Journal of Economic Analysis

Homepage: https://www.anserpress.org/journal/jea

Manufacturing Cost Estimation Using Piecewise Function Approaches

Eren Sakinc [a], Alice E. Smith [b,*]

[a] *Bayer Pharmaceuticals, New Jersey, USA*

[b] *Department of Industrial and Systems Engineering, Auburn University, Auburn, USA*

ABSTRACT

This paper describes two novel approaches to cost estimation of manufactured products where a data set of similar products have known manufactured costs. The methods use the notion of piecewise functions and are (1) clustering and (2) splines. Cost drivers are typically a mixture of categorical and numeric data which complicates cost estimation. Both clustering and splines approaches can accommodate this. Through four case studies, we compare our approaches with the often-used regression models. Our results show that clustering especially offers promise in improving the accuracy of cost estimation. While clustering and splines are slightly more complex to develop from both a user and a computational perspective, our approaches are packaged in an open-source software. This paper is the first known to adapt and apply these two well-known mathematical approaches to manufacturing cost estimation.

KEYWORDS

Artificial intelligence; cost estimation; non-parametric modeling; clustering; splines; cross-validation; dissimilarity index

* Corresponding author: Alice E. Smith
E-mail address: smithae@auburn.edu

ISSN 2811-0943
doi: 10.58567/jea02030007

Received 8 April 2023, Accepted 20 May 2023, Available online 22 June 2023, Version of Record 15 September 2023

1. Background

Poorly established product prices may cause two unfavorable consequences: (1) A potential loss of profit due to the gap between the expected cost and the actual cost, and (2) A loss of customers and goodwill due to higher prices than competitors in the market. Statistical tools have always been popular among executive planners when cost estimation effort takes place. Before proceeding forward into statistics, 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[1]. From a statistical perspective, cost drivers are explanatory variables that have a contribution to the manufacturing cost of products. Through this paper, synonyms for cost drivers are cost variables, design variables, design attributes or, simply, variables and attributes.

The main concern of our research is to predict the manufacturing cost of a product without dealing with probability density or mass function assignments or making strong assumptions concerning parameters. We will convert physical similarities of products into meaningful mathematical similarities and make product-by-product comparisons. When making product-by-product comparisons, the number of analogies is likely to grow as the number of products grows. Therefore, over a diverse product family, establishing only a single accurate estimation model is challenging and doubtful. This motivates us to make comparisons by dividing the database of products into neighborhoods until these neighborhoods become sufficiently homogenous and using piecewise functions. Using statistical terminology, we can call these neighborhoods, groups or clusters. We then develop cost estimation models for each cluster. There are many clustering techniques as we explain later but few applicable to the general task of cost estimation in manufacturing.

When cluster specific models are considered within their defined ranges, at the boundaries they are non-continuous but can form piecewise functions. Since the main concern of this research is to predict the manufacturing cost of a product with non-parametric methods, an alternative to clustering is to use splines. We can define a spline as a function that is constructed by piecewise polynomial functions where these polynomial segments connect. Our research also seeks the possibility of building spline models to accommodate cost estimation process with improved accuracy.

There are two issues rendering this cost estimation problem quite complicated: (1) incorporating qualitative and quantitative variables in a dataset simultaneously, (2) the number of variables in a dataset may be less than the number of products but still large relative to the number of products. We address the first issue by using applicable clustering and spline techniques and the second issue by removing irrelevant variables and leveraging the data set.

In this paper, we have collected four datasets from three manufacturing industries. The representative features have been selected according to the cost drivers for these specific manufacturing processes. The diversity of the manufacturer datasets shows that this study can be extended over different industries by including industry specific design variables.

This paper is the first known application of clustering and of splines to cost estimation in manufactured products. We show that these approaches can be relatively straightforward and can offer advantages over the often-used multiple regression models. Section 2 gives the relevant literature while section 3 details the clustering approach. Section 4 details the spline approach and Section 5 gives results and discussion. Section 6 describes our software system which is in the public domain. Section 7 wraps up with concluding remarks and future research.

2. The relevant literature

2.1. Manufacturing cost estimation

[1] According to Chartered Institute of Management Accountants (CIMA).

Layer et al. (2002) point out that manufacturing cost calculations are classified based on the timing of calculations: (1) Pre-calculation, (2) Intermediate calculation, and (3) Post-calculation. Pre-calculation estimates the potential costs before actually manufacturing the item. The price of a product is usually declared based on the pre-calculation values when a new unique design has been requested by a customer for a future manufacturing agreement. As a result, higher accuracy in the pre-calculation step is crucial to generate designs where low-cost and high-quality are maintained. On the other hand, the actual cost is the interest of the post-calculation phase. Instead of estimated cost drivers, incurred costs are included in the post evaluation step. Our research interest is the pre-calculation phase where we seek establishing the cost of a product accurately before actual production takes place. However, we need historical data of product costs previously recorded based upon the post-calculation for our methods.

Manufacturing cost estimation techniques are classified into two main categories consistently by authorities. Layer et al. (2002) and Dai et al. (2006) termed these two main categories qualitative and quantitative techniques. However, second-level classifications vary according to subjective opinions. Figure 1 has been regenerated from a literature survey of product cost estimation and gives an overview of the key advantages and limitations of the underlying product cost estimation techniques (Dai et al., 2006).

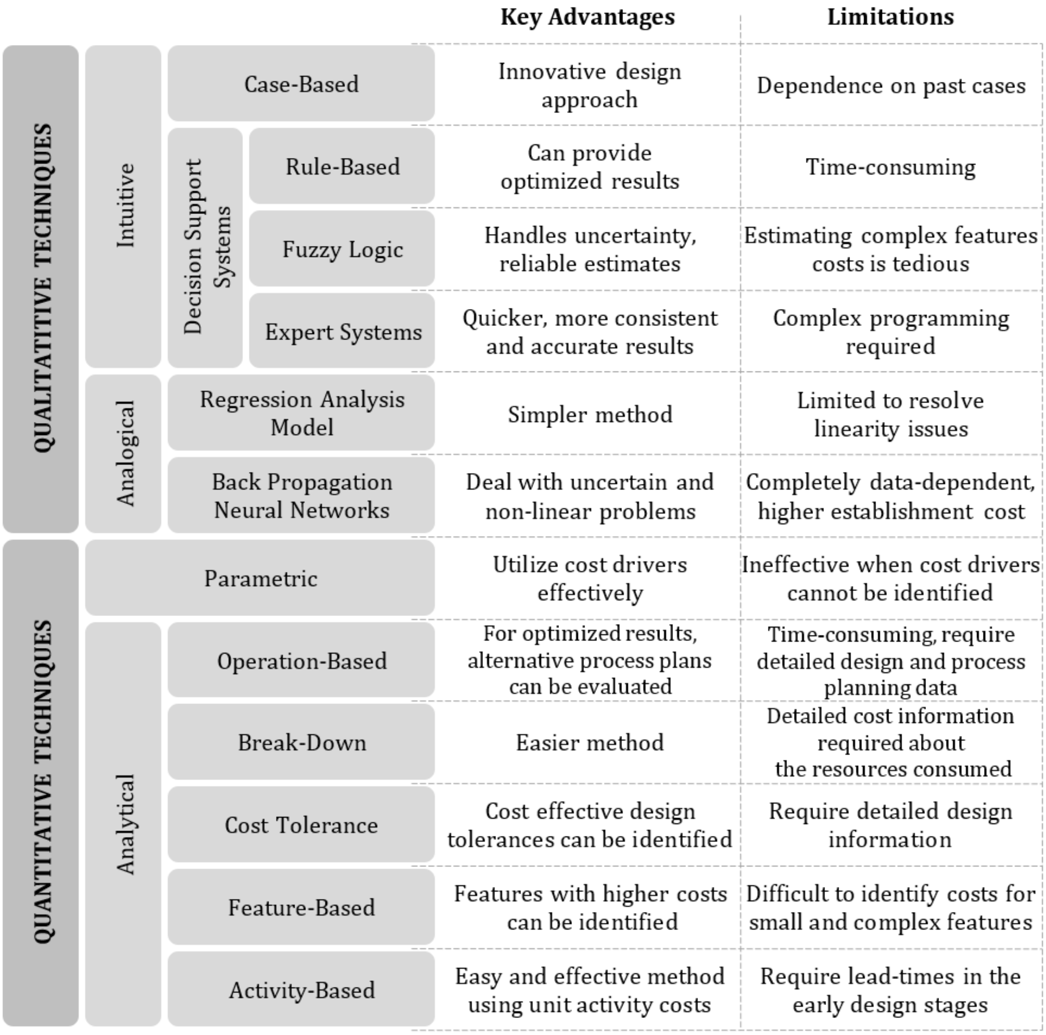
Our clustering-based cost estimation approach fits none of these classifications strictly but can be considered as a combination of several approaches, namely case-based systems, analogical parametric cost estimation techniques, operation-based, and feature-based models. In our study, manufacturing cost estimation uses historical data of similarities among previously manufactured products.

On the other side, spline functions have never been used as a manufacturing cost estimation tool in the literature. Our curiosity in using such a model motivated us to develop spline cost estimation models that can accommodate mixed categorical and numeric design attributes. Our spline-based cost estimation approach can also be considered a combination of several approaches, namely analogical non-parametric regression analysis along with operation-based and feature-based models.

2.2. Clustering approaches

Figure 2 shows the main approaches in clustering while Table 1 assesses the clustering methods according to their suitability for manufacturing cost estimation. A leading algorithm is the k-means (or c-means) clustering method. It was first introduced by MacQueen (1967) to allocate observations in a dataset into a pre-determined number of clusters – k. The logic behind the k-means algorithm is to find the content of k partitions by minimizing within cluster variances.

Two decades after the introduction of the k-means algorithm, the partitioning around medoids (PAM) paradigm was developed by Kaufman and Rousseeuw (1987). They called this method, the k-medoids algorithm. The objective of the method is not to minimize within cluster variability as in k-means. Unlike k-means approach, the method uses real observations as cluster centers and partitions the whole data around these cluster medoids. In other words, instead of devising the error sum of squares approach, the algorithm seeks cluster contents around representative objects based upon minimization of total dissimilarity. Allocating the observation points to the nearest medoid is advantageous in many aspects. Since the cluster centers are picked from appropriate elements in the actual dataset, the variables in that dataset do not solely need to be on an interval scale. Kaufman and Rousseeuw (1987) also proved that the k-medoids approach gives more robust results than methods based on variance minimization, as with k-means. Additionally, the existence of outliers does not perturb the k-medoids clustering progress.



				Key Advantages	Limitations
QUALITATIVE TECHNIQUES	Intuitive	Case-Based		Innovative design approach	Dependence on past cases
		Decision Support Systems	Rule-Based	Can provide optimized results	Time-consuming
			Fuzzy Logic	Handles uncertainty, reliable estimates	Estimating complex features costs is tedious
			Expert Systems	Quicker, more consistent and accurate results	Complex programming required
	Analogical	Regression Analysis Model		Simpler method	Limited to resolve linearity issues
		Back Propagation Neural Networks		Deal with uncertain and non-linear problems	Completely data-dependent, higher establishment cost
QUANTITATIVE TECHNIQUES	Parametric			Utilize cost drivers effectively	Ineffective when cost drivers cannot be identified
	Analytical	Operation-Based		For optimized results, alternative process plans can be evaluated	Time-consuming, require detailed design and process planning data
		Break-Down		Easier method	Detailed cost information required about the resources consumed
		Cost Tolerance		Cost effective design tolerances can be identified	Require detailed design information
		Feature-Based		Features with higher costs can be identified	Difficult to identify costs for small and complex features
		Activity-Based		Easy and effective method using unit activity costs	Require lead-times in the early design stages

Figure 1. Overview of product cost estimation techniques with advantages and limitations. Adapted from Dai et al. (2006).

2.3. Clustering similarity measures

Most clustering techniques require an assignment of a similarity (or dissimilarity) measure in the very initial step. Table 2 gives a comprehensive summary of ten similarity measures. The attributes included in the table are specifically chosen considering the scope of our application problems. These are the aspects of correlation consideration, handling only numeric data, handling only categorical data, handling mixed numeric and categorical data, non-negativity requirement, scaling for ranges of variable and elliptical shaped data, modifiable weights, sensitivity to outliers, unitless measure and metric properties, and, lastly, but most importantly, compatibility of these measures with typical manufacturing cost estimation. As you can see from this table, none of the existing similarity measures are completely compatible with our requirements in their original forms. Notice that, a plus sign (+) points out the presence of the feature for a similarity measure.

Unfortunately, existing similarity measures cannot handle mixed numeric and categorical variables. Using Gower's index to construct a proximity matrix is a good alternative for the clustering analysis because it enables us to transform outcomes of different types of variables into a single mathematical value including categorical and numeric variables (Kaufmann and Rousseeuw, 1990). The original form of Gower's index handles interval, nominal,

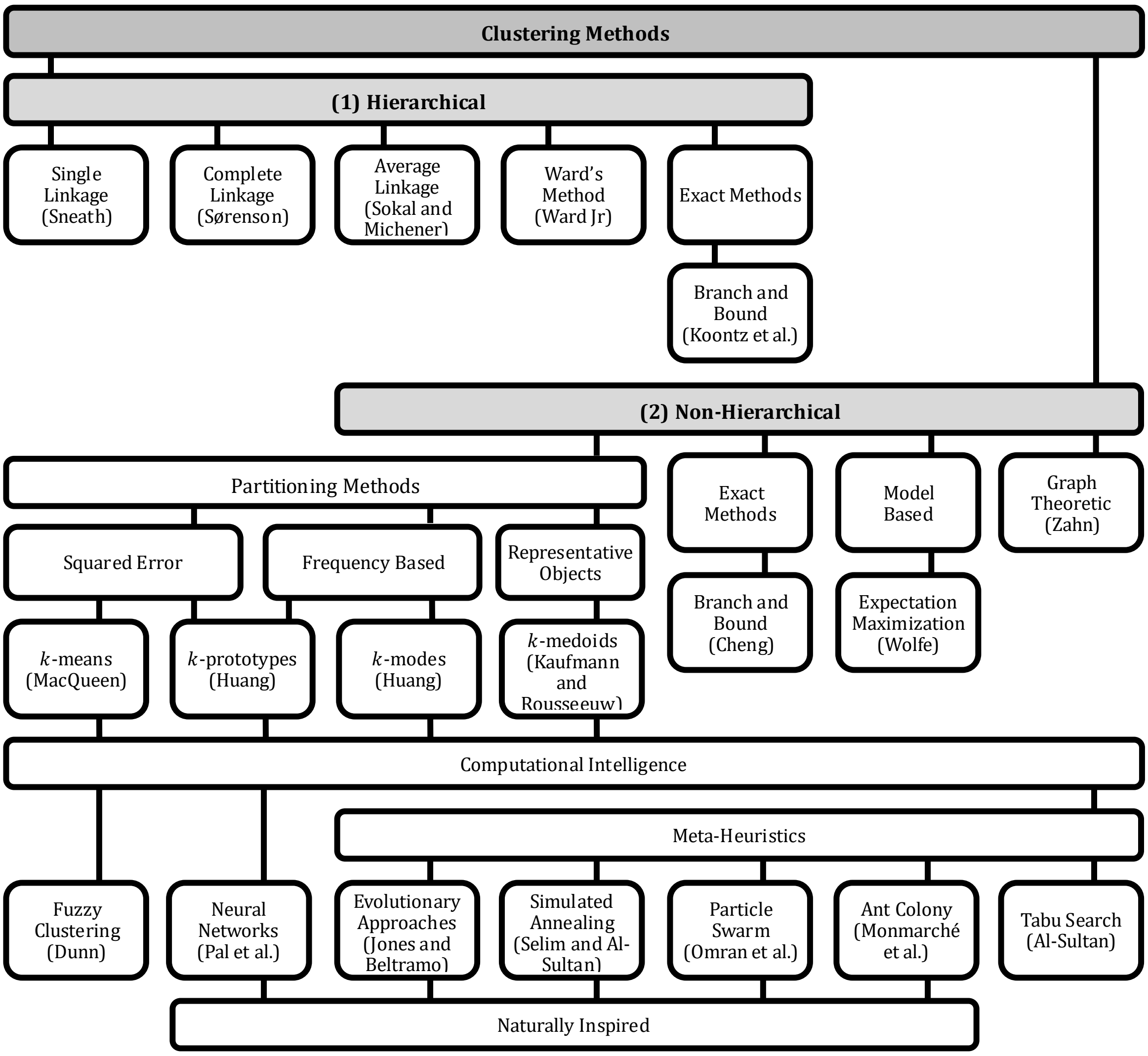


Figure 2. Extended classification of clustering methods.

and binary data as a similarity coefficient between 0 and 1. Kaufmann and Rousseeuw (1990) described a slight generalization of this coefficient which covers ordinal and ratio variables in addition to the ones mentioned for the original index. With a simple transformation, Gower's original similarity coefficient (Gower, 1971) can be converted into a dissimilarity value between 0 and 1. Kaufmann and Rousseeuw (1990) also transformed the similarity coefficients into dissimilarities. The only downside for Gower's index is that the index is linear. The discrimination capacity of the index might not be as powerful as a quadratic or a higher degree polynomial expression.

2.4. Splines

Splines constitute a reasonable approach for nonparametric estimation of manufacturing cost functions. A spline is a piecewise polynomial (or other functional form) with different polynomials located between "knots" in the cost driver hyperspace. Unfortunately, commonly known splines are restricted to continuous predictors (attributes). This is a disadvantage when it comes to the generalization of using splines for manufacturing cost estimation problems since we may encounter mixed categorical and numeric predictors.

A numerically stable representation of splines can be written as linear combinations of a set of basis functions called B-splines. B-splines was a major development in spline theory and is now the most used in spline applications and software. The term "B-spline" was introduced by Curry and Schoenberg (1947). B-spline is a generalization of

Table 1. Overview of the most common clustering methods.

Clustering Technique	Computational Complexity[2]		Type of Data[3]			Sensitivity to Outliers	Best Data Set Size[4]	Initial Seed Dependence	Comments
	Time	Space	C	N	M				
Enumeration[5]		$C(N,K)$	+	+	+	No	S	No	Impractical / prohibitive
Enumeration[6]		$K^N/K!$	-	+	-	No	S	No	Impractical / prohibitive
Single Linkage	$O(N^2)$	$O(N^2)$	+	+	-	Yes	S	No	Good for taxonomy
Complete Linkage	$O(N^2)$	$O(N^2)$	+	+	-	No	S	No	Not sensitive to outliers
Average Linkage	$O(N^2)$	$O(N^2)$	+	+	-	No	S	No	Good for taxonomy
Ward's Method	$O(N^2)$	$O(N^2)$	-	+	-	Yes	S	No	Sensitive to normality
k-means	$O(NKd)$	$O(N+K)$	-	+	-	Yes	L	Yes	Easy to implement
k-medoids	$O(Kd(N-K)^2)$	$O(N+K)$	+	+	+	No	S	No	Relatively complex
k-modes	$O(NKd)$	$O(N+K)$	+	-	-	No	S – L	Yes	Best for binary data
k-prototypes	$O(NKd)$	$O(N+K)$	+	+	+	Yes	S – L	Yes	Efficient as k-means
Branch & Bound	N/A	Varies	-	+	-	No	S	No	Gives exact solution
Model Based	$O(N \log N)$	N/A	+	+	+	No	S – L	No	Non-arbitrary similarity
Graph Theoretic	$O(N^2)$	$O(N^2)$	-	+	-	No	S	No	For irregularly shaped clusters
Meta-Heuristics	Varies	Varies	+	+	+	No	L	Possibly	Gives solutions fast
Cluster Ensemble	Varies	Varies	+	+	+	No	S	Varies	Consolidation issues

Table 2. Summary of the most common similarity measures.

	Consider Correlations	Handle Numeric Data	Handle Categorical Data	Handle Mixed Numeric and Categorical Data	Non-negativity Requirement	Scale for Elliptical Data	Scale for Range	Modifiable Weight for Differences	Sensitive to Outliers	Unitless Measure	Distance Metric	Compatibility to Our Work
Euclidean Distance		+							+		+	
Scaled Euclidean Distance		+				+			+	+	+	
Minkowski Metric		+						+	+		+	
Mahalanobis Distance	+	+				+				+	+	
Canberra Metric		+			+		+			+		
Czekanowski Coefficient		+			+		+			+		
Chebychev Distance		+									+	
Pearson Correlation	+	+				+				+		
Cosine Similarity		+				+	+					
Similarity Coefficients			+				+			+		

Bézier curves using the de Boor recursion formula (Boor, 1976). B-splines are attractive for non-parametric modeling but choosing the appropriate number of knots with their locations is a significant issue. Eilers and Marx (1996) proposed a roughness penalization procedure by starting with a relatively large number of knots, but still less than one per observation. This method combines the reduced knots of regression splines with the roughness penalty of smoothing splines where the coefficients are determined partly by the data to be fitted and partly by an additional penalty function that aims to avoid over-fitting.

The method of tensor product splines is an extension to the one-dimensional spaces of polynomial splines over a space of multi-dimensional splines by taking tensor products. Because of the outer product nature of the multi-dimensional space, many properties of polynomial splines in one dimension are retained, such as working with single dimension B-spline functions (Schumaker, 2007). Tensor product models consider interaction terms between univariate spline functions. We will use an approach which takes the tensor products of spline functions into account to handle multiple predictors.

[2] N: Number of objects, K: Number of clusters, d: Number of variables (dimension)

[3] C: Categorical, N: Numerical, M: Mixed Categorical and Numerical

[4] S: Small, L: Large

[5] Enumeration expression is written for combinatorial problems where K objects are chosen out of N observations as cluster centers

[6] Enumeration expression is written for combinatorial problems where N observations are allocated into K clusters with the nearest mean

2.5. Cost estimation approaches using clustering and splines

One of the most relevant studies that have been conducted so far is the work of Angelis and Stamelos (2000) concerning software cost estimation. Angelis and Stamelos developed a non-parametric bootstrap simulation tool to investigate the accuracy of the underlying estimation methodology which is constructed on Euclidean, Manhattan, and Chebyshev distances between an active project and historical projects. Although this work specifically uses similarities between historical projects and an active project in the development phase with an emphasis on Gower's index, it does not employ any clustering technique nor an estimation model such as regression models or neural networks.

Lee et al. (1998) proposed a two-phase software cost estimation method which is based on clustering analysis and neural networks for mixed numerical and categorical data. For quantitative attributes, they used average Euclidean distance. On the other hand, for nominal attributes, the Jaccard coefficient is calculated. A neural network which is trained using the output of clustering analysis promises higher accuracy than a non-cluster-integrated neural network. As a downside, their work was limited to single linkage hierarchical clustering without the existence of ordinal and binary variables. Van Hai et al. (2022) considered several alternative clustering methods to estimate effort (not cost) of software projects. They used five categorical variables and clustered using k-means, both for the variables collectively and separately to compare those approaches with not clustering the data.

Xu and Khoshgoftaar (2004) extended software cost estimation efforts with a fuzzy c-means (k-means) clustering approach. Because software experts define the level of complexity according to their subjective opinions, using cost associated variables which take certain numerical values does not reflect the true nature of software cost estimation efforts. Hence, this research accounts for the imprecision and vagueness of expert knowledge with linguistic variables and fuzzy rules. Although the whole method appears to handle mixed numerical and categorical variables, in fact, the clustering module itself is only limited to numerical data.

The performance of multivariate adaptive regression splines (MARS) for software cost estimation efforts was investigated by Pahariya et al. (2009). The real challenge in our methodology is dealing with mixed numeric and categorical variables, and Pahariya et al.'s work is not very helpful as it mandates unreasonable simplifications in the data preparation phase such as discretization of numerical data into ordinal variables.

Michaud et al. (2003) conducted research on estimating total direct medical costs of people with rheumatoid arthritis. These medical costs include physician and healthcare worker visits, medications, diagnostic tests and procedures, and hospitalization where the effect of age on the total cost indicated a V-shaped scatter. To model this relatively complex age vs. cost relationship, they used linear splines with a single interior knot. Even though Michaud et al. implemented an approach to estimate the cost based on categorical and numeric demographic predictors, they only used an integer scale numeric variable, age, to develop the spline models.

Another cost estimation related research was done by Almond et al. (2005) about the hospitalization costs of low birth weight on heavier and lighter infants from twin pairs born in the United States. To quantify the health status of a newborn, among these five variables, only birth weight factor is used to build a piecewise linear spline model. However, no categorical factors have been considered in the spline model. Almond et al. calculated hospitalization cost by adding generic expenses for each treatment performed on a newborn. The research lacks two aspects compared with our cost estimation methodology: (1) Not utilizing categorical variables in the spline model, and (2) Not using actual cost values to evaluate the performance of the underlying parametric model.

Carides et al. (2000) presented a procedure for estimating the mean cumulative cost of long-term treatment on two clinical studies: (1) Heart failure clinical trial of left ventricular dysfunction, and (2) Ulcer treatment. A two-stage estimator of survival cost with parametric regression, and a non-parametric regression with cubic smoothing splines are devised to exploit the underlying relationship between total treatment cost and survival time. However, only continuous covariates are used in the two-stage model and the effect of both categorical and numeric attributes associated with each of these clinical studies was not considered.

Valverde and Humphrey (2004) developed translog, Fourier, and cubic spline models to predict the cost effects of 20 individual bank mergers. The motivation behind this research was to accurately estimate the decrease in unit costs due to the merger. The underlying performance metric was the actual cost changes affecting all merging banks. Only two numeric variables were under consideration in the cubic spline models: (1) Value of loans, and (2) Value of securities (and other assets) while categorical merger bank attributes were not implemented in the cost estimation efforts.

Table 3 highlights the most relevant cost estimation literature using clustering techniques/splines and type of data. A "+" sign indicates that the underlying research is in which specific area of application, what kind of approach is devised, and what type of data is used. For instance, Carides et al. (2000) implemented spline models to estimate clinical costs by using numeric data. As you notice, clustering techniques or spline models have not been used in manufacturing cost estimation efforts because of the complex relationships between categorical and numeric design attributes.

Table 3. Overview of the most relevant research.

Article	Area of Application[7]			Estimation Approach		Type of Data[8]			Comments
	SCE	CCE	MCE	Clustering	Splines	C	N	M	
Angelis and Stamelos (2000)	+							+	Analogical relationships used
Lee at al. (1998)	+			+				+	No ordinal or binary variables
Xu and Khoshgoftaar (2004)	+			+			+		Subjective attribute assignments
Pahariya et al. (2009)	+				+		+		Omitted majority of variables
MIchaud et al. (2003)		+			+		+		Considered one variable in splines
Almond et al. (2005)		+			+		+		Used estimated medical costs
Carides et al. (2000)		+			+		+		Promising estimation results
Valverde and Humphrey (2004)					+		+		Limited data with poor accuracy

3. Clustering cost estimation approach

3.1. Grouping products

Our clustering cost estimation approach is a two-phase process. In the first phase, we use all historical products to evaluate possible clustering formations and to build a cost estimation model for each cluster. The second phase is the cost prediction phase in which a new design is assessed for the best cluster fit and then the corresponding cost estimation model is used. According to design similarities between a new design and the existing clusters established in the first phase, we select the best cluster to which the new design should be assigned. Once the best cluster is found, the remaining part is to use the cluster specific cost estimation model to predict the manufacturing cost of the new design.

3.2. Determining the number of clusters

Unfortunately, there is no definitive methodology for determining the number of clusters (SAS Institute Inc., Cary, 2008). In a practical sense, graphically assessing the data scatter is a good start but when there are more than two or three dimensions (i.e., variables), this is not as practical as it first appears. Also, when the data is mixed with categorical and numeric values, it is very hard to identify clusters visually.

Even though it is possible to have an idea of how many product groups exist in a database based on experts'

[7] SCE: Software Cost Estimation, CCE: Clinical Cost Estimation, MCE: Manufacturing Cost Estimation

[8] C: Categorical, N: Numeric, M: Mixed Categorical and Numerical

opinions in a company, the groups are usually not distinct, or the given opinions do not represent the similarities among products perfectly. The distinction power of a similarity measure becomes very crucial in this phase because it forms the basis of these comparisons among products or products with clusters. During the cluster analysis stage, we need to choose the appropriate number of clusters. This is directly linked with how many cost estimation models are required to be built at the end of the first phase.

There are few methods appropriate for mixed data but among these are Dalrymple-Alford's C-index (1970), Baker and Hubert's Gamma (1975) or Rousseeuw's silhouette width (1987). These three statistics operate on a dissimilarity matrix and a vector of integers indicating the cluster number to which each observation is assigned.

The C-index uses the sum of all the within cluster distances. The number of clusters which minimizes the C-index should be chosen. Baker and Hubert (1975) devised an index called Gamma which was adopted from Goodman and Kruskal's gamma (γ) (1954) to use in clustering applications. The index basically compares within cluster distances with between cluster distances (Everitt, 2010) where a pair of distances is considered consistent (inconsistent) if the within cluster distance is less (greater) than the between cluster distance (Li and Racine, 2007). Gamma was found to be one of the best performing statistics among the 30 considered by Milligan and Cooper (Milligan and Cooper, 1985). Another index which is applicable for mixed numeric and categorical data is Rousseeuw's silhouette width (1987). It was devised to assess how well each object lies within its assigned cluster. Even though the silhouette width was first developed for partitioning around medoids, it is possible to use it in any context for which a distance matrix can be derived. The fundamental procedure behind this approach is plotting the average silhouette widths for the entire dataset that are obtained from different choices for the number of clusters and selecting the number of clusters which maximizes the index.

Our methodology of selecting the appropriate number of clusters is neither deterministic nor arbitrary, but it is consistent with and also as simple as the one defined in the user manual of SAS for numeric data (SAS Institute Inc., Cary, 2008). We look for consensus among three statistics, namely C-index, Gamma, and silhouette width, and these statistics can be applied regardless of the type of data. Thus, there is no absolute optimal choice of number of clusters but rather the narrowing of possible choices to a few (or sometimes only one), superior numbers of clusters for the analyst to choose from.

3.3. Choice of clustering algorithm

The k-medoids algorithm was found to be more robust than any clustering technique that uses the error sum of squares (Kaufmann and Rousseeuw, 1987). Instead of minimizing the error sum of squares, it finds a set of representative observations (medoids) for each cluster and then allocates all other remaining observations to these clusters according to the closest distance to each medoid. This is advantageous in three aspects: (1) Possibility of clustering mixed data when a dissimilarity matrix can be derived, (2) Possibility of handling outliers, and (3) Elimination of making assumptions about underlying distributions such as multivariate normality.

We employ the k-medoids clustering algorithm as described in Kaufmann and Rousseeuw (1990). They implemented the k-medoids algorithm in a program called "PAM". PAM consists of two phases. These phases are called BUILD and SWAP. The first phase, BUILD, constructs an initial solution of k representative objects and the second phase, SWAP, attempts to improve the set of representative objects. The objective function of the algorithm is to minimize the sum of distances (dissimilarities) of each object to their closest representative object.

3.4. Regression models

For each cluster, a regression model is developed. In a regression model for the manufacturing cost estimation problem, the outcome (or dependent) variable is the manufacturing cost, and independent (explanatory) variables are the cost drivers (design attributes in this case). We assume a 5% confidence level for determining the significance of independent variables and their interactions. Checking interactions between variables is crucial because some variables create antagonistic or synergetic effects which may significantly impact the cost of a product. The variables and interaction terms are eliminated if these are irrelevant or have statistically non-significant contribution on the cost value.

To reduce the computational load and to avoid over parameterization we developed linear regression models. However, the performance of quadratic regression models was also assessed without much effect on results. We constructed k regression models where k represents the number of clusters.

4. Spline cost estimation approach

Our spline cost estimation approach is also a two-phase process. In the first phase, we use all historical products to build a spline cost estimation model. There are several different spline functions available for practitioners to use for estimation purposes. However, the main concern is handling mixed numeric and categorical data. The second phase is the cost prediction phase in which the manufacturing cost of a new design is assessed.

In this research, we need to model complex relationships of categorical and numeric variables. A range of kernel regression methods have been proposed to model such relationships (Ma et al., 2014). We used the same approach as described in Racine et al. (2014) to accommodate the existence of categorical and numeric design attributes since the method demonstrates robust performance on both simulated and real world data without breaking the data into subsets of continuous only and categorical only variables. Racine et al. (2014) proposed tensor-product polynomial splines weighted by kernel functions method to estimate the unknown conditional mean in the location-scale model.

Racine et al. (2014) implemented their work in R with a package called “crs” (Nie and Racine, 2012). The package is appealing for applied researchers because it uses a framework for nonparametric regression splines to address the existence of categorical and numeric variables. We used the same package in R and applied it to our cost estimation problems.

There are two common approaches to determine the location of knots (Audet et al., 2009): (1) Knots can be placed based on equally spaced quantiles where the number of observations in each segment is equal or (2) Knots can be placed at equally spaced intervals. The “crs” package has the flexibility to use either option but most significantly, it chooses the knot placement strategy automatically based on whichever method provides better output.

The package “crs” offers two search options to optimize the number of interior knots along with the value of bandwidths – smoothing parameter for categorical variable: (1) Exhaustive search or (2) Non-smooth optimization by mesh adaptive direct search, NOMAD (Audet et al., 2009). The number of interior knots for each continuous predictor is an integer value and the ranges for each categorical predictor is a value between [0,1]. Clearly, using an enumeration-based method such as exhaustive search might be computationally expensive for large datasets considering the number of categorical and numeric variables. In the “crs” package, the NOMAD approach was adopted to leverage recent advances in mixed-integer problems and to avoid the computational burden of using a brute-force method. Note that in some cases, the optimal spline degree is found to be zero, and the bandwidth is one. It means the corresponding variables are automatically removed from the model.

5. Test cases and results

In this section, we apply our manufacturing cost estimation methodology on four datasets from three different industries. We present these real-world problems from least to most complexity according to their sizes in terms of number of numeric and categorical variables and observations. The data was collected from socks, electromagnetic parts, and plastic tools manufacturing factories in Ankara and Konya, Turkey. Mixed numeric and categorical design attributes, cost drivers, or other variables comprise in these datasets. Due to the confidentiality agreements that were signed with these companies, we cannot state any brand names or product codes. Note that these data sets are diverse and representative but do not cover the realm of cost estimation possibilities. Therefore, the results presented herein cannot be assumed to be fully generalizable.

Because of the relative smallness of the data sets, we leverage the data fully. We use leave-one-out cross-validation in our study to validate the performance of the estimation models that are being constructed. An observation is left out to test a cost estimation model that is built or trained with the remaining observations in the dataset. The observation being left out for every replication can be considered as an external test data point since it is not used in the cluster analysis nor model building phases.

For clusters, first, we conduct a cluster analysis and then build cluster specific cost estimation models based on the entire data except the left-out observation. Second, we find the cluster in which the left-out observation falls. Finally, we test the corresponding cluster specific estimation model with the left-out data point. With the same logic, first we build a spline model leaving one product out of the data sample. Second, we evaluate the spline model validity with the left-out observation point.

Figure 3 gives the overall structure of our proposed approaches to manufacturing cost estimation. Next, we describe the case studies and data sets.

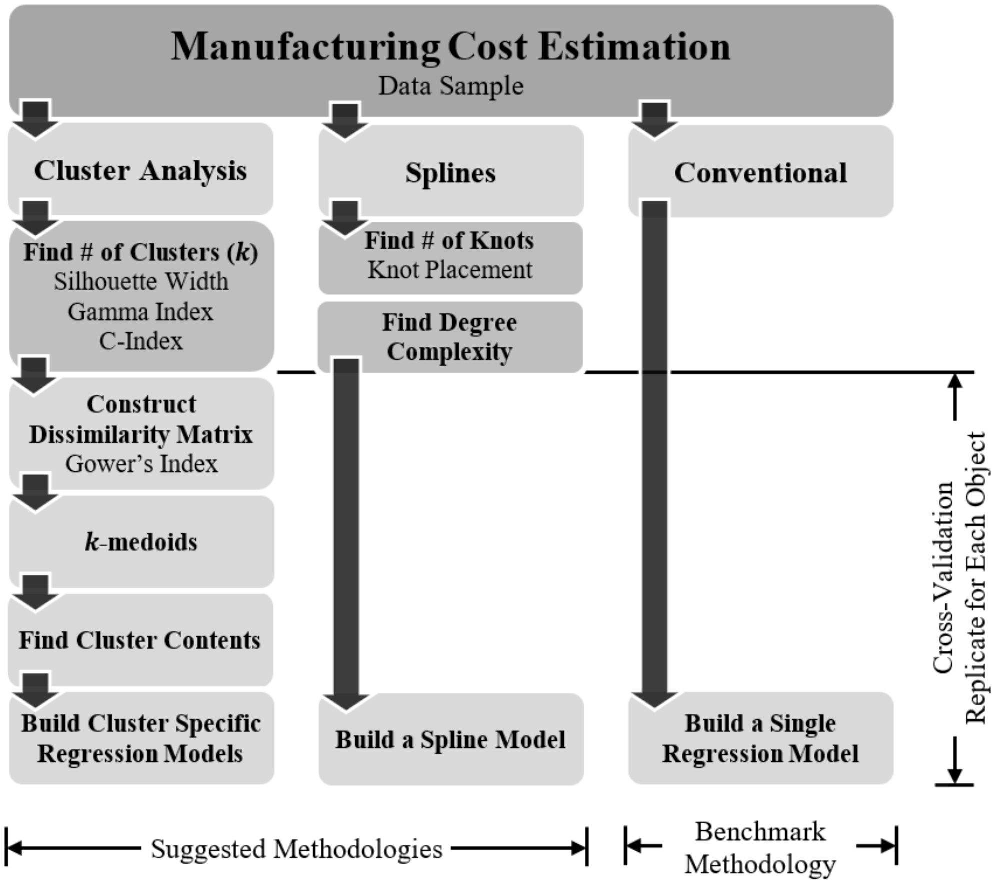


Figure 3. Summary of the proposed manufacturing cost estimation methodologies.

5.1. Company and dataset descriptions

We thought it is very important to validate and demonstrate our proposed methods on actual cost estimation data rather than simulated data sets. Actual data can be imprecise and sparse. These are qualities that complicate cost estimation, and our data sets reflect this.

5.1.1. Socks manufacturing data

The first application problem dataset was collected from a socks manufacturer which produces copyrighted and licensed socks for some major brands in Europe and USA. Their range of products consists of sports, casual, and formal/dress socks for women, men, children, and infants. The manufacturing processes include pattern design, knitting, toe seam, washing-softening, pattern printing, final quality control, and packaging. Steam, silicon, and antibacterial washing are the types of washing-softening operations. In the printing department, the company can apply lithographs, holograms, and heat transfer, embroidery, rubber, acrylonitrile butadiene styrene (ABS), and caviar bead prints.

The dataset that we collected from the company's database contains information for 76 products of women's and men's socks. There are nine variables associated with these products, and eight of these variables are qualitative (categorical), namely raw material, pattern, elasticity, woven tag, heel style, leg style, fabric type, and gender. The only quantitative variable measured on a continuous scale in this dataset is the actual cost which is recorded in Turkish Lira (TL) money units. Table 4 is the summary of the dataset and associated attributes. The columns of the table are variable name, data type, variable type, and categories (for categorical data) or range (for numeric data) from left to right, respectively. For nominal variables, the order of categories is not important since there is no logical transition between categories. However, for ordinal variables, categories represent the order of the labels from the lowest to the highest category in its ordinal scale. For instance, elasticity is an ordinal variable that can take a value from "None" to "Double". In this case, "None" represents the lowest elasticity level and "Double" represents the highest elasticity level of the sock material.

Table 4. Summary of the socks manufacturing dataset.

Variable Name	Data Type	Variable Type	Categories/Range
Raw Material	Categorical	Nominal	Bamboo Lycra Cotton Lycra Cotton Coolmax Lycra Organic Cotton Lycra Modal Lycra
Pattern	Categorical	Symmetric Binary	Yes No
Elasticity	Categorical	Ordinal	None Plain Derby Curly Double
Woven Tag	Categorical	Symmetric Binary	None Label
Heel	Categorical	Symmetric Binary	None Plain
Leg Style	Categorical	Ordinal	None Short Medium Long
Fabric Type	Categorical	Symmetric Binary	Plain Towel
Gender	Categorical	Symmetric Binary	Women Men
Actual Cost	Numeric	Interval Scale	$[0, \infty)$

5.1.2. Electrical grounding parts data – tubular cable lugs

The second application problem dataset was collected from an electromagnetic parts manufacturer which produces lightening protection elements, grounding materials, metal masts for various purposes, and cabins for specific purposes. Steel, copper, stainless steel, aluminum, brass, bronze, cast iron, plastic, and concrete are the primary raw materials used to manufacture these static grounding systems. In the facility, they can coat these materials with electro galvanization, hot deep galvanization, electro copper coating, electro tin coating, electro chromium-nickel (Cr-Ni) coating, black insulation, and green-yellow insulation.

The dataset that we collected from the company's database contains information for various tubular cable lugs of 68 observations. There are 12 variables associated with these 68 observations, namely lug type, cross-section, hole diameter, number of holes, gap between holes, material weight, process time, inner diameter, outer diameter, coating type, coating time, and the actual cost. Ten of these variables are quantitative attributes and nine of them are recorded on continuous scales. These nine continuous valued variables are cross-section, hole diameter, gap between holes, material weight, process time, inner diameter, outer diameter, coating time, and the actual cost, and their units are recorded in mm2, mm, mm, kg, mm, mm, minutes, and TL, respectively. The remaining one quantitative variable takes integer values. The label of the strictly integer valued quantitative variable is the number of holes, and it does not have any measurement units. There are at most two holes on a lug and the minimum number of holes is zero. DIN, forend, long, standard, and forend standard are the categories of the variable lug type. Table 5 is the summary of the dataset and its associated attributes.

5.1.3. Lightening protection parts data – air rods

The third application problem dataset was collected from the same electromagnetic parts manufacturer as in the second problem and includes information about 197 air rods for lightening protection purposes. In the dataset, there are 10 variables associated with these 197 observations. Five of these variables take continuous numeric values and the remaining five are categorical labels. The numeric variables are rod diameter, rod length, screw size, material weight, and the actual cost. The values of these variables are measured with these units, respectively: mm, mm, mm, kg, and TLs. The screw size takes a value of zero when there is no screw used, and the actual minimum screw size is 8.5 mm. The categorical variables are screw type, main material, coating, raw material, and screw nut coating. In Table 6 the summary of the dataset and its associated attributes are shown.

Table 5. Summary of the tubular cable lugs manufacturing dataset.

Variable Name	Data Type	Variable Type	Categories/Range
Lug Type	Categorical	Nominal	DIN Forend Forend Standard Long Standard
Cross-section	Numeric	Interval Scale	$[0, \infty)$
Hole Diameter	Numeric	Interval Scale	$[0, \infty)$
Number of Holes	Numeric	Interval Scale	0, 1, 2, …
Gap b/w Holes	Numeric	Interval Scale	$[0, \infty)$
Material Weight	Numeric	Interval Scale	$[0, \infty)$
Process Time	Numeric	Interval Scale	$[0, \infty)$
Inner Diameter	Numeric	Interval Scale	$[0, \infty)$
Outer Diameter	Numeric	Interval Scale	$[0, \infty)$
Coating	Categorical	Nominal	None Tin
Coating Time	Numeric	Interval Scale	$[0, \infty)$
Actual Cost	Numeric	Interval Scale	$[0, \infty)$

Table 6. Summary of the air rods manufacturing dataset.

Variable Name	Data Type	Variable Type	Categories/Range
Rod Diameter	Numeric	Interval Scale	[16, ∞)
Rod Length	Numeric	Interval Scale	[150, 6000]
Screw Size	Numeric	Interval Scale	[8.5, 16]
Screw Type	Categorical	Nominal	None Interior Screw Exterior Screw
Material Weight	Numeric	Interval Scale	[0, ∞)
Main Material	Categorical	Nominal	Aluminum Copper Iron-Steel Bronze Gray Cast Iron Stainless Steel Brass Plastic
Coating	Categorical	Nominal	No Coating Electro-Galvanizing Hot Dip Galvanizing Electrodeposited Copper Electrodeposited Tin Electrodeposited Cr-Ni Black Insulation Yellow Green Insulation
Raw Material	Categorical	Nominal	Aluminum Rod Ø16 Aluminum Rod Ø20 Brass Rod Ø16 Brass Rod Ø20 Copper Rod 16 x 3000 Copper Rod 16 x 3500 Copper Rod 20 x 3000 Copper Rod 20 x 6000 Stainless Rod Ø16 Stainless Rod Ø20 Transmission Ø16 Transmission Ø20
Screw Nut Coating	Categorical	Nominal	No Screw Nut Non-Coated Galvanized Stainless Brass
Actual Cost	Numeric	Interval Scale	[0, ∞)

5.1.4. Plastic products data

The last dataset was taken from a plastic parts manufacturer which produces kitchenware, food and non-food storage containers, and salad, pastry, bathroom, and hanger accessories. In this dataset, there are many products with completely different physical shapes. However, we may group them according to their raw material types, manufacturing processes/operations, or some other factors. The dataset covers 51 variables for 130 plastic products. There are ten main categories of variables, raw material, press, vacuum, paint, sticker, wall plug, labor complexity, and actual cost. There are 13 variables under the raw material category where 12 of them are binary and one is numeric. These 12 variables represent the type of raw material such as anti-shock, acrylonitrile butadiene styrene (ABS), poly carbon, and carbon fiber. If a material is used in the main material mixture for a particular product, the value of the underlying material variable takes one, otherwise zero. The only variable measured on a continuous scale is mixture weight under the raw material subject. It is recorded in grams. The second variable category is press, which stands for the pressing process. There are three machine groups in the company that can perform press operations. Tederic, TSP, and Haitian are the names of these machine groups. There are 11, eight, and four different machines under the Tederic, TSP, and Haitian groups, respectively. Every machine corresponds to a variable in the dataset. There can be multiple alternative machines to perform the same operation; however, if a machine is used for any step of production for a particular product, its variable takes a numeric value representing the machining time. If the underlying machine is not used for that product, the value of that machine's variable takes a value of zero. The next variable category is for the vacuuming process. There are two variables under the vacuum topic: (1) Poly vinyl chloride (PVC) type for the vacuuming process and (2) the number of vacuums required. The PVC type is a categorical variable and the number of vacuums takes discrete numeric values. Under the boxing category, there are seven variables. Six of these variables are numeric variables and one of them is a categorical variable. These variables are number of items in a box, net weight, gross weight, length, width, depth of the box, and the type of the boxing material. Each remaining category corresponds to a single variable. Package, paint material weight, sticker, wall plug, labor complexity, and actual cost are, respectively, binary, numeric, binary, binary, ordinal, and numeric variables. The unit of the paint material weight is grams. Also, the actual cost is recorded in TLs. Furthermore, the labor complexity is tracked according to the complexity of the manufacturing and assembly operations and ranked from 1 (easiest) to 3 (most complex), sequentially. In Table 7, the summary of the dataset and its associated attributes are shown.

We termed the application problems dataset 1 (DS 1), dataset 2 (DS 2), dataset 3 (DS 3), and dataset 4 (DS 4) for the socks manufacturing, the tubular cable lugs, the air rods, and the plastic products problem sets, respectively.

5.2. Cluster analysis and the number of clusters

As discussed earlier we used Kaufmann and Rousseeuw's (2022) k-medoids algorithm as it was implemented in "PAM". The first target is to determine the appropriate number of clusters. The C-index, the Gamma, and the average silhouette width graphs are the primary tools to choose the appropriate number of clusters. We plotted the values of the underlying indices from 2 to 20 clusters. As expected, the value of Gamma and the average silhouette width increase as the number of clusters increases. The value of the C-index decreases as the number of clusters increases which is consistent with the pattern of the other two indices. The graphs of these three indices with respect to the number of clusters are given in Figures 4 through 7 for test cases DS 1 through DS 4, respectively.

Remember that our policy is to seek a consensus among these three graphs. For DS 1, a settlement point of the indices is seven clusters as shown in Figure 4 with the black points where a local trough is observed right before a dramatic jump in the C-index. Furthermore, at the point of seven clusters, local peaks can be observed one step before the sudden drops in Gamma and silhouette width trends. For DS 2, the silhouette width does not have any value higher than 0.5. However, a local peak is observed at 11 clusters. When we compare the performance of the other two indices with the silhouette width, 11 is a reasonable value as the appropriate number of clusters. Furthermore, after 11 clusters, the cluster contents become unbalanced where too many observations accumulated in some groups. For DS 3, we picked the point where the silhouette width goes above 0.5 for the first time because a value above 0.5 indicates a robust clustering structure. After 14 clusters, the value of silhouette width stagnates right below the 0.5 line. If we check the consistency of silhouette width with the other two statistics, we can see that 14 clusters are appropriate. For DS 4, the silhouette width never moves higher than 0.5, but there is a sudden drop in the C-index value at 10 clusters. When the Gamma index is considered, the value increases slowly to the point at 10 clusters and after that it becomes stable. Combining the information derived from these statistics, we can conclude that 10 is a proper value. There are several other possible points that these indices suggest, but 7, 11, 14, and 10 are the most conspicuous points for DS 1, DS 2, DS 3 and DS 4, respectively, when we monitor these graphs from left to right simultaneously.

Table 7. Summary of the plastic products manufacturing dataset.

Variable Name	Data Type	Variable Type	Categories/Range
Cristal	Categorical	Symmetric Binary	Yes, No
Anti-Shock	Categorical	Symmetric Binary	Yes, No
PP	Categorical	Symmetric Binary	Yes, No
ABS	Categorical	Symmetric Binary	Yes, No
Poly Carbon	Categorical	Symmetric Binary	Yes, No
NAT ABS	Categorical	Symmetric Binary	Yes, No
Randum	Categorical	Symmetric Binary	Yes, No
ESM	Categorical	Symmetric Binary	Yes, No
i20	Categorical	Symmetric Binary	Yes, No
Carbon Fiber	Categorical	Symmetric Binary	Yes, No
Stainless Steel	Categorical	Symmetric Binary	Yes, No
PVC	Categorical	Symmetric Binary	Yes, No
Weight	Numeric	Interval Scale	$[0,\infty)$
Tedeceric 100_1	Numeric	Interval Scale	$[0,\infty)$
Tedeceric 100_2	Numeric	Interval Scale	$[0,\infty)$
Tedeceric 110	Numeric	Interval Scale	$[0,\infty)$
Tedeceric 120	Numeric	Interval Scale	$[0,\infty)$
Tedeceric 140	Numeric	Interval Scale	$[0,\infty)$
Tedeceric 188_1	Numeric	Interval Scale	$[0,\infty)$
Tedeceric 188_2	Numeric	Interval Scale	$[0,\infty)$
Tedeceric 188_3	Numeric	Interval Scale	$[0,\infty)$
Tedeceric 230_1	Numeric	Interval Scale	$[0,\infty)$
Tedeceric 230_2	Numeric	Interval Scale	$[0,\infty)$
Tedeceric 280	Numeric	Interval Scale	$[0,\infty)$
TSP 120_1	Numeric	Interval Scale	$[0,\infty)$
TSP 120_2	Numeric	Interval Scale	$[0,\infty)$
TSP 150_1	Numeric	Interval Scale	$[0,\infty)$
TSP 150_2	Numeric	Interval Scale	$[0,\infty)$
TSP 220	Numeric	Interval Scale	$[0,\infty)$
TSP 250	Numeric	Interval Scale	$[0,\infty)$
TSP 360_1	Numeric	Interval Scale	$[0,\infty)$
TSP 360_2	Numeric	Interval Scale	$[0,\infty)$
Haitian 110	Numeric	Interval Scale	$[0,\infty)$
Haitian 150_1	Numeric	Interval Scale	$[0,\infty)$
Haitian 150_2	Numeric	Interval Scale	$[0,\infty)$
Haitian 250	Numeric	Interval Scale	$[0,\infty)$
PVC Type	Categorical	Ordinal	0, 15, 20
# of Vacuums	Numeric	Interval Scale	$[0,\infty)$
# in box	Numeric	Interval Scale	1, 2, 3, …
Net Weight	Numeric	Interval Scale	$[0,\infty)$
Gross Weight	Numeric	Interval Scale	$[0,\infty)$
Length	Numeric	Interval Scale	$[0,\infty)$
Width	Numeric	Interval Scale	$[0,\infty)$
Depth	Numeric	Interval Scale	$[0,\infty)$
Type	Categorical	Nominal	Blister, Polybag, Display Box, Bound, Card, PVC Shrink, Sticker, Box
Package	Categorical	Symmetric Binary	Yes, No
Paint Weight	Numeric	Interval Scale	$[0,\infty)$
Sticker	Categorical	Symmetric Binary	Yes, No
Wall Plug	Categorical	Symmetric Binary	Yes, No
Labor Complexity	Categorical	Ordinal	1, 2, 3
Actual Cost	Numeric	Interval Scale	$[0,\infty)$

Table 8 shows the number of observations allocated to each cluster using k-medoids based on the chosen number of clusters for each application dataset. When we analyze the individual observations in each cluster, it is easy to see that the categorical variables play an important role in forming the cluster contents. Also, we plotted the minimum (min), maximum (max), and average (mean) actual cost values of products allocated in each cluster in

Figures 8 through 11 for DS 1 through DS 4, respectively. These graphs are provided to illustrate how actual cost values strongly overlap among clusters for the most cases. It is interesting to observe that the similarity of products does not necessarily follow the same similarity pattern of the actual cost values. Since multiple cost drivers contribute to product cost, there is no single factor determining the cluster contents. The interactions of multiple cost drivers are more influential than a single one for each product.

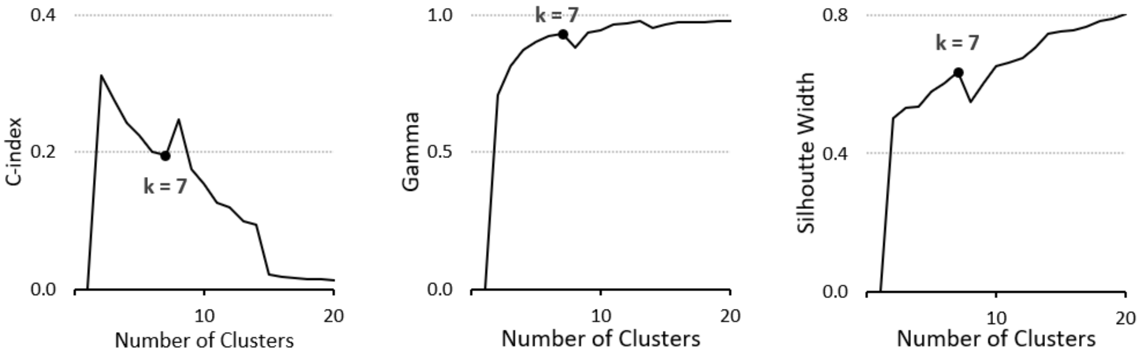


Figure 4. C-index, Gamma and silhouette width plots for DS 1 of 76 products.

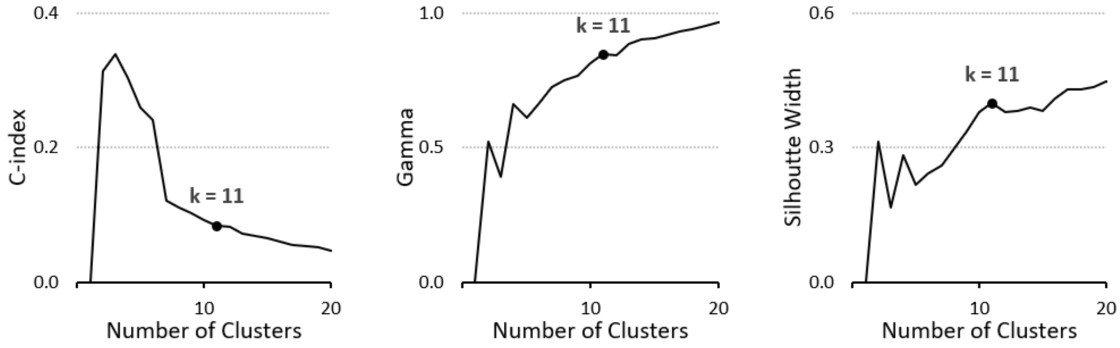


Figure 5. C-index, Gamma and silhouette width plots for DS 2 of 68 products.

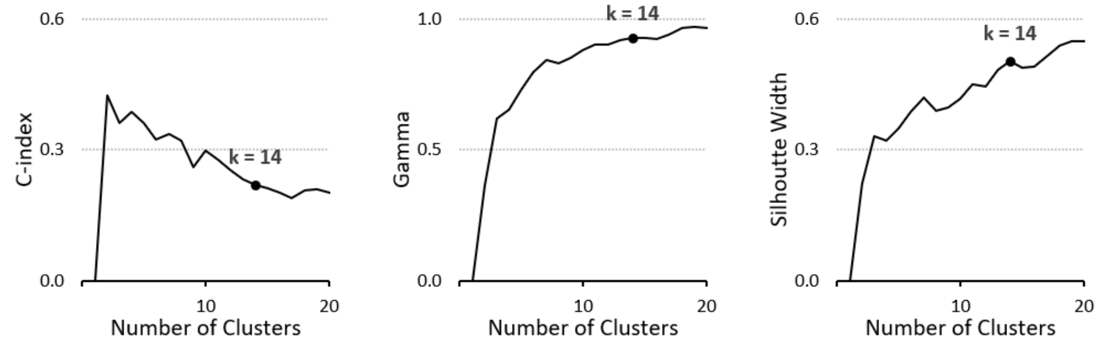


Figure 6. C-index, Gamma and silhouette width plots for DS 3 of 197 products.

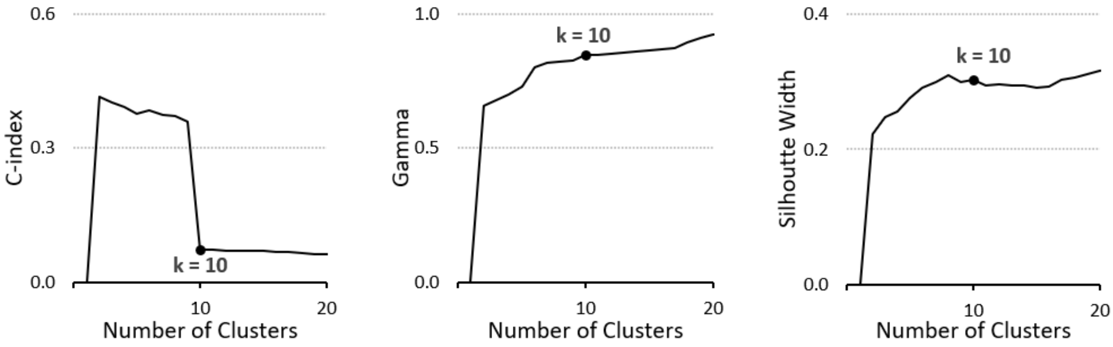


Figure 7. C-index, Gamma and silhouette width plots for DS 4 of 130 products.

Table 8. The number of observations in each cluster for the test cases.

Cluster No	DS 1	DS 2	DS 3	DS 4
1	37	10	26	24
2	11	9	23	20
3	11	8	23	17
4	6	8	17	16
5	5	7	16	15
6	3	5	16	10
7	3	5	14	8
8		5	13	8
9		4	9	7
10		4	9	5
11		3	8	
12			8	
13			8	
14			7	

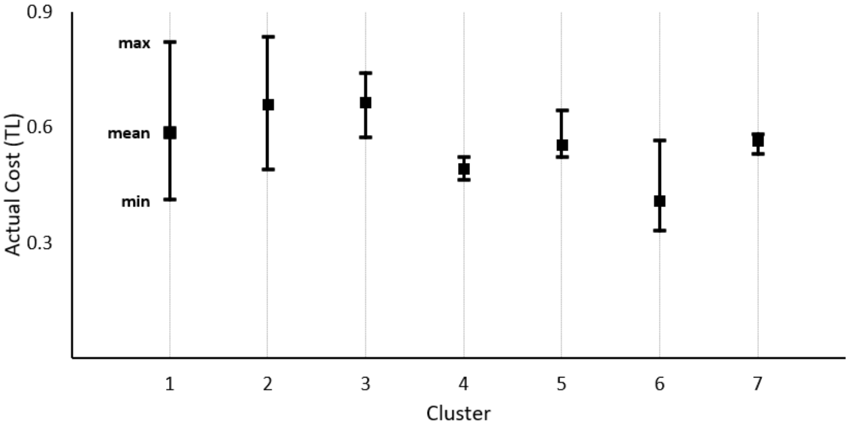


Figure 8. The minimum (min), maximum (max), and average (mean) actual cost values of objects allocated in each cluster for DS 1.

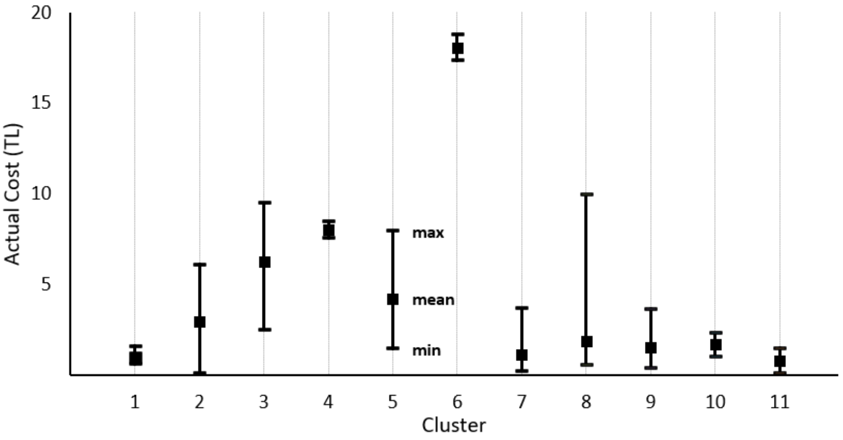


Figure 9. The minimum (min), maximum (max), and average (mean) actual cost values of objects allocated in each cluster for DS 2.

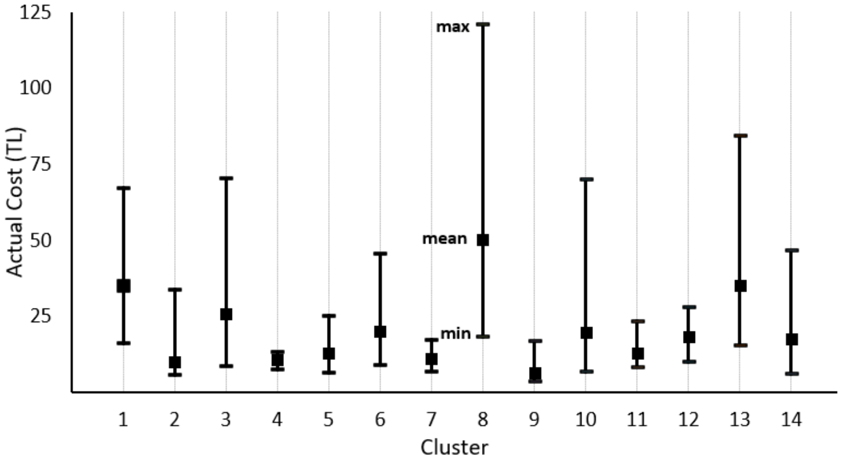


Figure 10. The minimum (min), maximum (max), and average (mean) actual cost values of objects allocated in each cluster for DS 3.

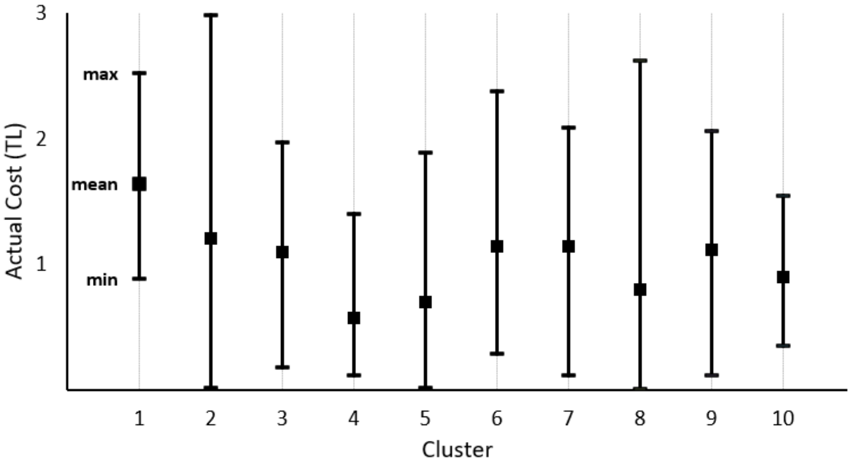


Figure 11. The minimum (min), maximum (max), and average (mean) actual cost values of objects allocated in each cluster for DS 4.

5.3 Spline model parameters

As discussed earlier we used the R package called "crs" to build spline models in the presence of categorical and numeric design attributes, but none of the continuous predictors came out to be a higher degree than cubic splines considering the cross-validated set of parameters. When the polynomial degree of a predictor is zero, the variable is automatically removed from the spline model due to its irrelevance. We ran the spline model script with both "additive" and "tensor" inputs initially. The results show that using tensor products (that is, including interaction terms) provided slightly more accurate results. For the final input parameter, "knots", we let the cross-validation decide the best knot placement strategy. See Table 9 for the complete set of parameters used for the spline models.

Table 9. The R "crs" function input parameters used to build the spline models.

Parameter	Value
degree.max	10
degree.min	0
segments.max	10
segments.min	1
cv	NOMAD
cv.func	cv.ls
complexity	degree-knots
basis	tensor
knots	auto

5.4 Results and discussion

As we discussed earlier, we used leave-one-out cross-validation to leverage the data for both validation and model building. Without proper validation, our methodology would not have credibility to be used in a real-life business environment. This validation module is fully integrated in the same R script.

In Table 10, we present the performance metrics of each cost estimation approach, termed CLU (clustering), SPL (splines), and REG (regression) for the four test cases, DS 1, DS 2, DS 3, and DS 4. These metrics are the mean absolute relative error (MARE) and the max absolute relative error (Max ARE) over the validated predictions for each product. Notice that SPL does not have error defined for DS 1 in the table since this dataset does not contain any continuous predictors to form a spline basis. The minimum values of MARE and Max ARE are depicted in bold in Table 10 for each dataset. According to the MARE values, the most accurate cost estimation approach is CLU based on overall performance. However, SPL generates slightly more accurate predictions for DS 3 compared to CLU. Clearly, REG was outperformed by both CLU and SPL. It is a little difficult to decide which cost estimation method is superior between CLU and SPL. SPL is not applicable to the first dataset (DS 1) since there are no continuous predictors to build a spline basis. This is a disadvantage for wholly categorical or qualitative datasets. A second aspect is that SPL was significantly bettered by CLU for DS 2. But, when we consider Max ARE values, SPL did better than CLU for two of the three test cases. Of paramount importance, both CLU and SPL were able to predict the manufacturing cost of products with good accuracy, especially compared to the often-used REG method. Figure 12 shows the performance of the cost estimation methods over the four data application problems in terms of the MARE values given in Table 10.

We also evaluated the performance of spline models by setting the maximum polynomial degree to 1 to make a fair comparison between SPL and CLU, and SPL and REG because CLU and REG are basically linear models in our test cases. Furthermore, we removed the interaction terms in the spline models by setting the "basis" input as "additive" to eliminate interaction terms. The performance difference between the default tensor product SPL model and the linear additive SPL model was minimal and these changes did not affect its overall accuracy. The linear additive SPL model still outperformed REG by far. We can conclude that even considering suboptimal spline model parameters, SPL is a better alternative than REG.

Table 10. Performance metrics of each cost estimation model for the application problems.

MARE			
	CLU	SPL	REG
DS 1	6.25%	N/A	8.54%
DS 2	4.98%	38.70%	49.82%
DS 3	5.81%	4.08%	15.42%
DS 4	12.39%	17.55%	33.83%
Max ARE			
	CLU	SPL	REG
DS 1	49.12%	N/A	49.82%
DS 2	46.67%	162.01%	429.52%
DS 3	56.04%	26.23%	64.36%
DS 4	203.54%	94.73%	233.79%

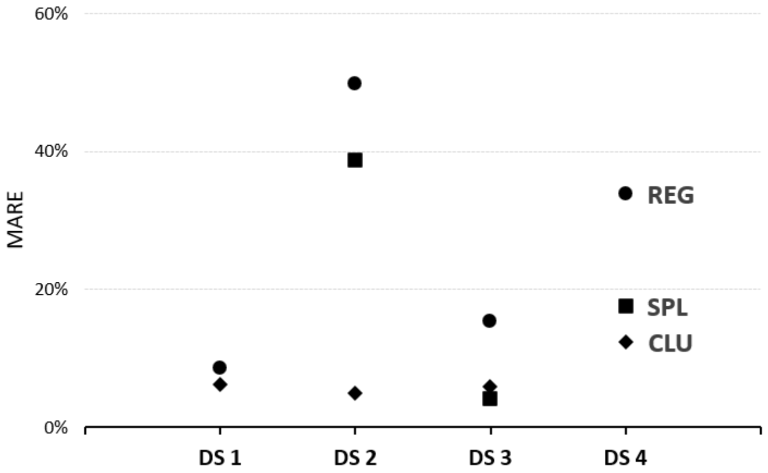


Figure 12. Performance of the cost estimation approaches in terms of MARE.

We used a paired t-test to evaluate the significance of the mean of the differences in AREs. In Table 11, p-values for the paired t-tests on the mean of the differences are given. All cost estimation approaches produce significantly different ARE results than each other at a 95% confidence level. Therefore, we can conclude that there is a clear dominance in the performance of CLU compared to REG and SPL compared to REG. However, for the CLU and SPL pair, we could not conclude if one of them is superior over the other because for only DS 2, CLU demonstrates a clear dominance when MARE values are considered. For DS 3, SPL turns out to be the best approach but very close in performance to CLU. For the last application dataset, DS 4, CLU finds slightly more accurate estimated values than SPL.

We also considered the sensitivity of MARE with respect to the number of clusters for CLU. As expected, MARE decreases as the number of clusters increases and finally it converges to a limit value. The limit MARE values are around 5%, 3%, 4%, and 11% for the test cases DS 1 through DS 4, respectively. Figure 13 shows the change in MARE values when the number of clusters increases for each application dataset. Even though increasing the number of clusters results in more accurate estimates, it might be likely to be over-parameterized which results in a less robust and less dependable model.

Table 11. p-values for the paired t-tests of the pairs of cost estimation approaches.

DS 1	REG	SPL
CLU	9.12×10^{-8}	N/A
SPL	N/A	
DS 2	REG	SPL
CLU	6.65×10^{-9}	2.68×10^{-9}
SPL	0.0003	
DS 3	REG	SPL
CLU	2.52×10^{-17}	3.44×10^{-15}
SPL	3.62×10^{-18}	
DS 4	REG	SPL
CLU	1.50×10^{-22}	1.33×10^{-18}
SPL	2.58×10^{-25}	

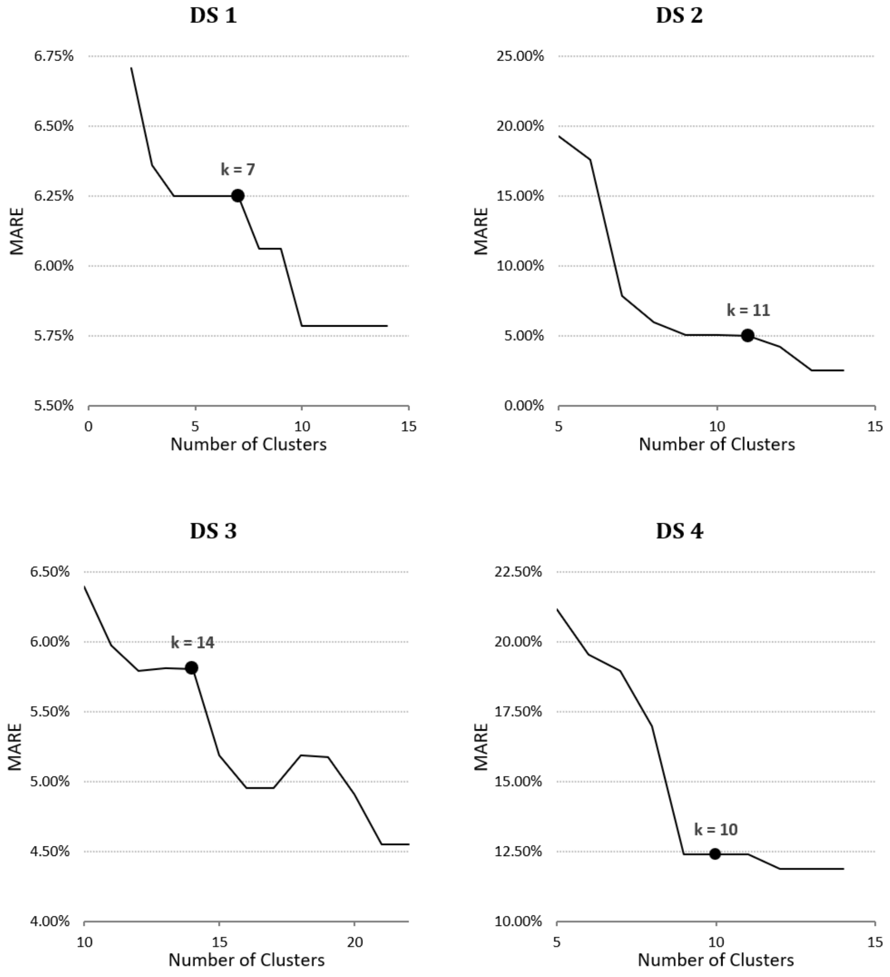


Figure 13. MARE vs. number of clusters of each application problem for CLU.

We provide the R^2 values for each cost estimation method in Table 12. The maximum R^2 (R-sq) value for each data set is in bold to show the best model fit among the three methods. The R^2 values of CLU and REG from the table show that finding a well-suited model for DS 1 is challenging due to lack of relevant continuous predictors in the dataset. Adding more variables to the cost estimation for DS 1 might increase the true explanatory power of the models but unfortunately the dataset was strictly limited to only eight categorical predictors. However, this dataset is atypical as most manufactured products include both numeric and categorical cost drivers. For the other datasets, each cost estimation approach can explain the total variability with a high R^2. For a better illustration of R^2 values, we plotted the fitted values (predicted cost) by observed values (actual cost) in Figures 14 through 17 for DS1 through DS 4, respectively.

Table 12. Coefficient of determination (R^2) values for the MCE approaches.

R^2	CLU	SPL	REG
DS 1	63.49%	N/A	53.19%
DS 2	99.94%	91.50%	84.63%
DS 3	96.83%	99.52%	90.49%
DS 4	93.69%	88.46%	76.47%

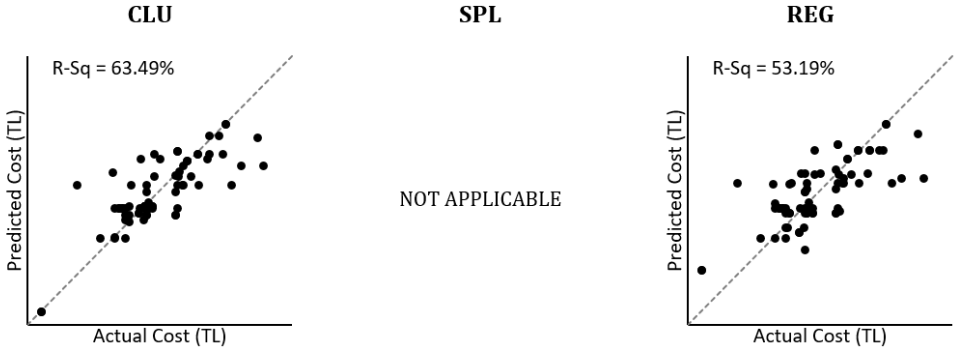


Figure 14. Fitted values (predicted cost) vs. observed values (actual cost) along with the R^2 values (R-Sq) for DS 1.

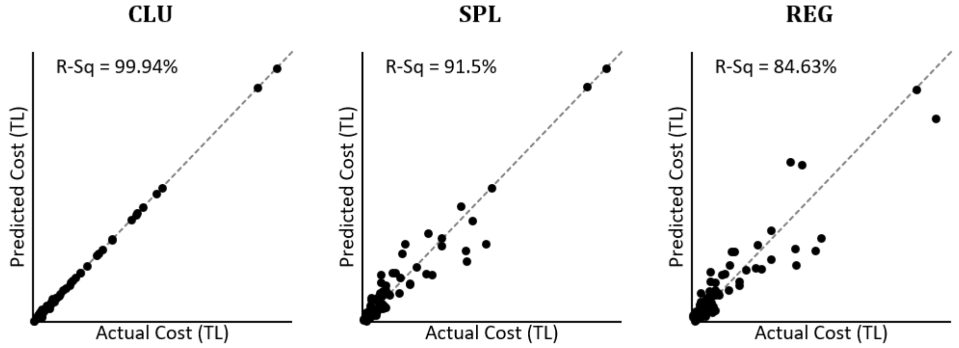


Figure 15. Fitted values (predicted cost) vs. observed values (actual cost) along with the R^2 values (R-Sq) for DS 2.

6. Manufacturing cost estimation user interface

This software system is available for open access at the link below: https://github.com/erensakinc/MCE.

This GitHub repository includes full directions for running the software and also includes the data sets we used in this paper. We built an interface using the R package called "shiny" (Chang, 2016). It is a web application framework to turn R scripts into interactive web applications. It has two main components: (1) Server-side component that is responsible for the computational tasks and rendering plots and tables, and (2) User interface component that is the actual interactive web interface with input entering elements such as check boxes, radio and

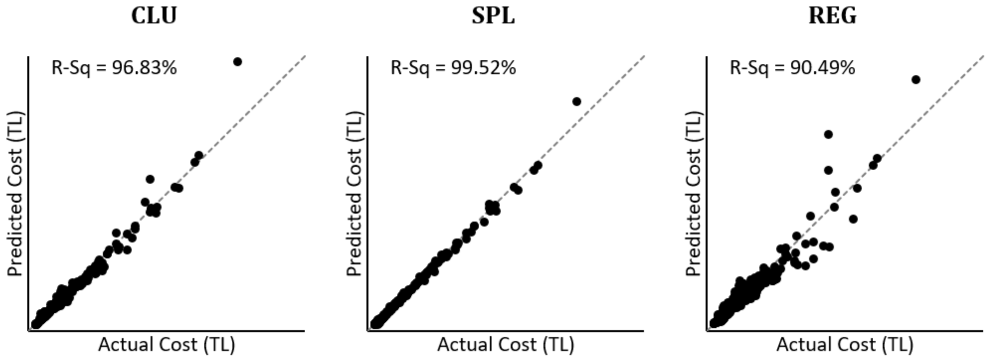


Figure 16. Fitted values (predicted cost) vs. observed values (actual cost) along with the R^2 values (R-Sq) for DS 3.

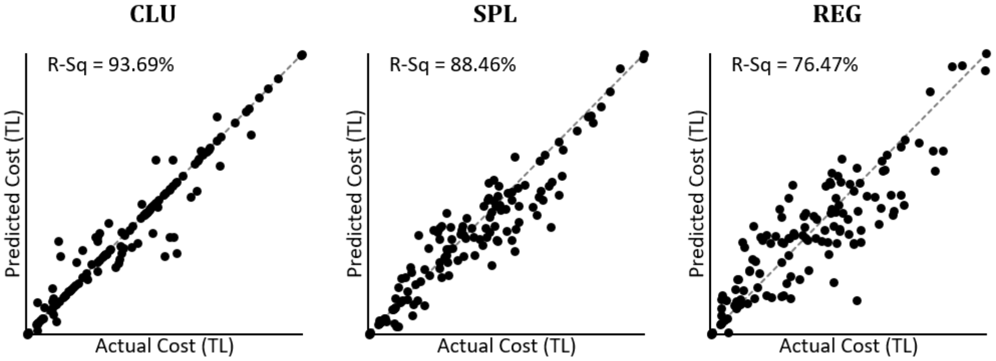


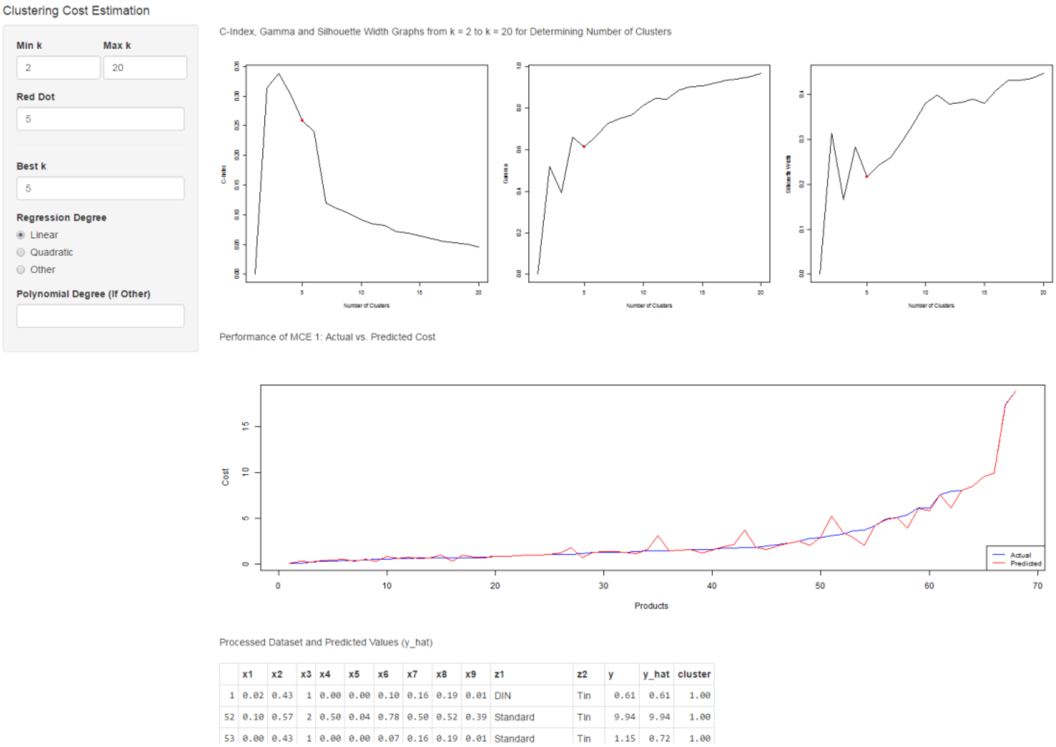
Figure 17. Fitted values (predicted cost) vs. observed values (actual cost) along with the R^2 values (R-Sq) for DS 4.

action buttons, and other numeric and text input boxes. The interface is a web-based application, and it is published online for cost estimation practitioners. It consists of four main tabs: (1) Load Data, (2) CLU, (3) SPL, and (4) REG.

The "Load Data" tab is for uploading a dataset to the system in a comma separated values format. In this tab, the user enters a vector representing the variable types as discussed earlier. The "CLU" tab is for the clustering-based cost estimation approach. It has two main parts. The first part has three inputs, namely the minimum number of clusters, the maximum number of clusters, and a red dot to mark the selected number of clusters on the graphs. The second part has two inputs, the best number of clusters and the polynomial regression model degree: linear, quadratic, or a higher degree. The interface passes the given information to the server and the server-side application renders the C-index, Gamma, and silhouette width graphs based on the minimum and maximum number of clusters. The user is required to enter the preferred number of clusters to proceed to the cost estimation step. When the selected number of clusters is entered, the application builds the final cluster contents and cluster specific estimation models and then produces the actual cost vs. predicted cost graph along with a table of predicted values (the column name is y_hat) for each data point. In this table, there is an extra column called "cluster" that shows in which cluster the specific data point is classified. A screenshot of the CLU tab after solving a cost estimation problem is given in Figure 18.

The "SPL" tab is for the spline-based cost estimation approach. The spline model inputs are maximum and minimum spline degrees, maximum and minimum number of segments, optimization complexity, knot placement strategy, spline basis, optimization algorithm, and the cross-validation function. All inputs are passed to the "crs" package and then a categorical spline regression model is constructed to predict manufacturing costs. The output is like the “CLU” tab’s output. It generates a graph of the actual vs. predicted costs and a table of predicted values.

The last tab, “REG”, represents the traditional cost estimation approach, a single polynomial regression model. It only has a single input for the regression degree. Once the regression degree is determined (selected) a similar output is generated where the actual vs. predicted cost graph and the table of predicted values are shown.



Clustering Cost Estimation

Min k: 2
Max k: 20
Red Dot: 5
Best k: 5

Regression Degree
● Linear
○ Quadratic
○ Other

Polynomial Degree (If Other):

C-Index, Gamma and Silhouette Width Graphs from k = 2 to k = 20 for Determining Number of Clusters

Performance of MCE 1: Actual vs. Predicted Cost

Processed Dataset and Predicted Values (y_hat)

	x1	x2	x3	x4	x5	x6	x7	x8	x9	z1	z2	y	y_hat	cluster
1	0.02	0.43	1	0.00	0.00	0.10	0.16	0.19	0.01	DIN	Tin	0.61	0.61	1.00
52	0.10	0.57	2	0.50	0.04	0.78	0.50	0.52	0.39	Standard	Tin	9.94	9.94	1.00
53	0.00	0.43	1	0.00	0.00	0.07	0.16	0.19	0.01	Standard	Tin	1.15	0.72	1.00

Figure 18. Clustering based cost estimation (CLU) application tab after analysis.

7. Conclusions

In this paper, we investigated ways of using piecewise functions formed by either clustering or splines to predict the manufacturing cost of a product prior to actually manufacturing it. In real applications, the most likely scenario is to have a set of data about the products and their cost related attributes (drivers) where these attributes are mixed categorical and numerical, as we consider. The accuracies of the two novel methodologies presented in this work are assessed in comparison to each other and to also a regression model with the absence of clustering approaches (this latter approach being common practice in industry). We did not compare with some other data driven alternatives such as neural networks for a few reasons. First, neural networks require large data sets to perform adequately on multi-variate prediction and for cost estimation, often data sets are quite small. Second, building and validating a neural network is quite artful requiring considerable experience and judgement on the part of the analyst.

Our results show that predictions are more accurate taking a clustering approach, which could translate into more profitability and sales to organizations because they could price their manufactured goods appropriately. This would avoid too low pricing which could result in less profitability or even losses or too high pricing which could deter sales by not being competitive. One limitation of our approach is that the future product to be manufactured is related in a cost manner to past products whose manufacturing costs are already known. The known cost data must be representative as this method is data driven and is largely dependent on the integrity of the data used. Another consideration is that the number of clusters must be ascertained. While using the metrics discussed in the paper simplifies this process, it is not automatic. Finally, the computational effort is quite modest for these small sized data sets but might be more of a concern for very large data sets.

One existing method of cost estimation is regression trees, and this does offer a useful future research focus. A regression tree is a variant of decision trees where real-valued functions are approximated. The regression tree methodology may be generalized to manufacturing cost estimation since it is not limited to continuous predictors only. That is, using mixed numeric and categorical data is allowed in the regression tree building process.

In this research, irrelevant predictors are removed from the CLU, SPL, and REG models as described earlier. Future research may consider the information gain criterion when deciding on the inclusion of a candidate predictor in the cost estimation model. This approach could yield an information rich but parsimonious set of cost drivers to be used in predicting cost using our clustering or spline approach. A further refinement may be to use a dimension reduction method such as principal component analysis in lieu of the cost drivers themselves.

Funding Statement

This research received no external funding.

Acknowledgment

Acknowledgments to anonymous referees' comments and editor's effort.

Conflict of interest

All the authors claim that the manuscript is completely original. The authors also declare no conflict of interest.

References

Almond, D., Chay, K. Y., & Lee, D. S. (2005). The cost of low birth weight. *The Quarterly Journal of Economics* 120(3), 1031-1084. https://doi.org/10.1093/qje/120.3.1031

Al-Sultan, K. S. (1995). A tabu search approach to the clustering problem. *Pattern Recognition* 28(9), 1443-1451. https://doi.org/10.1016/0031-3203(95)00022-R

Angelis, L. & Stamelos, I. (2000). A simulation tool for efficient analogy based cost estimation. *Empirical Software Engineering* 5(1), 35-68. https://doi.org/10.1023/A:1009897800559

Audet, C., Le Digabel, S. & Tribes, C. (2009). NOMAD user guide. Les cahiers du GERAD, *Technical Report* G-2009-37. https://www.gerad.ca/fr/papers/G-2009-37.pdf

Baker, F. B. & Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association* 70(349), 31-38. https://doi.org/10.2307/2285371

Carides, G. W., Heyse, J. F. & Iglewicz, B. (2000). A regression-based method for estimating mean treatment cost in presence of right-censoring. *Biostatistics* 1(3), 299-313. https://doi.org/10.1093/biostatistics/1.3.299

Chang, W. (2016). Package 'shiny': Web application framework for R, R Package version 0.13.2. Retrieved from https://github.com/rstudio/shiny/

Cheng, C.-H. (1995). A branch and bound clustering algorithm. *IEEE Transactions on Systems, Man and Cybernetics* 25(5), 895-898.

Curry, H. B. & Schoenberg, I. J. (1947). On spline distributions and their limits: The Polya distribution functions. *Bulletin of the American Mathematical Society* 53, no. 1114.

Dai, J. S., Niazi, A., Balabani, S. & Seneviratne, L. (2006). Product cost estimation: Technique classification and methodology review. *Journal of Manufacturing Science and Engineering* 128(2), 563-575. https://doi.org/10.1115/1.2137750

Dalrymple-Alford, E. C. (1970). Measurement of clustering in free recall. *Psychological Bulletin* 74(1), 32-34.

de Boor, C. (1976). A Practical Guide to Splines. New York: Springer-Verlag.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics* 3(3), 32-57.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* 11(2), 89-102. https://doi.org/10.1214/ss/1038425655

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2010). *Cluster Analysis*. Chichester: John Wiley & Sons.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* 49(268), 732-764.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27(4), 857-871.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining Conference*, 21-34.

Huang, Z. (1998). Extensions to the *k*-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery* 2(3), 283-304.

Jones, D. R., & Beltramo, M. A. (1991). Solving partitioning problems with genetic algorithms. *Proceedings of the Fourth International Conference on Genetic Algorithms*, 442-449.

Kaufmann, L., & Rousseeuw, P. (1987). *Clustering by means of medoids. In Statistical Data Analysis Based on the L1-norm and Related Methods*, 405-416. Amsterdam: Springer.

Kaufmann, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data*. New York: John Wiley & Sons.

Koontz, W. L. G., Narendra, P. M., & Fukunaga, K. (1975). A branch and bound clustering algorithm. *IEEE Transactions on Computers* 24(9), 908-915.

Layer, A., Brinke, E.T., Houten, F.V., Kals, H., & Haasis, S. (2002). Recent and future trends in cost estimation. *International Journal of Computer Integrated Manufacturing*, 15(6), 499-510. https://doi.org/10.1080/09511920210143372

Lee, A., Cheng, C.H., & Balakrishnan, J. (1998). Software development cost estimation: integrating neural network with cluster analysis. *Information & Management*, 34(1), 1-9. https://doi.org/10.1016/S0378-7206(98)00041-X

Li, Q., & Racine, J.S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Ma, S., Racine, J.S., & Yang, L. (2014). Spline regression in the presence of categorical design predictors. *Journal of Applied Econometrics*, 10(5), 705-717. https://www.jstor.org/stable/26609055

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297.

Michaud, K., Messer, J., Choi, H.K., & Wolfe, F. (2003). Direct medical costs and their predictors in patients with rheumatoid arthritis. *Arthritis and Rheumatism*, 48(10), 2750-2762. https://doi.org/10.1002/art.11439

Milligan, G.W., & Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2), 159-179.

Monmarché, N., Slimane, M., & Venturini, G. (1999). On improving clustering in numerical databases with artificial ants. *In Advances in Artificial Life* (pp. 626-635). Springer. https://link.springer.com/chapter/10.1007/3-540-48304-7_83

Nie, Z., & Racine, J.S. (2012). The crs package: Nonparametric regression splines for continuous and categorical predictors. *The R Journal*, 4.2, 48-56. https://doi.org/10.32614/RJ-2012-012

Omran, M., Salman, A., & Engelbrecht, A.P. (2002). Image classification using particle swarm optimization. *Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution and Learning*, 370-374. https://link.springer.com/chapter/10.1007/978-3-540-34956-3_6

Pahariya, J.S., Ravi, V., & Carr, M. (2009). Software cost estimation using computational intelligence techniques. *World Congress on Nature and Biologically Inspired Computing*, 849-854 https://doi.org/10.1109/NABIC.2009.5393534 .

Pal, N.R., Bezdek, J.C., & Tsao, E.C.-K. (1993). Generalized clustering networks and Kohonen's self-organizing scheme. *IEEE Transactions on Neural Networks*, 4(4), 549-557. https://doi.org/10.1109/72.238310

Racine, J.S., Nie, Z., & Ripley, B.D. (2014). Package 'crs': Categorical regression splines. *R Package version* 0.15-24. Retrieved from https://github.com/JeffreyRacine/R-Package-crs/

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

SAS/STAT 9.2 *User's Guide*. (2008). SAS Institute Inc.

Schumaker, L.L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618994

Selim, S.Z., & Al-Sultan, K.S. (1991). A simulated annealing algorithm for the clustering problem. *Pattern Recognition*, 24(10), 1003-1008. https://doi.org/10.1016/0031-3203(95)00022-R

Sneath, P.H.A. (1957). The application of computers to taxonomy. *Microbiology*, 17(1), 201-226.

Sokal, R.R., & Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 38, 1409-1438.

Sørenson, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 5, 1-34.

Valverde, S.C., & Humphrey, D.B. (2004). Predicted and actual costs from individual bank mergers. *Journal of Economics and Business*, 56, 137-157. https://doi.org/10.1016/j.jeconbus.2003.05.001

Van Hai, V., Nhung, H.L.T.K., Prokopova, Z., Silhavy, R., & Silhavy, P. (2022). Toward improving th e efficiency of software development effort estimation via clustering analysis. *IEEE Access*, 10, 83249-83264. https://doi.org/10.1109/ACCESS.2022.3185393 .

Ward Jr, J.H. (1996). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244.

Wolfe, J.H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5(3), 329-350.

Xu, Z., & Khoshgoftaar, T.M. (2004). Identification of fuzzy models of software cost estimation. *Fuzzy Sets and Systems*, 145(1), 141-163. https://doi.org/10.1016/j.fss.2003.10.008

Zahn, C.T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, 100(1), 68-86. doi: 10.1109/T-C.1971.223083