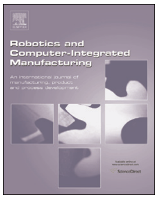
Contents lists available at ScienceDirect

Robotics and Computer-Integrated Manufacturing

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



Identifying nonconformity root causes using applied knowledge discovery

Michael Donauer [a,b], Paulo Peças [c], Americo Azevedo [a,b,*]

[a] INESC TEC (Formerly Inesc Porto), Porto, Portugal
[b] Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, Porto 4200-465, Portugal
[c] IDMEC, Instituto Superior Técnico, Technical University of Lisbon, Av. Rovisco Pais, Lisbon 1049-001, Portugal

ARTICLE INFO

Article history:
Received 4 February 2014
Received in revised form
9 October 2014
Accepted 29 December 2014
Available online 10 February 2015

Keywords:
Total Quality Management
Nonconformities
Root cause analysis
Quality tools
Knowledge Discovery in Databases
Herfindahl–Hirschman Index

ABSTRACT

Quality control, failure analysis and improvement are central elements in manufacturing. Total Quality Management (TQM) provides several quality oriented tools and techniques which, in the event of things, are not always applicable. The increased use of Information Technology (IT) in manufacturing means increased data availability and improved potential for knowledge extraction. Exploiting this knowledge requires data storage and processing facilities with demanding, time consuming sessions for interpretation. Without suitable tools and techniques, knowledge remains hidden in databases. This paper presents a method to help identify root causes of nonconformities (NCs) using a pattern identification approach. Hereby, a general framework, Knowledge Discovery in Databases (KDD), is adapted. This adaptation involves incorporating an economic concentration measure, the Herfindahl–Hirschman Index (HHI), as the data mining algorithm. After presenting the theoretical background, a new methodology is proposed. The suggested approach can be regarded as a quality tool to help make root cause identification of failures simpler and more agile. A case study from the automotive industry is examined using this tool. Results are obtained and presented in the form of matrix based patterns. They suggest that concentration indices help indicate possible root causes of NCs, warranting further investigation in this area.

1. Introduction

Controlling and maintaining quality of production processes is an important topic for manufacturing companies. The increased use of Information Technology (IT) in manufacturing means increased availability of data. Demanding activities of data processing, interpretation and visualization are necessary in order to extract knowledge from data.

Total Quality Management (TQM) can be understood as a philosophy consisting of values, tools and techniques to increase customer satisfaction and continuous improvement. Tools and techniques to control process parameters are a central element. Deviations from specified target values of the process parameters signal the need for action. Nowadays, ensuring the capacity needed for storing manufacturing data is not a problem. Likewise, central processing units are highly capable of treating the data. However, suitable tools are not always applicable and must be customized/tailored to solve specific problem statements. If suitable tools are not available, knowledge remains hidden in databases [1]. Quality tools and techniques offer a variety of methods to visualize and control process data. Applied statistics can help gather evidence to support hypotheses about cause effect relations [2]. These tools are remedies to numerous quality problems but might not always be effectual. For instance, if there are too many variables associated with a particular application, statistical analysis and visualizations of traditional quality tools are impractical evaluation instruments.

In mass production like environments, with numerous machines at several process stages, the traceability of products depends largely on the degree of IT implementation. Additional work must be done for data analysis, interpretation and visualization. Knowledge Discovery in Databases (KDD), for example, offers a general framework consisting of sub-elements to generate knowledge from a dataset [3]. Its core element is data mining (DM), a method designed to identify patterns [3]. Developers using this method have a high degree of freedom to choose what kind of element to employ in the DM step.

This paper presents a methodology to identify specific

[*] Corresponding author at: INESC TEC, Campus da FEUP, Rua Doutor Roberto Frias, Porto 4200-465, Portugal.
E-mail address: ala@fe.up.pt (A. Azevedo).

http://dx.doi.org/10.1016/j.rcim.2014.12.012
0736-5845/© 2015 Elsevier Ltd. All rights reserved.

machines within a manufacturing process as the possible origins of a given nonconformity (NC). In particular, the concentration of NCs is analyzed for single machines of a multi-stage and multi-machine manufacturing process. Once the relevant machine has been identified, it is possible to further investigate likely root causes. The study presented here relates to a real industrial problem. The methodology is validated using a case study from the automotive industry. The company operates in a mass production environment, producing similar products, which vary in size, composition and shape. Quality related data for two consecutive manufacturing process stages is evaluated. After data treatment, the result is visually depicted using colour highlighted matrices. The matrices serve to identify the source responsible for the NCs being investigated. In order to create the matrices, all NC occurrences are analyzed and their concentration is measured over the machines. The visualization takes into account production stages and production volume. Additionally, the specific types of NCs that occur at the machines within the process stages are considered.

The results obtained are of interest to both academia and industry. Practitioners, such as quality engineers, can integrate the approach into their set of quality tools in their quest for improving quality in manufacturing. Different disciplines such as IT, quality control and economics are brought together.

The paper is structured as follows. A literature review is given and the relevant background discussed. The methodology is presented and formalized for the defined set process stages to be analyzed. Defining these process stages is the first step presented in this paper, followed by the identification and collection of relevant data. In order to identify patterns, a methodology is used which is adapted from discovering knowledge in databases. After data treatment the core element of data mining is introduced and results can be obtained. The methodology's strengths and limitations are discussed and an application case is presented. The application case from the automotive industry is studied using this methodology. Results from identifying NC root causes are discussed and conclusions drawn.

2. Background and research development

This section provides the theoretical background in relevant subject areas and discusses the development of research on this topic. TQM and quality tools are reviewed and Knowledge Discovery in Databases (KDD) is presented, a specific method designed to identify patterns in datasets.

2.1. Quality tools in TQM

Total Quality Management (TQM) is a diverse and extensive subject, whose development has been influenced by a subjective body of thought. There is no global definition of TQM and its understanding and use varies across companies [4]. A generally accepted way of describing of TQM is as a philosophy that may also entail the use of certain tools and techniques. It aims to increase customer satisfaction and encourage continuous improvement.

Using tactics to change a company's culture leads to a mind-set focused on satisfying the needs of the internal customer [5,6]. This is endorsed through structured technical techniques. TQM is also understood as representing a management system consisting of values, techniques and tools as three independent components [7].

The tools of TQM described in literature have evolved over time. However, it is still generally accepted to include the seven quality control tools (Table 1) which were first selected by Ishikawa [8]. At a later date a new set of seven management tools was presented. These are more related to process mapping and problem-solving [9]. Only a few people were responsible for the development of the basic tools – Shewhart [10], Deming [11], Juran and Gyrna [12], Ishikawa [13], Ōno [14], Shingō [15] and Taguchi [16] – starting in the late 1930s. Since then, the evolution that has occurred reflects people's ability to bring the tools together programmatically in order to achieve company-wide benefits [17]. Common tools and techniques are summarized in Table 1 [18].

McQuater et al. [2] describe tools and techniques, as portrayed in Table 1, to be practical methods, skills, means or mechanisms used for a specific circumstance. Their purpose is to achieve positive change and improvement. However, many of the tools in Table 1 are not appropriate for root cause identification of nonconformities. Instead, these are more applicable for detecting change, e.g. control charts and Statistical Process Control [19,20]. Others provide a method of structured analysis for root cause identification, such as the cause and effect diagram [8] or a method designed to prevent process and product problems before they occur, such as Failure Mode and Effect Analysis (FMEA) [21]. A quality tool has also been suggested for prioritizing failure types to be selected for future quality improvement projects [22].

Nevertheless, none of the tools contains a systematic approach to identify root causes of failures from a production dataset. The absence of such a kind of a tool might be a reflection of the antiquated and formal essence of the TQM tools and methods. TQM tools are to be applied in industrial environments by a wide spectrum of stakeholders with distinct analytical and computing competences. As such, TQM tools must be easy to apply and simple to use, or at least easy to interpret. These essential characteristics of TQM tools tend to inhibit the transfer of scientific knowledge and of complex, developed methodologies to the industrial environment. In fact, the field of KDD has suffered a significant development in the last decade through the use of advanced computing algorithms, data mining strategies and the use of artificial intelligence related methods.

The contribution of the proposed research in this paper is to

Table 1
Quality tools and techniques. Adapted from [18].

The seven basic quality control tools	The seven management tools	Other tools	Techniques
Cause and effect diagram	Affinity diagram	Brainstorming	Benchmarking
Check sheet	Arrow diagram	Control plan	Department purpose analysis
Control chart	Matrix diagram	Flow chart	Design of experiments
Graphs	Matrix data analysis method	Force field analysis	Fault tree analysis
Histogram	Process decision programme chart	Questionnaire	FMEA
Pareto diagram	Relations diagram	Sampling	Poka yoke
Scatter diagram	Systematic diagram		Problem solving methodology
			Quality costing
			Quality function deployment
			Quality improvement teams
			Statistical process control

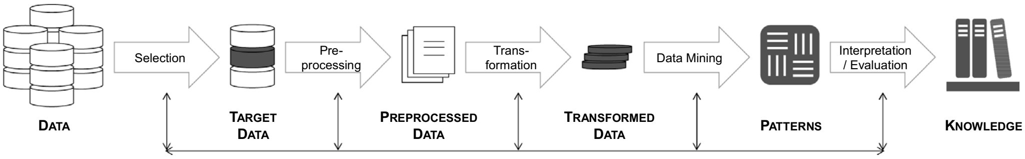


Fig. 1. The KDD process c.f. [3].

cover this gap. This is done by presenting a methodology (or tool) that is simpler, more agile and easier to interpret to support the identification of root causes of failures from a production dataset in a systemic and efficient way. The proposed methodology was developed based on the KDD field of knowledge, which the next section now discusses.

2.2. Pattern identification in KDD

KDD can be described as the complete process of discovering useful knowledge from data [3]. Knowledge is usually exposed by identifying patterns which can then be used for further analysis. Pattern identification is a core element of the methodology and referred to as data mining; it is one specific procedure of KDD [1]. Harding et al. [23] define data mining as a concept and algorithm mix consisting of machine learning, statistics, artificial intelligence and data management.

But terminology is ambiguous and one must be aware that different communities use different terms to mean the same thing. Fayyal et al. [3] compiled the following names for data mining across different communities: knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Data mining (DM) is the term used in this paper.

Fig. 1 portrays the process of KDD that is described by Fayyal et al. [3]. It starts by selecting a relevant target dataset from a database. The target data must be pre-processed to remove noise and outliers before the data can be processed further. The cleaned data can then be fed into the specially tailored DM algorithm to generate patterns. The produced patterns must then be interpreted and evaluated for the knowledge to be uncovered.

Fayyal et al. [3] mention that the DM algorithm can be composed of a specific mix of the model (the function of the model and the form in which it is presented), the preference criterion (some form of goodness-of-fit function of the model to the data) and the search algorithm (the specification of an algorithm for finding the path).

Recent reviews of the literature on the topic of KDD and DM report its popularity for use in manufacturing [1,23,24]. These reviews deal with KDD and DM surrounding the topic of quality improvement. One can say that DM functions can be categorized into clustering, association, classification and prediction [1,24]. Köksal et al. [24] refer to an increasing use of DM applications for quality related tasks, such as quality control, fault diagnosis, defect analysis, predicting quality and parameter optimization [1,24].

Of these tasks, the most widely used are applications for quality prediction followed by those used for classifying quality and parameter optimization [24]. There are plenty of applications or algorithms proposed for deploying DM. Some of these are maps for classification, regression or clustering of data, while others include summaries, dependency modelling of variables and sequence analysis [3]. Models may be represented in different ways, ranging from decision trees to linear and non-linear models, to case-based reasoning and beyond to include probabilistic graphical models of dependencies.

In addition, Du and Xi [30] state that three classes of methods are available for identifying root causes. The engineering-model-based methods draw on mathematical models from engineering that usually integrate product quality information and root cause information. The success of this class of methods depends on the availability of know-how related to the possible root causes and the accuracy of the model (knowledge that is in general very difficult to attain for a complex system) [30,31]. Another class of methods, known as knowledge-based methods [30], requires the understanding and establishment of root cause logics. It also requires the combination of data mining and knowledge discovery techniques to define the decision approach [32]. Besides the high level of effort necessary to build-up this class of method, its application field is usually restricted to the manufacturing systems under analysis. The third class is referred to as intelligent-learning-based methods [30]. These methods are usually based on neural networks, taking advantage of their self-learning potential and the fact that prior modelling or reasoning is not mandatory. The methods based on neural networks proceed to identify the root causes by pattern recognition in quality control charts and/or changes in the process performance [33,34]. Others have proposed the integration of neural networks with engineering rules and design characteristics [30]. Despite the improved accuracy and effectiveness of this last class of methods, the complexity involved in building up the methods and the intricate numerical outputs mean they lose out in terms of agility and ease of interpreting the results obtained.

So, one can conclude that quality tools and techniques provide a remedy for a wide range of problems. However, they do not sufficiently remedy the problem of root cause analysis. Although, data in manufacturing is omnipresent, it requires effort for treatment and analysis. If a suitable tool is not available it must be tailored to solve a specific problem, else the knowledge remains hidden in databases. In the field of data analysis, KDD provides an efficient methodology to make sense of a dataset. This general methodology is present in engineering and well established for quality related topics.

Thus a gap has been identified in TQM tools in relation to the problem of root cause identification. This gap can be filled by adapting KDD related knowledge and techniques to fit with TQM tools, favouring elements of simplicity and agility. This is the aim of the approach presented in this paper.

2.3. Research aims and scope

The main aim of the developed approach is to increase the agility and to facilitate the application of a systematic strategy for identifying the root causes of nonconformities. The method can be regarded as a contribution to the field of TQM and is targeted at both researchers and practitioners in quality management. Filling a gap among the existing TQM tools and methods, the proposed approach allows KDD related developments to be employed by applying KDD to “quickly” identify root causes of nonconformities in an industrial real-time setting and decision making environment. This is especially important for industries with multi-stage and multi-machine processes, 100% manual inspection and demanding customers. In this context, a novel approach to identifying root causes using a visual representation is an actual need.

The concentration of nonconformities (NC) is the measure used to assess the level of importance and hence to identify the path to determine the main root cause. Respecting the spirit of the TQM tools, the interpretation of the results is based on visualizing the results for the machines and processes involved. This helps ease the task of identifying the problem points and also facilitates communication among the several people that may be involved. Finally, the identified possible root causes must be proven as correct by investigating the data and other evidence on site.

This paper makes a significant contribution to extend existing methods for identifying nonconformity root causes in multi-stage and multi-machine production processes. This is done by adapting a general framework that is composed of concentration measure formulae. This framework can be regarded as a quality tool designed to reveal the individual machines contributing to nonconformities.

3. Nonconformity root causes and the identification matrix

This section presents the proposed methodology designed to serve as a quality tool. The functionality, the inputs required and the outputs generated are all explained and discussed.

3.1. Purpose and overview

The proposed methodology is of particular interest for manufacturing companies with multi-stage production processes each composed of multiple machines. Imperfect production processes result in a variety of NCs on a given product. The methodology proposes a visualization of concentration indices according to NCs allocated to machines. The integration of an economics concept (the concentration measure) into a KDD methodology means it can be used as a quality tool for identifying possibilities to improve processes in mass production type environments with diverse NCs. This enables the user to efficiently identify machines with a high concentration of nonconformities, potentially indicating the root cause of the NCs. The methodology can be applied by practitioners such as quality engineers of industrial companies with problems where various NCs occur at multiple machines distributed across multiple process steps. Moreover, the approach can be applied to any type of concentration measurement exercise.

Following the process steps of the methodology (please refer to Fig. 2) leads to the creation of the quality tool. An adapted methodology for pattern identification is suggested allowing manufacturing processes to be improved by learning from data. The methodology is similar to the KDD methodology (Fig. 1), adding a method for measuring concentration (taken from the field of economics) as the data mining sub-step. The resulting patterns provide the basis for interpretation and knowledge creation. Firstly, one can identify the NC from among a set of specific NC types that occur concentrated at individual machines. Secondly, one can identify at which individual machines a specific NC occurs the most. This additional knowledge serves to highlight possible origins of NCs.

In order to obtain the desired results, one must first gather quality related data for the manufacturing process. This data provides a list of NCs of the relevant manufacturing processes. Ideally, for each NC it includes a record of the exact machines that the product had passed through at every relevant manufacturing step.

In order to retrieve data in a reliable manner the format of the input data file must be defined. The input file can then be preprocessed to remove outliers and noise. This includes spreadsheet calculation which must be tailored or integrated with the previously obtained input data file. The next step is to define the DM method or algorithm. In this paper, the Herfindahl–Hirschman Index is the measure used as the data mining sub-step algorithm.

The result is visualized using a coloured table with cells ranging from red to green – green indicates no concentration and red indicates a high concentration of nonconformities. Dark red shading indicates critical machines where NCs occur highly concentrated. This high concentration suggests the associated machine to be the contributor of that particular NC. Further analysis is suggested and corrective action necessary, if applicable.

3.2. The Herfindahl–Hirschman Index as a concentration measure

The Herfindahl–Hirschman Index (HHI), also referred to as the Herfindahl Index, is a method used to measure concentration [25]. Unaware of Hirschman's published work Herfindahl developed a similar method of measuring concentration at a later date [25,26]. The equations are identical apart from the square root of Hirschman's index on Herfindahl's equation [26,27]. Herfindahl's equation is depicted in (1) and Hirschman's in (3).

$$\mathrm{HHI} = \sum_{i=1}^{N} a_i^2 \tag{1}$$

with

$$a_i = \frac{x_i}{\sum_j^N x_j} \tag{2}$$

$$I = \sqrt{\sum_{i=1}^{N} b_i^2} \tag{3}$$

with

$$b_i = \frac{x_i}{\sum_j^N x_j} \tag{4}$$

I, HHI is the concentration index.

The market share of participant i is given by

a_i, b_i. This is calculated as theshare that x_i represents relative to all the N market participants

$$\sum_{j=1}^{N} x_j \, .$$

In economics the index is the sum of the squared individual market shares of the participants in a specific market. Users apply

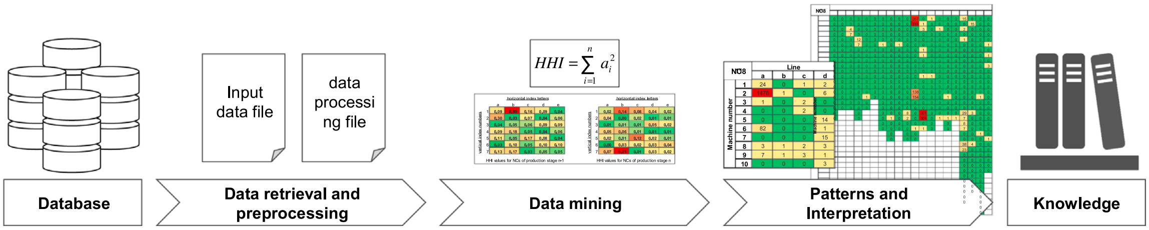

Input data file
data processing file
$HHI = \sum_{i=1}^{n} a_i^2$
Database
Data retrieval and preprocessing
Data mining
Patterns and Interpretation
Knowledge
NC8
Line
Machine number

Fig. 2. The methodology of the study. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

the equation to measure competition in the market [25]. For instance, it is used to analyze the effects on market competition after company mergers [25,28]. The US Department of Justice and Federal Reserve deploy the index in such cases. It is also used to measure income concentration in households [25].

3.3. Approach, strengths and limitations

The main strength of this approach is the ability to generate knowledge from production and quality related data. This knowledge is visually and intuitively presented, facilitating the introduction of quality improvement measurements. Moreover, the DM algorithm used is simple enough to be implemented and executed using basic spreadsheet calculation software. As such, the user does not have to invest in new software and only needs to replicate this paper's instructions to realize the potential benefits of the tool. Additionally, the results presented by the tool are highly illustrative. The tables use colour highlighting and signal high and low concentrations of NCs among machines.

A drawback of this approach is that it is essentially an offline tool. Thus it requires periodical updates and must be fed with recent input data files. If wanted, it can act in real time, requiring an upgrade with additional IT development to operate as an online tool according to the on-site IT environment. Additionally, this is a method to allocate NCs to machines. The root cause must be identified and may not originate with a machine failure. Thus, there may be cases of false positives that indicate a particular machine as responsible for a high concentration of NCs, without the machine being the root cause of the NC.

4. Application case study

The suggested methodology in Fig. 2 was applied using a mature, high technology, automotive parts producer. The company maintains a quality management system and is certified by quality standards such as DIN EN ISO 9000, DIN EN ISO 9004 and ISO/TS 16949.

4.1. Problem description

The high technology product of the company is produced in the form of a multi-stage production process operation.

The multi-stage production process of the company is a composition of different manufacturing processes. The difference stages require specialist knowledge in various fields of science. Among the processes are the mixing of raw materials, the assembly of subassemblies and a process similar to injection moulding. Each production stage consists of multiple machines. At every production stage a product passes through exactly one machine. Thus, there is no product that skips one stage and there is no product that passes two machines in one stage. Barcodes are attached to the product and every machine is equipped with a barcode scanner that saves the product-machine relationship to a database. This ensures that the information stored in the database makes the product traceable. Thus, the path that the product takes across the individual machines of the production stages can be recreated.

At the end of the manufacturing line an inspection station is located where product appraisal is performed by humans. They manually assess the product for conformance to requirements. Conforming products are accepted and forwarded to be shipped to customers. Nonconforming products are rejected and the type of NC is added to the information system. A decision follows the rejection regarding the recoverability of the product. Recoverable products are sent to be reworked and unrecoverable products are scrapped. Due to the complex nature of the product and its processes, there are a multitude of causes of nonconformities. These are attributable to process failures, machine stoppages, incorrect composition, quality of raw materials or human error. Additionally, NCs are often only detectable in the finished product and cannot be seen in the unfinished product in between stages. The NCs vary from minor cosmetic recoverable blemishes to severe imperfections that may not be recoverable. The company is focused on scrap rate reduction. Thus, they form multi-departmental teams to initiate studies to identify and eliminate root causes. This task has proven to be difficult, which is why the tool in this paper was developed.

4.2. Data retrieval and pre-processing

Fig. 3 illustrates the production stages and the input of information into the database.

Corresponding data is input into the database at the last two production stages before the inspection station. The barcode scanners attached to the machines scan the product with an

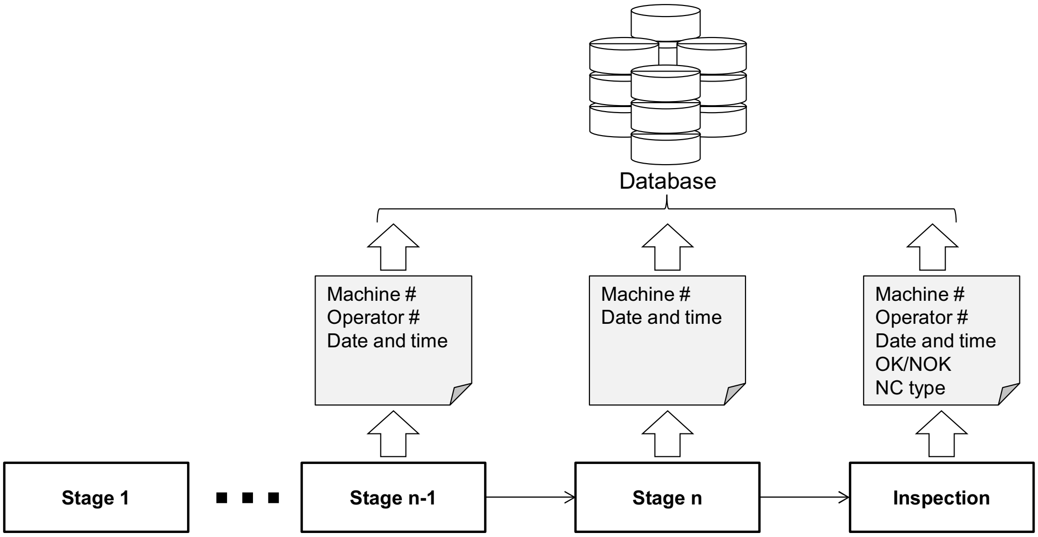


Fig. 3. The production flow and data input to the database.

Table 2
Retrieved data input file from database.

Barcode	Stage $n-1$			Stage n	Inspection			
	Machine	Date	Time	Machine	NC type	Decision	Date	Time
–	–			–	–			
1***622	A1	2011-02-11	16:32:14	B12	NC17	S	2011-02-01	06:10:55
1***799	C4	2011-02-11	18:17:54	Q34	NC24	S	2011-02-09	17:31:36
1***464	B7	2011-02-11	19:16:25	T07	NC4	R	2011-02-03	07:01:40
1***699	B9	2011-02-12	0:07:21	G06	NC1	R	2011-02-11	09:47:09
1***244	A5	2011-02-12	0:58:27	A19	NC31	R	2011-02-01	03:47:59
1***505	C2	2011-02-12	9:14:49	R23	NC12	R	2011-02-08	14:02:45
–	–			–	–			

Table 3
Pre-processing of data to identify the number of occurrences according to machines.

Stage $n-1$	NC1	NC2	–	NCn	Stage n	NC1	NC2	–	NCn
Machine 1	x	y	–	z	Machine 1	x	y	–	z
Machine 2	–	–	–	–	Machine 2	–	–	–	–
–	–	–	–	–	–	–	–	–	–
Machine n	u	v	–	w	Machine n	u	v	–	w

attached barcode. The data contains information about the specific production machine, the operator involved as well as time and date (please refer to Table 2). Lastly, the result of the appraisal decision provides a final input.

Basic calculations are required to compute the occurrences of NCs according to individual machines using the retrieved data as presented. This results in a matrix with machine number and NC type filled with the number of incidents as presented in Table 3.

Applying the statistical formula (1) and (2) to the tables, one can calculate the HHI for a specific NC for one production stage. Calculating the HHI for all NCs for each production stage provides information showing the level of concentration of NCs for individual machines. Visualizing the data in a specific way can help in identifying the NCs with the highest concentration for particular machines.

4.3. *Matrix output – results and discussion*

This section presents the matrix output. Firstly, those NCs with high HHI values are identified. Secondly, two examples are presented of cases with the highest HHI values for each production stage.

4.3.1. *Concentration indices for two production stages*

Following the methodology steps of the tool generates the data used in the result tables. Table 4a and b provide an overview of the concentration of all NCs for the two production stages. The values in each cell represent the HHI results for each NC for the two stages. Results for stage $n-1$ are illustrated in Table 4a and results for stage n are presented in Table 4b. Each cell represents the concentration of incidents of one specific NC within one production stage. This measure shows the concentration of NC occurrences among the production machines for one production stage.

In order to relate the NC type with the indicated HHI number it is necessary to consult Table 5. In this table the NC types are stored in the form of a matrix with horizontal index letters and vertical index numbers. Comparing these indices reveals the NC type that matches its corresponding HHI index number.

In Table 4a NC8, which is located in cell 1b, has an HHI of 0.8 in stage $n-1$ and an HHI of 0.14 in stage n. NC8 records by far the highest concentration with the manufacturing machines in stage $n-1$. This value is close to 1.0, which indicates a very high concentration. This same NC (NC8) does not have the highest HHI for stage n. However, compared to the other NCs in that stage, it does show a high concentration. One may hypothesize that only one or just a few machines in stage $n-1$ are completely responsible for NC8's occurrence.

In Table 4b, NC14 which is located in cell 7b, shows an HHI of 0.21 in stage n and an HHI of 0.17 in stage $n-1$. NC14 shows the highest concentration for the manufacturing machines in stage n. This same NC (NC14) does not show the highest HHI for stage $n-1$. Thus, one may hypothesise that only a few machines in stage n are completely responsible for NC14's occurrence.

The concentration numbers across the stages are not comparable to each other. This reflects the different total number of machines at each step. However, within one production stage they do become comparable with each other, since the number of manufacturing machines is the same.

In order to analyze the results, the strategy is followed to observe the results for the highest concentrations for both stages. With regard to Table 4a the highest concentration HHI value is indicated in cell 1b which relates to NC8. Table 4b records the highest concentration for NC14 in cell 7b.

4.3.2. *The highest concentration measure of a given NC of stage $n-1$*

The following shows the result table for NC8 according to the selection criterion, with the highest HHI of stage $n-1$ presented in Table 6a. For the sake of completeness the result table of NC8 for stage n is also presented in Table 6b.

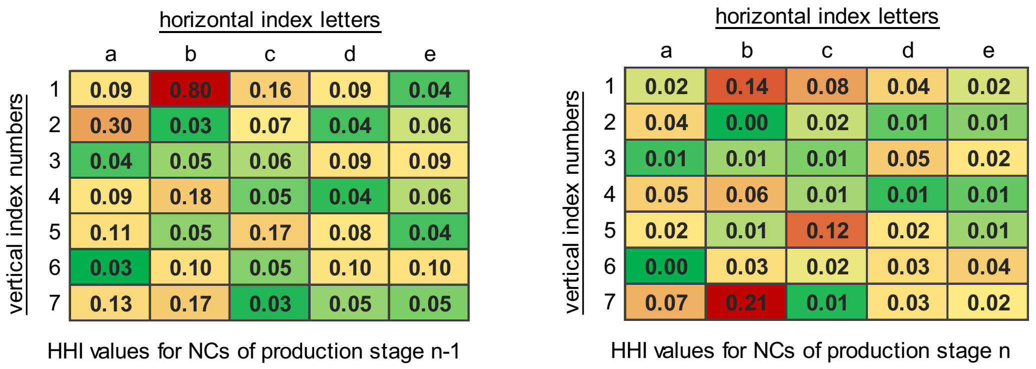
Production stage $n-1$ consists of fewer machines than in production stage n. The machines at each production stage operate in parallel and each production stage consists of exactly one machine processing the product being fabricated. The route that the product takes from one production stage to the other depends on the set-up configuration of the machines. Different configurations allow products varying in size, composition and shape.

Table 6a presents the number of occurrences of NC8 for every machine of production stage $n-1$. As one can see, machine number 2A is associated with 1478 NCs out of a total number of 1653 NCs. In other words, 89% of all NC8 occurrences are related with machine 2A for production stage $n-1$. All other machines of this production stage show numbers of occurrences between 0 and 82.

Table 6b presents the number of occurrences for every machine of production stage n. In comparison to stage $n-1$ the NCs are a bit more fragmented. Machine number K1, K2, L19 and L20 have a total number of 1187 occurrences. This means that 70% of all NC8 occurrences were associated with four machines in stage n. Of the remaining 30%, machines K15 and K16 relate to around 150 occurrences each, while all other machines are free of NCs or record only a very few.

The above results highlight the very high concentration of NCs associated with machine 2A of production stage $n-1$. Thus, the number of occurrences is very concentrated on one single

Table 4
(a) The HHI for production stage $n-1$ and (b) the HHI for production stage n according to the NCs.



horizontal index letters

vertical index numbers	a	b	c	d	e
1	0.09	0.80	0.16	0.09	0.04
2	0.30	0.03	0.07	0.04	0.06
3	0.04	0.05	0.06	0.09	0.09
4	0.09	0.18	0.05	0.04	0.06
5	0.11	0.05	0.17	0.08	0.04
6	0.03	0.10	0.05	0.10	0.10
7	0.13	0.17	0.03	0.05	0.05

HHI values for NCs of production stage n-1

horizontal index letters

vertical index numbers	a	b	c	d	e
1	0.02	0.14	0.08	0.04	0.02
2	0.04	0.00	0.02	0.01	0.01
3	0.01	0.01	0.01	0.05	0.02
4	0.05	0.06	0.01	0.01	0.01
5	0.02	0.01	0.12	0.02	0.01
6	0.00	0.03	0.02	0.03	0.04
7	0.07	0.21	0.01	0.03	0.02

HHI values for NCs of production stage n

Table 5
Corresponding NCs for the HHI in Table 4 for stage $n-1$ and stage n.

		Horizontal index letters				
		a	b	c	d	e
Vertical index numbers	1	**NC1**	**NC8**	**NC15**	**NC22**	**NC29**
	2	**NC2**	**NC9**	**NC16**	**NC23**	**NC30**
	3	**NC3**	**NC10**	**NC17**	**NC24**	**NC31**
	4	**NC4**	**NC11**	**NC18**	**NC25**	**NC32**
	5	**NC5**	**NC12**	**NC19**	**NC26**	**NC33**
	6	**NC6**	**NC13**	**NC20**	**NC27**	**NC34**
	7	**NC7**	**NC14**	**NC21**	**NC28**	**NC35**

production machine. This suggests that further steps are needed to analyze the root causes of NC8 at this specific machine (2A) at stage $n-1$. A possible tool to complement the analysis can be the cause and effect diagram from Table 1. This lists all possible contributing factors to look out for, such as operator error, machine failure, method errors or material problems. As such, the particular branch where the machine sits should be placed under scrutiny in order to methodically investigate the exact root cause, if there is one to find.

4.3.3 The highest concentration measures of a given NC of stage n

The result table for NC14 is presented in Table 7b, using the highest HHI of stage n as the selection criterion. Additionally, the result table of NC14 for stage $n-1$ is presented in Table 7a.

The same logic as presented in Section 4.3.2 can be followed in the analysis of Table 7a and b. However, the results in Table 7 are visibly very different to those seen in Table 6. The highest occurrences for NC14 are related with machine 32Q. This machine is responsible for 308 incidents, which represents 45% of all NCs produced in stage n. When looking at Table 7a, machine 2D shows 232 incidents, which represent 36% of all NCs produced in stage $n-1$. While the number of occurrences are by far the highest in stage n (Table 7b) there is a different picture for stage $n-1$ (Table 7a). Besides machine 2D, machine 6C and 4D also show moderately high incident rates.

Based on the information gleaned above, there is no strong reason to believe one machine to be the originator of NC14. In this case it is suggested that besides machine 32Q from stage n, machine 2D from stage $n-1$ should be selected for further analysis. However, when complementing the analysis with the cause and effect diagram there must be room for other possible contributor factors besides the machines, such as operators, methods or materials, among others.

When comparing the results in Table 6a and b a mismatch of the total numbers of NC occurrences becomes apparent. While the total number of NC8 in stage $n-1$ sums up to 1653, stage n shows 1699 occurrences of this NC. In theory both numbers should match since every product passes exactly one machine at each production stage. The reason for this inconsistency is incomplete datasets. These can be the result of technical defects with scanned barcodes or neglected data entry by operators, among other problems. The same type of inconsistency accounts for the discrepancies between Table 7a and b.

Similar matrices as presented in Tables 6a,b and 7a,b can be obtained for all other NCs presented in Table 5 [29] but including these figures is beyond the scope of this paper.

5. Conclusion

This paper proposed a methodology designed to help identify the root causes of nonconformities in a simpler and more agile manner. Following the methodology leads to the construction of tables providing a visual depiction of the problem areas. The general KDD framework provides a starting point which is then adapted to the problem in hand. This adaptation involved integrating a concentration measure (HHI) used in economics as the DM algorithm. A visual depiction of the knowledge generated helps identify the root causes in environments similar to mass production.

The proposed methodology can be used as a quality tool. Its use has been validated in an application to the automotive industry. The methodology produces data tables where cells are shaded according to the concentration of a specific incident linked to particular machines. These tables help single out the possible main contributors of NCs by making them visible to the user. With this information the root cause can be further investigated.

Results indicate that it is possible to identify individual machines as being possible sources for the NCs. The applied visualization technique reveals the machines that indicate a high concentration of a specific NC.

This quality tool serves as an instrument in the data analysis step. The information produced allows the user to perform selective analysis to identify the root cause and introduce corrective action. The highly visual results ease the interpretation and facilitate analysis to constantly improve manufacturing quality.

The methodology integrates different disciplines, namely IT, quality control and economics. Integrating a well-known economics concept into a general IT framework provides it with an additional field of application in the area of quality. Practitioners – such as quality engineers in industrial companies – can use the tool to identify root causes in high volume production processes

Table 6
(a) Result tables of NC8 allocated to machines of process stage $n-1$ and (b) stage n.

NC8		Line			
		A	B	C	D
Machine number	1	24	0	1	2
	2	1478	1	0	6
	3	1	0	2	0
	4	0	0	2	0
	5	0	0	0	14
	6	82	0	0	1
	7	0	0	0	15
	8	3	1	2	3
	9	7	1	3	1
	10	0	0	0	3

NC8		Line																			
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Machine numbber	1	0	0	0	0	0	0	0	0	0	0	267	0	1	0	0	0	15	0	0	0
	2	0	0	0	0	0	0	0	0	0	0	315	1	0	0	0	0	0	15	0	0
	3	0	0	4	0	0	0	0	0	0	0	0	3	0	1	0	0	2	0	0	0
	4	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	6	0	0	0	7	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1
	7	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	12	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	15	0	0	0	0	0	0	0	0	0	0	138	0	0	0	0	0	0	0	0	0
	16	0	0	0	0	1	0	0	0	0	0	154	0	0	0	1	0	0	0	1	0
	17	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0
	18	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0
	19	0			0			0	0	0	8	0	301	0	0	0	0	29	3	0	0
	20	0			0			0	1	0	2	0	304	0	1	1	1	7	0	0	0
	21							0	0	0	0	0	0	0		0		8	0	0	0
	22							0	0	0	0	1	1	0		0		7	0	0	0
	23								0		0	0						0	0	1	0
	24								0		0	0						0	0	0	1
	25																	38	4	0	0
	26																	23	0	0	0
	27																	0	0	0	0
	28																	0	0	0	0
	29																	0	0	0	0
	30																	0	0	0	0
	31																	0	0	0	
	32																	0	0	0	
	33																	0	0	0	
	34																	0	0	0	
	35																	0		0	
	36																	0		0	
	37																			0	
	38																			0	

Table 7
(a) Result presentation of NC14 allocated to machines of process stage $n-1$ and (b) stage n.

NC14		Line			
		A	B	C	D
Machine number	1	4	3	7	3
	2	2	3	4	232
	3	2	2	5	7
	4	17	1	7	74
	5	13	4	0	1
	6	4	6	103	7
	7	25	6	4	7
	8	3	7	10	6
	9	3	11	9	7
	10	4	4	6	23

NC14		Line																			
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Machine numbber	1	0	0	0	3	0	2	0	1	0	0	0	0	0	0	1	0	2	4	0	2
	2	1	0	1	0	0	1	1	0	0	0	0	1	0	0	4	1	0	2	0	0
	3	0	2	0	0	0	0	0	0	0	1	0	3	0	0	2	1	1	1	0	0
	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	5	0	0	3	0	0
	5	0	5	1	0	0	0	0	0	1	0	1	1	0	3	0	0	1	0	1	2
	6	0	9	0	2	0	1	0	0	0	0	0	1	1	1	0	1	0	1	0	1
	7	0	0	0	1	0	0	0	1	2	0	0	0	0	3	0	1	0	0	1	0
	8	0	0	1	0	1	2	0	0	0	0	0	1	1	2	0	0	1	0	1	1
	9	0	1	0	0	0	0	0	2	0	0	2	0	3	1	3	0	0	3	0	0
	10	0	2	2	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
	11	0	1	0	0	0	1	2	0	1	7	0	1	0	1	0	0	2	0	10	0
	12	0	0	0	1	0	0	0	0	0	5	0	0	0	5	1	1	0	0	24	0
	13	1	0	1	4	0	1	0	2	0	2	0	1	0	0	1	0	1	1	1	2
	14	0	0	1	2	0	0	0	1	0	5	2	3	0	2	0	1	0	0	3	2
	15	0	3	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	0	1	1
	16	0	2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
	17	0	0	0	0	0		0	1	0	5	1	1	1	1	0	0	0	1	0	1
	18	2	0	0	0	0		0	0	0	1	0	2	0	0	0	0	0	0	1	6
	19	0			2			0	0	1	2	1	0	0	0	0	1	0	2	1	2
	20	0			2			0	0	0	2	0	0	0	1	0	1	0	0	0	1
	21							0	1	0	2	0	0	0		0		2	0	0	0
	22							0	2	0	0	1	0	0		0		0	0	1	1
	23								0		0	0						1	1	0	0
	24								0		0	0						0	1	0	1
	25																	0	0	0	2
	26																	0	1	0	0
	27																	0	39	0	0
	28																	0	35	0	0
	29																	0	0	0	0
	30																	0	0	0	0
	31																	0	0	0	
	32																	308	0	0	
	33																	0	0	0	
	34																	0	2	0	
	35																	1	0	0	
	36																	0		0	
	37																			0	
	38																			0	

with numerous machines and diverse NCs. As such, the tool is applicable to industries other than the one in the application studied. However, it is advantageous if the database provides comparable data and information is similarly available. As with the results presented, the visual representation of the data helps in gaining a quick understanding of which NCs show the highest concentrations for machines across different production stages.

While initial findings are promising, further research is necessary. As a start, the success rate can be checked to further validate this tool. This can be done by recording the number of positive cases versus the false positive cases. In this respect, if the machine identified with this tool is the true origin of the specific NC, this constitutes a positive case. If the root cause is attributable to something else it would constitute a false positive case. This tool was developed to be used offline. However, with further developments the method can be integrated into an installed IT system of a company. After it is tailored to the specific conditions it can operate as an online tool. In this case, the initial data input step would become obsolete. Automated alerts can be set for critical values and time to react would be reduced.

As this paper demonstrates combining knowledge of different disciplines can result in the emergence of new methods, tools and knowledge. The authors highly encourage cross-discipline and interdisciplinary research.

Acknowledgements

The support of FCT Fundação para a Ciência e a Tecnologia (Portugal) under the PhD Grant SFRH/BD/33791/2009 is gratefully acknowledged.

The authors are also grateful to the anonymous company, their managers and employees for providing data, responding to questions and validating results.

References

[1] A.K. Choudhary, J.A. Harding, M.K. Tiwari, Data mining in manufacturing: a review based on the kind of knowledge, J. Intell. Manuf. 20 (5) (2009) 501–521.

[2] R.E. McQuater, C.H. Scurr, B.G. Dale, P.G. Hillman, Using quality tools and techniques successfully, TQM Mag. 7 (6) (1995) 37–42.

[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The KDD process for extracting useful knowledge from volumes of data, Commun. ACM 39 (11) (1996) 27–34.

[4] G.M. Bounds, Beyond Total Quality Management. Toward the Emerging Paradigm, McGraw-Hill, New York, 1994.

[5] J.L. Hradesky, Total Quality Management Handbook, McGraw-Hill, New York, 1995.

[6] J. Rampey and H. Roberts, Perspectives on total quality, in: Proceedings of the Total Quality Forum IV, Cincinnati, Ohio, 1992.

[7] U. Hellsten, B. Klefsjö, TQM as a management system consisting of values, techniques and tools, TQM Mag. 12 (4) (2000) 238–244.

[8] K. Ishikawa, Guide to Quality Control, Asian Productivity Organization, Tokyo, 1976.

[9] S. Mizuno, Management for Quality Improvement: The 7 New QC Tools, Cambridge, MA, Productivity Press, 1988.

[10] W. A. Shewhart, Economic Control of Quality of Manufactured Product, D. Van Nostrand Company, Inc., New York, 1931, 501.

[11] W.E. Deming, The New Economics: For Industry, Government, Education, The MIT Press, Cambridge, MA, 2000.

[12] J.M. Juran, F.M. Gyrna Jr, Quality Planning and Analysis, 2nd ed., McGraw-Hill, New York, 1980.

[13] K. Ishikawa, Introduction to Quality Control, Productivity Press, Taylor & Francis, 1990.

[14] T. Ōno, Toyota Production System: Beyond Large Scale Production, Productivitiy Pres, Cambridge, MA, 1988.

[15] S. Shingō, Zero Quality Control: Source Inspection and the Poka-Yoke System, Productivity Press, Productivity, Inc. Portland, OR, 1986.

[16] G. Taguchi, Introduction to Quality Engineering: Designing Quality into Products and Processes, 1986.

[17] S. Beckman, D. Rosenfield, Operations Strategy: Competing in the 21st Century, McGraw-Hill, 2008.

[18] B.G. Dale, R. McQuater, Managing Business Improvement and Quality: Implementing Key Tools and Techniques (Business Series), Blackwelll Business, Oxford, 1998.

[19] Q. Huang, S. Zhou, J. Shi, Diagnosis of multi-operational machining processes through variation propagation analysis, Robot. Comput.-Integr. Manuf. 18 (3) (2002) 233–239.

[20] D.C. Montgomery, Introduction to Statistical Quality Control, Vol. 2, Wiley, New York, 1991.

[21] R.E. McDermott, R.J. Mikulak, M.R. Beauregard, The Basics of FMEA, Productivity, Portland, 1996.

[22] M. Donauer, P. Peças and A.L. Azevedo, Nonconformity Tracking and Prioritization Matrix: An Approach for Selecting Nonconformities as a Contribution to the Field of TQM, Production Planning & Control (forthcoming).

[23] J.A. Harding, M. Shahbaz, Srinivas, A. Kusiak, Data mining in manufacturing: a review, J. Manuf. Sci. Eng. 128 (4) (2006) 969–976.

[24] G. Köksal, İ. Batmaz, M.C. Testik, A review of data mining applications for quality improvement in manufacturing industry, Expert Syst. Appl. 38 (10) (2011) 13448–13467.

[25] S.A. Rhoades, The Herfindahl–Hirschman index. (cover story), Fed. Reserve Bull. 79 (3) (1993) 188.

[26] A.O. Hirschman, The paternity of an index, Am. Econ. Rev. 54 (1964) 761–762.

[27] A.O. Hirschman, National Power and the Structure of Foreign Trade, University of California Press, Berkeley, 1969.

[28] S. Calkins, The new merger guidelines and the Herfindahl–Hirschman Index, Calif. Law Rev. 71.2 (1983) 402–429.

[29] M. Donauer, P. Peças, A. Azevedo, Nonconformity root causes analysis through a pattern identification approach, in: A. Azevedo (Ed.), Advances in Sustainable and Competitive Manufacturing Systems, Springer International Publishing, Switzerland, 2013, pp. 851–863.

[30] S. Du, L. Jun, X. Lifeng, A robust approach for root causes identification in machining processes using hybrid learning algorithm and engineering knowledge, J. Intell. Manuf. 23 (5) (2012) 1833–1847.

[31] J. Shi, S. Zhou, Quality control and improvement for multistage systems: a survey, IIE Trans. 41 (9) (2009) 744–753.

[32] J. Lian, X. Lai, Z. Lin, F.S. Yao, Application of data mining and process knowledge discovery in sheet metal assembly dimensional variation diagnosis, J. Mater. Process. Technol. 129 (2002) 315–320.

[33] N. Das, V. Prakash, Interpreting the out-of-control signal in multivariate control chart—a comparative study, Int. J. Adv. Manuf. Technol. 37 (2008) 966–979.

[34] R.S. Guh, On-line identification and quantification of mean shifts in bivariate processes using a neural network-based approach, Qual. Reliab. Eng. Int. 23 (2007) 367–385.