THE EFFECT OF DATA QUALITY ON DATA MINING – IMPROVING PREDICTION ACCURACY BY GENERIC DATA CLEANSING

(Practice-Oriented-Paper)

Jørgen Stang
Det Norske Veritas, Norway
Jorgen.Stang@dnv.com

Tore Hartvigsen
Det Norske Veritas, Norway
Tore.Hartvigsen@dnv.com

Joachim Reitan
FFI, Norway
Joachim.Reitan@ffi.no

Abstract: The recent advent and abundance of mainstream tools for performing business intelligence such as data mining and predictive analysis have made applications previously limited to highly trained specialists accessible to a wide range of users. At the same time, the amount of available data is rapidly increasing due to several factors such as government regulations, affordable hardware, mature master data and data warehouse sources and available 3rd party data for areas like demographics and financial analysis. As a consequence, advanced data mining techniques for predictive analysis find increasingly novel uses ranging from detecting mechanical failures in propeller shafts to forecasting customer retention. The usefulness of the data mining models will be highly dependent on the quality of the underlying data source and misrepresented data could distort the analysis and produce erroneous predictions which at best will be useless and at worst jeopardize operations. The data flaws could be hard to detect by manual inspections of the data mining results as these are frequently non intuitive. Therefore, the data quality should be carefully evaluated prior to implementing any data mining models and any findings should either be cleaned or used as an indicator for the confidence level of the resulting data mining model. This will in turn establish the effect of the data quality on the data mining as described in [7]. Also, the reverse approach have been suggested, where data mining is used as a successful means to detect data quality indicators [11] [16]. This article describes a practical case which attempt to assess the effect of improving the data quality on the prediction accuracy for a work order database.

Key Words: Data Quality, Data Mining, Work Orders, Analysis Accuracy.

INTRODUCTION

This article demonstrates the effect of data quality on data mining for a work order data system. The work order data describes tasks commissioned by an organization to perform a particular job (repair, maintenance, periodic service, calibration) on a specific physical item. Among other, the work order data contains temporal items (start/end dates, planned and actual), costs, description of the physical item and the commissioning body. Several data mining scenarios could be investigated for work order data, such as determining patterns leading to delayed deliveries, forecasting work order durations and time series analysis to predict future workload. Also, a wide range of data quality issues could be envisaged that would affect both the outcome of the data mining and the confidence of the outcome, ranging from syntactical errors and non schema compliance to faulty processes both upstream and downstream of the data capturing. The work presented here implements a relatively simple data mining structure which attempts to forecast a binary work order delay value (work order will be delayed / not delayed). Two data mining models are used, (1) *clustering* and (2) *decision tree*, implementing both discrete and continuous attributes. To asses the effect of the improved data quality on the data mining models, the generic component of three data quality metrics; *completeness*, *accuracy* and *integrity* are used, and the single and accumulate effects are measured for both the discrete and the continuous model. Initially, the data mining model is created for the original (dirty) source, subsequently the selected metrics are defined and measured and finally the data mining is performed for the cleansed dataset. Lift charts [21] are used to determine the accuracy of the data mining models created with both dirty and cleansed data and hence the effect of the data quality on the prediction accuracy can be determined. The results are compared to the theoretical study by Blake and Mangiameli [7].

BACKGROUND

Work order data systems are used to track and plan the installation, maintenance, service and repair of any type of equipment. Commonly, the systems are used to create reports for work in progress or completed work, however, historical data also provides a great potential for predictive analysis reports typically determining the probability for a new or progressing work order to complete on time. Also, clustering algorithms could be used to map out the factors influencing the delay. The work order data considered in this article consists of the entities *client*, *work order*, *external resources*, *internal resources*, *stock resources* and the *physical item* records, as shown in the figure 1.

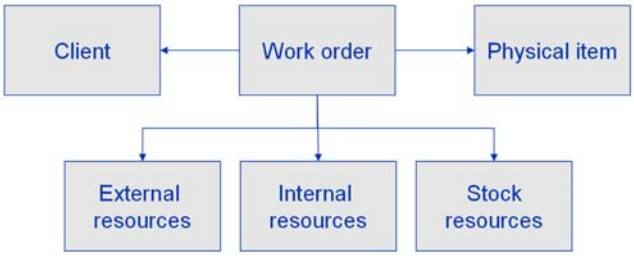

Figure 1 – Work order data tables

The dataset contains approximately 50 000 work orders created over a period of 15 years for the maintenance and repair of complex military equipment for the Norwegian Armed Forces. The data are currently used both for reporting purposes as well as for statistical analyses to determine trends or any anomalies in the data. It has been known that there are data quality issues in the work order data that could potentially adversely influence the quality of the reports and the statistical analyses. The ability to perform the work specified in the work order in a timely manner is extremely important as any delays will result in possibly significant down stream costs as delayed equipment could detain troops and large scale

live events such as military exercises or other major activities. Hence, predictive analyses to pinpoint bottlenecks and recurring patterns resulting in delayed work orders are performed continuously and the data quality needs to be monitored to assess the confidence level of the analysis. The work described here attempts to measure the amount of data quality problems for three well defined and generic data quality metrics (completeness, accuracy and integrity) and to subsequently measure the effect of the bad data on the quality of the data analysis. The effect is measured by comparing the accuracy of the lift charts for the trained model for both the original (dirty) data and for the cleansed data.

THE TOOLSET

The measurement of the effect of the data quality on the data mining accuracy is implemented by evaluating the Data Quality Index (DQI) [16] for data quality metrics relevant to the particular data mining scenario. The extracted work order data resides in a MS Access database and adapters are created to upload this data to the Data Quality Framework [13] to assess the data quality indices and to perform data mining. MS SQL Server (SS) Integration Services is used for data integration, profiling and cleansing, SS Analysis Service is used for data mining [12] and the data quality indices are formalized and measured by the tool described in [13]. The figure 2 shows the applied tools which can be adapted to any data source to measure effects of data quality on data analysis scenarios, often formalized as a Data Quality Deployment Matrix [18] or a Data Quality Scorecard [17] with DQI's plotted against analysis type or business requirements. This will yield the confidence level of the analysis result as a function of the relevant DQI.

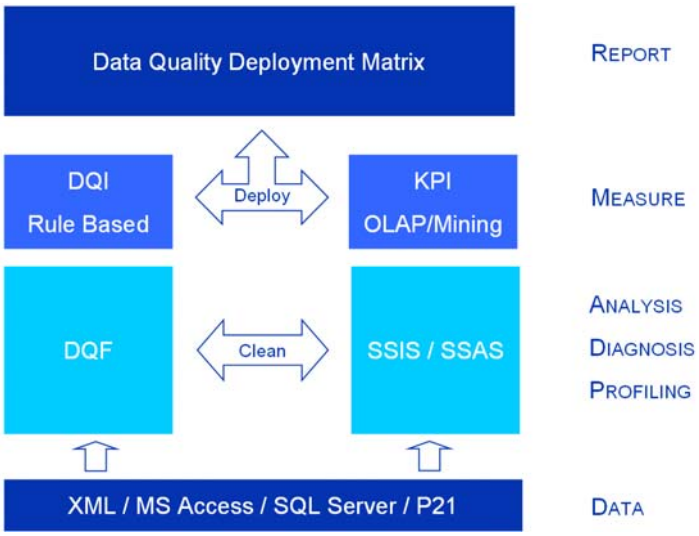

Figure 2 – The Data Quality Framework / Toolset

DATA QUALITY METRICS

A plethora of data quality metrics and supporting frameworks are defined in the literature, a good review is given in [20]. The initial approach for measuring data quality here follows the diagnostics approach described by Wang in [19]. The result of the diagnostics is subsequently used as indicators for how to define the formal data quality rules which should be applied to all data residing in the system. The formal rules should be applied and monitored repeatedly in a Total Data Quality Management (TDQM) [15] or Define-Measure-Analyze-Inspect-Control (DMAIC) cycle [2,5] framework to ensure a consistently high and predictable data quality. In general, data quality issues could be divided into *detectable* and *undetectable* errors. The detectable errors (also denoted *syntactic* in [4]) can be automatically identified and cleaned, whereas the undetectable (*semantic* [4]) must be evaluated by sampling and the effect

modeled by data quality matrices that can be incorporated into the data mining algorithms [22]. As described by Davidson and Tayi in [22], this will, in combination with bootstrapping (or bagging) [23], normally reduce the data variance and hence improve the prediction accuracy for the mining model. The data quality metrics described below (*completeness*, *accuracy* and *integrity*) can be applied to both types of data quality issue, however, the work described here is restricted to the detectable type. This implies that all invalid entries are identified by automatic means and subsequently removed from or replaced in the dataset. Further work could apply the methodology described by Davidson and Tayi in [22] for the work order data described in this article.

The definitions of *completeness* and *accuracy* used here leverage a simple rule definition of both metrics to conveniently measure the data quality index which again is used to quantify the amount of bad data. Each attribute is evaluated as valid or invalid according to the relevant metric and the DQI is calculated as

$$DQI = 1 - (\frac{A_{invalid}}{A_{evaluated}}),$$

where $A_{evaluated}$ denotes all present attributes and $A_{invalid}$ are all attributes failing the relevant data quality metric rule. A DQI equal to 1 indicates that all attributes meets the requirements whereas 0 means that all data are bad. The DQI can then be used as an indicative measure for the improvement of data and the resulting effect on the data mining accuracy.

Completeness

In this article adopt the definition for completeness as given by Naumann [3], yielding this shorthand for completeness C when coverage is considered as unity (ie. all data is required):

$$C = 1 - (\frac{c_{missing}}{c_{\max}}),$$

where $c_{missing} = |(e \in s \mid e.a = null)|$ *and* $c_{\max} = |s|$

where C denotes a value from 0-1 and $c_{missing}$ and c_{max} represent the missing data and the max data content respectively. The relevance of the attributes is considered as a function of the data mining scenario. Data used by the data mining model should be present and any missing values are substituted with 0, deteriorating the accuracy of the data mining. To evaluate the effect of the null values on the data mining outcome, all records containing null value attributes are eliminated from the original data to create the cleansed dataset.

Accuracy

The data quality metric accuracy is commonly defined as the closeness between the given value and the value of the real world item being represented [1]. In this article, the accuracy is estimated by locating relative outliers from a distribution chart and is assigned a binary value, accurate or not accurate. The distribution charts are manually inspected to locate possible outliers and rules are defined to calculate the attribute Data Quality Index [16]. Attributes found to be inaccurate are removed from the original dataset to produce the cleansed dataset. Please refer to the section titled *Data Quality Rules* for examples of outlier thresholds used here.

Integrity

The work described here implements the second of Codds [10] five integrity constraints; (1) entity integrity, (2) referential integrity, (3) domain integrity, (4) column integrity and (5) user defined integrity. Referential integrity measures the validity of foreign keys in any table, assuming every key should refer

to an existing and valid record in an external table. Similarly to the completeness and accuracy metric described above, the integrity measure is calculated as the normalized form of the ratio of invalid references to available keys, yielding linearly 0 for all bad and 1 for all good. The section titled *Data Quality Rules* gives examples of referential integrity measures used here.

DATA MINING

Data mining is the process of analyzing data with the purpose of discovering non intuitive patterns that can be used to gain valuable insights to improve the current application of the data. Frequently, data mining is one of several techniques implemented as a part of Business Intelligence initiatives. The algorithms deployed in data mining commonly range from simple statistical clustering techniques to advanced methods such as linear/non linear regression, time series predictions and neural networks. The availability of user friendly data mining tools in most mainstream database engines have over the recent years resulted in a more widespread use of a technology previously reserved for statistical experts. With respect to the fact that all experience indicates that all data sources contain bad data to some extent, the recent abundant use of data mining stress the importance of highlighting and measuring the effect of the bad data on the data mining results. Also, recent studies [11] illustrate the usefulness of data mining to assess data quality. This process is commonly coined Data Quality Mining (DQM).

In this article we employ two common data mining algorithms, clustering and decision trees. Data mining models are created for both algorithms using both clustered and continuous input data applied to both the original data (dirty) and the cleansed data. This gives us a comparison of the effect of the bad data both for data mining algorithms and for data mining structures.

Clustering

Clustering is commonly labeled as a classical data mining technique. Clustering has been in use for decades, however, clearly the advent of computers have made it more applicable to large datasets. Clustering performs segmentation through an iterative process to group records with similar characteristics. The technique can be used to detect anomalies (values outside main clusters) as well as for making predictions and classification schemes.

Decision trees

Decision trees have been used extensively since the late 1980s [14]. Decision tree is a classification and regression algorithm used to perform predictive analyses. The result is formalized as a tree structure where each leaf depicts a weighed relation between the characteristic it represents and the outcome defined by the root node.

Input structure

The input values are given as either discrete or continuous variables. Typically the predictable column is given as a state variable (true or false). Attributes with inherent discrete values (categories) are defined as discrete for all examples, whereas attributes of a continuous nature is defined both as discrete and continuous for providing comparisons between the two. When using discrete values for the decision trees, relations between column characteristics and the predictive column are established to create a tree structure indicating probable outcomes. For the continuous values, the decision tree algorithm will fit lines through the values and apply linear regression to each node created at the line discontinuities. The motivation for using discrete values when source data is continuous is twofold: (1) Some mining algorithms are restricted to discrete values and (2) the result could more easily be interpreted if a large value range is dived into intervals. The segmentation method to create discrete “buckets” from continuous variables is simply attempting to create groups with an equal number of elements. Incomplete input data represented as *null values* are defined as a separate state for all columns labeled *missing*. This implies that

the mining model will treat incomplete data as any other data and include the missing state as an outcome or effect with a calculated probability of occurring. This will prevent the model from degenerating when encountering null values, however, a high number of *missing* will obviously distort the result.

Lift charts

Lift charts [21] were devised by the marketing industry to be able to predict potential customer response to advertising campaigns. Target groups with a high likely response rate can be identified to facilitate efficient marketing. By holding out a given percentage of the population when training the model (a 30% holdout population was used in the case described here), the accuracy of the predictive model can be assessed. Figure 3 indicates the main elements of the lift charts; *target population*, *response*, *baseline*, *ideal line* and *actual prediction accuracy*.

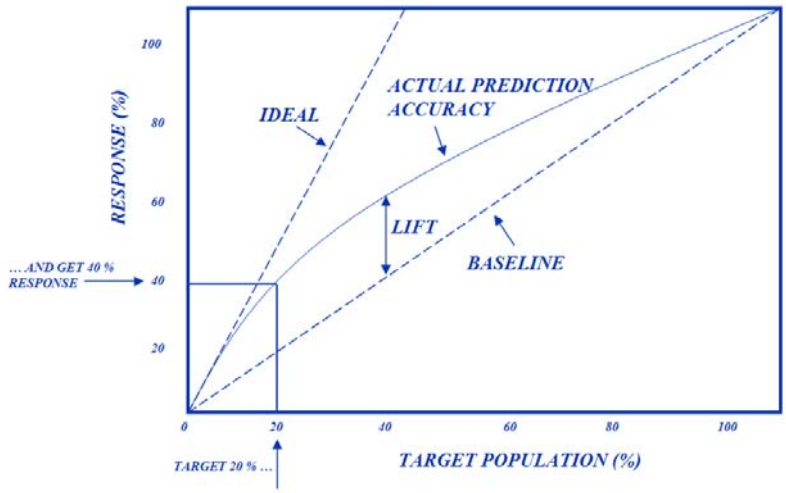

Figure 3 – Lift chart

In the above example, the *baseline* indicates the response rate for random selections, ie. targeting 20% of the population will reach 20% of the potential positive responses. The *ideal* line shows that for this population the best predictive model would be able to give 100% response for a target population of approximately 40%. The *actual prediction accuracy* is the accuracy of the trained model with the given holdout population and the *lift* is the improvement as compared to the baseline for any target population. In this particular case, a target population of 20% can be used to reach 40% of the potential positive responses.

The usefulness for lift charts as rating devises for unrelated analysis models is questionable [21]. However, for relative comparisons between different predictive algorithms for the same dataset, or, as in the case described here, as a ranking device for identical algorithms applied to the same dataset but with varying levels of data quality they provide a useful means to evaluate the relative performance.

CASE STUDY

The following section describes the case study where the previously defined methods are used to assess the data quality of a particular work order database and to measure the effect of the data quality on a defined predictive analysis.

Defining the Predictive Analysis

The work order database contains temporal values specifying start/end dates for the planned duration, the actual duration and the physical item availability. Also, the type of physical item, the manufacturer, the client and miscellaneous costs are given. One of the main concerns of the clients is to ensure that the work on the physical items is completed within the time the item is made available to the workshop. As mentioned previously, any unforeseen delays are likely to incur other costly downstream consequences. Hence, a delayed work order is defined as any activity where the actual end date exceeds the end date for item availability. To assist the client and the workshop in identifying work orders which are likely to be delayed, the predictive analysis should attempt to predict the probability of any work order exceeding item availability using planned days, delayed start, item type, costs and item currency. Item currency is included to detect any adverse effects of working on expired items, such as non availability of spare parts. To avoid discovering a high number of weak relationships, the number of input fields is kept at a minimum as suggested by McClanahan in [8].

Diagnostics

The diagnostics approach to data quality is well suited when a clearly defined exchange agreement specifying data quality requirements is unavailable. Commonly the diagnostic activity will produce a more formal understanding of the data quality problems present in the data source and hence this is frequently applied in the initial phase of the TDQM cycle. Generic data quality metrics such as completeness, accuracy, overlap and statistical distributions provides data profiles and indicators to potential problem areas. Once identified, formal data quality rules yielding data quality indicators can be defined and described in the exchange agreement. In this particular case, the diagnostics indicated an overall high completeness and the value distributions were largely evenly distributed making outliers easy to spot. Figure 4 shows the general completeness plot for the entire work order model as well as one example of value distributions with outliers. Due to non disclosure, the figure yields no actual data. However, the left hand side shows model completeness by table where the green color depicts null values. The right hand side illustrates an outlier plot where the red circles indicate definite outlier candidates for the *start date* attribute where the date exceeds the current date by more than 200 years.

The general diagnosis charts enable a rapid overview of the data and highlights potential anomalies. From the below completeness chart we can infer that the data is possibly well suited for data mining as a high ratio of the required data is present. Also, the relatively few outliers and generally even distribution of values gives no indication of extensive data contamination. Ideally, the semantic quality [4] of the data (ie. correspondence to real world) should also be considered. However, due to time restrictions, in this work this is only performed on an intuitive level. Physical items with zero cost or dates significantly outside the relevant range are considered inaccurate by inspection and rules are defined to quantify the particular indicators.

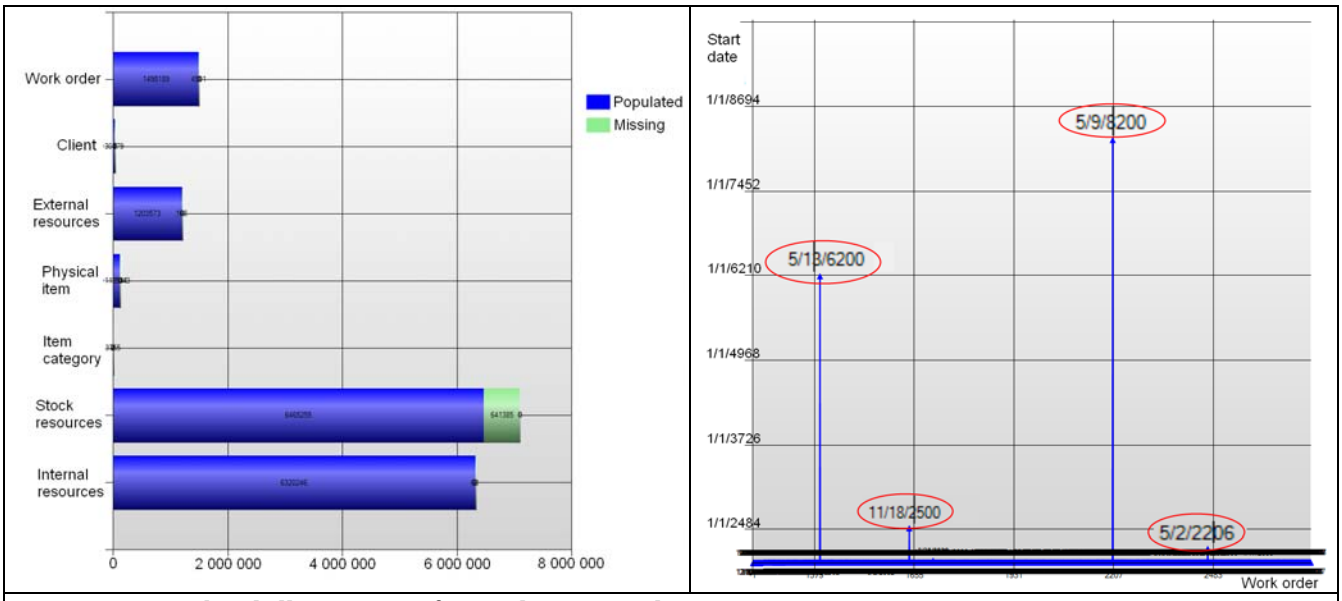

Figure 4 – Graphical Illustrations of Completeness and Accuracy

Data Quality Rules

Drilldowns in the general diagnostics result is performed to locate and measure the data quality issues related to the data mining model. Rules are defined to measure the completeness and accuracy of the temporal values for planning, execution and item availability, as well as rules considering the same metrics for item and resource costs. Figure 5 shows the results from the rule evaluation. Again, the actual values are illegible due to non disclosure, however, the general trend of high scores on data quality indices are shown. The red part of the column depicts the relative few errors in the dataset, typically giving data quality indices equal to approximately 0.92 - 0.98 (8 - 2% error) according to the described metrics. The rules shown in figure 5 provide measures for (1) work orders where end date is given as earlier than start data, (2) stock item unit price is zero and (3) the work order refers to non existing external or internal resources (ie. foreign key values referring to non existing primary keys). In addition, (4) the date outliers shown in figure 4 were detected by disallowing any date exceeding the current date plus 20 years. In the preceding, (1), (2) and (4) are labeled *inaccuracies*, whereas (3) is accounted for as referential integrity errors. Also, general *completeness* rules were applied (populated or missing).

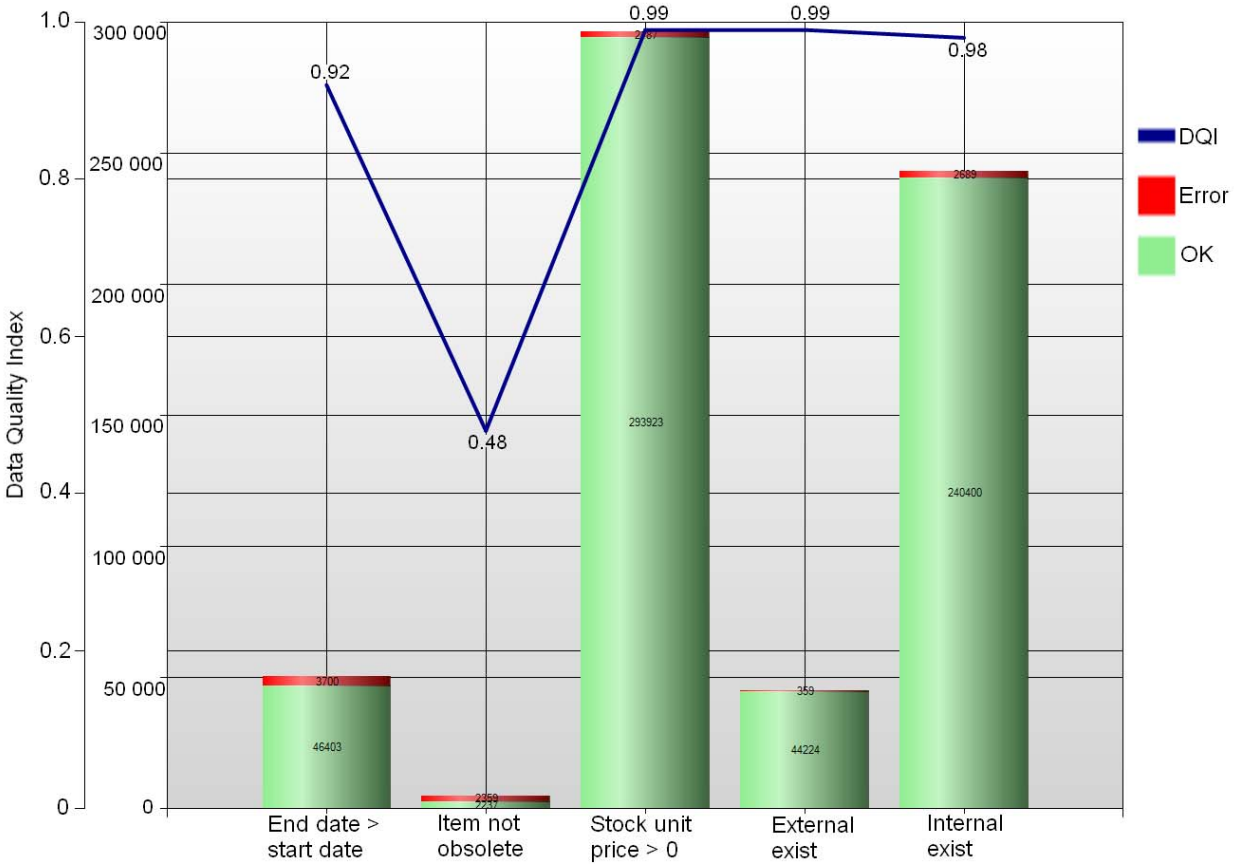

Figure 5 – Data Quality Indices for the Defined Rules

Data Cleansing

To facilitate the measurement of the effect of the data quality on the data mining accuracy, a cleansed data model is generated by resolving the issues reported by the data quality indices. The cleansing performed here simply discards any records with inaccurate values and auto completes missing values if possible. A large number of physical items were missing the expiry data, obviously because the item had not expired. This null value was corrected by entering current date adding 20 years, ensuring the item would not be excluded from the model or erroneously tagged as obsolete. Also, some date outliers for end of availability for the physical item was considered to represent items with infinite availability. These values were set equal to the actual end date to eliminate any model disturbance caused by assigning any significance to the wrong values. Further, some values that could benefit the model had to be disregarded due to suspicious values. The physical item cost distribution indicated a relatively high number of zeros and inspection revealed that the item cost was always filled in however it was frequently incorrect and had to be excluded from the mining model. This was unfortunate as the physical item cost could bear some impact on the mining outcome. By intuition the maintenance of costly items could be more complex and hence more prone to delays as compared to less costly items.

The data cleansing resulted in an overall reduction in the size of the dataset of approximately 6%, ideally the cleansing should not be intrusive and rather aim to repair values, however, resource restrictions dictates that this approach must suffice here.

Data Mining Model and Structure

The motivation for the data mining exercise is to establish a tentative model containing values known at the time of work order registration to predict the probability of the particular work order being delayed

and the most likely factors influencing the delay. We have defined 4 separate models using 2 different data mining algorithms, respectively discrete and continuous variables for original and cleansed data using both clustering and decision trees. The predictable value, *delayed*, is defined as true if the actual completion date exceeds the end of availability data and false otherwise. To determine the probability of the delayed value, 7 input values are defined: *planned days* (planned end ÷ planned start), *delayed start* (actual start ÷ planned start), *costs* (given directly), *item type* (categorical value, given directly) and *obsolete status* for the physical item (expired date < start date). Inherently discrete attributes (such as item name) is kept discrete for all models. On the other hand, continuous variables such as dates, durations and costs are treated as both continuous as well as discrete. To convert continuous variables as discrete the continuous values are divided into equal discrete buckets of variables. This is introduced to assess the effect of using continuous versus discrete variables on the data mining analysis result. The values are calculated and joined from the original tables to create the flattened and de-normalized table suitable for data mining.

Predicting Work Order Delay and Causes

All the defined models and selected data mining algorithms agreed on the predicted result. Hence, the comparisons are confined to the level of influence for the input values and the level of confidence (accuracy) of the analysis. Figure 6 depicts the predictable value in the centre and the input values at the perimeter. The number assigned to each arrow indicates the relative strength of the influence, where 1 is strong and 7 is weak. Figure 7 further elaborates on the results by showing the relative attribute discriminations indicating what value ranges are likely to predict false or true. Several interesting issues can be inferred from the two figures:

(1) The number of planned days has the strongest influence on the delay prediction, however, most would expect long work orders to be more susceptible to delays. The opposite is found here, where work orders planned to last less than 36 days are significantly more likely to be delayed as compared to work order with planned duration above 36 days.
(2) Some item types (item name) are more susceptible to delays than others. In particular, navigational items are less likely to be delayed than intendancy items.
(3) Work orders on obsolete physical items (past expired date) have no significant contribution to delay.

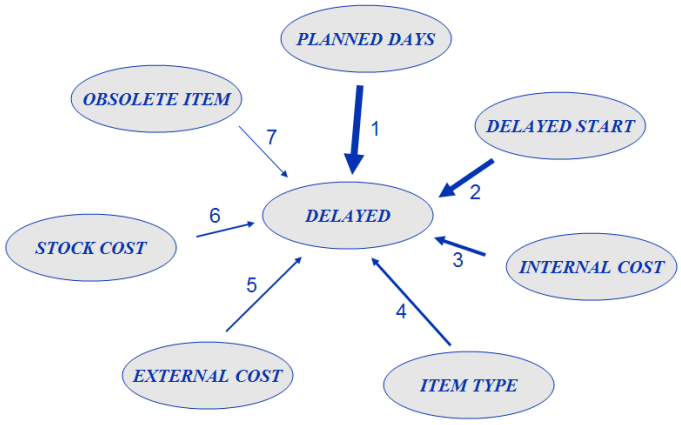

Figure 6 – Attribute level of influence on outcome

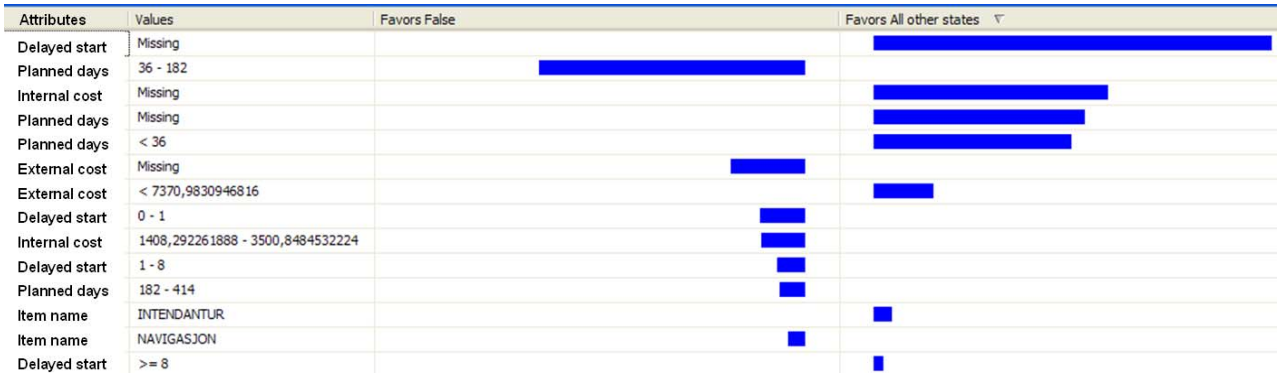

Figure 7 – Attribute discrimination

Estimating the Mining Model Accuracy

Lift charts are used to evaluate the accuracy of the mining models. The accuracy estimate is performed on a test set which is randomly selected from the input model. The test set consists of historical data which contains the delay state and hence this test set can be used as queries to the trained model to assess the accuracy of the mining model. The target population (true delay prediction) percentage is shown along the vertical axis whereas the overall population (all work orders) is shown along the horizontal axis. The straight line running diagonally from the lower left to the upper right corner of the charts shown in figures 8 and 9 represent random guess values for the delayed state, whereas the steepest line shows the ideal situation where all predictions for the work order delay state are correct. All other curves fall in between the ideal line and the random guess line (baseline) and represent the accuracy of the different mining models expressed as the "lift" from the random guess line. As can be seen, the best model is achieved by using continuous variables on the cleansed model, yielding a 17% lift. However it could be discussed if this is a good mining model or if it should be further cleansed or complemented, this article is only concerned with the relative differences in lift, not the overall usefulness of the mining model itself.

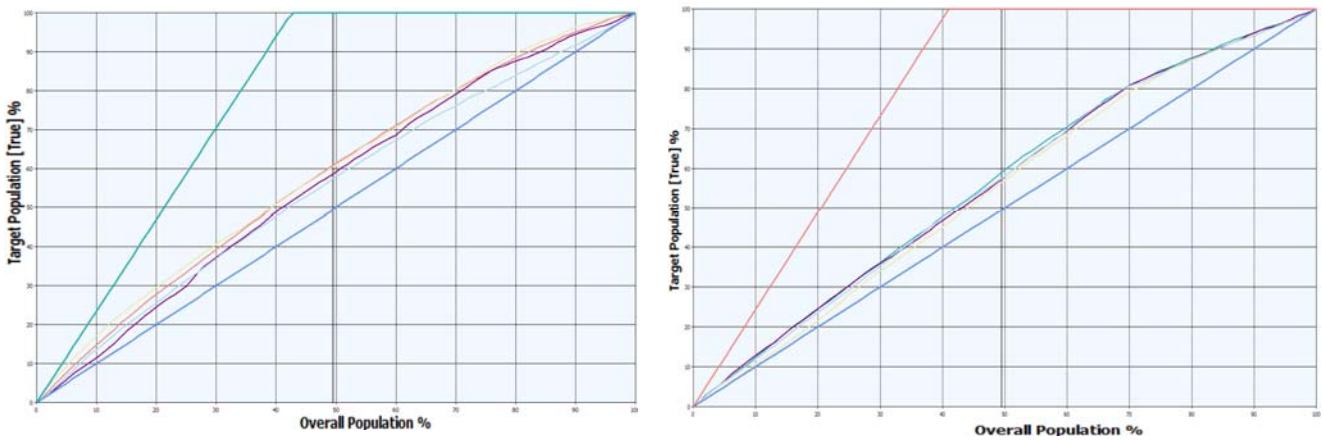

Figure 8 – Lift chart for discrete attributes for (a) original data and (b) cleansed data

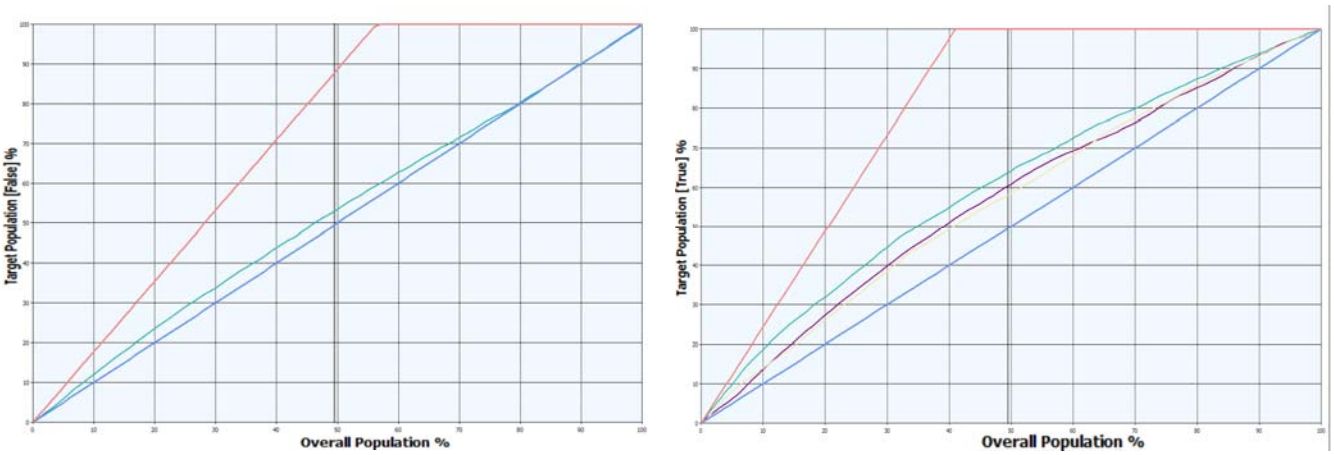

Figure 9 – Lift chart for continuous attributes for (a) original data and (b) cleansed data

Figures 9 illustrate clearly the significant positive effect of the data cleansing on the data mining outcome for the models using continuous variables. The original data set only yields a 5% lift whereas the cleansed model gives 17% for the decision tree model, slightly less for the clustering model. The discrete models shown in figure 8 which operate on less granular data are seen to be less influenced by the data quality issues as compared to the continuous models considering single values. Table 1 summarizes the maximum lift from each of the above figures.

	Original	Cleansed
Discrete	10%	10%
Continuous	5%	17%

Table 1 – Maximum lift for the respective models

As reported by Blake and Mangiameli in [7], the different data quality metrics will influence the accuracy of the data mining model to varying degrees. By breaking down the cleansed model according to the data quality metrics, the trend shown in [7] where consistency (integrity) and accuracy is found to have a more significant impact on the data mining result as compared to completeness, is confirmed. Table 2 shows the contributions from the three data quality metrics described previously in the article, namely *integrity*, *accuracy* and *completeness*. The figures are not normalized and can only be expected to serve as trend indicators. Also, the metric definitions may not be strictly in accordance with [7] and therefore should only be taken as an indication that this practical case seem to follow the theoretical scenarios presented in [7].

	Cleansed	Lift
Completeness	7%	13.8%
Accuracy	0.3%	14.5%
Integrity	1.5%	15.1%

Table 2 – The effect of data quality metrics on the data mining model

Verification

In addition to measuring the accuracy of the data mining model and the effect of data quality issues, the mining results should also be verified by comparisons to the underlying data. In the case presented in this article, the unexpected results showing that short projects are more likely to be delayed is particularly interesting to confirm by inspecting the source data. Figure 10 show planned duration for the work order

as vertical columns (labeled values 1-65 days) and the accumulated delayed days along the unlabeled line with triangles as data point markers. The accumulated delayed days are scaled according to the vertical axis on the right hand side (days). Also, the percentage of delayed work orders for the particular planned durations are shown as the unlabeled vertical columns scaled on the left hand side axis (%). The verification chart confirms the mining results by showing a clear trend for decreasing project delays as the work order planned duration is increasing. Also, the subsequent delay (number of days downstream projects must wait for the physical item) are only significant (>100 days) for projects planned to last approximately 15 days or less. There could be several feasible explanations for why short projects overrun whereas longer projects deliver on time, a more professional handling of high cost work orders could be one. Also, the significant peak for work orders planned to last 1 day could suggest some upstream data quality problem. However, it is outside the scope of this work to elaborate further into the underlying causes.

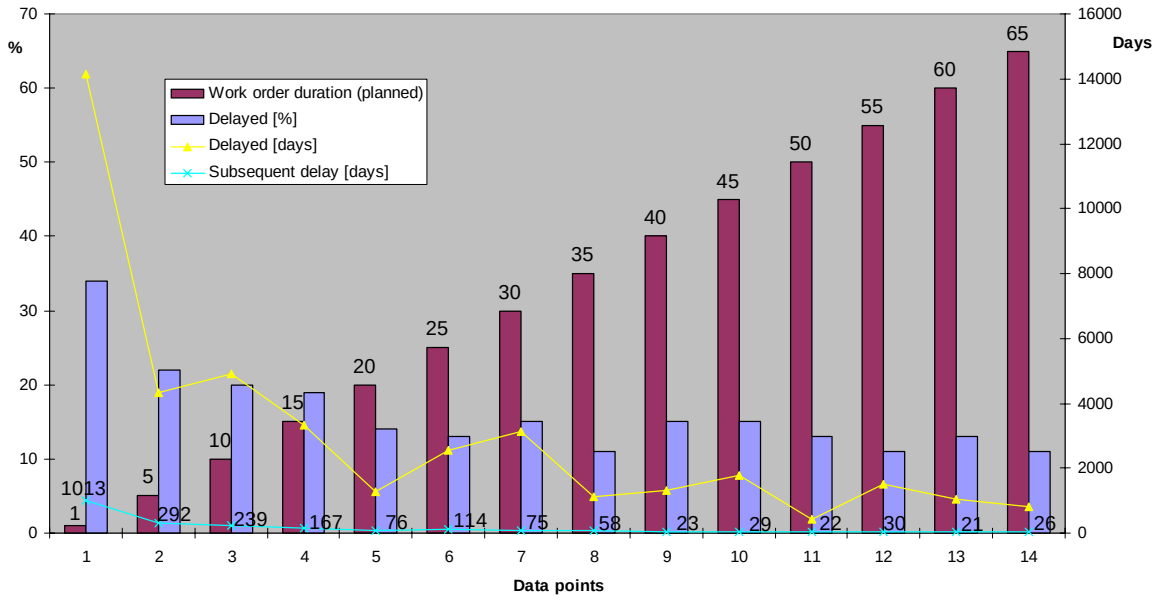

Figure 10 – Verification of the data mining result

Return of Investment for Data Quality in Work Order Data

The return of investment for data quality activities could be elusive and difficult to measure directly in a tangible manner. For the work order scenario described here, the ability to predict delays and hence mitigate the factors causing the delay will improve the ability to perform downstream activities as planned. Hence, the direct cost of data quality issues making the data mining less accurate is probably insignificant as compared to the indirect cost incurred when planned down stream activities have to be carried out without the physical item or the activity have to be postponed. Hence, the workshop itself will probably experience minimal gains by improving the planning phase for short projects, however, the units submitting the items for work should be keener to improve the predictability. Typically, military equipment is of dire importance in both exercise and real scenarios.

CONCLUSION / DISCUSSION

This article demonstrates the effect of data quality on data mining by measuring the accuracy of data mining models before and after cleansing the source data. Two interesting findings can be noted:

(1) Using discrete attribute models produce more robust models which will be less susceptible to data quality issues. For the particular case presented here, the cleansed dataset did not improve the accuracy of the discrete mining model by any measureable amount.
(2) Continuous attribute models perform poorly as compared to discrete models when using dirty data, however, they respond well to data cleansing and for this particular example the accuracy of the continuous model exceeded the discrete model for the cleansed dataset.

Both techniques used here, clustering and decision trees, showed the same characteristics as explained in (1) and (2), however decision trees performed marginally better than clustering for this case.

By using the generic component (ie. context independent) of the data quality metrics *accuracy, completeness* and *integrity,* cleansing was performed automatically by removing outliers and invalid references to improve accuracy and integrity, and eliminating nulls or filling in valid values to improve completeness. The actual cleansing was only required for approximately 6% of the data whereas the most significant improvement in accuracy for the continuous attribute and decision tree based mining model was found to be 12%. This indicates the importance of operating on clean data and also the relatively high impact data cleansing efforts will have on the confidence levels of the data mining outcome. Further, the data mining outcome itself has revealed other data quality issues which are context and process dependent. In this particular case the data mining revealed a high probability for short projects to be overdue whereas the more extensive projects was relatively well managed and so delivered on time. On the process level this indicates suboptimal routines for handling small projects, resulting in more downstream delays than for the big projects. On the data quality level, this also indicates that the temporal values for planning (start/end dates) were over represented by the value 1 (day) which is likely to be an inaccurate value in the planning context as these projects were predominantly delayed. This clearly illustrates how data mining can be efficiently used to detect anomalies in the data. Upstream process improvements should be undertaken to verify the correctness of all work order durations planned for 1 day. Also, a scorecard or dashboard could be employed to highlight all work order durations exceeding item availability to improve general predictability.

The accuracy of the data mining models were evaluated by lift charts [21] and the improved accuracy obtained by cleansing the model was measured both as an aggregate and for each of the data quality metrics *completeness, accuracy* and *integrity* separately. As reported by Blake in [7], this work also found the integrity and accuracy data quality metrics to have the most significant impact on the data mining accuracy. More work needs to be carried out to further establish this relation, especially for larger and more complex models, however, it is interesting to note the tentative correspondence between the practical case described here and the theoretical study performed in [7].

On a general note, the work order dataset exhibited a high completeness, which is indicative for a good case for data mining. However, the statistical profiles for the column values revealed several dummy value candidates, such as 0.0 for the price of some costly item. This will obviously impair the accuracy of the mining model and is onerous to detect and clean automatically. The high degree of completeness could be the result of using mandatory input values ”forcing” hurried users to enter the inaccurate data. Hence, this advocates for avoiding mandatory values as null (incomplete) values obviously are more easily detected generically than context dependent inaccuracies. Also, it suggests that domain expertise should be used to further cleanse and enrich the model to improve the prediction accuracy.

REFERENCES

[1] Batini, C. and Scannapieco, M., "*Data Quality, Concepts, Methodologies and Techniques*", Springer , 2006

[2] Brunson, D. and Frank, S. , "*Six Sigma Data Quality Processes*", http://www.b-eye-network.com/view/2756, May 2, 2006

[3] Naumann, F., Freytag, J.C. and Leser, U., "*Completeness of Integrated Information Sources*", Information Systems 29 (7), 583-615, 2004

[4] Price, R. and Shanks, G., "*Empirical Refinement of a Semiotic Information Quality Framework*", Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.

[5] Pyzdek, T., "*The Six Sigma Handbook*", McGraw-Hill, 2003

[6] Richter, J., "*CLR via C#*", Microsoft Press, 2006b

[7] Blake, R.H., and Mangiameli, P., "*The effects and interactions of data quality and problem complexity on data mining*", Proceedings of the 13th International Conference on Information Quality, ICIQ 2008, pp160-175

[8] McClanahan, "*Cleaning a formulation database using rule discovery techniques*", Proceedings of the 13th International Conference on Information Quality, ICIQ 2008, pp176-183

[9] Hassine, S.B., Clement, D., Laboisse, B., "*Using association rules to detect data quality issues*", Proceedings of the 13th International Conference on Information Quality, ICIQ 2008, pp184-197

[10] Codd, E.F., "*A relational model of data for large shared data banks*", Communications of the ACM 13(6), pp377-387

[11] Hipp, J.,Guntzer, U., Grimmer, U., "*Data quality mining – making a virtue of necessity*", Proceedings of the 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2001

[12] Langit, L., "*Smart business intelligence solutions with Microsoft SQL Server 2008*", Microsoft Press, 2009

[13] Stang, J., T. Christiansen, Skogan, D., Kvalheim, A., Ihrgens, T.A., "*A generic data quality framework applied to the product data for naval vessels*", Proceedings of the 13th International Conference on Information Quality, ICIQ 2008, pp264-396

[14] Berson, A., Smith, S., Thearling, K., "*Building data mining applications for CRM*", McGraw-Hill, 1999

[15] Wang, R., "*A product perspective on total data quality management*", Communications of the ACM, pp58-65, 1998

[16] Lee, Y.W., Pipinio, L.L., Funk, J.D., Wang, R.Y., "*Journey to data quality*", The MIT Press, 2006

[17] Nousak, P., Phelps, R., "*A scorecard approach to improving data quality*", Proceedings of SUGI27, 2002

[18] Stang, J., T. Christiansen, Skogan, D., Kvalheim, A.,"*Six Sigma applied to the product data for naval vessels*", Proceedings of the 13th International Conference on Information Quality, ICIQ 2008, Poster

[19] Pipino, L.L., Lee, Y.W., Wang, R.Y., "*Data quality assessment*", Communications of the ACM, vol. 45, 2002, pp211-218

[20] Fehrenbacher, D.D., Helfert, M., "*An empirical research on the evaluation of data quality dimensions*", Proceedings of the 13th International Conference on Information Quality, ICIQ 2008, pp230-245

[21] Coppock, David S., "*Why Lift?, Data Modeling and Mining*", Information Management Online, June 21, 2002

[22] Davidson, I., Tayi, G.K., "*Data preparation using data quality matrices for classification mining*", European Journal of Operational Research, 2009, pp 764-772

[23] Breiman, L., "*Bagging predictors*", Machine Learning, Kluwer Academic Publishers, 1996, pp 123-140