Proceedings of the Seventh International Conference on
Railway Technology: Research, Development and Maintenance
Edited by: J. Pombo
Civil-Comp Conferences, Volume 15, Paper 17.1
Civil-Comp Press, Edinburgh, United Kingdom, 2026
ISSN: 2753-3239, doi: 10.4203/ccc.15.17.1

Establishing Baseline Expectations for Track Defect Detection via Repeatability Mapping of Instrumented Railway Vehicle Signals

A. C. Pires, T. T. Freire and A. A. Santos

Department of Mechanical Engineering, University of Campinas, Brazil

Abstract

This paper presents an exploratory data analysis workflow that establishes baseline, expected time-series behaviour for instrumented railway vehicles by accounting for track location and context, such as turnouts, bridges, and curve transitions. A multi-sensor dataset spanning six months is analysed, focusing on bogie-frame vertical accelerations and secondary-suspension displacement estimates referenced to track position. A novel data quality score is applied to remove intervals with low supply voltage or physical impossibilities, filtering the dataset to retain only locations with sufficient passes for repeatability analysis. The study demonstrates that turnout regions exhibit significantly wider distributions and more frequent large responses compared to regular track segments, necessitating separate analytical treatment. The primary achievement is the generation of repeatability maps with contextual overlays. These maps facilitate systematic inspection and establish testable expectations for subsequent defect detection models in scenarios where validated defect labels are sparse or unavailable.

Keywords: condition monitoring, drive-by monitoring, exploratory data analysis, anomaly detection, repeatability analysis, instrumented railway vehicle

1 Introduction

Railways seek higher productivity and safety while reducing operating costs. Continuous track condition monitoring supports this goal by providing evidence of degradation and localized defects across the network [1]. When deterioration is detected early, maintenance can be planned before defects grow, extending service life and lowering the risk of service disruptions and, in severe cases, derailments.

Instrumented railway vehicles (IRVs) are an increasingly used option for network-wide monitoring. An IRV acts as a moving probe: sensors installed on an in-service vehicle record its dynamic response to track excitations. These measurements can be used to infer track quality and to flag locations that repeatedly trigger unusual responses. IRVs also offer practical advantages over dedicated track geometry cars (TGCs) [2, 3, 4]. They can collect data without dedicated inspection runs, often at higher temporal frequency (e.g., daily), reducing unobserved time gaps. They also provide a complementary view of the track by using the vehicle response rather than only geometric measurements.

Many railways already collect IRV data in day-to-day operations, but converting these measurements into actionable insights remains challenging. Continuous monitoring produces large data volumes that must be processed, summarised, and communicated in a timely manner [5]. In many operational settings, validated defect labels are sparse or unavailable, and maintenance findings cannot be matched reliably to IRV passes because the vehicle may not traverse the same segment at the relevant time. As a result, maintenance records often cannot be used to trace the early evolution of a defect, and some locations flagged by IRV responses may have no corresponding record.

In addition, the measured response reflects both the track condition and several confounding factors. Without a baseline for these effects, manual analysis can be inconsistent, and downstream models can overreact to operational variability or misinterpret benign signatures as defect-like behaviour. For these reasons, exploratory data analysis (EDA) is a necessary step before applying model-based assessment or defect detection [6]. EDA supports this workflow by characterizing typical signal behaviour, revealing recurring location-specific signatures, and providing interpretable visual evidence that informs later modeling choices and supports expectations about model failure modes.

The key difficulty is that the relationship between track excitations and measured response is not one-to-one. Operating conditions such as load state and travelling speed change the vehicle dynamics and, therefore, the signal distribution [7, 8]. Track curvature can further modulate the response, particularly for sensors sensitive to lateral dynamics [2, 4]. In addition, fixed infrastructure elements (e.g., turnouts and bridges), and data quality issues can generate characteristic patterns and transient spikes. If these effects are not accounted for, they can be confounded with defect-like responses.

This paper focuses on the exploratory data analysis of IRV measurements collected during regular operations that precedes model-based defect detection [9]. Recognizing that validated defect labels are often sparse or unavailable in this context, the analysis is based on the hypothesis that spatially localized, repeatable large responses observed across multiple sensors and multiple passes are indicative of track-related issues. Repeatability maps are used to examine this hypothesis while explicitly considering confounding factors and known sources of variability. The resulting expectations are intended to constrain and interpret subsequent model outputs by clarifying what a downstream model should detect reliably and which patterns it should treat cautiously. To summarize, the study seeks to answer the following research questions:

1. Which track locations consistently exhibit large responses, and how stable are these responses over time under nominal operating conditions?

2. Which locations show repeatable anomalous responses across passes and sensors, and how do these patterns change under different operating conditions and known confounding factors?

Answering these questions clarifies observed signal behaviour and reduces avoidable false positives when developing track defect detection methods based on IRV data. Finally, a visual diagnostics tool is provided to consolidate these analyses and support systematic inspection of large-scale IRV datasets.

2 Dataset description

The measurements analysed in this work were obtained from the same instrumented railway vehicle (IRV) used by Pires et al. [2] using the measurements from October 2023 to March 2024. The present section provides a brief description of the instrumentation for context. Comprehensive details regarding sensor layout, calibration, and signal-conditioning procedures can be found in the previously cited work.

Figure 1 shows the wagon and the placement of the main sensors considered in this study. The IRV records the following quantities:

- **Bogie-frame vertical acceleration (4 sensors):** Four uniaxial accelerometers are installed on the side frames above the first and last wheelsets of the leading and trailing bogies (one sensor per side frame). These channels capture local vertical impacts.

- **Carbody acceleration (1 triaxial sensor):** A triaxial accelerometer is mounted on the carbody above the trailing bogie to measure global wagon motion, including dynamics related to hunting.

- **Secondary suspension displacement (4 sensors):** Four load cells are installed in parallel with an auxiliary spring in the secondary suspension, on the left and right sides of both bogies. The measured force is converted into an equivalent vertical displacement using the known spring stiffness.

- **Position and speed (GPS, 1 sensor):** A GNSS receiver provides position and instantaneous speed, which are used to spatially reference the other sensor signals.
- **Coupler (drawbar) force (2 sensors):** Two redundant strain-gauge bridges are mounted on the drawbar between adjacent wagons to measure axial forces associated with longitudinal train dynamics.
- **Brake pipe pressure (1 sensor):** A pressure transducer measures brake pipe pressure, enabling identification of braking and release events.
- **Battery voltage (1 sensor):** The instrumentation supply voltage is logged continuously and later used as an indicator of potential data-quality issues related to power variations.

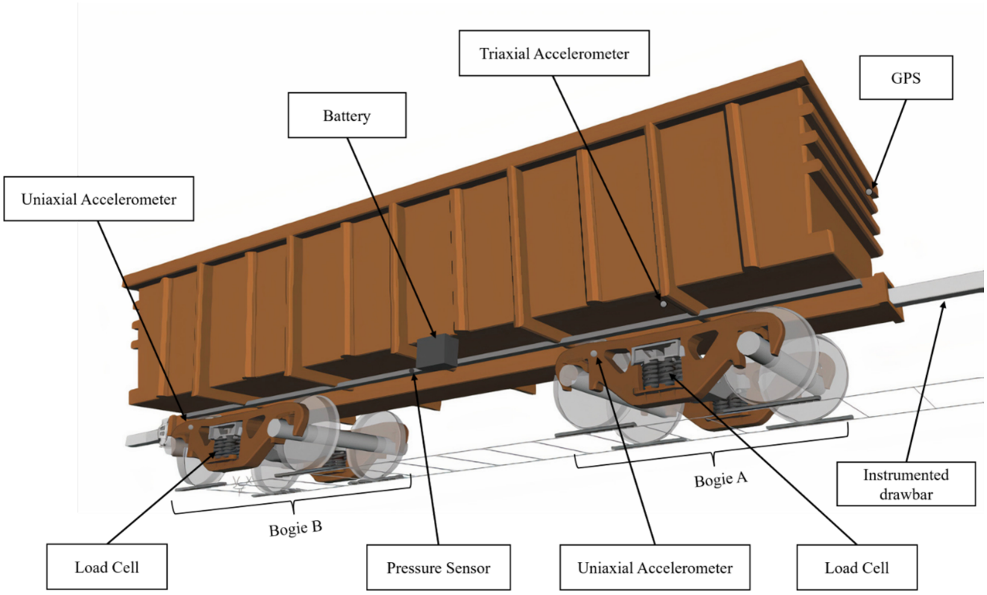

Figure 1: Sensor locations of the instrumented railway vehicle.

In this study, drawbar forces are not used in the analysis because they are not direct indicators of track condition. The analysis focuses mainly on the four bogie-frame vertical accelerations and four secondary-suspension displacement measurements, which are expected to be more sensitive to track-related excitation and defect-like responses. Table 1 summarises the variables used in this study.

2.1 Power supply and data-quality screening

The IRV instrumentation is powered by a rechargeable 12 V battery and a pneumatic turbine. While the wagon is moving, the turbine provides most of the electrical power and recharges the battery by drawing a small amount of pressure from the brake system. This pressure draw does not affect braking performance. When the wagon is

Source	Sensor	Name	Description	Units
IRV	**Displacement**	SS_Z_Filt_AL	Secondary suspension displacement (A,L).	mm
		SS_Z_Filt_AR	Secondary suspension displacement (A,R).	
		SS_Z_Filt_BL	Secondary suspension displacement (B,L).	
		SS_Z_Filt_BR	Secondary suspension displacement (B,R).	
	Uniaxial accelerometer	UA_Z_AL	Vertical acceleration (A,L).	m/s^2
		UA_Z_AR	Vertical acceleration (A,R).	
		UA_Z_BL	Vertical acceleration (B,L).	
		UA_Z_BR	Vertical acceleration (B,R).	
Notes: A/B = leading/trailing bogie; L/R = left/right; Z = vertical; Filt = filtered.				

Table 1: IRV sensor variables.

parked or moving slowly, the turbine cannot provide enough power and the battery supplies the system. As a consequence, changes in brake pressure and battery voltage can influence the sensor readings. These effects have been linked to recurring data-quality problems in the IRV measurements.

To handle these issues during exploratory analysis, a Data Quality Score (DQ_Score) was proposed to summarize how reliable each data interval is. The score ranges from 0 to 1, where 1 means no issues were detected and 0 indicates severe degradation. The DQ_Score combines checks based on battery voltage and brake pressure with simple signal-validity checks, such as rejecting secondary-suspension displacement estimates that are physically impossible. In practice, values below 0.7 are typically unreliable, while values between 0.7 and 0.85 often indicate data that required manual correction to avoid discarding usable segments. The DQ_Score is then used to filter or flag low-quality intervals so that visual diagnostics are not dominated by measurement artifacts.

2.2 Data filtering

The visual analysis targets spatially localized responses that are repeatable across sensors and across IRV passes. To support this objective, the dataset is filtered to retain locations with sufficient coverage and with trips that are close enough in time to assess repeatability. The filtering criteria are:

1. DQ_Score above 0.80 (see Section 2.1);

2. At least 6 trips available for the same location;

3. No gap larger than 2 months between successive trips;

4. At least 4 trips within any 2-month window.

After filtering, approximately 10% of the original dataset remains. Criterion 1 removes about 50% of the data, and Criteria 2-4 further reduce the retained subset to 10%. This screening increases confidence that the remaining data are both reliable and suitable for assessing repeatability.

The filtered dataset is then divided into two track segment types: regular track segments (EH) and turnout regions (RH). The EH group includes tangents and curves, while the RH group includes switch regions. After filtering, the EH group contains 24 540 samples and the RH group contains 2 851 samples. The exploratory analysis in this work focuses on these two groups.

3 Exploratory data analysis

This section presents the exploratory analysis performed on the dataset described in Section 2.

3.1 Track locations prone to defects

Figure 2 presents raincloud plots for the sensors listed in Table 1, comparing regular track segments (EH) and turnout regions (RH). A raincloud plot combines a boxplot, a probability density estimate, and a swarm plot of individual samples. Focusing on the swarm plots, it is evident that across all sensors, the RH distributions are more dispersed than the EH distributions and exhibit higher upper percentiles, indicating more frequent high-magnitude responses in turnout regions. While this behaviour is observable in the swarm plots of the combined distributions, it is not reflected in the density estimates. Instead, the combined distribution appears nearly identical to that of the EH group. This occurs because the EH group contains a significantly larger number of samples; consequently, aggregate views are dominated by EH behaviour, which effectively masks the RH response range. This reinforces the need to handle turnout regions separately from tangents and curves during defect-oriented analysis. Specifically, RH responses should be evaluated against other RH locations rather than the full dataset. Furthermore, turnouts are known to experience higher impact forces than regular tangent and curved tracks [10], increasing the likelihood of damage and wear.

If the EH group is further divided into tangents and curves, the difference relative to RH becomes more pronounced, as shown in Figure 3. The accelerometer distributions show clearer separation than the displacement distributions, primarily because the accelerometers exhibit more frequent large-magnitude responses. These results indicate

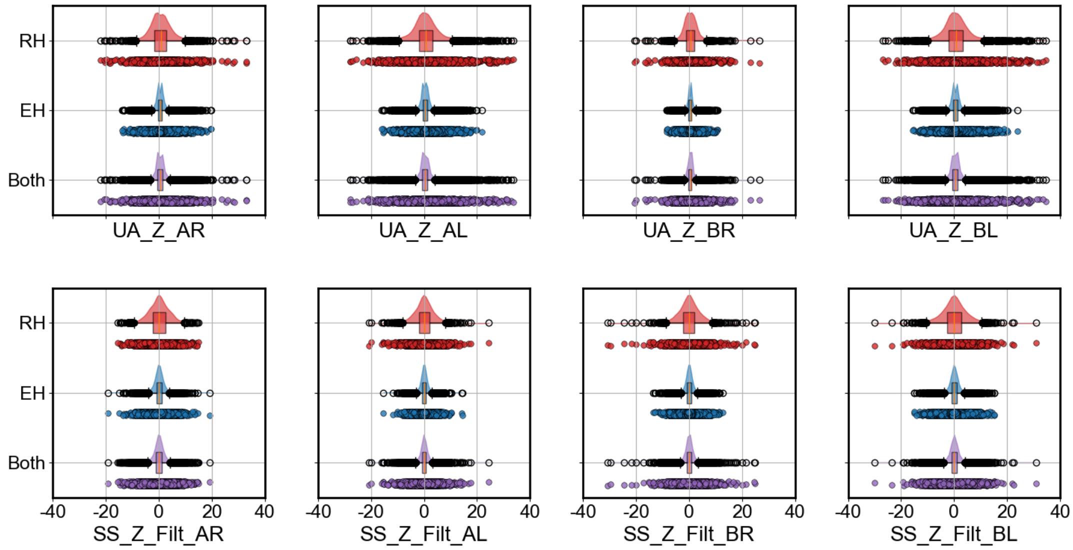

Figure 2: Raincloud plot of the sensors separated by EH and RH groups.

that turnout regions should be excluded from model training when the objective is to learn baseline behaviour for regular track.

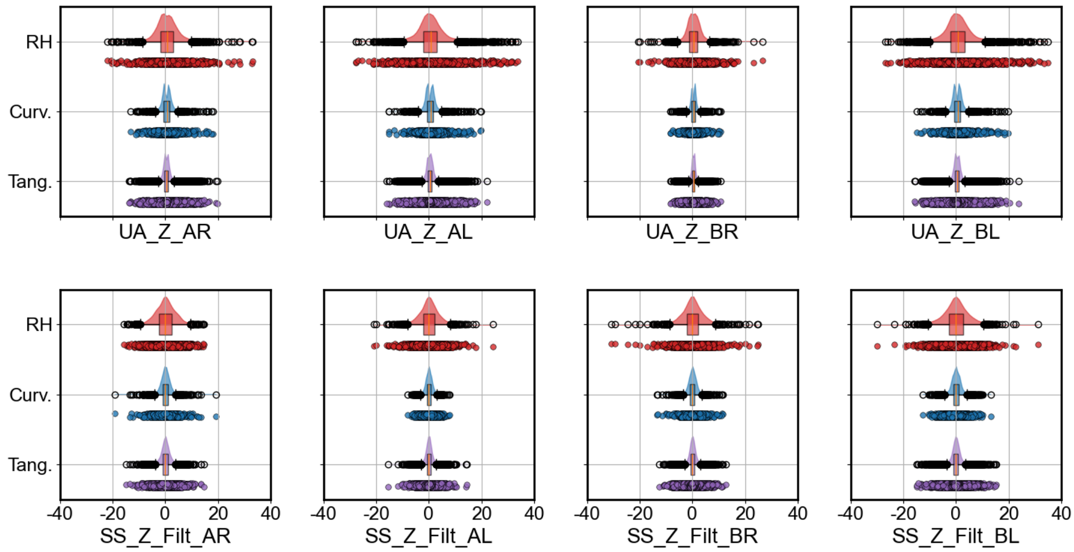

Figure 3: Raincloud plot of the sensors with EH split into tangents and curves, and RH shown separately.

4 Inspection of high response locations

With the data quality obtained from the pre-processing of section 2.1, visual analysis of the track locations with the highest responses can be performed. This analysis is important to understand if there are any specific characteristics that could explain the high responses, such as track defects or specific track features.

Figure 4 presents a repeatability heatmap for section EH 45/46 on line 2. Notably, starting from trip 7 (early December 2023), persistent large dynamic responses are observed in the central portion of the segment, with trips 14 to 18 showing consistent high-magnitude readings. This repeatability suggests the presence of a potential track defect. Additionally, large responses at the beginning and end of the segment are likely attributed to turnouts located near the section boundaries.

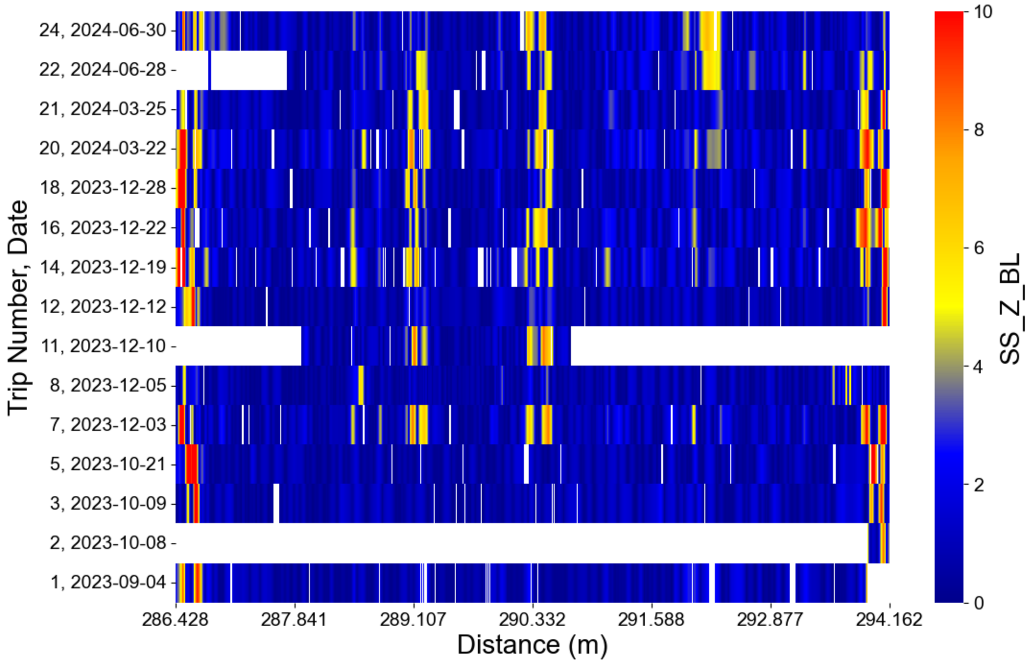

Figure 4: Repeatability heatmap of EH 45–46 and adjacent RH sections for all lines for wagon 2.

The heatmap also reveals several white (missing) regions, which occur due to two main factors:

1. The wagon initially travels along line 1 and switches to line 2 via a turnout, as seen in trips 1 and 2;

2. GPS signal loss during passage through EH sections.

While the heatmap provides a general intuition about the data, it lacks a detailed spatial context and global view of all sensors data, both necessary to interpret the causes of large responses. Figure 5 complements this view by showing time series data for the four secondary suspension sensors and location context, along with the maximum left-positioned vertical acceleration values `UA_Z_L` for each trip between December 2023 (trips 7–18) and March 2024 (trip 20).

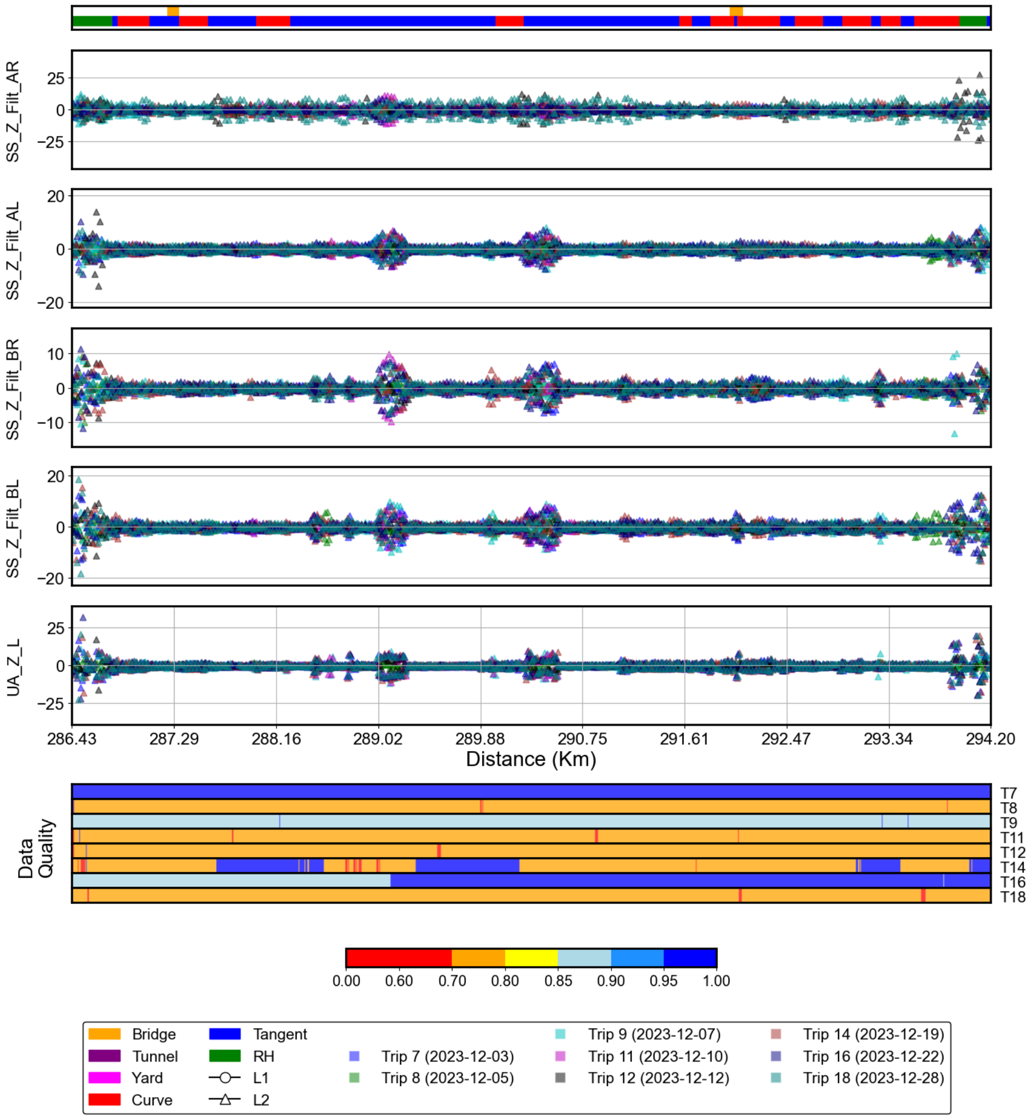

Figure 5: Time series view of Figure 4, enriched with contextual information for EH 45/46 Line 2.

The top bars in Figure 5 indicates the track context for each position, including the segment type (tangent, curve or RH - turnout), and the location of features such as bridges, tunnels or yards.

From the time series, the large responses at the segment boundaries correspond to turnouts, which is an expected behaviour. However, to assess whether these responses are abnormal, data from other turnouts must be analysed for comparison. Preliminary analysis suggests that high responses at turnouts are common.

Of particular interest are two locations in the middle of the segment, between km 288.99 and 290.71, which consistently show high responses across several sensors and multiple trips. The anomalies that occur under similar operating conditions are strong candidates for further inspection by the maintenance team.

However, some anomalies may be false positives caused by data quality issues rather than actual track defects. To assess this, the bottom plot of Figure 5 shows the normalized DQ_Score introduced in Section 2.1. Most trips exhibit moderate to significant data quality issues. For example, orange-colored regions reflect miscalibration in one of the secondary suspension sensors, while light blue indicates low battery voltage (below 10.5 V).

This is particularly evident in trip 18, where the sensor SS_Z_Filt_AR displays higher oscillations compared to the other three sensors, explaining its low DQ_Score. Nevertheless, the remaining sensors maintain sufficient quality to support defect detection in this case.

When filtering data quality to be above 0.80, the overwhelming majority of issues that affect the measurements are removed. Figure 6 shows the results after applying this filter. Note that most of the issues in SS_Z_Filt_AR were resolved and the segment identified earlier as potential anomalies remains, but trips 11 and 18 had to be removed from the dataset. Note that in the fourth curve, there are sensor responses that coincide with the transition between segments. This needs to be investigated further since this appears to be caused by a severe change in curvature that leads to impact.

5 Conclusions

This paper presented an exploratory data analysis workflow for instrumented railway vehicle measurements collected during regular operation, specifically addressing scenarios where validated defect labels are sparse or unavailable. The process involved establishing a baseline of expected time-series behaviour as a function of track location and context, and using this baseline to form testable expectations for subsequent model-based defect detection.

Reliable visual inspection requires data screening before comparing locations. Power-supply effects and simple signal-validity checks were summarised through a data quality score, enabling low-quality intervals to be flagged or removed before analyzing high-response regions. Additional filtering retained only locations with enough passes and temporal proximity, supporting repeatability analysis at the same track position.

The results show that turnout regions exhibit wider distributions and more frequent large responses than regular track segments, motivating separate analysis and separate baselines. When the objective is to learn baseline behaviour for tangents and curves, these findings support excluding turnout regions from the training set. Finally, repeatability maps with contextual overlays provide a practical way to inspect candidate locations, relate responses to track context (e.g., turnouts, bridges, and curve transitions), and identify cases where apparent anomalies are explained by measurement artifacts or known track features.

Although the workflow does not confirm defects without ground truth, it helps

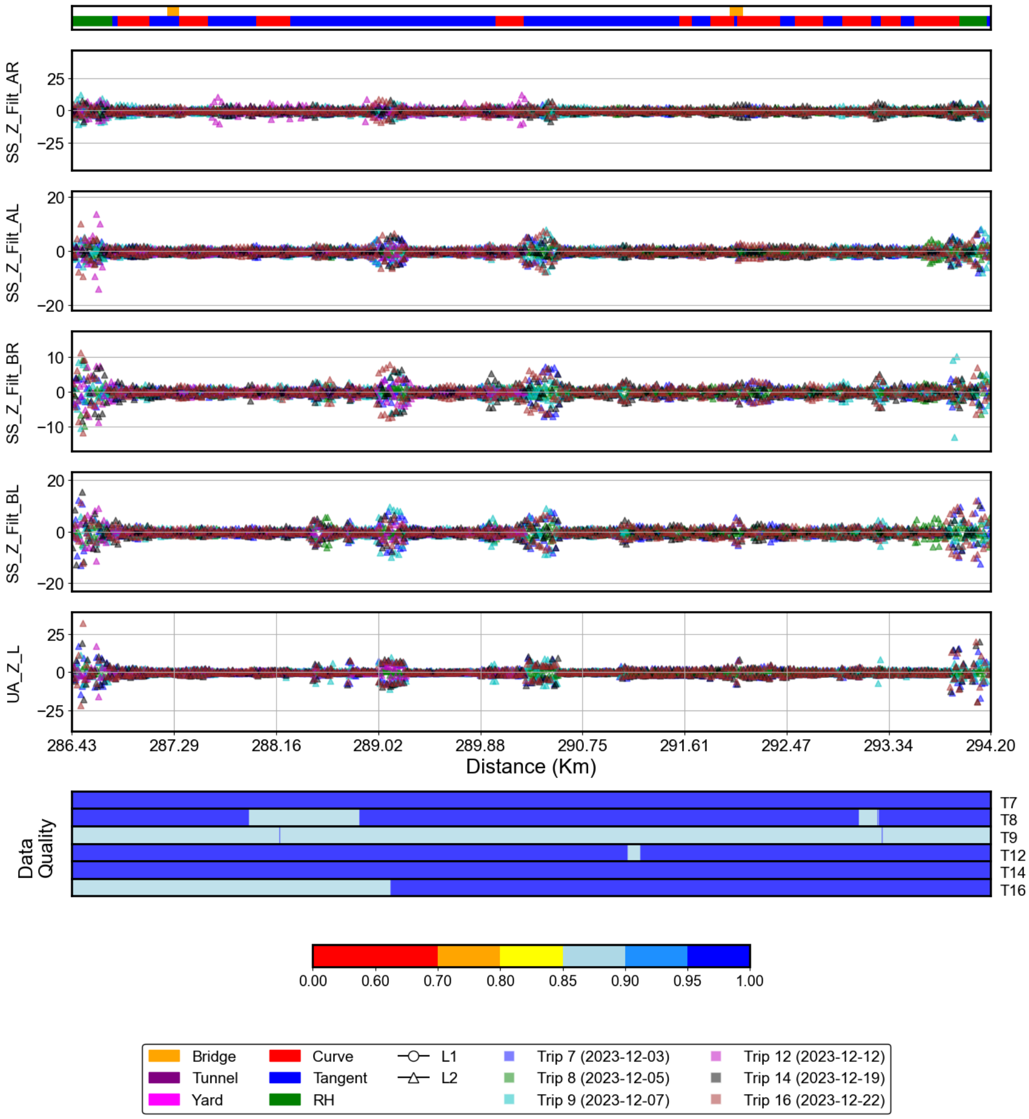

Figure 6: Time series view of the repeatability heatmap with `DQ_Score` above 0.80 for EH 45/46 Line 2.

reduce the risk of avoidable false positives in later modeling, helps prioritize candidate locations for inspection, and supports interpretation of subsequent defect detection model outputs under real operating conditions.

Finally, the answers to the research questions posed in Section 1 are summarised as follows:

RQ1: *Which track locations consistently exhibit large responses, and how stable are these responses over time under nominal operating conditions?*

Turnout locations consistently exhibit higher-magnitude responses compared to tangents and curves, as evidenced by the distributions in Figure 2 and the repeatability heatmap in Figure 4.

RQ2: *Which locations show repeatable anomalous responses across passes and sensors, and how do these patterns change under different operating conditions and known confounding factors?*

Using the proposed methods and visual representations, two track segments were identified with anomalous behaviour across multiple train passes and sensors. One of these segments coincides with a transition from a curve to a tangent, although further analysis is required to establish a definitive correlation. Notably, these regions were discernible even without filtering the data according to the proposed DQ score, though the application of the filter made the anomalies significantly more pronounced.

Acknowledgments

The authors would like to thank VALE S.A. for funding the project through Grant 03-P-18453/2019 and Grant 23068.037324/2019-32. Also, to CNPq (National Council for Scientific and Technological Development), Grant 303582/2023-5.

References

[1] I. Bravo, U. Alvarado, J. Nieto, P. Ciáurriz, "On-Board Accelerometers in Railway Track Condition Monitoring. A Systematic Review", Transportation Research Interdisciplinary Perspectives, 33, 101572, 2025.

[2] A.C. Pires, M.C.A. Viana, L.M. Scaramussa, G.F.M. Santos, P.G. Ramos, A.A. Santos, "Measuring Vertical Track Irregularities from Instrumented Heavy Haul Railway Vehicle Data Using Machine Learning", Engineering Applications of Artificial Intelligence, 127, 107191, 2024.

[3] T.L. Chong, M.N. Awad, N. Nadarajah, W.K. Chiu, S.N. Lingamanaik, G. Hardie, R. Ravitharan, H. Widyastuti, "Defining Rail Track Input Conditions Using an Instrumented Revenue Vehicle", Procedia Engineering, 188, 479-485, 2017.

[4] G.J. Yeo, "Monitoring Railway Track Condition Using Inertial Sensors on an In-Service Vehicle", PhD thesis, University of Birmingham, 2017.

[5] P. Reichl, H.V. Ribeiro, G.F.M. Santos, "Improving the Efficiency of Vale's EFC Line Through the Use of Continuously Measured Operational Data", in "International Heavy Haul Conference", International Heavy Haul Association, Perth, Australia, 2015.

[6] M. Komorowski, D.C. Marshall, J.D. Salciccioli, Y. Crutain, "Exploratory Data Analysis", Secondary Analysis of Electronic Health Records, Springer, Cham, Switzerland, 185-203, 2016, DOI:10.1007/978-3-319-43742-2_15.

[7] F. Balouchi, A. Bevan, R. Formston, ”Development of Railway Track Condition Monitoring from Multi-Train In-Service Vehicles”, Vehicle System Dynamics, 59, 1397-1417, 2021.

[8] H. Ono, H. Tsunashima, T. Takata, S. Ogata, ”Development and Operation of a System for Diagnosing the Condition of Regional Railways Tracks”, Mechanical Engineering Journal, 10, 22-00239, 2023.

[9] K. Shaukat, T.M. Alam, S. Luo, S. Shabbir, I.A. Hameed, J. Li, S.K. Abbas, U. Javed, ”A Review of Time-Series Anomaly Detection Techniques: A Step to Future Perspectives”, Advances in Information and Communication, Springer, Cham, Switzerland, 865-877, 2021, DOI:10.1007/978-3-030-73100-7_60.

[10] E. Zhang, C.L. Pun, A. Hiew, W. Yan, ”Dynamic Response and Wear Analysis of a Swing Nose Crossing in Heavy Haul Railways”, Railway Engineering Science, 33, 192-215, 2025.