Cleaning of Railway Track Measurement Data for Better Maintenance Decisions

Bjarne Bergquist
Quality Technology
Luleå University of Technology, Luleå, Sweden
Phone: +46 920 49 2137
bjarne@ltu.se

Peter Söderholm
Trafikverket and Quality Technology
Luleå University of Technology, Luleå, Sweden
Phone: +46 10 123 81 67
peter.soderholm@trafikverket.se

Osmo Kauppila
Industrial Engineering and Management
University of Oulu, Finland
Phone +358 40 825 7692
osmo.kauppila@oulu.fi

Erik Vanhatalo
Quality Technology
Luleå University of Technology, Luleå, Sweden
Phone: +46 49 1720
erik.vanhatalo@ltu.se

ABSTRACT

Data of sufficient quality, quantity and validity constitute a sometimes overlooked basis for eMaintenance. Missing data, heterogeneous data types, calibration problems, or non-standard distributions are common issues of operation and maintenance data. Railway track geometry data used for maintenance planning exhibit all the above issues. They also have unique features stemming from their collection by measurement cars running along the railway network. As the track is a linear asset, measured geometry data need to be precisely located to be useful. However, since the sensors on the measurement car are moving along the track, the observations' geographical sampling positions come with uncertainty. Another issue is that different seasons and other time restrictions (e.g. related to the timetable) prohibit regular sampling. Hence, prognostics related to remaining useful life (RUL) are challenging since most forecasting methods require a fixed sampling frequency.

This paper discusses methods for data cleaning, data condensation and data extraction from large datasets collected by measurement cars. We discuss missing data replacement, dealing with autocorrelation or cross-correlation, and consequences of not fulfilling methodological pre-conditions such as estimating probabilities of failures using data that do not follow the assumed distributions or data that are dependent. We also discuss outlier detection, dealing with data coming from multiple distributions, of unknown calibrations and other issues seen in railway track geometry data. We also discuss the consequences of not addressing or mishandling quality issues of such data.

Keywords

Track geometry, big data, railway, data quality, diagnostics, prognostics, maintenance, Sweden.

1. INTRODUCTION

The amount of asset condition data, as well as its availability for both practitioners and scientists, continues to grow. The eMaintenance concept has grown along, helping to solve hitherto unsolvable maintenance problems. The rapid increase in collected asset condition data is due to new possibilities made available by digitisation and accelerated technological development.

However, data do not serve a particular purpose in itself. Data need to be put into a context-dependent purpose, and issues such as the required levels of detail and aggregation depend on that purpose [1]. Quality data need to be "fit for purpose" [2]. As this fit depends heavily on the context, there is no single set of agreed dimensions for data quality. Accuracy, completeness, consistency and timeliness form one of the most frequently used sets [3].

The massive data streams come with associated challenges, e.g., in the management of big data due to its inherent properties: volume, variety, velocity, veracity, and value. For example, pre-processing activities to convert field data into a format compatible with the intended data analysis may consume the most analysis time. Issues such as missing data, heterogeneous data types, calibration problems, or non-normality often surface when analysts try to turn datasets related to operations and maintenance of technical systems into the desired format. Additionally, railway track geometry data, which we analyse in this paper, have unique features stemming from their collection method. Since the sensors on the measurement car are moving along the track, there is uncertainty in the geographical sampling position of the observations. The sampling intervals are affected by seasonal and other restrictions, and the irregular sampling intervals can be problematic in condition forecasting for maintenance purposes, see, e.g. Bergquist and Söderholm [4, 5].

In this paper, we study how railway geometry data can be processed to make them fit for prediction and maintenance planning. We investigate data cleaning, aggregation and extraction of information. Issues that we address include missing data, auto- and cross-correlation and the data not meeting requirements such as distributional and independence assumptions, as well as their consequences. We also investigate outlier detection, handling data from multiple distributions, calibration issues. Finally, we discuss the implications of not addressing data quality issues of track geometry data. Data and examples in this paper are based on measurements obtained on track section 119, which is part of the Swedish Iron ore line, and it connects the cities of Boden and Luleå by 35 kilometres of track.

2. RAILWAY TRACK DATA

Track measurement cars record the railway track measurement data that we will discuss in this paper, that is, measurement trains and trollies that regularly travel along the Swedish railway network to measure characteristics of different parts of the infrastructure. Both trollies and measurement trains measure several geometrical properties of the track, substructure and catenary system. These measurements can be used to analyse deviations from the designated geometry. The measurement train (IMV 200) consists of an engine and a measurement car. The measurement car obtains measurements through accelerometers mounted in the car body and linear variable differential transformers to relate the position of the car body to the axles. Each 5 cm of the track length is measured, but these data are post-processed into observations taken 25 cm apart before they are uploaded to the database. Track geometry measurements include track gauge, cross-level/cant, twist, and vertical and side alignment of the two rails.

3. DATA BINNING

The supplier regularly uploads measurement data from the measurement cars to the decision support system Optram [6]. The Optram system allows data exports for further analyses through comma separated files (.csv). One full run of the measurement train for track section 119 equates a .csv file with a size of around 330 MB. For some purposes require such high-resolution data. For many other purposes, such large datasets may become too bulky, such as when several measurement occasions are to be combined. There are also other reasons, elaborated later in the paper, to replace the 25 cm observations with other measures by binning the data into representative summary statistics. We have binned the data into 200 m track segments. In practice, this means replacing 800 observations of the 25 cm resolution (or 4 000 of the original 5 cm observations) by summary statistics for each measurement and segment. Examples of summary statistics include the maximum value or the standard deviation of a particular property within the track segment. The binning was performed in the Microsoft Power BI Desktop® software. Some summary statistics such as the average will also improve later analyses since the distribution of the average will be closer to the normal distribution due to the central limit theorem. The paper will, without loss of generality, from this point use the binned data for 200 m track segments. Any faults and peculiarities found in the binned data would be valid also for the original observations. However, in some cases, the binning procedure will hide outliers and other problems visible in the 25 or 5 cm data.

4. DATA OVERVIEW

Probably the best first step in finding data peculiarities and outliers, real or not, is to plot the data. Many analysis software allows for plotting several variables in a matrix of bivariate scatterplots. Such a matrix plot is useful since strange patterns as well as the correlation between variables become apparent and may not be evident in univariate plots. Bivariate plots produced one-by-one will time-consuming to produce if the data-set includes many variables. Figure 1 shows a matrix of bivariate scatter-plots of the largest obtained measurements of the variables in each segment. The software we use for all plots in this paper is JMP® version 14.1.0. The variables that are plotted in Figure 1 appear in the following order (the maximum values of): Twist (6m base), Twist (3m base), Side shortwave amplitude (right rail), Side shortwave amplitude (left rail), Height shortwave amplitude (right rail), Height shortwave amplitude (left rail), and Gauge. These data were obtained from 103 passes of the measurement cars on track section 119 between April in 2007 and February 2019.

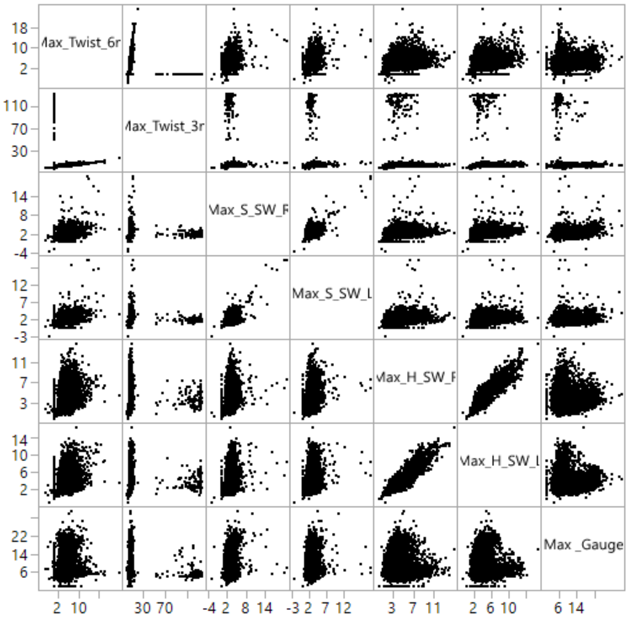


Figure 1. Matrix of scatter plots of maximum values on the seven variables.

Many methods for multivariate data, base the calculations on the assumption that the data follow a multivariate normal distribution. Such ideal data would display point swarms in the bi-variate scatter plots that are either circular or oval along a diagonal. An oval shape would indicate a positive or negative correlation between the two variables and a circular shape would indicate weak or no correlation. Patterns deviating from this expected behaviour indicate issues that the analyst should handle or at least consider the consequences of, before further analyses. One peculiarity in Figure 1 (e.g. first and second row) is that the data seem to separate into two groups for the twist variables. Figure 2 presents one of these bivariate scatterplots between the two twist measures for increased readability (row 2, column 1 in Figure 1). Note that the observations are maximum twist errors obtained for a 200 m segment. Any zero values would be an indication of measurement problems, and likewise, a negative value would indicate negative twist readings for a full 200 segment, which is not realistic as twist needs to sum to zero over a point defect or else there will be a constant lean of the track after the defect. A series of positive ones must thus follow a series of negative readings. Data cleaning therefore started by removal of 'strange' twist measurement observations before proceeding to further analyses.

Figure 2 shows that the 3 m twist maximum also seems to be split into two groups. The raw data for the 3 m twist maximum thus contained data from two distributions. Further analysis revealed that measurements of the measurement car for two dates (passages) had an average a hundred times lower than the remaining 101 indicating some measurement issue. Since these erroneous data made up only a small portion of the data, all zero values or lower on the 6 m twist maximum observations, and all 3 m twist maximum observations lower than 20 mm/m were removed from further analyses.

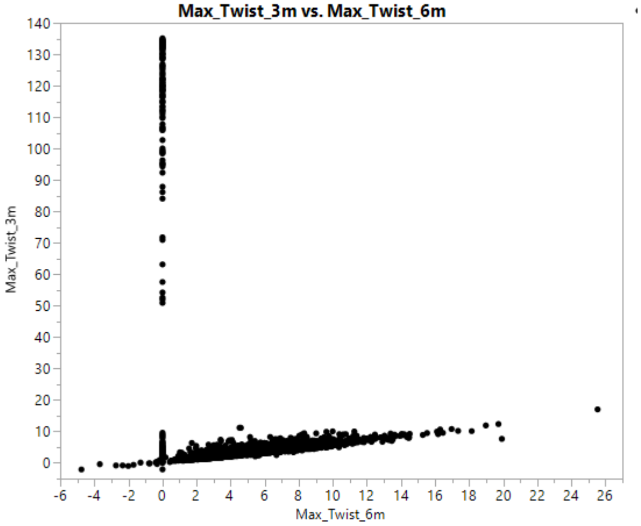


Figure 2. Bi-variate scatterplot between the maximum of the 6 m twist [mm/m] and the 3 m twist [mm/m].

Depending on the analysis software, the analysis to be done at the later stage, and the wealth of data, the analyst may select to keep or delete the observation. If the analyst chooses to keep the observation, with just deleting the faulty variable measurements, the other variables' measurements may still be valid and strengthen analyses where those variables are important. Some software and analyses do not tolerate incomplete datasets, while others perform well if the missing observations are not systematic or too many. Some of these use data imputation methods. Again, imputation methods still require caution and cannot reconstruct cases where there is a systematic structure to the missing data, or when there are too few remaining observations.

5. DISTRIBUTIONAL PROPERTIES

It is also good practice to study the variables in a univariate plot that reveals information about their distributions, such as a histogram. Data are never ideally normally distributed, which is a common assumption for many analysis methods in later stages. Often, the normal distribution describes the data reasonably well; other times some properly chosen transformation can make the transformed align better with the normal distribution. Without such transformations, inference based on a normal probability assumption, for instance, estimating the probability that a track segment would have a maximum twist above a certain limit, would be unreliable.

For example, the appearance of the histogram in Figure 3 with a long right tail suggests that a lognormal distribution would be similar and could approximate to the observed distribution. The lognormal distribution is reasonable, given that the shortwave has a natural zero limit. A log transformation often improves the maximum and the standard deviation, but the analyst is always well advised to study the transformed data. Outliers may, for instance, become more clear after a transformation, so we recommend revisiting univariate and matrix plots as in Figure 1 after transformations to identify and remove potential additional 'bad' observations. The variable in Figure 3 has undergone a log transformation, and Figure 4 shows the transformed data. The transformed data shows a few suspected outliers to the left that need more careful inspection, but the transformation did prove useful in making the variable more normally distributed.

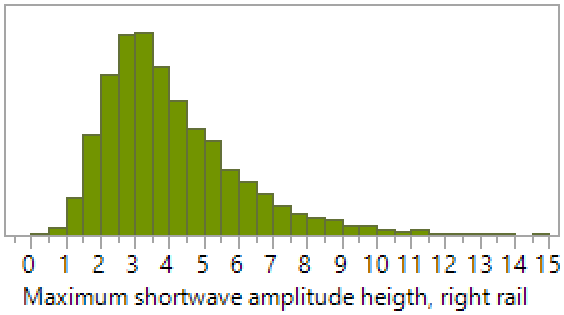


Figure 3. Histogram of the maximum amplitude of the shortwave of the height of the right rail [mm].

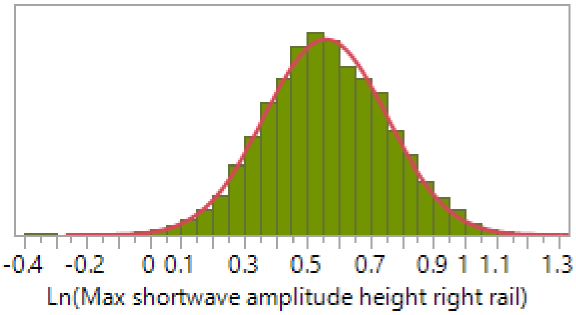


Figure 4. Histogram of log-transformed maximum amplitude of the shortwave of the height of the right rail. The red curve shows the curve fitting of a theoretical normal distribution.

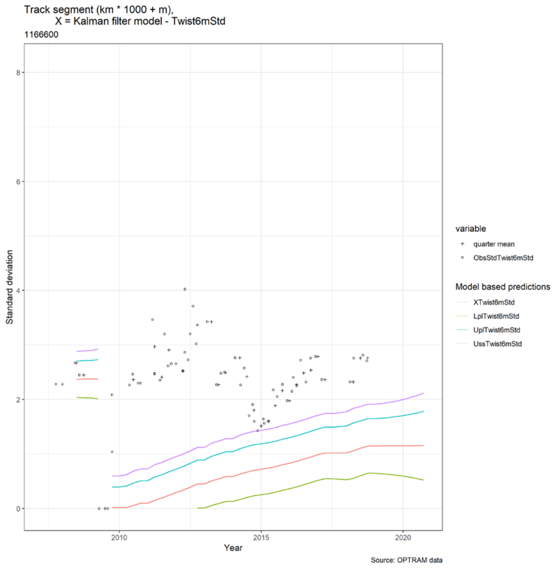
The log transformation belongs to a set of standard variance stabilising transformations. It is often the case that there is a multiplicative relation between the expected average value of a property and its variance. When the variable variation connects multiplicatively to the the average, that connection will induce analysis problems if not addressed. For multiplicative relations, the log transformation is a standard transformation, converting the multiplicative relation to an additive one. There are other standard transformations too. The standard deviation is χ^2 distributed, with a long right tail, and the log transformation will make the distribution of the variable closer to the normal. Other standard transformations include the square root for Poisson distributed data or the sine for binomially distributed data.

Box and Cox [7] suggested finding an appropriate power transformation by empirical testing. However, we do not recommend using the 'best' power transformation that the Box-Cox test produces without further consideration. Plotting of the data may, for instance, reveal that the reasons for the best fit were outliers or data stemming from multiple populations such as in Figure 2. Standard variance stabilising transformations are found in [8].

6. OUTLIER DETECTION

Outliers may lead to faulty conclusions if they are erroneous, but may also reveal relevant information. One should, therefore, not remove them carelessly. Figure 5 shows observations and quarterly averages of the segment max of the 6 m twist variable. Two consecutive observations have generated zero variance readings due to a fault of the measurement train. The figure also shows a model that tries to fit the data for prediction purposes, along with a 95% prediction interval and a three sigma upper prediction limit. The model includes some tolerance toward noisy input data, but the model is designed to restart if conditions have improved considerably. Such improvements are not natural but would be the results of maintenance actions. The model was also made robust versus outliers in that unreasonable condition decline was not allowed. The standard deviation may be more substantial for some measurements during spring thaw or autumn frost heave, but the variation amplitudes will shrink when that period has passed. The model, therefore, had a maximum tolerated increase rate of the standard deviation for automatic outlier removal due to frost.

As illustrated in Figure 5, the effect on the model of the two outliers in combination with the model's maximum tolerated increase rate of the standard deviation can be detrimental for the results. The model, based on the average of quarterly measurements, tries to restart due to the outliers, but the limit for the maximum tolerated standard deviation increase keeps the model from adjusting back to the real data. In this case, removal of zero standard deviation values would be a simple solution, but other outliers, for instance, due to instability due to frost heave or thaw will also need handling or a model that is programmed to disregard strange measurements obtained from such times. A difficulty for an automated model solution there is that frost depends on geography and altitude among other things.



Source: OPTRAM data

Figure 5. Observations (o) and quarter averages (+) of the standard deviation of the 6 m twist variable for one segment.

7. FAULT LOCALISATION

The coordination of fault diagnosis (i.e. fault detection, fault localisation and cause identification) at different maintenance echelons, in time, and space is central for eMaintenance. The result of an erroneous fault diagnosis on single maintenance echelons can be false alarms or undetected faults, while insufficient coordination of different maintenance echelons can lead to No Fault Found events. One example is when dealing with linear assets and combining the use of different measurement methods for fault diagnosis, e.g. measurement trains that identifies a failure that later on should be corrected by maintenance personnel that use manual inspection. Maintenance personnel may need to wait for a passing train to pinpoint the issues if, for instance, the fault localisation requires that the track is loaded to be visible. A precise geographical localisation is, therefore, essential from the standpoint of knowing where to locate track failures when faults are not evident from visual inspections by the maintenance crew.

The quality of performed fault diagnostics, on single or multiple maintenance echelons, at different occasions in time will also determine the power of fault prognostics. The time dimension of the data is crucial to establish causal relationships between measurements and maintenance actions. Both time and position are crucial for following deterioration progress over time and make predictions, e.g. regarding the remaining useful life.

Measurement trains possess the property that the instrument is travelling along the measured object, rather than being attached to it at a fixed spot. The lack of fixation means that the trains can assess the condition of assets such as a railway network, with measurement update frequencies set by the speed capacity and accessibility of the measurement train and staff, as well as its access to the track. High-speed measurement trains with top speeds of 400 km/h are in use [9], but significantly substantial assets can also be measured relatively often using the fastest Swedish measurement trains with top speeds of 200 km/h. Compared to fixed instruments, the length of track possible to survey is vast. Whereas equipment costs and costs for data connectivity are rapidly decreasing, an array of instruments attached to the railway would currently generate insurmountable costs for regular railways. Another benefit of using measurement trains, instead of fixed sensors, is that the trains are likely as inaccurate for the whole measurement sequence.

An array of stationary instruments would need to be calibrated using the same scales so that data from one position would be comparable to data from another. A downside of the moving instrument is that the instrument (the measurement train) relies on other measurements to connect the measured properties with a place along the track. Localisation would seem like a small obstacle, but in reality, it can be a significant problem. Figure 6 shows to measurements of a track twist point defect. The thirteen measurements were taken between April 2007 and June 2010. As it happens, the one (green) measurement placing the fault in the left-most position is also the first measurement, and the measurement that is placing the fault most to the right in the figure is the latest. The measurements are likely made by a now-retired measurement train (STRIX) although the data lacks the measurement train information for the oldest measurements. The fault localisation differs by 25 m.

Figure 7 shows the last measurement from Figure 6 taken by the STRIX measurement train (the curve that has increased most to the right, Figure 6) and the next measurement taken on the line.

Another measurement train, the IMV 100N measurement trolley, had obtained this measurement, measured at a later date.

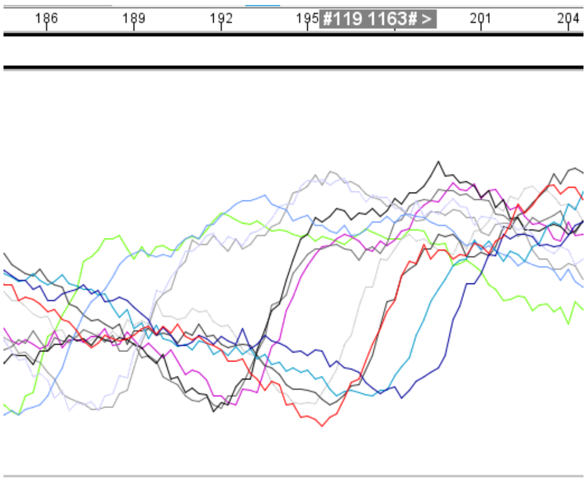


Figure 6. Varying localisation of a track twist (3m) point defect. Measurements between April 2007 and June 2010.

The method of using pattern matching and then use an average location for all pattern matched measurements has its merits. The matching would be a fairly straightforward process if data looked like in Figure 6. However, we presume that different post-processing and different measurement equipment makes a pattern matching calibration process complicated. Indeed, attempts to align measurements faces significant problems. The studied data also display what appears to be a systematic trend, placing the fault longer and longer westward (toward Luleå). The measurement train location was not based on GPS at this time although the measurement trains had such instruments installed. Instead, the positioning was produced using dead reckoning from known locations. Likely, wheel wear used for the dead reckoning is the culprit of the systematic localisation error trend.

Calibration of the position using known obstacles that would give a known pattern in the data, for instance, a switch, or other objects of known locations such as ballises could work to anchor the measurements to specific locations. This method is also used with some regularity already in the pre-processing before the upload to the Optram database. If the measurement data included GPS locations, localisation could also be improved using map-matching, see for instance [10].

For some purposes, such as for regular maintenance, the exact location of point defects are not as important as monitoring the decay of a more extended segment, since tamping machines require a certain tamp length to be efficient. By binning data into segments, the chances are fair that troublesome lengths of the track will remain within the segment when bins from different measurements are compared. Point defects located on the border of two segments will affect the variability of both segments, but in reality, the time series seem robust versus such difficulties, perhaps due to the slow drift of the localisation error as seen in Figure 6.

Another difficulty of the positioning is that the track kilometres that are used to locate objects in the one-dimensional space of the track are not always one kilometre in reality. The track location along the Swedish railway network uses the km-m system, which relates to how far any location is from the Stockholm station. Reinvestments may change the design of the track so a particular track kilometre may, therefore, be longer or shorter than 1000 m. The analyst must be aware of this discrepancy and have a plan for how to handle kilometre differences for the binning procedure.

8. IRREGULAR SAMPLING

Since the operation of measurement trains are affected by, e.g. staff and train schedules, weather conditions, vacations, measurement train repairs, the measurements are not obtained at equal sampling intervals. The Swedish Transport Administration classifies the tracks into inspection classes according to a combination of the maximum allowed speed and total yearly tonnage. These inspection classes regulate both the allowed geometrical tolerances of the track, but also how often the tracks need to be measured by the trains. Often, a significant purpose of eMaintenance is prognostics and prediction of remaining useful life, i.e. when can we expect that a particular asset needs maintenance? A typical analysis method for prognoses is to use time series modelling. However, time series methods assume that data are sampled at regular intervals. Interpolation methods can be used to overcome the irregularity issue in the sampling. The modelling can then use these interpolated values. Linear interpolation, using the last measurements as inputs or some splines models have been used for this purpose, see also [4]. Interpolation may underestimate the variation in the data or produce other unexpected results, so again, we advise caution and plotting before drawing conclusions based on such procedures.

9. REPRODUCIBILITY

Measurements taken repeatedly can be used to estimate the measurement reproducibility. By measurement reproducibility, we mean how close two or more measurement results obtained with the same equipment and the same operators on the same object. Another condition is that the measurements are performed too far apart in time. For measurement trains, sack-stations on single track lines mean that trains need to turn and travel in the other direction. For the Boden-Luleå case, the trains often return the next day. The expected real differences between such measurements can be considered negligible compared to the measurement noise. Hence, one can use these close measurements for reproducibility measurements for the specific trains. The procedures have most often been to measure the main track when travelling from Boden to Luleå, and measure side-tracks for the return trip, and not report measurements from the main track. However, once in a while, both trips are measured, and those measurements can be used for estimating the reproducibility uncertainty, see also [5].

Faulty calibrations will affect both measurements, and such faults will not be found using such comparisons. Calibration errors and differing measurement precisions between trains are important if the prognostic models need measurements from all trains since many prognoses methods rely on historical data. Note, however, that seeking differences between different measurement trains usually require longer waits between measurements, which increases the background noise for such calculations.

10. DISCUSSION AND CONCLUSION

In this paper, we discuss cleaning or pre-processing of track geometry measurement data used for maintenance planning in the Swedish railway network. We outline important steps and problems in the data pre-processing needed before proceeding to other advanced analysis methods. We also illustrate some real data examples from track section 119 that connects the cities of Boden and Luleå by 35 kilometres of track. The data cleaning steps involve some knowledge and understanding of the measurement systems as well as the measured properties, e.g., to erase outliers, to select proper transformations or to use binning and select suitable binning sizes. While such knowledge often is critical for generating valuable analysis results, we hope that we have described the procedure in such a way that the analyst will understand when subject-matter knowledge is needed, and when general data handling practices may be used.

We have summarised some of the challenges and available solutions Table 1.

Table 1. Examples of challenges and potential solutions in data cleaning.

Challenge	Solution
Skewness	Transformation
Non-normality, but symmetric	Empirical distributions with percentiles balancing alpha and beta risks
Dependence	Increase sampling interval Fit model and use residuals Adjust limits based on empirical percentiles
Positioning error	Data binning and use of distribution measures for distances
Uneven sampling intervals	Interpolation
Uneven sample sizes	Inter-measurement alignment and missing data treatment

Turning to the studied case, the geometry data together with inspection data and other data relevant for maintenance planning exhibit at least four out of the 5Vs that constitute challenges for big data processing: volume, variety, (velocity), veracity, and value, see, e.g. [11, 12]. The main issues that the analyst needs to handle in the pre-processing of track geometry data are many, and countermeasures involve many steps. These steps may comprise of data binning (e.g., into 200 m segments), data overview and outlier identification and handling (e.g. by use of univariate, bivariate and multivariate approaches), The steps may also include variable transformations, handling of spatial and temporal localisation issues (e.g. binning, pattern matching, and point asset fitting). The irregular sampling frequency may need handling (e.g. by using time series modelling in combination with models for interpolation and extrapolation): Finally, reproducibility issues may need attention (e.g. identification of measurements that one can treat as repeated measures). In this paper, we provide examples of how these issues can be handled to clean data for future use in enhanced diagnostic and prognostic models important for eMaintenance applications. The value of enhanced diagnostics on single and multiple maintenance echelons can be measured by reduced false alarm rates, lower degree of undetected faults, improved fault localisation, and a reduced number of No Fault Found events. Additionally, enhanced prognostics provides improved remaining useful life estimation valuable in maintenance planning.

ACKNOWLEDGEMENTS

We would extend our thanks to Trafikverket (the Swedish transport administration) and Vinnova (the Swedish Innovation Agency) for financial support for the study.

11. REFERENCES

[1] Gustavsson, M. & Jonsson, P. (2008). Perceived quality deficiencies of demand information and their consequences. *International Journal of Logistics: Research and Applications*, 11(4), 295-312.

[2] Otto, B., Hüner, K.M. & Österle, H. (2011). Toward a Functional Reference for Master Data Quality Management. *Information Systems and e-Business Management*, 10, 395-425.

[3] Batini, C., Cappiello, C., Francalanci, C. & Maurino, A. 2009. Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, 41 (3).

[4] Bergquist, B. & Söderholm, P. (2016). Predictive Modelling for Estimation of Railway Track Degradation, edited by Kumar, U. Ahmadi, A., Verma, A.K., Varde, P. in "*Current Trends in Reliability, Availability, Maintainability and Safety*", Springer International Publishing AG, Cham., 331-347.

[5] Bergquist, B. & Söderholm, P. (2016). Measurement System Analysis of Railway Track Geometry Data using Secondary Data Analyses. *eMaintenance 2016*. Luleå. June 14-16.

[6] Smith, A. (2016). *Optram Analysis and Forecasting: A Bentley White Paper*. Accessed on-line 2019-02-22 on https://www.bentley.com/

[7] Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211-243.

[8] Box, G. E., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for Experimenters*. John Wiley & Sons, New York, NY.

[9] Nielsen, J., Berggren, E., Lölgen, T., & Müller, R. (2013). Overview of methods for measurement of track irregularities. RIVAS Railway Induced Vibration Abatement Solutions Collaborative Project.

[10] Känsälä, K., Rantala, S., Kauppila, O. & Leviäkangas, P. (2018). Acceleration sensor technology for rail track asset condition monitoring. *Proceedings of the Institution of Civil Engineers - Management, Procurement and Law*, 171(1), 32-40.

[11] Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013). Big data challenge: a data management perspective. *Frontiers of Computer Science*, 7(2), 157-164.

[12] Demchenko, Y., De Laat, C., & Membrey, P. (2014). Defining architecture components of the Big Data Ecosystem. In *"2014 International Conference on Collaboration Technologies and Systems (CTS)"*, IEEE, 104 - 112.