infrastructures

MDPI

Article

Enhancing Predictive Maintenance Through Detection of Unrecorded Track Work

Jan Schatzl *, Florian Gerhold, Markus Loidolt and Stefan Marschnig

Institute of Railway Engineering and Transport Economy, Graz University of Technology, 8010 Graz, Austria; florian.gerhold@tugraz.at (F.G.); markus.loidolt@tugraz.at (M.L.); stefan.marschnig@tugraz.at (S.M.)
* Correspondence: jan.schatzl@tugraz.at; Tel.: +43-316-873-6716

Abstract: Predictive maintenance can help infrastructure managers to reduce costs and improve railway availability while ensuring safety. However, its accuracy depends on reliable data from various sources, especially track measurement data. When analysing track data over time, historical maintenance actions must be considered, as otherwise the interpretation of the data would be misleading. This research aims to address inconsistencies in recorded maintenance data by detecting unrecorded track works through track geometry evaluations. The main goal is to provide the foundations for accurate descriptions of track behaviour, supporting the implementation of effective predictive maintenance regimes. As part of the research, three different approaches are analysed and evaluated, whereby two of them are based on cross-sectional analyses and the third one detects track works in longitudinal track dimension. The results show that the CRAB algorithm produces the most statistically significant results. Conversely, the cumulative track geometry-based algorithm provides a homogeneous representation of past maintenance work and a result that is statistically only marginally inferior. Consequently, these two methods are best suited to build the foundation for making accurate cross-sectional conclusions about track geometry behaviour. This allows for the verification and enhancement of existing maintenance databases.

Keywords: railways; data analysis; detection of maintenance; track behaviour; predictive maintenance

Citation: Schatzl, J.; Gerhold, F.; Loidolt, M.; Marschnig, S. Enhancing Predictive Maintenance Through Detection of Unrecorded Track Work. *Infrastructures* **2024**, *9*, 204. https://doi.org/10.3390/infrastructures9110204

Academic Editor: Giuseppe Cantisani

Received: 16 October 2024
Revised: 11 November 2024
Accepted: 13 November 2024
Published: 16 November 2024

1. Introduction

The rapidly developing field of predictive maintenance is concerned with ensuring the reliability of infrastructure systems such as railway tracks. In the railway sector, predictive maintenance aims to anticipate track failures, predict the time to an exceedance of pre-defined threshold values, and plan maintenance measures. That helps to minimise operational disruptions, reduce maintenance costs, and preserve the requested quality [1]. This approach relies heavily on accurate and comprehensive data, including both measurement data for modelling and metadata such as asset information and maintenance records. However, the effectiveness of predictive maintenance systems is often compromised by incomplete or inaccurate metadata, particularly when maintenance activities are not recorded or incorrectly documented.

In the railway sector, maintenance planning is often empirical, based on the knowledge and experience of regional infrastructure managers. This knowledge is a highly valuable asset that, in the best case, is combined with data-driven predictive maintenance approaches. This requires prediction of track behaviour based on descriptive models. The detection of unrecorded or misrepresented maintenance activities is a critical challenge in this context. If maintenance activities are omitted from the data warehouse, the predictive models are fed with incomplete data, leading to incorrect degradation rate calculations and ultimately incorrect maintenance plans. This can result in either over-maintenance, which is costly, or under-maintenance, which compromises safety and reliability [2].

The deterioration of track quality is interrupted by maintenance work, as maintenance leads to a sudden enhancement of quality. Deterioration branches, being the foundation

of descriptive models, must therefore be described between two maintenance measures. Unavailable or incorrect maintenance data end in misleading descriptive models and are discussed in several publications dealing with predictive maintenance models, though they are not always considered. This is exemplified by the works of Wen et al. [3], Andrews et al. [4], Guler et al. [5], and Jovanovic et al. [6], who have incorporated maintenance data into their degradation models and maintenance planning yet have not discussed the potential impact of unrecorded or incorrect data.

In contrast, Caetano et al. state that in their research a model for detecting changes in geometry parameters, which is not further described, is used [7]. Furthermore, the predictive maintenance model established is based on fixed 200 m sections, which only permit the identification of the occurrence of a maintenance activity within the section, but not its precise location. Audley and Andrews also coped with incomplete maintenance and renewal records that did not meet the requirements necessary for developing track degradation models [8]. Consequently, they employed an algorithm to ascertain the precise date of maintenance and renewal activities, not merely the correct year. The algorithm combines existing maintenance and renewal records with track measurement data, thereby enabling the precise determination of the time at which the measure was undertaken within a 220-yard section.

Sedghi et al. identified the need for an empirical determination of unrecorded maintenance data for building a stochastic data-driven decision-support framework that integrates automated prediction of track geometry degradation with planning and scheduling optimisation modelling [9]. It was determined that an ad-hoc improvement of the standard deviation of the longitudinal level by 15% indicates an executed maintenance activity in a fixed 200 m section.

Neuhold et al.'s maintenance data was also incomplete, leading to the necessity of an algorithm to detect undocumented maintenance and renewal activities [10]. In their algorithm, they employ a combination of outlier detection based on data points from the modified standard deviation of the longitudinal level and a comparison of successive quality indices. They state that if the quality between two data points increases, which means that the modified standard deviation of the longitudinal level decreases and the data point cannot be classified as an outlier, then a maintenance or renewal action was executed between those two points in time.

One of the objectives of Fellinger's thesis was to ascertain the effect of tamping on the standard deviation of the longitudinal level in turnouts for setting up a prediction model of turnout behaviour [11]. As the maintenance record was incomplete, he decided to complete the input data for a dozen of turnouts by conducting exhaustive research and having discussions with regional managers. Based on this accurate input data, he constructed a model for detecting past maintenance activities. He uses a linear model with a prediction interval to forecast the standard deviation of the longitudinal level of every measurement run based on the preceding values. Maintenance is then detected with a probability in dependence on the prediction interval if the standard deviation of the longitudinal level was lower than the lower limit of the prediction interval. As with all of the aforementioned algorithms, this is a cross-section-based maintenance detection. Fellinger's method also forms the foundation for one algorithm described later in this paper.

The literature indicates a lack of knowledge with regard to reliable methods for the detection of performed, ballast-related maintenance measures and an objective performance evaluation of the methods. Only a few publications address the quality of maintenance data and the problem of undocumented actions on the correctness of analytical models. Most researchers deal with missing information by making well-founded manual data adjustments, but this is not possible for large data sets. Also, the majority of researchers employ the improvement of quality as a criterion for determining the efficacy of track work. However, this approach is not always sufficient, as improvements in quality within a time series can also be attributed to the presence of poorly synchronised data or data errors. This paper addresses the issue by proposing and comparing three methods for detecting

performed, ballast-related maintenance actions based on signal characteristics. By accurately identifying these unrecorded actions, the proposed methods allow for the calculation of more reliable degradation rates, enabling more accurate and stable maintenance plans. The significance of this work lies in its potential to improve the accuracy and effectiveness of predictive maintenance systems for railway tracks.

2. Methodology

All following considerations are based on data from the Austrian standard track recording car. This delivers several signals, such as the longitudinal level, the track gauge, and the alignment two to four times a year—depending on the importance of the track [12]. As mentioned before, the condition and behaviour of the track are mostly described by the development of the standard deviation of the longitudinal level. In order to determine accurate deterioration models, deterioration branches have to be bounded by maintenance actions that affect the longitudinal level. Therefore, the input data for this research is the longitudinal level in the wavelength range of 3 to 25 m, described as the D1 signal in the European Standards [13]. For all three algorithms, it is important that the input signals are synchronised, as they are only roughly positioned in the database. As described by Fellinger [14], this most effectively works by shifting the measurements run with the aim of minimising the Euclidean distance d (Formula (1)) between the measurement runs, whereby the latest valid measurement run before a renewal or the latest measurement run in the database, respectively, forms the reference signal.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

The calculation of the Euclidean distance between two measurement points also includes the positional value in the longitudinal track dimension described by x. As this term is the result of the synchronisation process, it is not relevant for the calculation; therefore, the one-dimensional distance of y between two measurement points is sufficient for the synchronisation process. The sum of the distances $D_{M1\,|\,M2}$ is then a kind of quality index for the synchronisation of two measurement signals, M_1 and M_2, with a length L, shown in Formula (2).

$$D_{M_1|M_2} = \sqrt{\sum_{i=1}^{L} \left(y_{i|M_1} - y_{i|M_2} \right)^2} \tag{2}$$

When shifting one of the two measurement signals, the shift with the minimum distance $D_{M1\,|\,M2}$ can be found. This shift represents the distance in longitudinal direction by which the signal has to be moved, most of the time lying in the range of a few meters. In the upper part of Figure 1, two unsynchronised signals are shown. Those signals are then synchronised through the described process, with the result displayed in the lower part of Figure 1.

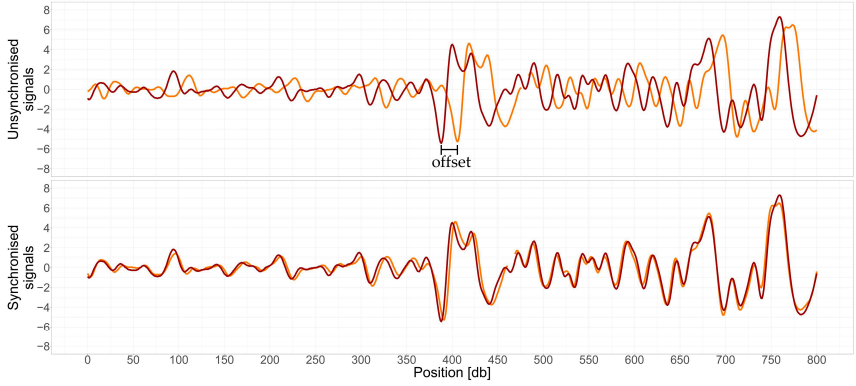


Figure 1. Signal synchronisation.

In the following, three different approaches to detect executed maintenance actions derived from the longitudinal level in the wavelength range of 3 to 25 m meters are presented. Hereby, the first two algorithms are cross-section based, whereas the third method uses the principle of the cumulative sum of the longitudinal level.

2.1. SEARCH Algorithm

The SEARCH algorithm, first described by Fellinger [11] and further developed for the aim of this comparison, can be employed to detect unrecorded tamping actions. The decision-making process is based on five conditions, all of which represent possible developments of the standard deviation of the longitudinal level D1. Moreover, the algorithm's precision can be enhanced by incorporating recorded tamping actions.

Basically, the SEARCH algorithm operates on a cross-sectional basis and can be implemented across the entire line by iterating over each cross-section. For every cross section, a loop is executed, whereby a new measurement point, including its corresponding date, is appended to a temporary data set in each iteration. Subsequently, the five conditions, presented in the following in numerical order, are checked to ascertain whether a tamping action may have occurred. If this is the case, the proposed date of the action is saved. The date is calculated via the mean of the measurement dates before and after the predicted maintenance. Furthermore, all data points before the predicted maintenance are deleted from the temporary data set. In the event that no condition is fulfilled, or the temporary dataset is cleared, a new point is added to said dataset. The loop is continued until the latest measurement is added to the temporary dataset and has been evaluated.

2.1.1. Condition/Rule 1

The application of Rule 1 is limited to cases where the temporary dataset comprises precisely two measurement points. In the event that the second measurement point exhibits a lower quality and a higher value than the first, no action is required, given that it is reasonable to expect a decline in track quality over time. Conversely, if the quality of the initial measurement exceeds the quality of the subsequent measurement by a defined value, a tamping action is identified. The threshold value is set at 0.25 mm, which allows for the reasonable assumption that a significant improvement can be attributed to track work and not to an issue with the data or other influences. Rule 1, like Rules 2–4, is illustrated graphically in Figure 2.

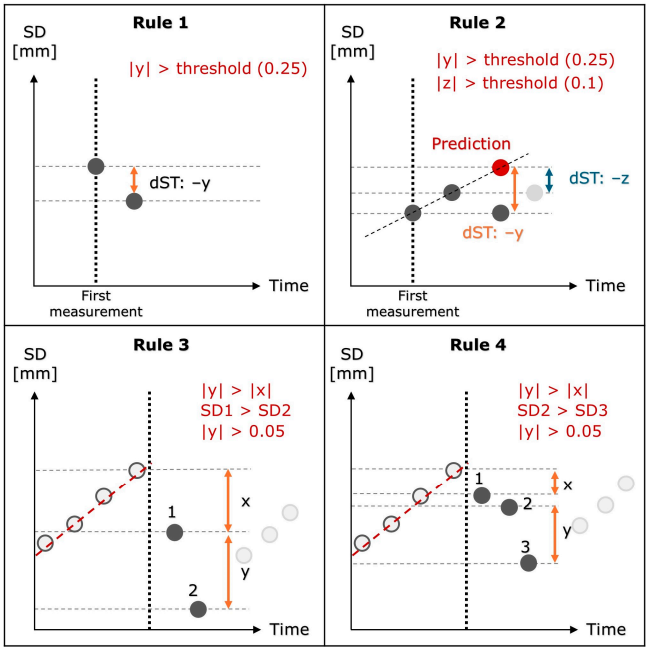


Figure 2. Rules 1–4 of the SEARCH algorithm.

2.1.2. Condition/Rule 2

Rule 2 is applied to a dataset comprising three to five data points. All data points, with the exception of the final one, are employed in the calculation of a linear regression, which is subsequently utilized to forecast the value of the last point. In the event that the final point exhibits a quality improvement of greater than 0.25 mm relative to the prediction, and the subsequent measurement point also demonstrates a quality enhancement of at least 0.1 mm in comparison to the same prediction, a tamping action is identified in advance of the last point of the provisional dataset. Incorporating the subsequent data point serves to reduce the probability of an outlier being erroneously identified as maintenance work.

2.1.3. Condition/Rule 3

The third rule can only be applied if a tamping action has been identified and the initial two data points of the provisional dataset have not fulfilled any of the specified conditions. The initial data point of the provisional set will be excluded if three conditions are met:

- The absolute increase in quality from temporary point 1 to temporary point 2 is greater than the absolute quality increase from the final point in the preceding deterioration branch to temporary point 1.
- Temporary point 2 exhibits higher quality than that observed at temporary point 1.
- The absolute increase in quality from temporary point 1 to temporary point 2 is greater than 0.05 mm.

Rule 3 is applied in order to eliminate outliers at the beginning of deterioration branches, thereby ensuring stable regressions.

2.1.4. Condition/Rule 4

Rule 4 is similar to Rule 3, as it also demands an already detected tamping and three measurement points in the new temporary dataset. Additionally, three conditions must be met:

- The absolute increase in quality from temporary point 2 to temporary point 3 is greater than the absolute quality increase from the final point in the preceding deterioration branch to temporary point 2.
- Temporary point 3 is of a higher quality than temporary point 2.
- The absolute increase in quality from temporary point 2 to temporary point 3 is greater than 0.05 mm.

The objective of this rule is to ensure stable regression for deterioration branches by eliminating outliers at the start of those branches. If the aforementioned conditions are met, the first two points of the temporary dataset will be excluded.

2.1.5. Condition/Rule 5

While Rules 1 through 4 are necessary for specific instances to ensure the proper functionality of Rule 5, Rule 5 can be regarded as the primary rule for detecting tamping actions. The temporary dataset must comprise a minimum of four data points, and no other rule must have identified a tamping action. All points, with the exception of the final one, are used to calculate a linear regression model. This is employed to forecast the value of the final point in the temporary dataset within a confidence interval with a statistical significance of 0.995. Consequently, it is possible to ascertain whether the measurement point is included in the linear regression.

Should the quality of the measurement exceed the predicted value, it may be indicative of either an outlier or the execution of a tamping action. Should the subsequent measurement point also exceed the predicted quality range (confidence interval), a tamping action will be recorded, and all points except the final one will be excluded from the temporary dataset. In the event that the final point is identified as an outlier, it is excluded from subsequent calculations.

Figure 3 illustrates an exemplary dataset with its linear regression and confidence interval. As is evident from this example, the predicted standard deviation is not significantly different from the actual value. Furthermore, the stability of the regression is of importance, as the widening of the confidence interval is dependent on the scattering of the data and the time span between the last and the predicted point.

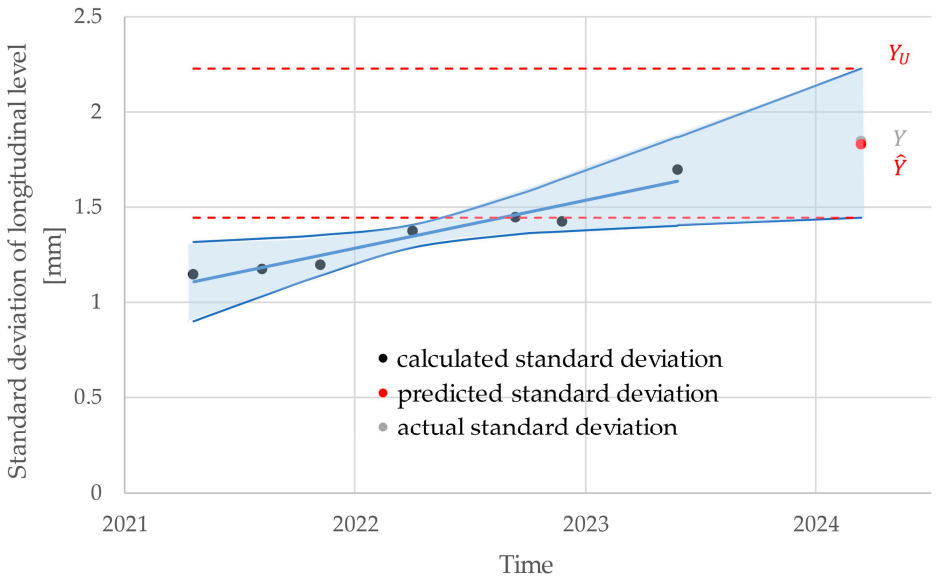


Figure 3. Application of Rule 5 of the SEARCH algorithm.

2.1.6. Condition/Rule 0—Adaptation of the Algorithm

After completion of the loop, the boundaries of all deterioration branches and outliers are known. If there are also known maintenance actions, these can be compared to the calculated ones. Furthermore, if there is a detected and a real tamping action amidst two measurement points, the calculated one will be overwritten by the recorded one. On the other hand, if no measurement point has been detected, a recorded one can be added.

In case the recorded data of tamping actions can be trusted, the algorithm will function more effectively. As Rule 0, already known maintenance actions will be included by default, instead of comparing and adding those afterwards. Upon the addition of a new measurement point to the temporary dataset, a verification is conducted to ascertain whether a recorded tamping action exists between the newly introduced point and the preceding one. If the aforementioned condition is satisfied, only the most recent data point will remain in the temporary data set, and the corresponding recorded maintenance action will be saved. The use of trusted tamping actions serves to enhance the algorithm, reducing the likelihood of missed outliers, unstable linear regressions, and slightly differing, not detectable behaviour of adjacent deterioration branches.

Figure 4 illustrates the application of the five rules to a fictional cross-section. While Rules 1 (yellow) and 2 (blue) are applied, they never result in a detection. Conversely, Rules 3 (and 4, green) do detect an outlier, and another outlier is identified by Rule 5 (red) in the first deterioration branch. As is the case with the majority of cross-sections, nearly all tamping actions are detected by Rule 5.

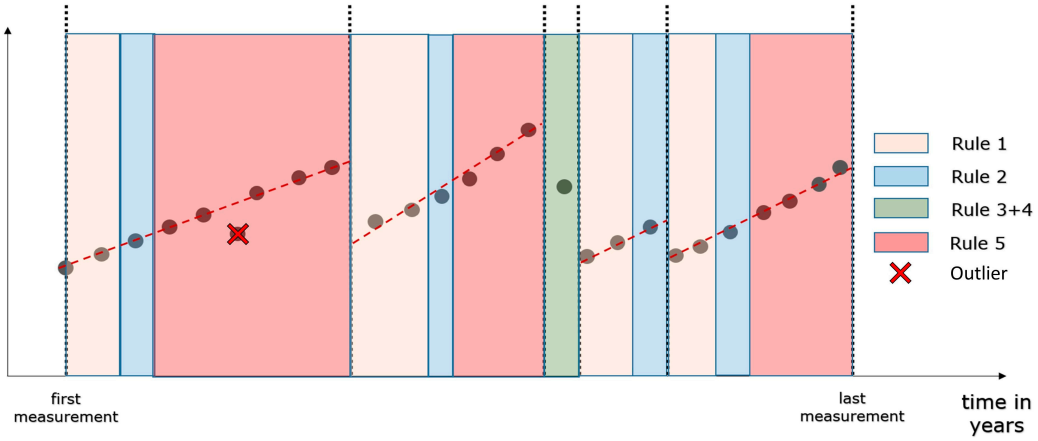


Figure 4. Application of all rules to one cross section.

For better understanding, the workflow of the SEARCH algorithm is depicted in a flow chart in Figure 5.

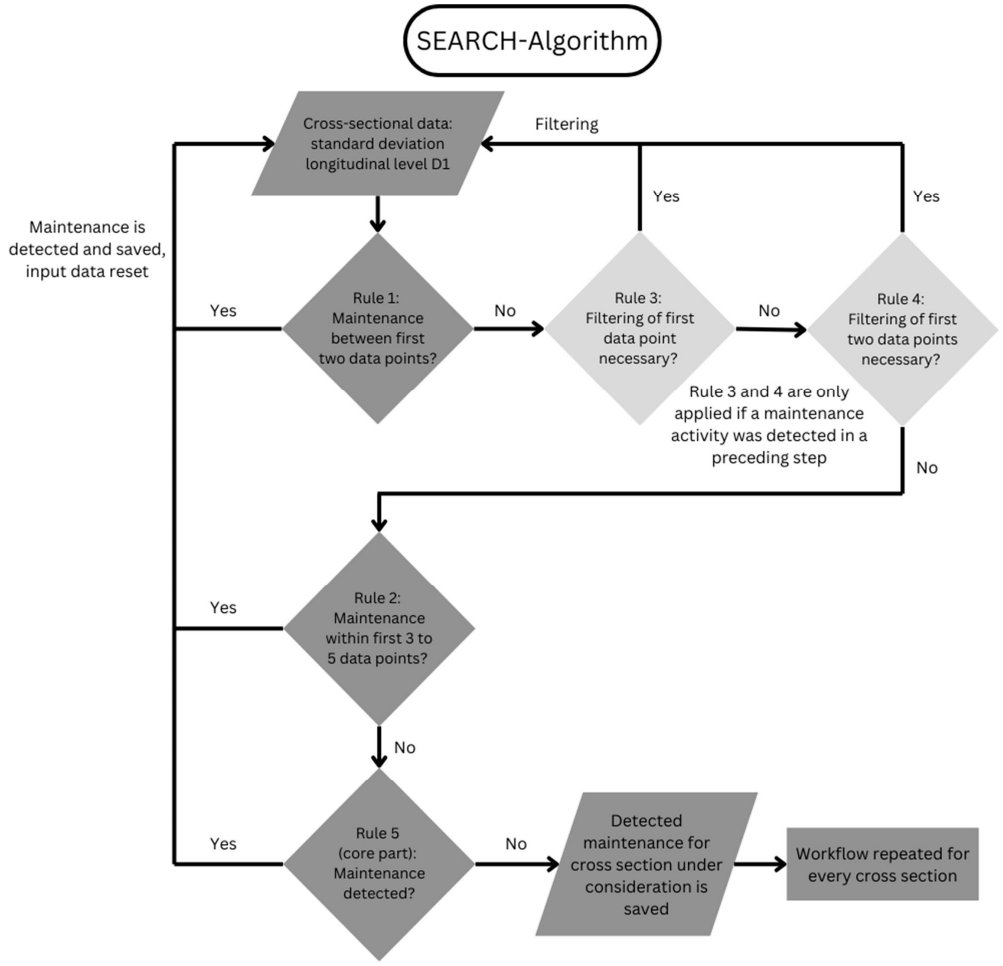


Figure 5. Flow chart of the SEARCH algorithm.

2.2. Cross-Section- and RANSAC-Based (CRAB) Algorithm

The second method provides the option of incorporating recorded maintenance data into the process, in a manner analogous to that of the SEARCH algorithm. If maintenance data is available, the period during which measurement data is available is divided into two or more rooms, with the number of rooms depending on the number of maintenance

activities. In the absence of maintenance data, all measurement data is treated as pertaining to a single room. The fundamental tenets of this algorithm are derived from the core principles of the RANSAC (Random Sample Consensus) algorithm [15], which is an iterative method employed to estimate the parameters of a mathematical model from a dataset that may contain outliers. The method operates by repeatedly selecting random subsets of the data, fitting a model to them, and evaluating which model has the greatest number of inliers. In this case, for each defined room bounded by maintenance actions or the beginning and end of data recording, respectively, every possible combination of two measurement points represented by the standard deviation of the longitudinal level D1 is selected iteratively. The primary objective is not to detect outliers but to identify individual deterioration branches, which occurs in two steps. In the first step, the two chosen data points establish a straight line around which an interval is traversed. The size of the interval is defined by the standard deviation of the data points in the respective room, whereby a value of 1/3 of the standard deviation has been found to be a sensible choice. Figure 6 shows that with an interval range of 1/3 (0.33) of the standard deviation of the longitudinal level D1, the highest F1 score can be reached by applying the algorithm to a calibration data set, which is further introduced in 2.3. The value by which the standard deviation of the longitudinal level D1 is multiplied is plotted on the x-axis.

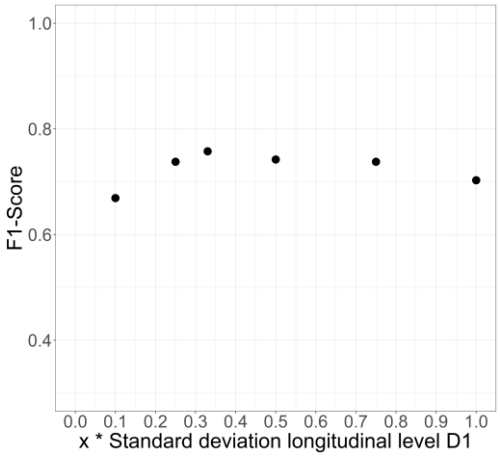


Figure 6. Calibration of interval setting.

All data points that fall within the interval are then labelled and saved in a list (green points in Figure 7). Once all potential combinations within the designated room have been processed, the set of data points that were most frequently identified as contiguous is defined as a set for a segregated deterioration branch.

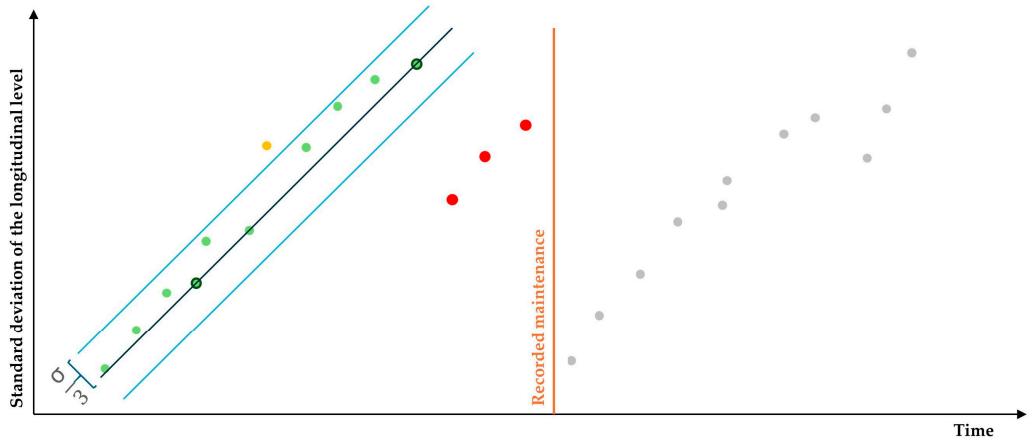


Figure 7. First step of the CRAB algorithm.

Data points that fall within the specified room and which are not part of the data set are discarded as outliers (orange points in Figure 7); those that fall outside the room (red points) are used for the subsequent determination of a deterioration branch. To prevent rooms with an excess of outliers from being erroneously designated as a segregated deterioration branch, the ratio of the time span of the identified room in relation to the labelled inliers must not exceed 1.5 times the overall inspection interval in this cross section. If no further measuring points are available for which room affiliation can be determined, the system will proceed with the next room initially defined by existing maintenance activities or the next cross-section. The procedure aims to fragment the cross-section into deterioration branches. Nonetheless, issues predominantly arise when the maintenance interval is shortened towards the end of the service life or the enhancement in the standard deviation of the longitudinal level subsequent to maintenance is minimal. Consequently, a further refinement is conducted in the second step, with the outcomes of the initial step serving as the basis for this process. In the second step, three measurement points are always employed across the room, with the first and third measurement points establishing a straight line. Subsequently, the vertical distance between the second data point and the representative point on the line is calculated, as shown in Figure 8.

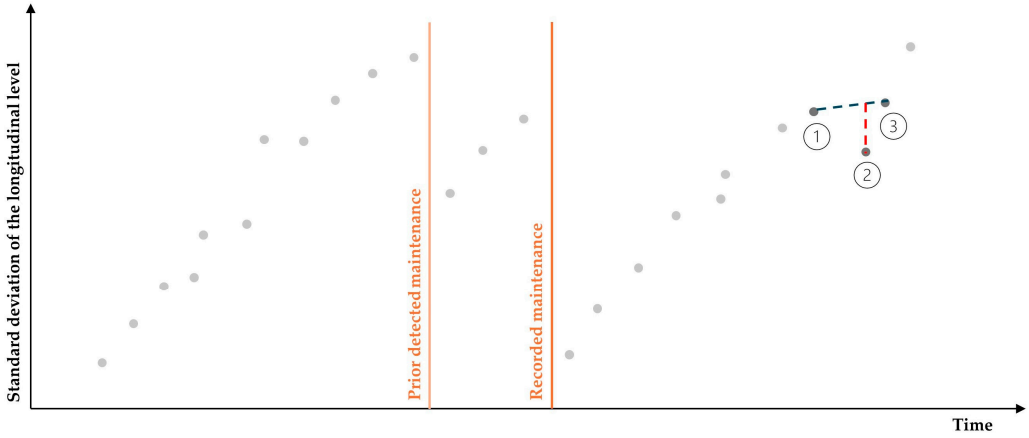


Figure 8. Second step of the CRAB algorithm.

It has been concluded empirically that if the distance is within half and double the entire room's standard deviation below the established line, maintenance was executed prior to the second data point. If the measuring point is situated below the lower limit, the data point is designated as an outlier. Upon completion of both steps, the algorithm returns the detected maintenance actions for each cross-section, while the date of the maintenance is defined as the midpoint between the two adjacent measurement runs.

For better understanding, the workflow of the CRAB algorithm is depicted in a flow chart in Figure 9.

2.3. Cumulative Track Geometry-Based Algorithm

The third algorithm employs the cumulative track geometry index, initially proposed by Loidolt for the assessment of turnout condition [16]. In the publication, the cumulative sum of the root mean squares (RMS) with an influence length of 3 m is used to represent the average track geometry quality of a turnout or parts of a turnout. For the aim of this paper, the approach is slightly modified, and the cumulative sum of the square roots instead of the root mean squares of the longitudinal level are used. The calculated index is called the Cumulative Index (CI) and is defined in Formula 3. The length L of the section can be selected arbitrarily, as will be demonstrated in the following explanations. CI is therefore

described as a function of the position and can be seen for multiple measurement runs in the upper part of Figure 10.

$$CI_i = \sum_{i=1}^{L} \sqrt{LL_i^2} \tag{3}$$

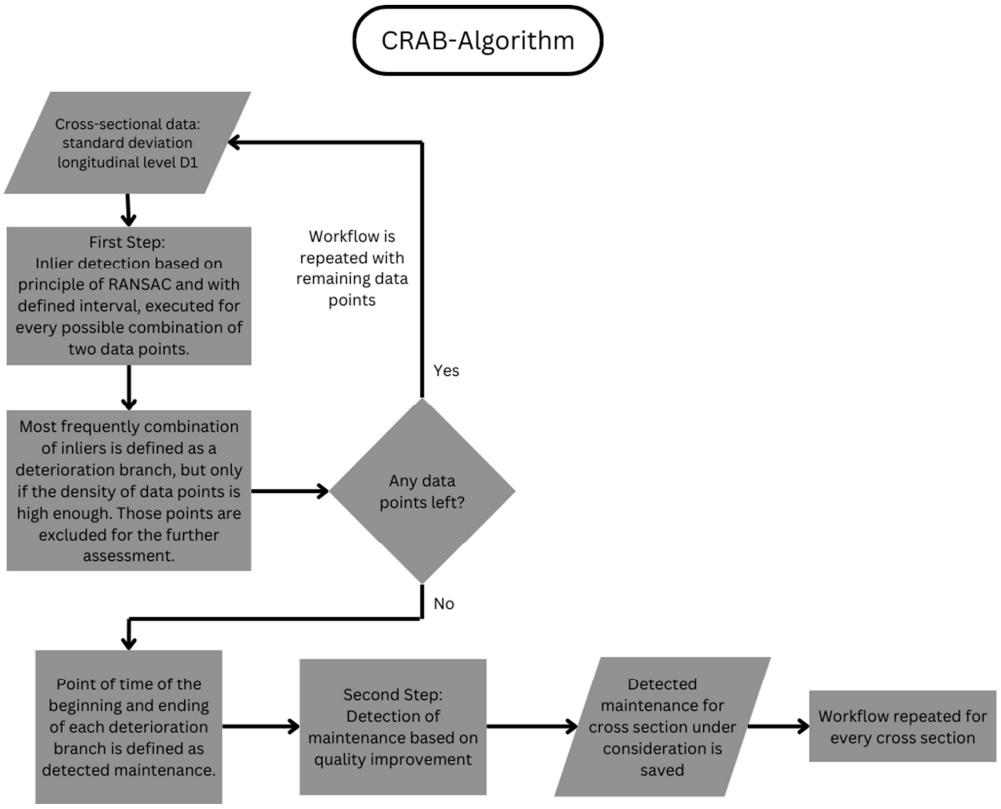


Figure 9. Flow chart of the CRAB algorithm.

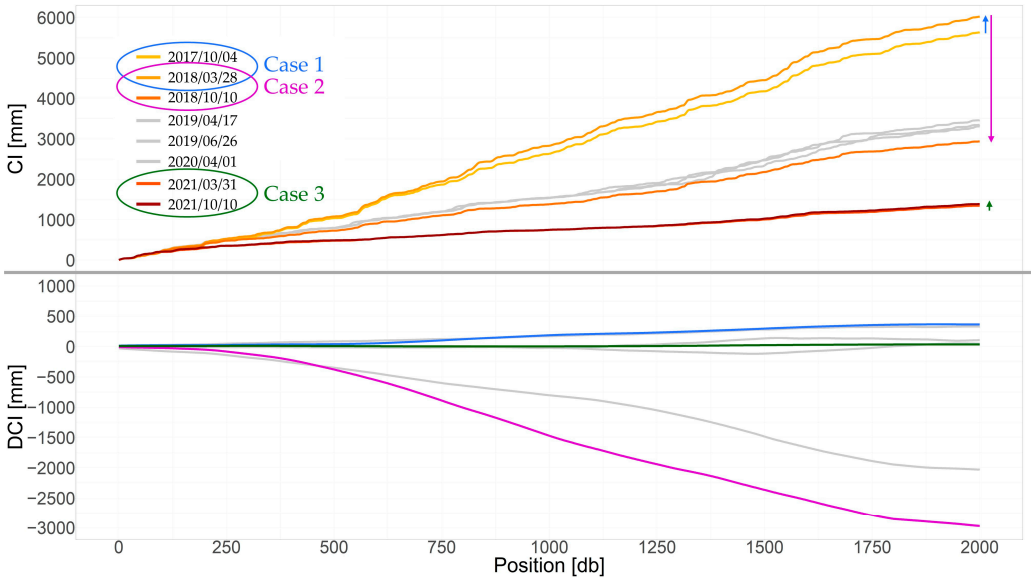


Figure 10. Visualisation of the CI and DCI signal for multiple measurement runs.

The local gradient of the cumulated curves reflects the track geometry quality of the respective location, with high gradients depicting poor quality. Deviations between two cumulated curves indicate either track deterioration or executed maintenance. In order

to capture the gradient differences across a range of CIs for detecting maintenance, the difference between each consecutive CI is calculated and referred to as the difference signal (DCI). Subsequently, the DCI for each position is smoothed by calculating the moving mean with a span of 100 metres with the objective of minimising excessive scattering. The DCI signal is depicted in the lower part of Figure 10.

Three scenarios are presented. In Case 1, no maintenance is performed between the two measurement runs, and track geometry deteriorates at a rapid rate as the track is potentially approaching the end of its service life. This accelerated deterioration results in an increase in the amplitude of the longitudinal level, which in turn leads to a positive gradient in the DCI signal (blue). In contrast, Case 2 involves a maintenance activity between the measurement runs, which serves to reduce the longitudinal level amplitudes. Therefore, the CI signal after the maintenance has a lower gradient than the CI signal of the measurement before the maintenance. Consequently, the gradient of the DCI signal in the area where maintenance has been carried out (magenta) is negative. After a few measurement runs, which are shown in grey, a track renewal was executed. As expected, the CI signal of the first measurement run after the renewal (31 March 2021) has a flat gradient. New, undamaged components result in a minimal deterioration of track geometry and no need for maintenance. Consequently, the CI signals display gradients that are almost identical (Case 3) and flatter than the gradients in Case 1. Furthermore, the DCI signal also has a low but positive gradient.

In this instance, the gradient is approximated via the secant of the DCI over a length of 100 m. The length of the secant was determined through an investigation in which secant lengths of 50 m, 75 m, 100 m, 125 m, 150 m, 200 m, and 250 m were analysed. The relation of true positive rate to false positive rate (Figure 11) reveals that a secant length of 100 m yields the best results, as then a good balance between a low false positive rate and a high true positive rate can be achieved. This analysis was conducted on four sections with comprehensive maintenance documentation, providing a ground truth that enabled the determination of the ideal secant length.

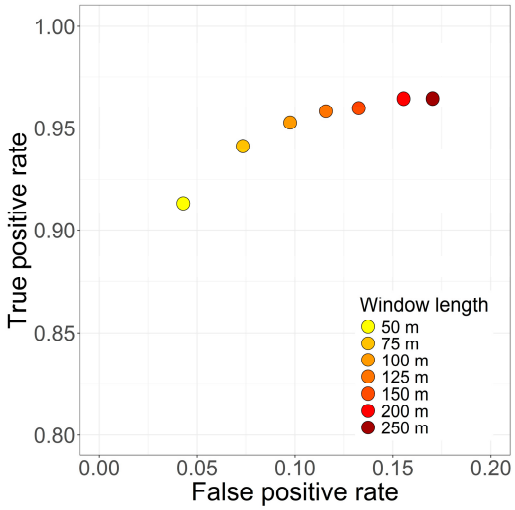


Figure 11. Determination of the secant length.

Given that the gradient of the DCI can now be described at each point using the secant gradient, the subsequent step is to ascertain in which areas the gradient of the DCI signal is negative. Given the varying gradients before and after the maxima and minima of the DCI (as depicted in Figure 12, where the gradient before the minimum is steeper than after), the selected range may be either too long or too short due to the secant length of 100 m. Accordingly, the precise location of the maxima and minima, which delineate the commencement and conclusion of the maintenance section, is subsequently determined (Figure 12).

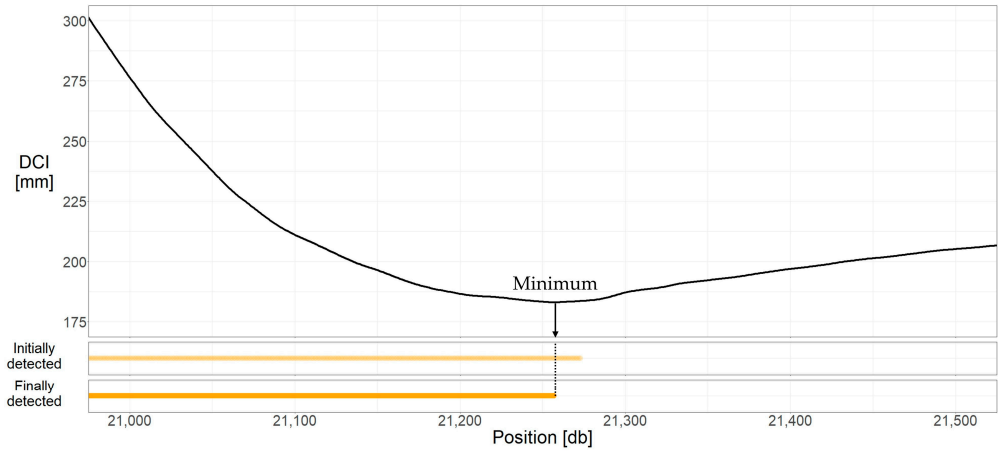


Figure 12. Consideration of asynchronous gradients.

If the algorithm is cancelled at this point, maintenance measures are incorrectly assigned to an excessive number of sections. The sections are therefore subjected to a more detailed analysis in two stages. The first step is to ascertain whether the negative slope of the DCI is due to potentially incorrectly synchronised data. For this purpose, a signal correlation analysis is conducted for each identified section. The final signal preceding the identified maintenance measure must exhibit a linear correlation of at least 0.7 with one of the two preceding measurement runs. It is assumed that no further maintenance was carried out during this period. The process is then repeated with the initial measurement signal following the identified maintenance activity, comparing it to the subsequent two signals. If either the correlation value before or after maintenance is too low, the detected section should be identified as an erroneous detection and subsequently be excluded from further consideration. The threshold value of 0.7 was determined using the same data set as the influence length and is based on the interpretation of Figure 13. It was ensured that the ratio of true negatives (correctly labelled as a section without maintenance; green curve in Figure 13) to false negatives (incorrectly labelled as a section without maintenance; red curve in Figure 13) is high and that the number of false negatives is low, so that the precision of the filtering is high. These requirements are best met by a correlation coefficient of 0.7, as the number of false negatives is small up to this point and increases sharply thereafter (red curve).

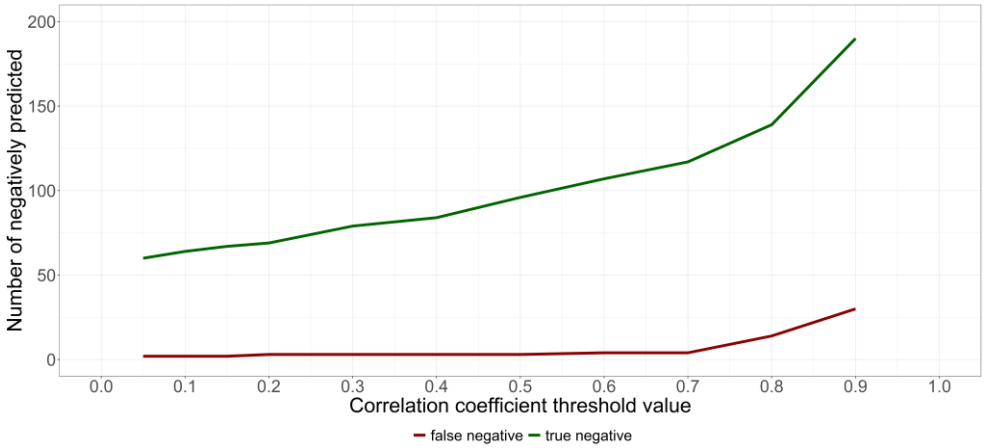


Figure 13. Determination of the correlation coefficient threshold value.

Secondly, the steepness of the DCI must also be taken into account when considering the result. To achieve this, the difference between the highest and lowest points of the DCI for the specific maintenance section is calculated and related to the length of the

maintenance section. This allows for the consideration of the overall reduction in the longitudinal level. The value of 700 has proven to be an appropriate choice when applied to the previously described test data set. This was determined using the ROC curve, the progressions of the negative and positive curves, and the F1 score (Figure 14). The ROC curve in Figure 14b illustrates that the optimal threshold value should be situated within the range of 600 to 1000, as evidenced by the comparable distances to the diagonal in this range. Additionally, Figure 14a demonstrates that the F1 score exhibits minimal growth from a threshold value of 700 onwards, whereby the enhancement in the F score up to 700 is considerable. This is further corroborated by the comparison of true negatives (dark green) and false positives (light red) in Figure 14c, which also indicate a flattening of both curves at this value. One evaluation alone would not allow a clear statement to be made about the most appropriate value, but when the results of all three evaluations are considered collectively, it becomes evident that the value of 700 is the most appropriate.

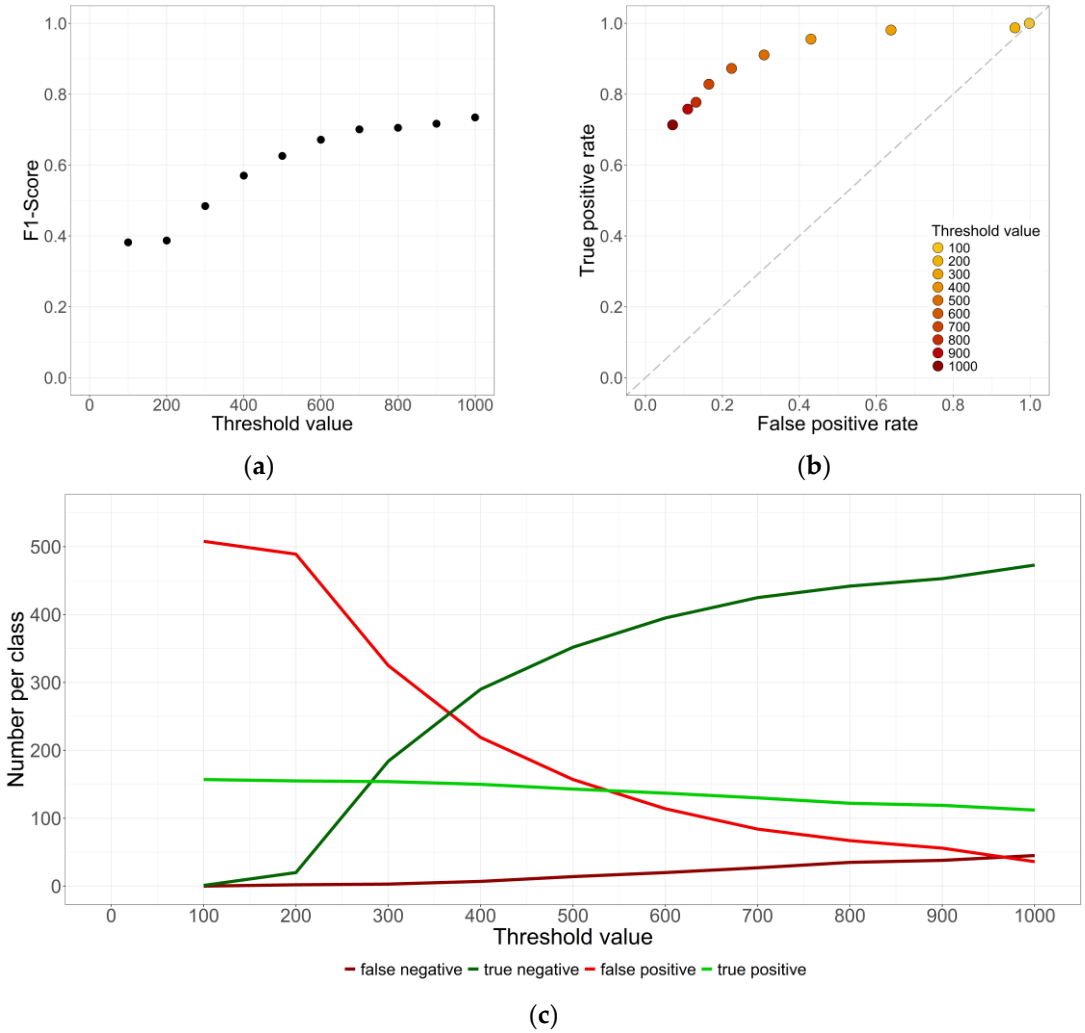


Figure 14. Determination of the threshold value (**a**) F1 score; (**b**) ROC curve; (**c**) Number of maintenance sections per class as a function of the threshold value.

Once all the mentioned steps have been completed, the algorithm generates a list of identified maintenance tasks, including the commencement and conclusion of each section, as well as the estimated date. In contrast to the preceding two algorithms, the definition of maintenance work is not cross-section based but rather line wide.

For better understanding, the workflow of the CTG-based algorithm is depicted in a flow chart in Figure 15.

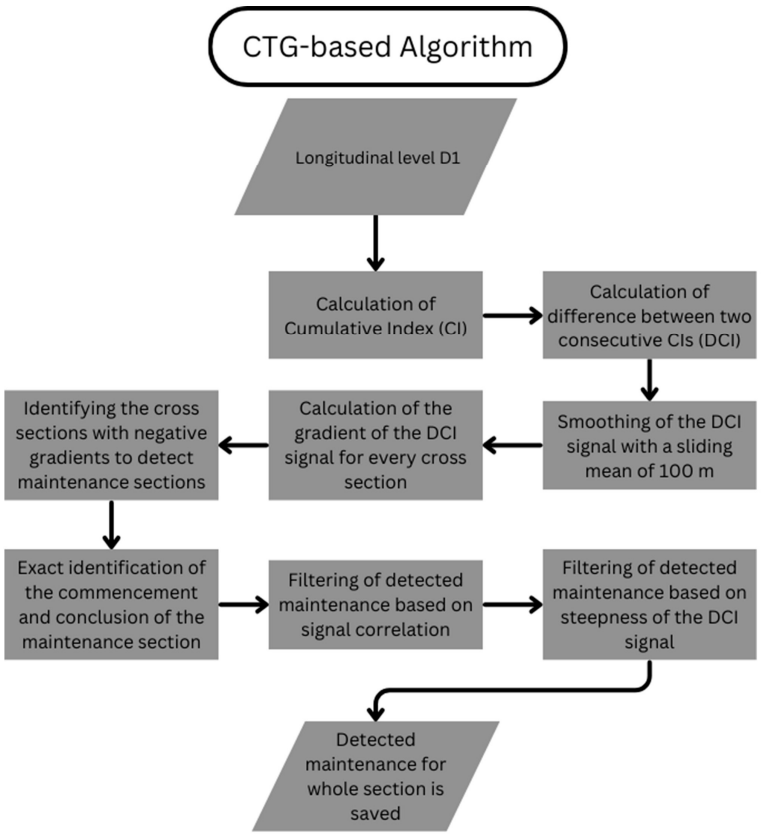


Figure 15. Flow chart of the CTG-based algorithm.

2.4. Method Performance Comparison

The three algorithms described are applied to four track sections in the network of the ÖBB-Infrastruktur AG, all of which have an average age of around 20 years. Apart from that, the sections have the following boundary conditions:

- Section 1: The first section has an average daily load of approximately 85,000 gross tons and is predominantly composed of concrete sleepers on 60E1 rails, extending for approximately 12 kilometres. The area encompasses 12 turnouts, 17 bridges, and two station areas.
- Section 2: In contrast to the first section, the second section has an average load of only 17,000 gross tons per day. Approximately 2/5 of the 60E1 rails are installed on concrete sleepers, while an equal number are installed on wooden sleepers. The remaining rails are installed on concrete sleepers with under sleeper pads. The section includes six turnouts, 26 bridges, two short tunnels, and four station areas. The total length of the section is 11 kilometres.
- Section 3: The third section, spanning approximately 10 kilometres, is primarily composed of 60E1 rails on concrete sleepers. The track is subjected to a gross tonnage of approximately 50,000 per day. The section incorporates 10 turnouts, 25 bridges, and 3 station areas.
- Section 4: The fourth section, which extends approximately five kilometres, is primarily composed of 60E1 rails on concrete sleepers. The track bears a load of approximately 67,000 gross tons per day and encompasses nine bridges, one station area, and no turnouts within the specified region.

The selection of sections is based on the consideration of enabling a comparison of sections with disparate loads and expected deterioration. The evaluation is based on data from the Austrian track recording car dating back to 2003 (Section 4), 2005 (Section 1), 2006 (Section 2), and 2012 (Section 3). The data necessary for the evaluation are the longitudinal level D1, which describes the vertical track geometry in the wavelength range from 3 to

25 m (used for the CTG-based algorithm), and the sliding standard deviation of this signal with an influence length of 100 m for the SEARCH and CRAB algorithm. For the recording of the track geometry, the Austrian track recording car utilises an inertial measurement unit (IMU) paired with an optical track gauge measurement system and a navigation system. The measuring principles and data output of the track recording car comply with European regulations (EN 13848) [13]. The modified maintenance database serves as the reference case for assessing the precision of the algorithms. Therefore, the recorded maintenance work was augmented and corrected with manually recorded sections through a process of visual inspection of the measurement signals and the TQIs derived from them.

In order to evaluate the three algorithms, the following metrics will be employed: precision, recall, and F-score. The classifications required for this analysis (true positive, true negative, false positive, false negative) are determined on a cross-sectional basis. Recall is defined as the number of true-positive results divided by the total number of elements that actually belong to the positive class. The precision for a class is the number of true positives divided by the total number of elements labelled as belonging to the positive class. As it is not reasonable to use recall and precision as the sole criteria, the two parameters are combined using the F-score. The formula for the F-score is:

$$F\text{-}Score = 2 * \frac{(precision * recall)}{(precision + recall)} \quad (4)$$

3. Results

Figure 16 illustrates that, with the exception of Section 2 (Figure 16b), the CRAB algorithm (shown in yellow) demonstrates the most optimal performance. This can be derived from the fact that the F-score of the CRAB algorithm is highest in Section 1 (Figure 16a: 0.78), Section 3 (Figure 16c: 0.85), and Section 4 (Figure 16d: 0.76) compared to the other methods. In Section 2 (Figure 16b), the suboptimal recall, in particular, results in a relatively low F-score (0.62) of the CRAB algorithm. This section has the lowest average loading and therefore the lowest expected deterioration rate. In all sections, the SEARCH algorithm (shown in red) exhibits the least favourable performance, particularly in the first two sections.

When all sections are considered together, a similar picture emerges (Figure 17). The performance of the CRAB algorithm is the best, closely followed by the cumulative track geometry-based algorithm (CTG, shown in orange). The SEARCH algorithm performs worst, achieving an F-score of just 0.53. The other two algorithms achieve an F-score of 0.74 and 0.76, respectively, whereby they differ primarily in terms of precision.

Given the suboptimal performance of the CRAB algorithm in Section 2, it is reasonable to devote further attention to this section. Figure 18a illustrates the maintenance procedures that were actually carried out in Section 2 as grey horizontal lines. Furthermore, the maintenance tasks identified by the three algorithms are displayed in the heat maps in Figure 18b–d as horizontal lines in magenta. The colours of the heat map represent the magnitude of the standard deviation of the longitudinal level D1 with an influence length of 100 m. As time progresses, the standard deviation of the longitudinal level D1 increases as track quality decreases. Following a tamping process, the standard deviation of the longitudinal level D1 exhibits a sudden drop, resulting in a colour shift in the heat map. The low recall rate of the CRAB algorithm can be attributed primarily to the fact that the algorithm only recognizes the continuous tamping measure in 2007 in specific sections. This is attributable to the slight enhancement in track geometry resulting from the tamping process, coupled with low deterioration rates after the tamping operation. This is where the most significant issue with the CRAB algorithm becomes evident: In cases where deterioration rates are constantly flat, it is not possible to differentiate between deterioration branches, given that minor changes in quality result in data points from multiple branches being classified as inliers. Consequently, it is not possible to distinguish between the deterioration areas. Therefore, especially for low-deterioration sections, the

performance would be improved if executed maintenance and renewal recordings would be included. Nevertheless, for comparing the algorithms, no recordings are considered. As can be observed, the CRAB algorithm also incorrectly identifies individual cross-sections and short sections with a detected maintenance action (Figure 18c). In contrast, the CTG-based algorithm provides a more uniform representation of the identified maintenance (Figure 18d). However, three major maintenance sections between data breaks 18,000 and 27,000 are not detected at all. In this section, the SEARCH algorithm primarily encounters difficulties in correctly detecting maintenance work at the beginning of the time series, which can be seen from the large number of erroneously detected track works from 2007 to 2010 (Figure 18b). Furthermore, in many isolated cross sections, maintenance is detected erroneously. This in turn resembles a very non-homogeneous and implausible picture of maintenance.

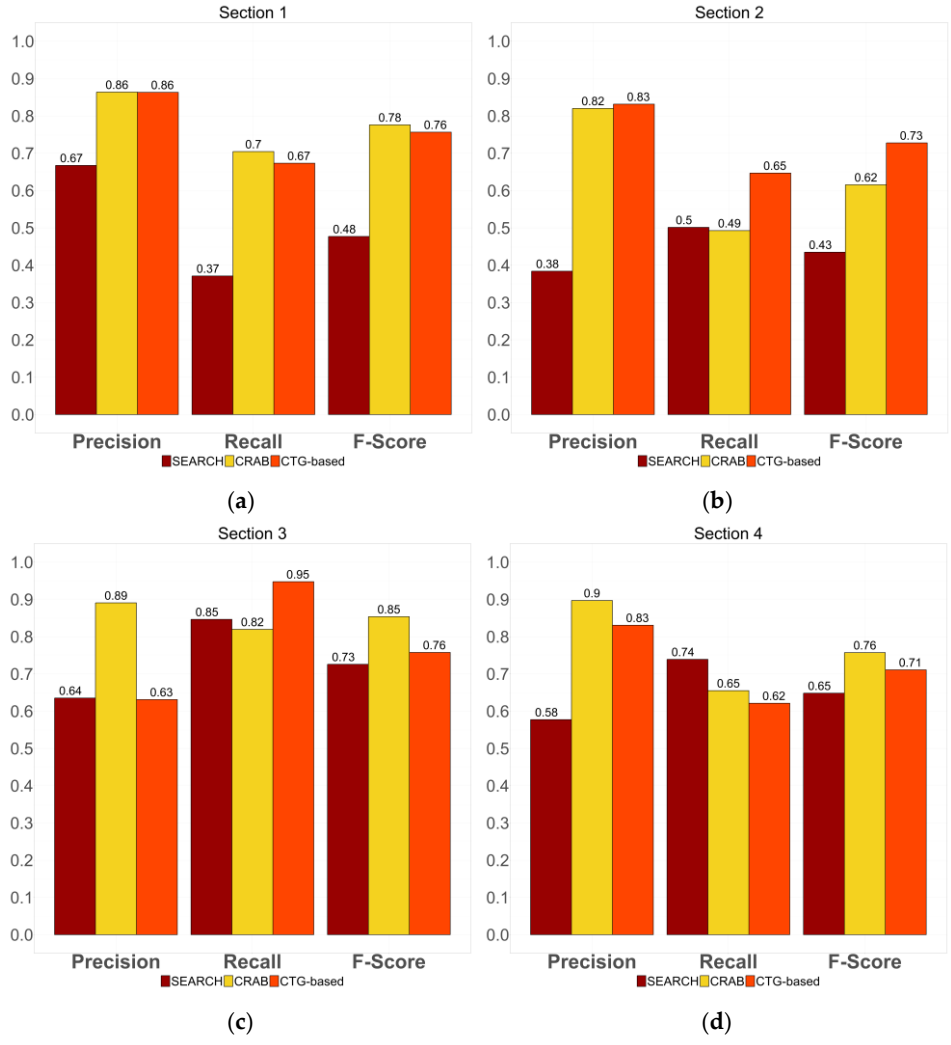
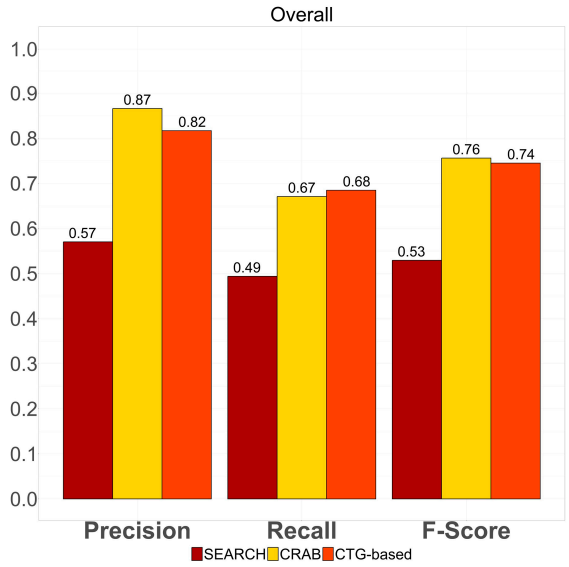


Figure 16. Statistical results of the three algorithms applied to the described four sections (**a**–**d**).

In contrast to the findings of Section 2, all algorithms in Section 3 (Figure 19b–d) demonstrate favourable performance values. This can be supported by comparing the results of the algorithms in Figure 19b–d with the actual track works in Figure 19a. While the SEARCH algorithm also incorrectly identifies maintenance work on numerous isolated cross-sections, it accurately recognises the bulk of the actually performed major maintenance works (Figure 19b). The CTG-based algorithm correctly identifies the obvious maintenance works (Figure 19c). However, it incorrectly identifies one long section as a maintenance section, as the measurement values are absent in this area (position 19,000 to 32,000, year 2015). The CRAB algorithm also correctly identifies all major maintenance

works (Figure 19c). However, the precise location of the commencement and conclusion of the maintenance work may not be accurately determined, which means that the precision is somewhat smaller than 1.



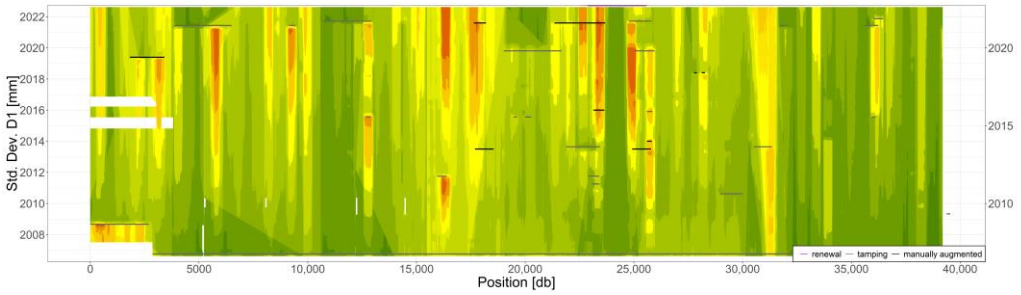
	SEARCH	CRAB	CTG-based
Precision	**0.57**	**0.87**	**0.82**
Recall	**0.49**	**0.67**	**0.68**
F-Score	**0.53**	**0.76**	**0.74**

Figure 17. Combined statistical result.

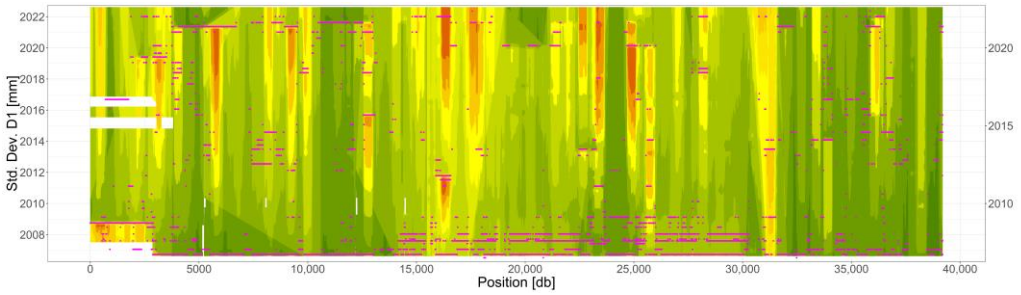
In order to transform the aforementioned findings into quantifiable data, it is necessary to ascertain whether the maintenance work recorded in each maintenance section can be accurately identified by the algorithm. An actual executed activity is deemed to be detected if track work is identified in at least 50% of the cross-sections within the maintenance section. The outcome of this assessment is presented in Figure 20, which illustrates the results for all algorithms and sections. In the first step, only the maintenance sections that were also part of the infrastructure manager's database are evaluated. This demonstrates that the CRAB algorithm effectively identifies the majority of track work in sections 3 (all sections detected) and 4 (22 of 24 sections detected), while the result in sections 1 (44 of 78 sections detected) and 2 (20 of 22 sections detected) is less optimal. Therefore, a further examination of Section 2 is conducted to ascertain how this outcome was attained: While the CRAB algorithm correctly identifies all evident track works, as described by Figure 19, numerous short track works are not discerned. Nevertheless, these are incorporated into Figure 20, irrespective of their length. This results in a considerable number of undetected track work sections. In comparison to the CRAB algorithm, the CTG-based algorithm is unable to identify a greater number of track works in any given section. However, across all sections, the CTG-based algorithm is still more effective than the SEARCH algorithm in detecting recorded track works. Once more, the significant discrepancy between Sections 2 and 3 is evident. The CTG-based algorithm identifies all track works in Section 3 in a manner analogous to the CRAB algorithm. The SEARCH algorithm misses one of 19 sections. In Section 2, however, both the SEARCH algorithm and the CTG-based algorithm are incapable of detecting more than half of the recorded track works. This is also due to the aforementioned numerous short track works in Section 2.

Nevertheless, it would be erroneous to consider the detection of recorded maintenance actions as the primary objective. Indeed, it is more prudent to concentrate on unrecorded track works. As illustrated in Figure 21, the number of unrecorded track works exhibits considerable variability across the sections. Nevertheless, the CRAB algorithm can be employed to identify the majority of unrecorded track works in each section, thereby enhancing the informative value of the existing database. The CTG-based algorithm demonstrates comparable performance. Furthermore, the SEARCH algorithm is also

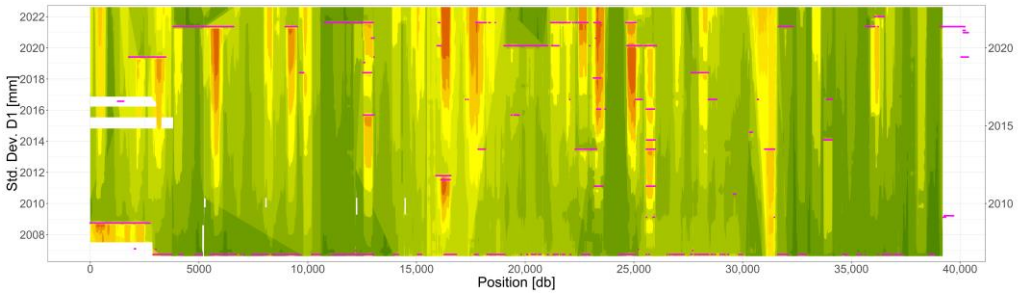
ranked last in this evaluation, as there is no section in which the SEARCH algorithm detects more track works than the other two algorithms.



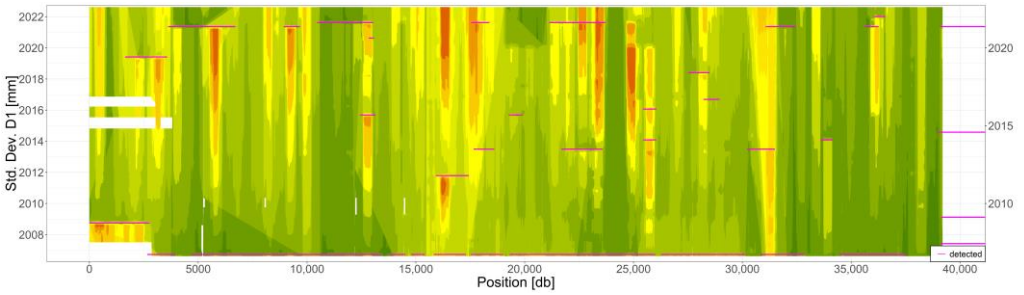
(**a**) Actually executed maintenance and renewals



(**b**) SEARCH algorithm

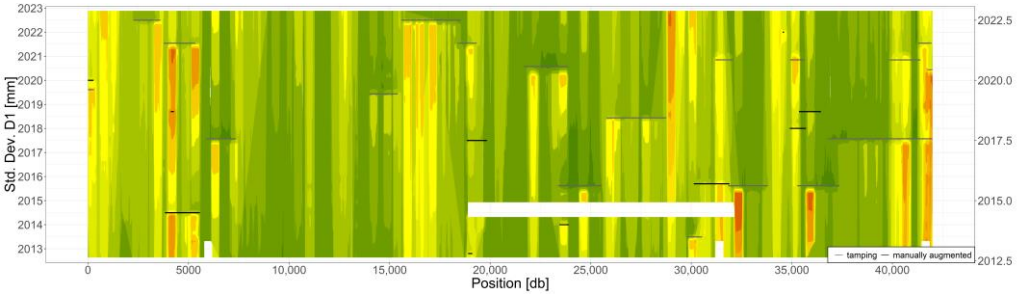


(**c**) CRAB algorithm

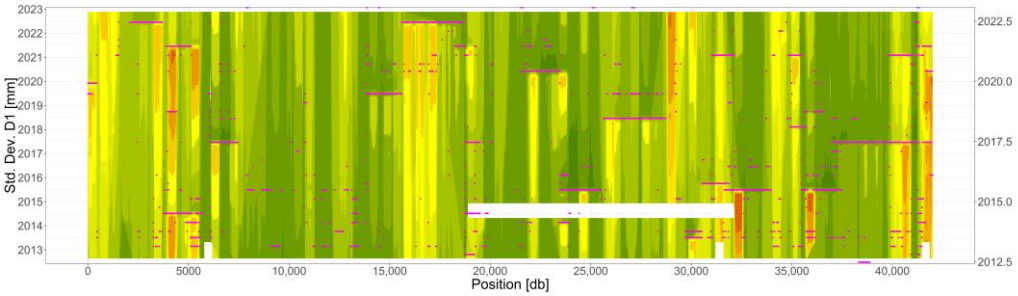


(**d**) CTG-based algorithm

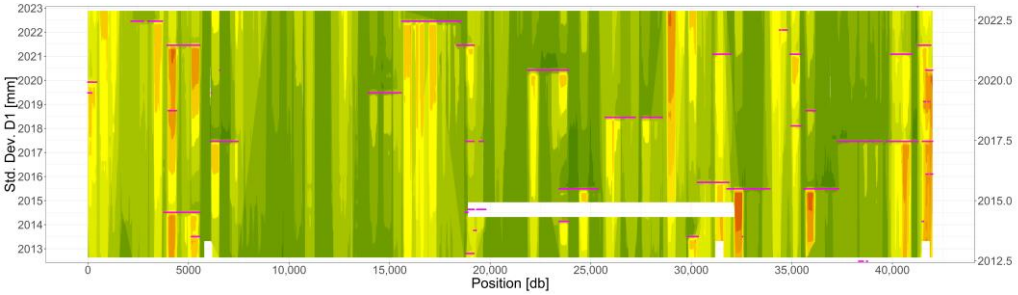
Figure 18. Recorded and detected maintenance in Section 2.



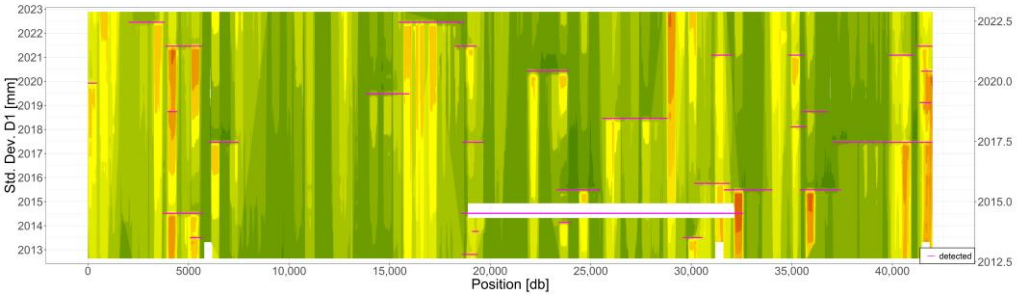
(**a**) Actually executed maintenance and renewals



(**b**) SEARCH algorithm



(**c**)CRAB algorithm



(**d**) CTG-based algorithm

Figure 19. Recorded and detected maintenance in Section 3.

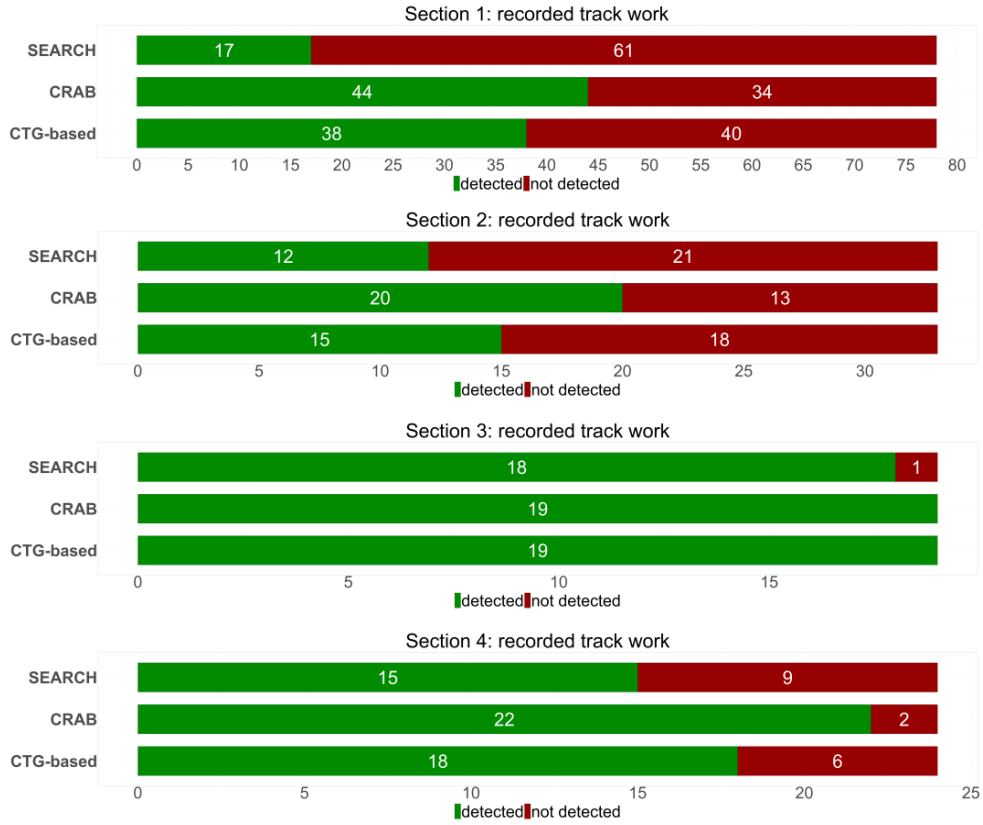


Figure 20. Amount of detected and not detected track work section for recorded track works.

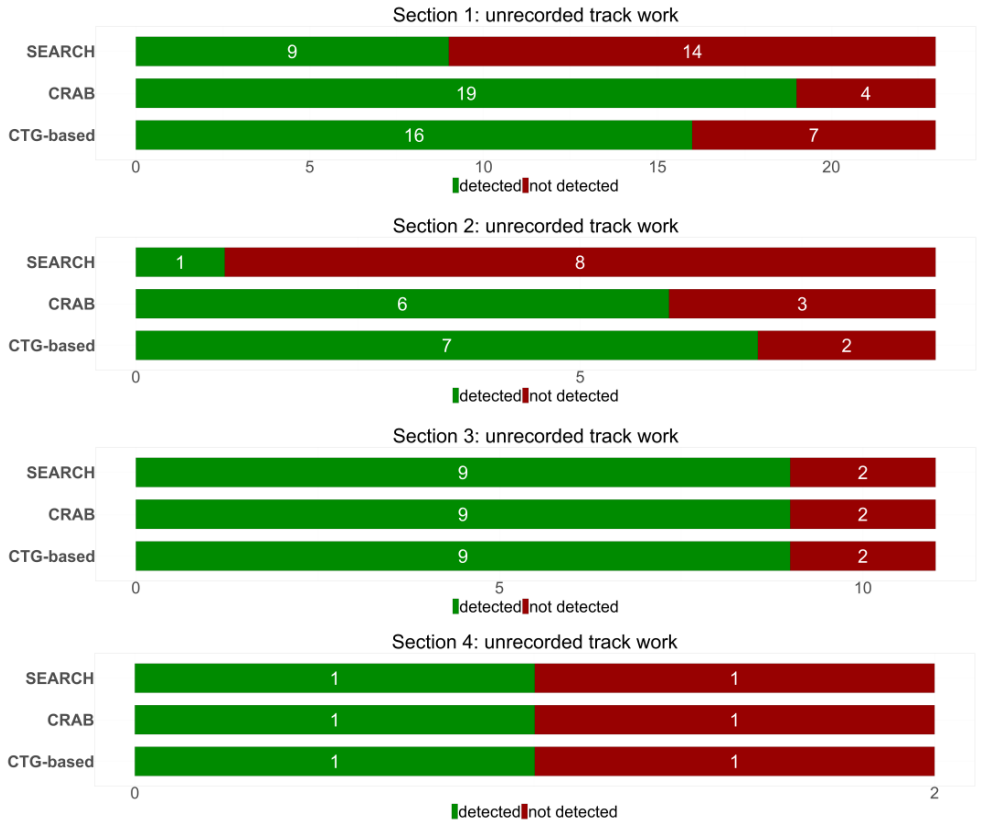


Figure 21. Amount of detected and not detected track work section for unrecorded track works.

4. Discussion

Up to now, the methodologies that have been developed are all based on cross-sectional, simplistic approaches that are unable to achieve the desired level of precision. Therefore, for enhancing predictive maintenance regimes, reliable methods to detect unrecorded maintenance works are required. As evidenced by the results, the CRAB algorithm and the CTG-based algorithm are the most effective of the proposed algorithms for the detection of unrecorded track works. The CRAB algorithm demonstrates superior performance in the majority of analyses. It bears repeating that the three algorithms are based on disparate principles. The SEARCH algorithm and the CRAB algorithm are capable of detecting track works on a cross-sectional basis. It should be noted, however, that the surrounding cross-sections are not included in the detection directly. However, both algorithms employ the standard deviation of the longitudinal level D1 with an influence length of 100 m to detect track works. This approach ensures that the development of the track geometry quality of neighbouring cross-sections is incorporated into the detection process over the influence length. Consequently, the detection of track works extends beyond their actual length, which in turn results in a reduction in the performance indicators described. The CTG-based algorithm differs from the other two algorithms in that it analyses the longitudinal level D1 longitudinally. As with the other two algorithms, it is often the case that the detected maintenance sections are longer than their true length. This is due to the asynchrony of the DCI signal at local minima and maxima and the lack of clarity regarding the definition of these points.

Since the performance of the CRAB algorithm and the CTG-based algorithm are decent but based on completely different principles and thus susceptible to different errors, a combination may further improve the results. In addition, recorded track work that has not been included in the evaluation of the algorithms can be considered as input variables. This improves the performance, especially for cross-section-based algorithms, as the data set is already divided in advance. The prerequisite for this is that the recorded track work data is trustworthy. Moreover, it is of paramount importance that the quality of the measurement data is of a high standard. It is of particular importance that the same measurement system, or at the very least a measurement system that is capable of reproducing the same results, is used for all measurements. Future research should address a meaningful combination of the approaches. One possibility would be to employ the CTG-based algorithm with stricter thresholds to identify a preliminary detection of the track works. Building on this, the CRAB algorithm can utilise these track works that have been detected with a high degree of probability as input parameters to achieve an even more precise result based on cross-section. Overall, the algorithms permit the identification of unrecorded maintenance activities with varying degrees of reliability. However, the specific type of maintenance performed is not currently considered. Future research may investigate whether the detected measures correspond to tamping, ballast bed cleaning, or a track renewal.

5. Conclusions

In conclusion, the CRAB algorithm is the most effective at identifying unrecorded tracks at the cross-sectional level. Conversely, the CTG-based algorithm offers the benefit of achieving a more 'homogeneous' mapping. This allows for the maintenance database to be updated more efficiently. Future research should aim to combine the two approaches in order to further improve the model accuracy and the results. As track geometry data and behaviour typically differ from country to country, testing the algorithms in different countries is another way to improve and deepen the research. Furthermore, more data sources could be integrated for enhancing the algorithms. For instance, other measurement signals may be employed for the purpose of detecting ballast-related or even other types of maintenance. In any case, both methods are able to significantly improve the quality of the input data for descriptive models and thus contribute to predictive maintenance regimes with a variety of advantages. To illustrate, the utilisation of increasingly brief

track closures can be employed in a more efficacious manner with the implementation of targeted maintenance measures. This yields notable benefits in terms of operational and economic efficiency.

Author Contributions: Conceptualization, J.S. and M.L.; methodology, J.S. and M.L.; software, J.S.; validation, J.S., F.G., M.L. and S.M.; formal analysis, J.S.; investigation, J.S.; resources, S.M.; data curation, J.S.; writing—original draft preparation, J.S. and F.G.; writing—review and editing, J.S., F.G., M.L. and S.M.; visualization, J.S. and F.G.; supervision, M.L. and S.M.; project administration, S.M.; funding acquisition, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to this data. Data were obtained from ÖBB-Infrastruktur AG.

Acknowledgments: Open Access Funding by the Graz University of Technology.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Binder, M.; Mezhuyev, V.; Tschandl, M. Predictive Maintenance for Railway Domain: A Systematic Literature Review. *IEEE Eng. Manag. Rev.* **2023**, *51*, 120–140. [CrossRef]
2. Marschnig, S.; Neuper, G.; Hansmann, F.; Fellinger, M.; Neuhold, J. Long Term Effects of Reduced Track Tamping Works. *Appl. Sci.* **2022**, *12*, 368. [CrossRef]
3. Wen, M.; Li, R.; Salling, K.B. Optimization of preventive condition-based tamping for railway tracks. *Eur. J. Oper. Res.* **2016**, *252*, 455–465. [CrossRef]
4. Andrews, J.; Prescott, D.; de Rozières, F. A stochastic model for railway track asset management. *Reliab. Eng. Syst. Saf.* **2014**, *130*, 76–84. [CrossRef]
5. Guler, H.; Jovanovic, S.; Evren, G. Modelling railway track geometry deterioration. *Proc. Inst. Civ. Eng.-Transp.* **2011**, *164*, 65–75. [CrossRef]
6. Jovanović, S.; Guler, H.; Čoko, B. Track degradation analysis in the scope of railway infrastructure maintenance management systems. *JCE* **2015**, *67*, 247–257. [CrossRef]
7. Caetano, L.F.; Teixeira, P.F. Predictive Maintenance Model for Ballast Tamping. *J. Transp. Eng.* **2016**, *142*, 04016006. [CrossRef]
8. Audley, M.; Andrews, J.D. The effects of tamping on railway track geometry degradation. *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit* **2013**, *227*, 376–391. [CrossRef]
9. Sedghi, M.; Bergquist, B.; Vanhatalo, E.; Migdalas, A. Data-driven maintenance planning and scheduling based on predicted railway track condition. *Qual. Reliab. Eng.* **2022**, *38*, 3689–3709. [CrossRef]
10. Neuhold, J.; Vidovic, I.; Marschnig, S. Preparing Track Geometry Data for Automated Maintenance Planning. *J. Transp. Eng. Part A Syst.* **2020**, *146*, 04020032. [CrossRef]
11. Fellinger, M.; Veit, P. *Sustainable Asset Management for Turnouts: From Measurement Data Analysis to Behaviour and Maintenance Prediction*; Verlag d. Technischen Universität Graz: Graz, Austria, 2020; ISBN 978-3-85125-776-2.
12. Hanreich, W. Moderne Fahrweginspektion mit dem Oberbautechnischen Messwagen EM250. *ZEVrail Glasers Annalen* **2004**, *126*, 18–27.
13. *EN 13848-5*; Railway Applications-Track-Track Geometry Quality-Part 5: Geometric Quality Levels-Plain Line, Switches and Crossings. European Standard: Brussels, Belgium, 2017.
14. Fellinger, M.; Wilfling, P.A.; Marschnig, S. CoMPAcT-Data Based Condition Monitoring and Prediction Analytics for Turnouts. In *Intelligent Quality Assessment of Railway Switches and Crossings*; Galeazzi, R., Kjartansson Danielsen, H., Kjær Ersbøll, B., Juul Jensen, D., Santos, I., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 129–148, ISBN 978-3-030-62471-2.
15. Fischler, M.A.; Bolles, R.C. Random sample consensus. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
16. Loidolt, M.; Marschnig, S.; Berghold, A. Track geometry quality assessments for turnouts. *Transp. Eng.* **2023**, *12*, 100170. [CrossRef]