Expert Systems with Applications 38 (2011) 13448–13467

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Review

A review of data mining applications for quality improvement in manufacturing industry

Gülser Köksal [a,*], İnci Batmaz [b], Murat Caner Testik [c]

[a] Department of Industrial Engineering, Middle East Technical University, Inonu Blvd., Ankara 06531, Turkey
[b] Department of Statistics, Middle East Technical University, Ankara 06531, Turkey
[c] Department of Industrial Engineering, Hacettepe University, Ankara 06800, Turkey

ARTICLE INFO

Keywords:
Knowledge discovery in databases
Data mining
Quality improvement
Six sigma
Design for six sigma
Quality description
Prediction
Classification
Parameter optimisation
Data mining software
Manufacturing

ABSTRACT

Many quality improvement (QI) programs including six sigma, design for six sigma, and *kaizen* require collection and analysis of data to solve quality problems. Due to advances in data collection systems and analysis tools, data mining (DM) has widely been applied for QI in manufacturing. Although a few review papers have recently been published to discuss DM applications in manufacturing, these only cover a small portion of the applications for specific QI problems (quality tasks). In this study, an extensive review covering the literature from 1997 to 2007 and several analyses on selected quality tasks are provided on DM applications in the manufacturing industry. The quality tasks considered are; product/process quality description, predicting quality, classification of quality, and parameter optimisation. The review provides a comprehensive analysis of the literature from various points of view: data handling practices, DM applications for each quality task and for each manufacturing industry, patterns in the use of DM methods, application results, and software used in the applications are analysed. Several summary tables and figures are also provided along with the discussion of the analyses and results. Finally, conclusions and future research directions are presented.

1. Introduction

Quality improvement (QI) of industrial products and processes requires collection and analyses of data to solve quality related manufacturing problems. While QI programs such as six sigma (6σ), design for six sigma (DFSS), and *kaizen* keep encouraging collection of data to attack quality problems, with advances in automation and computer systems, data from manufacturing processes is becoming more and more available. Although traditional data analysis tools have been successfully used in improving quality of products and processes, now better tools exist to mine massive data sets collected through computerized systems in the industry.

Quality problems may involve several input and output variables that are not easy to model and/or optimise. As Yang and Trewn (2004) point out, data mining (DM) and knowledge discovery in databases (KDD) have been successfully used for solving QI and control problems involving multivariate data in various stages of product/process life cycles. Yet, a review of such problem solving approaches is needed to guide practitioners.

The aim of this study is to extensively review DM applications in manufacturing industries to handle QI problems that are of wide interest to practitioners as well as to outline related future research directions and challenges for researchers. Recently, some other reviews of DM applications in manufacturing and/or product design have also appeared in the literature (see Braha, 2001; Choudhary, Harding, & Tiwari, 2008; Feng & Kusiak, 2006; Harding, Shahbaz, Srinivas, & Kusiak, 2006; Kusiak, 2006; Pham & Afify, 2005; Wang, Tong, Eynard, Roucoules, & Matta, 2007a). However, these reviews are concerned with DM applications in manufacturing in general, and cover only a small sample of the papers that are reviewed in this study. Furthermore, their context is too broad to answer the following open research questions in the scope of our interest:

1. What are the characteristics of data collected in these applications? Which data preparation and preprocessing approaches are used?
2. How many DM applications have been published (within the scope of this study) in different manufacturing industries and how does this number change over time? Which manufacturing industries have applied DM frequently and for what purpose (quality task)? Which DM methods are commonly used for each quality task and manufacturing industry?

* Corresponding author. Tel.: +90 312 2102285; fax: +90 312 2104786.
E-mail addresses: koksal@ie.metu.edu.tr (G. Köksal), ibatmaz@metu.edu.tr (İ. Batmaz), mtestik@hacettepe.edu.tr (M.C. Testik).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.063

3. Do these applications contribute to the literature in terms of new methods, new methodologies, or successful implementation of the existing ones? What are the reasons of selecting these methods? What are the reported benefits and shortcomings of these methods? What are the results of these applications?
4. Which typical patterns or sequences of DM functions and methods are used in these applications to solve the QI problems?
5. Which software has been used in these applications? What are the reasons of selecting these software products? What are the reported benefits and shortcomings of the software?

The review of the literature covers the publications from 1997 through 2007. Applications in the manufacturing sector in general are considered and the publications are selected for the review based on the selection criteria provided in Table 1.

Each publication selected for review is analysed in detail to fill out a table that summarises its basic characteristics which are of interest in this study as shown in Table 2. This table serves as the major source of information to answer the research questions. For this purpose, the applications are sorted according to a characteristic under consideration, and/or simple statistics are calculated for those that satisfy a certain property. The complete table is available in Köksal, Batmaz, and Testik (2008).

Classification of DM functions and methods for a particular application area is an important activity on its own. Even though DM functions are typically classified as data summarisation, classification, prediction, clustering and so on, these classes depend very much on the type of knowledge that can be discovered in databases (Choudhary et al., 2008; Wei, Piramuthu, & Shaw, 2003). Assigning the methods to the functions is also challenging, since some DM methods serve multiple DM functions. In this study, DM functions and methods used in the applications are classified based on the desired knowledge to be discovered from the quality data and the way they are used in the applications.

In the following sections, a classification of QI and control tasks within the scope of this study is given first. Next, the KDD process, relevant DM functions are briefly described. A comprehensive analysis of the reviewed literature from various points of view is provided subsequently. The analysis summarises data handling practices, discusses the application studies for each quality task and for each manufacturing industry, presents patterns of DM function and method usage, examines results of the applications, and surveys the software used in these applications. Finally, conclusions are presented and future research directions are discussed.

2. Quality improvement tasks

A classification of the QI and control tasks and the context of the review are provided in the following. The classification is based on the purpose of a study, and it does not necessarily yield disjoint classes, since the final purpose of a study can be achieved in a hierarchy of lower level purposes. In addition, the aim of this classification is not to identify disjoint classes, but to present an analysis of the DM applications considering different purposes in QI. Hence, before such a classification is presented, some background on various quality initiatives and commonly used quality problem solving approaches is given as well as the QI and control activities that occur in a life cycle of a product.

As quality gained more and more importance over time, many quality initiatives and concepts have emerged. Inspection (100%), statistical quality control (SQC), total quality control (TQC), zero defects, total quality management (TQM), *kaizen*, ISO 9000 quality standards, quality award programs (Malcolm Baldrige, European Quality Award and so on), 6σ, DFSS, lean six sigma have been among the most recognised ones (Fasser & Brettner, 2002; Montgomery, 2005). One apparent trend in the evolution of these initiatives is the emphasis on more proactive approaches and "upstream (design) processes" (Kolarik, 1995; Montgomery, 2005; Phadke, 1989; Taguchi, Chowdhury, & Taguchi, 2000). Furthermore, emphasis on the bottom-line results in shorter time periods has also gained importance and led to the 6σ quality programs. These programmes have received considerable recognition around the world due to their success, especially in improvement of "on-line" or downstream processes using the so called define-measure-analyse-improve-control (DMAIC) approach to reach six sigma (less than 3.4 part per million (PPM) defectives) quality levels (Brady & Allen, 2006; Fasser & Brettner, 2002). Realising that such high

Table 1
Literature selection criteria and assessment methods used in the review.

Criteria	Desired value	Assessment
Contextual criteria		
Purpose of the study	Quality characterization and/or quality prediction and/or quality classification and/or parameter optimisation	Analyzing the abstract, keywords; identifying types of quality tasks involved, relevance of the cited references to the domain
Existence of an industrial application of a DM approach	Yes	Analyzing the abstract, keywords; identifying types of DM functions and methods involved, existence of case specific results, relevance of the cited references to the domain
Industrial sector of the application	Manufacturing	Checking if the type of product or process that the application is concerned with is covered by ISIC (2009) Section C.
Bibliographical criteria		
Date of publication	January 1997–December 2007	Checking if the publication's date of print is within the desired time period
Type of publication	Full paper, conference proceeding, book, book chapter, thesis/dissertation (if a full paper or a conference proceeding is published about the thesis later, then ignored)	Analyzing the source of publication
Source of publication	Covered by the databases[a] searched by the keywords[b]; references cited by the reviewed papers	Reaching the sources by searching the selected databases, and references of reviewed papers

[a] Databases used: Science Direct, Taylor & Francis Online Journals, Academic Search Complete, Wiley Interscience Journals, Business Source Complete, IEL-IEEE/IEE Electronic Library, Emerald Management Extra, MathSciNet, Oxford Online Journals, Springer Link, Web of Science – Science Citation Index Expanded, Ebrary, SIAM, Dissertations and Thesis, Elsevier Electronic Books, Safari Books Online, Springer Book series, Springer Electronic Books, Book Index with Reviewers, ENGnet BASE, SCOPUS

[b] Keywords used: quality, quality improvement, quality control, quality assurance, manufacturing, intelligent manufacturing, artificial intelligence, data mining, knowledge discovery, machine learning, inductive learning, design, product design, process design, design optimisation, parameter optimisation, parameter design, process optimisation, optimisation, prediction, classification, product characterization, process characterization, data analysis, clustering, association, failure analysis, yield improvement, six sigma, neural network, genetic algorithm, Taguchi method, decision tree, fuzzy logic, regression

Table 2
Description of the data table for the reviewed application studies.

ID	Purpose	Product or process	ISIC	Data source	Input variables	Output variables	Quality tasks	Data handling and DM functions	DM methods	Software used	Results
Reference information for the application study	Description of the major purpose of the study	Description of the product and/or process involved in the study	Code of the application industry according to ISIC (2009) Section C	Data source as one of the following: Experimental, Observational, Simulated. If experimental, type of the experiment is also given. Number of records (train + test + verification) is also written.	Types of input variables are given as continuous, nominal, binary, ordinal as well as number of them for each type	Types of output variables are given as continuous, nominal, binary, ordinal as well as number of them for each type	Types of the quality tasks involved in sequence	Types of the data handling and DM functions involved in sequence corresponding to each quality task	DM methods used to perform the corresponding data handling or DM functions	Software tools used	Brief description of the results obtained, as successful, unsuccessful, and more successful than (something). Also the other comments are listed here

quality levels cannot be achieved without developing strong designs of products or processes, DFSS programs have been put in practice, which follow a similar approach such as define-measure-analyse-develop (optimise)-verify (DMADV) (Brady & Allen, 2006).

Many quality control and improvement activities take place throughout the life cycle of a product as summarized in Table 3. This summary is adapted from Phadke (1989), and it classifies the product development process into three phases: product and manufacturing process design, manufacturing of the product, and usage of the product by the customer. In this study, we focus on certain QI and control activities that occur during the first two stages as explained below.

Three critical stages of product and manufacturing process design phase are: conceptual design of the products/processes, setting values of their design parameters (i.e. parameter design, or design optimisation), and setting tolerances for the design parameters (i.e. tolerance design). During the conceptual design of a product the most appropriate architecture and technology for achieving the desired functions of the product are selected. Then, the best settings are determined for the controllable product/process design parameters. At this parameter design stage, it is typical to assume wide tolerances on uncontrollable noise factors, and search for parameter settings insensitive to adverse effects of the noise factors. If such settings cannot be found, then tolerances for variability of the product/process parameter values are selectively reduced, and use of higher grade materials and sophisticated control systems are considered for further reduction in the variability. In this study, we especially consider conceptual design and parameter design stages, where DM has found widespread use.

Major QI and control activities during the product manufacturing phase are; determining factors that significantly affect quality, modelling relationships between input and output characteristics of quality, and predicting quality levels for a given set of input parameters. Such quality analyses provide invaluable feedback to product/process design (or re-design) and other corrective actions for QI. Most common DM applications are expected to occur in such activities due to ease of collecting relevant data, performing the analysis, and interpreting the results. We also focus on such QI activities within the scope of this study.

Some quality control and improvement activities are not considered in this review. These activities as well as the reasons of excluding them are briefly explained in the following.

When it comes to controlling quality of products, the oldest approach is to inspect all units produced and discard/repair the defective ones. We have excluded the inspection activities in this study, since smart inspection systems utilizing DM techniques form a mature area that deserves a special attention on its own. For example, a branch of chemistry, called chemometrics, has grown into a well-established data collection and analysis method in areas such as multivariate calibration and pattern recognition in the chemical process industry (Yang & Trewn, 2004). In general, automated visual inspection (AVI) has become inevitable in modern manufacturing (Newman & Jain, 1995). In AVI, many conventional image processing functions have been employed, however, much recent work has also focused on incorporating techniques from the area of artificial intelligence (AI) and DM into the systems. Pham and Alcock (2003) provide a good coverage of many of these applications.

Other important QI and control activities that we have excluded are quality monitoring and process control. Process industries and discrete parts manufacturing industries have had a long history of these activities that aim to reduce variability. While quality monitoring tries to reduce variability by detection and removal of assignable causes, process control is based on the idea of process compensation and regulation to reduce variability. Although they

Table 3
Quality control and improvement activities and methods (adapted from Phadke, 1989).

Product development stage	Quality control and improvement activity	Examples of traditional methods used
Product design, manufacturing process design	Concept design	QFD, Pugh's concept selection, TRIZ, technological forecasting
	Parameter design	Design of experiments, ANOVA, response surface modelling and analysis, regression, optimisation
	Tolerance design	Statistical tolerancing, cost analysis
Manufacturing	Inspection/Screening	Pattern recognition, automated inspection
	Quality analysis	ANOVA, regression, classification, clustering
	Process control	Feedback control, feed forward control, manual adjustments
	Quality monitoring	Statistical process control (control charts), Principal component analysis
Customer usage	Warranty and repair/replacement	Replacement analysis

consist of important DM applications through on-line process monitoring and automatic process control, they make use of special methods collected under temporal data mining due to 'time' being an important variable of the models. Furthermore, the literature on these topics is too voluminous to be included in the scope of this review. For example, reviews of artificial neural networks (ANNs) and wavelet transforms for quality monitoring are provided in Zorriassatine and Tannock (1998) and Ganesan, Das, and Venkataraman (2004), respectively. Similarly, a review paper on process control using ANNs is provided in Hussain (1999). These reviews indicate the widespread use of DM in the area. Yet, there are many other DM methods in the literature, such as principal component analysis (PCA), partial least squares (PLS), support vector machines (SVM), and decision trees (DTs), that have found applications in quality monitoring and process control. A separate review of these might also be beneficial for future studies.

The final phase of product development process considers the usage of the product by the customers. Since the product is in the hands of the customer at this phase, quality perception can be improved through repair or replacement of the product under warranty, and other services beyond warranty. We have also excluded DM applications involving QI and control at after-sales-services, which typically utilise text mining approaches due to the textual nature of the associated data.

In summary, the QI and control activities we choose to consider in this study involve the following tasks, which we refer to as *quality tasks* from this point on:

- *Description (or characterisation) of product and process quality*: Quality of products or processes can be defined or characterised by performing the following tasks:
  - Identifying attributes/variables, which affect quality significantly.
  - Ranking the attributes/variables based on their significance.
  - Identifying how low, medium and high yielding products are naturally grouped in data, and finding the most probable causative factor(s) that discriminates between low and high yielding products.

  These activities are typically performed at earlier stages (define-measure-analyse) of DMAIC and DMADV for quality analysis and product/process design.
- *Predicting quality*: When quality output is a real valued variable, developing models that relate input characteristics of quality to the output, and using such models to predict what the resulting quality characteristic value will be for a given set of input parameter values. This quality task is typically performed at the Analyse stage of both DMAIC and DMADV cycles dealing with real valued quality output characteristics. The prediction models developed can be used later in Improve or Design (Optimise) – Verify stages, or directly in the Control stage.
- *Classification of quality*: Classifying a quality characteristic of interest for nominal, binary or ordinal outputs (such as defects). For a given set of input parameters, predicting the class of the quality output.

  Similar to predicting quality, this quality task is typically performed at the Analyse stage of both DMAIC and DMADV cycles.
- *Parameter optimisation*: Based on the learned characteristics of the cases yielding high quality, finding optimal levels of process/product parameters that consistently yield target quality performance. Note that this quality task is also performed at the Improve or Develop (Optimise) – Verify stages.

3. The knowledge discovery in databases process and data mining

Fayyad, Piatetsky-Shapiro, and Smyth (1996) define KDD as "the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data". It consists of the following main steps: (i) data preparation, (ii) data preprocessing, (iii) DM, (iv) evaluation and interpretation, (v) implementation. Note that DM is a step in the KDD process that consists of applying data analysis and discovery algorithms. DM tasks can be classified into two groups (Han & Kamber, 2006): descriptive and predictive. These tasks can be accomplished by using various methods based on DM functions, used to specify the types of patterns to be mined. These functions include summarisation (characterisation), clustering, association, classification, prediction and so on. Below, the steps of KDD process are described briefly.

3.1. Data preparation

DM functions mainly utilise available data sources such as data warehouses, marts, databases or files for gathering data (Pyle, 1999). In applications, the data sources are first located, accessed, and integrated. Next, selected data is put into a tabular format in which instances and variables take place in rows and columns, respectively (Giudici, 2003). If the data set built is very large, a representative reduced data set can be obtained by sampling. In certain situations all data may not be readily available for mining, where data *farming* process can help us "defining features that are the most appropriate for DM" (Kusiak, 2006). Then, one may collect necessary data recording the feature values directly from real time or experimental observations, or indirectly from simulation results.

3.2. Data preprocessing

Real-world data is generally dirty, incomplete and inconsistent. Redundancies may also occur due to integration of data from various sources. The main purpose of this step is to handle these kinds

problems to improve the data quality. In addition, transforming and reducing data can help to improve the accuracy and efficiency of DM function(s). Basic data preprocessing techniques are as follows (see Giudici, 2003; Pyle, 1999; Witten, 2005):

Data cleaning involves techniques for filling in missing values, smoothing out noise, handling outliers, detecting and removing redundant data.
Data transformation puts the data into appropriate forms for mining when necessary.
Data reduction is applied to reduce the data set to be mined. While 'dimension reduction' technique eliminates unnecessary attributes, 'data compression' and 'numerosity reduction' techniques provide other forms of reduced data representations.
Discretisation, a form of data reduction, reduces the number of levels of an attribute by collecting and replacing low-level concepts with high-level concepts.

3.3. Data mining

DM methods can be categorised based on various criteria. In this study, we classify them according to the 'types of knowledge mined' (DM functionalities) such as clustering, association, classification, and so on for achieving descriptive/predictive DM tasks (Dunham, 2003).

3.3.1. Descriptive data mining

Descriptive DM involves exploration of patterns and relationships that may exist in data. Basic descriptive functions are summarisation, clustering, association rule generation and sequence discovery (Dasu & Johnson, 2003; Giudici, 2003). These are also used for data exploration before a classification/prediction and/or optimisation DM function is implemented. In this review, sequence discovery is not observed due to the nature of problems within the scope of the study.

Summarisation is the presentation of general characteristics of a data set. Basic approaches are OLAP and attribute-oriented induction. There are numerous statistical methods available for data summarisation (Giudici, 2003). Descriptive statistics and graphical displays can effectively describe univariate data. For bivariate data, additional methods (e.g. correlation analysis (CA), scatter plots) can be used to determine the relationship that may exist between the variables. For describing multivariate data, however, dependency and association measures as well as multidimensional graphs such as scatter plot matrix and Andrews curves are needed (Martinez & Martinez, 2002).

Clustering is the process of grouping data into classes of similar objects. The similarity among objects is usually measured by distance measures. Major distance-based clustering methods can be organized in two categories (Han & Kamber, 2006). *Partitioning methods* classify the data into k parts in such a way that observations in each part are closely related to each other. *Hierarchical methods* group the data into a tree of clusters by either using bottom-up (agglomerative) or top-down (divisive) approaches. Besides, there are others classified as density-based, grid-based and model-based methods.

Association tries to identify groups of items that occur together. Assuming that a database consists of a set of records which contains a set of items, most algorithms accomplish the association task in two steps: finding frequent item sets, and then generating interesting if-then rules (Hand, Mannila, & Smyth, 2001).

3.3.2. Predictive data mining

Predictive modelling can be accomplished by performing classification or prediction functions to forecast future values of categorical or continuous type data, respectively. Major classification/prediction methods can be categorised into the following groups (Dunham, 2003): statistical based (S-based) methods use classical techniques which depend on statistical theory, and thus, provide statistical inference. DT-based algorithms construct DTs which look like a flowchart using a top-down recursive approach. DT-based algorithms automatically generate rules having 'if-then' type structures. ANN-based algorithms, on the other hand, consist of a set of connected input–output units each having a weight, which is updated by a learning algorithm such as backpropagation (BP) used with gradient descent (GD) optimisation technique, or the radial basis function (RBF). In addition, different classification methods can be combined to improve the model accuracy. Following a successful implementation of a classification/prediction function, or as a stand-alone DM function, optimisation can be performed to determine the settings of factors (or design parameters) that yield desired responses.

3.4. Evaluation and interpretation

The KDD process described above tries to uncover previously unknown structures that may reside in data. Depending on the data sets and research objectives, one can start at any step and continue with the others as long as there are research questions to be answered. Besides, one may try several methods for describing or modelling process/product quality. Therefore, assessing the utility and reliability, and then, interpreting the information discovered in the modelling should be the final stage in a DM process (Giudici, 2003). Evaluation of the DM methods to reach a final decision requires a comparison of results obtained from various DM methods using several measures including accuracy, time and resource requirements. It is obvious that to obtain reliable results, knowledge extracted should be evaluated and interpreted correctly (Dunham, 2003).

3.5. Implementation

The KDD process described above provides tools for better understanding the relations in quality data. The final KDD step involves implementation of the results obtained into the QI related decisions of the industry (Giudici, 2003). This is typically performed informally. Kusiak (2006) proposes a framework for more structured and transparent decision making based on decision making constructs called decision tables, decision maps, atlases, and library.

4. Analysis of the applications

In the following, we examine and discuss the reviewed literature from various points of view. Data handling or more specifically data preparation and data preprocessing before performing the DM functions in the quality tasks are discussed first. Next, DM applications in each of the quality tasks defined earlier are categorised with an emphasis on the manufacturing industries. Findings of these applications are presented subsequently, and then some DM patterns found in these applications are explained. Finally, software tools used in these applications are examined.

4.1. Data handling

Data preparation and data preprocessing are two essential steps that should be taken before performing any DM function. Nevertheless, these steps are either briefly explained or considerably less emphasis is given to them in presenting the applications. Below, we provide some analyses of the literature regarding data preparation and preprocessing steps.

4.1.1. Preparation of quality data

In most of the QI applications reviewed (37% of all data sets), data were collected through statistically designed experiments (DOE). Experimental design settings used in these applications are provided in Table 4. The number of records in the experimental data sets ranged from 9 to 1323. Observational data sets are also common in the reviewed applications (28% of all data sets) and these were either obtained through online sensor measurements or from *historical* logs. The number of records in the online-observational data sets ranged from 27 to 16,381. On the other hand, historical-observational data sets (20% of all data sets) were typically obtained from two different databases, one for production and the other for quality data. The number of records in such observational data sets ranged from 27 to 58,076. Alternatively, process simulations and simulated data sets (15% of all data sets) also existed and these were used to test the performance of the algorithms used. The largest number of records observed in the reviewed applications is 250,000 and this belongs to a simulated data set.

4.1.2. Preprocessing of quality data

Data cleaning techniques used in the reviewed applications are presented in Table 5. Most of the cleaning techniques were applied to the historical-observational type data sets. In the applications, data transformations were also considered in the data preprocessing step. As can be seen in Table 6, normalisation of the data was the most common transformation used. Preprocessing of data through data reduction was also common in the reviewed studies (see Table 7). Especially, dimension reduction is observed to be a widely used form of data reduction, and there is a long list of techniques applied for that purpose, analysis of variance (ANOVA), subjective evaluation (SE) and ANN being the most noticeable among them. In addition, data compression techniques such as PCA to obtain manageable sizes of data are also common for data reduction. Finally, discretisation techniques used in the preprocessing of data are provided in Table 8. These were applied extensively on continuous type of data, especially before DT-based methods were used, and mostly SE method was preferred.

Depending on the analysis of literature reviewed above, we can say that quality data can be either too small or too large in size. Besides, it may contain missing, outlying, inconsistent and incomplete observations (see Table 5). In addition, quality data may have several mixed type (discrete and continuous) of input and output variables (Köksal et al., 2008). Due to enormous number of input variables, correlations among them are frequently observed, and usually are handled by using dimension reduction techniques (see Table 7). Moreover, quality data is usually imbalanced in terms of the number of records contained for defective/nondefective items.

4.2. Analysis of DM applications according to quality tasks

The number of DM applications for the covered quality tasks has been increasing for the last 10 years (see Fig. 1). A total of 127 application studies (presented in 130 papers) within the scope of this review have been found. In this section, these applications are reviewed according to the quality tasks performed, giving special attention to the manufacturing industries involved.

An analysis of the usage of quality tasks (shown in Fig. 2) indicates that predicting quality is the most frequently performed task

Table 4
Design types used in the QI literature.

Design type	QI applications
2- and 3-level full factorials	Brinksmeier et al. (1998), Lin and Wang (2000), Chen and Ramaswamy (2002), Feng and Wang (2003), Jiao et al. (2004), Shi et al. (2004), Feng and Wang (2004), Kurtaran et al. (2005), Karnik et al. (2007) and Erzurumlu and Oktem (2007)
Fractional factorials	Feng and Wang (2002), Yang and Tsai (2002), Ip et al. (2003), Kim et al. (2003), Shi et al. (2004), Feng and Wang (2004) and Yang et al. (2005)
Taguchi orthogonal designs	Tay and Butlar (1997), Lin and Wang (2000), Shen et al. (2007), Xue et al. (2005), Krimpenis et al. (2006), Ozcelik and Erzurumlu (2006), Hung (2007), Lee and Dornfeld (2007), and Teng and Hwang (2007)
D-optimal design	Kim et al. (2001)
One factor at a time experiments	Wang et al. (2006) and Riverol and Cooney (2007)

Table 5
Data cleaning techniques used in the QI literature.

Data to be cleaned	Techniques	QI applications
Missing values	Deleted	Li et al. (2003b), Bakır et al. (2006), Huang et al. (2006), Batmaz (2007), Hsu and Chien (2007) and Chien et al. (2007)
	Replaced with a representative value (e.g. mean)	Cool et al. (1997), Batmaz (2007) and Maimon and Rokach (2001)
	Unspecified	Kang et al. (1999) and Ho et al. (2006)
Outliers	Removed	Dhond et al. (2000), Zhai et al. (2002), Bakır et al. (2006), Batmaz (2007), Zhou et al. (2006) and Karim et al. (2006)
	Unspecified	Kang et al. (1999)
Incomplete data	Filtered out using OLAP techniques	Ho et al. (2006)
	Filtered out using low filters	Rallo et al. (2002)
	Deleted	Chien et al. (2006)
Inconsistent data	Treated by filters and removed	Zhou et al. (2006)
Misspelled data	Eliminated	Chien et al. (2007)
Noise/unwanted values		Wang (2007)
Unspecified	Unspecified	Shahbaz et al. (2006)

Table 6
Transformations applied in the QI literature.

Transformation	QI applications
Normalisation	Cool et al. (1997), Lewis and Ransing (1997), Tsai et al. (1999), Cherian et al. (2000), Suneel et al. (2002), Vasudevan et al. (2002), Rallo et al. (2002), Li et al. (2003a), Li et al. (2003b), Zuperl and Cus (2003), Shi et al. (2004), Perzyk et al. (2005), Hou et al. (2003), Ho et al. (2006), Yin and Yu (2006), Karim et al. (2006), Zhou et al. (2006), Zhang et al. (2007), Karnik et al. (2007), Lee and Dornfeld (2007), Chen et al. (2007) and Hung (2007)
Smoothing	Li et al. (2003b)
Logarithm	Umbrello et al. (2007)
Transforming categorical data to numerical	Karim et al. (2006)
Centring data	Tsai et al. (1999)
Adjusting abnormal data	Dhond et al. (2000)
Unspecified	Shahbaz et al. (2006)

Table 7
Data reduction techniques used in the QI literature.

Data reduction approach	Methods	QI applications
Dimension reduction	MLR	Ali and Chen (1999), Tseng et al. (2005) and Yin and Yu (2006)
	ANOVA	Lin et al. (2000), Chiang et al. (2002), Shi et al. (2004), Ozcelik and Erzurumlu (2006), Krimpenis et al. (2006) and Lee and Dornfeld (2007)
	DT	Ali and Chen (1999), and De Abajo et al. (2004)
	CA	Lian et al. (2002), and De Abajo et al. (2004)
	ANN	Kang et al. (1999), Han et al. (1999), Dhond et al. (2000), Rallo et al. (2002) and De Abajo et al. (2004)
	RST	Kusiak (2000), Zhai et al. (2002), Huang et al. (2006) and Chang et al. (2007)
	PLS	Zhou et al. (2006), Chong et al. (2007) and Achiche et al. (2007)
	SE	Liao et al. (1999a), Liao et al. (1999b), De Abajo et al. (2004), Yin and Yu (2006) and Chong et al. (2007)
	GA	De Abajo et al. (2004) and Zhang et al. (2007)
	GP	Li et al. (2006)
	Experimentation	Braha and Shmilovici (2002)
	GSA and ANN	Yin and Yu (2006)
	Specialized techniques	Ordieres Meré et al. (2004) and Vosniakos et al. (2007)
	Unspecified	Lu (2001)
Data compression	PCA	Chang and Jiang (2002), Lian et al. (2002), Skinner et al. (2002), Li et al. (2003b) and Brudzewski et al. (2006)
	WT	Brudzewski et al. (2006)
	SE	Brudzewski et al. (2006)
	Desirability function	Li et al. (2003a) and Hung (2007)

Table 8
Discretisation methods used in QI literature.

Cases and methods		QI applications
Cases where discretisation is used before DT-based methods		Kang et al. (1999), Maimon and Rokach (2001), Braha and Shmilovici (2002), Georgilakis and Hatziargyriou (2002), De Abajo et al. (2004), Chien et al. (2006), Wang (2007), Hsu and Chien (2007), Shao et al. (2007) and Wang (2007)
Discretisation methods	SE	Kusiak and Kurasek (2001), Maimon and Rokach (2001), Braha and Shmilovici (2002), De Abajo et al. (2004), Hou and Huang (2004)
	DT	Georgilakis and Hatziargyriou (2002)
	DA	Tseng et al. (2005)
	k-means	Chang et al. (2007)
	CC	Chien et al. (2006)
	OD	Kusiak (2000)
	Unspecified	Kang et al. (1999), Last and Kandel (2001), Zhai et al. (2002), Perzyk et al. (2005), Huang et al. (2006), Hsu and Chien (2007), Shao et al. (2007) and Wang (2007)

(42%). Applications involving classification of quality (25%) and parameter optimisation (23%) are also considerable in number. However, description of quality is performed relatively less (10%).

A classification of the reviewed applications according to the manufacturing industry codes of "International Standard Industrial Classification (ISIC) of All Economic Activities, Revision 4, Section C – Manufacturing" is provided in Fig. 3. Many DM applications in the literature regarding the QI activities appear to be from the metal product manufacturing industries (with operations such as forming, welding, casting and machining) (37%), as well as computer and electronic product manufacturing industries and device manufacturers for those (30%). Plastics (7%), basic metals (7%), coke and refined petroleum (4%) and chemical products (4%) manufacturing industries observed significantly less usage of DM methods. Moreover, motor vehicles, trailers, food, beverages, textile, paper and paper products, nonmetallic mineral products, wood products except furniture, pharmaceutical products and electrical equipment manufacturing industries had rare (2–0.8%) DM applications.

For many other industries, DM applications within the scope have not been observed. These include tobacco products, wearing apparel, leather and related products, printing and reproduction of recorded media, machinery and equipment (other than motor

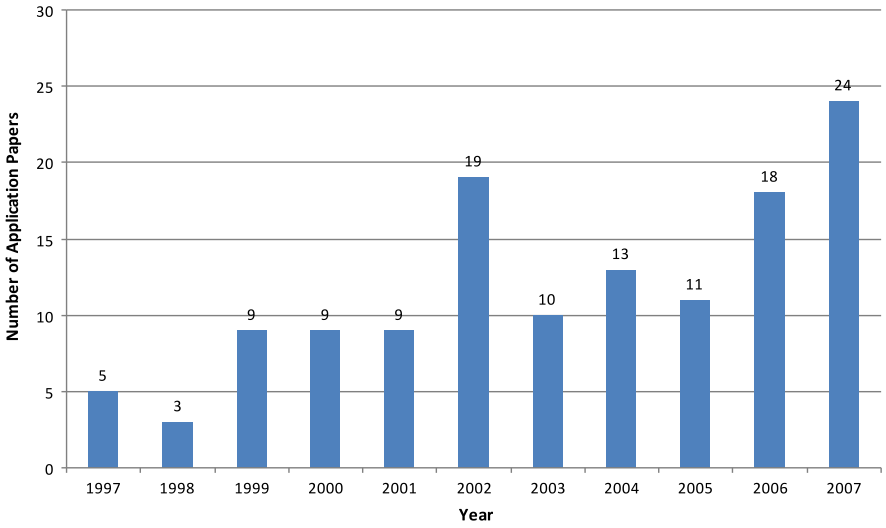

Fig. 1. Number of the application papers observed in the literature.

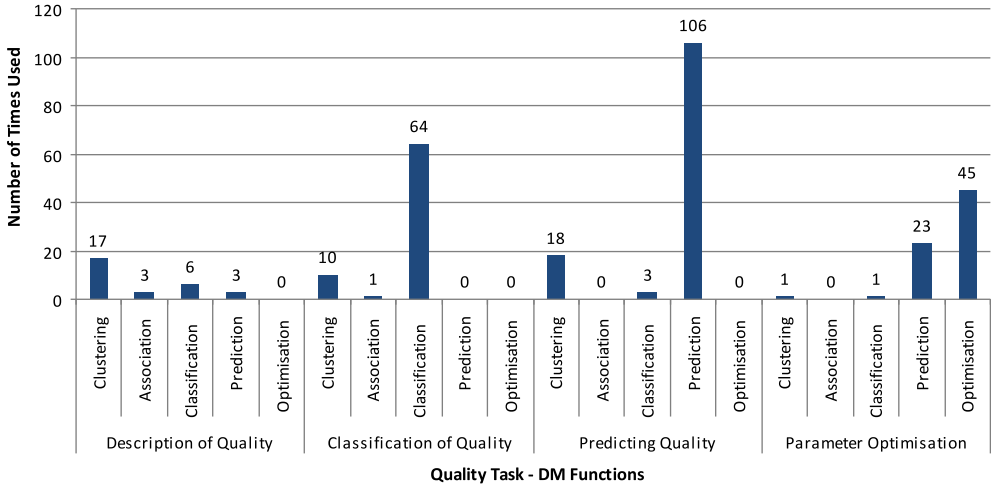

Fig. 2. Number of times DM functions are used for quality tasks in the application studies.

vehicles), and furniture manufacturing industries, which can be considered as potential application areas.

4.2.1. Description of product and process quality

Product and process quality description is generally the first stage of a QI study in an industrial application. In the computer and electronic product manufacturing industry, self organizing maps (SOMs) have been used to identify critical poor yield factors (Gardner & Bieker, 2000) and a variant of SOM has been used to improve yield (Karim, Halgamuge, Smith, & Hsu, 2006) in wafer manufacturing. Kang, Choe, and Park (1999) also considered SOM for generating better operating conditions in semiconductor manufacturing. On the other hand, Tseng, JothiShankar, and Wu (2004a), Tseng, JothiShankar, Wu, Xing, and Jiang (2004b) utilised rough set theory (RST) to identify the features that produce solder ball defects and affect the quality in printed circuit board (PCB) manufacturing. Moreover, rule induction has been utilized to select attributes that represent the quality of motherboard assemblies (Huang, Li, & Peng, 2006) and for eliminating waste by improving the quality of wafers in integrated circuit manufacturing (Kusiak, 2000). Bertino, Catania, and Caglio (1999) considered DT and association analysis for determining possible causes of faulty wafers, whereas Hsu and Chien (2007) considered DT and ANN to extract patterns from wafer bin map to improve yield. Association analysis was also used for identification of root-cause machine set (Chen, Tseng, & Wang, 2005). In order to infer possible causes of faults and manufacturing process variations in semiconductor manufacturing, Chien, Wang, and Cheng (2007) considered *k*-means clustering. Furthermore, hierarchical clustering was used for describing the relationships between machines and wafer yield rates (Hu & Su, 2004), for determining the cause of low quality wafers (Skinner et al., 2002), and for improving yield (Baek, Jeong, & Han, 2005).

Metal product manufacturing industry applications also include several product and process quality description studies, where product/process attributes or variables affecting quality were determined, ranked based on significance, or grouped based on similarities. In packaging manufacturing De Abajo, Diez, Lobato, and Cuesta (2004) developed a tinplate quality diagnostic model, where SE, genetic algorithm (GA), SOM, and DT were used. Ali and Chen (1999) used multiple linear regression (MLR) and DT in an injection moulding process for QI. For a cutting process, Chang and Jiang (2002) applied PCA for dimension reduction, and for a hot rolling process Cser et al. (2001) used SOM for clustering. ANN was used by Dhond, Gupta, and Vadhavkar (2000) in a steel mill blast furnace process and by Han, Han, and Liu (1999) in a pressurized

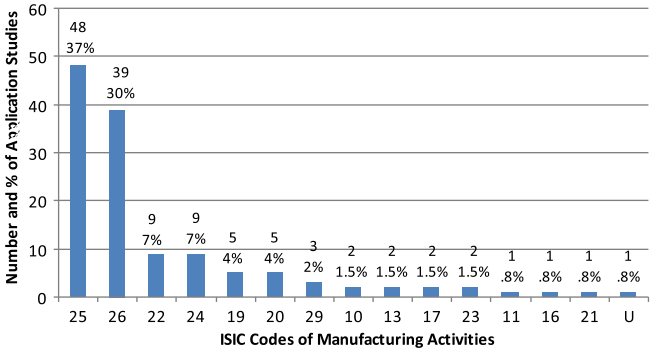

Fig. 3. Number (percentage) of the application studies with respect to ISIC manufacturing activities. (10 – Manufacture of food products, 11 – Manufacture of beverages, 13 – Manufacture of textiles, 16 – Manufacture of wood and of products of wood and cork, except furniture, 17 – Manufacture of paper and paper products, 19 – Manufacture of coke and refined petroleum products, 20 – Manufacture of chemicals and chemical products, 21 – Manufacture of basic pharmaceutical products and pharmaceutical preparations, 22 – Manufacture of rubber and plastics products, 23 – Manufacture of other nonmetallic mineral products, 24 – Manufacture of basic metals, 25 – Manufacture of fabricated metal products, except machinery and equipment, 26 – Manufacture of computer, electronic and optical products, 29 – Manufacture of motor vehicles, trailers and semi-trailers, U-unspecified).

cylinder manufacturing. Lewis and Ransining (1997) employed Bayesian belief network (BBN) and proposed a network architecture as a manufacturing diagnostic method in a pressure die casting process. RST was considered in Chang, Liu, Fan, and Chang (2007) for electrical discharge machining and in Tseng, Leeper, Banda, Herren, and Ford (2004c) for surface roughness in a machining process. CA, maximal tree method (MTM), and PCA were utilized by Lian, Lai, Lin, and Yao (2002) to control dimensional deviation of body-in-white in the assembly of sheet metal products. Tseng, Kwon, and Ertekin (2005) used MLR and rule induction for deriving rules for the process variables that affect the quality characteristics in machining processes.

There are also some product and process quality description studies in the other industries. ANOVA was used to determine the attributes of quality for a polymerisation process in silicon-filler manufacturing (Chiang, Su, Li, & Huang, 2002) and for warpage of thin shell plastic parts in a plastic injection moulding process (Ozcelik & Erzurumlu, 2006). For a better glass coating quality, Li, Feng, Sethi, Luciow, and Wagner (2003b) applied PCA to define process attributes. SOM for process variables was used in a low-density polyethylene process by Rallo, Ferre-Gine, Arenas, and Giralt (2002), and a variant of SOM was used in industrial conveyor belt manufacturing by Karim et al. (2006). Yin and Yu (2006) considered SE, MLR, and grey superior analysis (GSA) in order to select important variables in worsted spinning process. In the petrochemical industries, Zhou, Xiong, Zhang, and Xu (2006) considered PLS in an ethylene pyrolisis process, and Liao, Zhan, and Mount (1999a,1999b) considered an algorithm for inducing modular rules (PRISM) for identifying failure mechanisms of mechanical components. In order to minimize loss of product during system changeover in a refinery fluid catalytic cracking process, Sebzalli and Wang (2001) considered PCA and fuzzy *c*-means (FCM) clustering. Neuro fuzzy logic (FL) and DT were considered in Shao, Rowe, and York (2007) for pharmaceutical product formation.

4.2.2. Predicting quality

Several computer and electronic product manufacturing industry applications serving the goal of predicting quality have been published. Chen, Lee, Deng, and Liu (2007) employed ANN in a chemical vapour deposition process to establish a predictor of product quality, and Li, Su, and Chiang (2003a) used ANN to determine parameter settings in a silicon compound manufacturing process. For a robust design of very-large scale integration (VLSI) circuits, Ilumoka (1998) also considered ANN. Furthermore, Yang, Tsai, and Yeh (2005) used ANN to propose a quality control system for the solder stencil printing process, and Shi, Schillings, and Boyd (2004) used ANN for modelling nonlinear cause-effect relationships in PCB manufacturing, whereas general regression neural networks (GRNNs) were used in Tseng (2006) for predictions in a spot welding process. Prediction applications involve other methods, too. Ip, Kwong, and Wong (2003) used fuzzy regression (FR) to model transfer moulding for microchip encapsulation of electronic packages. For a silicon wafer manufacturing process and in order to model the relationships between surface roughness and cutting parameters in turning operations, Jiao, Lei, Pei, and Lee (2004) considered MLR and fuzzy adaptive network (FAN). MLR, parametric and nonparametric time series analysis (TSA), ANN, and case-based reasoning (CBR) were used to predict the manufacturing process in plastic optic fibres manufacturing (Kim & Lee, 1997). To specify the failure cause and to improve the yield, Mieno et al. (1999) employed DT. Skinner et al. (2002) considered MLR, generalized linear model (GLZ), and DT to determine the quality or yield of wafers. DT was also used to clarify the correlation between yield and various wafer parametrical data (Tsuda, Shirai, Takagi, & Take, 2000) and for prediction in integrated circuit manufacturing (Maimon & Rokach, 2001). Furthermore, Last and Kandel (2001) considered info-fuzzy network (IFN) for predicting line yield and flow times, and Li, Huang, and Wu (2006) considered genetic programming (GP) for improving the product yield in semiconductor manufacturing.

Metal product manufacturing industries observed many DM applications for prediction. Feng and Wang (2002, 2003) considered nonlinear regression (NLR) and computational ANN (Comput-NN) together with NLR to develop an empirical model for surface roughness. Sung, Kim, Xue, Kim, and Cha (2007) and Xue et al. (2005) used FR to predict the bead geometry in robotic welding process. On the other hand, MLR, NLR, and ANN were considered in Kim, Son, and Yarlagadda (2003) to present models for predicting top-bead width in a welding process. Ali and Chen (1999) also considered MLR, NLR, and ANN, but to present models for quality characteristics in an injection moulding process. Moreover, MLR and ANN were used to estimate mechanical properties of galvanized steel (Ordieres Meré, González Marcos, González, & Lobato Rubio, 2004), to model a cutting process (Chang & Jiang, 2002), to adjust influential factors of quality in steel plate production (Deng & Liu, 2002), to construct a prediction model for surface roughness in turning operation (Lin & Wang, 2000), to construct a prediction model for surface roughness of end milling (Tsai, Chen, & Lou, 1999), and to forecast hot metal temperature in a steel mill blast furnace. ANN together with NLR was considered for modelling a grinding process in Brinksmeier, Tönshoff, Czenkusch, and Heinzel (1998). Erzurumlu and Oktem (2007) considered ANN together with response surface methodology (RSM) to develop a model to predict surface roughness value error on mould surfaces, whereas Ho, Lee, Chen, and Ho (2002) applied ANN, fuzzy neural network (FNN), and neural fuzzy inference systems (ANFIS) to predict surface roughness in a turning process. ANN also had many other prediction applications; in a metal inert gas welding process (Meng & Butler, 1997; Tay & Butler, 1997), in steel welding processes (Cool, Bhadeshia, & Mackay, 1997; Vasudevan, Murugananth, & Bhaduri, 2002; Vasudevan, Rao, Venkatraman, Jayakumar, & Raj, 2005), in a laser welding process (Olabi, Casalino, Benyounis, & Hashmi, 2006), in a shot welding process (Hamedi, Shariatpanahi, & Mansourzadeh, 2007), in powder metal parts manufacturing (Cherian, Midha, & Pipe, 2000), in a knurling process (Feng & Wang, 2004), in a steel casting process (Perzyk, Biernacki, & Kochanski, 2005), in a pressure die casting process (Krimpenis, Benardos, Vosniakos, & Koukouvitaki, 2006), in a

computer numeric control (CNC) turning process (Suneel, Pandle, & Date, 2002), in oil methods manufacturing (Aloudat, 2006), in stainless still drilling process (Karnik, Gaitonde, & Davim, 2007), and in a hard machining process (Umbrello, Ambogio, Filice, & Shivpuri, 2007).

To develop prediction models for quality, other industries also made use of DM methods. ANNs were used in all of these applications. Plastics manufacturing industries utilised ANN, especially in plastic injection moulding processes (Kurtaran, Ozcelik, & Erzurumlu, 2005; Ozcelik & Erzurumlu, 2006; Sadeghi, 2000; Shen, Wang, & Li, 2007). Chemical process industries have a variety of ANN applications; in a thermal spray process (Guessasma, Salhi, Montavon, Gougeon, & Coddet, 2004), in a plasma spraying process (Wang, Fang, Zhao, & Zeng, 2007b), in a low density polyethylene process (Rallo et al., 2002), in an oxidative dehydrogenation of propane process (Holena & Baerns, 2003), in an ethylene pyrolysis process (Zhou et al., 2006), in a diamond coating like carbon process (Ho, Lau, Lee, Ip, & Pun, 2006), and in a silicon compound manufacturing process (Li et al., 2003a). Other industry applications include a reaming process (Mathews & Shunmugam, 1999), a laser sintering process (Wang, Wang, Zhao, & Liu, 2006), a canned foods process (Chen & Ramaswamy, 2002), a beer fermentation process (Riverol & Cooney, 2007), a worsted spinning process (Yin & Yu, 2006), and a particle board manufacturing process (Cook, Ragsdale, & Major, 2000). Also for a glass coating process, Li et al. (2003b) used both ANN and DT.

4.2.3. Classification of quality

Classification of a quality characteristic or predicting the class of a quality output is also a common task in DM applications. Computer and electronic product manufacturing industries exploited several DM methods with the purpose of classification. RST rule induction was used in integrated circuit manufacturing (Kusiak, 2000), in PCB printed circuit board manufacturing (Kusiak & Kurasek, 2001), and in motherboard assembly (Huang et al., 2006). Furthermore, Yang and Tsai (2002) employed neuro-fuzzy systems (NFSs) in a surface mount technology assembly process and Georgilakis and Hatziargyriou (2002) utilized DT, ANN, and entropy network (EN) for increasing the classification success rate of transformer iron losses. In an effort to improve wafer cleaning process Braha and Shmilovici (2002) used SOM, DT, ANN, and combination of multiple classifiers (CMC). To analyse final product quality ANN was used in Chen et al. (2007). DT was used in Kang et al. (1999) to generate better operating conditions and to improve the yield (Baek et al., 2005; Chien, Li, & Jeang, 2006; Chien et al., 2007; Li et al., 2006). Moreover, DT was also used for drop test analysis of portable electronic products (Zhou, Nelson, Xiao, Tripak, & Lane, 2001) and for classification in integrated circuit manufacturing (Maimon & Rokach, 2001; Rokach & Maimon, 2006). Lu (2001) proposed ways of first reducing the massive data sets into smaller size data, then using traditional methods such as ANN for identifying and classifying semiconductor and electronics process quality problems.

Classification applications in the metal product manufacturing industries were not as many as those of prediction. DTs were used in an aluminium coating process (Baek, Kim, & Kim, 2002), in assembly of sheet metal products (Lian et al., 2002), in casting processes (Bakır et al., 2006), and in assembly of automobiles (Wang, 2007). Furthermore, Shahbaz, Srinivas, Harding, and Turner (2006) considered DT and association rules in a fan blade manufacturing process. ANN had also some applications; in hot rolling (Cser et al., 2001), in tinplate manufacturing (De Abajo et al., 2004), in pressure die casting (Krimpenis et al., 2006), and in a cutting process (Lee & Dornfeld, 2007). On the other hand, Perzyk et al. (2005) employed naïve Bayesian classifier (NBC) in a steel casting process and Tseng et al. (2005) employed RST rule induction in a machining process. Kusiak (2002) used RST to derive associations among control parameters and the product quality in metal forming process. RST was also employed together with fuzzy set theory (FST) (Hou & Huang, 2004) and together with ANN (Hou, Liu, & Lin, 2003) in conveyor belts manufacturing for rule induction.

Classification applications also have been observed in other industries to improve quality. DT was used in ultra precision manufacturing (Huang & Wu, 2005). Jemwa and Aldrich (2005) considered an integration of SVM and DT for a continuous stirred tank reactor. ANN were considered in diagnosis of the causes of prestressed concrete damages (Tam, Tong, Lau, & Chan, 2004), and in predicting the quality of injected plastic parts (Sadeghi, 2000). Sarimveis, Doganis, and Alexandridis (2006) considered ANN and FST for classifying the product quality in paper manufacturing. Achiche, Baron, Balazinski, and Benaoudia (2007) characterized wood chip quality online to optimise a thermo mechanical pulp process through use of fuzzy decision support system (FDSS) based on GA learning.

4.2.4. Parameter optimisation

Computer and electronic product manufacturing industries made use of ANN for simultaneously optimising multiple responses in an ion implantation process (Hsieh & Tong, 2001). GA was used for parameter optimisation of multiple quantum well avalanche photodiodes (Kim, Oh, Lee, Lee, & Yun, 2001). Taguchi method (TM) was considered in Teng and Hwang (2007) to develop more accurate predictions of the package warpage in electronic packages manufacturing. To optimise multiple quality responses in a surface mount electronic assembly operation, Lu and Antony (2002) utilised fuzzy rule based TM. Hung (2007) employed TM in combination with ANN and GA to optimise wire bond design parameters. Moreover, RSM was used for robust design of VLSI circuits (Ilumoka, 1998) and for optimising the manufacturing parameters in light-emitted diode (LED) packaging processes, where association rules were also used, for dynamic parameter adjustments (He, Li, & Qi, 2007).

Metal product manufacturing industries employed GA for optimisation of a grinding process (Brinksmeier et al., 1998), optimisation of spot welding parameters (Hamedi et al., 2007; Tseng, 2006), for determination of cutting parameters in machining operations (Cus & Balic, 2003), for obtaining optimal process conditions in pressure die casting (Krimpenis et al., 2006), and for optimising burr size in drilling stainless steel (Karnik et al., 2007). Chong, Albin, and Jun (2007) used patient rule induction method (PRIM) based rule induction for determining optimal settings of process variables in a steel making process. ANN was used to control a hot rolling mill (Cser et al., 2001), to optimise a metal inert gas welding process (Meng & Butler, 1997; Tay & Butler, 1997), to optimise cutting parameters in a turning process (Zuperl & Cus, 2003), to optimise a steel manufacturing process (Liu, Tang, Fan, & Deng, 2004), and to optimise a sintering process (Zhang, Xie, & Shen, 2007). On the other hand, Lin and Wang (2000) used simulated annealing (SA) for a surface roughness and error of roundness study in turning operations, Olabi et al. (2006) used TM for optimising a laser welding process, Lee and Dornfeld (2007) used TM in a cutting process. In optimising electrical discharge machining processes, Lin, Wang, Yan, and Tarng (2000) and Lin, Lin, and Ko (2002) used fuzzy based TM. Furthermore, Lou and Huang (2003) employed FL in an automotive coating process.

Other studies optimising process parameters for use in QI studies are as follows: GA was used to optimise process parameters of a particle board manufacturing process (Cook et al., 2000), to determine the optimal values of process parameters in plastic injection moulding (Kurtaran et al., 2005; Shen et al., 2007), to minimize warpage of thin shell plastic parts (Ozcelik & Erzurumlu, 2006), and to determine parameter settings in a silicon compound manu-

facturing process (Li et al., 2003a). Other optimisation methods used are, ANN for optimising multiple quality characteristics of a polymerization process (Chiang et al., 2002) and for optimising process parameter values in a wax model rapid prototyping process (Vosniakos, Maroulis, & Pantelis, 2007). For a slider manufacturing process, Ho et al. (2006) also considered ANN together with fuzzy rule sets (FRS), and expert systems (ESs). Holena and Baerns (2003) considered sequential quadratic programming (SQP) for approximating the dependency of propane yield on catalyst composition. MLR and nonlinear programming (NLP) for optimising a textiles batch dyeing process was also considered in Köksal, Smith, Fathi, Lu, and McGregor (1998).

4.3. Reported performance of the DM methods

The reviewed applications can be categorised into three groups according to the performance of the DM methods reported by them:

1. In the majority of the studies, one or more of the method(s) are used to accomplish the aimed quality task(s). In such studies, usually successful results are reported. These typically involve a DM method such as ANN (e.g. Cool et al., 1997; Suneel et al., 2002; Yang et al., 2005), DT (e.g. Mieno et al., 1999; Tsuda et al., 2000), NLR (e.g. Feng & Wang, 2002), FR (e.g. Ip et al., 2003; Xue et al., 2005), Bayesian neural network (BNN) (e.g. Vasudevan et al., 2002) or IFN (e.g. Last & Kandel, 2001) for predicting quality; DT (e.g. Chien et al., 2006; Kang et al., 1999), probabilistic neural network (PNN) (e.g. Tam et al., 2004), RST (e.g. Kusiak & Kurasek, 2001) or NFS (e.g. Yang & Tsai, 2002) for classifying quality; TM (e.g. Chang et al., 2007; Teng & Hwang, 2007), GA (e.g. Cus & Balic, 2003) or ANN (e.g. Hsieh & Tong, 2001) for parameter optimisation; apriori (e.g. Da Cunha, Agard, & Kusiak, 2006) and an agglomerative clustering method (e.g. Hu & Su, 2004) for quality description.
   Alternatively, a method is used for achieving several quality tasks. For example, DT is used both for describing and classifying quality (e.g. Baek et al., 2005); ANN is used for both predicting quality and parameter optimisation (e.g. Chiang et al., 2002) and also used for both classifying and predicting quality (e.g. Sadeghi, 2000).
   In some other studies, several DM methods are used together to accomplish the quality tasks. For example, while ANN is used for predicting quality as a part of a parameter optimisation study, GA (e.g. Karnik et al., 2007; Shen et al., 2007), TM (e.g. Olabi et al., 2006), RSM (e.g. Ilumoka, 1998) or SQP (e.g. Holena & Baerns, 2003) is used for finalising the optimisation. In another application, after SOM is used for classifying quality, ANN is used for parameter optimisation (e.g. Cser et al., 2001).
2. In another group of studies, several well-known DM methods are used as alternatives for accomplishing the same function and their performances are compared. The most frequently used DM method in this group is the ANN. Its performance is mostly compared to the performance of the classical statistical modelling method MLR. These comparison studies indicate that ANNs are more successful than MLR modelling for prediction (e.g. Deng & Liu, 2002; Kim et al., 2003; Rallo et al., 2002). Better performance of ANNs can naturally be observed in multidimensional data since these are powerful tools in modelling nonlinear relationships (Fu, 1994). Yet, there is another study where both ANN and MLR perform equally well (Dhond et al., 2000). We should note here that both S-based and computational methods, such as ANN and DT, have their own advantages and disadvantages. S-based methods used in the classification/predicting quality tasks take the advantage of statistical theory which leads to 'statistical inference'. Nevertheless, the theory they are based on requires distributional assumptions, which may be hard to be validated for multidimensional data. This may necessitate the help of a human expert. In contrast to the S-based methods, computational methods provide 'computational inference' without expressing any probability. In spite of these shortcomings, it is easier with computational methods to develop automatic models without a significant human intervention.
   Other methods that are found to be more successful than MLR are FAN (Jiao et al., 2004), RST (Tseng et al., 2005), TSA, CBR (Kim & Lee, 1997), and GSA combined with ANN (Yin & Yu, 2006). Findings of the other comparative studies for prediction related to the ANN can be listed as follows:
   - ANN with multi-layer perceptron (MLP) and BP outperforms RBF (Hamedi et al., 2007), but requires long computing time.
   - ANN with multiple inputs is more successful than ANN with a single input (Mathews & Shunmugam, 1999).
   - According to Guessasma et al. (2004) for an ANN, the most adequate architecture, learning paradigm, transfer function and the error function are the multilayer normal feedforward quick propagation, sigmoid functions and the MSE, respectively.
     The most apparent finding based on a comparative study for classification related to the ANN is stated by Perzyk et al. (2005) as follows:
   - NBC, an S-based method improved through Bayesian classifier (BC) by assuming class conditional independence, performs better than ANN.
     In addition to ANN, performance of DT is also compared to those of other methods. DT is found to be more successful than MLR/GLZ for prediction/classification (Bakır et al., 2006; Skinner et al., 2002). The commonly used DT methods (such as C4.5, C5.0) can deal with missing and continuous type data. In addition, these methods generate if-then type of rules that are not provided by ANN methods. However, they also suffer from 'overfitting' that may occur due to modelling the noise in the training data. Other methods compared to DT for classification and their respective performances are stated as follows:
   - GP is more successful than DT (Li et al., 2006).
   - Attribute decomposition approach (ADA), NBC, and DT have the best, better and the worst performances, respectively (Maimon & Rokach, 2001).
   - Breadth-oblivious-wrapper (BOW) is more successful than NBC and DT (Rokach & Maimon, 2006).
     All DT algorithms mentioned above perform well for small to medium sized databases. There are also algorithms developed particularly for large databases such as SLIQ and SPRIT, which might be used in future QI studies.
     Other findings of comparative studies for prediction can be summarised as follows:
   - MLR is more successful than SVM for prediction (Brudzewski, Kesik, Kolodziejczyk, Zborowska, & Ulaczyk, 2006).
   - FNN outperforms polynomial network (PN) and ANFIS for prediction (Ho et al., 2002).
     Findings of comparative studies related to optimisation are as follows:
   - ANN with BP learning algorithm is more accurate, but more time consuming than ANN with RBF (Zuperl & Cus, 2003).
   - PRIM-based rule induction is more robust compared to the functional approach (Chong et al., 2007).

- Grey relational analysis (GRA) is more straightforward than fuzzy-based TM method (Lin, Lin, & Ko, 2002).
- ANN (Gaussian, RBF, GD) is successful, but quite time consuming (Tay & Butler, 1997).

3. In a relatively smaller group of studies, either a new method is developed to obtain a better performing one (e.g. Baek et al., 2002; Liao et al., 1999a; Liao et al., 1999b) or two or more known methods are used in combination to improve the performance. Findings related to the combined methods are given as the following:
   - FST and RST combined is more successful than RST alone for classification (Hou & Huang, 2004).
   - SE is more successful than spatial statistics integrated with adaptive resonance theory neural network (ARTNN) for clustering (Hsu & Chien, 2007).
   - Support vector classifier (SVC) and DT combined is more rapid and SVC has better generalisation properties for classification (Jemwa & Aldrich, 2005).
   - GSA in combination with ANN is more successful than MLR for prediction (Yin & Yu, 2006).
   - CMC (boosting and stacked generalisation) is more successful than the performances of individual methods DT, ANN and SOM for classification (Braha & Shmilovici, 2002).
   - Hybrid DT and ANN classifier (HDTNNC) performs better in terms of optimal time and accuracy compared to EN, DT and ANN for classification (Georgilakis & Hatziargyriou, 2002).
   - Fuzzy-rule based TM is found successful for optimisation (Lu & Antony, 2002).
   - ANN used in combination with GA (e.g. Hamedi et al., 2007; Krimpenis et al., 2006; Vosniakos et al., 2007), used in combination with TM (Olabi et al., 2006) and used in combination with RSM (Ilumoka, 1998) produce good results.

In spite of the long list of DM methods used for different DM functions, there are still some methods not found a use in QI applications. Among these, robust regression (RR) and multivariate adaptive regression splines (MARS) worth mentioning, since they are found to be successful, especially for modelling complex relationships (Kartal, 2007).

Knowledge obtained by DM applications might be hard to interpret and put in use, especially by industry people (Harding, Shahbaz, Srinivas, & Kusiak, 2006; Kusiak, 2006; Wang et al., 2007a). However, this has not been a concern raised strongly by the applications, since most of these applications were made by "experts" as can be deduced from the authors' academic identities.

4.4. Patterns of DM function and method usage in quality tasks

The ordering (or hierarchy) of DM functions as well as DM methods used to perform a quality task is referred to as 'patterns'. Different patterns followed in the applications are analysed and findings are presented in this section.

An analysis of DM function usage for each quality task (Fig. 2) indicates the following:

- Prediction is the most common (44%) DM function used in the applications.
- Classification is the second most commonly (25%) used DM function, and it is typically used (86%) for classification of quality. It is rarely considered for description of quality (8%), predicting quality (4%), and for parameter optimisation (1%).
- Clustering is used in 15% of all of the used DM functions, and it is commonly used for quality description (37%) and predicting quality (39%). It is used relatively less for classification of quality (22%). Besides, there is only one study (2%) that utilises the clustering function for parameter optimisation.
- Optimisation is only used in performing the quality task parameter optimisation.
- Association is the most rarely used DM function (1%) compared to the others, and this is used for both description and classification of quality.

Table 9 presents the number of times each DM method is used to perform a DM function for a quality task in either a stand-alone or a combined manner. If a DM method used as a part of a combined/hybrid approach to accomplish a DM function indirectly serving to the eventual goal of the study, this method is counted in parentheses in the table. With the help of this type of display, one can easily see which (how many) DM functions (methods) are involved in performing each quality task directly or indirectly.

Table 9 does not contain the summarisation DM function, since it is explicitly stated in only two of the applications (Chien et al., 2007; Ordieres Meré et al., 2004) where summarisation techniques such as descriptive statistics and scatter plot matrix are used to explore the data.

Regarding the description of quality task, it can be observed from Table 9 that clustering, especially by SOM, PCA, agglomerative clustering, and *k*-means (e.g. Batmaz, 2007; Sebzalli & Wang, 2001; Skinner et al., 2002) and classification, by DT and RST (e.g. Bertino et al., 1999; Huang et al., 2006) were relatively common.

Furthermore, this task is generally followed by a classification of quality task using, for example DT or SVM (e.g. Baek et al., 2005; Brudzewski et al., 2006), or by a predicting quality task using ANN or DT (e.g. Chiang et al., 2002; Chien et al., 2007). Predicting quality and classification of quality are the most frequent quality tasks observed in the literature. Typical classification/prediction methods are generally used correspondingly for the tasks classification/predicting quality (e.g. Baek et al., 2002; Bakır et al., 2006, etc./ Cherian et al., 2000; Deng & Liu, 2002, etc.). However, it is observed that clustering and/or association methods are also used for classifying/predicting quality (Chien et al., 2007; Shahbaz et al., 2006; Skinner et al., 2002). Additionally, some other applications used clustering methods with classification methods for classifying quality. More specifically; Brudzewski et al. (2006) used a clustering method to determine the number of classes in data and then used this number in the SVM classification method. Hsu and Chien (2007) and Lian et al. (2002) used a clustering method to determine special patterns (groups) in the data, and then used DT methods to identify the root causes that lead to the identification of these groups. Tsai, Chiu, and Chen (2005) used a clustering algorithm in relation with CBR classification method to decrease processing time of the classification, and Sarimveis et al. (2006) used FCM to allocate the input data into different clusters and then an ANN structure is automatically formulated by assigning each cluster to a hidden layer node. Clustering methods had also some use with prediction methods for the predicting quality task. For instance, Ordieres Meré et al. (2004) determined nonhomogenous groups by clustering methods and then developed different models such as MLR and ANN, for each group separately, whereas Rallo et al. (2002) buried a clustering method into the prediction methods (i.e. fuzzy predictive adaptive resonance theory neural network (Fuzzy ARTMAPNN), dynamic radial basis function neural network (Dynamic RBFNN)). On the other hand, some DM methods were used in combination for classification of quality. For example, ANN and DT (e.g. Georgilakis & Hatziargyriou, 2002), DT and SVM (Jemwa & Aldrich, 2005), RST and linear programming (LP) (Kusiak, 2000), and RST and FST (Hou & Huang, 2004) were used in combination.

Some patterns have also been observed in the parameter optimisation studies. For instance, several studies directly used ANN

Table 9
Distribution of reviewed application studies with respect to DM functions and quality tasks involved.

			Quality task			
DM functions	Approach	Method	Description of quality	Classification of quality	Predicting quality	Parameter optimisation
Clustering	S-Based	PCA	(1) 1	(1)	(2)	
	NN-Based	SOM	(1) 1			(1)
	Others	Agglomerative	2	1	(4)	
		BIRCH		(1)		
		FCM	(1)	(2)	(2)	
		GSOM	1			
		k-means	2	(1)	(3)	
		Modified *k*-means	1		(2)	
		MTM	1	(1)		
		PAM	1		(2)	
		Rule Induction	(1)			
		SE	(1)	(1)		
		Single linkage	1			
		Spatial Stats integrated with ART ANN	(1)	(1)		
		Unspecified		1	(3)	
Association	Others	Apriori	1			
		Unspecified	2	1		
Classification	S-Based	GLZ		2		
		NBC		3		
	DT-Based	C4.5	1	9		
		C5.0		1		
		CART	1	1		
		ID5R		1		
		Statistical Batch Based		1		
		Unspecified	1	7	(1)	
	NN-Based	BNN		2		
		LVQ		1		(1)
		MLP with BP (GD)		5		
		MLP with BP(LM)		1		
		PNN		1		
		RBF		2		
		SOM		1		
	Others	ADA		1		
		Boosting		1		
		BOW		1		
		CBR			1	
		DNN-Based		1		
		DT and ANN		2		
		EN		1		
		FDSS based on GA		1		
		FST and RST		1		
		FRS		1		
		GA and RST		1		
		GP			(1)	
		MIG		1		
		NFS		1		
		PRISM	1	2		
		RST	2	6		
		RST and LP		1		
		SVM		2		
		SVM and DT		1		
		TM and ANN		1		
Prediction	S-Based	ANOVA			1	
		BBN	1		1	
		MLR			16	(1)
		NLR			6	(1)
		RSM			2	
		TSA			1	
	DT-Based	CART			4	
		ID3	1			
		Unspecified			2	
	NN-Based	ADNN			1	
		BNN			4	
		ComputNN			1	
		Fuzzy ARTMAPNN			1	
		GRNN			3	(2)
		MANN			1	(1)
		MLP with BP(GD)			29	(2)
		MLP with BP (SA)			1	
		MLP with BP(CG)			1	
		MLP with BP(LM)			3	(3)

Table 9 (*continued*)

DM functions	Approach	Method	Quality task			
			Description of quality	Classification of quality	Predicting quality	Parameter optimisation
		MLP with BP (unspecified)				(6)
		PN			1	
		RBF			5	(2)
		Recurrent NN Feedforward BP			1	
		Unspecified			6	(3)
	Others	AN			1	(1)
		ANFIS			1	
		BANN			1	
		CBR			1	
		FAN			1	
		FNN			1	
		FR			3	
		FST				(1)
		FST and ANN			1	
		FRS			1	
		GP			1	
		IFN			1	
		NFS	1			
		SO and MLP			1	
		TM and ANN			1	
Optimisation	S-Based	RSM				1
		TM				3
	NN-Based	ADNN				1
		MLP with BP				4
		RBF				2
		Unspecified				3
	Others	ANN, Fuzzy Rule Sets and ES				1
		ANN and GA				13
		ANN and TM				1
		TM, ANN and GA				1
		TM and FL				3
		RSM and MANN				1
		Fuzzy Rule Sets				1
		GA				1
		GA and NLR				1
		GA and FST				1
		FMM				1
		GRA				1
		NLP and MLR				1
		NLP and ANN				1
		PRIM-Based rule induction				1
		SA and ANN				1
		SQP				1

for parameter optimisation (e.g. Hsieh & Tong, 2001; Meng & Butler, 1997; Zhang et al., 2007), while some others considered ANN for predicting the fitness function values in optimisation by GA (e.g. Brinksmeier et al., 1998; Hamedi et al., 2007; Kim et al., 2001; Ozcelik & Erzurumlu, 2006). In addition, ANNs were also used for predicting the response variable in optimisation by RSM (Ilumoka, 1998), SA (Lin & Wang, 2000), NLP (Lou & Huang, 2003), and TM (Hung, 2007; Olabi et al., 2006). Similarly, it was observed that TM for parameter optimisation was either used alone (e.g. Chang et al., 2007; Teng & Hwang, 2007) or combined with FL (e.g. Lin, Wang, Yan, & Tarng, 2000; Lu & Antony, 2002).

The use of more efficient clustering methods such as CLARA, or CLARANS for handling large data sets has not been observed in the reviewed applications. Likewise, advanced clustering methods, such as density-based (e.g. DBSCAN, DENCLUE), grid-based (e.g. STING, WaveCluster), and combined methods (e.g. CLIQUE that adopts both grid- and density-based methods) have not been used. This might be due to the fact that these advanced methods require more elaborate approaches.

4.5. Software used for mining QI data

Different computational tools shown in Table 10 have been used with various purposes in mining quality data. These include spreadsheets, databases, statistical software, DM software, special purpose software, and high-level languages.

Well known statistical software packages; MINITAB™, SAS™, SPSS™ and STATISTICA™ were preferred for implementing S-based methods such as MLR, GLZ, ANOVA and PCA. Similarly, for data preparation and preprocessing, some studies preferred the well known spreadsheet-application Excel™, database management system MS ACCESS™ or ORACLE™. However, the other studies mostly utilised the general purpose software MATLAB™, high level languages such as C/C++, or various special purpose software packages. DM software packages such as SPSS Clementine™ and SAS/EM™ were only used in a few of the applications.

MATLAB™ was used in various applications for clustering, prediction, classification and optimisation, due to its appropriateness for developing domain specific solutions, supported by several open source toolboxes. Even though scalability might not be achieved with MATLAB™ (Kwon, Omitaomu, & Wang, 2008), the sizes of data in the QI applications were not too large to present this as an issue.

In some applications of DM methods, e.g. as ANN and DT, the programs were developed from scratch by using high level languages such as C/C++ for flexibility in handling and analysing the data according to the particular methodologies followed by the authors.

Special purpose software, on the other hand, may be beneficial and more affordable, if the quality problem is clearly defined and

Table 10
Software used for accomplishing DM tasks in the QI literature.

Type of software	Name of software	Used by	Used for
Spreadsheet-application	Excel™	Shahbaz et al. (2006) and Chien et al. (2007)	Data preparation, data preprocessing or EDM
Database management system	MS ACCESS™	Liao et al. (1999a) and Liao et al. (1999b)	
	ORACLE™	Shahbaz et al. (2006)	
Statistical software package	MINITAB™	Feng and Wang (2002), Skinner et al. (2002), Feng and Wang (2003), Tseng et al. (2005), Ozcelik and Erzurumlu (2006) and He et al. (2007)	S-Based methods such as MLR, GLZ, NLR, ANOVA, PCA, Taguchi Design, etc.
	SAS™	Skinner et al. (2002), Kim et al. (2003), Shi et al. (2004) and Chien et al. (2007)	
	SPSS™	Tsai et al. (1999)	
	STATISTICA™	Tseng et al. (2005)	
DM software package	SPSS Clementine™ (currently IBM™ SPSS™ Modeler)	Wang (2007), Bakır et al. (2006) and Batmaz (2007)	Data preprocessing, classification, prediction
	TANAGRA	Da Cunha et al. (2006)	Association
	SAS/EM™	Deng and Liu (2002)	Prediction
General purpose software	MATLAB™	Brudzewski et al. (2006)	Data preprocessing
		Batmaz (2007	Clustering
		Kim et al. (2001), Holena and Baerns (2003), Kim et al. (2003), Kurtaran et al. (2005), Krimpenis et al. (2006), Ozcelik and Erzurumlu (2006), Yin and Yu (2006) and Erzurumlu and Oktem (2007)	Prediction
		Jemwa and Aldrich (2005)	Classification
		Ip et al. (2003), Xue et al. (2005), Ho et al. (2006), Sung et al. (2007), Shen et al. (2007) and Vosniakos et al. (2007)	Optimisation
Special purpose software	BrainMaker™	Feng and Wang (2003),), Feng and Wang (2004)	ANN
	NeuroShell™ Predictor	Chen and Ramaswamy (2002)	
	NeuralWorks™ Predict	Cook et al. (2000)	
	NeuralWorks™ Professional II/Plus	Hsieh and Tong (2001), Yang et al. (2005)	
	FuzzyTECH™	Yang and Tsai (2002)	Neurofuzzy applications
	FormRules™	Shao et al. (2007)	
	Fuzzy-Flou	Achiche et al. (2007)	FDSS
	Mine Set™	Bertino et al. (1999)	Association
	Q-Yield™		
	Rosetta	Huang et al. (2006)	Clustering
	c5.0™	Shao et al. (2007)	DT
	ATLaS	Kim et al. (2001)	Simulation
	ProCAST™	Krimpenis et al. (2006)	
	DataTools	De Abajo et al. (2004)	Data preprocessing/classification
High-level language	C/C++	Han et al. (1999), Tsai et al. (1999), Suneel et al. (2002), Shi et al. (2004), Tseng et al. (2005) and Shahbaz et al. (2006)	Various purposes such as ANN, DT etc.
	Visual Basic™	Liao et al. (1999a), Liao et al. (1999b), Perzyk et al. (2005) and Vasudevan et al. (2005)	Classification, prediction
	JAVA™	Chen et al. (2005)	Association

understood. For example, BrainMaker™ is a popular ANN tool which can be used with different data sources. It provides an optional package based on GA to assist users to find the best network. NeuralWorks™ and NeuroShell™ are the other special purpose software preferred for ANN modelling and analysis. For association analysis, Mine Set™ and Q-Yield™; for fuzzy applications, FuzzyTECH™, FormRules™ and Fuzyy-Flou; for clustering, Rosetta, and for DT applications C5.0 were used. Furthermore, some special purpose software such as ProCAST™ and ATLaS allowed collecting data through simulation of the experimental conditions. These data were then analysed using some other appropriate software for the chosen DM methods.

Use of commercial or open source DM software seems to be limited compared to the general purpose and special purpose ones. Among them, SPSS Clementine™, SAS/EM™ and TANAGRA were used. SPSS Clementine™ and SAS/EM™ are the most complete packages and both dominate the market (Dunham, 2003; Goebel & Gruenwald, 1999; Haughton et al., 2003; Rexer Analytics, 2008). Their limited use observed within the scope of this review can be attributed to their being less affordable compared to general purpose and special purpose software packages, as well as their limitations in handling the quality problems in an unconventional way, which might be preferred by expert users.

5. Conclusions and future research directions

This paper presents the results of a comprehensive review of DM applications in manufacturing industries to the selected QI problems. Based on these results, the following concluding remarks can be stated, each remark addressing a research question of this study listed in the first section:

1. *Data characteristics and data handling:* An interesting observation regarding the DM applications in the literature is the amount of records used in the studies. Although some studies reported applications using over 10,000 records, in many cases the number of records was less than 1000, typically collected through statistically designed experiments. Even though effectiveness of most DM methods depends on the size of the data set, this has not been reported as a concern in the applications based on small data. Furthermore, in considerable number of the cases, separately stored production and quality data had to be combined to obtain an appropriate data set. Manufacturing organisations with well established and integrated data collection systems benefit to a larger extend from DM.
In many of the studies reviewed, data preprocessing and the use of some descriptive DM methods are not reported at all, even

though the success of a DM study depends heavily on the success in preprocessing of the data. The comprehensive summary we provide on data handling practices can be used as a reference for future studies.

2. *DM applications according to quality tasks and manufacturing industries*: It has been observed that there is an increasing trend in the use of DM algorithms for QI. This can be attributed to the availability of massive data sets in some manufacturing domains and the need to improve processes with the intense competition in the industries.
   Analysis of the literature also indicates that DM applications within the context of this study were mostly encountered in the metal, computer and electronic products manufacturing industries, and relatively less observed in plastics, glass, paper, food processing and chemical manufacturing industries. DM applications can find more application domains in the other industries. However, technical training requirements to successfully implement a DM study and costs of software may be two major factors that prevent widespread use of the methods.
   In the literature regarding product and process quality description, extracted knowledge from these studies were frequently used to accomplish the eventual goals such as quality classification, prediction, and/or optimisation. Many applications that involve predicting quality task were again from the metal, computer and electronic products manufacturing. Furthermore, ANN was the most widely used DM method in those domains. Applications involving classification of quality were not as many as those in the predicting quality category. ANN and DT were common methods in those applications. Parameter optimisation tasks were generally performed after predictive models were developed for a quality output. These tasks mostly aimed for finding optimal levels of process/product parameters that consistently yield target quality. ANN, GA, and TM were commonly used in parameter optimisation.
3. *Reported performance of the DM methods*: Performance comparisons of a few DM methods with each other, as well as with traditional statistical methods, are also common in the literature. With a relatively small number of records and variables, traditional statistical methods may still provide valuable information.
   In many of the comparative studies, for prediction ANN, DT, FAN, RST, TSA, CBR and GSA used in combination with ANN were found more successful compared to MLR. On the other hand, for classification of quality, NBC, GP, BOW, SVM, ADA and CMC outperformed ANN or DT. A few studies comparing performances of ANN and DT indicated that ANN performed better than DT for both prediction and classification. Moreover, findings support the fact that individual classifiers' accuracy can be improved by combined/hybrid use of different DM methods. Furthermore, for parameter optimisation, ANN used in combination with GA were found as successful.
   In QI studies, robustness (to noise in data, mixed types of variables and complexity) and ease of modelling are important considering that the practitioners are typically not experts in data handling and DM. Even though there exist potentially useful approaches such as RR and MARS, which are strong in these respects, they have not found a use in the reviewed literature, yet. Although appeared only in a few studies, FR also seems to be a promising approach for modelling quality data based on measurements involving uncertainty or semantic scale usage due to SE involved in their collection procedure.
   Interpretation and usage of the knowledge obtained by DM applications might be hard. However, the application studies, mostly performed by academic experts have not given this problem due consideration.
4. *Patterns of DM function and method usage*: DM methods are either used as stand-alone or more interestingly as combined (hybrid). Prediction is the most frequently used DM function observed in the literature, alone or as a prior/integral part of optimisation. In most of the applications ANNs were used for prediction purposes as well as MLR in considerable number of the others. DTs or ANNs, on the other hand, were preferred for classification. For optimisation, several studies directly used ANNs while some others considered ANN for predicting the fitness function values in optimisation by GA.
5. *Software tools:* Among various computational tools used in the reviewed applications well known statistical software packages such as MINITAB™, SAS™, SPSS™ and STATISTICA™ were preferred for statistical-based methods. Similarly, well known spreadsheet-applications (Excel™) and database management systems (MS ACCESS™, ORACLE™) were utilised for data handling. The other studies mostly preferred the general purpose software MATLAB™, high level languages such as C/C++, or various special purpose software packages, for their flexibility and adoptability.
6. *Other conclusions and directions for future studies:* An important necessity for industrial applications is speed in learning and developing the DM models and their solutions, due to continuously changing customer and technical requirements in manufacturing industries. 6σ projects have been largely successful mainly because of successful training of the industry people for the use of the methods with the help of affordable and user friendly software. For more effective and wide-spread use of the DM approaches, DM software with comparable qualities needs to be developed. This, in turn, necessitates availability of more robust, easy to learn and implement data handling and DM approaches for QI problems. Such software should also have the capability to help users to select the most appropriate methods for the problem, and to interpret the results obtained from the applications. This study can guide researchers and software producers in their effort to develop/further improve their methods and tools by providing them with critical information on typical characteristics of QI data collected, necessary/most preferred DM functions and methods, and expected results.

Acknowledgements

This study is supported by Turkish Scientific and Technological Research Institute through the contract 105M138 and by Middle East Technical University through the contract BAP-2006-07-02-06. The authors thank both project teams especially to Fatma Güntürkün and Berna Bakır for their contributions to finding and classifying the relevant survey papers.

Appendix A. Abbreviations

ADA	attribute decomposition approach
ADNN	adaptive neural networks
AI	artificial intelligence
AN	abductive network
ANFIS	neural fuzzy inference systems
ANN	artificial neural network
ANN-based	artificial neural network based
ANOVA	analysis of variance

(*continued on next page*)

ARTNN	adaptive resonance theory neural network
BANN	boosting and bagging (bootstrap aggregation) neural network
BBN	Bayesian belief network
BC	Bayesian classifier
BIRCH	balanced iterative reducing and clustering using hierarchies
BNN	Bayesian neural network
BOW	breadth-oblivious-wrapper
BP	backpropagation
CA	correlation analysis
CART	classification and regression tree
CBR	case-based reasoning
CC	composite classifier
CG	conjugate-gradient
CMC	combination of multiple classifiers
ComputNN	computational artificial neural network
DA	discriminant analysis
DFSS	design for six sigma
DM	data mining
DMADV	define-measure-analyse-develop (optimise)-verify
DMAIC	define-measure-analyse-improve-control
DNN-based	decision neural network based
DOE	design of experiments
DT	decision trees
DT-based	decision tree based
Dynamic RBFNN	dynamic radial basis function neural network
EN	entropy network
ES	expert system
FAN	fuzzy adaptive network
FCM	fuzzy *c*-means
FDSS	fuzzy decision support system
FL	fuzzy logic
FMM	fuzzy min-max
FNN	fuzzy neural network
FR	fuzzy regression
FRS	fuzzy rule sets
FST	fuzzy set theory
Fuzzy ARTMAPNN	fuzzy predictive adaptive resonance theory neural network
GA	genetic algorithm
GD	gradient descent
GLZ	generalized linear model
GP	genetic programming
GRA	grey relational analysis
GRNN	general regression neural network
GSA	grey superior analysis
GSOM	growing self organizing map
HDTNNC	hybrid decision tree and artificial neural network classifier
IFN	info-fuzzy network
KDD	knowledge discovery in databases
LM	Levenberg–Marquart
LP	linear programming
LVQ	learning vector quantization
MAE	mean absolute error
MANN	modular artificial neural network
MARS	multivariate adaptive regression splines
MIG	modified breath-first search of an interest graph
MLP	multi-layer perceptron
MLR	multiple linear regression
MSE	mean squared error
MTM	maximal tree method
NBC	naïve Bayesian classifier
NFS	neuro-fuzzy system
NLP	nonlinear programming
NLR	nonlinear regression
OD	object decomposition
OLAP	on-line analytical processing
PAM	partitioning around medoids
PCA	principal components analysis
PLS	partial least squares
PN	polynomial network
PNN	probabilistic neural network
PPM	part per million
PRIM	patient rule induction method
PRISM	an algorithm for inducing modular rules
QFD	quality function deployment
QI	quality improvement
RBF	radial basis function
RR	robust regression
RSM	response surface methodology
RST	rough set theory
SA	simulated annealing
S-based	statistical based
SE	subjective evaluation
SO	self organization
SOM	self organizing map
SQC	statistical quality control
SQP	sequential quadratic programming
6σ	six sigma
SVC	support vector classifier
SVM	support vector machines
TM	Taguchi method
TQC	total quality control
TQM	total quality management
TSA	time series analysis
WT	wavelet transform

References

Achiche, S., Baron, L., Balazinski, M., & Benaoudia, M. (2007). Online prediction of pulp brightness using fuzzy logic models. *Engineering Applications of Artificial Intelligence, 20*(1), 25–36.

Ali, O. G., & Chen, Y. (1999). Design quality and robustness with neural networks. *IEEE Transactions on Neural Networks, 10*(6), 1518–1527.

Aloudat, M. (2006). *A methodology for quality conformance analysis and prediction*. Thesis (PhD). University of Houston.

Baek, D. H., Jeong, I. J., & Han, C. H. (2005). Application of data mining for improving yield in water fabrication system. In O. Gervasi et al. (Eds.). *Computational science and its applications – ICCSA 2005, 9-12 May 2005 Singapore* (Vol. 3483, pp. 222–231). Berlin: Springer-Verlag.

Baek, J. G., Kim, C. O., & Kim, S. S. (2002). Online learning of the cause-and-effect knowledge of a manufacturing process. *International Journal of Production Research, 40*(14), 3275–3291.

Bakır, B., Batmaz, I., Güntürkün, F. A., Ipekci, İ. A., Köksal, G., & Özdemirel, N. E. (2006). Defect cause modeling with decision tree and regression analysis. In *Proceedings of XVII international conference on computer and information science and engineering, 8–10 December 2006*. Cairo: World Enformatica Society (Vol. 16, pp. 266–269).

Batmaz, I. (2007). Data mining applications on manufacturing data: A casting quality improvement case. In H. O. Ayhan & I. Batmaz (Eds.), *Recent advances in statistics* (pp. 197–206). Ankara, Turkey: TUIK.

Bertino, E., Catania, B., & Caglio, E. (1999). Applying data mining techniques to wafer

manufacturing. *Principles of Data Mining and Knowledge Discovery, 1704*, 41–50.
Brady, J. E., & Allen, T. T. (2006). Six sigma literature: A review and agenda for future research. *Quality and Reliability Engineering International, 22*, 335–367.
Braha, D., & Shmilovici, A. (2002). Data mining for improving a cleaning process in the semiconductor industry. *IEEE Transactions on Semiconductor Manufacturing, 15*(1), 91–101.
Braha, D. (Ed.). (2001). *Data mining for design and manufacturing: Methods and applications*. Dordrecht: Kluwer Academic Publishers.
Brinksmeier, E., Tönshoff, H. K., Czenkusch, C., & Heinzel, C. (1998). Modeling and optimization of grinding processes. *Journal of Intelligent Manufacturing, 9*(4), 303–314.
Brudzewski, K., Kesik, A., Kolodziejczyk, K., Zborowska, U., & Ulaczyk, J. (2006). Gasoline quality prediction using gas chromatography and FTIR spectroscopy: An artificial intelligence approach. *Fuel, 85*(4), 553–558.
Chang, D. S., & Jiang, S. T. (2002). Assessing quality performance based on the on-line sensor measurements using neural networks. *Computers and Industrial Engineering, 42*(2–4), 417–424.
Chang, P. C., Liu, C. H., Fan, C. Y., & Chang, H. C. (2007). Data clustering and fuzzy neural network for sales forecasting in printed circuit board industry. In *Proceedings of the 2007 IEEE symposium on computational intelligence and data mining (CIDM 2007)* (pp. 107–113). Honolulu, HI: IEEE.
Chen, C. R., & Ramaswamy, H. S. (2002). Modeling and optimization of variable retort temperature (VRT) thermal processing using coupled neural networks and genetic algorithms. *Journal of Food Engineering, 53*(3), 209–220.
Chen, W. C., Lee, A. H. I., Deng, W. J., & Liu, K. Y. (2007). The implementation of neural network for semiconductor PECVD process. *Expert Systems with Applications, 32*(4), 1148–1153.
Chen, W. C., Tseng, S. S., & Wang, C. Y. (2005). A novel manufacturing defect detection method using association rule mining techniques. *Expert Systems with Applications, 29*(4), 807–815.
Cherian, R. P., Midha, P. S., & Pipe, A. G. (2000). Modeling the relationship between process parameters and mechanical properties using bayesian neural networks for powder metal parts. *International Journal of Production Research, 38*(10), 2201–2214.
Chiang, T. L., Su, C. T., Li, T. S., & Huang, R. C. C. (2002). Improvement of process capability through neural networks and robust design: A case study. *Quality Engineering, 14*(2), 313–318.
Chien, C., Li, H., & Jeang, A. (2006). Data mining for improving the solder bumping process in the semiconductor packaging industry. *Intelligent Systems in Accounting, Finance and Management, 14*(1–2), 43–57.
Chien, C. F., Wang, W. C., & Cheng, J. C. (2007). Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Systems with Applications, 33*(1), 192–198.
Chong, I. G., Albin, S. L., & Jun, C. H. (2007). A data mining approach to process optimization without an explicit quality function. *IIE Transactions, 39*(8), 795–804.
Choudhary, A. K., Harding, J. A., & Tiwari, M. K. (2008). Data mining in manufacturing: A review based on the kind of knowledge. *Journal of Intelligent Manufacturing*, published online, ISSN:0956-5515.
Cook, D. F., Ragsdale, C. T., & Major, R. L. (2000). Combining a neural network with a genetic algorithm for process parameter optimization. *Engineering Applications of Artificial Intelligence, 13*(4), 391–396.
Cool, T., Bhadeshia, H. K. D. H., & Mackay, D. J. C. (1997). The yield and ultimate tensile strength of steel welds. *Materials Science and Engineering A, A223*(1–2), 186–200.
Cser, L., Gulyas, J., Szucs, L., Horvath, A., Arvai, L., & Baross, B. (2001). Different kinds of neural networks in control and monitoring of hot rolling mill. In *Proceedings of the 14th international conference on industrial and engineering applications of artificial intelligence and expert systems: engineering of intelligent systems, IEA/AIE 2001, 4–7 June 2001 Budapest Hungary*. In L. Monostori, J. Vancza, & M. Ali (Eds.). *Lecture notes in computer science* (Vol. 2070, pp. 791–796). London: Springer-Verlag.
Cus, F., & Balic, J. (2003). Optimization of cutting process by GA approach. *Robotics and Computer Integrated Manufacturing, 19*(1–2), 113–121.
Da Cunha, C., Agard, B., & Kusiak, A. (2006). Data mining for improvement of production quality. *International Journal of Production Research, 44*(18/19), 4027–4041.
Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning*. New York: Wiley-Interscience.
De Abajo, N., Diez, A. B., Lobato, V., & Cuesta, S. R. (2004). ANN quality diagnostic models for packaging manufacturing: an industrial data mining case study. In R. Kohavi et al. (Eds.). *KDD-2004: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, 22–25 August 2004 Seattle Washington* (pp. 799–804). New York: ACM Press.
Deng, B., & Liu, X. (2002). Data mining in quality improvement. *SUGI27: Proceedings of the twenty-seventh annual SAS® users group international conference*, 14–17 April 2002 Orlando, Florida [online]. Available from: http://www2.sas.com/proceedings/sugi27/Proceed27.pdf Accessed 23.03.2008.
Dhond, A., Gupta, A., & Vadhavkar, S. (2000). Data mining techniques for optimizing inventories for electronic commerce. In *S.J. proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining, KDD 2000* (pp. 480–486), 20–23 August. Boston. New York: ACM.
Dunham, M. H. (2003). *Data mining introductory and advanced topics*. New Jersey: Prentice Hall/Pearson Education.
Erzurumlu, T., & Oktem, H. (2007). Comparison of response surface model with neural network in determining the surface quality of moulded parts. *Materials and Design, 28*(2), 459–465.
Fasser, Y., & Brettner, D. (2002). *Management for quality in high-technology enterprises*. New York: Wiley-Interscience.
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communication of ACM, 39*(11), 27–34.
Feng, C. X. J., & Wang, X. F. (2003). Surface roughness predictive modeling: Neural networks versus regression. *IIE Transactions on Design and Manufacturing, 35*(1), 11–27.
Feng, C. X. J., & Kusiak, A. (2006). Data mining applications in engineering design, manufacturing and logistics. *International Journal of Production Research, 44*(14), 2689–2694.
Feng, C. X. J., & Wang, X. (2002). Development of empirical models for surface roughness prediction in finish turning. *International Journal of Advanced Manufacturing Technology, 20*(5), 348–356.
Feng, C. X. J., & Wang, X. F. D. (2004). Data mining techniques applied to predictive modeling of the knurling process. *IIE Transactions, 36*(3), 253–263.
Fu, L. M. (1994). *Neural networks in computer intelligence*. New York: McGraw-Hill.
Ganesan, R., Das, T. K., & Venkataraman, V. (2004). Wavelet-based multiscale statistical process monitoring: A literature review. *IIE Transactions, 36*(9), 787–806.
Gardner, M., & Bieker, J. (2000). Data mining solves tough semiconductor manufacturing problems. In *S.J. Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining, KDD 2000* (pp. 376–383), 20–23 August, Boston. New York: ACM.
Georgilakis, P., & Hatziargyriou, N. (2002). On the application of artificial intelligence techniques to the quality improvement industrial processes. In *Proceedings of the second hellenic conference on AI: Methods and applications of artificial intelligence , 11–12 April, Thessaloniki, Greece*. In I. P. Vlahavas & C. D. Spyropoulos (Eds.). *Lecture notes in computer science* (Vol. 2308, pp. 473–484). London: Springer-Verlag.
Giudici, P. (2003). *Applied data mining: Statistical methods for business and industry*. New York: J. Wiley.
Goebel, M., & Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. *SIGKDD Explor. Newsl., 1*(1), 20–33.
Guessasma, S., Salhi, Z., Montavon, G., Gougeon, P., & Coddet, C. (2004). Artificial intelligence implementation in the APS process diagnostic. *Materials Science and Engineering B, 110*(3), 285–295.
Hamedi, M., Shariatpanahi, M., & Mansourzadeh, A. (2007). Optimizing spot welding parameters in a sheet metal assembly by neural networks and genetic algorithm. *Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture, 221*(7), 1175–1184.
Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann Publishers.
Han, L., Han, L., & Liu, C. (1999). Neural network applied to prediction of the failure stress for a pressurized cylinder containing defects. *International Journal of Pressure Vessels and Piping, 76*(4), 215–219.
Hand, H., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, Mass: MIT Press.
Harding, J. A., Shahbaz, M., Srinivas, S., & Kusiak, A. (2006). Data mining in manufacturing: A review. *Journal of Manufacturing Science and Engineering-Transactions of ASME, 128*(4), 969–976.
Haughton, D., Deichmann, J., Eshghi, A., Sayek, S., Teebagy, N., & Topi, H. (2003). A review of software packages for data mining. *The American Statistician, 57*(4), 290–309.
He, S.-G., Li, L., & Qi, E.-S. (2007). Study on the quality improvement of injection molding in LED packaging processes based on DOE and data mining. In *Proceedings of the wireless communications, networking and mobile computing WICom 2007, 21–25 September 2007 Shanghai China* (pp. 6625–6628). IEEE Con. Pub.
Ho, G. T. S., Lau, H. C. W., Lee, C. K. M., Ip, A. W. H., & Pun, K. F. (2006). An intelligent production workflow mining system for continual quality enhancement. *Intelligent Journal of Advanced Manufacturing Technology, 28*(7–8), 792–809.
Ho, S., Lee, K., Chen, S., & Ho, S. (2002). Accurate modeling and prediction of surface roughness by computer vision in turning operations using an adaptive neurofuzzy inference system. *International Journal of Machine Tools and Manufacture, 42*(13), 1441–1446.
Holena, M., & Baerns, M. (2003). Feedforward neural networks in catalysis – A tool for the approximation of the dependency of yield on catalyst composition, and for knowledge extraction. *Catalysis Today, 81*(3), 485–494.
Hou, T. H., & Huang, C. C. (2004). Application of fuzzy logic and variable precision rough set approach in a remote monitoring manufacturing process for diagnosis rule induction. *Journal of Intelligent Manufacturing, 15*(3), 395–408.
Hou, T. H., Liu, W. L., & Lin, L. (2003). Intelligent remote monitoring and diagnosis of manufacturing processes using an integrated approach of neural networks and rough sets. *Journal of Intelligent Manufacturing, 14*(2), 235–239.
Hsieh, K. L., & Tong, L. I. (2001). Optimization of multiple quality responses involving qualitative and quantitative characteristics in IC manufacturing using neural networks. *Computers in Industry, 46*(1), 1–12.
Hsu, S. C., & Chien, C. F. (2007). Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing. *International Journal of Production Economics, 107*(1), 88–103.
Hu, C., & Su, S. (2004). Hierarchical clustering methods for semiconductor manufacturing data. *Proceedings of the IEEE international conference on*

networking, sensing and control, 21–23 March 2004 Taipei, Taiwan (Vol. 2, pp. 1063–1068). IEEE.

Huang, C., Li, T., & Peng, T. (2006). Attribute selection based on rough set theory for electromagnetic interference (EMI) fault diagnosis. *Quality Engineering, 18*(2), 161–171.

Huang, H., & Wu, D. (2005). Product quality improvement analysis using data mining: A case study in ultra-precision manufacturing industry. In *Fuzzy systems and knowledge discovery, proceedings of the second international conference on fuzzy systems and knowledge discovery, FSKD 2005, 22–29 August 2005 Changsha China*. In L. Wang & Y. Jin (Eds.). *Lecture notes in computer science* (Vol. 3614, pp. 577–580). Berlin: Springer.

Hung, Y. H. (2007). Optimal process parameters design for a wire bonding of ultra-thin CSP package based on hybrid methods of artificial intelligence. *Microelectronics International, 24*(3), 3–10.

Hussain, M. A. (1999). Review of the applications of neural networks in chemical process control — Simulation and online implementation. *Artificial Intelligence in Engineering, 13*(1), 55–68.

Ilumoka, A. A. (1998). A modular neural network approach to microelectronic circuit yield optimization. *Microelectronics Reliability, 38*(4), 571–580.

International Standard Industrial Classification of All Economic Activities, Revision 4, Section C – Manufacturing (10 February (2009). http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=27, United Nations.

Ip, K. W., Kwong, C. K., & Wong, Y. W. (2003). Fuzzy regression approach to modeling transfer moulding for microchip encapsulation. *Journal of Materials Processing Technology, 140*(1–3), 147–151.

Jemwa, G. T., & Aldrich, C. (2005). Improving process operations using support vector machines and decision trees. *AICHE Journal, 51*(2), 526–543.

Jiao, Y., Lei, S., Pei, Z. J., & Lee, E. S. (2004). Fuzzy adaptive networks in machining process modeling: Surface roughness prediction for turning operations. *International Journal of Machine Tools and Manufacture, 44*(15), 1643–1651.

Kang, B. S., Choe, D. H., & Park, S. C. (1999). Intelligent process control in manufacturing industry with sequential processes. *International Journal of Production Economics*, 583–590.

Karim, M. A., Halgamuge, S., Smith, A. J. R., & Hsu, A. L. (2006). Manufacturing yield improvement by clustering. In *Neural information processing: 13th international conference proceedings Part III, 3–6 October 2006 Hong Kong China*. In I. King et al. (Eds.). *LNCS* (Vol. 4234, pp. 526–534). Berlin: Springer.

Karnik, S. R., Gaitonde, V., & Davim, J. P. (2007). Integrating Taguchi principle with genetic algorithm to minimize burr size in drilling of AISI 316L stainless steel using an artificial neural network model. *Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture, 221*(12), 1695–1704.

Kartal, E. (2007). *Metamodeling complex systems using linear and nonlinear regression methods*. Unpublished Master of Science Thesis, Middle East Technical University, Ankara, Turkey.

Kim, E. S., Oh, C., Lee, S., Lee, B., & Yun, I. (2001). Modeling and optimization of process parameters for GaAs/AlGaAs multiple quantum well avalanche photodiodes using genetic algorithms. *Microelectronics Journal, 32*(7), 563–567.

Kim, I., Son, J., & Yarlagadda, P. K. D. V. (2003). A study on the quality improvement of robotic GMA welding process. *Robotics and Computer Integrated Manufacturing, 19*(6), 567–572.

Kim, S. H., & Lee, C. M. (1997). Nonlinear prediction of manufacturing systems through explicit and implicit data mining. *Computer and Industrial Engineering, 33*(3-4), 461–464.

Kolarik, W. J. (1995). *Creating quality: Concepts, systems, strategies, and tools*. New York: McGraw-Hill.

Köksal, G., Smith, W. A., Jr., Fathi, Y., Lu, J. C., & McGregor, R. (1998). A case study in off-line quality control: Characterization and optimization of batch dyeing process design. *International Journal of Technology Management: Special Issue on Total Quality Management, 16*(4/5/6), 358–382.

Köksal, G., Batmaz, İ., & Testik, M. C. (2008). *Data mining processes and a review of their applications for product and process quality improvement in manufacturing industry*. Technical Report No: 08-03, Industrial Engineering Department, Middle East Technical University, Ankara.

Krimpenis, A., Benardos, P. G., Vosniakos, G. C., & Koukouvitaki, A. (2006). Simulation-based selection of optimum pressure die-casting process parameters using neural nets and genetic algorithms. *International Journal of Advances Manufacturing Technology, 27*(5–6), 509–517.

Kurtaran, H., Ozcelik, B., & Erzurumlu, T. (2005). Warpage optimization of a bus ceiling lamb base using neural network model and genetic algorithm. *Journal of Materials Processing Technology, 169*(2), 314–319.

Kusiak, A., & Kurasek, C. (2001). Data mining of printed-circuit board defects. *IEEE Transactions on Robotics and Automation, 17*(2), 191–196.

Kusiak, A. (2000). Decomposition in data mining: An industrial case study. *IEEE Transactions on Electronics Packaging Manufacturing, 23*(4), 345–354.

Kusiak, A. (2002). A data mining approach for generation of control signatures. *Journal of Manufacturing Science and Engineering-Transactions of the ASME, 124*(4), 923–926.

Kusiak, A. (2006). Data mining: Manufacturing and service applications. *International Journal of Production Research, 44*(18/19), 4175–4191.

Kwon, Y., Omitaomu, O. A., & Wang, G. N. (2008). Data mining approaches for modeling complex electronic circuit design activities. *Computers & Industrial Engineering, 54*, 229–241.

Last, M., & Kandel, A. (2001). Data mining for process and quality control in the semiconductor industry. In D. Braha (Ed.), *Data mining for design and manufacturing* (pp. 207–234). Dordrecht: Kluwer Academic Publishers.

Lee, S. H., & Dornfeld, D. A. (2007). Prediction of burr formation during face milling using an artificial neural network with optimized cutting conditions. *Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture, 221*(12), 1705–1714.

Lewis, R. W., & Ransing, R. S. (1997). A semantically constrained bayesian network for manufacturing diagnosis. *International Journal of Production Research, 35*(8), 2171–2187.

Li, T. S., Huang, C. L., & Wu, Z. Y. (2006). Data mining using genetic programming for construction of a semiconductor manufacturing yield rate prediction system. *Journal of Intelligent Manufacturing, 17*, 355–361.

Li, T. S., Su, C. T., & Chiang, T. L. (2003a). Applying robust multi-response quality engineering for parameter selection using a novel neural-genetic algorithm. *Computers in Industry, 50*(1), 113–122.

Li, M., Feng, S., Sethi, I. K., Luciow, J., & Wagner, K. (2003b). Mining production data with neural network and CART. In *Proceedings of the third IEEE international conference on data mining (ICDM 2003), 19–22 November 2003 Melbourne FL* (pp. 731–734). IEEE.

Lian, J., Lai, X. M., Lin, Z. Q., & Yao, F. S. (2002). Application of data mining and process knowledge discovery in sheet metal assembly dimensional variation diagnosis. *Journal of Materials Processing Technology, 129*(1–3), 315–320.

Liao, T. W., Zhan, Z. H., & Mount, C. R. (1999a). An integrated database and expert system for failure mechanism identification: Part I – Automated knowledge acquisition. *Engineering Failure Analysis, 6*(6), 387–406.

Liao, T. W., Zhan, Z. H., & Mount, C. R. (1999b). An integrated database and expert system for failure mechanism identification: Part II – The system and performance testing. *Engineering Failure Analysis, 6*(6), 407–421.

Lin, C. L., Lin, J. L., & Ko, T. C. (2002). Optimization of the EDM process based on the orthogonal array with fuzzy logic and grey relational analysis method. *International Journal of Advanced Manufacturing Technology, 19*(4), 271–277.

Lin, J. L., Wang, K. S., Yan, B. H., & Tarng, Y. S. (2000). Optimization of the electrical discharge machining process based on the Taguchi method with fuzzy logics. *Journal of Materials Processing Technology, 102*(1–3), 48–55.

Lin, W. S., & Wang, K. S. (2000). Modelling and optimization of turning processes for slender parts. *International Journal of Production Research, 38*(3), 587–606.

Liu, X., Tang, H., Fan, Z., & Deng, B. (2004). Quality improvement modeling and practice in baosteel based on KIV-KOV analysis. In *Advances in web-age information management: 5th international conference, WAIM 2004, 15–17 July 2004 Dalian, China*. In Q. Li, G. Wang, & L. Feng (Eds.). *Lecture Notes in Computer Science* (Vol. 3129, pp. 720–725). Berlin: Springer-Verlag.

Lou, H. H., & Huang, Y. L. (2003). Hierarchical decision making for proactive quality control: System development for detect reduction in automotive coating operations. *Engineering Applications of Artificial Intelligence, 16*(3), 237–250.

Lu, J. C. (2001). Methodology of mining massive data sets for improving manufacturing quality/efficiency. In D. Braha (Ed.), *Data mining for design and manufacturing: Methods and applications* (pp. 255–288). Dordrecht: Kluwer Academic Publishers.

Lu, D. W., & Antony, J. (2002). Optimization of multiple responses using a fuzzy-rule based inference system. *International Journal of Production Research, 40*(7), 1613–1625.

Maimon, O., & Rokach, L. S. (2001). Data mining by attribute decomposition with semiconductor manufacturing case study. In D. Braha (Ed.), *Data mining for design and manufacturing* (pp. 311–336). Dordrecht: Kluwer Academic Publishers.

Martinez, W. L., & Martinez, A. R. (2002). *Computational statistics handbook with MATLAB*. Boca Raton: Chapman and Hall.

Mathews, P. G., & Shunmugam, M. S. (1999). Neural-network approach for predicting hole quality in reaming. *International Journal of Machine Tools & Manufacture, 39*(5), 723–730.

Meng, T. K., & Butler, C. (1997). Solving multiple response optimisation problems using adaptive neural networks. *International Journal of Advanced Manufacturing Technology, 13*(9), 666–675.

Mieno, F., Sato, T., Slubnya, Y., Odagiri, K., Tsuda, H., & Take, K. (1999). Yield improvement using data mining system. In *Semiconductor manufacturing conference proceedings, 1999 IEEE international symposium on, 11–13 October Santa Clara CA* (pp. 391–394). IEEE.

Montgomery, D. (2005). *Introduction to statistical quality control*. Hoboken, NJ: John Wiley.

Newman, T. S., & Jain, A. K. (1995). A survey of automated visual inspection. *Computer Vision, Graphics and Image Processing, 61*(2), 231–262.

Olabi, A. G., Casalino, G., Benyounis, K. Y., & Hashmi, M. S. J. (2006). An ANN and Taguchi algorithms integrated approach to the optimization of CO_2 laser welding. *Advances in Engineering Software, 37*(10), 643–648.

Ordieres Meré, J. B., González Marcos, A., González, J. A., & Lobato Rubio, V. (2004). Estimation of mechanical properties of steel strip in hot dip galvanising lines. *Ironmaking and Steelmaking, 31*(1), 43–50. 8.

Ozcelik, B., & Erzurumlu, T. (2006). Comparison of the warpage optimization in the plastic injection molding using ANOVA, neural network model and genetic algorithm. *Journal of Materials Processing Technology, 171*(3), 437–445.

Perzyk, M., Biernacki, R., & Kochanski, A. (2005). Modeling of manufacturing processes by learning systems: The naive bayesian classifier versus artificial neural networks. *Journal of Materials Processing Technology, 164–165*, 1430–1435.

Phadke, M. S. (1989). *Quality engineering using robust design*. Englewood Cliffs, NJ: Prentice-Hall.

Pham, D. T., & Afify, A. A. (2005). Machine-learning techniques and their applications in manufacturing. *Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture, 219*(5), 395–412.

Pham, D. T., & Alcock, R. J. (2003). *Smart inspection systems: Techniques and applications of intelligent vision*. Amsterdam, Boston: Academic Press.

Pyle, D. (1999). *Data preparation for data mining*. San Francisco, CA: Morgan Kaufmann Publishers.

Rallo, R., Ferre-Gine, J., Arenas, A., & Giralt, F. (2002). Neural virtual sensor for the inferential prediction of product quality from process variables. *Computers and Chemical Engineering, 26*(12), 1735–1754.

Rexer Analytics (2008). Available from: http://www.rexeranalytics.com/Data-Miner-Survey-Results-2008.html.

Riverol, C., & Cooney, J. (2007). Estimation of the ester formation during beer fermentation using neural networks. *Journal of Food Engineering, 82*(4), 585–588.

Rokach, L., & Maimon, O. (2006). Data mining for improving the quality of manufacturing: A feature set decomposition approach. *Journal of Intelligent Manufacturing, 17*(3), 285–299.

Sadeghi, B. H. M. (2000). A BP-neural network predictor model for plastic injection molding process. *Journal of Materials Processing Technology, 103*(3), 411–416.

Sarimveis, H., Doganis, P., & Alexandridis, A. (2006). A classification technique based on radial basis function neural networks. *Advances in Engineering Software, 37*(4), 218–221.

Sebzalli, Y. M., & Wang, X. Z. (2001). Knowledge discovery from process operational data using PCA and fuzzy clustering. *Engineering Applications of Artificial Intelligence, 14*(5), 607–616.

Shahbaz, M., Srinivas, S., Harding, J. A., & Turner, M. (2006). Product design and manufacturing process improvement using association rules. *Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture, 220*(2), 243–254.

Shao, Q., Rowe, R. C., & York, P. (2007). Comparison of neurofuzzy logic and decision trees in discovering knowledge from experimental data of an immediate release tablet formulation. *European Journal of Pharmaceutical Sciences, 31*(2), 129–136.

Shen, C., Wang, L., & Li, Q. (2007). Optimization of injection molding process parameters using combination of artificial neural network and genetic algorithm method. *Journal of Materials Processing Technology, 183*(2–3), 412–418.

Shi, X., Schillings, P., & Boyd, D. (2004). Applying artificial neural networks and virtual experimental design to quality improvement of two industrial processes. *International Journal of Production Research, 42*(1), 101–118.

Skinner, K. R., Montgomery, D. C., Runger, G. C., Fowler, J. W., McCarville, D. R., Rhoads, T. R., et al. (2002). Multivariate statistical methods for modeling and analysis of wafer probe test data. *IEEE Transactions on Semiconductor Manufacturing, 15*(4), 523–530.

Suneel, T. S., Pandle, S. S., & Date, P. P. (2002). A technical note on integrated product quality model using artificial neural networks. *Journal of Materials Processing Technology, 121*(1), 77–86.

Sung, B. S., Kim, I. S., Xue, Y., Kim, H. H., & Cha, Y. H. (2007). Fuzzy regression model to predict the bead geometry in the robotic welding process. *ACTA Metallurgica Sinica, 20*(6), 391–397.

Taguchi, G., Chowdhury, S., & Taguchi, S. (2000). *Robust engineering*. New York: McGraw-Hill.

Tam, C. M., Tong, T. K. L., Lau, T. C. T., & Chan, K. K. (2004). Diagnosis of prestressed concrete pile defects using probabilistic neural networks. *Engineering Structures, 26*(8), 1155–1162.

Tay, K. M., & Butler, C. (1997). Modelling and optimizing of a MIG welding process – A case study using experimental designs and neural networks. *Quality and Reliability Engineering International, 13*(2), 61–70.

Teng, S. Y., & Hwang, S. J. (2007). Predicting the process induced warpage of electronic packages using the P–V–T–C equation and the Taguchi method. *Microelectronics Reliability, 47*(12), 2231–2241.

Tsai, Y. H., Chen, J. C., & Lou, S. (1999). An in-process surface recognition system based on neural networks in end milling cutting operations. *International Journal of Machine Tools and Manufacture, 39*(4), 583–605.

Tsai, C. Y., Chiu, C. C., & Chen, J. S. (2005). A case-based reasoning system for PCB defect prediction. *Expert Systems with Applications, 28*, 813–822.

Tseng, H. Y. (2006). Welding parameters optimization for economics design using neural approximation and genetic algorithm. *International Journal of Advanced Manufacturing Technology, 27*(9/10), 897–901.

Tseng, T. B., Kwon, Y., & Ertekin, Y. M. (2005). Feature-based rule induction in machining operation using rough set theory for quality assurance. *Robotics and Computer-integrated Manufacturing, 21*(6), 559–567.

Tseng, T. L., JothiShankar, M. C., & Wu, T. (2004a). Quality control problems in printed circuit board manufacturing – An extended rough set theory approach. *Journal of Manufacturing Systems, 23*(1), 56–72.

Tseng, T. L., JothiShankar, M. C., Wu, T., Xing, G., & Jiang, F. (2004b). *Applying data mining approaches for defect diagnosis in manufacturing industry*. IERC 2004, Houston, TX, 7.

Tseng, T. L., Leeper, T., Banda, C., Herren, S. M., & Ford, J. (2004c). *Quality assurance in machining process using data mining*. IERC 2004, Houston, TX, 6.

Tsuda, H., Shirai, H., Takagi, O., Take, R. (2000). Yield analysis and improvement by reducing manufacturing fluctuation noise. In Proceedings of International Symposium on Semiconductor Manufacturing (ISSM 2000), 26-28 September Tokyo. 249-252.

Umbrello, D., Ambogio, G., Filice, L., & Shivpuri, R. (2007). An ANN approach for predicting subsurface residual stresses and the desired cutting conditions during hard turning. *Journal of Materials Processing Technology, 189*(1–3), 143–152.

Vasudevan, M., Rao, B. P. C., Venkatraman, B., Jayakumar, T., & Raj, B. (2005). Artificial neural network modeling for evaluating austenitic stainless steel and Zircaloy-2 welds. *Journal of Materials Processing Technology, 169*(3), 396–400.

Vasudevan, M., Murugananth, M., & Bhaduri, A. K. (2002). Application of bayesian neural network for modeling and prediction of ferrite number in austenitic stainless steel welds. In H. Cerjak & H. K. D. H. Bhadeshia (Eds.), *Mathematical modeling of weld phenomena-VI* (pp. 1079–1099). London: The Institute of Materials.

Vosniakos, G. C., Maroulis, T., & Pantelis, D. (2007). A method for optimizing process parameters in layer-based rapid prototyping. *Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture, 221*(8), 1329–1340.

Wang, K. (2007). Applying data mining to manufacturing: the nature and implications. *Journal of Intelligent Manufacturing, 18*, 487–495.

Wang, K., Tong, S., Eynard, B., Roucoules, L., & Matta, N. (2007a). Review on application of data mining in product design and manufacturing. In Lei, J., Yu, J., Zhou, S. (Eds.), *Proceedings of the fourth international conference on fuzzy systems and knowledge discovery (FSKD 2007), 24–27 August 2007Haikou Hainan, China*. IEEE (pp. 613–618).

Wang, L., Fang, J. C., Zhao, Z. Y., & Zeng, H. P. (2007b). Application of backward propagation network for forecasting hardness and porosity of coatings by plasma spraying. *Surface and Coating Technology, 201*(9–11), 5085–5089.

Wang, R., Wang, L., Zhao, L., & Liu, Z. (2006). Influence of process parameters on part shrinkage in SLS. *Intelligent Journal of Advanced Manufacturing Technology, 33*(5–6), 498–504.

Wei, C. P., Piramuthu, S., & Shaw, M. J. (2003). Knowledge discovery and data mining. In C. W. Holsapple (Ed.). *Handbook on Knowledge Management* (Vol. 2, pp. 157–189). Berlin: Springer-Verlag.

Witten, I. H. (2005). *Data mining: Practical machine learning tools and techniques*. Boston, MA: Morgan Kaufman.

Xue, Y., Kim, I. S., Son, J. S., Park, C. E., Kim, H. H., Sung, B. S., et al. (2005). Fuzzy regression method for prediction and control the bead width in the robotic arc-welding process. *Journal of Materials Processing Technology, 164*, 1134–1139.

Yang, K., & Trewn, J. (2004). *Multivariate statistical methods in quality management*. New York: McGraw-Hill.

Yang, T., & Tsai, T. (2002). Modeling and implementation of a neurofuzzy system for surface mount assembly defect prediction and control. *IIE Transactions, 34*(7), 637–646.

Yang, T., Tsai, T., & Yeh, J. (2005). A neural network-based prediction model for fine pitch stencil printing quality in surface mount assembly. *Engineering Applications of Artificial Intelligence, 18*(3), 335–341.

Yin, X., & Yu, W. (2006). The virtual manufacturing model of the worsted yarn based on artificial neural networks and grey theory. *Applied Mathematics and Computation, 185*(1), 322–332.

Zhai, L., Khoo, L., & Fok, S. (2002). Feature extraction using rough set theory and genetic algorithms—An application for the simplification of product quality evaluation. *Computers and Industrial Engineering, 43*(4), 661–676.

Zhang, J. H., Xie, A. G., & Shen, F. M. (2007). Multi-objective optimization and analysis model of sintering process based on BP neural network. *Journal of Iron and Steel Research, International, 14*(2), 01–05.

Zhou, C., Nelson, P. C., Xiao, W., Tripak, T. M., & Lane, S. A. (2001). An intelligent data mining system for drop test analysis of electronic products. *IEEE Transactions on Electronics Packaging Manufacturing, 24*(3), 222–231.

Zhou, Q., Xiong, Z., Zhang, J., & Xu, Y. (2006). Hierarchical neural network based product quality prediction of industrial ethylene pyrolysis process. In *Proceedings of the third international symposium on neural networks, advances in neural networks (ISNN 2006)*. In J. Wang et al. (Eds.). *Lecture Notes in Computer Science, 28 May – 1 June, 2006 Chengdu China* (Vol. 3973, pp. 1132–1137). Berlin: Springer-Verlag.

Zorriassatine, F., & Tannock, J. D. T. (1998). A review of neural networks for statistical process control. *Journal of Intelligent Manufacturing, 9*(3), 209–224.

Zuperl, U., & Cus, F. (2003). Optimization of cutting conditions during cutting by using neural networks. *Robotics and Computer-Integrated Manufacturing, 19*(1–2), 189.