Proceedings of the Seventh International Conference on
Railway Technology: Research, Development and Maintenance
Edited by: J. Pombo
Civil-Comp Conferences, Volume 15, Paper 17.9
Civil-Comp Press, Edinburgh, United Kingdom, 2026
ISSN: 2753-3239, doi: 10.4203/ccc.15.17.9

Composite Severity Score for Heavy-Haul Railway Track Assessment Using Dynamic, Geometric, and Contextual Information

W. Queiroz[1], P. Pizzigatti Correa[1], R. Motta[2], P. Pereira[2], O. Gogliano[1] and L. Bernucci[2]

[1] Department of Computer Engineering, Polytechnic School, University of São Paulo (USP), Brazil
[2] Department of Civil Engineering, Polytechnic School, University of São Paulo (USP), Brazil

Abstract

We propose a composite severity score for track condition assessment on heavy-haul railways that combines multisensor dynamic response, structural context, geometric non-compliance, and recurrence in a formulation that remains applicable when some data sources are unavailable. The method is evaluated on approximately 11,000 track segments from August 2020 using instrumented-vehicle measurements, control-car geometry data, and a ground-truth log of confirmed defects and operational restrictions from the Estrada de Ferro Vitória a Minas (EFVM), a Brazilian heavy-haul railway. Compared with the legacy fixed-threshold method, the proposed score increases recall for expert-confirmed critical segments from

52.3% to 81.6% and reduces the tendency to overclassify isolated signal peaks as high severity. In practical terms, the combined use of dynamic, geometric, and contextual information produces a more stable ranking for maintenance prioritisation.

Keywords: railway infrastructure, condition monitoring, severity classification, composite score, predictive maintenance, track geometry, instrumented vehicle, dynamic response.

1 Introduction

Condition assessment and maintenance prioritisation are central tasks in railway infrastructure management, particularly on heavy-haul routes where high axle loads accelerate geometry deterioration and local defects. Many operational screening workflows still rely on fixed thresholds applied to individual dynamic signals. Although such rules are easy to implement, they are sensitive to isolated peaks and make limited use of interactions among vehicle response, track geometry, and asset context.

Track quality indices such as the Track Quality Index (TQI) effectively summarise geometric deviations, and onboard monitoring systems provide valuable condition information from in-service vehicles [3, 7]. However, these approaches do not by themselves provide an interpretable severity measure that explicitly combines multisensor dynamic response, structural criticality, geometric noncompliance, and recurrence. This gap is directly relevant to railway monitoring, defect screening, and maintenance decision support.

The paper therefore proposes a composite severity score for track-segment assessment that integrates these four dimensions in a modular formulation. We deliberately favour an interpretable structure, because maintenance teams need to understand why a segment is ranked as severe. The method also remains usable when some information sources are unavailable, which is a common constraint in operational datasets.

The evaluation uses approximately 11,000 segments from August 2020 on the Estrada de Ferro Vitória a Minas (EFVM), a Brazilian heavy-haul railway, combining instrumented-vehicle measurements, control-car geometry data, and a ground-truth log of confirmed defects and operational restrictions. Compared with the legacy fixed-threshold workflow, the proposed score increases recall for confirmed critical segments from 52.3% to 81.6%.

The contribution of the paper is threefold: (i) an interpretable composite severity formulation tailored to railway track monitoring; (ii) a validation strategy based

on confirmed field records rather than proxy labels alone; and (iii) an empirical comparison against the legacy threshold-based workflow used on the case-study railway.

2 Background and Prior Approaches

The current classification workflow examined in this case study has its origins in threshold-based logic. Before proposing a new formulation, we briefly describe the original method and the exploratory data-driven enhancements evaluated previously.

2.1 Original Threshold-Based Model

The initial severity classification model was based on fixed rules and absolute thresholds for each sensor [7, 11]. Thresholds were empirically defined, and the final severity score for each segment was determined by the highest individual parameter. As a result, an isolated peak in bounce could trigger a high severity score even when the remaining signals were stable. The method did not consider correlation between variables or structural context.

2.2 Data-Driven Exploratory Models

We also evaluated exploratory data-driven alternatives using statistical analysis, k-means clustering [10], and Random Forest classifiers [9]. The purpose of k-means was to explore signal patterns rather than to generate operational classifications. The Random Forest model reproduced the labels generated by the original rule-based method, but it did not provide a new decision structure. Similar machine-learning strategies have been investigated in railway predictive maintenance and track degradation studies [5, 6]. In our setting, however, neither approach incorporated recurrence, geometry, or explicit auditability in the decision logic.

Part of the instrumentation and data infrastructure used in this work was originally described in Silva et al. [2]. In that study, k-means clustering was applied to identify behavioural patterns preceding reported failures, and severity was inferred indirectly from the rarity of cluster occurrences. However, this approach lacked objective criteria for severity and did not provide a directly operable decision structure.

The Instrumented Vehicle (VI) is equipped with triaxial and uniaxial accelerometers, load cells, pressure sensors, onboard GPS, and an autonomous

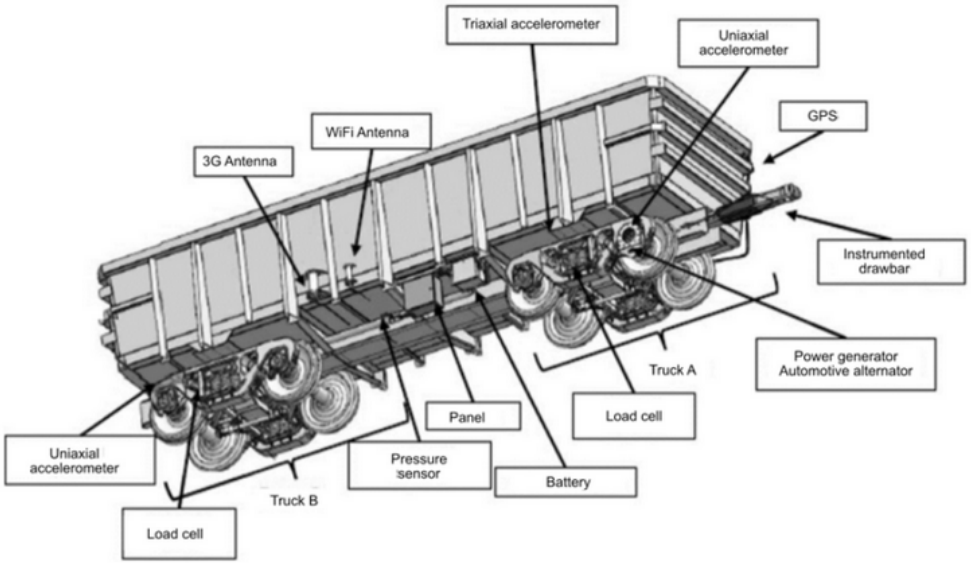

Figure 1: Functional diagram of the Instrumented Vehicle (IV), showing the location of acceleration, pressure, load, and GPS sensors.

power system. Data are sampled at 10 Hz and aggregated for later analysis.

2.3 Limitations of Previous Models

The absolute-threshold model has important limitations. Each sensor was evaluated independently, without considering the correlation between signals or their contextual relationship. The classification ignored the type of track segment (e.g., bridge, curve, turnout, transition) and lacked integration with geometric data from the Control Car (CC). Severity scores remained static even as new data became available, and thresholds were fixed and rigid, with no tolerance to signal variation or operational noise. Additionally, the method lacked multivariate structure, contextual awareness, and auditability — making it incompatible with modern learning-based approaches.

Although previous studies have explored threshold-based classification, data-driven modelling, and track condition monitoring using individual data sources, most approaches treat dynamic signals, geometry, and structural context in isolation or rely on black-box predictive models. In contrast, the formulation proposed in this work explicitly integrates multisensor dynamic response, geometric non-compliance, structural criticality, and recurrence into a unified and interpretable framework. This combination is particularly relevant for operational use, where maintenance decisions require not only detection performance but also

transparency and robustness under partial data availability.

3 Methods – Proposed Composite Score Model

The proposed model defines a continuous severity score, calculated as:

$$S = R \times C \times W \times G \tag{1}$$

In this formulation, S denotes the final severity score, R the recurrence factor, C the multisensor dynamic component, W the structural weighting factor, and G the geometry penalty. The composite score adopts a multiplicative structure to reflect the engineering expectation that concurrent risk factors amplify severity more strongly in combination than in isolation. We chose this structure to keep the score interpretable in engineering terms, rather than to maximise predictive performance through a black-box model. Alternative aggregation rules, such as additive or geometric-mean formulations, remain topics for future comparative assessment.

3.1 Recurrence factor

Reflects the repetition of anomalies across multiple voyages over the same segment. Here, n_{rec} denotes the number of repeated anomalous occurrences observed for the same segment across different passes. The factor acts as an amplification term even when individual sensor values remain below critical thresholds.

$$R = 1 + \log(1 + n_{\text{rec}}) \tag{2}$$

In the absence of multiple passes in the dataset analysed here, we assume $R = 1$. The logarithmic growth avoids inflation of S, ensuring stability while capturing meaningful recurrence.

3.2 Multisensor dynamic component

Represents the combined intensity of VI signals. It accounts for individual sensor behaviour, z-score standardisation, variable-specific weights, and cross-signal coherence.

$$C = \sum_{i=1}^{N} w_i \cdot z_i + \lambda \cdot \text{Var}(PC_1) \tag{3}$$

- w_i: weight assigned to sensor i
- z_i: standardised value (z-score) of sensor i
- PC_1: first principal component from PCA, i.e., the dominant direction in the data capturing the largest share of variance
- λ: adjustment factor controlling the contribution of the dominant principal-component mode

This structure captures both isolated spikes and distributed anomalies. When several signals rise together, C increases accordingly. The second term adjusts the aggregate score according to the variance associated with the first principal component, so the formulation remains sensitive to cases in which one mode dominates the response.

In the current implementation, we adopted $\lambda = 0.5$ based on preliminary sensitivity tests using the 2020 dataset. This value provided a good trade-off between coherence enforcement and false positive control. As with other parameters, λ can be adjusted through supervised learning as more failure labels become available.

3.3 Structural weighting

Represents the physical context of each segment. Curves, bridges, and transitions are structurally more prone to instability, vibration, and wear. The initial weights follow structural prioritisation logic commonly used in heavy-haul railways. However, the model supports automatic adjustment of these weights over time as more data become available. In this way, W may evolve from fixed rules to a calibrated parameter via supervised learning.

Turnouts (track switches) and transitions are zones where dynamic impacts are expected due to the physical geometry and switching components. These responses do not necessarily indicate degradation. Therefore, a slight attenuation factor is applied:

$$W = \begin{cases} 1.3 & \text{if bridge} \\ 1.2 & \text{if curve} \\ 1.1 & \text{if transition} \\ 0.95 & \text{if turnout (track switch)} \\ 1.0 & \text{if tangent} \end{cases} \tag{4}$$

The initial weights used in W reflect common risk prioritisation in heavy-haul freight railways, where bridges and curves present greater structural vulnerability

to vibration and wear [8]. Future iterations may adjust these weights through supervised calibration using validated failure records.

3.4 Geometry penalty

Represents the penalty associated with non-conformity with the NBR 16.387/2016 Brazilian standard [1]. Parameters considered include gauge, alignment, longitudinal level, twist, and superelevation. The more parameters exceed the allowed limits, the higher the value of G:

$$G = 1 + \gamma \cdot \left(\frac{n_{\text{exceeded}}}{n_{\text{total}}} \right) \tag{5}$$

- γ: scaling factor (typically 0.2)
- n_{exceeded}: number of geometric parameters exceeding the standard limits
- n_{total}: total number of parameters evaluated

If no geometric limits are violated in a segment, then $G = 1$ and the geometry has no impact on the severity score. The default value of $\gamma = 0.2$ was defined based on expert judgement and empirical calibration, providing moderate geometric influence while maintaining score stability. This magnitude is also consistent with previous studies that model geometric penalties in a non-dominant but additive manner (e.g., Weston et al. [3]). The factor can be fine-tuned using operational data or optimised through supervised learning as labelled failures become available.

3.5 Severity classes and adaptability

The severity score S increases with anomaly recurrence over time (R), multivariate dynamic deviations (C), structural criticality of the segment (W), and geometric non-conformity (G). This continuous score can be used for ranking, prioritisation, or decision support without relying on fixed thresholds [5, 6].

For reporting and validation, the continuous score S is mapped into three classes using empirical tertiles computed from the August 2020 dataset. This is a reporting convention rather than an optimised decision boundary. Let P_{33} and P_{66} be the 33rd and 66th percentiles of the observed S distribution. Then:

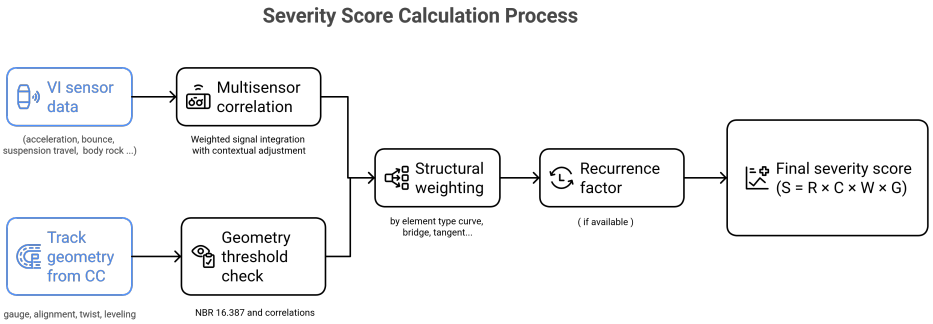

Figure 2: Severity score calculation process for the proposed formulation $S = R \times C \times W \times G$.

$$\text{Class}(S) = \begin{cases} \text{Low} & \text{if } S \leq P_{33} \\ \text{Moderate} & \text{if } P_{33} < S \leq P_{66} \\ \text{High} & \text{if } S > P_{66} \end{cases} \tag{6}$$

The proposed score is modular. Parameters such as sensor weights w_i in C, structural weights in W, and penalty factors (γ, λ) can be recalibrated as additional labelled data become available while preserving the same overall structure.

4 Results and Validation

This section presents the application of the proposed severity score to operational data collected from the case-study railway. The methodology was tested using synchronised dynamic and geometric measurements from the Instrumented Vehicle (IV) and the Control Car (CC), both aligned to August 2020.

4.1 Description of the Dataset

The broader project accumulated a large volume of operational data. However, only the data from August 2020 were used here because they are the only reliable and complete VI records currently available. This dataset also offers temporal and spatial compatibility with Control Car data, enabling consistent cross-referencing.

4.1.1 Instrumented Vehicle (IV)

The data correspond to a complete run along the monitored corridor, with only minor gaps. Measurements were recorded every 25 centimetres, yielding more than 4 million raw data points before aggregation. The dataset used here was aggregated by analysis segment, resulting in approximately 11,000 valid records. Each record includes triaxial acceleration, suspension travel, bounce, body rock, and the corresponding segment metadata.

4.1.2 Control Car (CC)

Geometric data were extracted from the quarterly inspection carried out during the same month. The variables included `bit_tras`, `emp_20`, `emp_100`, `al_10m`, `niv_long`, and `supr`, among others. These were evaluated against NBR 16.387/2016 and used in the calculation of the geometric factor G. Cross-referencing between IV and CC data was performed by kilometre, segment type, and track line.

The Control Car is a continuous inspection vehicle similar in structure to the EM80 model, widely used in North and South American railways. Its onboard systems enable automated extraction of structural parameters, which are then compared against the NBR 16.387/2016 limits and used in the computation of the geometric factor G.

Complementary to geometry cars, onboard inertial measurements (e.g., axle box acceleration) have been widely investigated as a scalable source of information for detecting localized defects and monitoring track condition [4].

The restriction to the August 2020 dataset was a deliberate choice to ensure data integrity.

4.2 Preliminary Results

The results presented in this section are exploratory and were obtained by applying the proposed severity score to the aligned dataset from August 2020. The model was executed with $R = 1$ because repeated passes over the same segments were unavailable. The geometry factor G was applied where Control Car data were available and aligned.

Comparisons were made against the original fixed-threshold labels and against replicated classifications derived from statistical percentiles. Figures 3 and 4 highlight how the proposed score improves classification granularity and reduces high-severity inflation.

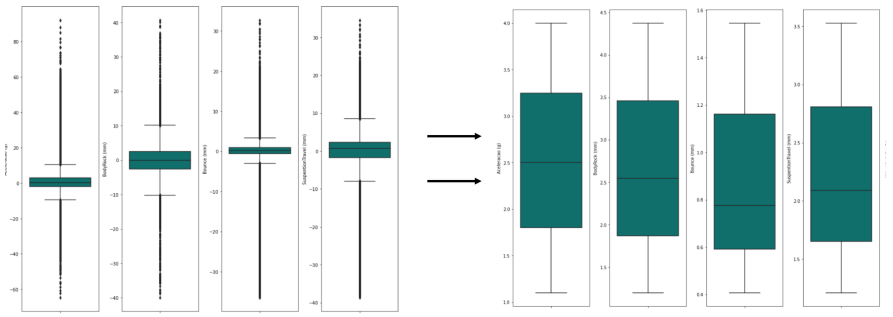

Figure 3: VI sensor signals before and after normalisation, illustrating the reduction of scale disparities before computation of C.

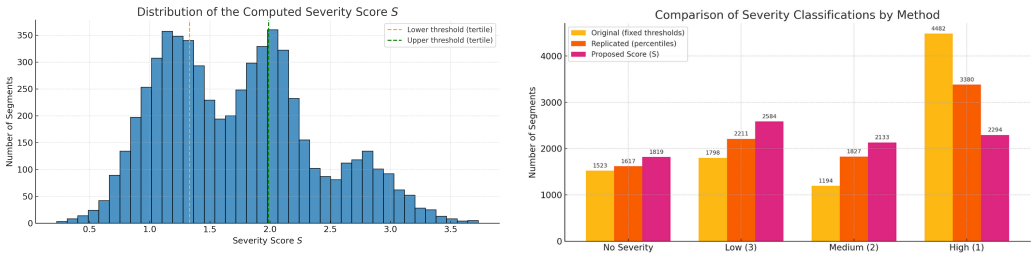

Figure 4: Left: distribution of the computed severity score S, with dashed lines indicating empirical tertiles. Right: severity level distribution across the original fixed-threshold rule, a percentile-based replica, and the proposed score model.

4.3 Validation Against Ground Truth

To assess the model's ability to reflect operational severity, we validated its outputs against a curated ground truth dataset containing confirmed failures and operational restrictions recorded in August 2020. Each record specifies the affected track segment, its location (km), track line (L1/L2), classification (e.g., defect, restriction), and physical context (e.g., curve, bridge).

The ground truth log includes infrastructure-related incidents confirmed by expert review. These incidents are based on actual interventions, inspection reports, and operational restrictions, making them a reliable source of validation.

We cross-referenced this ground truth with our unified dataset, matching line, segment type, and proximity in kilometres (tolerance ± 1.0 km). For each matched segment, we extracted the computed score S, the severity class assigned by our model, and the classification from the original fixed-threshold method. The dynamic inputs came from the Instrumented Vehicle (IV), while geometric parameters used in G were taken from Control Car (CC) data aligned to the same inspection period.

Among ground truth segments classified as *critical* (failures or restrictions), 81.6% were classified as *High* by the proposed score, whereas the threshold-based method classified 52.3% as *High*. Under the definition *critical* = positive and *High* = positive prediction, these values correspond to recall. Because the available log records confirmed issues rather than an exhaustive set of unaffected segments, the quantitative comparison is centred on recall; precision and false-alarm rates require further assessment with more complete maintenance outcome records.

Table 1 summarises a representative sample of these validated comparisons.

Table 1: Sample of validated segments comparing the proposed model, the original method, and ground truth.

KM	Line	Type	Score S	Prop. Class	Fixed Threshold	Ground Truth
12.0	L1	Bridge	2.57	High	Moderate	Confirmed Restriction
13.0	L2	Tangent 8	3.47	High	High	Confirmed Failure
35.0	L1	Tangent 23	3.99	High	None	Confirmed Defect
136.0	L1	Curve 2	3.23	High	Moderate	Confirmed Restriction
211.0	L1	Curve 4	3.30	High	High	Confirmed Restriction
311.0	L1	Tangent 1	3.10	High	Low	Confirmed Defect
326.0	L2	Tangent 13	2.97	High	High	Confirmed Failure
365.0	L1	Tangent 4	2.56	High	Low	Confirmed Defect
383.0	L1	Curve 6	2.80	High	High	Confirmed Defect

Taken together, these results show that the model produces a more informative severity ranking and reduces high-severity inflation. Segments such as KM 365.0

and KM 136.0 were underestimated by the original model, while KM 35.0 was not flagged at all despite a confirmed defect. In contrast, the proposed score identified these segments as high severity once geometry, dynamic anomalies, and structural context were considered together.

Although a full precision analysis is not possible because the available log does not include an exhaustive set of negative cases, the ranking behaviour of the proposed score still provides additional evidence of practical value. From an operational standpoint, maintenance actions are typically guided by prioritised inspection lists rather than by binary decisions alone. In this setting, the concentration of confirmed critical segments among the high-severity cases, together with the qualitative inspection of flagged segments not present in the ground-truth log, suggests that the score is capturing structurally sensitive locations and coherent multivariate anomalies rather than merely inflating classifications through isolated signal peaks. This is consistent with the observed increase in recall, which is not accompanied by indiscriminate expansion of high-severity classifications. This interpretation is also consistent with the contrast observed against the legacy threshold-based method.

5 Discussion

The case study highlights a practical weakness of threshold-based screening: it tends to react strongly to isolated peaks and, at the same time, to miss segments where moderate but coherent anomalies coincide with adverse geometry or structural context. This matters in maintenance planning, because inspection resources are limited and missed critical defects carry a higher operational cost than additional follow-up inspections.

The clearest quantitative result is the increase in recall for confirmed critical segments from 52.3% to 81.6%. Just as important is the qualitative reduction in inflated high-severity classifications, which suggests that the score is less sensitive to isolated signal excursions and better aligned with the underlying engineering context.

At the same time, the evidence should be interpreted within the limits of the dataset. The evaluation covers one month of measurements on one heavy-haul railway; recurrence was inactive ($R = 1$); and the validation log does not provide a complete set of negative cases. The current results therefore support the use of the ranking for screening and prioritisation, but they do not yet support a definitive assessment of network-wide false-alarm performance.

Future work should focus on calibrating the weights and scaling parameters with larger labelled datasets, repeated passes over the same segments, and more

complete maintenance outcome records. This would enable a more comprehensive assessment of precision-recall trade-offs, sensitivity to traffic and environmental conditions, and transferability to other railway networks.

6 Conclusions

This paper presented an interpretable composite severity score for railway track monitoring that combines multisensor dynamic response, structural context, geometric non-compliance, and recurrence. Applied to approximately 11,000 segments from a Brazilian heavy-haul railway, the method improved recall for confirmed critical segments from 52.3% to 81.6% relative to the legacy fixed-threshold workflow.

From a maintenance standpoint, the main practical value of the approach is not only higher sensitivity to critical cases, but also a more stable ranking under noisy operating conditions. By combining dynamic and geometric evidence with segment context, the score reduces dependence on isolated peaks and provides a more credible basis for maintenance prioritisation.

The present results correspond to a single case study and should be interpreted accordingly. Wider validation with repeated runs, more complete maintenance outcome logs, and transfer tests on additional networks is still required. Even so, the proposed formulation offers a transparent and operationally relevant alternative to purely threshold-based severity assessment.

Acknowledgements

The results presented in this paper are part of an Under-Rail Chair project funded by Vale S.A. The authors acknowledge Vale S.A. for its financial and technical support. They also thank Vale engineer Luciano Oliveira for his essential support in the resource and data acquisition activities that enabled the results presented in this paper.

References

[1] Associação Brasileira de Normas Técnicas (ABNT), "NBR 16.387:2016 — Geometria da Via Férrea: Desvios Padrão e Limites", Rio de Janeiro, 2016.

[2] T.C.G. de M. Silva, W.D. Queiroz, et al., "IoT and Big Data Analytics: Under-Rail Maintenance Management at Vitória–Minas Railway", Proceed-

ings of the IEEE International Conference on Machine Learning and Applications, 2023.

[3] P. Weston, C. Roberts, G. Yeo, E. Stewart, "Perspectives on railway track geometry condition monitoring from in-service railway vehicles", Vehicle System Dynamics, 53(7), 1063–1091, DOI:10.1080/00423114.2015.1034730, 2015.

[4] M. Molodova, Z. Li, A. Nunez, R. Dollevoet, "Automatic detection of squats in railway infrastructure", IEEE Transactions on Intelligent Transportation Systems, 15(5), 1980–1990, DOI:10.1109/TITS.2014.2307955, 2014.

[5] H. Guler, "Prediction of railway track geometry deterioration using artificial neural networks: a case study for Turkish state railways", Structure and Infrastructure Engineering, 10(5), 614–626, DOI:10.1080/15732479.2012.757791, 2014.

[6] I. Cárdenas-Gallo, C.A. Sarmiento, G.A. Morales, M.A. Bolívar, R. Akhavan-Tabatabaei, "An ensemble classifier to predict track geometry degradation", Reliability Engineering & System Safety, 161, 53–60, DOI:10.1016/j.ress.2016.12.012, 2017.

[7] J.M. Sadeghi, H. Askarinejad, "Development of track condition assessment model based on visual inspection", Structure and Infrastructure Engineering, 7(12), 895–905, DOI:10.1080/15732470903194676, 2011.

[8] Z. Shi, K. Wang, D. Zhang, Z. Chen, G. Zhai, D. Huang, "Experimental investigation on dynamic behaviour of heavy-haul railway track induced by heavy axle load", Transport, 34(3), 351–362, DOI:10.3846/transport.2019.10325, 2019.

[9] L. Breiman, "Random forests", Machine Learning, 45(1), 5–32, 2001.

[10] J. MacQueen, "Some methods for classification and analysis of multivariate observations", in "Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability", 1, 281–297, 1967.

[11] S. Galván-Núñez, N. Attoh-Okine, "A threshold-regression model for track geometry degradation", Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 232(10), 2459–2469, DOI:10.1177/0954409718777834, 2018.