JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Assessment and analysis of data quality systems for the production process

[1]Md Zakir Hossan, [2]Ahnaf Aiman Abdi, [3]KR Hossain

[1]Post Graduate Degree, [2]Graduate Degree, [3] PhD Researcher.
[1]Department of Mathematics, National University, Dhaka, Bangladesh.
[2]School of Information Engineering, Zhengzhou University, Henan, China.
[3]Chinese Academy of Sciences, Beijing, China.

Abstract: Manufacturing enterprises rely on information systems such as manufacturing execution systems to achieve efficient production planning and control, but enterprises are a reality. In the information system implementation process, many obstacles are often encountered, and the implementation of information systems fails. Among these failures, inaccuracies in data and poor data quality in a broader sense are considered to be significant factors. To address this issue, a production-control-oriented data quality assessment and analysis framework are proposed that is divided into four steps: building a data model, selecting data and metrics, implementing metrics, and analyzing results. This framework helps enterprises identify the causes of data quality problems by assessing the implementation quality of production management information systems and the current production control data quality level. Finally, we have realized this framework and verification using a Chinese auto parts manufacturer as an example. The results show that the framework can provide effective guidance for quality assessment and analysis of production control data.

Index Terms **- production planning and control; data quality, manufacturing enterprises, MES, quality and analytical framework.**

1.0. INTRODUCTION

To achieve efficient production planning and control, manufacturing enterprises need to be stored in MES and other information management systems as support. Poor data quality means that product managers can not accurately understand the current production situation of the enterprise through data, which will lead to production managers' challenge in finding production site problems in time and making adjustments, ultimately affecting the production performance of enterprises. To solve the problem of data quality management in production planning and control scenarios, this paper proposes a data quality assessment and analysis framework for production control to help manufacturing enterprises evaluate the current production data quality and find the root causes that may lead to poor data quality.

2.0. RELATED LITERATURE

The MIT team proposed total data quality management in 1998 [1]. Data quality management is divided into four parts: defining, measuring, analyzing, and improving. Definitions involve data quality concepts and dimensions, generally defined as "the degree to which data is fit to use." In contrast, data quality is considered to be a multi-dimensional concept, and the often mentioned data quality dimensions include accuracy, completeness, consistency, and timeliness [2]. Measurement is the quantification of the performance of data in various quality dimensions with specific values and the need to build data quality. Metrics are used to implement measurements, and standard measures include subjective scores and objective measures such as ratios, maximum and minimum values, and weighted averages. The analysis involves analyzing the root causes of data quality issues and low costs due to data quality. Promotion is the adoption of various measures to improve data quality, generally divided into two categories: data-driven, focusing on correcting data, and process-driven,[3] the latter focusing on controlling or redesigning the data generation process. Researchers have proposed many data quality frameworks after comprehensive data quality management was first proposed. The main content of these frameworks can basically be covered in the four parts of total data quality management, but the different frameworks in the degree of detail.

The phases and steps to focus on, the policies and techniques adopted, and the dimensions and measures defined vary. English L [7] and LOSHIN proposed that COLDQ pay particular attention to the possible costs to businesses from poor data quality. The DQA proposed by PIPINO et al. [5] focuses on a comparative analysis of data quality measurements. LEE et al. proposed that AIMQ has built a set of subjective scales for assessing enterprise data quality. The CDQ proposed by BATINI et al. [6] introduces the concept of data quality objectives to guide data quality improvement. SEBASTIAN [10]. The proposed DQAF constructs multiple types of measurements for data quality. The TBDQs proposed by VAZIRI et al. focus on the process by which data is generated.

Some researchers are concerned about the data quality problem in the production control scenario, Gao Zhiyong et al. found in the analysis of the production data of the chassis workshop of a sewing equipment factory that the actual feeding data of the plant was inconsistent with the theoretical feeding data; HAUG et al. [12] found that there were generally inaccurate, untimely, incomplete, and duplicate product master data in the enterprises; CAO and ZHU [13] analyzed inconsistent inventory turnover data and work order completion delays of an air conditioning equipment manufacturer through the normal accident theory quality issues; SCHUH

et al.[14] Found, medium-sized machining in Germany Quality problems in process data are widespread in enterprises; ZONG et al.[15] used the IP-MAP modeling method to report on the missing and inaccurate BOM data of an auto parts manufacturer, inaccurate production plans and reported labor problems such as untimely data were analyzed.

As can be seen from the literature, the existing data quality framework is at a highly abstract level and does not provide details on how it will be implemented; Research on the quality of production control data has not been adequately studied. Concerned that there is still a lack of a complete data quality framework for production control, especially for how to define and measure the quality of production control data, although researchers generally agree that root cause analysis is a necessary step in data quality management, there is a problem with how to proceed with the root causes the analysis still lacks sufficient research.

3.0. DATA QUALITY ASSESSMENT AND PRODUCTION CONTROL ANALYTICAL FRAMEWORK

This paper presents a data quality assessment and analysis framework for production governance, as shown in Figure 1, which consists of four parts: building a data model, selecting data and measures, implementing measurements, and analyzing measurement results. Build a data model to model the data to understand the semantics of the data, such as for relational databases, to clarify the meaning and relationships of each data table and field. Select data and measures to pick key datasets to work with study objects and construct data quality metrics to measure data quality. Implementation measurements are based on constructed data quality metrics, with data quality measurements queried through a database. Analyze the metric results. On the one hand, descriptive analysis of the current data quality level based on the measurement results, on the other hand, for the abnormal data of the quality measurement results, through the investigation and analysis of the process of its generation, collection and entry, it is found to the root cause that may cause its quality measurements to be abnormal.

3.1. SELECT DATA AND MEASURES

For the definition of production control content, this paper follows the internal production planning and control part of the Aachen production planning and control model in this part. Production control is to control production factors such as people, machines, materials, laws, and rudiments at the production site to overcome possible production interruptions and other problems and ensure that the final production plan can be completed. Accurate data is necessary to achieve this goal, which can generally be divided into production master data, planning data, and feedback (reporting) data [20].

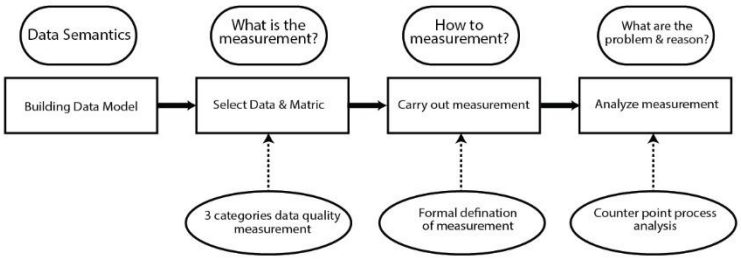

Figure 1: Data Quality Assessment and Analysis Framework

The production master data stores information such as materials, products, equipment, processes, BOMs, etc., which define how the product is produced. Planning data and feedback (job reporting) are stored concerning each other. Information on which production will be carried out and what product has been carried out. This paper focuses on program data and feedback (job reporting) data above.

After you identify the critical data that needs to be evaluated for quality, you need to determine from which dimensions to evaluate data quality. As described in this article, the commonly used data quality dimensions are accuracy, completeness, consistency, and timeliness. With BATINI and SCANNAPIECO [6] Definitions:

(1) Accuracy can be divided into grammatical accuracy and semantic accuracy. The former refers to the proximity of the data value to its corresponding value range, and the latter, exponentially. How close the data value is to the actual value represented by its target.

(2) Integrity refers to the degree to which data or records are missing; the degree of deletion is called column integrity, and the degree of record deletion is called overall integrity.

(3) Inconsistency refers to data that violates predefined semantic rules.

(4) Timeliness refers to the fact that changes in the state of the natural world are reflected in the correspondence. The degree of timeliness of data updates. Figure 2 shows the data quality assessment space for production control.

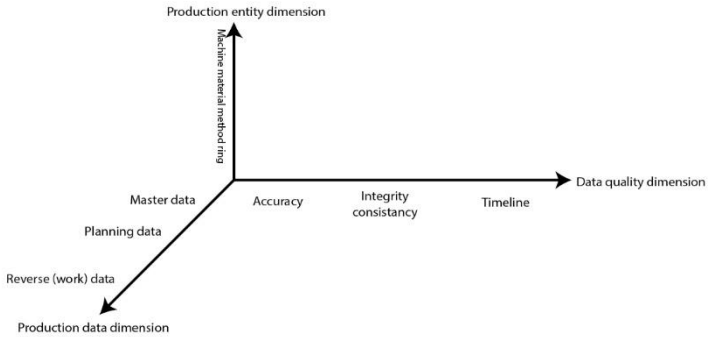

Figure 2: Data Quality Assessment Space for Production Control

The ability to construct an effective data quality measure to measure the performance of the data in each dimension depends on whether there is sufficient prior knowledge of the measured data. The semantic accuracy, overall completeness, and timeliness of the measured data require knowledge of the actual value, the valid overall quantity, and the time of real-world change. The grammatical accuracy, column integrity, and consistency of the measurement data require knowing the value range of the data, the missing values of the representation data, and the semantic rules that the data should follow. In practical application scenarios, reference data is often unknowable or requires a high cost to be known, resulting in data quality not always being measured in semantic accuracy, overall completeness, and timeliness. In contrast, the data's range, the missing values that characterize the data, and the semantic rules the data should follow can be learned through the data dictionary, the data user's expectations for data quality, and the formed business rules. This paper mainly explores the measurement of data syntax accuracy, column integrity, and consistency, especially for data consistency, and will focus on what aspects of production control-related data should be consistent and how to portray consistency as the semantic rules that data should follow. For data semantics, Measurements of accuracy, overall completeness, and timeliness are beyond the scope of this article. In summary, this article proposes the following three types of data quality measures.

(1) Column integrity measurement and grammatical accuracy of data. Given a relational data schema R, which contains a collection of properties denoted {A,...,A_1l}, DA_i is denoted as the domain of attribute A, so that R is an instance of the relationship, t ∈ R is the tuple of the relationship instance. Its number is denoted as | R |, and remember t. Ai is the value of the tuple t on attribute A, then the column integrity and syntactic accuracy of the relational instance R on a specific attribute Ai. It can be calculated by equation (1) and equation (2), respectively:

$$\frac{|\{t\in R|null(t.A_i)=0\}|}{|R|} \quad \ldots\ldots\ldots\ldots (1)$$

$$\frac{|\{t\in R|t.A_i\in D^A\}|}{|R|} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where null(t.A_i) = 0 indicates that the tuple t is not null on attribute A.

(2) The internal consistency of the plan data and the feedback (reporting) data were different. The planning data records the production activities that will be carried out and what should be completed in terms of production objectives and factors of production such as man, machine, material, law, and ring. Arrangements, data between each other, differences, or contradictions may occur, implying potential data quality issues. Plan data between should compliance can be divided into the following three categories:

a) Timing consistency between planned production activities. This consistency can be measured based on the timing differences between planned production activities. Typical examples, such as continuously using the same equipment in the production planning, must have no overlap in time.

b) The consistency of production goals at different levels of production plans. The measure of consistency can be based on the differences between production goals of different levels of production plans. Typical examples are as follows: The production plan goals of the following layers should be in line with the upper production planning goals to match the product type and quantity.

c) The consistency of planned production targets and planned input resources; the measure of this consistency can be based on the difference between production plan targets and planned input resources; typical examples include planned input of raw materials and meters. The type and quantity of the product to be produced should match the corresponding BOM data.

Feedback (reporting) data records the production activities that have been carried out, the realized production results, and the information on production factors such as people, machines, materials, laws, and rings in the production process. Similarly, the consistency that should be followed between these datasets can be divided into the following three categories: consistency of production results at different levels of recording production plans; consistency of recorded production results with recorded input resources.

(3) The consistency of the plan data and the feedback (work report) data. Combining planning data and feedback (job reporting) data, it is possible to calculate KPIs related to production control, such as product delivery rate and planning delay rate, through which enterprises can monitor production and evaluate product performance. An anomaly in a KPI reflects a production anomaly that the business is concerned about. Still, given the underlying data quality issues, the anomaly can also result from the poor quality of the data on which the KPI is calculated. In this regard, based on the KPIs commonly used in production control, this paper proposes the mutual consistency of planning data and feedback (work reporting) data, which can be divided into the following three categories:

a) Consistency of timing between planned production activities and producing activities that have been carried out, and the measure of this consistency can be based on the difference in timing between activities; typical examples include the planning end of a production plan. The interval should be close to the actual end time of the record.

b) The consistency of planned production targets with recorded production results. This consistency can be measured based on the difference between production targets and production results. Typical examples include the actual production volume of a product that should be counted when production volume is close.

c) Consistency of planned and recorded resources invested can be measured based on differences in resources invested. Typical examples, such as the planned use of equipment, should be consistent with the actual use of equipment. Figure 3 shows data quality metrics for production governance.

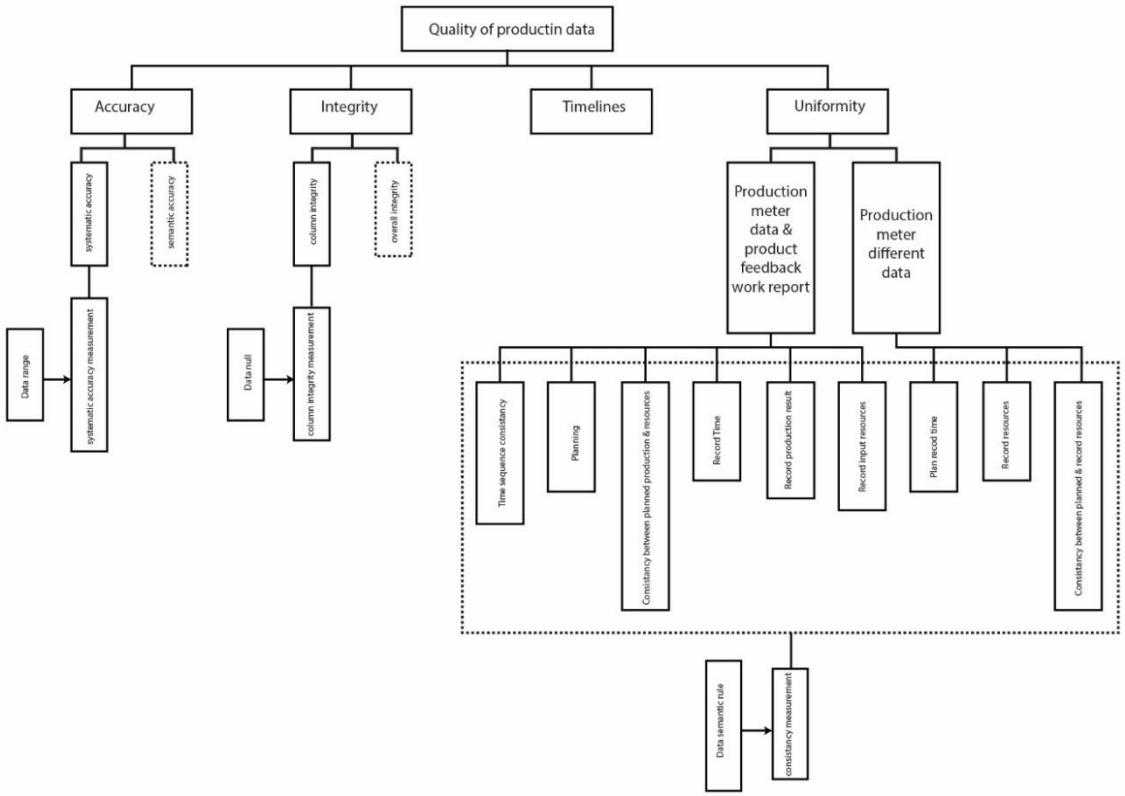

Figure 3: Data Quality Metrics for Production Control

3.2. IMPLEMENT MEASUREMENTS

First, the data quality metrics need to be formally defined to ensure that the meaning of the data quality metrics constructed is clear; second, you need to query through a database based on the data quality metrics you've built. Data quality measurements portray the consistency that data quality managers think data should maintain as a set of data quality rules C; for each rule c $\in$ C, the extent to which the data violates this rule is taken as a measure of data quality on this measure of conformity. Considering that production control data is often numerical data, the data involved in the rules may be cross-record or cross-table, and the use of data often involves aggregation operations, this paper combines negative constraints with aggregate integrity constraints [24-26] to formalize the definition of data quality rules.

A data mass consistency measure is defined by a triplet <c,f ($\bar{v}$), w ($\bar{u}$) > , where c is a rule , shaped like $\forall\bar{x} \neg ((P_1 \wedge ... \wedge P_m) \wedge (Q_1 \wedge .... \wedge (Q_n))$ for any $\bar{x}$, $P_1,....P_m$, $Q_1,...,Q_n$ [23])), this rule states that it cannot be true at the same time. (1) $P_1,...,P_m$ is a database atomic proposition to filter out the data that needs to be consistent, such as R ($A_1 = v_1$, $A_2 = v_2$,...), where R is a relational pattern in the database, A_i can be a constant or a variable in $\bar{x}$. R (A is a property of the relational pattern) True, if and only if the data The library has an instance of R in the Terre inside the tuple t, which is equal to the value on attribute A_1, A_2,…. in variable V_1, V_2…..

1) $Q_1,..., Q_n$ is shaped atomic proposition of $v_1 \varphi v_2$, used for determine whether the filtered data conforms to the rules, where: φ taken from the set {=,≠,<,>,≤,≥}; v_i can be a constant or a variable in $\bar{x}$, an aggregate or an arithmetic expression based on the above triad and the arithmetic symbol {+,-,×, ÷}.
(2) A polygon such as Agg_R(exp:fc_1,…,fc_0), where : Agg Yes taken from {COUNT,SUM,AVERAGE,MIN,MAX} aggregate operator numbers can be R-based an expression or another aggregation in the attribute and arithmetic symbols constructed; $fc_1,...,$ fc_o is shaped like Aφv or A = A′ filter, A is a property in R, v can be a constant or $\bar{x}$ variable, A′ is another property in relational mode R', if and only if an aggregate $Agg_R(\cdot)$ The parameter exp when the second form of filtering condition is in Agg as another aggregate $Agg'_{R'}(\cdot)$ appears in . An aggregation is to convert the data R(·) library. All tuples within an instance of R that meet the filter criteria perform the declared aggregation operation, such as SUM_R (A1+ A2:A3 =3)
(3) Represents all the properties within the instance to R Tuples with a value of 3 on A_3 sum the values of A_1 and A_2. One union can be nested within another, such as $SUM_{R'}$(SUM) represents the sum of all tuples in an instance of R′that have a value of 3 on the attribute B′, and the aggregate operation of these tuples on $SUM_{R'}(SUM_R(B:A = A'):B'=3)$ $sum_R(B:A = v)$, where v is the value of these tuples on the attribute A′.

For Rule c, give the discretion After $\bar{x}$, it can be judged that $P_1,...,P_m$, $Q_1,...,$ Q_n Yes no true if $P_1 \wedge ... \wedge P_m$ True, then $\bar{x}$ is called the pair rule C is triggered once, all are triggered. The quantity is denoted as I, if triggered at the same time $Q_1 \wedge ... \wedge Q_n$ is true, it is called the trigger violates the rules; Otherwise, the trigger is said to conform to the rule. To differentiate between the degree of violation of the rule that is triggered, the function f($\bar{v}$) is introduced $Q_1,...,Q_n$), where $\bar{v}$, consists of the term in. For arbitrary triggers, the function calculates a value on [0,1] based on the value of $\bar{v}$, the larger the value means that the trigger is more consistent with the rule, 1 means that the trigger is fully compliant with the rule, and 0 means that the trigger violates the rule to the maximum extent.

The form f($\bar{v}$) can be customized according to your needs, a special form is f($\bar{v}$) = d(v_1, v_2). In this case v_1 and v_2 some absolute or relative difference between the two terms is calculated and standardized to [0,1] , as may be used $d(v_1,v_2)= 1-|v_1-v_2|/\max\{v_1,v_2\}$ calculates the difference between two positive real numbers. w($\bar{u}$) is used to calculate the weight of each trigger, where $\bar{u} \subseteq \bar{x}$. The shape of w($\bar{u}$). When the consistency measure is defined as < c, f ($\bar{v}$)> means that all triggers have the same weight, one touch The

weight of $w(\bar{u})/\sum_{j=1}^{1} w(\bar{u}_j)$. Custom $f(\bar{v})$ Specific forms need to focus on the sensitivity of consistency measures so that measures can react when data of greater importance violate the rules to a lesser extent, and tolerate heaviness to some extent
To a lesser extent the data is a violation of the rules. Finally by equation (3) calculates the weighted average violation degree of all triggers for the rule as a data quality measurement based on that consistency measure, wherein is the value of $\bar{v}$ under trigger DQ-x_i.

$$DQ(<c,f(\bar{v}),w(\bar{v})>)= \sum_{i=1}^{1} f(\bar{v}_i) w(\bar{u}_i)/\sum_{i=1}^{1} w\,\bar{u}_i) \quad \ldots\ldots\ldots\ldots (3)$$

3.3. Analyze metric results

After you have built a data quality measure and implemented a data quality measurement based on the measurement, you need to analyze the measurement results. The analysis consists of two phases, the first of which is based on the importance of the data and the measurement results to locate the data quality problems that need to be solved first. The judgment of the measurement results can be different from the expectations of the data quality management personnel and the measurement results of different periods. In the second stage, according to process-driven thinking, the processes of relevant data generation, collection, and entry are investigated; the nodes that may affect the data quality in the process are located, and the data input, specific operations, and data output of the system or personnel on the node are analyzed to find the root causes that may cause data quality problems.

4.0. Case Studies

For an auto parts manufacturing company M in Dhaka, Bangladesh, the MES system operation of two workshops was adopted using a data quality assessment and analysis framework for production control sets of data for quality assessment and analysis.

4.1. Data Model

Figure 4 The E-R diagram shows the core part of the system data model. Figure 4 is divided into three parts, and the first part is the entity and relationship related to product manufacturing, including product (Product), product segment (Product Segment), material (Material), equipment (Equipment), and process clearance (Process Bill) and manufacturing list (Manufacturing Bill); The second part is the entity and relationship related to the production plan, including the production needs of the product (Production Requirement and Segment Requirement); The third part is related to production feedback Entities and relationships, including Production Response, Segment Feedback, Workpiece Completion, and Material Consumption. Figure 5 shows the other relationships between the entities of the above sections.

4.2. Data Quality Assessment and Analysis Results

Table 1 shows the quality measurements of key data from two shop floors of Enterprise M, where Ii is a consistency measure < c_i, f_i> triggers, DQ_i its measurement. The consistency measure < c> meaning as follows.

(1) < c_1, f_1>, the data quality rules of this consistency measure as shown in equation (4), which requires that the actual start time of production feedback (Segment Response. Act Start Time) of any product segment should not be later than the corresponding record time (Work piece Completion. Trans Time), data quality measurement. The quantity result is calculated by the function f_1, as shown in equation (5), for either trigger. The rule's data, which maps the difference between the above two times to [0, 1] Above.

$$C_1{:}\forall x, y_1, y_2 \neg\,(SegmentResponse(ID = x, actStartTime = y_1) \wedge WorkpieceCompletion (segmentResponeID = x, transTime = y_2) \wedge y_1 > y_2) \quad \ldots\ldots\ldots\ldots (4)$$

$$f_1 = \begin{cases} 1, & y_1 < y_2 \\ 0, & y_1 - y_2 \geq 1h \\ 1 - \frac{y_1 - y_2}{1h}, & o.w. \end{cases} \quad \ldots\ldots\ldots (5)$$

2) < c_2,f_2> The data quality rules for this consistency measure C_2 As shown in Equation (6), which requires the actual end time of production feedback for any product segment, Segment Response. Act End Time) should not be earlier.

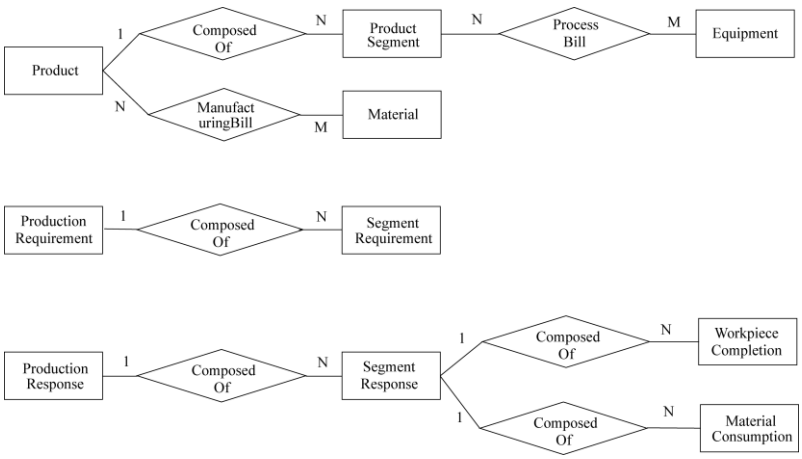

Figure 4: Case Enterprise Data Model

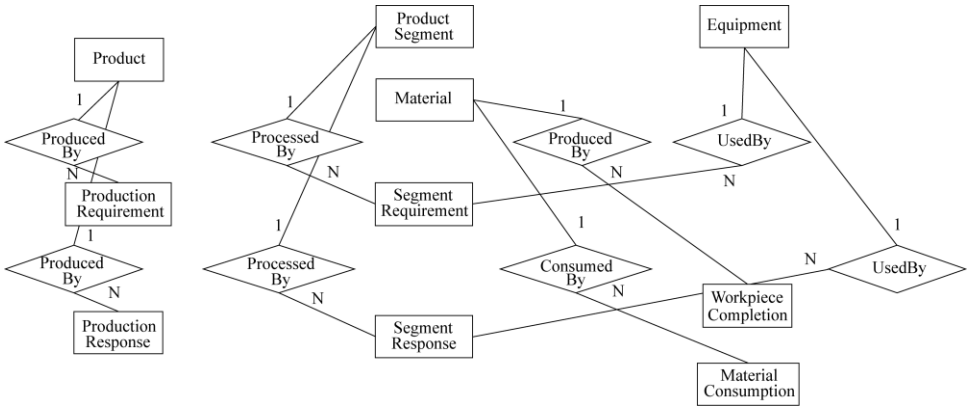

Figure 5: Case Enterprise Data Model

The recording time of any workpiece corresponding to the work report, the measurement of data quality. The quantity result is calculated by the function f_2, as shown in equation (7), for either trigger. The data of the rule, which maps the difference between the above two times to [0,1] above.

$$C_2{:}\forall x, y_1, y_2 \neg (SegmentResponse(ID = x, actEndTime = y_1) \wedge WorkpieceCompletion (segmentResponeID = x, transTime = y_2 \wedge y_1 < y_2) \quad \ldots\ldots\ldots\ldots\ldots (6)$$

$$f_2 = \begin{cases} 1, & y_2 < y_1 \\ 0, & y_2 - y_1 \geq 1h \\ 1 - \frac{y_1 - y_2}{1h}, & o.w. \end{cases} \quad \ldots\ldots\ldots (7)$$

3) <c_3, f_3>, the data quality rule c_3 of the consistency measure c_3 is shown in equation (8), which requires that the start time (ProductionRequirement. startTime) of any product production requirement should not be corresponding to it the beginning of the feedback of the production of all products (ProductionResponse. startTime) is a number of minimum values that do not equal. According to the measurement of the mass by the function f_3 calculation, as shown in equation (9), is correct for either of the triggering rules, which maps the difference between the two times described above [0,1] on.

$$C_3{:}\forall x, y \neg (ProductionRequirement(ID = x, startTime = y) \wedge MIN_{ProductionResponse} (startTime: productionRequiremtnID = x) \neq y) \quad \ldots\ldots\ldots\ldots\ldots (8)$$

$$f_3 = \max\{1 - |y - MIN_{ProductionResponse}(\text{startTime:productionRequirementID=x})|/12h, 0\} \quad \ldots\ldots\ldots (9)$$

4) < c_4, f_4>, the data quality rule c_4 for this consistency measure c_4, as shown in equation (10), which requires an end time for the production requirements of either product (Production Requirement.endTime) should not correspond to it. The maximum value of the end time (ProductionResponse.endTime) with product production feedback is unequal, and the measurement of data quality is calculated by the function f_4, as shown in equation (11), which maps the difference between the above two times to [0,1] above.

$$C_4{:}\forall x, y \neg (ProductionRequirement(ID = x, endTime = y) \wedge MAN_{ProductionResponse} (endTime: productionRequiremtnID = x) \neq y) \quad \ldots\ldots\ldots\ldots\ldots (10)$$

$$f_4 = \max\{1 - |MAX_{ProductionResponse} (\text{endTime:productionRequirementID=x})\text{-}y|/12h, 0\} \quad \ldots\ldots\ldots (11)$$

5) < c_5, f_5>, the data quality rules for this consistency measure c_5 are shown in Equation (12), which requires that the demand for the production needs of any one product (ProductionRequirement.num) should not correspond to all quality inspection results (WorkpieceCompletion. prodCode) as qualified ('QUALIFIED') of the number of work reported by the work (WrokpieceCompletion. qtyVal) sum unequal, data quality. The measurement result of the is calculated by the function f_5, as shown in equation (13), maps the relative difference between the above two quantities to [0,1] for either of the data that triggers the rule above.

$$C_5{:}\forall x, y \neg (ProductionRequirement(ID = x, num = y) \wedge v \neq y) \quad \ldots\ldots\ldots\ldots\ldots (12)$$

$$f_5 = \max\left(1 - \frac{|v-y|}{\max\{|v|,|y|\}}, 0\right) \quad \ldots\ldots\ldots (13)$$

$$V = SUM_{SegmentRequirement}(SUM_{SegmentResponse} (SUM_{WorkpieceCompletion}(qtyVal: segmentResponseID = ID, prodCode = ' QUALIFIED'): segmentRequirementID = ID): productionRequirementID = x)$$

It can be seen that although the two workshops carry out the same production activities and use the same MES system, there are differences in the data quality measurements. In terms of grammatical accuracy, both workshop measurements were made 1.000, that is, the demand for all records in the production demand of the two workshops and the number of declarations in all records in the workpiece reporting meet the positive integer value range constraints. In terms of column integrity, Workshop 2 products produce inverse most records in feed have start time (0.998) and end time (0.998), while some records in Workshop 1 have missing start time (0.818) or end time (0.684). In addition, there are a large number of missing records in the workpiece reports on both

workshops that refer to the production feedback of the product segment (WorkpieceCompletion.segmentResponseID）（0.600，0.150） In terms of internal consistency, relative to other measurements. The amount of results, between cars 2 in < c_2, f_2> the lowest measurement results (0.049), which means that there is a large number of productions in the product segment of Workshop 22. After the confirmation is over, there is still an unreasonable situation of the workpiece being reported. In terms of mutual consistency, Workshop 1 generally measured lower in mutual consistency than In Workshop 2, especially in < c_5, f_5> (0.43) is significantly different from Shop Hall 2 (0.91), which means Shop Hall 155 There is a large number of qualified workpieces reported to the number of workers and product production demand is a large difference.

Based on the importance of the data and the measurement results, the primary concerns are as follows

Data quality issues:

(1) The measurement results of the data of the two workshops on the consistency measurement < c_2,f_2> are quite different;

(2) The measurement results of the data of the two workshops on the consistency measurement < c_5,f_5> are quite different.

Table 1 Partial data quality measurements for case enterprises

Data quality dimension		Data quality measurement rules	Data quality measurement	
			Workshop 1	Workshop 2
Systematical Accuracy		$\frac{\lvert\{t \in ProductionRequirement \lvert t.num \in N^+\}\rvert}{\lvert PoductionRequirement\rvert}$	1.000	1.000
		$\frac{\lvert\{t \in WorkingCompletion \lvert t.qtyVal \in N^+\}\rvert}{\lvert WorkingCompletion\rvert}$	1.000	1.000
Column Integrity		$\frac{\lvert\{t \in ProductionResponse \lvert null(t.startTime) = 0\}\rvert}{\lvert PoductionResponse\rvert}$	0.818	0.998
		$\frac{\lvert\{t \in ProductionResponse \lvert null(t.endTime) = 0\}\rvert}{\lvert PoductionResponse\rvert}$	0.684	0.998
		$\frac{\lvert\{t \in WorkpieceCompletion \lvert null(t.segmentResponseID) = 0\}\rvert}{\lvert WorkpieceCompletion\rvert}$	0.600	0.150
Uniformity	Internal Consistency	<c_1,f_1>	I_1=2418 057 DQ_1=0.996	I_1=18 443 DQ_1=0.999
		<c_2,f_2>	I_2=2418 057 DQ_2=0.853	I_2=18 443 DQ_2=0.049
	Maturity Consistent different	<c_3,f_3>	I_3=4 056 DQ_3=0.742	I_3=10 353 DQ_3=0.994
		<c_4,f_4>	I_4=3 398 DQ_4=0.685	I_4=10 349 DQ_4=0.839
		<c_5,f_5>	I_5=5 990 DQ_5=0.43	I_5=10 740 DQ_5=0.91

Data quality issues (1) and (2) are both related to the workpiece reporting in the workshop, and Figures 6 and 7 show the different reporting processes in the two workshops.

On Shop Floor 1, employees enter information and print part barcodes on the barcode printer based on the part model and machine model of the downline. Whenever a part is off the line, the employee passes the bar code on the part and scans it with the terminal equipment. The system automatically recognizes the scanned bar code information to generate a work report record. In Workshop 2, after the completion of the plan, the employee collects the production data from the machine control panel and records it on the paper production tracking table, which is submitted to the production management personnel, and after verification, the production management personnel enter the data on the table into the system to form a work report record.

Further analysis of the above processes can identify the root causes of data quality issues (1) and (2). In workshop 1, whenever the parts go offline, the employee enters the report record by scanning the bar code, and in workshop 2, the report record is uniformly collected and reported to the production management personnel by the employee after the production is completed, and then checked by the latter to enter the system, which leads to a large delay in the report, so that the measurement results of the two workshop data on the consistency measurement < c_2, f_2 > are quite different. In addition, in Workshop 2, the data recorded in the work report comes from the machine's data acquisition system, and the 22 data is checked by the production management before entering the system; In Workshop 1, the printing, pasting and scanning of barcodes is done manually by employees.

Employee errors can lead to missing barcodes. At the same time, there is a certain limitation on the mechanism for the system to generate work report records based on the scanned barcode information, which may lead to the foreign key Work Completion of the generated work report records. Segment Response ID is missing, which in turn causes the reporting record to be uncorrelated with the product production demand record, resulting in a large difference in the measurement results of the two workshop data on the consistency measurement < c_5, f_5>.

.

4.3. Measurement Sensitivity Discussion

In Section 3.2, the definition of the specific forms f ($\bar{v}$) and w ($\bar{u}$) in the consistency measure requires heavy consideration of sensitivity.

This section discusses the effect of metric sensitivity on measurement results based on the consistency measures < c_5, f_5> in Section 4.2. Constructing a new degree of consistency Measure < c_5, f_5' , w_5>, where c_5 is shown in equation (12), f_5 ' is shown in equation (14), f_5 is a special case of f'_5 at α = 1, w as shown in equation (15):

$$f_5' = \max\left(1 - \frac{|v-y|}{\max\{|v|,|y|\}}, 0\right) \text{..... (14)}$$

$$w_5 = y \ldots\ldots (15)$$

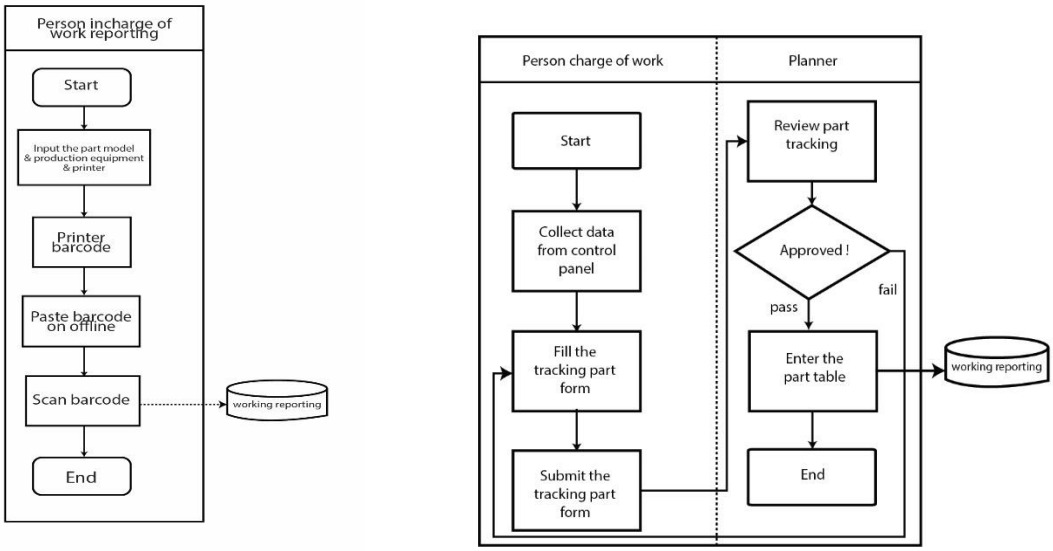

Figure 6: Workshop 1(left), Workshop 2(right) Workpiece Reporting Process

The new consistency measure pays more attention to the production needs of the more demanding products and the data quality of the corresponding workpiece reports, and setting a larger (small) $\alpha \in R^+$ will make the measurement more (cannot) tolerate the violation of rule c by the data. Table 2 shows the weighted consistency measures $< c_5, f'_5, w_5 >$ and weightless consistency measures $< c_5, f'_5 >$ on The Shop 1 and Shop 2 data under different α values. It can be seen that although the data quality of Shop 2 is higher than that of Shop 1 under all parameters, the relative difference between the two decreases with the increase of α. In addition, for Shop Floor 1, the measured results with weights under the same α are higher than the metric measurements without weights, which to some extent indicates that the data quality problem is relatively concentrated on the production needs of products with small demand and their corresponding workpiece reports.

5.0. SUMMARY

To solve the problem of data quality management in the scenario of production planning and control, this paper proposes a data quality evaluation and analysis framework for production control, which is divided into four steps: building a data model, selecting data and measurements, implementing measurements, and analyzing measurement results. Finally, this paper applies the proposed framework to the quality assessment and analysis of the production control data of an auto parts manufacturing company in Dhaka, Bangladesh. The results show that the framework helps IT and production managers locate key data quality issues. And from the possible operational errors of personnel, unreasonable data creation and update mechanisms of the system, etc., we can find the root causes of data quality problems.

Table 2 Data quality measurements for each parameter

α	Workshop 1		Workshop 2	
	weightless	weighted	weightless	weighted
0.01	0.02	0.02	0.16	0.13
0.1	0.11	0.12	0.40	0.40
0.3	0.24	0.28	0.68	0.71
0.5	0.32	0.38	0.81	0.83
0.7	0.38	0.44	0.87	0.89
0.9	0.41	0.49	0.90	0.92
1.0	0.43	0.51	0.91	0.93
1.1	0.44	0.51	0.92	0.94
1.3	0.46	0.55	0.93	0.95
1.5	0.48	0.58	0.94	0.96
1.7	0.49	0.59	0.94	0.96
1.9	0.50	0.60	0.95	0.96
2.0	0.51	0.61	0.95	0.96

ACKNOWLEDGMENT

This research was not financially funded.

CONFLICT OF INTEREST

The authors declared that there is no conflict of interest for this research.

REFERENCES

[1] GÜNTHER L C,COLANGELO E,WIENDAHL H H,et al. Data quality assessment for improved decision-making:a methodology for small and medium-sized enterprises[J]. Procedia Manufacturing,2019,29(3):583-591.
[2] SCHUH G,REUTER C,PROTE J P,et al. Increasing data integrity for improving decision making in production planning and control[J]. CIRP Annals,2017,66(1):425-428.
[3] WANG R Y. A product perspective on total data quality management[J]. Commun ACM,1998,41(2):58-65.
[4] WANG R Y,STRONG D M. Beyond accuracy:What data quality means to data consumers[J]. Journal of Management Information Systems,1996,12(4):5-33.
[5] PIPINO L L,LEE Y W,WANG R Y. Data quality assessment[J]. Commun ACM,2002,45(4):211-218.
[6] BATINI C, CAPPIELLO C, FRANCALANCI C, et al. Methodologies for data quality assessment and improvement[J]. ACM Computing Surveys,2009,41(3):1-52.
[7] ENGLISH L P. Improving data warehouse and business information quality:methods for reducing costs and increasing profits[M]. Hoboken:John Wiley & Sons,Inc.,1999:199-235.
[8] LOSHIN D. Enterprise knowledge management:The data quality approach[M]. San Francisco:Morgan Kaufmann,2001:73-99.
[9] LEE Y W,STRONG D M,KAHN B K,et al. AIMQ:a methodology for information quality assessment[J]. Information & Management,2002,40(2):133-146.
[10] SEBASTIAN-COLEMAN L. Measuring data quality for ongoing improvement:a data quality assessment framework [M]. San Francisco:Morgan Kaufmann,2013:182-245.
[11] VAZIRI R,MOHSENZADEH M,HABIBI J. TBDQ:A pragmatic task-based method to data quality assessment and improvement[J]. PLOS ONE,2017,11(5):e0154508.
[12] HAUG A,ARLBJØRN J S,PEDERSEN A. A classification model of ERP system data quality[J]. Industrial Management & Data Systems,2009,109(8):1053-1068.
[13] CAO L,ZHU H. Normal accidents:Data quality problems in ERPenabled manufacturing[J]. Journal of Data and Information Quality (JDIQ),2013,4(3):1-26.
[14] SCHUH G,THOMAS C,HAUPTVOGEL A,et al. Achieving higher scheduling accuracy in production control by implementing integrity rules for production feedback data[J]. Procedia CIRP, 2014,19(7):142-147.
[15] ZONG W,WU F,FENG P P. Improving data quality during ERP implementation based on information product map[J]. Enterprise Information Systems,2019,13(9):1275-1291.
[16] WOODALL P,KORONIOS A,GAO J,et al. An investigation into data quality root cause analysis[C]//Proceedings of ICIQ 2012:17th International Conference on Information Quality. Cambridge:MIT Press,2012:193-205.
[17] Harel Z, Silver SA, McQuillan RF, Weizman AV, Thomas A, Chertow GM, Nesrallah G, Chan CT, Bell CM: How to diagnose solutions to a quality of care problem. Clin J Am Soc Nephrol 11: 901–907, 2016.
[18] Langley GL, Nolan KM, Nolan TW, Norman CL, Provost LP: The Improvement Guide: A Practical Approach to Enhancing.
[19] Perla RJ, Provost LP, Murray SK: The run chart: a simple analytical tool for learning from variation in healthcare processes. BMJ Qual Saf 20: 46–51, 2011.
[20] Olmstead Pl. Distribution of sample arrangements for runs up and down. Ann Math Stat 1945;17:24e33.