Equivalent: Jurnal Ilmiah Sosial Teknik
Volume 8, Issue 5, 1978-1992
e_ISSN: 2775-0833
https://jurnalequivalent.id/index.php/jequi/index
DOI: doi.org/10.59261/jequi.v8i5.477

Validation of a Low-Cost Multisensor Ultrasonic Road Roughness Measurement System for International Roughness Index Assessment

Eko Wahyu Utomo
Universitas Langlangbuana, Indonesia

Siegfried
Universitas Langlangbuana, Indonesia

***Corresponding author:**
Eko Wahyu Utomo, Universitas Langlangbuana, Indonesia.
✉utomoeko042@gmail.com

Article Info:

Article history:
Received: September 19, 2026
Revised: October 19, 2026
Accepted: October 21, 2026
Available Online: October 26, 26

Keywords:
Inertial Measurement;
International Roughness Index;
Method-Comparison Validation;
Pavement Roughness;
Ultrasonic Sensor;

This is an open access article under the CC BY-SA license.

Copyright © 2026 by Author. Published by Politeknik Siber Cerdika International

Abstract

Background: Pavement roughness assessed through the International Roughness Index (IRI) is a key functional indicator for road maintenance prioritization, yet certified inertial profilers remain financially inaccessible for many road authorities, particularly in developing countries.
Objective: This study evaluates the measurement agreement of a low-cost vehicle-mounted Ultrasonic Surface Roughmeter (USR) that integrates ultrasonic distance sensing, an accelerometer, and GPS, against certified reference IRI measurements.
Methods: System-level agreement was assessed using 15 paired 100-m observations drawn from five 300-m road sections, benchmarked against a Gap Man Working Profilometer and a Vehicle Mounted IRI Measurement System, and analyzed using Pearson correlation, linear regression, absolute and squared error metrics, and Bland-Altman limits of agreement.
Results: USR IRI ranged from 2.79 to 7.54 m/km against a reference range of 2.88-7.50 m/km. The analysis produced Pearson r = 0.9973, R^2 = 0.9946, MAE = 0.112 m/km, RMSE = 0.133 m/km, MAPE = 2.45%, and a mean bias of +0.043 m/km, with a regression of USR IRI = 1.017(reference IRI) - 0.038 and Bland-Altman limits of agreement of -0.212 to +0.298 m/km.
Conclusion: The USR shows strong agreement with certified reference measurements within the tested range, supporting its use as a low-cost pavement-roughness screening technology, although the limited number of road sections means that full interchangeability with certified profilers cannot yet be claimed; broader multi-speed, multi-road, and repeated-run validation is recommended before operational deployment. The evidence is limited to the five tested road sections and does not establish full interchangeability with certified profilers.

To cite this article: Utomo, E. W. & Siegfried (2026). Validation of a low-cost multisensor ultrasonic road roughness measurement system for International Roughness Index assessment. *Equivalent: Jurnal Ilmiah Sosial Teknik, 8*(5), 1978-1992. https://doi.org/10.59261/jequi.v8i5.477

INTRODUCTION

Pavement roughness is one of the most consequential functional indicators in road infrastructure management because longitudinal surface irregularity directly affects ride quality, vehicle operating costs, fuel consumption, and, ultimately, the prioritization of scarce maintenance budgets across an entire road network (Dutta et al., 2026; Loprencipe & Zoccali, 2017; Nguyen et al., 2019; Žuraulis et al., 2021). The International Roughness Index (IRI) has become the de facto standard for quantifying this irregularity because it is derived through a standardized quarter-car simulation of longitudinal profile data and is, in principle, comparable across different profiling instruments, provided that the underlying profile has been measured with adequate accuracy. Certified inertial profilers remain the reference technology for collecting

longitudinal profiles at network speeds, and a synthesis by the National Academies of Sciences, Engineering, and Medicine (2018) makes clear that reliable IRI measurement depends not merely on possessing such an instrument but also on sustained attention to repeatability, reproducibility, and formal certification procedures.

The Federal Highway Administration (2022) similarly identifies inertial profilers as an accepted operational standard for network-level pavement condition assessment in the United States. This challenge is particularly acute for decentralized or resource-constrained road authorities across much of the developing world, where certified inertial profiling systems commonly cost tens to hundreds of thousands of U.S. dollars, placing routine, network-wide roughness monitoring beyond the practical reach of many provincial and municipal agencies, including a substantial share of Indonesia's regional road authorities. Although affordability provides a practical motivation for the present study, Indonesian procurement budgets are not quantified herein; therefore, affordability is treated as a practical motivation rather than an established national cost statistic. IRI was selected as the principal outcome measure because it is a standardized, profile-based index designed to support comparability across road-roughness measurement systems, whereas other pavement condition indicators capture different dimensions of pavement distress rather than longitudinal ride quality (ASTM International, 2022; ISO, 2016; Sayers et al., 1986).

This financial barrier has motivated a rapidly growing body of research into low-cost alternatives that replace dedicated profiling hardware with distributed, comparatively inexpensive sensing technologies. One prominent line of research exploits the accelerometer and GPS receiver already embedded in ordinary smartphones. Zhang et al. (2022) developed a smartphone-based methodology that relates vehicle vertical acceleration to IRI through a quarter-car representation, while Bai et al. (2025) further refined smartphone-based roughness evaluation methods intended for routine agency use in China. These response-type approaches are attractive precisely because the sensing hardware is already ubiquitous, but they share a structural limitation: because they infer roughness from a vehicle's dynamic response rather than directly measuring pavement geometry, their accuracy is inherently affected by vehicle suspension characteristics, tire condition, loading, and travel speed. Several smartphone-based studies explicitly acknowledge these factors as sources of measurement variability. A second, increasingly prominent line of research applies machine learning either to predict IRI from indirect signals or to improve the robustness of response-type estimates.

Kaloop et al. (2023) applied novel machine-learning techniques to IRI prediction for flexible pavements; Jalili et al. (2024) trained an artificial neural network using smartphone-based crowdsourcing data; Sang et al. (2024) proposed a data-driven, smartphone-based IRI estimation framework validated using distributed vehicle fleets; and Chen et al. (2025) introduced a physics-guided neural network that explicitly incorporates prediction uncertainty and stability rather than treating IRI estimation as a purely black-box regression problem. Wu et al. (2024) pursued a complementary approach by proposing a rapid vehicle-dynamics calibration procedure that allows a vibration-based roughness measurement system to be recalibrated across different vehicles without fully re-deriving the underlying quarter-car model.

Taken together, this literature demonstrates that indirect, response-type sensing can approximate IRI under controlled conditions, but it also repeatedly highlights the same underlying tension: accuracy depends on the specific vehicle, its dynamic state, and the quality of the model used to distinguish genuine road-induced excitation from vehicle-induced noise. Across these studies, the key distinction is methodological: smartphone-based and machine-learning systems infer roughness from vehicle response and therefore require vehicle- and speed-sensitive modeling or calibration, whereas direct geometric sensing measures the distance between the sensor and the road surface more directly. The present study of the Ultrasonic Surface Roughmeter (USR) is positioned in the latter category and evaluates measurement agreement rather than predictive-model accuracy.

A comparatively smaller body of research instead investigates direct, noncontact geometric sensing as an alternative to inferring roughness from vehicle response. Behera et al. (2021) developed a low-cost ultrasonic roughometer (Us-R) that directly measures the distance between the sensor and the road surface along the wheel path using an ultrasonic transducer

mounted on a bicycle or motorcycle. They reported IRI values within 5% of those obtained using a reference calibrated with the Machine for Evaluating Roughness using Low-cost Instrumentation (MERLIN). To the best of the present author's knowledge, this remains the closest prior analogue to the system evaluated in this study. However, its validation relied primarily on comparisons with a single reference instrument rather than a comprehensive method-comparison analysis of bias and limits of agreement. More recently Alatoom & Smadi (2026) advanced this approach by integrating an inertial measurement unit (IMU) with pulsed coherent radar to separate actual pavement geometry from vehicle-induced motion in the frequency domain.

Their system achieved strong agreement with reference measurements across 12 road segments at a total component cost of approximately US$200. Their sensor-fusion architecture illustrates a broader shift in low-cost sensing research away from single-sensor response measurement and toward multisensor systems that combine a direct geometric measurement channel with inertial and spatial context. Múčka (2016), in a widely cited methodological review, cautions that regardless of the sensing architecture adopted, the quality of the resulting IRI value depends on the quality of the profile acquisition, spatial referencing, and signal processing performed before the IRI calculation. This caution is particularly relevant to ultrasonic- and radar-based systems, whose accuracy can be affected by sensor alignment, surface texture, acoustic or electromagnetic reflection characteristics, and target geometry in ways that purely inertial systems do not experience.

Compared with the Us-R developed by Behera et al. (2021), which employed an ultrasonic transducer mounted on a bicycle or motorcycle and was evaluated primarily against a MERLIN-calibrated reference through correlation and error comparisons, the present prototype combines ultrasonic distance measurements with accelerometer and GPS data on a vehicle-mounted platform. It evaluates measurement agreement against two reference systems using bias estimates, error metrics, regression diagnostics, and Bland–Altman limits of agreement. This comparison is restricted to features explicitly reported in the cited studies and does not imply superior hardware performance.

A further methodological gap runs through much of this literature and motivates the analytical approach adopted in the present study. Many validation studies of low-cost roughness measurement devices report a single Pearson correlation coefficient or, at most, a coefficient of determination ((R^2)) from linear regression as evidence of validity. Bland and Altman, in the paper that introduced what has since become a standard statistical framework for comparing two measurement methods, demonstrated that a high correlation coefficient does not, by itself, establish agreement or interchangeability between two instruments because correlation is sensitive to the range of values sampled and insensitive to systematic bias.

Botha et al. (2025) recently extended this argument beyond its original clinical measurement context. Through a chemical engineering case study, they showed that Bland–Altman analysis can reveal an engineering instrument's true agreement characteristics in situations where a parity plot or a simple correlation coefficient would be misleading, particularly when the underlying measurements are unevenly distributed across their range. Despite this well-established statistical caution, Bland–Altman limits-of-agreement analysis remains uncommon in the published validation literature on low-cost ultrasonic and multisensor pavement roughness measurement devices. Most such studies continue to rely on correlation or regression coefficients as their principal evidence of validity. The methodological rationale for the present study follows the original Bland–Altman framework, which distinguishes association from agreement and cautions that correlation alone cannot establish interchangeability between two measurement methods (Martin Bland & Altman, 1986).

Against this background, the present study addresses two related research gaps. First, although direct-sensing ultrasonic roughness measurement devices have been proposed Behera et al. (2021) and multisensor fusion architectures have been validated with encouraging results in other national contexts Alatoom & Smadi (2026), a low-cost ultrasonic device that additionally integrates accelerometer and GPS channels and undergoes full system-level validation against reference instrumentation—the GAPMAN Walking Profilometer and a Vehicle-Mounted IRI Measurement System—has not, to the best of the present author's knowledge, been reported with

validation using the comprehensive statistical toolkit recommended in method-comparison research. Second, the study deliberately treats Pearson correlation as only one component of validation evidence and recognizes that it is insufficient on its own. Following the statistical principles established by Bland and Altman and reaffirmed in engineering contexts by Botha et al. (2025), the study combines regression diagnostics, absolute and relative error metrics, subgroup performance assessments across two IRI ranges, and Bland–Altman limits-of-agreement analysis.

This combined approach enables reviewers and prospective users to assess not merely whether the USR tracks reference IRI measurements but also how closely it does so, in which direction any systematic discrepancies occur, and what practical implications those discrepancies may have for screening-level pavement management decisions. Accordingly, the objective of this paper is to quantify the agreement between USR measurements and reference IRI measurements using this combined statistical approach and to determine, with appropriate caution regarding sample size and study scope, whether the evidence supports the use of the USR as a low-cost screening and monitoring instrument for road authorities that cannot presently justify the capital expenditure required for a certified inertial profiler. The technical contribution is the validation of an ultrasonic–accelerometer–GPS measurement chain, whereas the methodological contribution lies in applying a multimetrics method-comparison framework rather than relying on correlation alone. Operationally, the study has two objectives: (1) to quantify agreement between the USR and the reference systems across the 15 paired road segments using correlation, error, bias, and limits-of-agreement statistics; and (2) to determine whether the observed error structure is sufficiently promising to justify the USR as a preliminary screening technology under the tested conditions, without claiming certified interchangeability.

METHOD

This study employed a quantitative method-comparison design to evaluate the measurement agreement between a prototype low-cost Ultrasonic Surface Roughmeter (USR) and certified reference instrumentation for International Roughness Index (IRI) measurement. This approach is consistent with the international literature on validating alternative pavement-roughness sensing technologies. The USR integrates an ultrasonic distance sensor, an accelerometer, a GPS receiver, a microcontroller, and a data-acquisition system mounted on a survey vehicle. The ultrasonic sensor provides a noncontact estimate of the vertical distance between the sensor and the pavement surface along the wheel path; the accelerometer records vehicle motion to provide synchronization and dynamic context; and the GPS receiver supplies spatial referencing and independent vehicle-speed information.

The prototype was designed for dynamic operation without requiring the survey vehicle to stop, and field testing in the source study was conducted at approximately 30–40 km/h, consistent with typical low-speed network-screening surveys reported elsewhere in the low-cost sensing literature (Behera et al., 2021). The validation workflow comprised five stages: sensor data acquisition during field traversal; synchronization of ultrasonic, inertial, and GPS records; conversion of the retained surface-profile signal into segment-level IRI values; one-to-one matching of each 100-m USR result with the corresponding reference IRI measurement; and statistical method-comparison analysis. The retained manuscript records document the functional sensor types and mounting configuration but do not preserve complete manufacturer and model information, resolution, measurement range, or sampling-frequency metadata for every component. These specifications are therefore not reconstructed here to avoid introducing unsupported details.

Before the system-level validation reported in this study, the USR's individual sensing components were characterized separately. Repeated ultrasonic distance measurements across 14 experimental series, comprising 1,400 observations in total, yielded relative errors ranging from approximately −2.10% to +4.84%. Accelerometer–GPS synchronization achieved a reported correlation of 100% in the tested dataset, while comparisons between GPS-derived coordinates and speed measurements from a speedometer yielded correlations of 99.84% and 99.80%, respectively. These sensor-level results are presented as supporting evidence of component reliability rather than as the primary validation outcome. The focus and principal contribution of the present study are the system-level agreement between the final IRI values calculated by the

USR and the corresponding reference IRI measurements under field conditions. These component checks are reported exclusively as prevalidation characterization. The retained records do not provide sufficient procedural detail to reconstruct the exact repetition conditions or the operational definition underlying the reported 100% synchronization correlation. Consequently, this value is not used as evidence of system-level measurement agreement. System-level validity is assessed on the basis of the paired road-section comparisons reported below.

Field validation was conducted on five 300-m road sections, each subdivided into three consecutive 100-m segments, yielding 15 paired observations in total. Three sections, designated SP2–SP4, were compared against a GAPMAN Walking Profilometer, while two sections, designated SRG1 and SRG2, were compared against a Vehicle-Mounted IRI Measurement System. Both were treated as reference technologies within the validation framework, consistent with the general principles of reliable IRI measurement discussed by the National Academies of Sciences, Engineering, and Medicine. The instrument-to-section mapping used in this study follows the most recent calculation record available at the time of analysis. Because an earlier project description proposed a different mapping, this correspondence must be reverified against the original field log before the manuscript is finalized for submission. Any discrepancy should be resolved in favor of the primary field record rather than a subsequently reconstructed summary document.

This study did not involve human or animal participants and therefore did not require approval from a research ethics committee on that basis. All field data were nevertheless collected and archived in accordance with the author's institutional data-management standards. The retained dataset identifies the road sections only by the codes SP2–SP4 and SRG1–SRG2 and does not preserve verified geographic coordinates, pavement age, pavement type, or detailed surface-condition descriptors. These characteristics are therefore treated as unavailable rather than inferred. The reference-instrument allocation used in the present analysis follows the final retained paired dataset. Because independent verification against the original field log is unavailable in the manuscript record, subgroup comparisons are interpreted descriptively and are not used to claim that one reference instrument or road group performed better than the other.

Agreement between the USR and the reference instruments was evaluated using a combination of statistical techniques selected specifically to address the limitations of correlation-only validation identified in the Introduction. Pearson's correlation coefficient and the coefficient of determination ((R^2)) were calculated as baseline measures of linear association. Mean absolute error (MAE) and root-mean-square error (RMSE) were calculated to quantify the average absolute deviation and the influence of larger individual deviations, respectively. Mean absolute percentage error (MAPE) was calculated to express the absolute error relative to the reference values, thereby facilitating comparisons between the two subgroups, which had different IRI ranges. Ordinary least-squares linear regression of USR-derived IRI values on reference IRI values was used descriptively to characterize scaling and offset, rather than to establish an independently validated calibration equation. Finally, following the method-comparison framework established by Bland and Altman and its documented applicability to nonclinical engineering measurement contexts (Botha et al., 2025), Bland–Altman limits of agreement were calculated from the mean and standard deviation of the paired differences to characterize systematic bias and the dispersion of individual measurement differences across the range of tested IRI values.

Let $d_i = USR_i - Reference_i$ denote the difference for each of the *n = 15* paired observations. MAE was calculated as

$$MAE = (1/n) \sum_{i=1}^{n} |d_i|,$$

RMSE as

$$RMSE = \sqrt{[(1/n) \sum_{i=1}^{n} d_i^2]},$$

and MAPE as

$$MAPE = (100/n) \sum_{i=1}^{n} |d_i / Reference_i|.$$

Bland–Altman bias was calculated as the mean of the paired differences:

$$\text{Bias} = \bar{d} = (1/n) \sum_{i=1}^{n} d_i,$$

and the 95% limits of agreement were calculated as

$$\text{LoA}_{95}\% = \bar{d} \pm 1.96 \, s_d,$$

where *s_d* is the sample standard deviation of the paired differences (Martin Bland & Altman, 1986). All 15 retained pairs were complete and included in the analysis; no observations were excluded as outliers. Regression is reported solely as a descriptive diagnostic of scaling because both instruments are subject to measurement error, and any calibration equation would require independent calibration and validation datasets.

RESULTS AND DISCUSSION

Results

Table 1 presents the complete paired-observation dataset. Across all 15 100-m segments, USR-derived IRI values ranged from 2.79 to 7.54 m/km, closely tracking the reference IRI values, which ranged from 2.88 to 7.50 m/km. The absolute differences between paired USR and reference values ranged from 0.01 to 0.24 m/km, with corresponding absolute percentage errors ranging from 0.35% to 4.89%. Signed differences (USR − reference) ranged from −0.17 to +0.24 m/km. Positive values indicate overestimation by the USR, whereas negative values indicate underestimation, providing the directional information summarized by the overall bias.

Table 1

Paired 100-m Segment Observations of USR and Reference IRI

Segment	USR IRI (m/km)	Reference IRI (m/km)	|Error| (m/km)	Absolute % Error
SP2 0-100	5.88	5.80	0.08	1.38%
SP2 100-200	5.75	5.70	0.05	0.88%
SP2 200-300	7.54	7.50	0.04	0.53%
SP3 0-100	6.04	5.80	0.24	4.14%
SP3 100-200	6.84	6.70	0.14	2.09%
SP3 200-300	7.33	7.50	0.17	2.27%
SP4 0-100	5.01	4.80	0.21	4.38%
SP4 100-200	5.75	5.70	0.05	0.88%
SP4 200-300	4.93	4.70	0.23	4.89%
SRG1 0-100	2.87	2.97	0.10	3.37%
SRG1 100-200	2.79	2.88	0.09	3.12%
SRG1 200-300	3.16	3.31	0.15	4.53%
SRG2 0-100	3.04	2.97	0.07	2.36%
SRG2 100-200	2.87	2.88	0.01	0.35%
SRG2 200-300	3.36	3.31	0.05	1.51%

Table 1 shows that the two segments with the largest absolute errors, SP3 0–100 and SP4 200–300, both belong to the SP2–SP4 subgroup, which covers the higher end of the tested IRI range. The SRG1–SRG2 segments, which cover a lower IRI range, exhibit smaller absolute errors but, in some cases, comparatively larger percentage errors because their reference values are lower. Because the two subgroups cover different IRI ranges and use different reference instruments, the contrast between their absolute and percentage errors is descriptive only and should not be interpreted as evidence that either subgroup or reference system is intrinsically more accurate.

Table 2 summarizes overall system-level agreement across all 15 paired observations. The recalculated statistics differ from certain descriptive values reported in the original source calculation document. That document lists a standard deviation of 19.16075, which is mathematically inconsistent with an observed IRI range of only 2.79–7.54 m/km and was therefore excluded from the present analysis as an apparent data-entry error. The source document also reverses the USR and reference means. Direct recalculation from the 15 paired observations in Table 1 yields a mean USR IRI of 4.877 m/km and a mean reference IRI of 4.835

m/km. For internal consistency, all summary statistics reported in Table 2 were recalculated directly from the paired observations in Table 1. Only values reproducible from the retained paired dataset are used in the analysis.

Table 2
Overall System-Level Agreement Statistics (n = 15)

Metric	Result
n	15 paired observations
Mean USR IRI	4.877 m/km
Mean reference IRI	4.835 m/km
Mean bias (USR - reference)	+0.043 m/km
MAE	0.112 m/km
RMSE	0.133 m/km
MAPE	2.45%
Pearson r	0.9973
R^2	0.9946
Regression	USR IRI = 1.017 × Reference IRI − 0.038
95% Bland-Altman limits of agreement	−0.212 to +0.298 m/km

Table 2 shows that the mean bias of +0.043 m/km is small relative to the mean reference IRI of 4.835 m/km, representing less than 1% of the mean reference value. The estimated 95% Bland–Altman limits of agreement range from −0.212 to +0.298 m/km, indicating the range within which approximately 95% of individual paired differences would be expected to fall under the assumptions of the method. However, no external acceptance tolerance was prespecified; therefore, the magnitude of the bias and the limits of agreement are reported descriptively rather than characterized as universally acceptable.

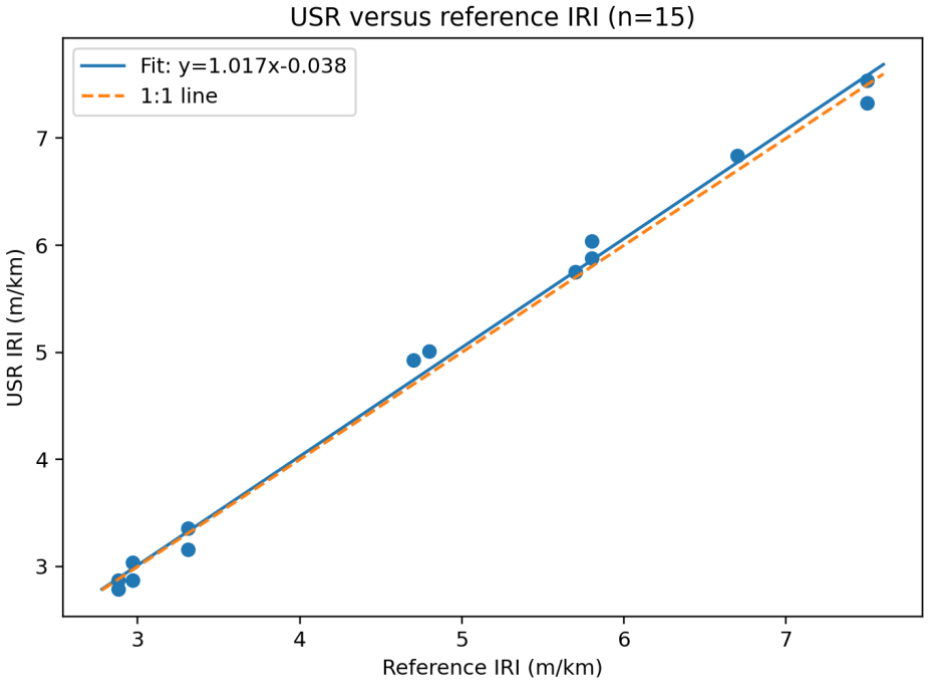


Figure 1. USR versus reference IRI with fitted regression and 1:1 line.

Figure 1 shows that the fitted regression line (solid) lies almost directly on top of the 1:1 reference line (dashed) across the tested range, visually confirming the near-unit slope and small intercept reported in Table 2. The paired observations also lie close to the two lines. The graphical proximity of the fitted and identity lines is supportive but does not, by itself, establish equivalence or interchangeability; these must be evaluated alongside MAE, RMSE, bias, and the limits of agreement.

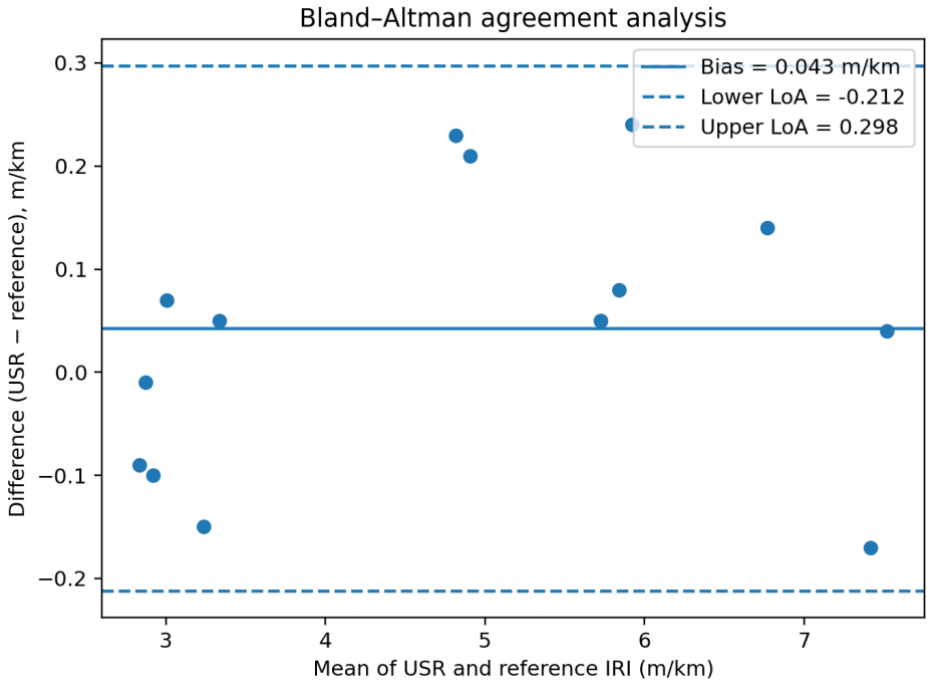


Figure 2. Bland-Altman agreement analysis for the 15 paired observations.

Figure 2 plots the difference between each paired USR and reference measurement against the mean of the two measurements. All 15 points fall within the estimated upper and lower limits of agreement, and the points are distributed around the mean-bias line without an obvious trend toward increasing or decreasing disagreement as the mean IRI rises. This pattern suggests no strong proportional bias across the tested range. However, with only 15 paired observations, the limits are sample estimates subject to substantial uncertainty. The fact that all observed points lie within these limits does not, by itself, establish acceptable agreement for regulatory or contractual applications.

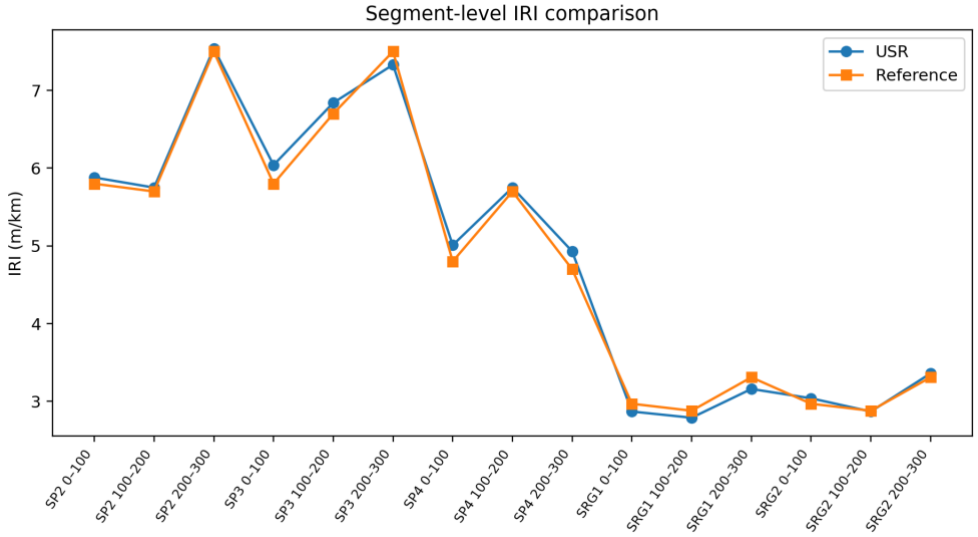


Figure 3. Segment-level comparison of USR and reference IRI.

Figure 3 compares USR and reference IRI values segment by segment. The two lines track each other closely across all 15 segments, with the largest visible separation occurring at SP3 0–100, consistent with the largest absolute error reported in Table 1.

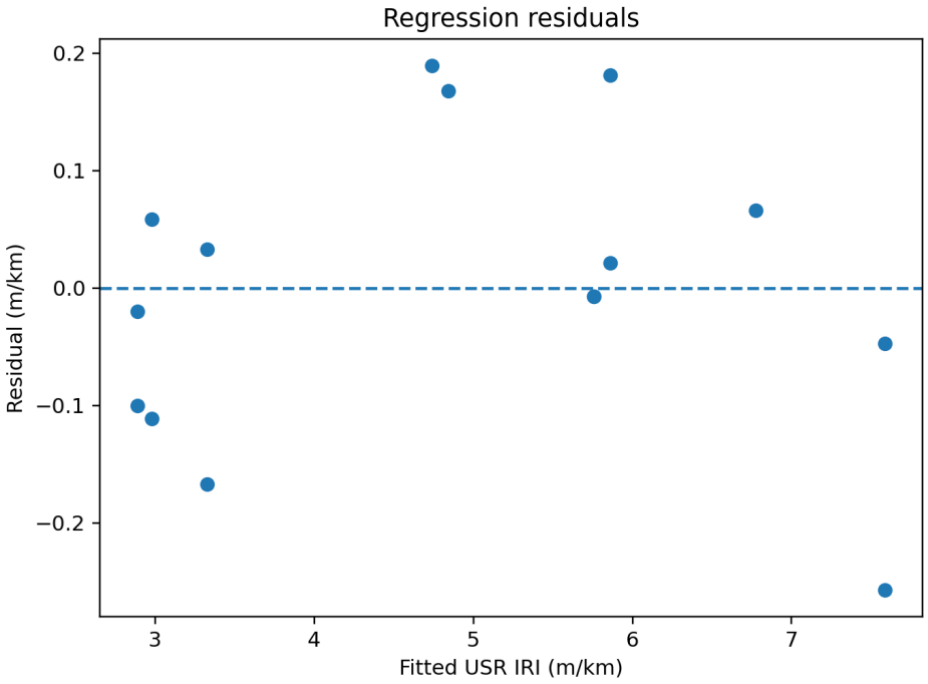


Figure 4. Regression residuals.

Figure 4 shows the residuals from the fitted regression across the observed IRI range. The residuals are small and do not display an obvious funnel-shaped or curved pattern, which could otherwise indicate heteroscedasticity or a nonlinear relationship, respectively. Nevertheless, the small sample size limits the ability to detect such patterns reliably.

Table 3 disaggregates agreement statistics by reference-instrument subgroup. The SP2–SP4 subgroup, benchmarked against the GAPMAN Walking Profilometer, showed a Pearson correlation coefficient of r = 0.9955 and an RMSE of 0.155 m/km across its IRI range of 4.93–7.54 m/km. The SRG1–SRG2 subgroup, benchmarked against the Vehicle-Mounted IRI Measurement System, showed a lower correlation coefficient of r = 0.9123 but a smaller RMSE of 0.090 m/km across its narrower IRI range of 2.79–3.36 m/km. The largest absolute difference across all 15 pairs was 0.24 m/km, recorded for segment SP3 0–100 and corresponding to a 4.14% absolute percentage error. The next-largest absolute differences were 0.23 m/km for SP4 200–300 and 0.21 m/km for SP4 0–100; all remaining absolute differences were 0.17 m/km or lower.

Table 3
Subgroup Agreement Statistics by Reference Instrument

Group	n	Mean USR	Mean ref.	Bias	MAE	RMSE	MAPE	r
SP2-SP4	9	6.119	6.022	0.097	0.134	0.155	2.38%	0.9955
SRG1-SRG2	6	3.015	3.053	-0.038	0.078	0.090	2.54%	0.9123

Table 3 shows that although the SRG1–SRG2 subgroup has a lower correlation coefficient than the SP2–SP4 subgroup, it nonetheless has a smaller RMSE and a comparable MAPE. This illustrates that the lower correlation coefficient in this case may reflect the subgroup's narrower IRI range rather than greater absolute measurement disagreement. This subgroup pattern should not be used for inferential comparisons because both the sampled IRI range and the reference instrument differ between groups.

Discussion

Interpreting Overall Agreement Against Established Standards

The principal finding of this study is that the USR closely tracks reference IRI

measurements across the tested range, with a near-unit regression slope of 1.017, a small intercept of −0.038, a small positive mean bias of +0.043 m/km relative to a mean reference IRI of 4.835 m/km, and Bland–Altman limits of agreement spanning −0.212 to +0.298 m/km. This pattern is consistent with the theoretical premise that IRI, when properly derived from measured longitudinal profiles, can be compared across different profile-measurement instruments. However, the synthesis by the National Academies of Sciences, Engineering, and Medicine (2018) makes clear that practical comparability depends on sustained attention to equipment certification, repeatability, and reproducibility rather than being an automatic property of any device that produces a value labeled "IRI." Accordingly, the present results should be interpreted as validation evidence specific to the tested USR configuration, road sections, and operating conditions, rather than as a general certification of interchangeability with the reference instruments used. Consistent with the historical IRI framework, the transferability of roughness values depends on controlled profile acquisition and calibration rather than on the index label alone (ASTM International, 2022; ISO, 2016; Sayers et al., 1986).

Positioning the USR Within the Low-Cost Sensing Literature

Against the broader movement toward low-cost pavement-roughness sensing, the USR's architecture occupies a distinct position. Response-type approaches that infer roughness from vehicle acceleration—whether processed through a quarter-car model Zhang et al. (2022), a refined smartphone application Bai et al. (2025), or increasingly sophisticated machine-learning pipelines Sang et al. (2024) —share a structural characteristic: their signals originate in the vehicle's dynamic response to the road rather than in direct measurement of the road surface itself. This dependence helps explain why Wu et al. (2024) developed a rapid recalibration procedure for situations in which the vehicle or its dynamic parameters change. The USR's ultrasonic channel, by contrast, measures the distance between the sensor and the road surface directly and noninvasively. It belongs to the same broad conceptual category as Behera et al. (2021) ultrasonic roughometer and, more recently, Alatoom & Smadi (2026) radar-based sensor-fusion system.

Direct geometric sensing of this kind may reduce dependence on vehicle-specific dynamic recalibration, although it introduces other potential sources of error, including sensor misalignment, acoustic reflection characteristics, and surface texture. Múčka (2016) identifies these considerations as relevant to profile-acquisition technologies more broadly. The present results, obtained without a vehicle-specific recalibration step, are consistent with this expected trade-off: the USR's agreement with reference IRI measurements appears comparable in order of magnitude to the performance reported by Alatoom & Smadi (2026) for their respective direct-sensing architectures. However, formal quantitative comparisons across studies are not appropriate because the road networks, operating speeds, and reference instruments differ. Rather than positioning the USR as a competitor to the machine-learning-based response models reviewed above, this study's contribution is better understood as the method-comparison validation of a multisensor physical measurement system. This is a distinct and necessary evidentiary step from the predictive-model validation that characterizes much of the smartphone-based and machine-learning literature.

The Value of Bland-Altman Analysis Over Correlation Alone

A methodological contribution of this study, beyond the specific USR results, lies in its explicit demonstration of why correlation coefficients alone are insufficient for validating a low-cost roughness-measurement instrument. This limitation was emphasized by Bland and Altman and more recently reaffirmed in the context of nonclinical engineering measurements by (Botha et al., 2025). The subgroup results in Table 3 illustrate this point. The SRG1–SRG2 subgroup has a substantially lower Pearson correlation coefficient (r = 0.9123) than the SP2–SP4 subgroup (r = 0.9955), which, if interpreted in isolation, might suggest materially poorer performance on the SRG sections. Yet the SRG1–SRG2 subgroup simultaneously achieves a smaller RMSE (0.090 m/km, compared with 0.155 m/km for SP2–SP4) and a comparable MAPE.

This pattern may be partly explained by the narrower underlying IRI range of the SRG1–SRG2 subgroup (2.79–3.36 m/km, compared with 4.93–7.54 m/km for SP2–SP4), which can

reduce the correlation coefficient without indicating a corresponding increase in absolute disagreement. A reviewer or practitioner relying on correlation alone could therefore reach a different conclusion from one relying on absolute error. This is consistent with Bland and Altman's warning that high correlation does not establish agreement and that a comparatively lower correlation does not, by itself, establish disagreement when the range of sampled values is restricted. This finding reinforces the broader argument presented in the Introduction: method-comparison studies of low-cost roughness sensors should report Bland–Altman limits of agreement and absolute error metrics alongside correlation coefficients, rather than relying on correlation alone.

Data Integrity and Reproducibility

Transparent reporting of the underlying data is an important component of a rigorous validation study. As noted in the Results, the source calculation record for this dataset contained an internally inconsistent standard-deviation entry and a reversed mean-value row. Both issues were identified through direct recalculation from the retained 15-pair dataset and corrected or excluded, as appropriate, rather than propagated into the reported statistics. This type of pre-submission data audit is a necessary companion to any method-comparison claim because the credibility of a validation study rests partly on reviewers' ability to reproduce its principal statistics independently. Retaining the complete paired-observation table, rather than reporting only aggregate coefficients, serves the same purpose: it enables readers to recompute MAE, RMSE, MAPE, the regression coefficients, and the Bland–Altman limits directly from Table 1 without access to the original data-acquisition software. This approach is preferable for a measurement-validation article of this kind and is consistent with Múčka (2016) broader caution that the credibility of any IRI value depends on transparent and auditable profile processing.

Practical Acceptance Versus Universal Interchangeability

A further interpretive distinction concerns practical suitability for a specific screening purpose versus general interchangeability with certified profiling equipment. An MAE of 0.112 m/km and a MAPE of 2.45% may be acceptable for network-level screening and maintenance prioritization, where the practical question is often whether a road section falls into a broad condition category. However, the same results might be insufficient for contractual acceptance testing or regulatory certification, where a predefined tolerance and a documented reference-method uncertainty budget are normally required before a pass/fail threshold can be justified. This study deliberately avoids declaring a universal acceptance threshold and instead frames its findings as sufficient to justify broader validation, rather than as sufficient on their own to certify the USR for a specific regulatory or contractual application.

This distinction also affects how the regression equation reported in Table 2 should be used. Because the equation was fitted and evaluated using the same 15 observations, it is reported as a descriptive diagnostic of scaling and offset rather than as an independently validated calibration equation. It should not be used to correct future USR readings without first being estimated using an independent calibration dataset and evaluated against a separate validation dataset, consistent with standard practice for avoiding optimistic performance estimates in method-comparison research. Accordingly, “screening suitability” in this article means that the observed errors and bias are sufficiently low to motivate broader field validation; it does not mean that a predefined regulatory tolerance, condition-classification accuracy threshold, or interchangeability criterion has already been met.

International and Cross-Context Relevance

Although the present dataset was collected on Indonesian road sections, the underlying measurement problem and the rationale for low-cost roughness-screening technology are international in scope rather than locally confined. Road authorities in many middle- and lower-income countries face the same structural mismatch identified in the Introduction between the need for network-wide roughness data and the capital cost of certified inertial profiling equipment. The broader research landscape to which this study contributes—including smartphone-based sensing in China Zhang et al. (2022), machine-learning-based estimation in the

Middle East and elsewhere Kaloop et al. (2023), data-driven smartphone frameworks Sang et al. (2024), and radar–IMU sensor fusion Alatoom & Smadi (2026) —reflects a global movement toward distributed, lower-cost pavement-condition monitoring rather than a phenomenon specific to any one national context.

The USR's contribution to this international literature is best framed as a physical-sensing complement to the predominantly response-type and machine-learning-based approaches that currently feature prominently in the field. Because its ultrasonic channel measures the distance between the road surface and the sensor directly, it offers a measurement principle that is less dependent on vehicle-specific dynamics. The present validation, which combines correlation, regression, absolute-error, and Bland–Altman diagnostics within a single method-comparison design, also offers an approach that other low-cost sensing studies could adopt to strengthen the evidentiary basis of their validation claims. In this sense, the study's potential contribution lies not only in validating a specific prototype but also in demonstrating a comprehensive validation approach for a multisensor ultrasonic architecture.

Limitations

Several limitations qualify these conclusions and should be stated explicitly. The validation sample is small: 15 segment-level pairs drawn from only five 300-m sections. Measurements were collected using a single vehicle at an operating speed of approximately 30–40 km/h, and the study did not employ a repeated-run design that would allow repeatability and reproducibility to be quantified using coefficients of variation or intraclass correlation coefficients, as recommended by established profiler-verification literature (National Academies of Sciences, Engineering, and Medicine, 2018). The study does not evaluate performance across multiple pavement types, weather conditions, or longitudinal grades. Moreover, the sample is too small to establish reliably whether error variance changes systematically at the extremes of the IRI range, although the contrasting error patterns between the two subgroups in Table 3 suggest that this question warrants investigation using a larger dataset.

The instrument-to-section mapping used in this analysis also relies on the most recent available calculation record, which conflicted with an earlier project description. This discrepancy should be resolved against the original field log before the findings are used to inform operational decisions. It is identified here as a matter of academic transparency rather than being silently resolved in favor of the present interpretation. The subgroup allocation reported here follows the final retained paired dataset. Because the manuscript archive does not contain an independently verifiable field log with geographic and pavement-characteristic metadata, the subgroup findings are treated as descriptive. The absence of these records limits site-level reproducibility and generalizability.

Directions for Future Validation

Building on the findings and limitations discussed above, several concrete extensions would strengthen the evidentiary basis for the USR's practical deployment. Table 4 summarizes the principal limitations of the current evidence and the corresponding experimental designs that could address them. Most importantly, a future validation program should prespecify its primary endpoint before data collection begins. If the objective is to establish interchangeability with certified profilers, the design should emphasize bias and limits of agreement estimated from a larger and more diverse sample. If the objective is screening or condition ranking, classification agreement and ranking performance become more relevant endpoints. If a calibration correction is proposed, it should be estimated using one dataset and independently evaluated using a separate validation dataset, rather than fitted and tested on the same observations, as was done for the descriptive regression in the present study. A repeated-run design under controlled, multiple-speed conditions and across independent sites, different vehicles, and wet and dry surface conditions would allow the reproducibility claims that this study cannot yet support to be tested directly. Finally, because the present validation dataset does not assess automated data-quality control, a natural extension would be to incorporate automated flagging of segments with abnormal GNSS position jumps, ultrasonic sensor saturation, or implausible profile discontinuities before IRI calculation. Building on the ultrasonic–inertial–GNSS measurement

chain established here, this would provide a quality-assurance layer broadly analogous to the certification safeguards described for certified profilers by the National Academies of Sciences, Engineering, and Medicine (2018), adapted to the resource constraints of a low-cost multisensor device.

Table 4
Recommended Next Validation Design

Factor	Current evidence	Recommended next experiment
Road sections	5 sections	At least several pavement types and condition ranges
Paired observations	15 x 100 m	Larger repeated dataset across routes
Speed	Dynamic operation at ~30-40 km/h	Controlled multi-speed runs
Runs	Limited repeatability evidence	Repeated same-day and multi-day runs
Reference	Gap Man + vehicle-mounted IRI system	Independent reference with documented uncertainty
Environment	Limited controlled conditions	Dry/wet, temperature and surface-condition comparison

Table 4 aligns the limitations of the present study with the specific experimental conditions that a follow-up validation should incorporate. The most consequential gaps are the small number of road sections, the limited operating-speed range, and the absence of a repeated-run design.

CONCLUSION

This study evaluated the system-level measurement agreement of a low-cost multisensor Ultrasonic Surface Roughmeter (USR) against reference IRI instrumentation using 15 paired observations from 100-m road segments across five 300-m road sections. The analysis combined correlation, regression, absolute and relative error metrics, and Bland–Altman limits of agreement rather than relying on correlation alone. Within the tested sections and operating conditions, the USR produced IRI values that were numerically close to the reference measurements, with a Pearson correlation coefficient of r = 0.9973, an R^2 of 0.9946, an MAE of 0.112 m/km, an RMSE of 0.133 m/km, a MAPE of 2.45%, and a mean bias of +0.043 m/km. The fitted regression had a near-unit slope, and the estimated Bland–Altman limits of agreement ranged from −0.212 to +0.298 m/km, indicating relatively small paired differences across the tested range. The high Pearson correlation describes association and is not, by itself, evidence of instrument equivalence. The assessment of agreement therefore rests on the combined error, bias, and Bland–Altman results.

The academic contribution of this study is twofold. Substantively, it provides validation evidence for a direct-sensing, ultrasonic-based multisensor architecture that remains comparatively underrepresented in a low-cost roughness-sensing literature dominated by response-type and machine-learning approaches. Methodologically, it illustrates, through its subgroup results, why Bland–Altman-based agreement analysis should complement rather than be replaced by correlation coefficients when validating low-cost pavement-monitoring instruments. Practically, the findings support further investigation of the USR as a promising low-cost screening and monitoring technology for road authorities that cannot presently justify the capital cost of a certified inertial profiler. However, the small validation sample does not support a claim of universal interchangeability with certified profiling equipment. Future validation should prioritize additional sites and pavement types, repeated same-day and multiday measurements, controlled variation in vehicle and speed, independent calibration and validation datasets, and automated data-quality checks before operational interchangeability is considered.

Future research should extend this validation along the specific dimensions identified in the Discussion: larger and more diverse road samples, repeated runs at multiple speeds and with

multiple vehicles, an independently estimated and separately tested calibration model, and an automated data-quality-control layer integrated into the ultrasonic–inertial–GNSS measurement chain. Together, these extensions would help advance the USR from a validated prototype toward a field-ready, low-cost complement to certified profiling technology, with potential relevance beyond the Indonesian road network from which the present dataset was collected.

ACKNOWLEDGEMENT

This research received no specific grant from any public, commercial, or not-for-profit funding agency. The author declares no conflict of interest. The manuscript was prepared from the author's 2024 doctoral dissertation dataset.

AUTHOR CONTRIBUTION STATEMENT

The article is the work of a single author, who was responsible for its conceptualization, data analysis, and writing.

REFERENCES

Alatoom, Y. I., & Smadi, O. (2026). Low-cost pavement roughness evaluation using Inertial Measurement Unit (IMU)-Pulsed Coherent Radar (PCR) sensor fusion. *Automation in Construction*, *182*. https://doi.org/10.1016/j.autcon.2025.106713

ASTM International. (2022). ASTM E950/E950M-22: Standard test method for measuring the longitudinal profile of traveled surfaces. ASTM International.

Bai, S., Guoqi, Y., Rong, A., Guo, G., & Wang, J. (2025). Research on Pavement Roughness Evaluation Method Based on Smartphones. *Journal of China and Foreign Highway*, *45*(6). https://doi.org/10.14048/j.issn.1671-2579.2025.06.007

Behera, H. K., Pradhan, S., & Das, S. S. (2021). Low cost ultrasonic roughometer for pavement roughness measurement. *Innovative Infrastructure Solutions*, *6*(3), 168. https://doi.org/10.1007/s41062-021-00521-0

Botha, Mrs. C. M., Liebenberg, S. C., Conradie, F. H., & van der Merwe, Mr. A. F. (2025). The potential of the Bland-Altman method in chemical engineering. *South African Journal of Chemical Engineering*, *54*, 348–356. https://doi.org/10.1016/j.sajce.2025.08.012

Chen, K., Torbaghan, M. E., Thom, N., & Faramarzi, A. (2025). Physics-guided neural network for predicting international roughness index on flexible pavements considering accuracy, uncertainty and stability. *Engineering Applications of Artificial Intelligence*, *142*, 109922. https://doi.org/10.1016/j.engappai.2024.109922

Dutta, M., Trivedi, A., Soriya, R., Verma, S., & Rajput, P. (2026). Road user-centric pavement management system: ride quality assessment using whole-body vibration exposure. *International Journal of Pavement Engineering*, *27*(1). https://doi.org/10.1080/10298436.2026.2632114

Federal Highway Administration. (2022). *Inertial profiler – Pavement.* FHWA InfoTechnology.

International Organization for Standardization. (2016). ISO 8608:2016 Mechanical vibration—Road surface profiles—Reporting of measured data. ISO.

Jalili, F., Ghavami, S. M., & Afsharnia, H. (2024). An Artificial Neural Network approach to assess road roughness using smartphone-based crowdsourcing data. *Engineering Applications of Artificial Intelligence*, *138*, 109308. https://doi.org/10.1016/j.engappai.2024.109308

Kaloop, M. R., El-Badawy, S. M., Hu, J. W., & Abd El-Hakim, R. T. (2023). International Roughness Index prediction for flexible pavements using novel machine learning techniques. *Engineering Applications of Artificial Intelligence*, *122*, 106007. https://doi.org/10.1016/j.engappai.2023.106007

Loprencipe, G., & Zoccali, P. (2017). Ride quality due to road surface irregularities: Comparison of different methods applied on a set of real road profiles. *Coatings*, *7*(5). https://doi.org/10.3390/coatings7050059

Martin Bland, J., & Altman, DouglasG. (1986). STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. *The Lancet*, *327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Múčka, P. (2016). Current approaches to quantify the longitudinal road roughness. *International*

Journal of Pavement Engineering, *17*(8). https://doi.org/10.1080/10298436.2015.1011782

National Academies of Sciences, Engineering, and Medicine. (2018). *Inertial profiler certification for evaluation of International Roughness Index* (NCHRP Synthesis 526). The National Academies Press. https://doi.org/10.17226/25207

Nguyen, T., Lechner, B., & Wong, Y. D. (2019). Response-based methods to measure road surface irregularity: a state-of-the-art review. In *European Transport Research Review* (Vol. 11, Number 1). https://doi.org/10.1186/s12544-019-0380-6

Sang, Y., Yu, Q., Fang, Y., Vo, V., & Wix, R. (2024). Smartphone-Based IRI Estimation for Pavement Roughness Monitoring: A Data-Driven Approach. *IEEE Internet of Things Journal*, *11*(11), 19708–19720. https://doi.org/10.1109/JIOT.2024.3369109

Sayers, M., Gillespie, T., & Queiroz, C. (1986). The international road roughness experiment: establishing correlation and a calibration standard for measurements. In *The World Bank* (Number 45).

Wu, D., Liu, C., Qin, B., Zhong, S., Zhang, X., & Du, Y. (2024). Fast calibration for vibration-based pavement roughness measurement based on model updating of vehicle dynamics. *International Journal of Pavement Engineering*, *25*(1). https://doi.org/10.1080/10298436.2023.2287688

Zhang, Z., Zhang, H., Xu, S., & Lv, W. (2022). Pavement roughness evaluation method based on the theoretical relationship between acceleration measured by smartphone and IRI. *International Journal of Pavement Engineering*, *23*(9), 3082–3098. https://doi.org/10.1080/10298436.2021.1881783

Žuraulis, V., Sivilevičius, H., Šabanovič, E., Ivanov, V., & Skrickij, V. (2021). Variability of gravel pavement roughness: An analysis of the impact on vehicle dynamic response and driving comfort. *Applied Sciences (Switzerland)*, *11*(16). https://doi.org/10.3390/app11167582