Manufacturing Cost Prediction in the Presence of Categorical and Numeric Design Attributes

By
Dr. Eren Sakinc and Dr. Alice E. Smith
Industrial and Systems Engineering, Auburn University

8th Annual SERC Sponsor Research Review
November 17, 2016
20 F Street NW Conference Center
20 F Street, NW
Washington, DC

www.sercuarc.org
Introduction

Critical Questions:
- What is the negotiation power over the underlying product price? Is the expected cost accurate?
- Is it possible to know the cost of a new and unique design before it is actually manufactured?

MANUFACTURER’S COST ESTIMATION STRATEGY

Step 1. Before Manufacturing
The expected cost is $ per item.

Step 2. During Manufacturing
The cost is $$ per item now.

Step 3. After Manufacturing
We cannot make it less than $$$$ per item!
• When manufacturing a new unique design, the focal point is to establish a price which maximizes customer value while being profitable.

• Since an irreversible and large amount of capital is tied up in production elements, estimating manufacturing costs accurately is critical.

• Final decisions about the product price should be based on analytical approaches, instead of intuitive expectations.
“Cost plus pricing” or “Cost based pricing”

Poorly established product prices that are a function of product cost may cause two unfavorable consequences:

— (1) A potential loss of profit due to the gap between the expected cost and the actual cost
— (2) A loss of customers and goodwill due to higher prices than necessary
Design Attributes (Cost Drivers)

• We need to know the cost structure of a product which consists of a collection of cost drivers.

• A cost driver is defined as any factor which changes the cost of an activity (according to Chartered Institute of Management Accountants – CIMA).

• From a statistical perspective, cost drivers are explanatory variables that have a contribution to the manufacturing cost of products.

\[
\text{cost drivers} = \{ \text{cost variables, design variables, design attributes, variables, attributes} \} 
\]
Type of Variables

• Categorical (Qualitative / Discrete) Variables
  ― Nominal
  ― Ordinal
  ― Binary – Symmetric and Asymmetric Binary

• Numeric (Quantitative / Continuous) Variables
  ― Interval Scaled
  ― Ratio Scaled
Cost Estimation Challenge vs. Competitive Pricing

Manufacturer A
$5 /piece

Customer

Tubular Cable Lugs
Cost Estimation Challenge vs. Competitive Pricing

Customer

Tubular Cable Lugs

Manufacturer A

$5 /piece

Manufacturer B

$4.50 /piece
Cost Estimation Challenge vs. Competitive Pricing

Manufacturer A
$5 /piece

Manufacturer B
$4.50 /piece

Manufacturer C
$4 /piece

Customer

Tubular Cable Lugs
# Alternative Approaches

<table>
<thead>
<tr>
<th>Decision Support Systems</th>
<th>Key Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Based</td>
<td>Innovative design approach</td>
<td>Dependence on past cases</td>
</tr>
<tr>
<td>Rule-Based</td>
<td>Can provide optimized results</td>
<td>Time-consuming</td>
</tr>
<tr>
<td>Fuzzy Logic</td>
<td>Handles uncertainty, reliable estimates</td>
<td>Estimating complex features costs is tedious</td>
</tr>
<tr>
<td>Expert Systems</td>
<td>Quicker, more consistent and accurate results</td>
<td>Complex programming required</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Analogical</th>
<th>Simpler method</th>
<th>Limited to resolve linearity issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression Analysis Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Back Propagation Neural Networks</td>
<td>Deal with uncertain and non-linear problems</td>
<td>Completely data-dependent, higher establishment cost</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quantitative Techniques</th>
<th>Key Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric</td>
<td>Utilize cost drivers effectively</td>
<td>Ineffective when cost drivers cannot be identified</td>
</tr>
<tr>
<td>Operation-Based</td>
<td>For optimized results, alternative process plans can be evaluated</td>
<td>Time-consuming, require detailed design and process planning data</td>
</tr>
<tr>
<td>Break Down</td>
<td>Easier method</td>
<td>Detailed cost information required about the resources consumed</td>
</tr>
<tr>
<td>Cost Tolerance</td>
<td>Cost effective design tolerances can be identified</td>
<td>Require detailed design information</td>
</tr>
<tr>
<td>Feature-Based</td>
<td>Features with higher costs can be identified</td>
<td>Difficult to identify costs for small and complex features</td>
</tr>
<tr>
<td>Activity-Based</td>
<td>Easy and effective method using unit activity costs</td>
<td>Require lead-times in the early design stages</td>
</tr>
</tbody>
</table>

## Alternative Approaches

<table>
<thead>
<tr>
<th>Qualitative Techniques</th>
<th>Key Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intuitive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case-Based</td>
<td>Innovative design approach</td>
<td>Dependence on past cases</td>
</tr>
<tr>
<td>Rule-Based</td>
<td>Can provide optimized results</td>
<td>Time-consuming</td>
</tr>
<tr>
<td>Fuzzy Logic</td>
<td>Handles uncertainty, reliable estimates</td>
<td>Estimating complex features costs is tedious</td>
</tr>
<tr>
<td>Expert Systems</td>
<td>Quicker, more consistent and accurate results</td>
<td>Complex programming required</td>
</tr>
<tr>
<td><strong>Decision Support Systems</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression Analysis Model</td>
<td>Simpler method</td>
<td>Limited to resolve linearity issues</td>
</tr>
<tr>
<td>Back Propagation Neural Networks</td>
<td>Deal with uncertain and non-linear problems</td>
<td>Completely data-dependent, higher establishment cost</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quantitative Techniques</th>
<th>Key Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Analytical</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parametric</td>
<td>Utilize cost drivers effectively</td>
<td>Ineffective when cost drivers cannot be identified</td>
</tr>
<tr>
<td>Operation-Based</td>
<td>For optimized results, alternative process plans can be evaluated</td>
<td>Time-consuming, require detailed design and process planning data</td>
</tr>
<tr>
<td>Break-Down</td>
<td>Easier method</td>
<td>Detailed cost information required about the resources consumed</td>
</tr>
<tr>
<td>Cost Tolerance</td>
<td>Cost effective design tolerances can be identified</td>
<td>Require detailed design information</td>
</tr>
<tr>
<td>Feature-Based</td>
<td>Features with higher costs can be identified</td>
<td>Difficult to identify costs for small and complex features</td>
</tr>
<tr>
<td>Activity-Based</td>
<td>Easy and effective method using unit activity costs</td>
<td>Require lead-times in the early design stages</td>
</tr>
</tbody>
</table>

Parametric / Non-parametric Simulation

- Monte Carlo Simulation
  - Parametric distribution assignments to cost drivers
  - Assignments are usually arbitrary

- Efron’s Non-parametric Bootstrapping
  - Empirical distributions
  - No benchmark comparison for validity
At most companies use linear regression models but more often rely on intuition and other ad hoc approaches.
Our Approach

• We would like to predict the manufacturing cost of a product quickly and accurately.

• We investigate ways of using clustering methods to predict the manufacturing cost of products in the presence of complex numeric and categorical design attributes.

• The accuracy of the methodology is assessed in comparison to a traditional approach, a polynomial regression model in absence of a clustering approach.
Motivations

- Many cases, costs are estimated based on primitive intuitive approaches that are far from reality and accuracy.

- Making parametrical distribution assumptions for design attributes can be arbitrary.

- Over a diverse product family, establishing only a single accurate estimation model is challenging and doubtful.
Objectives

• To accurately and quickly estimate the cost of a particular product before it is manufactured

• To deploy clustering techniques to achieve improved accuracy in the prediction

• To find appropriate number of clusters for a given case and series of products
Contributions

• First to introduce a manufacturing costs estimation approach for mixed categorical and numeric variables using clustering methods

• Implemented a simple heuristic to determine the appropriate number of clusters when there is no prior knowledge about the number of product groups
Assumptions and Limitations

- New products are based on some modifications or variations to existing or historical products
- The clustering contents are not necessarily optimized due to using a clustering heuristic
- Limited to non-parametrical approaches to avoid making assumptions concerning statistical distributions. We assume that all variables come from empirical distributions.
- We assume commodity production where the size of a batch is not important.
Suggested Methodology

1. Data Collection
   - Variable Pre-processing and Determining Number of Clusters
   - Interval-Scaled Variables
   - Ratio-Scaled Variables
   - Nominal Variables
   - Ordinal Variables
   - Binary Variables

2. Clustering Analysis
   - Cluster 1
   - Cluster 2
   - Cluster 3
   - Cluster 4

   - $EM_1$
   - $EM_2$
   - $EM_3$
   - $EM_4$

4. Available Cluster Contents
   - New Design

5. Find the Best Cluster for New Design
   - Predict the Manufacturing Cost of New Design
Suggested Methodology

Manufacturing Cost vs Design Variable

Cluster #1
Cluster #2
Cluster #3

EM₁
EM₂
EM₃

CA - CA’
Choice of Clustering Algorithm

- \( k \)-means
  - Squared error based
  - Limited to continuous variables only
  - Result is dependable on the initial random solution

- \( k \)-prototypes: Modified \( k \)-means
  - Frequency and Squared error based
  - Euclidean distance and simple matching coefficient
  - Weighting factor is arbitrary
  - Combining a quadratic expression with a linear expression
• \( k \)-medoids
  — Operates on a dissimilarity matrix
  — No randomness: Initial solution (BUILD), Moves (SWAP)
  — Handles outliers
## Choice of Distance Metric

<table>
<thead>
<tr>
<th></th>
<th>Consider Correlations</th>
<th>Handle Numeric Data</th>
<th>Handle Categorical Data</th>
<th>Handle Mixed Data</th>
<th>Non-negativity Requirement</th>
<th>Scale for Elliptical Data</th>
<th>Scale for Range</th>
<th>Modifiable Weight</th>
<th>Sensitive to Outliers</th>
<th>Unitless Measure</th>
<th>Compatibility to Our Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean Distance</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scaled Euclidean Distance</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minkowski Metric</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mahalanobis Distance</td>
<td>+ +</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canberra Metric</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czekanowski Coefficient</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chebychev Distance</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson Correlation</td>
<td>+ +</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosine Similarity</td>
<td>+</td>
<td>+ +</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similarity Coefficients</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gower’s Dissimilarity Index</td>
<td>+ + +</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Choice of Number of Clusters

• Top 6 performing indices (Milligan and Cooper):
  — Calinski and Harabasz’s PSF
  — Duda and Hart’s $J_e(2)/J_e(1)$ or PST2
  — *Dalrymple-Alford’s $C$-index
  — *Baker and Hubert’s Gamma
  — Beale’s F-ratio
  — Sarle’s CCC

• *Rousseeuw’s average silhouette width

• Consensus among Gamma (local peaks), silhouette width (local peaks & > 0.5), $C$-index (local troughs)
Choice of Predictive Model

The most complicated practice in the industry is regression models.

• Regression Models
• Splines
• Neural Networks
• Kriging
Summary of the Methodology

Manufacturing Cost Estimation

Data Sample

Cluster Analysis

Find # of Clusters ($k$)
- Silhouette Width
- Gamma
- C-Index

$k$-medoids

Find Cluster Contents

Build Cluster Specific Regression Models

MCE 1

Conventional

MCE 2

Build a Single Regression Model

Cross-Validation
Replicate for Each Object

Suggested Methodology

Benchmark Methodology
Performance Metrics

\begin{align*}
ARE_i &= \left| \frac{(Actual \ Cost)_i - (Estimated \ Cost)_i}{(Actual \ Cost)_i} \right| \\
MARE &= \frac{1}{n} \sum_{i=1}^{n} ARE_i \\
SE_i &= [(Actual \ Cost)_i - (Estimated \ Cost)_i]^2 \\
MSE &= \frac{1}{n} \sum_{i=1}^{n} SE_i \\
RMSE &= \sqrt{MSE}
\end{align*}
Real World Applications

• Electromagnetic and lightening protection parts manufacturer
  — DS1 Tubular cable lugs: 12 variables
  — DS2 Air rods: 10 variables

• Plastic kitchen and household products manufacturer
  — DS3 Plastic parts: 51 variables
Determining the Number of Clusters

**DS1**
- C-index
  - Number of Clusters: 0.0 to 0.4
  - Number of Clusters: 10 to 20
  - k = 11

**DS2**
- C-index
  - Number of Clusters: 0.0 to 0.6
  - Number of Clusters: 10 to 20
  - k = 14

**DS3**
- C-index
  - Number of Clusters: 0.0 to 0.6
  - Number of Clusters: 10 to 20
  - k = 10
### Results – Performance Metrics

#### MARE

<table>
<thead>
<tr>
<th></th>
<th>MCE 1</th>
<th>MCE 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS 1</td>
<td>4.98%</td>
<td>49.82%</td>
</tr>
<tr>
<td>DS 2</td>
<td>5.81%</td>
<td>15.42%</td>
</tr>
<tr>
<td>DS 3</td>
<td>12.39%</td>
<td>33.83%</td>
</tr>
</tbody>
</table>

#### RMSE

<table>
<thead>
<tr>
<th></th>
<th>MCE 1</th>
<th>MCE 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS 1</td>
<td>8.86%</td>
<td>140.26%</td>
</tr>
<tr>
<td>DS 2</td>
<td>355.72%</td>
<td>615.92%</td>
</tr>
<tr>
<td>DS 3</td>
<td>17.71%</td>
<td>34.20%</td>
</tr>
</tbody>
</table>

#### Min ARE

<table>
<thead>
<tr>
<th></th>
<th>MCE 1</th>
<th>MCE 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS 1</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>DS 2</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>DS 3</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

#### Max ARE

<table>
<thead>
<tr>
<th></th>
<th>MCE 1</th>
<th>MCE 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS 1</td>
<td>46.67%</td>
<td>429.52%</td>
</tr>
<tr>
<td>DS 2</td>
<td>56.04%</td>
<td>64.36%</td>
</tr>
<tr>
<td>DS 3</td>
<td>203.54%</td>
<td>233.79%</td>
</tr>
</tbody>
</table>
Results – Model Fit

DS1
R-Sq = 99.94%

MCE1

DS2
R-Sq = 96.83%

MCE2

DS3
R-Sq = 93.69%

Predicted Cost (TL)
Actual Cost (TL)

Predicted Cost (TL)
Actual Cost (TL)

Predicted Cost (TL)
Actual Cost (TL)
Results – Sensitivity of Number of Clusters

**DS1**

- MARE
- Number of Clusters: 5, 10, 15
- k = 11

**DS2**

- MARE
- Number of Clusters: 10, 15, 20
- k = 14

**DS3**

- MARE
- Number of Clusters: 5, 10, 15
- k = 10
## Results – Sensitivity to Polynomial Model

<table>
<thead>
<tr>
<th></th>
<th>DS1</th>
<th>DS2</th>
<th>DS3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MCE1Q</td>
<td>MCE2Q</td>
<td>MCE1Q</td>
</tr>
<tr>
<td>MARE</td>
<td>4.37%</td>
<td>43.22%</td>
<td>2.31%</td>
</tr>
<tr>
<td>Min ARE</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Max ARE</td>
<td>81.12%</td>
<td>440.98%</td>
<td>23.26%</td>
</tr>
<tr>
<td>MSE</td>
<td>0.42%</td>
<td>93.49%</td>
<td>188.45%</td>
</tr>
<tr>
<td>RMSE</td>
<td>6.50%</td>
<td>96.69%</td>
<td>137.28%</td>
</tr>
</tbody>
</table>
• We investigated ways of using clustering methods to predict the manufacturing cost of a product without actually manufacturing it.

• The accuracy of the methodology is assessed in comparison to a simple regression model with the absence of clustering approaches.

• The main concern is to predict the manufacturing cost of a product without dealing with arbitrary assignments of statistical distributions to cost related attributes.
• In real production systems often a variety of products are being manufactured under a single facility roof.

• Over a diverse product family, establishing only a simple accurate estimation model is challenging and even questionable.

• This motivated us grouping products according to their design features, common manufacturing operations or some other factors by dividing the whole database of products into neighborhoods.

• Then for each group of products (clusters), a cost estimation model is developed to predict the manufacturing cost of a new product with using the cluster specific model.
Direction of Future Research

• Developing a comprehensive similarity measure that demonstrates high inter-cluster variability while being able to handle mixed categorical and numeric design attributes.

• A deterministic model such as a mixed integer programming model can be implemented to obtain the optimal cluster results.

• Information gain criterion can be considered when deciding on the inclusion of a candidate predictor (design attribute) in the cost estimation model.
Manufacturing Cost Prediction in the Presence of Categorical and Numeric Design Attributes

By
Dr. Eren Sakinc and Dr. Alice E. Smith
Industrial and Systems Engineering, Auburn University

8th Annual SERC Sponsor Research Review
November 17, 2016
20 F Street NW Conference Center
20 F Street, NW
Washington, DC

www.sercuarc.org