0096***	0.9672	0.0769	0.5104	0.2329	***0.0190***	0.9018	0.1950

Note for Table 6-7 through Table 6-9:

1. The shade grid in the first row means the highest accuracy; 2. The shade grids except in the first row denote p-value <0.025 (significant different) via Wilcoxon Signed Rank Test.

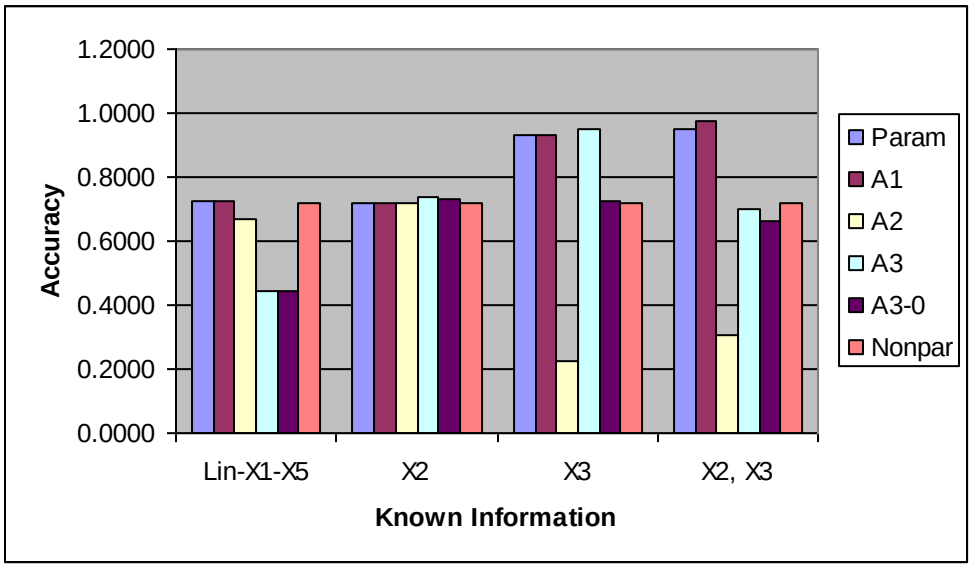


Figure 6-1 Comparisons under Test Case 1

Note:
1. The abbreviated name of each column is listed as the following table.

Series Name	Lin-X1-X5	X2	X3	X2, X3
Param	Param-L	Param-2	Param-3	Param-23
A1	A1-L	A1-2	A1-3	A1-23
A2	A2-L	A2-2	A2-3	A2-23
A3	A3-L	A3-2	A3-3	A3-23
A3-0	A3-L	A3-02	A3-03	A3-0203
Nonpar	Nonpar	Nonpar	Nonpar	Nonpar

2. "Nonpar" means the nonparametric approach based on SVR.

6.3.3.2 Comparison between Semiparametric Algorithms Based on SVR and Nonparametric Approach Based on SVR

As mentioned previously, it is also not wise to ignore the knowledge and only use a pure nonparametric approach based on SVR to estimate cost. From the Table 6-7 through Table 6-9 and Figure 6-1 through 6-3, appropriate semiparametric algorithms with known function forms on some cost drivers are significantly better than a corresponding pure nonparametric approach.

Under the test case 1, semiparametric A1 is close to or significantly better than its corresponding pure nonparametric approach. With known function form on "x_2", the performance of semiparametric A1 is not significantly different with that of the pure nonparametric approach. However, with known function forms on "x_3" or "x_2 and x_3",

semiparametric A1 attained estimating accuracy 93.21% and 97.57%, which respectively are significantly better than that of the corresponding nonparametric approach (accuracy: 71.81%). Semiparametric A3 is close to or significantly better than the corresponding pure nonparametric approach. The known function forms on "x_2" or "x_2 and x_3" for semiparametric A3 did not bring any significant improvement comparing to the nonparametric approach. Moreover, with known function form on "x_3", A3 has significant improvement, whose accuracy is 94.93%. Semiparametric A2 did not show any improvement on performance with known information.

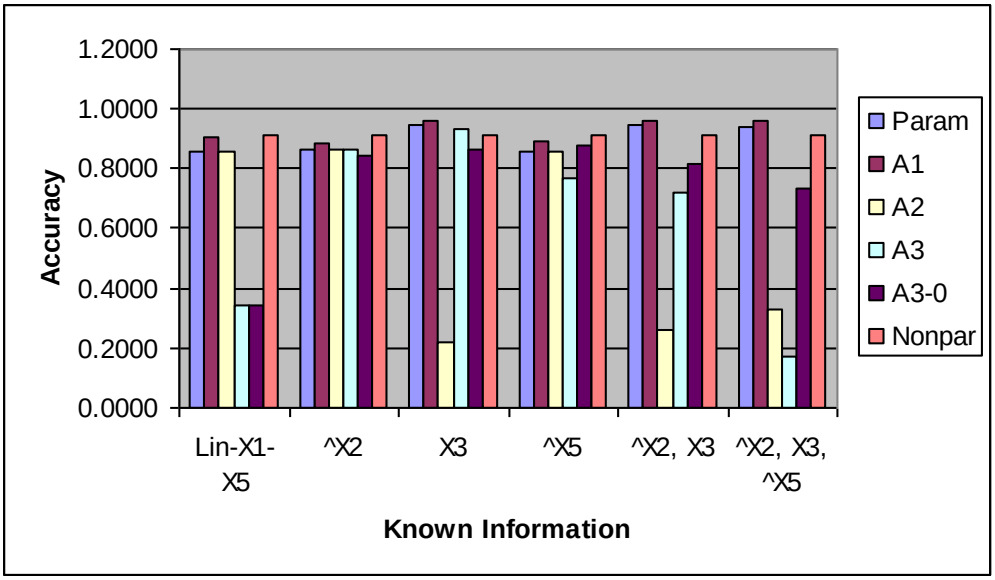


Figure 6-2 Comparisons under Test Case 2

Note:

1. The abbreviated name of each column is listed as the following table.

Series Name	**Lin-X1-X5**	**^X2**	**X3**	**^X5**	**^X2, X3**	**^X2, X3, ^X5**
Param	Param-L	Param-2	Param-3	Param-3	Param-23	Param-235
A1	A1-L	A1-2	A1-3	A1-3	A1-23	A1-235
A2	A2-L	A2-2	A2-3	A2-3	A2-23	A2-235
A3	A3-L	A3-2	A3-3	A3-3	A3-23	A3-235
A3-0	A3-L	A3-02	A3-03	A3-03	A3-0203	A3-020305
Nonpar	Nonpar	Nonpar	Nonpar	Nonpar	Nonpar	Nonpar

2. "Nonpar" means the nonparametric approach based on SVR.

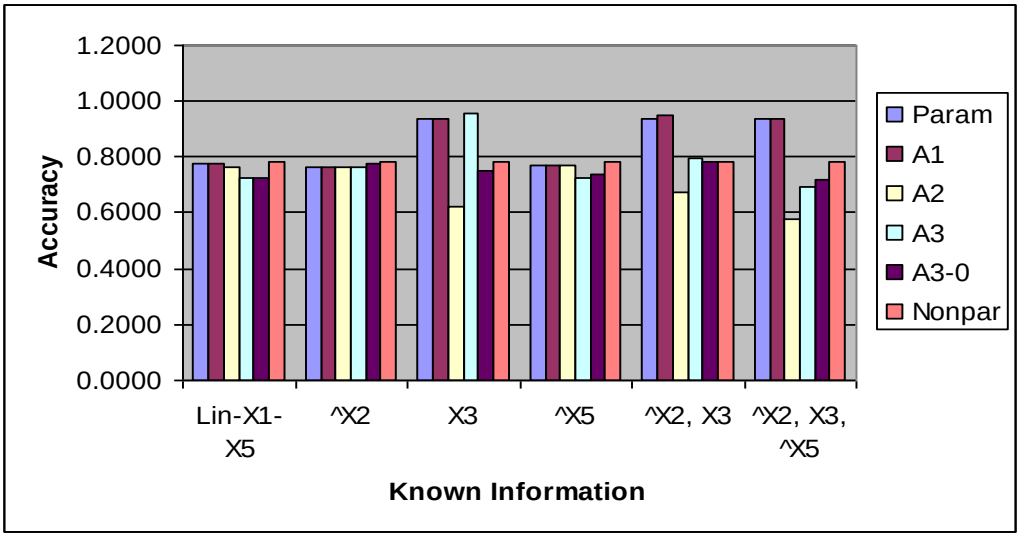


Figure 6-3 Comparisons under Test Case 3

Note: They are same as notes for Figure 6-2.

Under the test case 2, semiparametric A1 is mostly close to or significantly better than its corresponding nonparametric approach. With partial known function form on "x_5", the performance of semiparametric A1 is not significantly different with that of the nonparametric approach. The partial known function form on "x_2" worsened the performance a little bit comparing to the pure nonparametric approach. However, with known function forms on "x_3", "x_2 and x_3", or "x_2, x_3 and x_5", semiparametric A1 attained estimating accuracy 95.92%, 96.04% and 96.07%, which respectively are significantly better than the corresponding nonparametric approach (accuracy: 91.17%). Under most situations semiparametric A3 has worse performance than the pure nonparametric approach does. Moreover the known function forms on "x_3" brought significant improvement comparing to the nonparametric approach. This is also true under test case 1. The semiparametric A2 did not show any improvement on performance with known information.

Under the test case 3, semiparametric A1 is also mostly close to or significantly better than the nonparametric approach. With known function form on "x_2" or "x_5", the performance of semiparametric A1 is not significantly different with that of the pure nonparametric approach. However, with known function forms on "x_3", "x_2 and x_3", or

"x_2, x_3 and x_5", semiparametric A1 attained estimating accuracy 93.38%, 94.99% and 93.32%, which respectively are significantly better than the corresponding nonparametric approach (accuracy: 78.04%). Semiparametric A3 is close to or significantly better than the nonparametric approach. The known function forms on "x_2", "x_5", "x_2 and x_3", or "x_2, x_3 and x_5" for semiparametric A3 did not bring significant improvement comparing to the pure nonparametric approach. However, with known function form on "x_3", A3 has significant improvement, which has best accuracy 95.78%. Semiparametric A2 did not show any improvement on performance with known information.

6.3.3.3 Comparison among Three Semiparametric Algorithms Based on SVR

From the graphs in Figure 6-1 through Figure 6-3, the performance of semiparametric A1 is much better than A2 and A3 with different amount and type of known information. Moreover the performance of A1 is much more stable than that of A2 and A3: even inexact function forms or partial function forms did not significantly worsen the performance of A1 comparing to that of the corresponding pure nonparametric (see A1-L vs. Nonpar in Table 6-7 through Table 6-9). Additionally for A1, there is a trend: more information brings higher accuracy (see A1-2 vs. A1-23, A1-3 vs. A1-23 under the test case 1; or see A1-2 vs. A1-23, A1-3 vs. A1-23, A1-5 vs. A1-23 under the test case 2;). Furthermore, with known function forms of "x_3" A3 had very good performance, whose the estimating accuracy under three test cases respectively attained "94.93%", "93.59%", and "95.78%". They were obviously significantly better than corresponding pure nonparametric approach and pure parametric approach. Semiparametric A2 did not show any improvement under three test cases.

6.3.3.4 The Results with Inexact Function Forms

When there is not knowledge of function forms, the first-order forms are used as the function form of cost drivers. In Table 6-7 through Table 6-9 and Figure 6-1 through Figure 6-3, "A1-L", "A2-L", "A3-L", "A3-02", "A3-03", "A3-05", "A3-0203", and "A3-020305" represent this situation.

These tables and figures show inexact function forms for semiparametric approach cannot bring improvement of performance over the pure parametric approach and nonparametric approach. When the nonparametric approach have good performance for the data as test case 2, A1-L have significant better performance than the corresponding parametric approach. However, A1-L is not significant different with the nonparametric approach.

Additionally, these inexact forms increase the semiparametric model complexity, especially for A2 and A3, and even possibly worse the performance. A3-L under the test case 1, 2 has very bad performance. Under test case 2 (A3-02, A3-03, A3-0203, A3-020305), inexact forms make the performance semiparametric approach significantly worse than that of the corresponding pure nonparametric approach.

6.3.3.5 Discussion

These results suggest that when there may be some knowledge about the parametric form it cannot be wise to ignore the knowledge and only use the pure nonparametric approach. The semiparametric approach would be a good way for cost estimation in this situation. It can improve the estimating accuracy by combining a parametric component, which is based on the researcher's knowledge of the underlying model, with a nonparametric component, which is designed to capture any structure in the data that the parametric fit fails to explain. It can provide noticeable improvements over the two approaches when used individually.

Below are listed some important issues when using the semiparametric approach.

1. Based on Section 6.3.2.1 and Section 6.3.2.2, known exact function forms would certainly bring improvement for estimating accuracy. At least it could not worsen performance for the semiparametric A1. The result can be obtained according to A1-2 vs. Param-2, A1-3 vs. Param-3, and A1-23 vs. Param-23 under three test cases; or A1-5 vs. Param-5, A1-235 vs. Param-235 under the test case 2 and 3. But the semiparametric A2 and A3 cannot guarantee it.
2. Semiparametric A3 may have better performance when it works with a single cost driver with known exact function form. This was also verified by Smola, Frieb, et

al.[94]. Under three test cases, with known function forms of "x_3", A3 bring significant improvement comparing to the pure nonparametric method. But with inexact function form or partial function form of single cost driver, the estimating accuracy is not significantly different with the pure nonparametric approach. This does not conflict with the result presented in the paper [94]. However, semiparametric A3 performance has easily been worsened by model complexity which grows as the number of cost drivers increase, whatever the function forms of these cost drivers are exact or not (see A3-23, A3-235, A3-0203, A3-020305 in Table 6-7 through Table 6-9 and Figure 6-1 through Figure 6-3). Therefore, Semiparametric Algorithm 2 and 3 based on Tabu-SVR for cost estimates is sensitive to model complexity.

3. Semiparametric Approach 2 did not show any improvement with known function form (see A2 under three test cases)
4. Partial known function form mostly cannot solely bring the improvement (see A1-2 or A1-5, A3-2 or A3-5 under the test case 2, 3). Also exact function forms of x_2 associated with step function have not brought any improvement (see A1-2, A3-2 under three test cases) over the corresponding parametric approach (Param-2) and the nonparametric approach under these test cases.
5. Unknown interaction among cost drivers would worsen the estimating accuracy. (see A1-2, A1-5, A3-2, and A3-5 under the test cases 2, 3)

6.4 Conclusions

By way of conclusion, it is expected that semiparametric approach will be used when there is some knowledge about the parametric form but the form is not adequately known throughout the entire range of data or not to reflect true attribute of data. The semiparametric approach would be a good way for cost estimation in this situation. The semiparametric approach is able to combine a parametric component based on the researcher's knowledge of the underlying model, with a nonparametric component designed to capture any structure in the data that the parametric fit fails to explain. In the experiments, three test cases were produced based on five common basic cost

characteristics. All comparisons showed semiparametric A1 had much better and more stable performance than the corresponding parametric approach and the pure nonparametric approach.

Chapter 7 Sensitivity Analysis Based on Support Vector Regression for Cost Control

7.1 Introduction

The cost model for complex products design is not only used to provide accurate cost estimation but also can be used to explain complex, and often non-linear, relationships between input variables and cost behavior. This can easily be used by designers and decision makers in the design stage.

The objectives of sensitivity analysis based on SVR is to:

- determine the contribution of cost drivers;
- answer "what-if" question for cost trade-off study;
- determine the effect of each variables x_i (the absolute maximum change when the value of variable x_i is varied in its allowable range and all other variables are kept at their designated value) on cost, as Figure 7-1;
- establish the profile of each variable x_i when all other variables are kept at the designated set, which mean the outputs at its minimum value, then successively at their first quartile, median, third quartile and maximum of the variable x_i, as Figure 7-2, or outputs at a number of points with equal intervals in the whole range; and

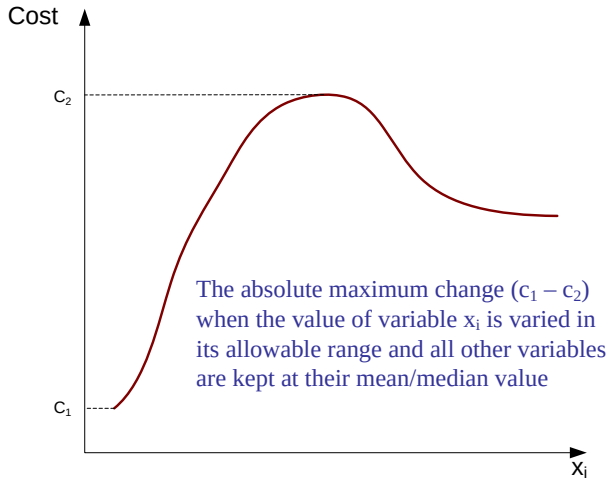


Figure 7-1 The Sensitivity of Variable x_i

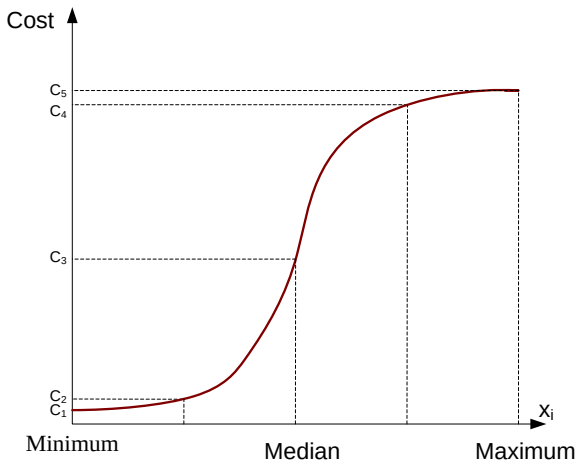


Figure 7-2 The Profile of Variable x_i

- establish a monotonic (non-decreasing or non-increasing) range for a particular variable x_i when other variables are kept at the designated set (see Figure 7-3).

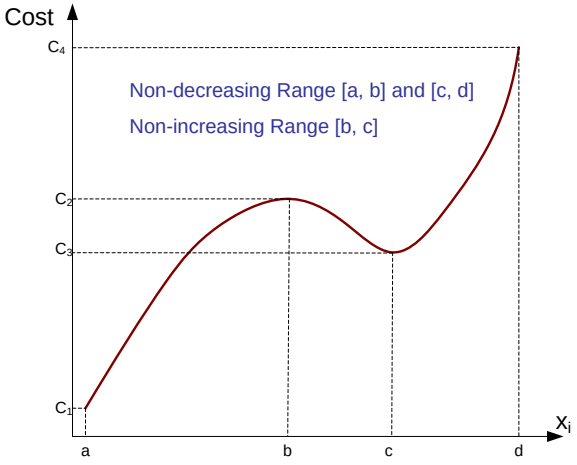


Figure 7-3 The Monotonic Range of Variable x_i

The performance of Tabu-SVR for cost estimation has been discussed in previous chapter. Tabu-SVR approach has a great potential to accurately estimate cost for complex products during the early design phases. However, Tabu-SVR is a nonparametric method, which cannot directly give the explanation of those complex and nonlinear relationships. This section would introduce two existing methods to solve this problem. These two

methods can provide a way to attain the above objectives of sensitivity analysis based on SVR. Finally they can be used by designers and decision in the design stage.

7.2 Methods

Sensitivity analysis is used to study the influence of cost drivers on cost. The method is using cost estimating approach based on SVR as an estimator (simulator). In the different scenarios, the outputs then are organized to obtain the above goals.

To attain the objectives in Section 7.1, there are two methods.

Method 1: A certain number of points with equal interval are produced in the range of the studied cost driver. For example, if five points are produced, they would be minimum, first quartile, median, third quartile and maximum. Other cost drivers are kept as the designated values (see Figure 7-4).

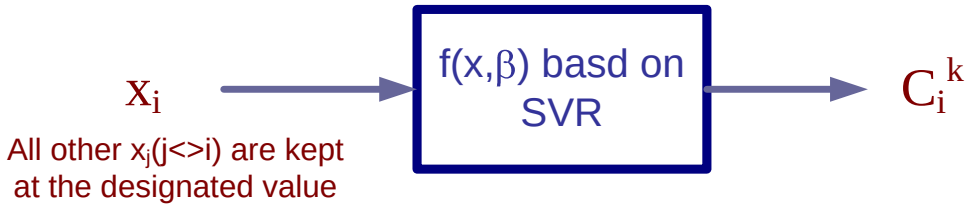


Figure 7-4 The Method 1 for Sensitivity Analysis Based on SVR

In Figure 7-4, the studied cost driver is x_i and the desired number of points is *k*. The value of cost driver x_i is varied in its allowable range and all other cost drivers are kept at their designated value. The maximum and minimum of *k* points would then determine the absolute maximum change. When *k* is greater, the effect of cost driver x_i would be more accurate. The effect of all cost drivers determines the contribution of cost drivers. Also these *k* points form the profile of x_i when all other cost drivers are kept at the designated set. Furthermore, these *k* points can establish a monotonic (non-decreasing or non-increasing) range for a particular variable x_i when other variables are kept at the designated set.

Method 2: The cost estimating approach (nonparametric approach and semiparametric approach) based on SVR adjusts the input values of one variable while keeping all the others untouched (see Figure 7-5). The changed costs against each change

in the cost drivers are noted. The cost driver whose changes affect the cost most is the one that has the most relative influence. These changes can take the form of $x_i=x_i+\delta$ where x_i is the selected input variable and δ is the change. The variable δ can be increased and decreased in designated percentage of the input value in the allowable range.

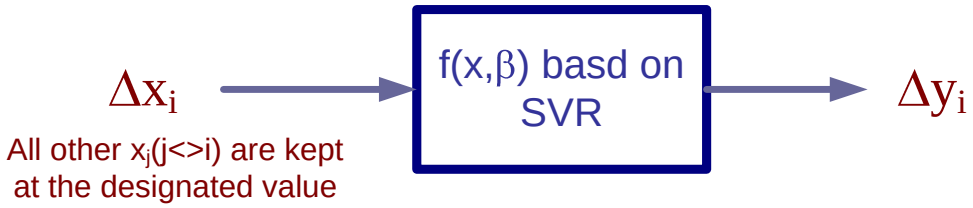


Figure 7-5 The Method 2 for Sensitivity Analysis Based on SVR

The methods can easily help conduct trade-off study to answer "what-if" question. When other cost drivers are kept at the designated values, small intentioned change on the studied cost driver would bring the cost change and direct the designers. When small change happenes at the sequenced points on the studied cost driver as other cost drivers keep untouched, MSE for this studied cost driver can be calculated. After this kind of MSE of all cost drivers, the contribution of cost drivers can be compared.

7.3 Numerical Example

7.3.1 Data Description

The training data used in this section were produced by FLOPS cost module as previous Section 5.2.2. Five input variables were chosen for cost drivers. They are *WTS_25*, *NENG*, *THRMAX*, *SMACH*, *QMAX*. *WTS_25* is total weight of engines. It is assumed between 10000 lbs and 70000lbs. *NENG* is the number of engines per aircraft. In this example, *NENG* is set 2 or 4. *THRMAX* is the maximum thrust per engine ranging from 20000 lbs to 90000 lbs. *SMACH* is the maximum Mach number at best altitude and ranges between 0.7 Mach and 1 Mach. *QMAX* is the maximum dynamic pressure ranging from 200 lb/ft^2 to 600 lb/ft^2. After training, the nonparametric model based on SVR was

constructed. Then the sensitivity analysis based on SVR for these five cost drivers can be preformed via the above two methods.

7.3.2 Results and Discussion

The cost driver, NENG, only has two values: 2 and 4. For method 1, the rest of cost drivers were divided by equal intervals to produce 20 points. For each cost drivers, the cost estimate was calculated while the other cost drivers were kept at the median value. The resulting problems are displayed in Figure 7-6 and Figure 7-7.

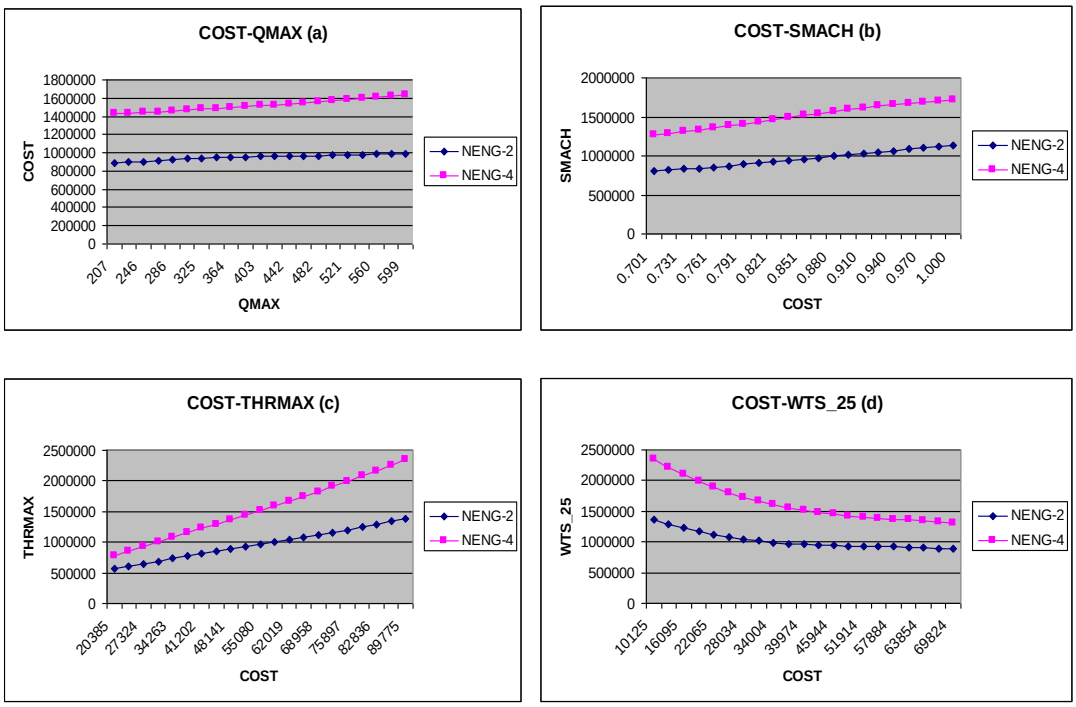


Figure 7-6 The Profiles of Four Cost Drivers: (a) QMAX; (b) SMACH; (c) THRMAX; (d) WTS_25

In Figure 7-6 has shown all cost drivers are monotonic for the range of interest. Except the cost is decreasing as WTS_25 drops, the costs are increasing as all other cost drivers: NENG, QMAX, SMACH, and THRMAX, rise. Figure 7-7 indicates the influence of all cost drivers. For NENG=2, THRMAX is most significant cost driver

because its range is biggest. The second one is WTS_25. The least important cost driver is QMAX, whose range is smallest. For NENG=4, there are same results as NENG=2.

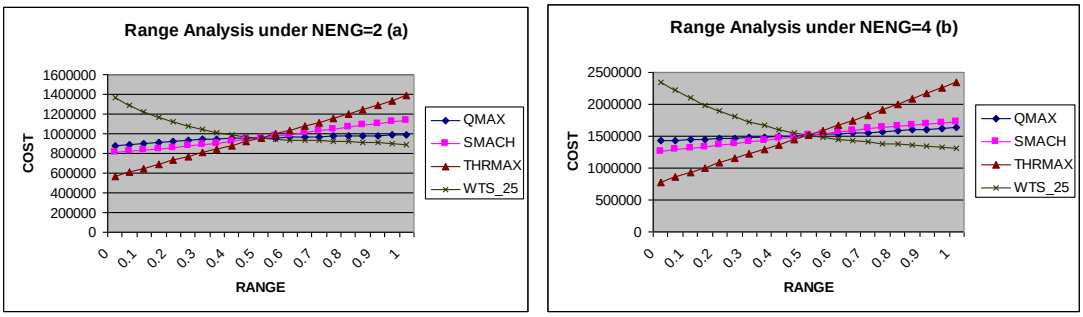


Figure 7-7 Contribution of the Cost Drivers by Method 1: (a) NENG=2; (b) NENG=4

For method 2, these 20 points were used as sequence points. Small change, 5%, was chosen to add to the sequence points. The MSE of output before and after can reflect the influence of cost drivers as Table 7-1. It has shown when NENG =2 THRMAX has biggest MSE, 3.43E+10. The cost driver THRMAX is most significant. The second significant cost driver is WTS_25. QMAX is the least important. The results are same as NENG =4.

Therefore, method 1 and method 2 showed the same results. In this example, THRMAX has most significant impact on cost. The following sequence of importance is: WTS_25, SMACH, and QMAX.

Table 7-1 MSE of Cost Drivers When Small Changes (5%) Are Added and Other Cost Drivers Are Kept at the Median

MSE	QMAX	SMACH	THRMAX	WTS_25
NENG=2	7.26E+08	5.46E+09	3.43E+10	2.06E+10
NENG=4	2.26E+09	1.05E+10	1.24E+11	8.19E+10

In this example, all cost drivers are monotonic. It hence is unnecessary to do monotonic analysis.

7.4 Conclusions

After the cost was predicted accurately, knowing the contribution of each cost driver is also very important. Method 1 and Method 2 presented in this chapter for sensitivity analysis would improve the explanation capability of cost estimating approach based on SVR. Also they can help designer and decision-maker to perform trade-off study during the design phases.

This chapter illustrates how to use method 1 and method 2 to make sensitivity analysis based on SVR for the aircraft engine cost. In the numeric example, the method 1 and method 2 showed the contribution of cost drivers and established the profiles of cost drivers. These results show these two methods could easily help designer make trade-off study and answer “what-if” questions.

Chapter 8 Conclusions and Future Research

During the design phases, designers and decision-makers often need to know accurate cost information to assess and compare multiple alternatives and to determine preferred design. They need to identify cost reduction opportunities and tradeoffs to meet aggressive targets (requirement, performance, and schedule). They also need to evaluate cost reduction ideas and alternatives affecting system performance factors for their impact and compare the results with the original "baseline" design. Therefore, a cost estimating model must be reasonably accurate, robust, and capable of operating on data of the detail typically available in the related phase, to support cost trade-off studies for designers and decision makers.

This study first focused on identifying and selecting cost drivers, and then on nonparametric and semiparametric cost approach based on support vector regression to improve cost estimating accuracy. The methods of sensitivity analysis were introduced to determine the contributions and profiles of cost drivers, which can work to support cost tradeoffs by designers and decision makers. The study is concluded as following sections.

8.1 Conclusions

A generic cost model was first presented in Chapter 2. It consists of three components: output (cost), $f(x;\beta)$, and an input space x. Based on the generic model, there are two types of cost estimating approaches: approaches based on functional relationship and approaches based on inputs and structural relationship. The approaches based on the inputs and structural relationship (feature-based cost approach, activity-based cost estimating approach, process-based cost approach and simulation) are often applied in the preliminary or later phase and respectively do a good job under a certain situation and scenario. The approaches based on functional relationship are building blocks for the approaches based on structural relationship and inputs. The performance based on inputs and structural relationship depends on the identification of inputs, structure and the approaches based on functional relationship. If there is not a good approach based on

functional relationship, there would be no accurate estimation. The approach based on functional relationship is corner stones for cost estimation.

Based on functional relationship, the approaches were classified into: expert judgment, parametric method, neural network approach and case-based reasoning approach. The significant drawback of expert judgment is its subjective nature. Parametric method needs apriori knowledge of the functional form and it is very difficult to deal with nonlinearity and discontinuity. Case-based reasoning method is hard to define a similarity measure and adjusting methods and thus cannot guarantee the accuracy of estimating cost. For artificial neural network approach, it is hard to produce near optimal neural network models and overcome over fitting problems when there are lots of historical data. The artificial neural network approach was thought to lack explanation capabilities.

Cost estimation has always been difficult at the early stage of product development when only a few conceptual attributes of the product are known for complex product. The relationship between these attributes and cost is very hard to obtain. And the discontinuity and nonlinearity may often exist in these relationships. Therefore, this study focused on a new way to identify and select cost drivers and to estimate cost based on support vector regression, which can be applied in the entire life cycle, especially for complex products during the early design phases. And the methods of sensitivity analysis could make it overcome the “black box” problem and be able to provide guide to designers.

Chapter 3 presented the research framework of this study. This study was composed of four parts:

- identifying cost drivers via Causal-Associated (CA) method and eliminating the insignificant cost drivers using Tabu-Stepwise method;
- estimating cost via Tabu-SVR, a nonparametric approach based on support vector regression with a tabu search algorithm;
- estimating cost using semiparametric approach based on support vector regression (SVR); and

- indicating the effect of cost drivers on cost for cost modeling based on SVR via sensitivity analysis.

In Chapter 4, a new method, Causal-Associated (CA) approach, was proposed to identify the cost drivers, which is different with traditional methods for identifying cost drivers. CA approach would bring more complete and correct understanding and explanation to the cost analysis, and help avoid missing some cost drivers. It therefore results in an improved predictive capacity. After that, Tabu-Stepwise selection technique was presented to eliminate the unrelated or insignificant cost drivers under nonlinear situation and reduce the variance in the model output and the cost of collecting the data.

A case study in Chapter 4 was employed to illustrate the feasibility, the procedure of Causal-Associated method and Tabu-Stepwise method. From this case study, it was seen that the Causal-Associated method helps avoid missing some cost drivers. When using associated cost drivers to represent some root cost drivers, the assumptions and preconditions were easily obtained. After that, Tabu-Stepwise selecting method based on Tabu-SVR was used to select the cost drivers. The test data showed it improved the accuracy of the cost estimating prediction by eliminating irrelevant variables, also reduced expenditure of the collection, storage, and computation load in the process of cost estimation.

A nonparametric approach based on support vector regression was introduced in Chapter 5. It includes three subparts: the methodology of Tabu-SVR, test cases and experiments. The procedure of Tabu-SVR, a nonparametric cost estimating approach based on SVR, was presented. The tabu search algorithm for choosing parameters of SVR was proposed. For validation and verification of performance on Tabu-SVR, the test cases were generated. The test cases include simulated data sets and pilot data set. The simulated data sets were produced based on five summarized basic common cost characteristics: accumulation; linear function; power function; step function; and exponential function. The pilot data set was produced by the FLOPS cost module (aircraft engine part). The cost models were constructed respectively based on SVR, parametric method, neural networks and case-based reasoning. The cost estimating approach based on SVR was studied. The performance of all methods was compared.

From the results, Tabu-SVR significantly improved the performance comparing to SVR based on empirical study via choosing appropriate parameters. The RBF and polynomial kernels show better performance over linear kernel under most data sets. Moreover, RBF kernel is much more robust. This means RBF kernel is less dependent on the sample data used. When an a apriori CER (functional form) is unknown, Tabu-SVR cost modeling yielded good performance over other cost modeling techniques: parametric method, case-based reasoning, and neural networking. The Tabu-SVR was able to capture these nonlinearities and discontinuities, along with interactions among cost drivers, had strong predicable capability. Especially, when the cost data does not allow to be discerned the appropriate CER because no enough knowledge is obtained or finding appropriate function forms become more complex as the dimensionality of cost drivers grows. Therefore, the cost model based on Tabu-SVR has a great potential to accurately estimate cost for complex products during the early design phases.

In Chapter 6, the focus is on semiparametric cost estimating approach based on support vector regression (SVR). After presenting three semiparametric algorithms based on SVR, three data sets based on common basic characteristics were produced. The experiments showed that Semiparametric Algorithm 1 is the best approach under most situations and Algorithm 3 might have better performance under some situation. It often had better performance over the pure nonparametric approach and the pure parametric approach. The model complexity would influence the estimating accuracy for Algorithm 2 and Algorithm 3. The inexact function forms of some cost drivers would not bring the improvement of cost estimating accuracy and even worsen the performance.

Sensitivity analysis for cost modeling was discussed in Chapter 7. Two existing methods introduced in this chapter for sensitivity analysis would improve the explanation capability of cost estimating approach based on SVR. Also they can help designer and decision-maker to perform trade-off study during their design phases. This Chapter illustrates how to use method 1 and method 2 to make sensitivity analysis based on SVR for aircraft engine cost. The numeric example showed the method 1 and method 2 were able to determine the contribution of cost drivers and establish the profiles of cost drivers. Therefore, the method 1 and method 2 could easily help designers make trade-off study and answer “what-if” questions.

8.2 Future Research

Identifying and selecting cost drivers are very important to cost estimation. The CA method proposed in this study is a good way to identify the cost drivers. Tabu-Stepwise method based on Tabu-SVR can eliminate the insignificant and irrelevant cost drivers under nonlinear situation. But it has two drawbacks: first it is very time consuming; second it could not guarantee the best candidate sets. Future research could focus on how to reducing irrelevant cost drivers with less time and finding the better candidate sets as possible.

Tabu-SVR has better performance than SVR based on empirical study and other traditional methods: parametric method, case-based reasoning, and neural networking under most situations. From this study, appropriate parameters and kernel would significantly impact on the performance of cost estimation. The tabu search algorithm was proposed to solve parameters choosing problems. It is still time consuming to choose appropriate parameters to construct a cost model based on SVR. Therefore, choosing appropriate kernel and parameters for SVR can still be studied further in the future.

This study showed inexact and incomplete function forms would not bring any benefit for estimating performance and even sometimes worsen the performance. It is still worth studying how to use incomplete function forms to improve the performance. Additionally, the influence of interaction of cost drivers to performance in semiparametric approach could be a topic of future research.

Appendices

Appendix A: Software Development

In this study, the Tabu-Stepwise selection method in Chapter 4, the Tabu-SVR, and Semiparametric Algorithm 1, 2, 3 were implemented by the software developed by the author. The software was developed in Microsoft Visual C++ 2003 under Windows XP. The main framework of software is as Figure 5-9.

The procedures and methods of Tabu-Stepwise, Tabu-SVR, and Semiparametric Algorithm 1, 2, 3 have been discussed in Chapter 4, 5, and 6. The data structure of software was designed by referencing LIBSVM [96] and SVM^{Light} [97]. The optimization module was developed based on ILOG CPLEX 9.0.

In order to demonstrate the validity of the software, dummy data sets were created to test each module in the software. The middle outputs of matrix calculation realized by software were verified by being compared with the results of Matlab. A couple of functions were created to test the tabu search algorithm for choosing parameters for SVR, such as:

$$y = a_0 + (x_1 - a_1)^{\beta_1} + (x_2 - a_2)^{\beta_2} + (x_3 - a_3)^{\beta_3} + (x_4 - a_4)^{\beta_4} + (x_5 - a_5)^{\beta_5}$$

where a_i, i=0, 1, 2 ,3 ,4 ,5; β_j, j=1, 2, 3, 4, 5; a_i and β_j are constant. x_j is values searched by the tabu search algorithm.

Appendix B: The Partial Output of Software (Tabu-Stepwise Method)

The method and analysis about result of Tabu-Stepwise see Section 4.2.2 and Section 4.3.4.

Start from the Result of Adjusted R-square (15)

OS time: 23:57:30
OS date: 03/08/07

Current OS time: 00:42:01
Var:100111111110111 *Index:30713 *gamma:0.052810875 *C:7.7668482e+008 *e:567.72013 Value:1175042927.758595
Current OS time: 01:25:44
Var:000111111110111 *Index:30712 *gamma:0.023796549 *C:6.3782715e+008 *e:1644.4411 Value:1374664340.997489
Current OS time: 08:10:24
Var:110111111110111 *Index:30715 *gamma:0.010688677 *C:6.2225299e+008 *e:425.6524 Value:1552956573.694922
Current OS time: 09:30:08

Var:111011111110111 *Index:30711 *gamma:0.02239864 *C:2.5133743e+008 *e:6553.8658 Value:1490856550.715772
Current OS time: 10:14:14
Var:111101111110111 *Index:30703 *gamma:0.021285461 *C:4.0030664e+008 *e:1396.9057 Value:1195978998.023215
Current OS time: 11:33:17
Var:111110111110111 *Index:30687 *gamma:0.012904201 *C:9.9607283e+008 *e:4678.8819 Value:1245110090.179428
Current OS time: 12:33:08
Var:111111011110111 *Index:30655 *gamma:0.020879212 *C:5.9978692e+008 *e:446.63711 Value:1330770859.65468
Current OS time: 13:18:42
Var:111111101110111 *Index:30591 *gamma:0.0055869241 *C:4.9638682e+008 *e:260.61523 Value:1392342933.541083
Current OS time: 14:35:51
Var:111111110110111 *Index:30463 *gamma:0.008514491 *C:2.2421014e+008 *e:3904.2058 Value:1506839501.653724
Current OS time: 15:19:48
Var:111111111010111 *Index:30207 *gamma:0.0064658684 *C:5.8921178e+008 *e:54.95376 Value:1583886532.58673
Current OS time: 16:10:48
Var:111111111100111 *Index:29695 *gamma:0.018772303 *C:2.7411525e+008 *e:73.349337 Value:1494388142.995594
Current OS time: 17:47:50
Var:111111111110111 *Index:30719 *gamma:0.018574108 *C:5.0550009e+008 *e:23.574596 Value:1343286945.438213
Current OS time: 18:35:24
Var:111111111111011 *Index:28671 *gamma:0.024208341 *C:3.3508086e+008 *e:38.647055 Value:**761342910.2086661**
Current OS time: 18:19:45
Current OS date: 03/10/07

Start from the Result of Cp (15)

OS time: 00:26:21
OS date: 03/09/07

Current OS time: 02:42:50
Var:110001111110011 *Index:26595 *gamma:0.044072428 *C:9.7811969e+008 *e:716.03517 Value:**579987291.0513141**
Current OS time: 08:34:06
Var:010001111110011 *Index:26594 *gamma:0.03020724 *C:6.6311213e+008 *e:270.93373 Value:693538760.5838952
Current OS time: 09:59:59
Var:100001111110011 *Index:26593 *gamma:0.05274831 *C:9.5668995e+008 *e:98.415288 Value:756624098.3907197
Current OS time: 12:54:44
Var:111001111110011 *Index:26599 *gamma:0.029657512 *C:9.3807308e+008 *e:42.097124 Value:763940017.6385477
Current OS time: 15:19:00
Var:111101111110011 *Index:26607 *gamma:0.050607916 *C:9.6131403e+008 *e:250.93535 Value:732529223.8408792
Current OS time: 16:40:12
Var:111110111110011 *Index:26591 *gamma:0.011577547 *C:7.0424173e+008 *e:1562.5843 Value:854630423.9268196
Current OS time: 17:31:05
Var:111111011110011 *Index:26559 *gamma:0.011212986 *C:7.5999129e+008 *e:93.500146 Value:864496281.4449811
Current OS time: 19:16:35
Var:111111101110011 *Index:26495 *gamma:0.012671489 *C:8.0258272e+008 *e:304.63891 Value:983949355.3646636
Current OS time: 20:16:11
Var:111111110110011 *Index:26367 *gamma:0.030613986 *C:9.1647587e+008 *e:22.181785 Value:912192509.8520589
Current OS time: 21:13:01
Var:111111111010011 *Index:26111 *gamma:0.038851661 *C:8.5702377e+008 *e:107.22442 Value:1046047127.069876
Current OS time: 22:02:37
Var:111111111100011 *Index:25599 *gamma:0.03688848 *C:5.8823888e+008 *e:154.1143 Value:652171464.0421363
Current OS time: 09:58:01
Var:111111111110011 *Index:26623 *gamma:0.024645471 *C:5.9611198e+008 *e:420.36258 Value:735977417.6815288
Current OS time: 11:37:56
Var:111111111111011 *Index:28671 *gamma:0.014944798 *C:5.9669329e+008 *e:1026.1681 Value:730285283.6970077
Current OS time: 16:16:13
Current OS date: 03/11/07

Start from First Variable (15)

OS time: 11:40:47
OS date: 03/17/07

Current OS time: 12:23:40
Var:100000000000000 *Index:1 *gamma:4.3879513 *C:7.0583209e+008 *e:10132.145 Value:2003086893.195135
Current OS time: 13:43:54
Var:111000000000000 *Index:7 *gamma:4.7479872 *C:33920638 *e:2459.6004 Value:**250747646.9397545**
Current OS time: 22:40:52
Current OS date: 03/17/07

Start from the Result of Adjusted R-square (20)

OS time: 23:10:35
OS date: 03/16/07

Current OS time: 23:31:52
Var:10010111111011100001 *Index:554985 *gamma:0.018856142 *C:6.2471956e+008 *e:140.15949 Value:1427580085.650467
Current OS time: 23:50:24
Var:00010111111011100001 *Index:554984 *gamma:0.039943672 *C:5.8764228e+008 *e:113.4089 Value:1240999446.779702
Current OS time: 00:52:39
Var:11100111111011100001 *Index:554983 *gamma:0.026018204 *C:5.7362659e+008 *e:523.33981 Value:1313703275.508232
Current OS time: 01:40:09
Var:11111011111011100001 *Index:554975 *gamma:0.014438349 *C:6.494865e+008 *e:4203.5336 Value:1223761224.341701
Current OS time: 04:30:07
Var:11111111111101100001 *Index:552959 *gamma:0.019056498 *C:4.2863589e+008 *e:617.50788 Value:**1178473513.340879**
Current OS time: 06:45:52
Current OS date: 03/17/07

Start from the Result of Cp (20)

OS time: 11:51:38
OS date: 03/17/07

Current OS time: 12:17:43
Var:10011001111001100000 *Index:26521 *gamma:0.036593017 *C:7.4843016e+008 *e:233.55167 Value:661887569.5358445
Current OS time: 12:43:33
Var:00011001111001100000 *Index:26520 *gamma:0.033535666 *C:9.9073074e+008 *e:1823.9649 Value:641909134.0751731
Current OS time: 14:10:16
Var:11110001111001100000 *Index:26511 *gamma:0.021390641 *C:7.0383319e+008 *e:766.69016 Value:**637918642.460373**
Current OS time: 16:06:07
Var:11111111110001100000 *Index:25599 *gamma:0.039823188 *C:6.3686396e+008 *e:558.23428 Value:716098169.7540585
Current OS time: 18:58:48
Current OS date: 03/17/07

Start from the First Variable (20)

OS time: 00:14:57
OS date: 03/18/07

Current OS time: 01:01:33
Var:10000000000000000000 *Index:1 *gamma:4.3879513 *C:7.0583209e+008 *e:10132.145 Value:2003086893.195135
Current OS time: 02:24:01
Var:11100000000000000000 *Index:7 *gamma:4.7479872 *C:33920638 *e:2459.6004 Value:**250747646.9397545**
Current OS time: 16:08:05
Current OS date: 03/18/07

Note:

*gamma, *C, and *e are the parameters of SVR model with RBF kernel. They are γ, C, and ε(see Section 2.5.2 and Section 5.1.2).

Bibliography

1. Blanchard, B.S. and W.J. Fabrycky, 1998, Systems Engineering and Analysis. 3rd ed: Prentice-Hall, Inc.

2. Creese, R.C. and L.T. Moore, 1990, "Cost Modeling for Concurrent Engineering". Cost Engineering, 32(6): p. 23-27.

3. Prince, F.A., 2002, "Why NASA's management doesn't believe the cost estimate". Engineering Management Journal, 14(1): p. 7.

4. Johnson, V.S., 1990, "Minimizing life cycle cost for subsonic commercial aircraft". Journal of Aircraft: p. 139-145.

5. Johnson, V.S., 1989, "Life cycle cost in the conceptual design of subsonic commercial aircraft (volumes 1 and 2) ", Ph.D. Dissertation, University of Kansas

6. McCullers, L.A.A. and NASA Langley Research Center, 1995, "Flight Optimization System (FLOPS)".

7. Harwick, W.T., 1997, "Space economics measurement - A survey of space cost models", in AIAA Defense and Space Programs Conference and Exhibit - Critical Defense and Space Programs for the Future, Huntsville, AL, Sept. 23-25, 1997.

8. Bashir, H.A. and V. Thomson, 1999, "Estimating design complexity". Journal of Engineering Design, 10(3): p. 247-257.

9. PRICE Systems, 2004, "Price H". http://www.pricesystems.com/.

10. Galorath Incorporated, 2004, "SEER H". http://www.galorath.com/.

11. Solverson, R.R., 1993, "Design to cost with PRICE H", in AIAA/AHS/ASEE, Aerospace Design Conference: Irvine, CA.

12. Smith, K., 2002, "NASA/Air Force Cost Model", in SCEA's National Conference: Scottsdale, AZ.

13. Vision Spaceport Synergy Team, 1999, "Spaceport Cost Model Research Report". Command and Control Technologies Corporation. p. 1-25.

14. Marx, W.J., D.N. Mavris, and D.P. Schrage, 1998, "A knowledge-based system integrated with numerical analysis tools for aircraft life-cycle design". AI EDAM

(Artificial Intelligence for Engineering, Design, Analysis and Manufacturing), (12): p. 211-229.

15. Seo, K.K., J.H. Park, D.S. Jang, and D. Wallace, 2002, "Prediction of the life cycle cost using statistical and artificial neural network methods in conceptual product design". International Journal of Computer Integrated Manufacturing, 15(6): p. 541-554.

16. Department of Defense, 1999, Parametric Estimating Handbook. Second Edition ed: (http://www.ispa-cost.org/PEIWeb/newbook.htm).

17. MileHam, A.R., G.C. Currie, A.W. Miles, and D.T. Bradford, 1993, "A Parametric Approach to Cost Estimating at the Conceptual Stage of Design". Journal of Engineering Design, 4(2): p. 117-125.

18. Seo, K., J.H. Park, D.S. Jang, and D. Wallace, 2002, "Approximate estimation of the product life cycle cost using artificial neural networks in conceptual design". International Journal of Advanced Manufacturing Technology, 19(6): p. 461-471.

19. Hughes, R.T., 1996, "Expert judgement as an estimating method". Information and Software Technology, 38(2): p. 67-75.

20. Heemstra, F.J. and R.J. Kusters, 1991, "Function point analysis: evaluation of a software cost estimation model". European Journal of Information Systems, 1(4): p. 229-237.

21. Rush, C. and R. Roy, 2001, "Expert judgement in cost estimating: Modelling the reasoning process". http://citeseer.ist.psu.edu/525669.html.

22. Canada, J.R., W.G. Sullivan, and J.A. White, 1996, Capital Investment Analysis for Engineering and Management: Prentice Hall.

23. Dean, E.B.1989, "Parametric Cost Estimating: A Design Function". in Transactions of the 33rd Annual Meeting of the American Association of Cost Engineers. San Diego CA.

24. Boren, H.E., 1976, "A Computer Model for Estimating Development and Procurement Costs of Aircraft (DAPCA III)". http://www.rand.org/pubs/reports/2007/R1854.pdf.

25. Sengupta, J.K., 2004, "The survivor technique and the cost frontier: A nonparametric approach". International Journal of Production Economics, 87(2): p. 185-193.

26. Lawson, B., 2003, "MAXIM Periscope Module". PRICE H Cost Modeling. Available at http://www.pricesystems.com/index.html.

27. Mosher, T., M. Barrera, and N. Lao.1998, "Integration of Small Satellite Cost and Design Models for Improved Conceptual Design-to-Cost". in Eight Annual International Symposium of the International Council on Systems Engineering. Vancouver, Canada: Available at: http://dutlsisa.lr.tudelft.nl/seinternet/Lectures/PDCpapers/paper04.pdf.

28. Mosher, T.J., N.Y. Lao, E.T. Davalos, and D.A. Bearden, 1999, "A comparison of NEAR actual spacecraft costs with three parametric cost models". IAA International Conference on Low-Cost Planetary Missions, 3rd, Pasadena, CA, Apr. 27-May 1, 1998, Acta Astronautica, 45(4-9): p. 457-464.

29. Duverlie, P. and J.M. Castelain, 1999, "Cost estimation during design step: Parametric method versus case based reasoning method". International Journal of Advanced Manufacturing Technology, 15(12): p. 895-906.

30. Weustink, I.F., E. ten Brinke, A.H. Streppel, and H.J.J. Kals, 2000, "A generic framework for cost estimation and cost control in product design". Journal of Materials Processing Technology, 103(1): p. 141-148.

31. Liao, T.W., Z.M. Zhang, and C.R. Mount, 1998, "Similarity measures for retrieval in case-based reasoning systems". Applied Artificial Intelligence, 12(4): p. 267-288.

32. Herrmann, J.W., S. Balasubramanian, and G. Singh, 2000, "Defining specialized design similarity measures". International Journal of Production Research, 38(15): p. 3603-3621.

33. Daengdej, J., D. Lukose, E. Tsui, P. Beinat, and L. Prophet, 1997, "Combining case-based reasoning and statistical method for proposing solution in RICAD". Knowledge-Based Systems, 10(3): p. 153-159.

34. Rehman, S. and M.D. Guenov, 1998, "A methodology for modelling manufacturing costs at conceptual design". Computers & Industrial Engineering, 35(3-4): p. 623-626.

35. Brinke, T.E., E. Lutters, T. Streppel, and H.J.J. Kals, 2000, "Variant-based cost estimation based on Information Management". International Journal of Production Research, 38(17): p. 4467-4479.

36. ten Brinke, E., E. Lutters, T. Streppel, and H. Kals, 2004, "Cost estimation architecture for integrated cost control based on information management". International Journal of Computer Integrated Manufacturing, 17(6): p. 534-545.

37. El-Mehalawi, M.E.L.S., 1999, "A Geometric Similarity Case-Based Reasoning System for Cost Estimation in Net-Shape Manufacturing", Ph.D. Dissertation, Ohio State University.

38. Finnie, G.R., G.E. Wittig, and J.M. Desharnais, 1997, "A comparison of software effort estimation techniques: Using function points with neural networks, case-based reasoning and regression models″. Journal of Systems and Software, 39(3): p. 281-289.

39. Shepperd, M. and C. Schofield, 1997, "Estimating Software Project Effort Using Analogies″. IEEE Transactions on Software Engineering, 23(12): p. 736-743.

40. Delany, S.J. and P. Cunningham, 2000, "The Application of Case-Based Reasoning to Early Software Project Cost Estimation and Risk Assessment". (http://www.cs.tcd.ie/publications/tech-reports/reports.00/TCD-CS-2000-10.pdf).

41. Idri, A., A. Abran, and T.M. Khoshgoftaar, 2004, "Fuzzy Case-Based Reasoning Models for Software Cost Estimation". Springer-Verlag (http://www.lrgl.uqam.ca/publications/pdf/803.pdf).

42. Khanna, T., 1990, Foundations of Neural Networks: Addison-Wesley, Reading, MA.

43. Funahashi, K.-I., 1989, "On the approximate realization of continuous mappings by neural networks″. Neural Networks, 2(3): p. 183-192.

44. Hornik, K., M. Stinchcombe, and H. White, 1989, "Multilayer feedforward networks are universal approximators″. Neural Networks, 2(5): p. 359-366.

45. Zhang, Y.F. and J.Y.H. Fuh, 1998, "A neural network approach for early cost estimation of packaging products″. Computers & Industrial Engineering, 34(2): p. 433-450.

46. Bode, J., 2000, "Neural networks for cost estimation: simulations and pilot application″. International Journal of Production Research, 38(6): p. 1231-1254.

47. Bode, J., 1998, "Decision support with neural networks in the management of research and development: Concepts and application to cost estimation″. Information & Management, 34(1): p. 33-40.

48. Smith, A.E. and A.K. Mason, 1997, "Cost estimation predictive modeling: Regression versus neural network″. The Engineering Economist, 42(2): p. 137-161.

49. Shah, J.J., 1991, "Assessment of Features Technology". Computer-Aided Design, 23(5): p. 331-343.

50. Feng, C.X., A. Kusiak, and C.C. Huang, 1996, "Cost evaluation in design with form features". Computer-Aided Design, 28(11): p. 879-885.

51. Jung, J.-Y., 2002, "Manufacturing cost estimation for machined parts based on manufacturing features". Journal of Intelligent Manufacturing, 13(4): p. 227.

52. OuYang, C. and T.S. Lin, 1997, "Developing an integrated framework for feature-based early manufacturing cost estimation". International Journal of Advanced Manufacturing Technology, 13(9): p. 618-629.

53. Haffner, S.M., 2002, "Cost Modeling and Design for Manufacturing Guidelines for Advanced Composite Fabrication", Ph.D. Dissertation, Massachusetts Institute of Technology.

54. Mabson, G.E., et al., 1996, "Cost Optimization Software for Transport Aircraft Design Evaluation (COSTADE)". National Aeronautics and Space Administration, Langley Research Center, Hampton, Va.

55. Ilcewicz, L.B., et al., 1996, "Cost Optimization Software for Transport Aircraft Design Evaluation (COSTADE)". National Aeronautics and Space Administration, Langley Research Center, Hampton, Va.

56. Park, C.S. and G.-T. Kim, 1995, "An economic evaluation model for advanced manufacturing systems using activity-based costing". Journal of Manufacturing Systems, 14(6): p. 439.

57. Ong, N.S., 1995, "Manufacturing Cost Estimation for Pcb Assembly - an Activity-Based Approach". International Journal of Production Economics, 38(2-3): p. 159-172.

58. Ben-Arieh, D. and L. Qian, 2003, "Activity-based cost management for design and development stage". International Journal of Production Economics, 83(2): p. 169-183.

59. Velcu, O., 2002, "Practical Aspects in the Implementation of an ABC Model", M. Sc Thesis, Swedish School of Economics and Business Administration.

60. Ozbayrak, M., M. Akgun, and A.K. Turker, 2004, "Activity-based cost estimation in a push/pull advanced manufacturing system". International Journal of Production Economics, 87(1): p. 49-65.

61. Steele, M.J. and D. Cope.2004, "An Approach to Estimating Reusable Launch Vehicle Cost with Simulation". in NASA Cost Symposium (March 9 – 11, 2004).

62. Asiedu, Y., R.W. Besant, and P. Gu, 2000, "Simulation-based cost estimation under economic uncertainty using kernel estimators". International Journal of Production Research, 38(9): p. 2023-2035.

63. Forrester, J.W., 2003, "Dynamic models of economic systems and industrial organizations". System Dynamics Review, 19(4): p. 329.

64. Forrester, J.W., 1989, "The Beginning of System Dynamics". http://sysdyn.clexchange.org/sdep/papers/D-4165-1.pdf.

65. Forrester, J.W., 1994, "Learning through System Dynamics as Preparation for the 21st Century". at http://web.mit.edu/sdg/www/D-4434-3.21stCent.pdf.

66. Forrester, J.W., 1991, "System Dynamics and the Lessons of 35 Years". at http://sysdyn.clexchange.org/sdep/papers/D-4224-4.pdf.

67. Sterman, J.D., 2000, Business Dynamics: Systems Thinking and Modeling for a Complex World: Irwin McGraw-Hill. 982.

68. Abdel-Hamid, T.K., 1993, "Adapting, correcting, and perfecting software estimates: a maintenance metaphor". Computer, 26(3): p. 20-29.

69. Abdel-Hamid, T.K. and S.E. Madnick, 1991, Software Project Dynamics: An Integtated Approach: Prentice Hall, Englewood Cliffs, New Jersey. 264.

70. Damle, P.H., 2003, "A system dynamics model of the integration of new technologies for ship systems", Master's Thesis, Virginia Tech.

71. Scott III, J.M., 2003, "A System Dynamics Model of the Operations, Maintenance and Disposal Costs of New Technologies for Ship Systems", Master's Thesis, Virginia Tech.

72. Monga, P., 2001, "A System Dynamics Model of the Development of New Technologies for Ship Systems", Master's Thesis, Virginia Tech

73. Glover, F., 1986, "Future paths for integer programming and links to artificial intelligence". Computers & Operations Research 13(5): p. 533–549.

74. Glover, F., 1993, "A user's guide to tabu search". Annals of Operations Research, 41(1-4): p. 3-28.

75. Vapnik, V.N., 1998, Statistical Learning Theory: Wiley-Interscience.

76. Smola, A.J. and B. Schölkopf, 2004, "A tutorial on support vector regression". Statistics and Computing, 14(3): p. 199-222(24).

77. Cristianini, N. and J. Shawe-Taylor, 2000, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods: Cambridge University Press.

78. Suykens, J.A.K., T.V. Gestel, J.D. Brabanter, B.D. Moor, and J. Vandewalle, 2002, Least Squares Support Vector Machines: World Scientific Publishing Company.

79. Muller, K.R., S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, 2001, "An introduction to kernel-based learning algorithms". Neural Networks, IEEE Transactions on, 12(2): p. 181-201.

80. Chalimourda, A., B. Scholkopf, and A.J. Smola, 2004, "Experimentally optimal [nu] in support vector regression for different noise models and parameter settings". Neural Networks, 17(1): p. 127-141.

81. Scholkopf, B., A.J. Smola, R. Williamson, and P. Bartlett, 1998, "New Support Vector Algorithms", in NeuroCOLT2 Technical Report Series. Available at: (http://users.rsise.anu.edu.au/~williams/papers/P115.pdf).

82. Kwon, Y., M.K. Jeong, and O.A. Omitaomu, 2006, "Adaptive support vector regression analysis of closed-loop inspection accuracy". International Journal of Machine Tools and Manufacture, 46(6): p. 603-610.

83. Cherkassky, V. and Y. Ma, 2004, "Practical selection of SVM parameters and noise estimation for SVM regression". Neural Networks, 17(1): p. 113-126.

84. Scheines, R., P. Spirtes, and C. Glymour, 2003, "Automatic Causal Discovery". Dept. of Philosophy & CALD, Carnegie Mellon.

85. Curran, R., S. Raghunathan, and M. Price, 2004, "Review of aerospace engineering cost modelling: The genetic causal approach". Progress in Aerospace Sciences, 40(8): p. 487.

86. El-Hami, M. and S. Abu-Sharkh, 1999, "A General Design Model for Electric Motors". Available at (http://ieeexplore.ieee.org/iel5/6235/16654/00769066.pdf).

87. Yeadon, W.H. and A.W. Yeadon, 2001, Handbook of Small Electric Motors: McGraw-Hill.

88. Lin, B. and D.C. Miller, 2004, "Tabu search algorithm for chemical process optimization". Computers & Chemical Engineering, 28(11): p. 2287-2306.

89. Gendreau, M. 2002, An Introduction to Tabu Search. Available from: http://www.ifi.uio.no/infheur/Bakgrunn/Intro_to_TS_Gendreau.htm.

90. Marx, W.J., D.N. Mavris, and D.P. Schrage, 1998, "Cost/time analysis for theoretical aircraft production". Journal of Aircraft, 35(4): p. 637-646.

91. Hutcheson, G.D. and J.D. Hutcheson, 1996, "Technology and Economics in the Semiconductor Industry". Scientific American, 274(1): p.

92. Johnson, V.S.1989, "Optimizing conceptual aircraft designs for minimum life cycle cost". in Recent Advances in Multidisciplinary Analysis and Optimization, Part 3, p.1195-1217.

93. Hsu, C.-W., C.-C. Chang, and C.-J. Lin, 2004, "A Practical Guide to Support Vector Classification". Department of Computer Science and Information Engineering, National Taiwan University. Available at http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf: Taipei 106, Taiwan.

94. Smola, A.J., T.T. Frieb, and B. Scholkopf, 1998, "Semiparametric Support Vector and Linear Programming Machines", in NeroCOLT2 Technical Report Series. www.neurocolt.com/tech_reps/1998/98024.ps.gz.

95. Pai, P.-F. and C.-S. Lin, 2005, "A hybrid ARIMA and support vector machines model in stock price forecasting". Omega, 33(6): p. 497.

96. Chang, C.-C. and C.-J. Lin, 2005, "LIBSVM : a library for support vector machines". Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

97. Joachims, T., 1999, "Making large-Scale SVM Learning Practical", in Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, Editors. MIT-Press.