Exploratory Data Analysis for Investigating the Effect of Source Receiver Distance on Noise Modelling

DOMENICO ROSSI[1], DALJEET SINGH[2], CLAUDIO GUARNACCIA[1]
[1]Department of Civil Engineering,
University of Salerno,
Via Giovanni Paolo II 132, Fisciano (SA),
ITALY

[2]Mechanical Engineering Department,
Thapar Institute of Engineering and Technology,
Patiala 147004,
INDIA

Abstract: - Nowadays, Machine Learning (ML) techniques are ubiquitous. They are applied to many contexts, both in production and research fields. Recently, they have shown strong performance in predicting road traffic noise levels, even when used at different sources and receiver distances. Due to their nature, they use the distance between source and receiver as any other feature, instead of considering it as one of the most important ones. This raises an important question: does distance really have a minor role, or does it influence the relationships between input variables and noise outputs in a non-easily perceivable way? In this contribution, the authors pursue an exploratory data analysis (EDA) approach to examine if and how source-receiver distance affects an ML model for road traffic noise estimation. This study highlights how the relationships between input variables and output sound levels change at different distances from the source. Statistical distributions, correlation patterns, and feature relevance indicators are investigated for two sensors recording the same noise events at different distances from the source. The results show that the relationship between variables and noise levels is different between the two cases. These findings emphasize the need to supplement performance-based evaluations with exploratory and interpretative analyses, providing new insights into the role of spatial factors in data-driven noise modeling.

Key-Words: - Road traffic noise, Machine Learning, Exploratory Data Analysis, Noise modelling, Source–receiver distance, Feature importance, Data-driven modelling, Environmental acoustics.

Received: February 8, 2026. Revised: April 13, 2026. Accepted: August 3, 2026. Published: September 17, 2026.

1 Introduction

In recent years, Machine Learning (ML) techniques have been more and more frequently used in the research context of prediction and modelling of road traffic noise, demonstrating promising capabilities, [1], [2], [3], [4], [5], [6], [7]. Such model often achieves predictions that are comparable (or even better) than those obtained by traditional physically based models. Surely, they demonstrated their strength point in non-linear scenarios, i.e., where the free flow regime is not assured and/or in case of intersections or roundabouts, [8], [9], [10], [11], [12]. An interesting and not yet fully investigated aspect of these approaches is the role of the source–receiver distance. In classical propagation models, in fact, distance is a pivotal parameter, and since it rules the propagation formulas, it typically has the dignity of a primary independent variable. On the contrary, in ML approaches (or in general in data-driven ones), source-receiver distance is considered as one of many features, and its influence is not explicitly enforced by the modelling framework. This aspect raises important questions regarding the interpretation of such predictions. The interpretation of results, in fact, is not straightforward, and it is not clear if distance plays a minor (or no) role in the modelling process, [13], [14], [15], [16], [17].To fill this gap, the authors used exploratory analysis techniques on a well-known environmental noise dataset to highlight the intrinsic complexity of the relationships between acoustic outputs and input variables and to determine whether the relationship between input variables and noise levels remains the same across space.

Therefore, after applying a Random Forest (RF) regressor on the data, rather than focusing on model performance (which has already been demonstrated in other works) [4], [5], [6], the authors investigated how the structure of the dataset and the relationships between variables evolve with the source–receiver distance. In this preliminary approach, the analysis is carried out by only comparing two representative scenarios: a near-field receiver located near the road (6.9 m), and a far-field receiver positioned at a significantly greater distance (270 m circa). The idea is that, while predictive performance may remain stable (as some studies not yet published demonstrate), the contribution of input variables to the generated noise levels may vary. This approach contributes to improving the interpretability of data-driven noise models and provides additional information on this research field.

The remaining of the contribution is organized as follows. In Section 2, the case study and the used data are presented, contextualizing the distance as the main key point of the analysis. Section 3 details the methodological approach used. Section 4 reports the results of the analysis, organized into four main steps: comparison of noise level distributions at near and far distances, evaluation of correlations between input variables and acoustic outputs, assessment of feature relevance through interpretable metrics, and synthesis of the observed differences in data structure as a function of distance. Finally, in Section 5, results are critically discussed and a last section is reported in Section 6, where the main conclusions of the study are summarized together with the limitations of the study itself and a brief description of possible future implementation.

2 Case Study and Data Used

The data used in this study comes from a publicly available dataset collected within the framework of a long-term monitoring research campaign conducted in proximity to a highway infrastructure in the city of Saint Berthevin, France. The monitoring campaign was designed for long-term environmental acoustics investigations and consisted of multiple acoustic and traffic sensors installed in the surrounding area, allowing for the simultaneous acquisition of noise levels and traffic parameters over an extended period. Acoustic measurements were performed using a set of sound level meters positioned at different distances from the main noise source, represented by a 4-lane highway connecting two important cities. The sensors were organized in measurement towers positioned at variable distances from the road, collecting A-weighted equivalent noise levels over 10-second intervals, which were subsequently cleaned and aggregated into 15-minute observations by the dataset authors.

In parallel, a series of meteorological features (air and ground temperature, wind speed and direction, solar insulation and precipitation quantity) collected dedicated data. Such meteorological sensors were mounted on different towers. Such data have not been included in this work. Figure 1 represents a realistic scaled positioning of all the acoustic and meteorological towers. Traffic information was obtained from sensors positioned along the highway infrastructure. A more detailed description of the sensors' spatial organization can be found in [18]. Data, then, are available at a 15-minute base, and for each 15-minute step, four main traffic data are used: number of light vehicles passing, number of heavy vehicles passing, average speed of light vehicles (km/h), and average speed of heavy vehicles (km/h). These variables constitute the input features considered in the present analysis, while the acoustic ones represent the outputs. Specifically, A-weighted equivalent continuous sound level (L_{Aeq}) over 15 minutes is used as input. The dataset also has output features isolating single frequencies, but for the purpose of this work, they were neglected. For the sake of simplicity and clarity, the authors performed such exploratory and preliminary analysis only for two representative receiver positions: a near-field sensor located near the source, and a far-field sensor positioned at a significantly greater distance. This selection allows for a direct comparison between the most contrasting spatial conditions. The choice to investigate only two receivers was a conscious methodological decision. The objective of this exploratory study, in fact, was to investigate if differences in source-receiver distance can reflect substantial changes in data structure, correlation patterns, and feature relevance. A scaled representation of the spatial configuration of the monitoring system and the relative position of the acoustic sensors is illustrated in Figure 1 (derived from the information reported in [19].

In Figure 1 all the sensors used in the original long-term monitoring campaign are shown: acoustic, meteorological, and traffic ones. In this contribution, the authors narrowed the used data to four input parameters and two output ones, neglecting the remaining acoustic information and all the meteorological ones (see Section 3 for more details).

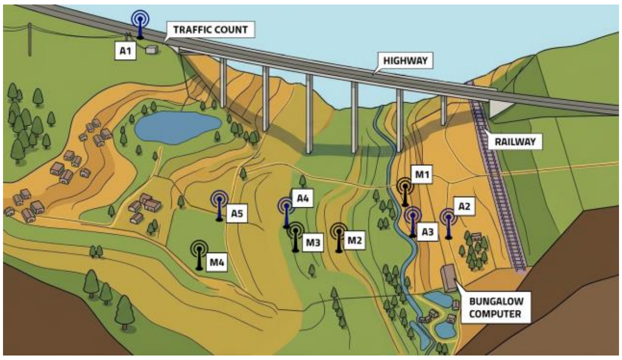

Fig. 1: Spatial configuration of the monitoring system, illustrating the relative positions of acoustic sensors at varying distances from the traffic source.
Source: created by the authors from the geospatial information in [19]

3 Materials and Methods

The present study adopts an exploratory data analysis (EDA) approach to investigate the influence of source–receiver distance on the structure of road traffic noise data. Unlike what could be expected and in contrast to conventional studies focused on model calibration and validation of ML prediction models, the main focus of this work is to analyze how the relationships between input variables and acoustic outputs evolve under different spatial conditions rather than optimize the prediction performance of the model itself. The whole methodology is based on a comparative analysis between two representative configurations: a near-field receiver, located at a short distance from the road traffic source (acoustic sensor 1, named " A1M5_LQA" at 6.9 meters from the road), and a far-field receiver ("A2M5_LQA") positioned at a significantly greater distance of 270 m. De facto, two sub-datasets were built: a "near" one having input data and output of A1M5_LQA, and a "far" one made of input features and A2M5_LQA output. To simplify the description, authors from now on will refer to such datasets simply as "near" and "far". The choice of this dual structure is a simplification intended to allow an easier and clearer interpretation of the effect of distance, avoiding potential ambiguities related to intermediate configurations. The input variables were the same for both data configurations, ensuring consistency in the comparison. In detail, the number of light and heavy vehicles and the average speed of light and heavy vehicles for every 15 minutes are present. More clearly, the input parameters are:
(i) TOT_VL_NB (total number of light vehicles);
(ii) TOT_PL_NB (total number of heavy vehicles);
(iii) VL_VIT (average speed of light vehicles); and
(iv) PL_VIT (average speed of heavy vehicles).

For readability, these variables are reported in figures together with their descriptive names.

The output variable is the A-weighted equivalent continuous sound level (L_{Aeq}), measured at the selected receiver locations. The analysis is carried out independently on the two datasets corresponding to the near and far receivers, and the results are then compared to identify differences in data structure and variable relationships.

3.1 Data Preparation

Before starting the exploratory analysis, the two datasets are preprocessed to ensure consistency and comparability, according to a procedure already described in detail in [18]. The most important aspect of this preprocessing is the removal of rows with missing values, since incomplete records in either input or output variables may bias statistical indicators and correlation measures. No additional feature engineering or dimensionality reduction techniques are applied, as the aim of the study is to preserve the original structure of the data and to observe naturally occurring relationships. Similarly, no normalization is strictly required for the statistical analyses performed, except where needed for specific methods such as feature relevance estimation. In previous works with the same data [4], the authors performed a normalization procedure before starting the ML calibration. Especially for some algorithms, in fact, normalization is essential to provide reasonable results. In this application, authors applied algorithms to the non-normalized data since the focus was not the prediction output results but, as already mentioned, the correlation between features.

3.2 Distribution Analysis

The first step of the exploratory analysis focuses on the statistical characterization of the output variable. The distribution of L_{Aeq} values is analyzed separately for the near and far datasets. Descriptive statistics such as mean, median, standard deviation, and interquartile range are computed, and graphical representations (violin plot) have been used to assess differences in central tendency, variability, and the presence of outliers. This exploratory analysis has already been extensively carried out in previous works [4]; therefore, in this work it is limited to the essential graphs useful for the prosecution of the correlation analysis. For the purposes of this work, anyway, in addition to standard descriptive statistics, higher-order moments of the distributions, namely skewness and kurtosis, were evaluated to detect possible asymmetry and tail behavior. Moreover, statistical tests were performed to verify whether the differences between near-field and far-field data were statistically significant. This step provides an initial understanding of how noise levels are distributed at different distances and whether signal variability changes across space.

3.3 Correlation Analysis

In the second step, the authors investigated the relationship between input variables and the acoustic output [19], [20], [21] Linear correlations between each traffic-related parameter and L_{Aeq} are computed independently for the two receiver configurations.

The comparison of correlation coefficients allows for the identification of potential shifts in the strength and direction of relationships as a function of distance, providing insight into whether the same variables maintain a consistent explanatory role through distance or if their influence changes when moving from near to far sensitive receivers. Correlation analysis was performed using the sample Pearson correlation coefficient. Given two samples X and Y containing n observations, the coefficient is computed according to equation (1)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2}\sqrt{(y_i - \bar{y})^2}} \qquad (1)$$

where x_i and y_i are the observations of the two variables, $\bar{x}$ and $\bar{y}$ are their means, and n is the number of observations, [20]. Pearson correlation was preferred over Spearman correlation because the aim was to quantify the linear relationships between traffic-related variables and acoustic outputs and also to allow a direct comparison with the literature where it is widely used. The resulting coefficient is dimensionless and ranges between −1 and +1: values close to +1 indicate a strong positive linear relationship between two variables, while values close to −1 suggest a strong negative relationship. If the coefficient is close to zero, then it is reasonable to believe that there's a weak or no linear dependence between the variables. Besides, for a more detailed assessment of the relevance of the differences, the correlation coefficients were also compared using Fisher's z-transformation.

3.4 Feature Relevance Analysis

Additionally, in addition to the correlation-based analysis, a study on feature relevance was performed, [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]. The approach was simple: an RF regression algorithm was applied to both the near and far datasets. Please note that the sole purpose of the regressor was to estimate the relative importance of each input variable; therefore, it is important to remark that the model is not optimized for predictive performance (hyperparameter tuning is intentionally kept minimal [4]). The goal, in fact, was not to achieve the best possible prediction accuracy, but rather to obtain an indication of how the model considers the different input variables and weights them. In detail, the Random Forest regressor was implemented using 50 decision trees and standard algorithm settings. Preliminary tests and also previous studies [6],[7] have shown that such a configuration provides stable and sufficiently accurate results (while maintaining reasonable computational cost). Feature importance values obtained from the model are then compared between the near and far datasets. In the context of RF models, feature importance is a relative measure of each input feature's contribution to the overall prediction. Such measurement is obtained by evaluating the reduction in node impurity produced by each feature across all trees. For better interpretability, the sum of all feature importances is reported to 1 through a normalization process. As an example, a feature importance value of 0.6 means that the corresponding variable contributes to the prediction more than another variable having an importance value equal to 0.1. The analysis is therefore useful for ranking variables according to their relative influence within the model.

4 Results

The exploratory analysis was carried out by comparing the datasets associated with the near-field and far-field receivers. The results are presented in terms of distributional properties, correlation structures, and feature relevance, with the aim of identifying potential variations in the role of input variables as a function of distance.

4.1 Distribution of Noise Levels

The distribution of equivalent noise levels (L_{Aeq}) shows a clear difference between the two receiver configurations. As already stated before (see section 3.2), the authors only reported the most essential graph to start the description of the correlation analysis, since the description of the dataset was already carried out in detail in other works, [5],[11]. To facilitate the interpretation of the statistical properties of the data, Figure 2 combines boxplot and violin plot representations, providing information not only on central tendency and variability but also on the overall shape of the distributions. As expected, results show that the near dataset exhibits overall higher values, due to its proximity to the noise source. Mean values of the two distributions are in fact significantly different (71.6 vs 49.9 dBA), as are median values (72.3 and 50.1 dBA). Since the median and mean values are very similar, a certain level of symmetry in the distributions is suggested. The standard deviations of the distributions are very similar in absolute value (2.9 vs 3.0 dBA), but the interquartile range is significantly different (2.6 vs 3.8 dBA – near and far). Therefore, the near distribution appears slightly more concentrated, with a narrower spread around the central tendency, while the far dataset is characterized by lower noise levels and marginally higher variability (although the difference in dispersion is not particularly pronounced). A noticeable aspect is the presence of outliers: the near-field distribution shows a higher number of low-value outliers, while the far-field case presents fewer extreme values, primarily located on the lower side of the distribution. To make the investigation more complete and detailed, additional distributional indicators were computed on the two datasets (near and far). Outputs show that the near-field configuration has a markedly higher negative skewness (-2.041) compared to the far-field dataset (-0.552), suggesting a more pronounced asymmetry towards lower noise levels.

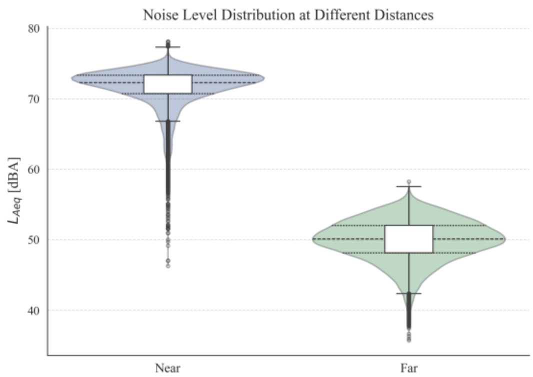

Fig. 2: Distribution of equivalent noise levels (L_{Aeq}) for near-field and far-field receiver positions represented through combined boxplot and violin plot visualizations. Dashed lines represent the median values of the distributions; dotted lines represent the quartiles.
Source: created by the authors.

Similarly, kurtosis values are significantly different between the two distributions (6.113 and 0.350 – near and far, respectively), indicating a more peaked distribution with heavier tails in the near-field case.

Finally, to quantitatively assess whether the two datasets could be considered statistically different, a Kolmogorov–Smirnov (KS) test was performed. The resulting KS statistic was equal to 0.998, indicating an almost complete separation between the two distributions. The associated p-value was lower than 0.001, confirming that the observed differences between the near-field and far-field distributions are statistically significant and cannot be attributed to random sampling variability.

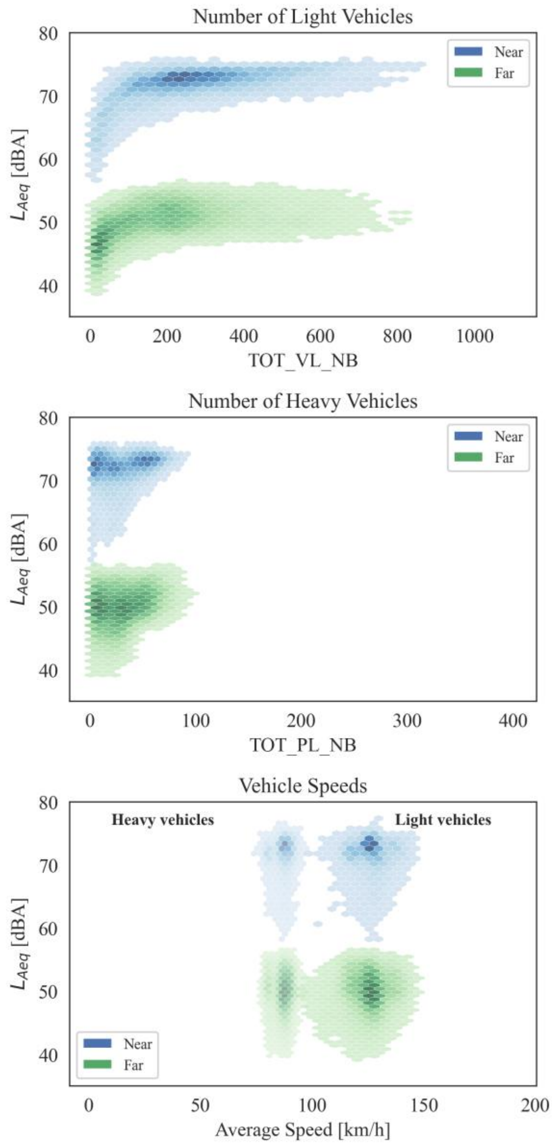

Fig. 3: Joint distributions between traffic variables and equivalent noise levels (L_{Aeq}) for near-field and far-field configurations obtained through two-dimensional hexagonal density maps. The figure highlights how the structure of the traffic–noise relationship changes with source–receiver distance, particularly for the total number of light vehicles (TOT_VL_NB).
Source: created by the authors.

The statistical data relating to the distributions of equivalent noise levels for the near and far configurations is presented in Table 1, while in Table 2 the results of KS test are reported. A further insight into the dataset structure can be obtained by analyzing the joint distributions of traffic variables and acoustic levels. Figure 3 reports two-dimensional density maps obtained through hexagonal binning for the four traffic-related features considered in this study.

The graphical representation highlights the regions with the highest observation density and allows a direct comparison between the near-field and far-field configurations. The most evident differences emerge for the total number of light vehicles (TOT_VL_NB), which already showed the strongest correlation with L_{Aeq} and the highest feature importance values. In this case, the near-field and far-field datasets exhibit markedly different density patterns, suggesting that the relationship between traffic flow and noise levels changes considerably with distance. Similar although less pronounced effects can also be observed for the remaining traffic variables. These observations further support the hypothesis that source–receiver distance does not simply induce a shift in noise levels but also affects the underlying structure of the data and the relationships between explanatory variables and acoustic outputs.

Table 1. Statistical characterization of the acoustic datasets for near-field and far-field configurations, including higher-order statistical indicators

	A1M5LQA	A2M5LQA
count	30347	30347
mean [dBA]	71.6	49.9
std [dBA]	2.9	3.0
min [dBA]	46.3	35.8
25% [dBA]	70.8	48.2
50% [dBA]	72.3	50.1
75% [dBA]	73.4	52.0
max [dBA]	78.2	58.3
Skewness	-2.0	-0.6
Kurtosis	6.1	0.4

Source: created by the authors.

Table 2. Outputs of the Kolmogorov–Smirnov test for the assessment of statistical difference between near and far field dataset

KS Statistic	0.998
p-value	< 0.001

Source: created by the authors.

These observations suggest that, while the general attenuation of noise with distance follows expected physical behavior, subtle differences in the variability and distribution of the data can be identified between the two datasets.

4.2 Correlation Analysis

Results of correlation analysis show a regular reduction in the strength of the relationship between traffic variables and noise levels when moving from the near to the far receiver. The comparison of correlation coefficients between traffic variables and noise levels is shown in Figure 4. The total number of light vehicles (TOT_VL_NB) has the strongest correlation with L_{Aeq} in both the near and far datasets, with a value of 0.53 in the near dataset and 0.42 in the far dataset. This confirms its role as the primary driver of traffic noise: the more vehicles are present, the greater the impact of noise. Speed-related variables, on the other hand, have a minor impact in terms of correlations. In particular, the average speed of light vehicles (VL_VIT) has a correlation coefficient of 0.07 and 0.04 near and far. A similar trend is observed for the average speed of heavy vehicles (PL_VIT), whose correlation with L_{Aeq} becomes almost negligible in the far field: 0.05 and -0.00 (near and far). The only variable showing a slightly different behavior is the number of heavy vehicles (TOT_PL_NB), for which the correlation appears marginally higher in the far-field dataset (0.26 and 0.27, near and far). Table 3 reports the values of correlation for all the input features.

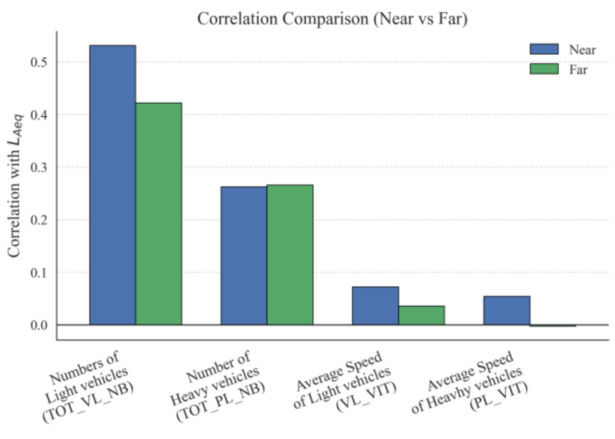

Fig. 4: Correlation coefficients between traffic variables and equivalent noise levels (L_{Aeq}) for the near-field and far-field configurations.

Source: created by the authors.

Table 3. Values of correlation coefficients between input traffic variables and output equivalent noise levels (L_{Aeq})

	TOT_VL_NB	TOT_PL_NB	VL_VIT	PL_VIT
near	0.53	0.26	0.07	0.05
far	0.42	0.27	0.04	-0.00

Source: created by the authors.

However, this difference is minimal and does not represent a substantial deviation from the general trend. Overall, the results indicate that the direct linear relationship between traffic parameters and noise levels weakens with increasing distance, particularly for variables related to vehicle speed.
To evaluate whether the observed differences between near-field and far-field correlation coefficients were statistically significant, Fisher's z-transformation test was performed for each traffic variable. The results indicate that the reduction of correlation observed for TOT_VL_NB, VL_VIT and PL_VIT is statistically significant ($p < 0.001$). Conversely, the difference associated with TOT_PL_NB was not significant ($p = 0.605$), suggesting that the relationship between heavy vehicle counts and noise levels remains essentially unchanged across the two investigated receiver positions. These results – reported in Table 4 - provide additional statistical support to the exploratory observations, confirming that source–receiver distance affects some traffic–noise relationships more strongly than others.

Table 4. Fisher's z transformation test for the statistical significance of correlation coefficients

Feature	Fisher z	p-value
TOT_VL_NB	19.49	<0.001
TOT_PL_NB	-0.52	0.605
VL_VIT	5.06	<0.001
PL_VIT	7.85	<0.001

Source: created by the authors.

4.3 Feature Relevance Analysis

The feature relevance analysis, performed using an RF regressor, provided insight into the role of input variables and how their impact changes through space. Results are reported in Table 5 and Figure 5. In the near dataset, the total number of light vehicles (TOT_VL_NB) has the highest feature ranking (0.66), confirming its primary role in determining noise levels close to the source. This means that the chosen model strongly depends on this feature for its prediction. The other features, on the contrary, have a marginal contribution (0.12, 0.12, and 0.11 for PL_VIT, VL_VIT, and TOT_PL_NB, respectively). The far dataset, on the contrary, has a more balanced distribution of feature importance. The contribution of TOT_VL_NB decreases (0.50), and the importance of PL_VIT and VL_VIT increases (0.18 and 0.19). The importance of TOT_PL_NB is very similar to that of the near dataset (0.12). Even more interesting is the observation of the difference between the importance of each feature in near and far datasets. The distance of TOT_VL_NB is remarkable (0.15), while the distances of the other features are close to zero (-0.06, -0.07, and -0.02 for PL_VIT, VL_VIT, and TOT_PL_NB, respectively).

Table 5. Feature importance values for near and far dataset

	Near	Far
TOT_PL_NB	0.108	0.123
VL_VIT	0.117	0.190
PL_VIT	0.118	0.180
TOT_VL_NB	0.657	0.506

Source: created by the authors.

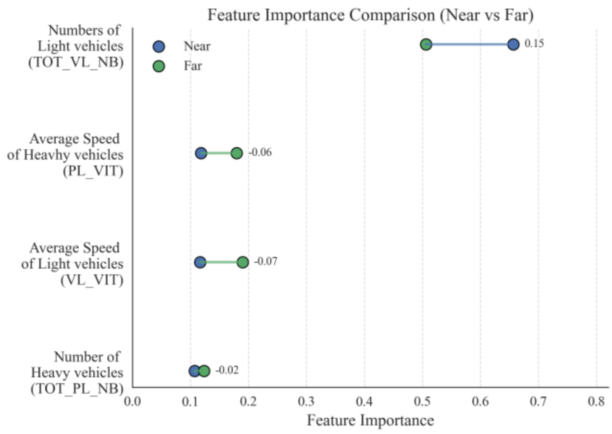

Fig. 5: Comparison of feature importance values for near-field and far-field configurations obtained through a Random Forest regressor.
Source: created by the authors.

An interesting observation is the consistent behavior of the velocity-related variables (VL_VIT and PL_VIT), which exhibit very similar importance patterns in both configurations. This suggests that their contribution is not only limited in magnitude but also structurally comparable. Such an observation clearly addresses the original question of this work: does the distance between source and receiver influence the feature importance in a data-driven prediction model? It appears so. A different approach to this observation is reported in Figure 6, where the shift in feature influence is shown from the features' point of view. Figure 6, in fact, displays how the near-far distance gives more or less importance to a feature on the whole prediction procedure and, moreover, which feature gains or loses importance when the source-receiver distance increases.

From this recapitulating figure, it can be easily seen how the speed-related features become more important in the far field, as long as the number of heavy vehicles increases. On the contrary, the number of light vehicles becomes predominant in the prediction capabilities when the source-receiver distance decreases.

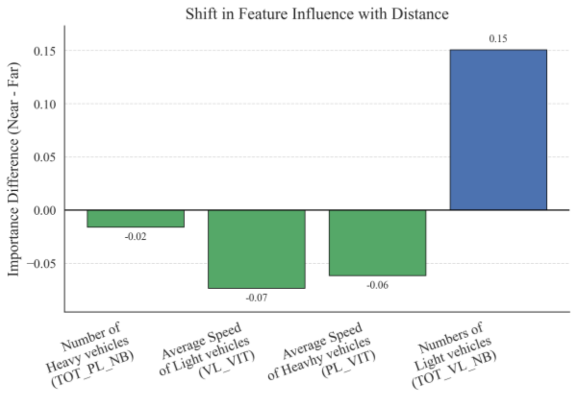

Fig. 6: Difference in feature importance between near-field and far-field configurations, highlighting the shift in variable influence with increasing distance.
Source: created by the authors.

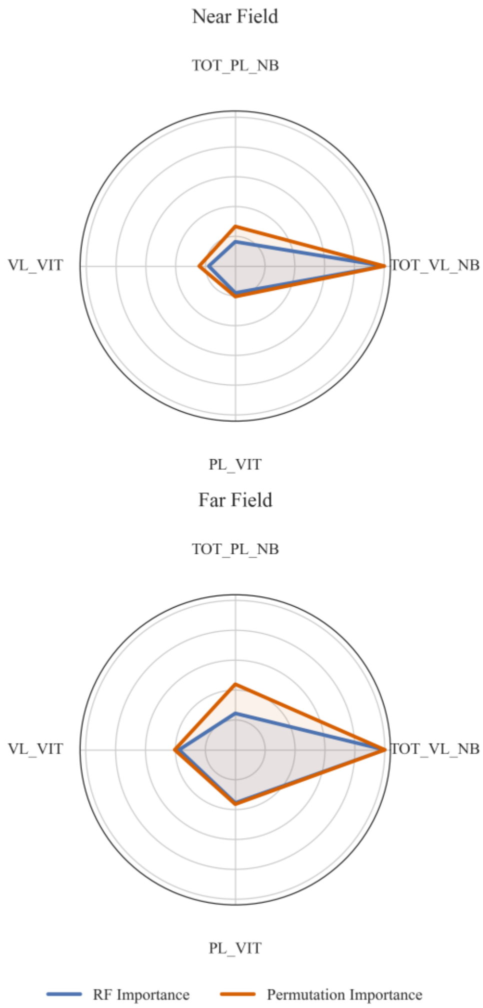

Fig. 7: Comparison between impurity-based Random Forest feature importance and permutation importance for the near-field and far-field datasets. Values were normalized with respect to the most influential feature in each configuration to facilitate the comparison of relevance patterns.
Source: created by the authors.

To ensure the robustness and validity of the feature relevance analysis, a permutation importance evaluation was performed. Figure 7 compares the normalized relevance profiles obtained through impurity-based RF importance and permutation importance for both near-field and far-field configurations. The strong similarity between the resulting profiles indicates that the observed feature relevance patterns are largely independent of the adopted importance metric. Finally, as an additional verification of the stability of the reported feature relevance patterns, a bootstrap-based confidence interval analysis was performed on the feature importance values. Outputs confirm that both in the near and in the far data, TOT_VL_NB has the highest mean importance (0.658 and 0.507, respectively). More interestingly, in both cases the 95% confidence intervals were extremely narrow ([0.650, 0.666] and [0.502, 0.512]), confirming the robustness of the reported rankings with increasing source-receiver distance.

5 Conclusions

The presented study sets itself in a recent research field investigating data-driven approaches to the assessment and forecast of noise levels from road traffic. As already demonstrated (contribution in press), in general the forecast capability of ML models, when applied to road traffic noise, tends to remain stable at variable source–receiver distances, therefore suggesting that this feature (distance) can be considered like one of the other input features. Nonetheless, since it represents a pivotal physical aspect for the propagation of noise, the authors investigated in detail the relationships existing between the various input variables and output noise levels, to discover if they remain stable across distance. In this contribution, an exploratory data analysis approach has been performed for the study of the effects of source–receiver distance when evaluating the output of a road traffic noise data-driven model. Whereas previous studies focused on predictive performance, the presented analysis specifically addressed how the relationships between input variables and acoustic outputs change across distance. The findings are interesting, since the results show that the influence of traffic-related variables on the model outputs weakens as the receiver moves away from the source. By means of correlation analysis and feature relevance estimation, the authors found that variables related to vehicle speed, for example, lose their explanatory power across distance. More in detail, when source-receiver distance grows, ML models maintain their predictive capability by shifting from a dominance of a single input variable (light vehicle flow) to a more distributed contribution of all the remaining ones. This result interestingly suggests how ML models can adapt their output results when one of them significantly varies. This aspect has important implications for the interpretation of data-driven models applied to road traffic noise estimation.

It appears that in near configurations the prediction is more directly governed by the input features reporting traffic intensity, while in far scenarios it is driven by a more complex interaction of all the features. An additional permutation importance analysis was performed to assess the robustness of the feature relevance results. The obtained rankings were found to be consistent with those derived from the impurity-based Random Forest importance measure, supporting the reliability of the observed trends. Overall, the obtained results suggest that distance is not an input feature just like the other ones, and its importance remains high just like in the classic physical models, even if its influence does not directly emerge from performance metrics, since it remains apparently inaccessible within the inner relationship of features used by the model. The author would like to underline, moreover, that due to the exploratory nature of the study, methodological choices have been consciously maintained at their minimum (only two receivers and a very simple tree-based ML model have been tested and evaluated). At this stage of analysis, therefore, no advanced modelling techniques like hyperparameter optimization, complex feature engineering, or interpretability frameworks (e.g., SHapley Additive exPlanations (SHAP) analysis) have been used, apart from a complementary permutation importance assessment that was considered to verify the robustness of the obtained rankings. This extreme simplification is, of course, one of the limitations of this study. More sophisticated analyses of feature relevance can be implemented, and the evaluation of distance influence can be extended to more ML models. Since they do not share all the basic mathematical functions, they could react differently to variations in source-receiver distance. Future improvements of this research will then be devoted to the expansion of this study to other source-receiver distances, as well as to other models. Additional future developments may include validating the proposed methodology using datasets collected in different geographical contexts and under different traffic conditions. Moreover, integrating meteorological variables and additional acoustic descriptors may provide a more comprehensive understanding of the factors influencing noise generation and propagation. Such extensions could further improve the interpretability and generalizability of data-driven noise prediction models.

Acknowledgements:
Authors want to thank dr. Andrea di Filippo for valuable help in producing Figure 1 and for fruitful scientific discussion.

References:
[1] Botteldooren D., Fast noise mapping: A machine learning approach for predicting traffic noise indicators, The Journal of the Acoustical Society of America, Vol.153, No.3_Supplement, 2023, pp. A261–A261, DOI: 10.1121/10.0018786..
[2] European Environment Agency, Environmental Noise in Europe 2020, Publications Office of the European Union, Luxembourg, 2020, ISBN: 978-92-9480-209-5.
[3] Ahmed A. A., Pradhan B., Chakraborty S., Alamri A., Developing Vehicular Traffic Noise Prediction Model Through Ensemble Machine Learning Algorithms with GIS, Arabian Journal of Geosciences, Vol.14, 2021, Article 1729 , DOI: 10.1007/s12517-021-08114-y.
[4] Zhang X., Kuehnelt H., De Roeck W., Traffic Noise Prediction Applying Multivariate Bi-Directional Recurrent Neural Network, Applied Sciences, Vol.11, No.6, 2021, pp. 2714, DOI: 10.3390/app11062714.
[5] Rossi D., Mascolo A., Singh D., Guarnaccia C., Machine Learning Regressors Calibrated on Computed Data for Road Traffic Noise Prediction, Machine Learning and Knowledge Extraction, Vol.7, No.4, 2025, pp. 133, DOI: 10.3390/make7040133.
[6] Wang J., Wang X., Yuan M., Hu W., Hu X., Lu K., Deep Learning-Based Road Traffic Noise Annoyance Assessment, International Journal of Environmental Research and Public Health, Vol.20, No.6, 2023, pp. 5199, DOI: 10.3390/ijerph20065199.
[7] Yin X., Fallah-Shorshani M., McConnell R., Fruin S., Franklin M., Predicting Fine Spatial Scale Traffic Noise Using Mobile Measurements and Machine Learning, Environmental Science & Technology, Vol.54, No.20, 2020, pp. 12860–12869, DOI: 10.1021/acs.est.0c01987
[8] Ruškić N., Mirović V., Marić M., Pezo L., Lončar B., Nićetin M., Ćurčić L., Model for

Determining Noise Level Depending on Traffic Volume at Intersections, Sustainability, Vol.14, No.19, 2022, Article 12443, DOI: 10.3390/su141912443

[9] Acosta O., Montenegro C., González Crespo R., Road Traffic Noise Prediction Model Based on Artificial Neural Networks, Heliyon, Vol.10, No.17, 2024, e36484, DOI: 10.1016/j.heliyon.2024.e36484.

[10] Singh D., Nigam S. P., Agrawal V. P., Kumar M., Modelling of Urban Traffic Noise Using a Systems Approach, Proceedings of the Institution of Civil Engineers – Transport, Vol.171, No.4, 2018, pp. 235–244, DOI: 10.1680/jtran.16.00094.

[11] Rossi D., Pascale A., Mascolo A., Guarnaccia C., Coupling Different Road Traffic Noise Models with a Multilinear Regressive Model: A Measurements-Independent Technique for Urban Road Traffic Noise Prediction, Sensors, Vol.24, No.7, 2024, pp. 2275, DOI: 10.3390/s24072275.

[12] Klimke M., Völz B., Buchholz M., Automatic Intersection Management in Mixed Traffic Using Reinforcement Learning and Graph Neural Networks, Proceedings of the 2023 IEEE Intelligent Vehicles Symposium (IV), 2023, pp. 1–8, DOI: 10.1109/IV55152.2023.10186800.

[13] Tursunalieva A., Alexander D. L. J., Dunne R., Li J., Riera L., Zhao Y., Making Sense of Machine Learning: A Review of Interpretation Techniques and Their Applications, Applied Sciences, Vol.14, No.2, 2024, Article 496, DOI: 10.3390/app14020496.

[14] Barrigón Morillas J. M., Gómez Escobar V., Méndez Sierra J. A., Vílchez-Gómez R., Trujillo Carmona J., An environmental noise study in the city of Cáceres, Spain, Applied Acoustics, Vol.63, No.10, 2002, pp. 1061–1070, DOI: 10.1016/S0003-682X(02)00030-0.

[15] Breiman L., Statistical Modeling: The Two Cultures, Statistical Science, Vol.16, No.3, 2001, pp. 199–231, ISSN: 08834237.

[16] Rudin C., Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, Nature Machine Intelligence, Vol.1, No.5, 2019, pp. 206–21, DOI: 10.1038/s42256-019-0048-x.

[17] Molnar C., Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd Edition, Independently Published, 2022, ISBN: 9798411463330.

[18] Lundberg S. M., Lee S. I., A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, Vol.30, 2017, DOI: 10.48550/arXiv.1705.07874.

[19] Gauvreau B., Long-Term Experimental Database for Environmental Acoustics, Applied Acoustics, Vol.74, No.7, 2013, pp. 958–967 , DOI: 10.1016/j.apacoust.2013.01.008.

[20] Teng J., Zhang C., Gong H., Liu C., Machine learning-based urban noise appropriateness evaluation method and driving factor analysis, PLOS ONE, Vol.19, No.12, 2024, pp. e0311571, DOI: 10.1371/journal.pone.0311571.

[21] Pearson K., Notes on Regression and Inheritance in the Case of Two Parents, Proceedings of the Royal Society of London, Vol.58, 1895, pp. 240–242, 10.1098/rspl.1895.0041.

[22] Gerstorfer Y., Hahn-Klimroth M., Krieg L., A Notion of Feature Importance by Decorrelation and Detection of Trends by Random Forest Regression, Data Science Journal, Vol.22, No.1, 2023, Article 42, DOI: 10.5334/dsj-2023-042.

[23] Yuan X., Liu S., Feng W., Dauphin G., Feature Importance Ranking of Random Forest-Based End-to-End Learning Algorithm, Remote Sensing, Vol.15, No.21, 2023, Article 5203, DOI: 10.3390/rs15215203.

[24] Archer K. J., Kimes R. V., Empirical Characterization of Random Forest Variable Importance Measures, Computational Statistics & Data Analysis, Vol.52, No.4, 2008, pp. 2249–2260, DOI: 10.1016/j.csda.2007.08.015.

[25] Chen R.-C., Dewi C., Huang S.-W., Caraka R. E., Selecting Critical Features for Data Classification Based on Machine Learning Methods, Journal of Big Data, Vol.7, No.1, 2020, Article 52, DOI: 10.1186/s40537-020-00327-4.

[26] Breiman L., Random Forests, Machine Learning, Vol.45, No.1, 2001, pp. 5–32, DOI: 10.1023/A:1010933404324.

[27] Guyon I., Elisseeff A., An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, Vol.3, 2003, pp. 1157–1182

[28] Bellinger C., Jabbar M. S. M., Zaïane O., Osornio-Vargas A., A systematic review of data mining and machine learning for air pollution epidemiology, Environmental Research, Vol.151, 2016, pp. 822–832, DOI: 10.1186/s12889-017-4914-3.

[29] Rudin C., Chen C., Chen Z., Huang H., Semenova L., Zhong C., Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges, Statistics Surveys, Vol.15, 2021, pp. 1–85, DOI: 10.1214/21-SS133.

[30] Umar I. K., Adamu M., Mostafa N., Riaz M. S., Haruna S. I., Hamza M. F., Ahmed O. S., Azab M., The state-of-the-art in the application of artificial intelligence-based models for traffic noise prediction: a bibliographic overview, Cogent Engineering, Vol.11, No.1, 2024, pp. 2297508, DOI: 10.1080/23311916.2023.2297508

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Conceptualization: CG, DR; Data curation: CG, DR; Formal Analysis: DR, DS: Funding acquisition: CG; Investigation: CG, DR, DS; Methodology: CG, DR, DS; Resources: CG; Software: DR; Supervision: CG; Validation: CG, DR, DS; Visualization: CG, DR, DS; Writing – original draft: DR; Writing – review and editing: CG, DR, DS.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.