A data quality metrics hierarchy for reliability data

Ralf Gitzel*, Simone Turrin**, Sylvia Maczey***, Shaomin Wu**** and Björn Schmitz*****

**ABB Corporate Research, ralf.gitzel@de.abb.com*
*** ABB Corporate Research, simone.turrin@de.abb.com*
**** ABB Corporate Research, Sylvia Maczey, sylvia.maczey@de.abb.com*
***** University of Kent, Shaomin Wu, S.M.Wu@kent.ac.uk*
****** Karlsruhe Institute of Technology, Björn Schmitz, bjoern.schmitz@kit.edu*

Abstract. In this paper, we describe an approach to understanding data quality issues in field data used for the calculation of reliability metrics such as availability, reliability over time, or MTBF. The focus lies on data from sources such as maintenance management systems or warranty databases which contain information on failure times, failure modes for all units. We propose a hierarchy of data quality metrics which identify and assess key problems in the input data. The metrics are organized in such a way that they guide the data analyst to those problems with the most impact on the calculation and provide a prioritized action plan for the improvement of data quality. The metrics cover issues such as missing, wrong, implausible and inaccurate data. We use examples with real-world data to showcase our software prototype and to illustrate how the metrics have helped with data preparation. This way, analysts can reduce the amount of wrong conclusions drawn from the data due to mistakes in the input values.

1 Introduction

Many critical business decisions in equipment-heavy industries require a good understanding of the reliability of these assets. However, if the reliability is calculated using field data, any data quality problems will affect the quality of the results. In this paper, we briefly describe the key data quality issues in reliability data (Section 2) and show that there is currently no good way to detect them (Section 3). We propose a series of metrics (Section 4) as well as a 3-tiered hierarchy which simplifies the discovery of the most relevant issues (Section 5). We use an example to illustrate our approach and discuss its practical use. The paper concludes with an outlook on future work (Section 6).

2 Reliability data and its data quality issues

In order to estimate the reliability of a product, a series of key data elements is needed. These are often found in one or more tables. However, for the sake of generalization we assume that data consists of multiple properties associated with events and units (e.g. individual motors, machines, etc.). The events in turn are associated with units. The *core failure data* needed are start date, failure date, and an ID characterizing the unit for tying the two dates together. Also, the product type or version needs to be known to properly group the failure data. Additionally, we need *censoring data* for products which have not failed. Sometimes, we can impute missing core data through *substitution data* (e.g. a delivery or production date instead of a start date). Finally, there is a series of data sets which provide *added value* such as failure modes and information that allows a clustering such as application, customer, or even stress levels.

Missing and *recognizably wrong data* are the most obvious data quality issues. Data is recognizably wrong, if it violates some universally accepted rule, natural law (absolute zero temperature) etc. Both missing and recognizably wrong data, most importantly, cause information loss as events or even whole units have to be removed. Wrong data which cannot be detected can distort the results of the reliability calculation. *Implausible data* is data which is not necessarily wrong but there is a strong suspicion that it might be. It is no easy decision what to do with such data but in general a lot of implausible data means that the data quality is poor. If the *data lacks in richness* of information, there will be a lack of detail that makes the results of a reliability analysis less accurate. For example, accurate failure dates are preferable to just knowing the year of failure. Finally, there are several statistical properties that identify a good *sample selection*, avoiding problems such as bias and small sample size. For a detailed discussion see Gitzel et al 2015.

3 Related work

There is rather large body of papers that addresses the topic of data quality. Data quality is a multi-dimensional problem (Kahn et al 2002, Balou and Pazer 1985) and several dimensional frameworks (see Borek at al 2014, pg. 13 for a survey of frameworks) as well as an ISO standard (Benson 2008) have been developed. Most of these frameworks (with the exception of Ban et al 2008) are based on relatively simple metrics.

However, there is little work which targets data quality in the industrial context, especially with a focus on reliability data analysis. Since data quality is a highly domain-specific topic, this means that there is a large gap in the prior art which needs to be filled. While there are papers which identify data quality problems in industrial data (e.g. IEEE 2007, Vadlamani 2007, Bendell 1988, Montgomery and Hodkiewicz 2014) and there is a supposition that the results of an analysis performed on poor reliability data is "severely misleading" (Bendell 1988), there is little work on quantifying these problems. A notable exception besides our own prior work (Gitzel 2014, Gitzel et al 2015) is a paper on assessing reliability data quality by Montgomery and Hodkiewicz which focuses on the "availability of specific data fields or individual records, the relevance of the data in specific fields, and the accuracy of the data" (Montgomery and Hodkiewicz 2014). Also of note should be algorithms which are designed to deal with "coarse" data, i.e. data of poor quality (for a review, see Wu 2013). A common approach is to use estimations to correct errors (e.g. Ala, and Suzuki 2009). These algorithms can be used in conjunction with our proposed metrics to address the problems discovered.

4 Data quality metrics

In order to understand the data quality problems of a particular case, we use a series of metrics, which represent an updated version of the metrics developed in Gitzel et al 2015. All metrics range from 0 to 100% where a high value means a good quality. Note that some of the names do not follow the literature at the request of our stakeholders.

4.1 Completeness metrics

A common problem is *incomplete data*. Missing data can either be empty properties such as "N/A" or "unknown". For our data quality analysis we have created a list of such terms. In a database, incomplete data takes the form of empty cells. For each property of each event and unit (see section 2), we need a completeness metric. The metric reflects the percentage of properties which are not empty.

4.2 Free-of-error

Errors in the data are one of the most common data quality problems. While not all errors can be discovered, there are some rules which can be used to discover them. For the purpose of our discussion, we distinguish between *logical*, *set-membership* and *syntactical errors*. For each rule there is a metric which covers one or more properties. The metric is the percentage of units or events that violate a particular rule.

Logical errors can be discovered through rules based on one or more data elements, many of which depend on the domain. Set membership errors occur if a data element should be one of several possibilities but the actual value is not one of those. Syntactical errors are typos that let a value deviate from a given pattern. For example, an event type could be failure or censoring but not "break-down" (set-membership error). As another example, the start date of a product cannot occur after the failure date (logical error). Providing the string "Q2.02.2015" as a date is a syntax error.

4.3 Inconsistency

Inconsistency is a data quality problem of lesser importance. It mostly refers to consistency of notation (e.g. format of serial numbers) or units (cm vs. inches). Inconsistency is not a problem if it does not hide duplicates (e.g. the same unit registered under different serial numbers) or leads to mixing different units of measurement. However, it does increase the probability of such errors occurring. Metrics measuring inconsistency can count the number of elements which conform to a specific pattern, e.g. a predefined structure for serial numbers.

4.4 Plausibility

Like free-of-error metrics, plausibility metrics are based on a series of rules. Many of these rules are highly domain-specific but there are some recurring plausibility issues. For example, there are some basic sanity checks associated with events. Normally, we know the release date of a product, so we can check whether there are any start dates before the corresponding release date, which is not impossible (e.g. pilot installations) but improbable in larger numbers. In the context of events, a common problem is the *"double tap"*, where a failure occurs very shortly (ca. 24h) after another one. In many cases, this is a reporting artifact and needs to be removed.

Finally, there should be one or more rules to check the *distribution of dates*. If there are dates which occur with great frequency, this could be an assumed date for old data. Our metric for this rule is 1 minus the highest percentage of occurrences. Also, if there are large areas devoid of failures, it could be that no records exist for that period of time. If there is a suspicion about such an area, a simple metric can measure the percentage of units which start and fail before or after the gap and are thus not affected. More advanced metrics could discover such voids.

Plausibility can also be applied to *maximum and minimum values* for a particular property. It is possible to have one free-of-error metric which convers impossible values (running hours that exceed the time between installation and failure) and one plausibility metric to identify unlikely values (running hours above x% of the time span as appropriate for the product).

A special kind of plausibility rule covers *sub-population membership*. If you want to base your subpopulation on certain properties (e.g. a specific customer and failure mode) there is a problem if these properties are affected by poor data quality. We use a metric which is 1 minus the percentage of units that might or might not be part of the fleet. This includes empty values as well as names that are similar (as identified by means of a Damerau-Levenshtein algorithm). Often the latter problem is quick to fix once identified.

Data that is *not trustworthy* comes from dubious sources, e.g. data collected by people we do not know for purposes other than reliability analysis. Unfortunately, the metadata required to judge the trustworthiness of data is rarely available. If such data is available, metrics can count all units and events coming from trustworthy sources.

4.5 Richness of information

If there are no details in the information, we consider the information to *lack in richness*. For example, a temperature could be described as "hot" or as "38°C". Both statements can describe the same temperature but the second description is far more accurate. A typical issue in reliability data is inaccurate dates (production month instead of exact day). In a few cases, we do not have information about individual failures, i.e. we know the number of units that failed in a time period but not which units. In many cases, richness is a static assessment. For example, if there is only a property "manufacturing year" instead of "manufacturing date", the richness is low throughout the sample because the exact day within the year is not known. However, it is also possible to search for properties that are suspicious (e.g. "1.1.20xx" or "6'000"). If the majority of running hours entries is multiples of 10, 100, or 1000, the information is probably less rich than numbers which are non-zero in multiple digits. In this case, a metric would measure the percentage of units that are not as accurate as the units of a property would suggest.

Information richness also covers machine readability. For example free-text failure descriptions are less accurately interpretable to a software-based algorithm than failure codes. An issue related to richness is the *excessive use of a category "other".* While using "other" to cover exotic cases is perfectly acceptable, a property that is "other" in 90% of the cases is quite useless for the purpose of identifying sub-populations.

4.6 Sample selection

In statistical analysis, sample selection is an important quality factor. It is not possible to measure all aspects of sample selection, however, there are a few factors which can be quantified easily. Typically, the **sample size** is of great importance for the quality of the calculated reliability. What makes a good sample size depends on the standard deviation of the population the sample is taken from (which can be estimated from

the standard deviation of the sample). We propose to use a metric which is reduced by the relative standard error of the mean for sample sizes of 30 or more and 0% for sample sizes below 30.

Furthermore, the selection of the sample must be *unbiased*, i.e., the sample should represent the population from which the sample is collected from. In other words, the composition of the sample regarding environmental factors, maintenance policy, applications, industry etc. should be very close or identical to that of the total population. For example, if a sample contains 30% assets from the chemical industry and 70% from food & beverage while the installed base for a product is 50/50, the sample will be unbiased. The metric to measure unbiasedness can be based on the Chebychev distance between the percentages in the sample and the percentages in the population, calibrated to give 100% if the distance is 0.

A special form of bias is *warranty bias*, i.e. samples where the majority of products come from the warranty period. A strong warranty bias implies that the reliability analysis focuses on those products with the poorest performance. A simple metric just counts the percentage of data originating from the warranty period.

Another problem occurs if we only know about failures and not about non-failures. In order to get an unbiased sample, we also have to include non-failures as censored data. However, in some cases (e.g. if we only know warranty data) the censoring date has to be based on an assumption for many or all of the units. We call this phenomenon *unconfirmed censoring* and have found that it can greatly impact on warranty data analysis and needs to be addressed with additional data collection (e.g., questionnaire surveys).

4.7 Substitution quality

Ideally, our reliability calculation uses start dates and failure dates. However, in some cases, this information is not available and we substitute dates which are correlated to these, e.g. delivery dates. If there is a high *level of substitution*, this affects the quality of the calculation. We use a metric that tracks the percentage of units which use the real start date instead of substitutes like the manufacturing date. A second metric tracks the same for failure dates as opposed to reporting dates.

5 Data quality metrics hierarchy

When using our first version of the metrics (Gitzel et al 2015), we found that it was difficult to identify the key problems if there was a lot of less important information in the database. For example, an implausible company name is not as problematic as a missing failure date. Using a hierarchy of metrics allows data analysts to quickly understand the key issues in order to react to them.

5.1 Importance of metrics

Our hierarchy of metrics is based on different levels of importance for data quality problems. We use 5 different levels of importance, as described in the following.

"Critical" metrics address problems which will remove an event or unit from the calculation, thus reducing the sample size, or which might falsify the results (an implausible failure date). If any of the critical metrics is low, action is required before the results of the reliability calculation can be used.

"Substitution for Critical" metrics identify the degree to which the real critical values are replaced by inaccurate values that need to be used because the real value is not known (see Section 0). These problems are often not worth the effort to fix but improvements in the data collection process can improve the accuracy of the reliability calculation.

"Subfleet" metrics cover problems which only become relevant if we want to examine subpopulations. For example, without good data about failure modes, we can hardly identify the reliability curve for a particular one such as "explosion". Subfleet metrics should be treated as critical metrics once we actually focus on a particular sub-population but only require action if such calculations are desired.

"Added Value" metrics cover information that was included but is not directly used in neither the calculation nor the creation of subpopulations. However, such information might be useful to a human scanning the data for some reason, e.g. to identify a unit for further investigation.

The category of *"Other/Unspecified" metrics* only exists for the purpose of "debugging" as it contains any metric that was not assigned to a specific category. In a correctly designed metrics hierarchy, this category is empty.

5.2 Hierarchical approach based on box-plots

In order to quickly track down key problems, we suggest a hierarchy of criticality (see Section 5.1), category (see Section 4), and individual metrics. We represent each level as a boxplot of all metrics in each category.

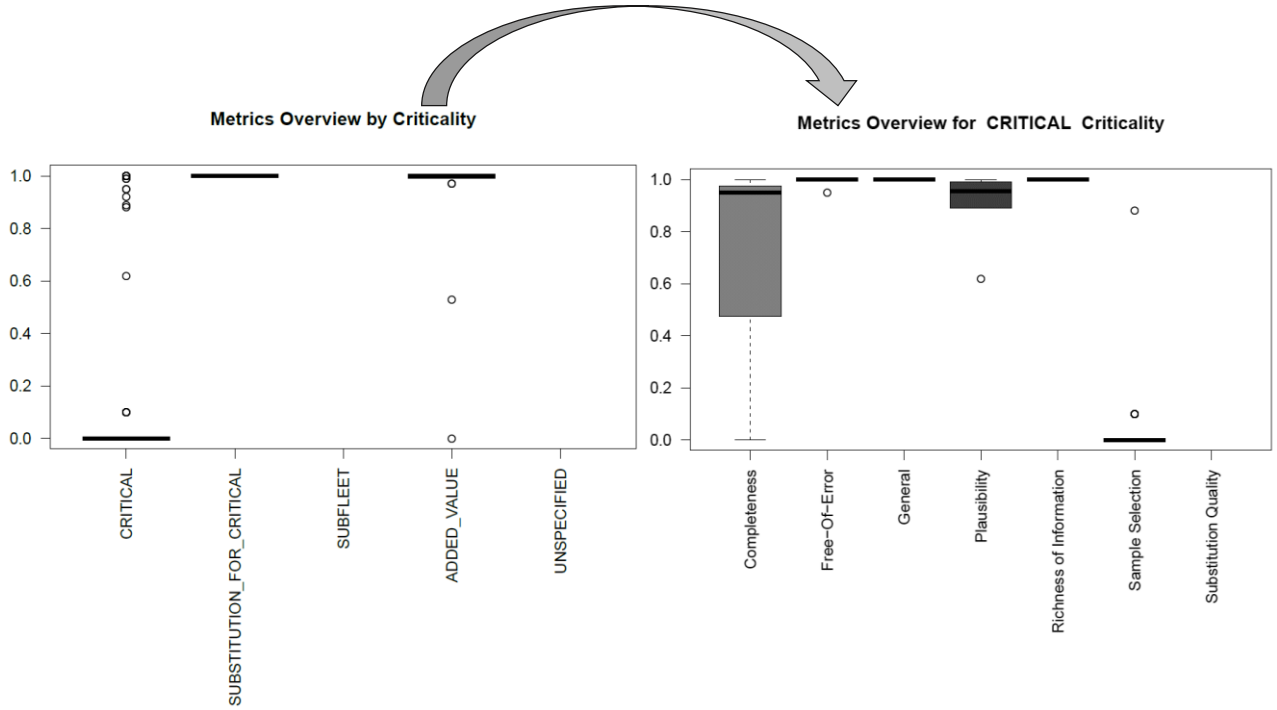

Figure 1: Level 1 and Level 2 Boxplots

Consider the anonymized example shown in Figure 1. The chart on the left shows all metrics, grouped by criticality. (In the example, there are no subfleet and unspecified metrics, so these columns are empty.) We can see at a glance, that the average critical metric is quite poor (as indicated by the bar) with a few metrics being quite good (the outlier dots). If we drill down into a view of all CRITICAL metrics (shown on the right), we find that the key problems are associated with the sample selection and the completeness, with most other metrics being quite good. Sample selection is quite poor on average, completeness metrics on the other hand cover a wide range as indicated by the box plot and its rather long whisker.

A closer investigation of all critical sample selection metrics brings us to the level of individual metrics for a particular category and criticality (see Figure 2). As we can quickly discern, the main problem in this scenario is that for each of the product types the sample size is too small (left). Completeness of data is quite good except for the missing disposal date which is missing for all cases (right).

5.3 Practicality of use

As can be seen by the example, the hierarchy of data quality metrics can be used to quickly identify the key problems in a data set. However, each data set has its unique problems. We have therefore developed a software framework, which can be used to calculate the metrics for a particular data set. Including the configuration of the metrics, it took us two person days to implement a set of metrics tailored to the particular problem. (This does not include the initial assembly of the data.) Using several iterations, we were able to fix multiple problems in the data up to the quality level shown in the example graphics. For example, in the first go, many of the product types where missing, leading to a poor performance of the critical completeness metric.

6 Conclusions and future work

In this paper, we have presented a hierarchy of metrics which allows the quick identification to data quality problems in a data set used for survival analysis. The metrics and their classification are based on several test data sets we had available in our company. The metrics have been implemented in a prototypical API based on Java and R which can be used to quickly program report generators for a particular data set.

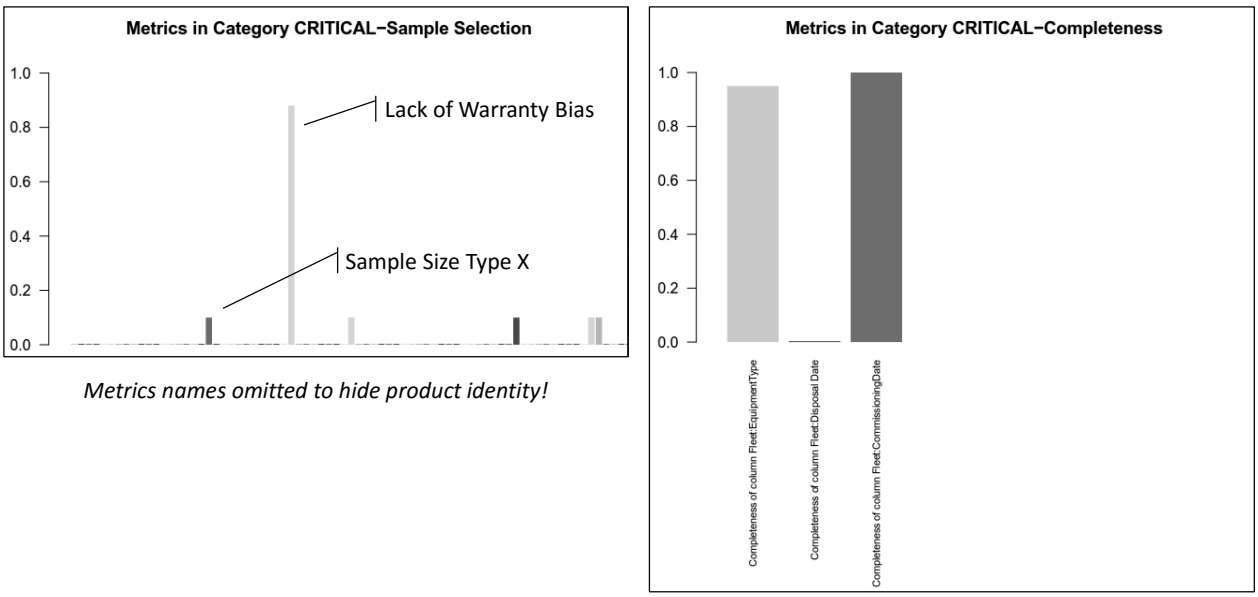

Figure 2: Level 3 Showing Individual Metrics

We plan to further refine our metrics by gaining a better understanding of the impact the data quality has on the results. Currently, the metrics measure the extent of the problem within the data (e.g. how many properties are empty) but not the impact (by how much is the reliability curve distorted by this). A low hanging fruit in this context is to weave the estimated power of the test into the sample selection metrics.

7 References

Gitzel, R. (2014) *Industrial Services Analytics*. Presentation at the 1. GOR Analytics Tagung, Munich, 11.4.2014

Gitzel, R., Turring, S., Maczey, S. (2015, July). A Data Quality Dashboard for Reliability Data. In *2015 IEEE 17th Conference on Business Informatics (CBI*), Vol. 1, pp. 90-97.

Kahn, BK.; Strong, Diane M.; Wang, Richard Y. (2002): Information quality benchmarks: product and service performance. *Communications of the ACM* 45 (4), pp. 184–192

Ballou, D.P., Pazer, H.L. (1985): Modeling data and process quality in multi-input, multi-output information systems. Management Science 31, 2, (1985), 150–162.

Borek, A.; Parlikad, AK.; Webb, J.; Woodall, P. (2014): *Total information risk management – maximizing the value of data and information assets*. Morgan Kaufmann

Benson, P. (2008): ISO 8000 the International Standard for Data Quality, in *MIT Information Quality Symposium*, July 16-17, 2008

Ban, X., Ning, S., Xu, X.; Cheng, P. (2008): Novel method for the evaluation of data quality based on fuzzy control. *Journal of Systems Engineering and Electronics*, 19 (3), 606–610.

IEEE (2007): IEEE Standard 493 - IEEE Recommended Practice for the Design of Reliable Industrial and Commercial Power Systems.

Vadlamani, R. (2007): Modified Great Deluge Algorithm versus Other Metaheuristics in Reliability Optimization, *Computational Intelligence in Reliability Engineering, Studies in Computational Intelligence* **40**, 2007, 21-36

Bendell, T. (1988): An overview of collection, analysis, and application of reliability data in the process industries. *IEEE Transactions on Reliability* **37**(2), 132–137.

Montgomery, N.; Hodkiewicz, M. (2014): Data Fitness for Purpose. In: *Proceedings of the MIMAR 2014 Conference*, Oxford, UK.

Wu, S. (2013): A review on coarse warranty data and analysis, *Reliability Engineering & System Safety*, **114**, 1-11

Alam, M.M.; Suzuki, K. (2009): Lifetime Estimation Using Only Failure Information From Warranty Database, *IEEE Transactions on Reliability*, **58**(4), 573--582