Available online at www.sciencedirect.com
ScienceDirect
Procedia CIRP 19 (2014) 142 – 147
www.elsevier.com/locate/procedia

Robust Manufacturing Conference (RoMaC 2014)

Achieving higher scheduling accuracy in production control by implementing integrity rules for production feedback data

Günther Schuh[a], Christina Thomas[a], Annika Hauptvogel[a], Felix Brambring[a]*

[a]*Laboratory for Machine Tools and Production Engineering (WZL) of RWTH Aachen University, Steinbachstr. 19, 52074 Aachen, Germany*

* Corresponding author. Tel.: +49-241-80-28241; fax: +49-241-80-22293. *E-mail address:* F.Brambring@wzl.rwth-aachen.de

Abstract

Excellent production planning and control (PPC) processes are a prerequisite for accomplishing a high adherence to promised delivery dates. Despite enormous efforts which are put into achieving high scheduling accuracy, manufacturing companies still regularly struggle in meeting their logistic targets. In consequence, these companies deal with high stocks, long lead times and ultimately only achieve a bad adherence to promised delivery dates. An important reason for this discrepancy are data inconsistencies, which occur in data collected on the shop floor, because these data are used to update the near- and middle-term scheduling of current production jobs. In this paper, the impact of several data inconsistencies in real-world production feedback data sets are investigated. Integrity rules for selected data inconsistencies are proposed and tested for their effects on a number of logistic targets in a simulation study.

Selection and peer-review under responsibility of the International Scientific Committee of "RoMaC 2014" in the person of the Conference Chair Prof. Dr.-Ing. Katja Windt.

Keywords: Robust production planning and control; Data inconsistency; Scheduling accuracy

1. Introduction

Companies of the manufacturing industry, especially in high-wage countries like Germany, are facing multiple interacting challenges: in response to increasing competition by companies mainly located in emerging countries and steadily rising customers' demands regarding highly customized product features and short delivery times, manufacturing companies need to produce a high number of individualized product variations with complex material flows while maintaining a high adherence to promised delivery dates [1].

These challenges call for excellent production planning and control (PPC) processes which are regularly supported by Advanced Planning and Scheduling (APS) systems for the near-term and middle-term scheduling of production jobs. For this detailed scheduling, APS-systems incorporate master data which can be retrieved from Enterprise Resource Planning (ERP) systems as well as a variety of high-resolution production feedback data concerning manufacturing and assembly processes which are gathered on the shop floor by so-called Production Data Acquisition (PDA) systems [2]. Additionally, most manufacturing companies still rely on manual reporting by employees for process steps where installing the necessary IT infrastructure is not economically reasonable [3]. This can either be realized via terminals, which are installed specifically for this purpose, or the data can still even be recorded in handwriting.

Obviously, collecting data through multiple sources is prone to inaccuracies and inconsistencies. Data inconsistencies in production feedback data have various root causes such as inaccurate or completely missing manual feedback due to human error, inconsistent aggregation of the data in databases, machine-tool-integrated sensors which fail dirty or even unplanned interferences in the material flow. However, the negative impact of poor production feedback data quality on the scheduling accuracy in production control is often neglected. Consequently, the quality of these production feedback data is regularly impaired by various data inconsistencies which neg-

Selection and peer-review under responsibility of the International Scientific Committee of "RoMaC 2014" in the person of the Conference Chair Prof. Dr.-Ing. Katja Windt.
doi:10.1016/j.procir.2014.05.003

atively effect detailed scheduling results and lead to high inventory, long lead times and ultimately to a bad adherence to promised delivery dates.

The approach presented in this paper does not aim at healing the mentioned root causes for the occurrences of data inconsistencies. This procedure might seem counter-intuitive at first, but even though the reasons for data inconsistencies in production control have been known for decades, they still exist and might become even more relevant in future. In order to realize cyber-physical production systems with self-controlling production jobs which will be produced without a centralized planning entity, even more data will have to be collected by additional sensors. Where events are tracked simultaneously by multiple sensors, the risk for occurring data inconsistencies increases. Therefore, this new approach of making these data less error-prone by taking the inconsistencies into account is deemed much more resilient against outer influences or changed circumstances.

In this paper, typical data inconsistencies in production feedback data are analyzed for their rate of occurrence in four real-world data sets of German manufacturing companies and evaluated regarding their negative impact on scheduling accuracy. Customized integrity rules build the basis for developing algorithms in order to clean the data from the detected data inconsistencies. The specific algorithms lean on methodologies from various fields such as market research and statistics. In conclusion, a process of validation is proposed and conducted exemplarily for one of the data sets.

First simulation runs have shown that the derived integrity rules provide promising results and can considerably improve near- and middle-term scheduling accuracy in production control.

2. State of the art

In general, missing or inconsistent data lead to several problems concerning their analysis. The most important impediment is the reduction of the sample size which decreases the validity of any statistical data analysis. Additionally, when considering practical applications, complete and consistent data sets are a prerequisite for computational operations since inconsistent data might lead to faulty assumptions about their inherent information or might even cause runtime errors of utilized software [4]. Therefore, the negative implications of missing or inconsistent data in databases and ways of dealing with them have been widely discussed by researchers and practicians of different fields: Wilkins proposes an approach for updating incomplete information for missing values and values which are unknown itself but lie in a known domain with a specific certainty [5, 6]. Morrissey presents two methods of estimating the uncertainty which is introduced by missing or imprecisely known data [7]. Owei introduces methodologies for handling and resolving issues concerning data which is uncertain, incomplete, imprecise, vague or inconsistent [8]. Packowski conducted a study on the negative implications of low master data quality and how companies can manage the arising problems strategically [9].

The influence of the master data quality on different planning activities in manufacturing companies is a topic in which the majority of manufacturing companies are already investing considerable effort [9]. However, little attention has been paid to the quality of production feedback data which is by far less static than master data and therefore more complicated to monitor in terms of integrity and quality.

Obviously, simply deleting data sets with missing or inconsistent data would be the easiest approach but can also lead to the afore-mentioned negative implications. Additionally, if the inconsistencies do not occur at random, there is a considerable risk of introducing a serious bias into the data. Therefore, instead of simply excluding data sets with missing or inconsistent data from further analysis by deleting the affected data sets, the goal is to replace or substitute the missing or inconsistent data by imputation [10].

For this end, assumptions about the true values of missing or inconsistent data have to be derived from historical data. A vast number of research exist on improving data quality from simple data analysis to more sophisticated Data Mining (DM) techniques which find extensive application in banking, insurance, marketing and related fields [11]. In the field of production engineering, DM methodologies have most commonly been applied in the context of parameter optimization for production processes, product quality improvement and engineering design [12,13,14]. However, there has only been scarce research on possibilities to improve data quality of production feedback data by imputing missing or inconsistent values. A DM approach has been developed by Kwak and Kim in order to optimize semi-conductor processes with the explicit consideration of missing data [15]. Wang and Wang have proposed a framework for handling missing data without explicitly considering production feedback data [16].

In this paper, integrity rules for several typical data inconsistencies in production feedback data are presented which fulfil two asks: First of all, these rules can be used to check if a specific inconsistency exists in the data by evaluating in how many data sets the corresponding integrity rule has been broken. Secondly, the integrity rules build the basis for deriving algorithms for imputing the missing or inconsistent data.

3. Quality of production feedback data

3.1. Typical data inconsistencies in production feedback data

In this paper, the focus lies on production feedback data which usually comprise information concerning the current statuses of all active production jobs such as start- and end-dates, utilized work stations as well as set-up and processing durations for all completed process steps. This information is used by APS-systems for updating the near- and middle-term scheduling [17,18]. However, these data is often corrupted by a number of typical data inconsistencies which lower the quality of the data.

Table 1 gives an overview of seven typical data inconsistencies and their specific rate of occurrence in four real-world production feedback data sets which are analyzed in this paper. The data inconsistencies are clustered by their relations to work stations and time stamps: data inconsistency (a) deals with missing information concerning utilized work stations while data inconsistencies (b) to (g) represent different cases of overlapping timespans of process steps or missing information concerning the start- and end-dates of these process steps. A more detailed explanation of all described data inconsistencies can be found in section four and in the appendix.

The production feedback data sets were provided by four German mid-sized mechanical engineering companies which produce highly customized products in a job-shop production environment with manual as well as automatic gathering of production feedback data. For each company, the investigated data sets represent a timespan between 11 and 24 months, covering the years 2010 to 2013, with information concerning between 13.000 and 64.000 production jobs. Accordingly, the data covers between 82.000 and 273.000 process steps. Additionally, the data sets include between 53 and 253 work stations which have been utilized in the respective production environments during the given time spans.

Table 1. Rate of occurrence for typical data inconsistencies.

Cluster	Data inconsistency	Co. 1 [%]	Co. 2 [%]	Co. 3 [%]	Co. 4 [%]	Average rate of occurrence [%]
Work stations	(a)	0.25	0.00	6.10	0.01	1.59
Time stamps	(b)	6.46	5.85	0.78	0.16	3.31
	(c)	0.74	2.88	0.20	0.33	1.04
	(d)	0.32	0.37	0.02	0.07	0.2
	(e)	0.36	0.53	0.02	4.20	1.28
	(f)	0.26	0.06	0.02	0.07	0.1
	(g)	6.91	14.19	5.69	3.12	7.48

When taking these statistics into consideration, the relatively low percentage values in Table 1 still represent a considerable number of observations of the specific data inconsistencies. Except for data inconsistency (a) in the second data set, all described inconsistencies have been present in the analyzed data sets. Especially, data inconsistency (g) which stands for missing start- and end-dates has been detected extensively.

In conclusion, the results presented in Table 1 show that data inconsistencies in production feedback data constitute a problem for detailed scheduling accuracy since their existence prevents a realistic calculation of completion dates for current production jobs.

3.2. Severeness of data inconsistencies in production feedback data

The data inconsistencies shown in Table 1 can be distinguished by their differing severeness in regard to their negative effects on simulation run results. The severeness of a data inconsistency is defined as a function of two parameters: its average rate of occurrence and its impact on the simulation results.

Regarding the rate of occurrence in the four analyzed data sets, the data inconsistencies can be separated in three groups: The highest rates of occurrence have been observed for data inconsistency (g), representing missing start- and end-dates, and (b), meaning that the start of the following process step was reported before the completion of the preceding process step. The second group consists of data inconsistencies (a), as well as data inconsistencies (c) and (e), which represent different cases of overlapping time stamps of two consecutive process steps. Finally, on average data inconsistencies (d) and (f) have the lowest rates of occurrence within the four investigated data sets and therefore can be considered the least lethal in this dimension of measuring the severeness.

The impact on the simulation results can be estimated by qualitatively deriving the influence of the various data inconsistencies on key figures in production control. These key figures are work-in-progress (WIP), utilization, lead time and adherence to promised delivery dates.

Data inconsistency (a), which denotes missing information concerning the utilized work station, has only a limited effect on the WIP unless the affected work station was the first or last one in the production sequence. However, the other three key figures can be seriously influenced by missing feedback concerning the work station since the production duration of the process step in question cannot be accounted for. This leads to a decrease in utilization and lead time and will also affect the adherence to promised delivery dates compared to the actual situation on the shop floor.

The impact of all other data inconsistencies can be considered rather similar since they all belong to the same cluster: Unless the first or last process steps are affected, the overlapping or missing time stamps will not have a high effect on the WIP. However, due to the overlapping time stamps, utilization of work stations and lead times of the respective production jobs will seem to be smaller than they actually are. The highest impact can be expected for the adherence to promised delivery dates, since in this key figure the initially planned completion dates are compared to the actual completion dates which might not have been reported at all.

4. Implementing integrity rules in production control

In database management, integrity rules denote logical expressions which are used in order to assure the soundness of data. If an integrity rule is breached, a data inconsistency has been detected [19,20]. Furthermore, integrity rules build the basis for deriving algorithms which can be used for imputing missing values or solving inconsistencies within the data. For production control purposes, we propose that scheduling accuracy can be increased through the implementation of integrity rules for typical data inconsistencies and the application of algorithms which are specifically designed to clean the data from these inconsistencies.

For data inconsistency (a), where information concerning the utilized work station is missing in the data, a well-known approach from market basket analysis called association rule induction has been adapted to fit the requirements of this production control application and has been described extensively in a previous paper [21]. Therefore, in this paper the focus lies on data inconsistencies (b) to (g).

Data inconsistency (b) is a case of overlapping time stamps of consecutive process steps which makes it seem as if the following process step has been started before the preceding process step has even been finished, which cannot be the case in a production environment with fixed lot sizes. Therefore, it is assumed that the beginning of the following process step is actually correct and the preceding process step must have been finished accordingly. By taking into account an average transport and waiting duration of the material between two production steps, the actual finishing time of the preceding process step can be estimated and will be used to patch data inconsistency (b). The average transport and waiting duration can be retrieved from the available data not only company-specific but even for each work station individually by taking the relevant time stamps within the data sets into account.

Data inconsistencies (c) to (f) have in common that either the start time stamp of the following process step occurred before the start time stamp of the preceding process step or that the same is true analogically for the end time stamps of the two process steps. For data inconsistencies (d) and (e) both is even true at the same time. Therefore, one algorithm can be proposed in order to clean all these inconsistencies. For all four inconsistencies, these configurations of time stamps raise the question if the sequence of the two affected process steps has even been reported correctly. So the first step in all these cases has to be a determination if the proposed sequence makes technologically sense, e.g. a milling process step should occur before surface-hardening. If this review comes to the conclusion that the process steps should in fact be switched, data inconsistency (b) resembles (f) and vice versa, (d) becomes (b) and can be treated accordingly while (e) does not even pose an inconsistency any more. In case, the sequence can be retained, the end time stamp of the preceding process step will become the new proposed start time stamp of the following process step while adding the average transport and waiting duration on top. Simultaneously, the end time stamp of the following process step will be adapted in order to reflect the processing duration of this process step. This approach can lead to a newly occurring data inconsistency (b) which should then be treated accordingly.

For data inconsistency (g) no data is given either for the start time nor the end time of a process step. In this case, the surrounding process steps constitute the anchor which can be used to estimate the start and end time stamps by taking into account the transportation and waiting duration as well as the set-up and processing time for the affected process step. In case the affected process step represent the first or the last process step of the production job, only one adjacent process step can be used to estimate the time stamps. However, this reduces the possibility to inflict another inconsistency by adjusting the time stamps.

5. Validating the approach

5.1. Validation process

For proving the applicability of the proposed approach, a data set of perfect integrity is needed in order to act as a basis of comparison. Unfortunately, due to the described conditions on the shop floor and the data acquisition process, such a data set does not exist. Therefore, a validation process is proposed in Figure 1, where a basis data set is produced from scratch by cleaning a real-life production feedback data set and making this new basis data the starting point of the validation. By following this process, it is possible to evaluate if the implementation of integrity rules leads to better results in terms of the approximation to the unknown true information than cleaning the data by simply deleting faulty data.

As described above, the raw data is gathered on the shop floor through different acquisition processes and afflicted with inconsistencies due to several system-inherent shortcomings. One of the most frequently used data cleaning methods is to simply discard of all data sets which are incomplete or inconsistent and to keep only the data with full integrity. This way the new basis data is created with some considerable manual effort. By analyzing the raw data regarding the relative rates of occurrence of the different data inconsistencies beforehand, the basis data can be purposely manipulated in order to feature the same inconsistencies with the respective rates of occurrence. This way, the new raw data is generated which possess the same qualities and characteristics as the original raw data with the important difference that a perfectly consistent basis data exists now for this new raw data set.

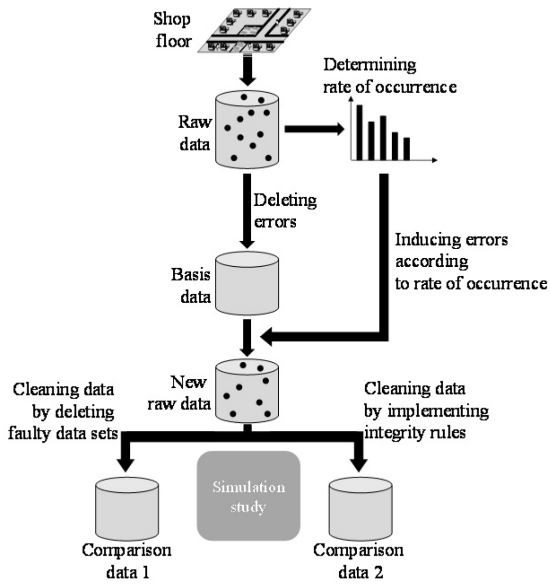


Fig. 1. Validation process

By assuring the integrity rules are not breached by utilizing the derived algorithms onto the newly generated raw data set, the inconsistencies can be remedied. In order to validate the applicability of the integrity rules, a simulation study is conducted through which the concurrence between the basis data and newly generated comparison data can be evaluated. Since the integrity rules will never retrieve the real data with perfect accuracy, a bias is introduced but knowing its level does not yet allow an assessment concerning the achievement of a higher scheduling accuracy. Therefore, instead of directly comparing both data sets line by line, this indirect approach of comparing simulation results has been chosen because of a higher conclusiveness regarding the ultimate goal to achieve a higher scheduling accuracy in production control. By additionally including the general approach of simply deleting the faulty data into the simulation study, the advantages of the integrity rules can be estimated.

5.2. Validation results

The validation process described in the previous paragraph has been conducted exemplarily for the data set of company 3.

As shown in Table 1, the data set of company 3 has been primarily affected by data inconsistency (a) and (g), although all other data inconsistencies have been found as well. The original data set has been prepared according to the process described in the previous section which resulted in three data sets with complete consistency called basis data, comparison data 1 and comparison data 2. All three data sets have then been used as the basis for a simulation run with the same simulation model, representing the circumstances on the shop floor of company 3. The results of these three simulation runs are shown in Figure 2.

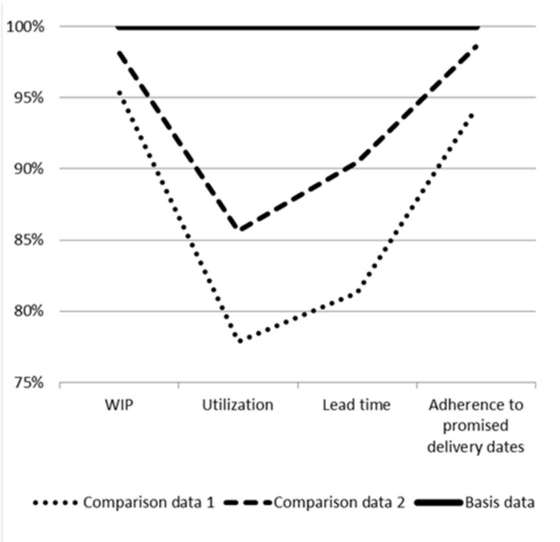


Fig. 2. Validation results

The results are measured in the four key figures described in section 3.2 and are computed relatively to the results of the basis data in order to make the deviations between comparison data 1 and comparison data 2 easily accessible.

The results of this simulation study can be deemed promising: comparison data 2 dominates comparison data 1 in respect of depicting “reality” since the results of comparison data 2 in all four key figures are closer to the results obtained with the basis data. As expected, the work-in-progress has not been dramatically affected by the observed data inconsistencies but still has improved slightly due to the application of the integrity rules.

Equally promising is the increased concordance in the adherence to promised delivery dates, since improving this key figure is usually among the most important logistic targets of mechanical engineering companies.

The highest deviations for both comparison data sets can be observed in respect to utilization and lead time. Although, the application of integrity rules did not achieve a such high concordance with the basis data as in the previously discussed key figures, there has been a considerable approximation to the results of the basis data simulation run. In conclusion, the application of integrity rules has lead to an approximation to the basis data in all four key figures.

6. Conclusions and further research

In this paper, typical data inconsistencies in production feedback data have been assessed towards their severeness regarding simulations in production control. Afterwards, algorithms have been derived based on integrity rules in order to clean affected data sets from any data inconsistencies which have been investigated in this paper. The overall goal is to improve detailed scheduling accuracy by increasing data integrity in production feedback data. First results of the validation look promising.

However, further research still needs to be undertaken in several directions: First of all, the assessment of the severeness has to be further examined and taken from the current qualitative evaluation to a more quantitative one in order to focus future research on the most important data inconsistencies. Moreover, there are a vast number of additional possible algorithms to be derived based on the integrity rules which might deliver even better results in terms of recovering the original data. Hence, future research will also take alternative algorithms into consideration.

Acknowledgements

This new approach of improving scheduling accuracy by reducing data inconsistencies in production control is being investigated by the Laboratory of Machine Tools and Production Engineering (WZL) within the publicly funded research and development project: "Cluster of Excellence - Integrative Production Technology for High Wage Countries” (German Research Foundation, DFG).

Appendix A. Data inconsistencies in production feedback data

The appendix gives an overview of the data inconsistencies investigated in this paper.

A.1. Inconsistencies associated with work stations

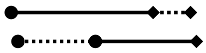
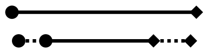
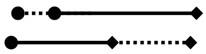
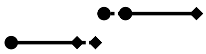
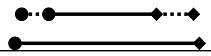
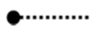
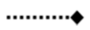
In this paper, only data inconsistency (a) is associated with work stations:

(a) No feedback about utilized work station: occurs when no information is given concerning the work station which has been utilized for a reported process step.

A.2. Inconsistencies associated with time stamps

Data inconsistencies (b) to (g) are associated with start and end time stamps of process steps. Data inconsistencies (b) to (f) represent different cases of overlapping time stamps which can occur for consecutive process steps as shown in Table A:

Table A. Illustrations of data inconsistencies (b) to (f)

Data incon-sistency	Inconsistent time stamps	Illustration
(b)	Case #1	
(c)	Case #2	
(d)	Case #3	
(e)	Case #4	
(f)	Case #5	
Key to symbols	Explanation	
	Timespan covered by process step	
	Possible beginning of process step in order to still be covered by respective case	
	Possible ending of process step in order to still be covered by respective case	

(g) Process step with missing start and end date: occurs when neither start nor end date of a reported process step are given.

References

[1] Wiendahl HP. Betriebsorganisation für Ingenieure (Enterprise organisation for engineers). Munich:Hanser; 2008.
[2] Buzacott JA, Corsten H, Gössinger R, Schneider, HM. Production Planning and Control: Basics and Concepts. Munich: Oldenbourg; 2013.
[3] Schuh G, Stich V (eds.). Produktion am Standort Deutschland (Production in Germany). Survey. Aachen; 2013.
[4] Date CJ. An Introduction to Database Systems. 7th ed., Reading, Mass.: Addison-Wesley; 2000.
[5] Wilkins MW. A model-theoretic approach to updating logical databases. In: Proceedings of the fifth ACM SIGACT-SIGMOD symposium on Principles of database systems. New York:ACM; 1986. p.224-234.
[6] Winslett M. A model-based approach to updating databases with incomplete information. Transaction on Databse Systems 1988; 13(2): 167-196.
[7] Morrissey JM. Imprecise information and uncertainty in information systems. ACM Transactions on Information Systems 1990; 8(2): 159-180.
[8] Owei V. An intelligent approach to handling imperfect information in concept-based natural language queries. ACM Transaction on Information Systems 2002; 20(3): 291-328.
[9] Packowski J (ed.). Strategisches Management von Stammdatenqualität (Strategic management of master data quality). Mannheim: Camelot; 2013.
[10] Schneidermann ED, Kowalski CJ, Willis SM. Regression Imputation of Missing Values in Longitudinal Data Sets. International Journal of Bio-Medical Computing 1993; 32: 121-133.
[11] Bramer M. Principles of Data Mining. 2nd ed., London: Springer;2013.
[12] Harding JA, Shahbaz M, Srinivas, Kusiak A. Data Mining in Manufacturing – A Review. Journal of Manufacturing Science and Engineering 2006; 128(4): 969-976.
[13] Choudhary AK, Harding JA, Tiwari MK. Data mining in manufacturing – A Review. Journal of Intelligent Manufacturing 2008; 20: 501-521.
[14] Wang K, Tong S, Eynard B, Roucoules L, Matta N. Review on Application of Data Mining in Product Design and Manufacturing. Fourth International Conference on Fuzzy Systems and Knowledge Discovery. Haikou; 2007. p. 613-618.
[15] Kwak DS, Kim KJ. A data mining approach considering missing values for the optimization of semiconductor-manufacturing processes. Expert Systems with Applications 2012; 29: 2590-2596.
[16] Wang H, Wang S. Discovering patterns of missing data in survey databases – An application of rough sets. Expert Systems with Applications 2009; 36: 6256-6260.
[17] Schuh G, Gottschalk S, Höhne T. High Resolution Production Management. CIRP Annals – Manufacturing Technology 2007; 56(1): 439-442.
[18] Perzyk M, Biernacki R, Kozlowski J. Data mining in manufacturing: significance analysis of process parameters. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture 2008; 222: 1503-1516.
[19] Imielinski T, Lipski Jr W. Incomplete Information in Relational Databases. Journal of the ACM 1984; 31(4): 761-791.
[20] Wilks J. Querying and updating constraint databases with incomplete information. Proceedings of the First East-European conference on Advances in Databases and Information systems. Swinton: BCS; 1997. p. 40-40.
[21] Schuh G, Potente T, Thomas C, Brambring F. Approach for Reducing Data Inconsistencies in Production Control. Zäh M (ed.) Proceedings of the Fifth International Conference on Changeable, Agile, Reconfigurable and Virtual Production. Munich:Springer; 2013. p. 347-351.