(18th ICPR) Data mining for improvement of product quality

Catherine M. da Cunha, Bruno Agard, A Kusiak

▶ **To cite this version:**

Catherine M. da Cunha, Bruno Agard, A Kusiak. (18th ICPR) Data mining for improvement of product quality. International Journal of Production Research, 2006, 44 (18-19), pp.4027-4041. 10.1080/00207540600678904 . hal-00512901

HAL Id: hal-00512901
https://hal.science/hal-00512901

Submitted on 1 Sep 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

(18th ICPR) Data mining for improvement of product quality

Journal:	*International Journal of Production Research*
Manuscript ID:	TPRS-2005-IJPR-0485.R1
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	28-Feb-2006
Complete List of Authors:	Da Cunha, Catherine; Institut National Polytechnique d eGrenoble, GILCO Agard, Bruno; Ecole polytechnique de Montréal, MAGI Kusiak, A; University of Iowa, Mechanical and industrial engineering
Keywords:	MASS CUSTOMIZATION, ASSEMBLY LINES, DATA MINING
Keywords (user):	

For Peer Review Only

scholarONE™
Manuscript Central

Revised February, 2006-02-28 *Special issue of IJPR for selected papers from 18th ICPR 2005*

Data Mining for Improvement of Product Quality

C. DA CUNHA†, B. AGARD‡ and A. KUSIAK§

†Laboratoire GILCO-INPG, 46 av Félix Viallet, 38031 Grenoble Cedex 1, France

‡Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal (Québec), H3C3A7, Canada

§ Intelligent Systems Laboratory, 3131 Seamans Center, Department of Mechanical and Industrial Engineering, The University of Iowa, IA 52242 - 1527, USA

For Peer Review Only

Abstract

The assemble-to-order strategy delays the final assembly operations of a product until a customer order is received. The modules used in the final assembly operation result in large product diversity. This production strategy reduces the customer waiting time for the product. As the lead-time is short, any product rework may violate the delivery time. Since quality tests can be performed on the stocked modules without impacting the assembly schedule, the quality of the final assembly operations should be the focus. The data mining approach presented in this paper uses the production data to determine the sequence of assemblies that minimizes the risk of producing faulty products. The extracted knowledge plays important role in sequencing modules and forming product families that minimize the cost of production faults. The concepts introduced in the paper are illustrated with numerical examples.

Keywords: Assemble-to-order, quality, data mining, mass customization, production strategy.

AMS Subject Classification: 90B50; 68P10; 90B25

1. Introduction

Faulty products lead to unnecessary expenses due to rework, repairing, recycling, and wasted time. “Zero fault” is an objective that industries are eager to reach. A variety of methods aims to achieve such goal, e.g., six sigma (Breyfogle 1999, Harry et al. 2000) and total productive maintenance (TPM) (Nakajima 1988). The study of past performance of production systems is necessary. The difficulty is in finding pertinent information as the data is stored in numerous forms and at different locations.

In this paper, data mining is used to extract knowledge from large data sets to improve the production quality. The emphasis is on the role of knowledge extraction in manufacturing quality in an assemble-to-order (ATO) environment. A methodology is proposed and validated on a pilot study. This stage of validation is a necessary work, a preliminary to tests on real instances.

Assemble-to-order is a production strategy that is particularly well suited when the customer tolerance for product delivery (waiting) time is low. The lower limit on the customer waiting time is the final assembly time of the product.

Any unexpected event or additional operations could violate the product delivery time constraint. Rework of a faulty product reduces the cost of faults when its cost is smaller than the cost of the lost material and labor. Nevertheless, when the product delivery time is a contractual requirement, overdue payment is added to the cost of rework. This type of faults should be particularly avoided.

The content of this paper is structured as follows. First, the background of the topic discussed in the paper is provided. Then, the use of information in assembly sequencing is addressed. The proposed methodology is described in Section 4. The paper concludes with computational results.

For Peer Review Only

2. Background

2.1. Diversity: An industrial example

To meet the customer needs, product diversity tends to grow and therefore it should be managed. The cost of offering a large product portfolio should not exceed the gains obtained by satisfying the range of customer needs (Child *et al.* 1991). It is then essential to find the range of diversity that minimizes the total cost (see Figure 1).

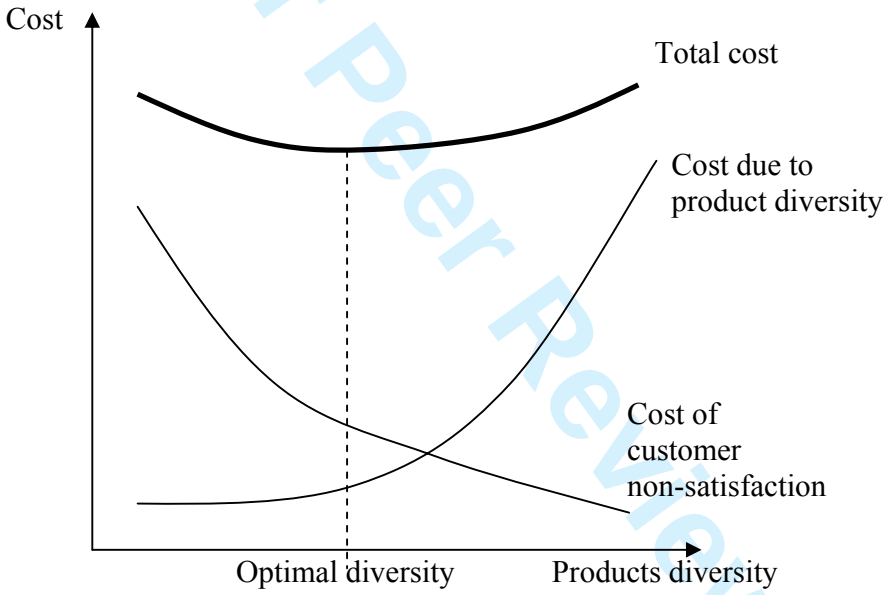

Figure 1: Diversity costs (Tarondeau 1998).

Different approaches have been used to address the product diversity challenge, e.g. design of product families, modular design, and delayed product differentiation. The assemble-to-order strategy links modular design and delayed product differentiation. Modules are built from basic parts and stocked, lastly the final assembly is performed after an actual order has been received. The product diversity is accomplished by a combinatorial association of basic parts.

For Peer Review Only

In this paper, an industrial example of electrical wire harnesses is discussed. It constitutes a major component of a vehicle as wires and connectors transmit electricity and information between different devices (see Figure 2).

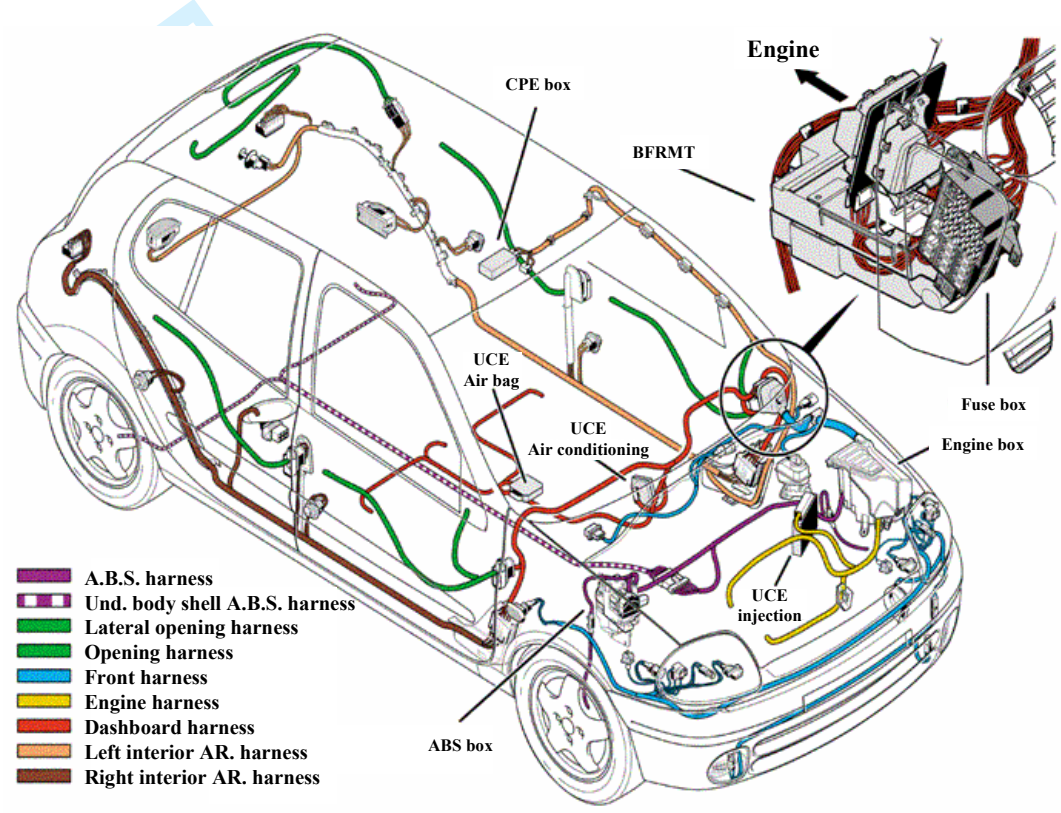

Figure 2: Wire harnesses in a car.

Control and activation of airbags, electrical windows, headlights, and so on are performed by wire harnesses. To illustrate the diversity of this product, consider a standard wire harness in a middle range performing 15 different functions. Depending on the silhouette and the engine type, these functions appear in different versions (up to 9). Potential diversity is then about 7 million of different wire harnesses for a unique car model (Agard and Tollenaere 2002).

In addition, there are inclusive and exclusive relations between the functions, e.g., the function “passenger air-bag” requires the function “driver air-bag”. Those relations reduce the actual diversity.

2.2. Costs

Evaluating the cost of product diversity is difficult. Even if direct costs, such as investment in new equipment or material costs, can be measured, indirect costs are difficult to estimate. Martin and Ishii (1997) proposed metrics to compare design alternatives based on the costs they induce, however this evaluation is difficult to perform in an industrial setting.

Product quality and complexity are interdependent. As the number of tasks performed by a worker increases, the number of errors may increase. The negative impact of product complexity seems unavoidable, unless the diversity is managed.

McDuffie *et al.* (1996) presented results of an international study in automotive industry. This study stressed the relationship between product diversity, productivity, and quality. The data included in the study indicated that when plants are adequately equipped to manage diversity, productivity is not significantly impacted by the scope of the product mix.

Like diversity costs, which are difficult to evaluate, savings due to process redesign are not easy to quantify particularly by the impact on the product quality. Besides, quantitative evaluation is recommended for criteria such as reduction of the time-to-market and flexibility improvement.

There are two main sources of fault costs:

- The loss of material when the faulty product is identified as such and discarded;
- The loss of image due to the non-conformance of customer requirements, when the faulty product reaches the market.

Let π be the added value by the rework operation. The term *added value* includes different sources of costs of a faulty product: the material, the equipment, and the labor associated

For Peer Review Only

with production and quality tests. The product rework is preferred over discarding of the faulty product, if and only if π is positive.

In a traditional production strategy the evaluation of rework cost is quite direct as it includes the cost of rework and the additional material needed. The balance is then:

$$\pi = product\ value - rework\ cost$$

In the case of assemble-to-order strategy such evaluation is more complex. When the delivery time is contractually required, the overdue payment needs to be considered, and the balance is then:

$$\pi = product\ value - rework\ cost - overdue\ payment$$

A trade-off between the cost of faults and rework is then needed. Furthermore, if savings due to quality improvements can be partially measured, in terms of reduction of the mean assembly-time and reduced material usage, the savings (or possibly gains) due to improved customer image of the product can not be directly evaluated.

2.3. Diversity management: Modularity and assembly-to-order strategies

Postponement strategy aims at reducing the risk associated with product diversity by delaying its differentiation, i.e. the stage after which the products assume their unique identities (Zinn 1990, Aviv and Federgruen 1999).

The modularity concept has been used for management of product diversity. APICS [1]defined modular production as the capacity to design and manufacture a set of modules that can be combined in a maximal numbers of ways (APICS 1998).

The choice of modular design implies rethinking the design process (Kusiak 1999). The modules created can be independent or not (i.e., that they can be assembled without requiring another module or not). Figure 3(a) shows modules that are interdependent, i.e., modules 1 and 3 can not be assembled unless module 2 has been installed, while the modules in Figure 3(b) can be assembled in any order.

[1] Advancing Productivity, Innovation, and Competitive Success. The association of operation management.

For Peer Review Only

One of the advantages of having independent modules is that re-sequencing of assembly sequence can be done without redesign the modules.

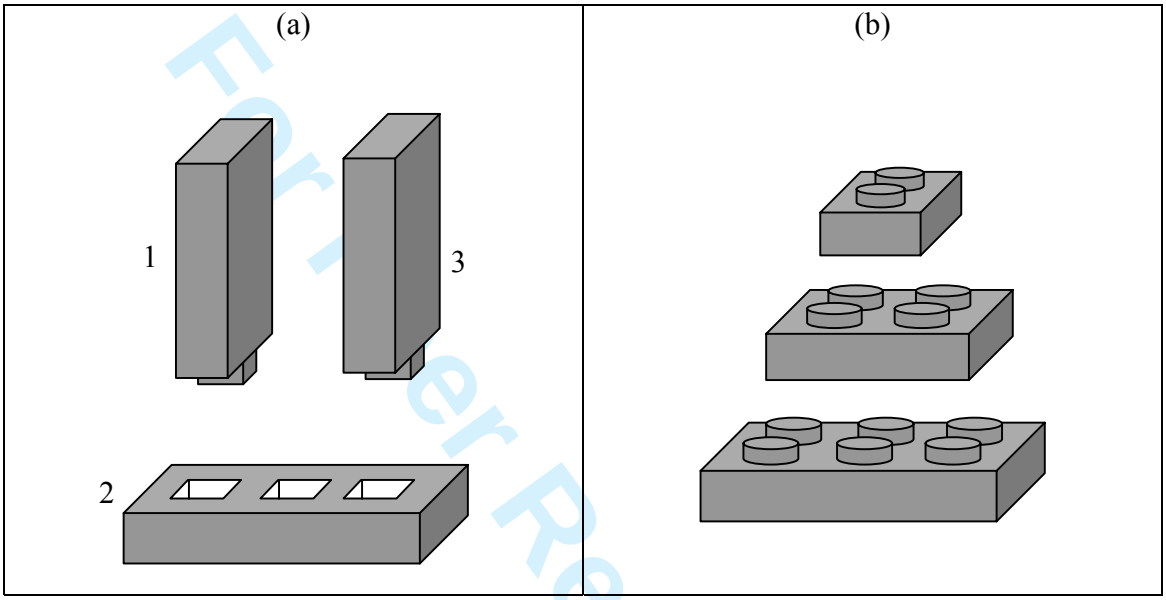

Figure 3: Examples of modular products.

3. Quality analysis

3.1. Fault identification

A product is considered as faulty when it does not meet its specifications. Inspection serves the purpose of identifying faults. When performed at the end of the assembly process, it comes at a significant cost of rework. In-process inspection detects faults as it occurs and thus the rework cost decreases.

Nevertheless, inspection is expensive and it may not be possible to test the product after each assembly operation. It is then essential to carefully determine the location and timing of inspection tests.

Inspection may involve different types of tests, e.g., exhaustive, random. Its role is to identify faulty products and control the quality of products. Statistical process control

For Peer Review Only

(SPC) detects variations in the product specifications and applies corrective actions to the process, mostly adjustment of equipment, before producing inadequate products (Quesenberry 1997).

The selection of the test protocol is generally driven by the characteristics of the product and the process. Consider production of needle syringe that it is monitored to avoid quality problems during its use. A faulty product can lead to medical problem such as bad penetration of the injected substance or breaking the needle. Therefore, the fixing system of the needle and barrel is controlled exhaustively (the inspection is performed visually either by a human or a video control device). However, the length of the plunger is not crucial and therefore the inspection is performed randomly.

3.2. Data mining applications

Anand and Büchner (1998) defined data mining as the discovery of non-trivial, implicit, previously unknown, and potentially useful and understandable patterns from large data sets. Data-mining algorithms have been applied in areas such as marketing (Berry and Linoff 1997), medicine (identification of genes impacting drug development, Shah and Kusiak 2004), and industrial design (Agard and Kusiak 2004).

The patterns extracted from production data could assist production managers in improving quality of products. Therefore, it is important to locate data enabling identification patterns of interest. Such data should be understandable for the data miner and the domain expert. It is also absolutely necessary that the extraction and transformation of the operational data can be done automatically and rapidly for the analysis to take place.

Production data can contain errors, e.g., due to data entry. In the problem discussed in the paper, the term parasite noise is not limited to this kind of errors; it also includes the other sources of faults, for example: faulty material, non-adapted equipment.

One of the aims of this research is to validate that data mining techniques can perform in presence of a non-negligible “errors rate”

For Peer Review Only

The main goal of this research is to improve the quality of assembly process with data mining. The challenge is in the extraction of associations between faults and assembly sequence in presence of parasite noise. Faults may come from sources such as assembly sequence, power outage, raw material, and other random phenomena.

When faulty patterns are identified, it becomes then possible to re-sequence the assembly operation or to rethink the tests policy (Kusiak and Huang 1997) in order to reduce the number of operations that have to be redone. It is therefore important to remember that rework operations are more expensive.

The following three-step approach is used to improve the product quality:

1. Identification of assembly sequences having an impact on quality;
2. Generation of a new sequence;
3. Generation of a new test policy.

4. Computational results

4.1. Input Data

The data mining approach used in this research was prototyped on a randomly generated data set. Consider here an example of a workshop; 6 different assembly operations can be performed. This production is monitored and production data is stored. This case will be used to stress the pertinence of our hypothesis that mining production data could provide production managers good information to improve the quality of their products.

A test instance was constituted of 250000[2] product's operation routes. The data was randomly generated with respect to the following constraints:

- An operation can be performed as a normal task (i.e., non-rework task) at most once per product

[2] TANAGRA's limit for the association rules extraction, see Section 4.3.

For Peer Review Only

- To reproduce the behavior of a real production system, random faults are generated. Because of a randomly phenomenon, 5% of the products need rework. When the first test proved the product to be faults, rework has to be done.
  The random data represent products with faults due to sources other than the assembly sequence. This characteristic of the input data is needed to validate the hypothesis that data mining can perform in a noisy production context.
- If the rework operations can not restore the quality characteristics, the product is considered as faulty and is destroyed.

Furthermore, systematic faulty sequences are introduced. The challenge of the data mining process will be to extract such patterns.

Examples of two always faulty operation sequences are:

- If operation 5 is not the last task, this operation has to be redone;
- If operation 2 precedes operation 4, operation 4 has to be redone.

The historical production records contain assembly sequences and quality tests. Depending on the results of the tests some assembly operations could have to be redone.

An example of a product route would be:

5-4-2-1-6-Tf-4-Tok

This route (represented in Figure 4.) indicates that this product after assembly operations 5, 4, 2, 1 and 6 was detected as faulty (Tf). It was then necessary to redo assembly task 4 (4). The final test proved the product to conform to quality standards (Tok).

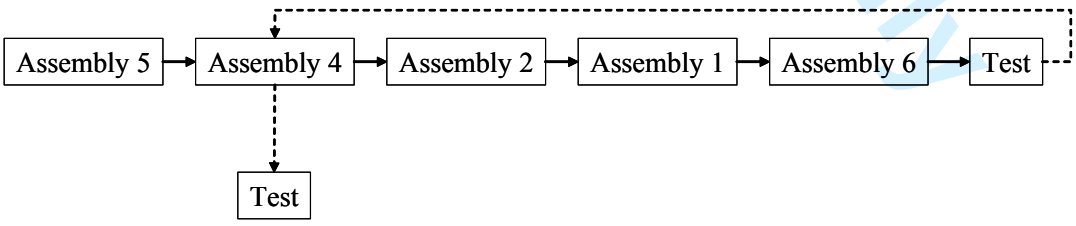

Figure 4: Example of a product route.

For Peer Review Only

The historical data is represented in a tabular form where each row contains a product route. For example the first route (top row) is in Table 1.

Table 1: Product routes.

Product route	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Test	Rework	Final test
#1	5	4	2	1	6		Tf	4	Tok
#2	1	4	5				Tok		
#3	6	2	5	3			Tf	5	Tok
#4	2	1	4				Tok		
#5	2	4					Tf	4	Tok

The six columns (Task 1 – Task 6) in Table 1 represent different assembly operations of the product. The sequence of the operations is indicated by the sequence of columns. Column 7 represents the status of the first quality test, column 8 the possible rework operation and the last column the final quality status of the product.

For each product, the number of operations is randomly affected from [[1,6]] thus the first line of Table 1 represents the product #1 requiring 5 operations while the second line represents the product #2 that only requires 3 operations. The first test operation follows the final operation. The sequences are randomly generated.

There are two kinds of faults, the ones generated by faulty sequences and the others that are the result of random events (e.g. power outage).

The results of the first test of the generated products follow these constraints:

If operation 5 is not the last task, this operation has to be redone;

If operation 2 immediately precedes operation 4, operation 4 has to be redone; } **Faulty sequences**

There are about 5% of faulty products resulting from random events.

The rework is efficient in 80% of the cases. } **Noise**

For Peer Review Only

4.2. Data mining implementation

TANAGRA (data mining software for research and education[3]) was chosen as a knowledge extraction tool. The association rule algorithm was used to extract rules (Agrawal and Srikant 1994).

One of the challenges of data mining is to manage different types of data (Chen *et al.* 1996). Here, the difficulty was to find a technique to identify the sequences causing faults. Therefore, the structure of the input data had to be modified. As the frequency of faulty products is low, the support of the rules is low as well. Therefore, it is interesting to study two data sets: one with all routes and one describing only faulty product routes. The first set allows identifying sequences leading to quality products and the second to understand the sources of faults. The data of Table 1 has been separated in two tables (Table 2 and Table 3).

Table 2: Routes of quality products.

Product route	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Test	Rework	Final test
#2	1	4	5				Tok		
#4	2	1	4				Tok		

Table 3: Routes of faulty products.

Product route	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Test	Rework	Final test
#1	5	4	2	1	6		Tf	4	Tok
#3	6	2	5	3			Tf	5	Tok
#5	2	4					Tf	4	Tok

[3] http://eric.univ-lyon2.fr/~ricco/tanagra/index.html

For Peer Review Only

Patterns extracted from Table 2 could be used to improve the motivation of the assembly line operators. For example, if an operation never needs rework, the operators performing it could be rewarded.

Since the focus of this paper is the identification of faulty patterns, only information presented in the format of Table 3 will be considered for further analysis.

Sequences such as: 1-4-2 and 4-2-5 should be recognized has sharing the sub-sequence 4-2 of two operations. Such patterns could not be easy to extract when dealing directly with the data presented as in Table 1, Table 2 or Table 3.

In order to perform such identification, another representation of the routes is required.
The example sequence **5-4-2-1-6-Tf-4-Tok** represented in the first row of Table 3, is represented as the set **{B5, 5_4, 4_2, 2_1, 1_6, 6F, R4}** in Table 4. The precedence information is contained in the data itself. The routes can then be encoded in a binary form Table 4 represents the same information as Table 3, however, in a different form. Column B1 states that operation 1 is the first task (beginning of the assembly process), respectively column B5 states that operation 5 is the first task. Column 1F states that operation 1 is the final task. The last column "Rework" represents the rework operations. Rework = 4 states that operation 4 has to be redone.
Note that the fact that the first test indicates that the product is faulty, is included in the route in Table 4 as the information of column "Rework". Only faulty products are selected and quality products would just have the Rework column that would be empty.

For Peer Review Only

Table 4: New representation of faulty product routes.

B1	B2	B3	B4	B5	B6	1_2	1_3	1_4	1_5	1_6	1F	2_1	2_3	2_4	2_5	2_6	2F	3_1	…	3F	4_1	4_2	…	4F	…	5_3	5_4	…	6_2	…	6F	Rework
				1						1		1										1				1				1	4
					1										1					1						1			1			5
	1													1										1								4

A post-processing of the rules extracted (from the data in the format of Table 4) is required to eliminate redundancies. Furthermore, as the process quality is of interest, only the rules with the "test" as an outcome are considered.

It is common in data mining that human expertise is needed to assess and validate the extracted knowledge. The latter is possible due to the knowledge acquired by experience (Craven *et al*. 2000). Thus an expert may be able to detect rules that do not ally to a particular context. For example, consider the following two rules:

Rule a: Any product requiring operation 6 is reworked more often than any other product.

Rule b: Operation 6 never needs rework.

An expert could interpret these rules as follows: The worker performing operation 6 is highly skilled; however, s/he distracts his co-workers by singing while working.

4.3. Results: Extracted rules

Association rules

The extracted association rules have the format shown in Table 5. The quality of the rules is measured with support and confidence (the last two columns of Table 5), both defined in (Agresti 1990). Let's recall how these two metrics work, support corresponds to the probability that an individual in the tested population possesses both the "antecedent" and the "consequent" of the association rule, the confidence is the probability that a given individual possesses the consequent knowing that it possesses the "antecedent". Consider a simple example: upon 100 people 50 are women and 8 have red hair; only two women have

For Peer Review Only

red hair. The association rule (woman ⇒ red hair) has support of 0.02 (2/100) and confidence of 0.01 (2/50).

Table 5: Extracted association rules.

No.	Antecedent	Consequent	Support	Confidence
1	"2_4 = 1"	"rework = 4"	0.165	0.961
2	"5_4 = 1"	"rework = 5"	0.217	1
3	"5_1 = 1"	"rework = 5"	0.162	0.98
4	"5_3 = 1"	"rework = 5"	0.158	0.979
5	"B5 = 1"	"rework = 5"	0.368	0.948
6	"5_2 = 1"	"rework = 5"	0.185	0.895
7	"B2 = 1"	"rework = 5"	0.195	0.886
8	"2_4 = 0"	"rework = 5"	0.725	0.877
9	"1F = 1"	"rework = 5"	0.182	0.832
10	"3F = 1"	"rework = 5"	0.163	0.817
11	"2F = 1" and "B2 = 1"	"rework = 2"	0.012	1

A human expert could make the association rules more understandable, e.g.:

Rules 2, 3, 4, 5, and 6 have similar meaning: If operation 5 is not the last task, it needs to be redone.

Rule 1: If operation 2 precedes operation 4, the product has to be reworked.

Rule 11: If operation 2 is the only operation, it needs to be redone.

The rules with confidence lower than 1 may represent patterns that are not always true. It is important that improving the quality of a system be not limited to the search for systematic failures, but also considers sources of faults. Faults may have multiple sources and finding the exact root causes requires synthesis of information residing at various data bases. “Partial truth” may be a source of potential improvements.

To improve product quality, it is of paramount importance to master the production processes, it is thus essential to consider in the “analysis and decision” step information that describes systematic faults and the correction of theses faults should have priority over all

For Peer Review Only

the other quality actions. Therefore the association rules that have a support lower than the noise ratio (here 5%) are not considered in the "analysis and decision" step. Nevertheless that kind of information may be important in a further stage: when all internal and systematic causes of faults are under control, the random causes should then be considered

Summary results

Consider an example where a data mining algorithm finds two major links between rework operations and the assembly operations. The pertinence of the rules was evaluated using the Tschuprow's T indicator[4] as shown in Table 6.

Table 6: Summary results.

Row	Column	Tschuprow's T	Cross-tabulation
Rework	2_4	0.717	"Rework = 4 and 2_4 = 1", 99 occurrences
Rework	6F	0.657	"Rework = 6 and 6F = 1", 2 occurrences

Strong associations were found between rework of operation 4 and the operations sequence **operation 2-operation 4** (supported with 99 occurrences). This validates rule 1 listed in Section 4.1.

The source of the rework of operation 6 was also clarified as operation 6 is performed at the end of the assembly line. Nevertheless, this information should be handled carefully because of the rare frequency of this event (only two products needed the rework of operation 6).

One of the limits of data mining techniques is that the relationships identified between two (or more) events prove the existence of an association between the members of the rule. Nevertheless, the existence of causality relationships can not be established and the identification of the cause and consequence between the (antecedent, consequent) couple can not be automated. This identification can only be done by the domain experts.

[4] Tschuprow's T varies between 0 and 1, T = 0 states the independence (in the mathematical sense) of the two variables, T = 1 states that the link between the variables explicates the entire observation.

For Peer Review Only

4.4. Analysis and decisions

The rules determine new sequences of assembly operations. For example, **association rule 1** indicates re-sequencing of operations 2 and 4:

The product requiring operations 2 and 4 should always pass through operation 4 first.

Rules 2, 3, 4, 5, and 6 point out to assembly re-sequencing; operation 5 has to be performed as the last one to minimize the risk of inducing faulty products.

The incorporation of **Rule 11** will not lead to re-sequencing of operations, however, this operation should be dealt with as it often leads to faults, e.g., different tools could be used.

Introducing an additional test operation after the second task could be considered. Products needing rework would be detected sooner and the rework would be less costly.

Of course, any reorganization of the process has its consequences, and to predict them it is not easy. Therefore the whole process of data-analysis and system reorganization should be on a regular basis to determine the best configuration (Figure 5).

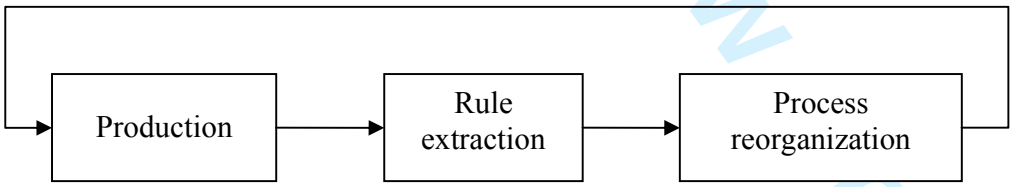

Figure 5: Improvement scheme.

Furthermore, production data could be temporal, e.g., due to seasons. Therefore, it is critical to consider a possible link between the production quality and time period.

To incorporate this characteristic, data, other than operations and quality status should be considered, e.g., studying the impact of the work shift could lead to a redesign of break schedule (e.g., 2 breaks of 15 minutes instead of 1 break of 30 minutes).

For Peer Review Only

5. Limits and extensions of the model

The proposed model organizes the data to represent the assembly sequences that immediately precede or succeed another. The original data describing the assembly operations (presented in Table 1) are filtered (in Table 2 and Table 3) and transformed (the new representation is shown in Table 4) in order to highlight the relationships between sequences and faults. This transformation only stresses the immediate precedent of an operation.

In different applications, other operations that precede (not necessarily the immediate precedent) are also likely to have significant impact on the quality. A different data transformation is required. Two cases are described:

Extension 1 – A critical operation, adding a massive component for example, may make all the subsequent assemblies more difficult to operate: the workers may have to work around the component, the access to the product may be more difficult and the quality of all the assemblies that follow may be endangered.
It is then necessary to represent that operation for each subsequent operation.
Consider the critical operation is operation 3. Implications of that operation on the subsequent operations and faults may be identified with the following transformation.
For example, the sequence 2-**3**-5-7-6-9 induces a supplementary data set {3_5, 3_7, 3_6, 3_9}. That additional data is included in Table 4 with the previously described one. It may emphasize quality problems that come from 3_9 for example, even if operation 3 is not an immediate predecessor for operation 9.

Extension 2 – Many operations may be critical operations. It is then necessary to apply *Extension 1* to all critical operations that need to be evaluated.

For Peer Review Only

It may be not necessary and even not efficient to consider each operation as if it was a critical operation.

In order to identify what are the critical operations it is easy to transform Table 3 to represent for each product route what are the operations concerned (see Table 7).

Table 7: Table for detection of critical operations.

Product route	Operation 1	Operation 2	Operation 3	Operation 4	Operation 5	Operation 6	Test
#1	1	1	1	1	1	1	Tf
#3		1	1		1	1	Tf
#5		1		1			Tf
Total	**1**	**3**	**2**	**2**	**2**	**2**	

It is then better to start with the operations corresponding to the highest total.

6. Conclusion

This paper discussed the application of data mining for quality improvement of assembly operations. The focus was on the impact of assembly sequences on the quality of products. The initial hypothesis was that the data mining algorithms were effective in identifying the potential quality issues in presence of noise (that comes in part from unexpected events). Therefore, the data considered in this research included random events that occur in production systems.

The computational results confirmed that the source of assembly sequence faults can be detected with association rules, even in the presence of noise. The rules extracted with data mining algorithms can be used to improve production quality by avoiding "risky" sequences.

Further investigation will focus on using the method described in this article on a real case data, for example the electrical wire harness. This stage will permit to study the robustness of this approach.

For Peer Review Only

The selection of an assemble-to-order strategy is not limited to the technical solutions. Any change in production strategy has to be completed with organizational rethinking (Benson *et al*. 1991). Further research should consider data about other sources of faults, including human factors and material data.

7. References

Agard B. and Tollenaere M., 2002, Design of wire harnesses for mass customization, 4th international conference on integrated design and manufacturing in technical engineering, IDMME 2002, Clermont-Ferrand, France, CD-ROM.

Agard B. and Kusiak A., 2004, A Data-Mining Based Methodology for the Design of Product Families, International Journal of Production Research, Vol. 42, No. 15, pp. 2955-2969.

Agrawal R. and Srikant R., 1994, Fast algorithms for mining association rules in large databases, in Proc. International Conference on Large Databases, pp. 478-499.

Agresti A., 1990, Categorical Data Analysis, Wiley, New York.

Anand S. and Büchner A., 1998, Decision Support Using Data Mining, Financial Times Pitman Publishers, London, UK.

APICS, 1998, APICS Dictionary, 9th Edition, Falls Church, VA.

Aviv Y. and Federgruen A., 1999, The Benefits of Design for Postponement, Chap. 19 in Quantitative Models for Supply Chain Management, pp. 553-584, in Tayur S., Ganeshan R., and Magazine M. (Eds), Kluwer Academic Publishers, Boston, MA.

Berry M. and Linoff G., 1997, Data Mining Techniques: For Marketing, Sales, and Customer Support, Wiley, New York.

Benson P., Saraph J., and Schroeder R., 1991, The effects of organizational context on quality management: an empirical investigation, Management Science, Vol. 37, No. 9, pp. 1107–1124.

For Peer Review Only

Breyfogle F.W., 1999, Implementing Six Sigma: Smarter Solutions Using Statistical Methods, John Wiley, New York.

Chen M.-S., Han J., and Yu P., 1996, Data mining: An overview from a database perspective, IEEE Transactions on Knowledge and Data Engineering Vol. 8, No. 6, pp. 866-883.

Child P., Diederichs R. and Sanders F.-H., Wisniowski S., 1991, The management of complexity, Sloan Management Review, Vol. 33, No. 1, pp. 73-80.

Craven M., DiPasquo D., Freitag D., McCallum A., Mitchell T., Nigam K. and Slattery S., 2000, Learning to construct knowledge bases for the world wide web, Artificial Intelligence, Vol. 118, No. 1-2, pp. 69-113.

Harry M., Schroeder R., and Linsenmann D.R., 2000, Six Sigma: The Breakthrough Management Strategy Revolutionizing the World's Top Corporations, Currency, New York, NY.

Kusiak A. and Huang C.C., 1997, Design of Modular Digital Circuits for Testability, IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part C, Vol. 20, No. 1, pp. 48-57.

Kusiak A., 1999, Engineering Design: Products, Processes, and Systems, Academic Press, San Diego, CA.

MacDuffie J., Sethuraman K., and Fisher M., 1996, Product variety and manufacturing performance: evidence from the international automotive assembly plant study, Management Science, Vol. 42, No. 3, pp. 350-369.

Martin M. and Ishii K., 1997, Design for Variety: Development of Complexity Indices and Design Charts, ASME Design Engineering Technical Conferences, DETC.

Nakajima S., 1988, Introduction to TPM: Total Productive Maintenance, Productivity Press, Cambridge, MA.

Quesenberry C.P., 1997, SPC Methods for Quality Improvement, Wiley, New York.

Shah S. and Kusiak A., 2004, Data Mining and Genetic Programming Based Gene/SNP Selection, Artificial Intelligence in Medicine, Vol. 31, No. 3, pp. 183-258.

For Peer Review Only

Tarondeau J-C., 1998, Stratégie Industrielle, 2nd Edition, Vuibert, France.

Zinn W., 1990, Should you assemble products before an order is received? Business Horizons, No. 5, pp. 70-73.

Acknowledgements

The authors wish to acknowledge the support of the Natural Science and Engineering Research Council of Canada (NSERC). The research project has been also supported by the Fonds de Recherche sur la Nature et les Technologies (FQRNT).

For Peer Review Only

Data Mining for Improvement of Product Quality

C. DA CUNHA†, B. AGARD‡ and A. KUSIAK§

†Laboratoire GILCO-INPG, 46 av Félix Viallet, 38031 Grenoble Cedex 1, France

‡Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal (Québec), H3C3A7, Canada

§ Intelligent Systems Laboratory, 3131 Seamans Center, Department of Mechanical and Industrial Engineering, The University of Iowa, IA 52242 - 1527, USA

For Peer Review Only

Abstract

The assemble-to-order strategy delays the final assembly operations of a product until a customer order is received. The modules used in the final assembly operation result in large product diversity. This production strategy reduces the customer waiting time for the product. As the lead-time is short, any product rework may violate the delivery time. Since quality tests can be performed on the stocked modules without impacting the assembly schedule, the quality of the final assembly operations should be the focus. The data mining approach presented in this paper uses the production data to determine the sequence of assemblies that minimizes the risk of producing faulty products. The extracted knowledge plays important role in sequencing modules and forming product families that minimize the cost of production faults. The concepts introduced in the paper are illustrated with numerical examples.

Keywords: Assemble-to-order, quality, data mining, mass customization, production strategy.

AMS Subject Classification: 90B50; 68P10; 90B25

1. Introduction

Faulty products lead to unnecessary expenses due to rework, repairing, recycling, and wasted time. “Zero fault” is an objective that industries are eager to reach. A variety of methods aims to achieve such goal, e.g., six sigma (Breyfogle 1999, Harry et al. 2000) and

total productive maintenance (TPM) (Nakajima 1988). The study of past performance of production systems is necessary. The difficulty is in finding pertinent information as the data is stored in numerous forms and at different locations.

In this paper, data mining is used to extract knowledge from large data sets to improve the production quality. The emphasis is on the role of knowledge extraction in manufacturing quality in an assemble-to-order (ATO) environment. A methodology is proposed and validated on a pilot study. This stage of validation is a necessary work, a preliminary to tests on real instances.

Assemble-to-order is a production strategy that is particularly well suited when the customer tolerance for product delivery (waiting) time is low. The lower limit on the customer waiting time is the final assembly time of the product.

Any unexpected event or additional operations could violate the product delivery time constraint. Rework of a faulty product reduces the cost of faults when its cost is smaller than the cost of the lost material and labor. Nevertheless, when the product delivery time is a contractual requirement, overdue payment is added to the cost of rework. This type of faults should be particularly avoided.

The content of this paper is structured as follows. First, the background of the topic discussed in the paper is provided. Then, the use of information in assembly sequencing is addressed. The proposed methodology is described in Section 4. The paper concludes with computational results.

2. Background

2.1. Diversity: An industrial example

To meet the customer needs, product diversity tends to grow and therefore it should be managed. The cost of offering a large product portfolio should not exceed the gains obtained by satisfying the range of customer needs (Child *et al.* 1991). It is then essential to find the range of diversity that minimizes the total cost (see Figure 1).

For Peer Review Only

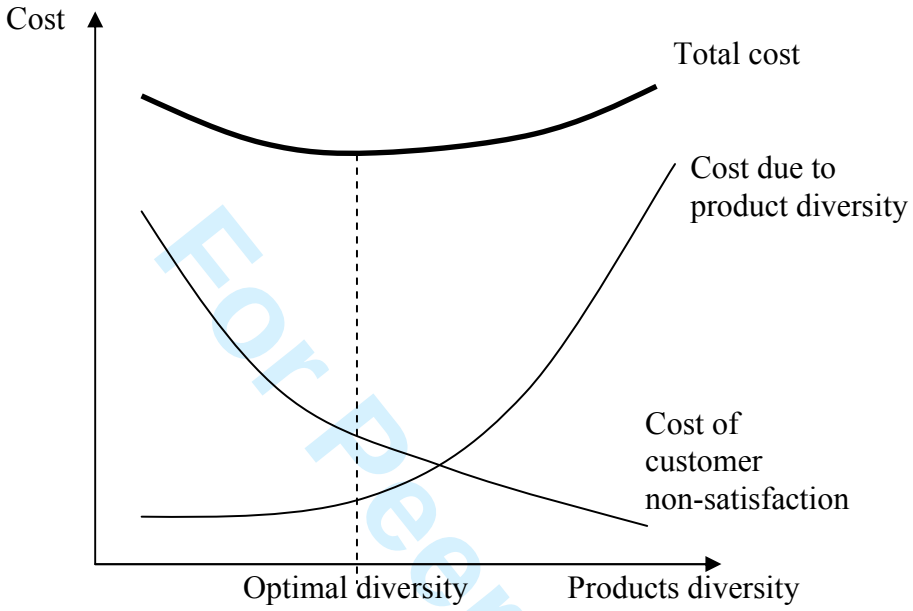

Figure 1: Diversity costs (Tarondeau 1998).

Different approaches have been used to address the product diversity challenge, e.g. design of product families, modular design, and delayed product differentiation. The assemble-to-order strategy links modular design and delayed product differentiation. Modules are built from basic parts and stocked, lastly the final assembly is performed after an actual order has been received. The product diversity is accomplished by a combinatorial association of basic parts.

In this paper, an industrial example of electrical wire harnesses is discussed. It constitutes a major component of a vehicle as wires and connectors transmit electricity and information between different devices (see Figure 2).

For Peer Review Only

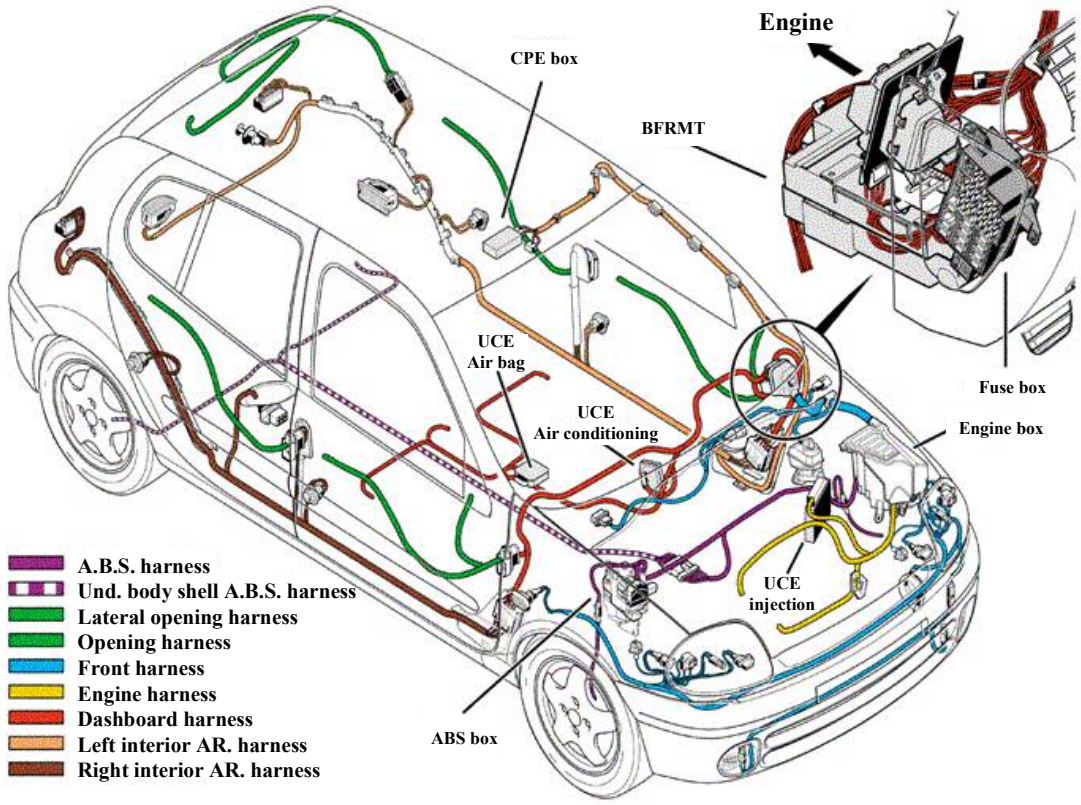

Figure 2: Wire harnesses in a car.

Control and activation of airbags, electrical windows, headlights, and so on are performed by wire harnesses. To illustrate the diversity of this product, consider a standard wire harness in a middle range performing 15 different functions. Depending on the silhouette and the engine type, these functions appear in different versions (up to 9). Potential diversity is then about 7 million of different wire harnesses for a unique car model (Agard and Tollenaere 2002).

In addition, there are inclusive and exclusive relations between the functions, e.g., the function “passenger air-bag” requires the function “driver air-bag”. Those relations reduce the actual diversity.

For Review Only

2.2. Costs

Evaluating the cost of product diversity is difficult. Even if direct costs, such as investment in new equipment or material costs, can be measured, indirect costs are difficult to estimate. Martin and Ishii (1997) proposed metrics to compare design alternatives based on the costs they induce, however this evaluation is difficult to perform in an industrial setting.

Product quality and complexity are interdependent. As the number of tasks performed by a worker increases, the number of errors may increase. The negative impact of product complexity seems unavoidable, unless the diversity is managed.

McDuffie *et al.* (1996) presented results of an international study in automotive industry. This study stressed the relationship between product diversity, productivity, and quality. The data included in the study indicated that when plants are adequately equipped to manage diversity, productivity is not significantly impacted by the scope of the product mix.

Like diversity costs, which are difficult to evaluate, savings due to process redesign are not easy to quantify particularly by the impact on the product quality. Besides, quantitative evaluation is recommended for criteria such as reduction of the time-to-market and flexibility improvement.

There are two main sources of fault costs:

- The loss of material when the faulty product is identified as such and discarded;
- The loss of image due to the non-conformance of customer requirements, when the faulty product reaches the market.

Let π be the added value by the rework operation. The term *added value* includes different sources of costs of a faulty product: the material, the equipment, and the labor associated with production and quality tests. The product rework is preferred over discarding of the faulty product, if and only if π is positive.

In a traditional production strategy the evaluation of rework cost is quite direct as it includes the cost of rework and the additional material needed. The balance is then:

$$\pi = \textit{product value} - \textit{rework cost}$$

For Peer Review Only

In the case of assemble-to-order strategy such evaluation is more complex. When the delivery time is contractually required, the overdue payment needs to be considered, and the balance is then:

$$\pi = product\ value - rework\ cost - overdue\ payment$$

A trade-off between the cost of faults and rework is then needed. Furthermore, if savings due to quality improvements can be partially measured, in terms of reduction of the mean assembly-time and reduced material usage, the savings (or possibly gains) due to improved customer image of the product can not be directly evaluated.

2.3. Diversity management: Modularity and assembly-to-order strategies

Postponement strategy aims at reducing the risk associated with product diversity by delaying its differentiation, i.e. the stage after which the products assume their unique identities (Zinn 1990, Aviv and Federgruen 1999).

The modularity concept has been used for management of product diversity. APICS [1]defined modular production as the capacity to design and manufacture a set of modules that can be combined in a maximal numbers of ways (APICS 1998).

The choice of modular design implies rethinking the design process (Kusiak 1999). The modules created can be independent or not (i.e., that they can be assembled without requiring another module or not). Figure 3(a) shows modules that are interdependent, i.e., modules 1 and 3 can not be assembled unless module 2 has been installed, while the modules in Figure 3(b) can be assembled in any order.

One of the advantages of having independent modules is that re-sequencing of assembly sequence can be done without redesign the modules.

[1] Advancing Productivity, Innovation, and Competitive Success. The association of operation management.

For Peer Review Only

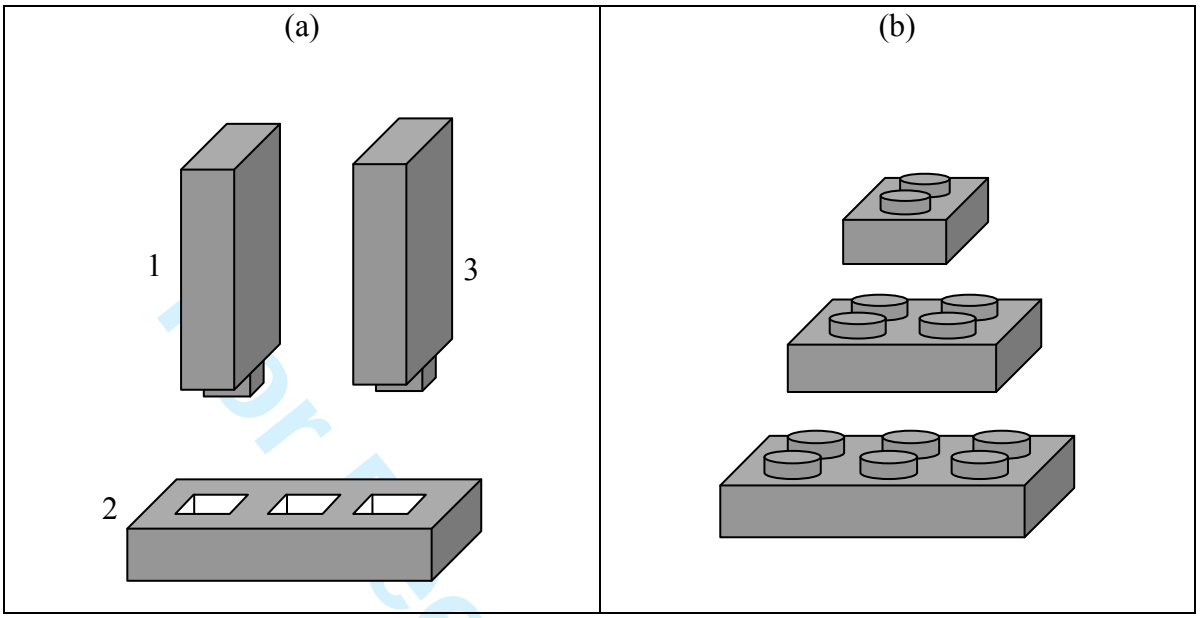

Figure 3: Examples of modular products.

3. Quality analysis

3.1. Fault identification

A product is considered as faulty when it does not meet its specifications. Inspection serves the purpose of identifying faults. When performed at the end of the assembly process, it comes at a significant cost of rework. In-process inspection detects faults as it occurs and thus the rework cost decreases.

Nevertheless, inspection is expensive and it may not be possible to test the product after each assembly operation. It is then essential to carefully determine the location and timing of inspection tests.

Inspection may involve different types of tests, e.g., exhaustive, random. Its role is to identify faulty products and control the quality of products. Statistical process control (SPC) detects variations in the product specifications and applies corrective actions to the process, mostly adjustment of equipment, before producing inadequate products (Quesenberry 1997).

For Peer Review Only

The selection of the test protocol is generally driven by the characteristics of the product and the process. Consider production of needle syringe that it is monitored to avoid quality problems during its use. A faulty product can lead to medical problem such as bad penetration of the injected substance or breaking the needle. Therefore, the fixing system of the needle and barrel is controlled exhaustively (the inspection is performed visually either by a human or a video control device). However, the length of the plunger is not crucial and therefore the inspection is performed randomly.

3.2. Data mining applications

Anand and Büchner (1998) defined data mining as the discovery of non-trivial, implicit, previously unknown, and potentially useful and understandable patterns from large data sets. Data-mining algorithms have been applied in areas such as marketing (Berry and Linoff 1997), medicine (identification of genes impacting drug development, Shah and Kusiak 2004), and industrial design (Agard and Kusiak 2004).

The patterns extracted from production data could assist production managers in improving quality of products. Therefore, it is important to locate data enabling identification patterns of interest. Such data should be understandable for the data miner and the domain expert. It is also absolutely necessary that the extraction and transformation of the operational data can be done automatically and rapidly for the analysis to take place.

Production data can contain errors, e.g., due to data entry. In the problem discussed in the paper, the term parasite noise is not limited to this kind of errors; it also includes the other sources of faults, for example: faulty material, non-adapted equipment.

One of the aims of this research is to validate that data mining techniques can perform in presence of a non-negligible “errors rate”

The main goal of this research is to improve the quality of assembly process with data mining. The challenge is in the extraction of associations between faults and assembly

For Peer Review Only

sequence in presence of parasite noise. Faults may come from sources such as assembly sequence, power outage, raw material, and other random phenomena.

When faulty patterns are identified, it becomes then possible to re-sequence the assembly operation or to rethink the tests policy (Kusiak and Huang 1997) in order to reduce the number of operations that have to be redone. It is therefore important to remember that rework operations are more expensive.

The following three-step approach is used to improve the product quality:

1. Identification of assembly sequences having an impact on quality;
2. Generation of a new sequence;
3. Generation of a new test policy.

4. Computational results

4.1. Input Data

The data mining approach used in this research was prototyped on a randomly generated data set. Consider here an example of a workshop; 6 different assembly operations can be performed. This production is monitored and production data is stored. This case will be used to stress the pertinence of our hypothesis that mining production data could provide production managers good information to improve the quality of their products.

A test instance was constituted of 250000[2] product's operation routes. The data was randomly generated with respect to the following constraints:

- An operation can be performed as a normal task (i.e., non-rework task) at most once per product

- To reproduce the behavior of a real production system, random faults are generated. Because of a randomly phenomenon, 5% of the products need rework. When the first test proved the product to be faults, rework has to be done.

[2] TANAGRA's limit for the association rules extraction, see Section 4.3.

For Peer Review Only

The random data represent products with faults due to sources other than the assembly sequence. This characteristic of the input data is needed to validate the hypothesis that data mining can perform in a noisy production context.

- If the rework operations can not restore the quality characteristics, the product is considered as faulty and is destroyed.

Furthermore, systematic faulty sequences are introduced. The challenge of the data mining process will be to extract such patterns.

Examples of two always faulty operation sequences are:

- If operation 5 is not the last task, this operation has to be redone;
- If operation 2 precedes operation 4, operation 4 has to be redone.

The historical production records contain assembly sequences and quality tests. Depending on the results of the tests some assembly operations could have to be redone.

An example of a product route would be:

5-4-2-1-6-Tf-4-Tok

This route (represented in Figure 4.) indicates that this product after assembly operations 5, 4, 2, 1 and 6 was detected as faulty (Tf). It was then necessary to redo assembly task 4 (4). The final test proved the product to conform to quality standards (Tok).

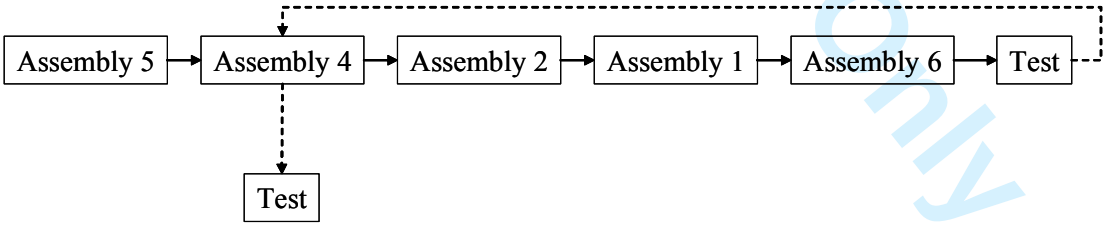

Figure 4: Example of a product route.

The historical data is represented in a tabular form where each row contains a product route. For example the first route (top row) is in Table 1.

For Peer Review Only

Table 1: Product routes.

Product route	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Test	Rework	Final test
#1	5	4	2	1	6		Tf	4	Tok
#2	1	4	5				Tok		
#3	6	2	5	3			Tf	5	Tok
#4	2	1	4				Tok		
#5	2	4					Tf	4	Tok

The six columns (Task 1 – Task 6) in Table 1 represent different assembly operations of the product. The sequence of the operations is indicated by the sequence of columns. Column 7 represents the status of the first quality test, column 8 the possible rework operation and the last column the final quality status of the product.

For each product, the number of operations is randomly affected from [[1,6]] thus the first line of Table 1 represents the product #1 requiring 5 operations while the second line represents the product #2 that only requires 3 operations. The first test operation follows the final operation. The sequences are randomly generated.

There are two kinds of faults, the ones generated by faulty sequences and the others that are the result of random events (e.g. power outage).

The results of the first test of the generated products follow these constraints:

If operation 5 is not the last task, this operation has to be redone;

If operation 2 immediately precedes operation 4, operation 4 has to be redone; } **Faulty sequences**

There are about 5% of faulty products resulting from random events. } **Noise**

The rework is efficient in 80% of the cases.

For Peer Review Only

4.2. Data mining implementation

TANAGRA (data mining software for research and education[3]) was chosen as a knowledge extraction tool. The association rule algorithm was used to extract rules (Agrawal and Srikant 1994).

One of the challenges of data mining is to manage different types of data (Chen *et al.* 1996). Here, the difficulty was to find a technique to identify the sequences causing faults. Therefore, the structure of the input data had to be modified. As the frequency of faulty products is low, the support of the rules is low as well. Therefore, it is interesting to study two data sets: one with all routes and one describing only faulty product routes. The first set allows identifying sequences leading to quality products and the second to understand the sources of faults. The data of Table 1 has been separated in two tables (Table 2 and Table 3).

Table 2: Routes of quality products.

Product route	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Test	Rework	Final test
#2	1	4	5				Tok		
#4	2	1	4				Tok		

Table 3: Routes of faulty products.

Product route	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Test	Rework	Final test
#1	5	4	2	1	6		Tf	4	Tok
#3	6	2	5	3			Tf	5	Tok
#5	2	4					Tf	4	Tok

[3] http://eric.univ-lyon2.fr/~ricco/tanagra/index.html

For Peer Review Only

Patterns extracted from Table 2 could be used to improve the motivation of the assembly line operators. For example, if an operation never needs rework, the operators performing it could be rewarded.

Since the focus of this paper is the identification of faulty patterns, only information presented in the format of Table 3 will be considered for further analysis.

Sequences such as: 1-4-2 and 4-2-5 should be recognized has sharing the sub-sequence 4-2 of two operations. Such patterns could not be easy to extract when dealing directly with the data presented as in Table 1, Table 2 or Table 3.

In order to perform such identification, another representation of the routes is required.
The example sequence **5-4-2-1-6-Tf-4-Tok** represented in the first row of Table 3, is represented as the set **{B5, 5_4, 4_2, 2_1, 1_6, 6F, R4}** in Table 4. The precedence information is contained in the data itself. The routes can then be encoded in a binary form Table 4 represents the same information as Table 3, however, in a different form. Column B1 states that operation 1 is the first task (beginning of the assembly process), respectively column B5 states that operation 5 is the first task. Column 1F states that operation 1 is the final task. The last column "Rework" represents the rework operations. Rework = 4 states that operation 4 has to be redone.
Note that the fact that the first test indicates that the product is faulty, is included in the route in Table 4 as the information of column "Rework". Only faulty products are selected and quality products would just have the Rework column that would be empty.

For Peer Review Only

Table 4: New representation of faulty product routes.

B1	B2	B3	B4	B5	B6	1_2	1_3	1_4	1_5	1_6	1F	2_1	2_3	2_4	2_5	2_6	2F	3_1	…	3F	4_1	4_2	…	4F	…	5_3	5_4	…	6_2	…	6F	Rework
				1						1		1										1					1				1	4
					1										1					1						1			1			5
	1													1										1								4

A post-processing of the rules extracted (from the data in the format of Table 4) is required to eliminate redundancies. Furthermore, as the process quality is of interest, only the rules with the "test" as an outcome are considered.

It is common in data mining that human expertise is needed to assess and validate the extracted knowledge. The latter is possible due to the knowledge acquired by experience (Craven *et al*. 2000). Thus an expert may be able to detect rules that do not ally to a particular context. For example, consider the following two rules:

Rule a: Any product requiring operation 6 is reworked more often than any other product.

Rule b: Operation 6 never needs rework.

An expert could interpret these rules as follows: The worker performing operation 6 is highly skilled; however, s/he distracts his co-workers by singing while working.

4.3. Results: Extracted rules

Association rules

The extracted association rules have the format shown in Table 5. The quality of the rules is measured with support and confidence (the last two columns of Table 5), both defined in (Agresti 1990). Let's recall how these two metrics work, support corresponds to the probability that an individual in the tested population possesses both the "antecedent" and the "consequent" of the association rule, the confidence is the probability that a given individual possesses the consequent knowing that it possesses the "antecedent". Consider a simple example: upon 100 people 50 are women and 8 have red hair; only two women have

For Peer Review Only

red hair. The association rule (woman ⇒ red hair) has support of 0.02 (2/100) and confidence of 0.01 (2/50).

Table 5: Extracted association rules.

No.	Antecedent	Consequent	Support	Confidence
1	"2_4 = 1"	"rework = 4"	0.165	0.961
2	"5_4 = 1"	"rework = 5"	0.217	1
3	"5_1 = 1"	"rework = 5"	0.162	0.98
4	"5_3 = 1"	"rework = 5"	0.158	0.979
5	"B5 = 1"	"rework = 5"	0.368	0.948
6	"5_2 = 1"	"rework = 5"	0.185	0.895
7	"B2 = 1"	"rework = 5"	0.195	0.886
8	"2_4 = 0"	"rework = 5"	0.725	0.877
9	"1F = 1"	"rework = 5"	0.182	0.832
10	"3F = 1"	"rework = 5"	0.163	0.817
11	"2F = 1" and "B2 = 1"	"rework = 2"	0.012	1

A human expert could make the association rules more understandable, e.g.:

Rules 2, 3, 4, 5, and 6 have similar meaning: If operation 5 is not the last task, it needs to be redone.

Rule 1: If operation 2 precedes operation 4, the product has to be reworked.

Rule 11: If operation 2 is the only operation, it needs to be redone.

The rules with confidence lower than 1 may represent patterns that are not always true. It is important that improving the quality of a system be not limited to the search for systematic failures, but also considers sources of faults. Faults may have multiple sources and finding the exact root causes requires synthesis of information residing at various data bases. “Partial truth” may be a source of potential improvements.

To improve product quality, it is of paramount importance to master the production processes, it is thus essential to consider in the “analysis and decision” step information that describes systematic faults and the correction of theses faults should have priority over all

For Peer Review Only

the other quality actions. Therefore the association rules that have a support lower than the noise ratio (here 5%) are not considered in the "analysis and decision" step. Nevertheless that kind of information may be important in a further stage: when all internal and systematic causes of faults are under control, the random causes should then be considered

Summary results

Consider an example where a data mining algorithm finds two major links between rework operations and the assembly operations. The pertinence of the rules was evaluated using the Tschuprow's T indicator[4] as shown in Table 6.

Table 6: Summary results.

Row	Column	Tschuprow's T	Cross-tabulation
Rework	2_4	0.717	"Rework = 4 and 2_4 = 1", 99 occurrences
Rework	6F	0.657	"Rework = 6 and 6F = 1", 2 occurrences

Strong associations were found between rework of operation 4 and the operations sequence **operation 2-operation 4** (supported with 99 occurrences). This validates rule 1 listed in Section 4.1.

The source of the rework of operation 6 was also clarified as operation 6 is performed at the end of the assembly line. Nevertheless, this information should be handled carefully because of the rare frequency of this event (only two products needed the rework of operation 6).

One of the limits of data mining techniques is that the relationships identified between two (or more) events prove the existence of an association between the members of the rule. Nevertheless, the existence of causality relationships can not be established and the identification of the cause and consequence between the (antecedent, consequent) couple can not be automated. This identification can only be done by the domain experts.

[4] Tschuprow's T varies between 0 and 1, T = 0 states the independence (in the mathematical sense) of the two variables, T = 1 states that the link between the variables explicates the entire observation.

For Peer Review Only

4.4. Analysis and decisions

The rules determine new sequences of assembly operations. For example, **association rule 1** indicates re-sequencing of operations 2 and 4:

The product requiring operations 2 and 4 should always pass through operation 4 first.

Rules 2, 3, 4, 5, and 6 point out to assembly re-sequencing; operation 5 has to be performed as the last one to minimize the risk of inducing faulty products.

The incorporation of **Rule 11** will not lead to re-sequencing of operations, however, this operation should be dealt with as it often leads to faults, e.g., different tools could be used.

Introducing an additional test operation after the second task could be considered. Products needing rework would be detected sooner and the rework would be less costly.

Of course, any reorganization of the process has its consequences, and to predict them it is not easy. Therefore the whole process of data-analysis and system reorganization should be on a regular basis to determine the best configuration (Figure 5).

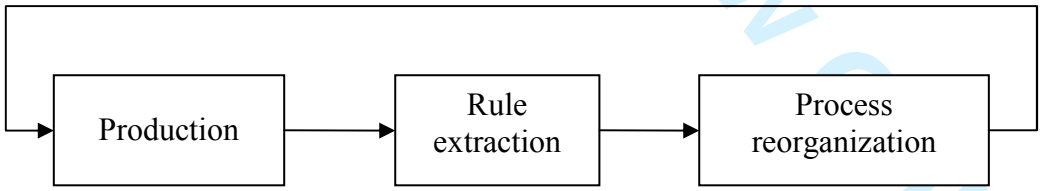

Figure 5: Improvement scheme.

Furthermore, production data could be temporal, e.g., due to seasons. Therefore, it is critical to consider a possible link between the production quality and time period.

To incorporate this characteristic, data, other than operations and quality status should be considered, e.g., studying the impact of the work shift could lead to a redesign of break schedule (e.g., 2 breaks of 15 minutes instead of 1 break of 30 minutes).

For Peer Review Only

5. Limits and extensions of the model

The proposed model organizes the data to represent the assembly sequences that immediately precede or succeed another. The original data describing the assembly operations (presented in Table 1) are filtered (in Table 2 and Table 3) and transformed (the new representation is shown in Table 4) in order to highlight the relationships between sequences and faults. This transformation only stresses the immediate precedent of an operation.

In different applications, other operations that precede (not necessarily the immediate precedent) are also likely to have significant impact on the quality. A different data transformation is required. Two cases are described:

Extension 1 – A critical operation, adding a massive component for example, may make all the subsequent assemblies more difficult to operate: the workers may have to work around the component, the access to the product may be more difficult and the quality of all the assemblies that follow may be endangered.
It is then necessary to represent that operation for each subsequent operation.
Consider the critical operation is operation 3. Implications of that operation on the subsequent operations and faults may be identified with the following transformation.
For example, the sequence 2-**3**-5-7-6-9 induces a supplementary data set {3_5, 3_7, 3_6, 3_9}. That additional data is included in Table 4 with the previously described one. It may emphasize quality problems that come from 3_9 for example, even if operation 3 is not an immediate predecessor for operation 9.

Extension 2 – Many operations may be critical operations. It is then necessary to apply *Extension 1* to all critical operations that need to be evaluated.

For Peer Review Only

It may be not necessary and even not efficient to consider each operation as if it was a critical operation.

In order to identify what are the critical operations it is easy to transform Table 3 to represent for each product route what are the operations concerned (see Table 7).

Table 7: Table for detection of critical operations.

Product route	Operation 1	Operation 2	Operation 3	Operation 4	Operation 5	Operation 6	Test
#1	1	1	1	1	1	1	Tf
#3		1	1		1	1	Tf
#5		1		1			Tf
Total	**1**	**3**	**2**	**2**	**2**	**2**	

It is then better to start with the operations corresponding to the highest total.

6. Conclusion

This paper discussed the application of data mining for quality improvement of assembly operations. The focus was on the impact of assembly sequences on the quality of products.

The initial hypothesis was that the data mining algorithms were effective in identifying the potential quality issues in presence of noise (that comes in part from unexpected events). Therefore, the data considered in this research included random events that occur in production systems.

The computational results confirmed that the source of assembly sequence faults can be detected with association rules, even in the presence of noise. The rules extracted with data mining algorithms can be used to improve production quality by avoiding "risky" sequences.

Further investigation will focus on using the method described in this article on a real case data, for example the electrical wire harness. This stage will permit to study the robustness of this approach.

For Peer Review Only

The selection of an assemble-to-order strategy is not limited to the technical solutions. Any change in production strategy has to be completed with organizational rethinking (Benson *et al*. 1991). Further research should consider data about other sources of faults, including human factors and material data.

7. References

Agard B. and Tollenaere M., 2002, Design of wire harnesses for mass customization, 4th international conference on integrated design and manufacturing in technical engineering, IDMME 2002, Clermont-Ferrand, France, CD-ROM.

Agard B. and Kusiak A., 2004, A Data-Mining Based Methodology for the Design of Product Families, International Journal of Production Research, Vol. 42, No. 15, pp. 2955-2969.

Agrawal R. and Srikant R., 1994, Fast algorithms for mining association rules in large databases, in Proc. International Conference on Large Databases, pp. 478-499.

Agresti A., 1990, Categorical Data Analysis, Wiley, New York.

Anand S. and Büchner A., 1998, Decision Support Using Data Mining, Financial Times Pitman Publishers, London, UK.

APICS, 1998, APICS Dictionary, 9th Edition, Falls Church, VA.

Aviv Y. and Federgruen A., 1999, The Benefits of Design for Postponement, Chap. 19 in Quantitative Models for Supply Chain Management, pp. 553-584, in Tayur S., Ganeshan R., and Magazine M. (Eds), Kluwer Academic Publishers, Boston, MA.

Berry M. and Linoff G., 1997, Data Mining Techniques: For Marketing, Sales, and Customer Support, Wiley, New York.

Benson P., Saraph J., and Schroeder R., 1991, The effects of organizational context on quality management: an empirical investigation, Management Science, Vol. 37, No. 9, pp. 1107–1124.

For Peer Review Only

Breyfogle F.W., 1999, Implementing Six Sigma: Smarter Solutions Using Statistical Methods, John Wiley, New York.

Chen M.-S., Han J., and Yu P., 1996, Data mining: An overview from a database perspective, IEEE Transactions on Knowledge and Data Engineering Vol. 8, No. 6, pp. 866-883.

Child P., Diederichs R. and Sanders F.-H., Wisniowski S., 1991, The management of complexity, Sloan Management Review, Vol. 33, No. 1, pp. 73-80.

Craven M., DiPasquo D., Freitag D., McCallum A., Mitchell T., Nigam K. and Slattery S., 2000, Learning to construct knowledge bases for the world wide web, Artificial Intelligence, Vol. 118, No. 1-2, pp. 69-113.

Harry M., Schroeder R., and Linsenmann D.R., 2000, Six Sigma: The Breakthrough Management Strategy Revolutionizing the World's Top Corporations, Currency, New York, NY.

Kusiak A. and Huang C.C., 1997, Design of Modular Digital Circuits for Testability, IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part C, Vol. 20, No. 1, pp. 48-57.

Kusiak A., 1999, Engineering Design: Products, Processes, and Systems, Academic Press, San Diego, CA.

MacDuffie J., Sethuraman K., and Fisher M., 1996, Product variety and manufacturing performance: evidence from the international automotive assembly plant study, Management Science, Vol. 42, No. 3, pp. 350-369.

Martin M. and Ishii K., 1997, Design for Variety: Development of Complexity Indices and Design Charts, ASME Design Engineering Technical Conferences, DETC.

Nakajima S., 1988, Introduction to TPM: Total Productive Maintenance, Productivity Press, Cambridge, MA.

Quesenberry C.P., 1997, SPC Methods for Quality Improvement, Wiley, New York.

Shah S. and Kusiak A., 2004, Data Mining and Genetic Programming Based Gene/SNP Selection, Artificial Intelligence in Medicine, Vol. 31, No. 3, pp. 183-258.

For Peer Review Only

Tarondeau J-C., 1998, Stratégie Industrielle, 2nd Edition, Vuibert, France.

Zinn W., 1990, Should you assemble products before an order is received? Business Horizons, No. 5, pp. 70-73.

Acknowledgements

The authors wish to acknowledge the support of the Natural Science and Engineering Research Council of Canada (NSERC). The research project has been also supported by the Fonds de Recherche sur la Nature et les Technologies (FQRNT).

For Peer Review Only