Pass-by Noise Modelling Applying Machine Learning

Xue Zhang, Helmut Kuehnelt, Wim De Roeck

Pass-by Noise Modelling Applying Machine Learning

Xue Zhang[1,2] **Helmut Kuehnelt[1]** **Wim De Roeck[2]**
[1] Center for Low-Emission Transport, AIT Austrian Institute of Technology GmbH, Vienna, Austria
[2] Department of Mechanical Engineering, KU Leuven, Leuven, Belgium
Xue.Zhang@ait.ac.at

ABSTRACT

According to World Health Organization (WHO), traffic noise is the second deadly environmental pollution in EU that affects health badly after air pollution. With the developing of information and communication technology (ICT), more and more complex traffic data are generated and collected. Meanwhile the emerging field of data mining and machine learning brings data-driven methods under spotlight.

In this paper, the standard pass-by noise data from Applus+ IDIADA is used for training and validating noise models that can predict the pass-by noise of a single vehicle. Different machine learning methods for regression, i.e. multiple linear regression, polynomial regression, decision trees, support vector machine, are applied to the IDIADA data for models training with KFold cross validation. Several models are trained separately based on different gears, driving conditions (wide open throttle and cruise driving mode), and different mass of vehicles (lighter or heavier than 3500 Kg).

A generalized model with integrating mass of vehicles as a categorical feature is also achieved. The accuracy (R^2, coefficient of determination) of noise prediction in the generalized model reaches up to 0.99 with applying support vector regression, decision trees and polynomial regression. These three models are then compared with applying statistical tests and proved to be not significantly different in the model performance with the available dataset.

In conclusion, this experiment of model training shows promising to apply machine learning in the field of transport, specifically in traffic noise study.

For the next step, the noise model will be extended to real traffic scenarios instead of a single vehicle. More features related to traffic will be introduced.

Keywords: pass-by noise, machine learning, polynomial regression, multiple linear regression, decision trees, support vector machine, KFold cross validation, statistical test, null hypothesis, p-value.

1. INTRODUCTION

With the open question, how big data and machine learning helps in automotive industry, specifically in pass-by noise reduction, the literature review is carried out in following two aspects, the applications of big data in transport field and traffic noise modelling.

In terms of big data in transport, public transport ridership prediction, travel demand prediction, travel pattern identification, traffic flow identification and prediction, traffic accident analysis, Electric vehicles (EV) and smart city related studies have been vastly investigated. There is increasing number of studies being performed in public transport sector in the last decades. The authors in [1] have depicted how public transport users can benefit from big data in the age of digitalization. Gradient boosting algorithm is proved to have big advantages in predicting short-term subway ridership among multimodal public transportation system [2]. The average travel demand of passengers to enter each station with machine learning and neural network models for both long and short term is predicted in [3]. In order to mitigate the congestion and improve the overall performance caused by increased number of vehicles, [4] has proposed a data-mining procedure to model travel patterns of passengers in Beijing. In the work of [5], the pattern extraction with battery data as well as energy management scheme in order to solve the issue of 'range anxiety' has been discussed. Regional traffic flow correlation analysis in both temporal and spatial domain has been analyzed in [6] and real time traffic flow prediction is achieved based on this optimized temporal-spatial-historic traffic model. Different machine learning models have been developed in [7] for automatic classification of injury severity types from different traffic accidents, in order to draw behavior roadway accident patterns. In [8] the author has explained data analytics for electric vehicle grid integration and shown an overview of smart grid and EV integration especially in charge planning and how EV sells power back to the grid. An architecture built on Hadoop for real time EV data management in terms of collection, storage and processing for knowledge discovery and decision making is clarified in [9]. Among all present studies, emission assessment is relatively rarely studied. The authors in [10] have applied an optimized energy system model to predict the CO2 emission by electric vehicles in Germany in 2030. A multi-purpose data processing platform has been introduced in [11] to investigate the mobility patterns in urbanized areas the potential impact in terms of gaseous emissions and energy demand by shifting vehicles fleet from conventional fuel vehicles to EVs. The implementation of smart city is proved to be able to significantly reduce the greenhouse gas emissions in [12].

To reduce the overall traffic noise in the environment for ensuring a better life quality, European Union has set severe regulations of noise limit for different vehicle categories. The pass-by noise measurement test is clarified in common international standards, such as ISO362. The measurement is mandatory for vehicle manufactures, who should provide a statement that noise level of their vehicles complies with requirements of the regulation. Due to above mentioned politic and environment concern, traffic

noise model from vehicle fleet has been since long investigated, mostly required by government authorities [13]. However, these models tend to predict the average noise level over a period of time as a relatively rough assumption. For example, in [13] six commonly used models are compared, according to different regions, FHWA TNM highway noise prediction model for USA, 01 db MITHRA highway and railway noise prediction model for France and Belgium, CoRTN highway noise prediction model for UK, Australia, Hong Kong and New Zeeland. RLS90 highway and car park noise prediction model for Germany, STL-86 highway, tram and light rail noise model for Switzerland, ASJ-1993 highway noise prediction model for Japan. Most of these models mentioned above are based on a constant speed assumption and relatively obsolescent. A more recent study has been performed in [14] to compare the new adopted traffic noise models in recent years. Japan has updated the ASJ-1993 to ASJ RTN-Model 2008, which can be applied further in different traffic conditions. A unified model for northern countries, Norway, Denmark, Sweden and Finland, called Nord 2000 has been developed for road and rail traffic. France has also improved the previous model and built a new one called NMPB-Routes-2008, which can be applicable to both highway and other road networks. HARMONOISE model has been proposed for all EU members for the purpose of assessment and management of environmental noise as well as road and railway traffic. The acceleration and deceleration related corrections are also brought into some of these models. Based on the existing European models, EU has again improved and developed the most state-of-art generalized noise model called CNOSSOS-EU [15], inherited calculation coefficients coming from large databases through the development of Nord2000, HARMONOISE ad IMAGINE projects [16][17].

2. DATA SOURCE

In this paper, advanced machine learning methods will be applied to build single vehicle pass-by noise model based on real data from standard pass-by noise testing, provided by Applus IDIADA, consisting of data from passenger car, B segment (PC) and light commercial vehicle over 3500kg, class N2 (LCV).

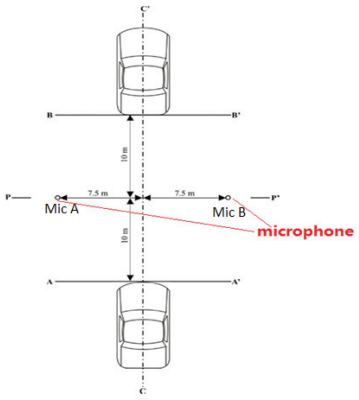

Figure 1. Standard pass-by noise test of passenger car setup by Applus IDIADA

Fig. 1 shows the standard pass-by noise test setup of passenger car. The test site and ambient conditions are in accordance with ISO 10844:2014 "Acoustics: Specification of test tracks for measuring noise emitted by road vehicles and their tyres". As shown in Fig. 1, two stationary microphones are located on the ground with the height of 1.2m ± 0.02m and the distance of 7.5m ± 0.05m to the center as tagged in line PP'. The testing vehicle drives from AA' to BB' which are parallel to line PP' with 10 meters distant to each other. CC' is the driving path, perpendicular towards AA' and BB'. The test consists of constant speed test with cruise control (CRS) and acceleration test with wide open throttle (WOT). For the acceleration test with WOT, the manufacturer shall define the position of the reference point in front of AA'. At the defined reference point the accelerator pedal should be fully depressed and kept this condition until the vehicle reaches line BB'. Then the accelerator shall be released immediately. The test speed is 50 km/h ± 1 km/h, the test speed shall be reached at the reference line PP'. The constant speed test shall be done with the same gear used in the acceleration test and at constant speed of 50 km/h with a tolerance of ±1 km/h between AA' and BB'. In terms of LCV, constant speed test is not required for vehicles with a PMR (power mass ratio) < 25. With regards to acceleration test, the vehicle speed shall be 35 km/h ± 5 km/h.

In this paper, the data from IDIADA consists of the acceleration test and constant speed test of PC with Gear3 and Gear4 and acceleration test of LCV with Gear3. The following numeric information is used in machine learning model training, including root mean square (RMS) sound pressure, vehicle instantaneous speed, location on track, distance to microphone and the corresponding time stamp. IDIADA's audio test equipment has calculated the root mean sound pressure based on unit driving distance of 0.1m, i.e. on non-constant time window.

2.1 Data Pre-processing

The categorical data in Dataset5 has been processed with label encoding and transformed into numeric data.

Many machine learning algorithms can be only applied to numeric data as they are mostly based on mathematic models such as linear regression, support vector regression, etc. although tree related algorithms can handle categorical data as they are not algebraic. One hot encoding and label encoding are two common ways for handling categorical data in machine learning. Label encoding is a simple approach to transfer each value to a number from 0 to (number of categories - 1). One hot encoding transforms each category value into a new feature and assigns 1 or 0 to this feature accordingly.

Comparing with one hot encoding, label encoding is very easy and intuitive, but it has the drawback of misinterpretation as having order hierarchy. One hot encoding will not weight a value improperly but will add many more additional features with regards to the number of categories and cause high dimensionality issue. In this paper, there is only one categorical variable, containing only two categories, therefore, the label encoding is applied, which has assigned 0 to PC and 1 to LCV.

2.2 Data Description

Experiment	Dataset1 (PC, gear3, WOT)	Dataset2 (PC, gear3, CRS)	Dataset3 (PC, gear4, WOT)	Dataset4 (LCV, gear3, WOT)	Dataset5(PC+LCV, gear3, WOT)
Size	311 samples * (4 attributes + 1 target)			373 samples* (4 attributes + 1 target)	684 samples* (4 attributes + 1 target)
Target	RMS sound pressure (Pa)				RMS sound pressure (Pa)
Features	Vehicle speed (m/s)				Vehicle speed (m/s)
	Distance to Mic (m)				Distance to Mic (m)
	Vehicle location (m)				Vehicle location (m)
	Relative time (s)				Vehicle category

Table 1. Basic information of the datasets from 5 different experiments

Tab. 1 shows the attributes information in all available datasets. 5 experiments based on 5 datasets are used for modelling and compared pair-wise in this paper. As Tab. 1 indicates, Dataset1, 2, 3 are all standard pass-by noise testing data from passenger car, with a different driving mode, Gear3 WOT, Gear3 CRC, Gear4 WOT respectively. Dataset4 is from LCV with Gear3 WOT and Dataset5 has the driving data of Gear3 WOT from PC and LCV altogether, which is a mixture of Dataset1 and Dataset4. All datasets have the target variable, RMS sound pressure (from Mic A). The first 4 datasets have the same feature variables: vehicle instantaneous speed, time stamp (the test starting time is always considered to be 0, the other time stamps are referring to the test start time), distance to Mic A, vehicle location on the track (the center of the test track is considered 0 as the reference point). In Dataset5, instead of the variable 'Relative time', another new feature variable, 'Vehicle category', is added, the categorical data is through label encoding transformed to numeric data as aforementioned.

The whole training and analysis are quite similar for each of these datasets. In this paper, the analysis focuses on Dataset5.

The correlation matrix states how the features as well as the target relate to each other. It is very necessary to understand the relationship of them before using the data for model training, especially in the case when the data is very complex with a high dimension. When two feature variables have high correlation, it means their trends are similar and possibly they carry repetitive information, which will lead to bad performance in training the models. Therefore, high correlation filter should be applied to achieve dimension reduction. When the correlation coefficient is beyond a threshold value, one of the two highly correlated features can be dropped. On the other hand, if the feature variable shows a high correlation with the target variable, it should be kept.

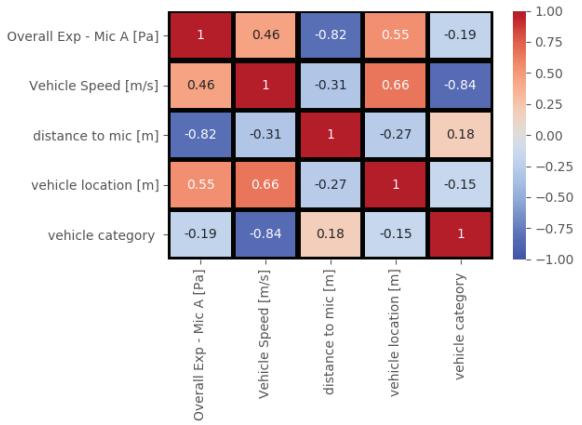

Figure 2. Correlation heatmap of all variables from Dataset5

Fig. 2 shows the correlation heatmap of the input data. The most correlated feature with the target variable, RMS sound pressure ('Overall Exp – Mic A [Pa]'), seems to be 'distance to mic'. 'Vehicle category' plays the least important role for affecting the sound pressure. 'Vehicle location' and 'vehicle speed' have intermediate effect on RMS sound pressure. On the other hand, the two feature variables, 'vehicle speed' and 'vehicle category' have the correlation coefficient of -0.84, which shows relatively high correlation since in the standard pass-by noise testing, the target vehicle speed depends on the vehicle category. However, here both variables will be kept as the test scenario is quite simple and the dimension of input data is already small.

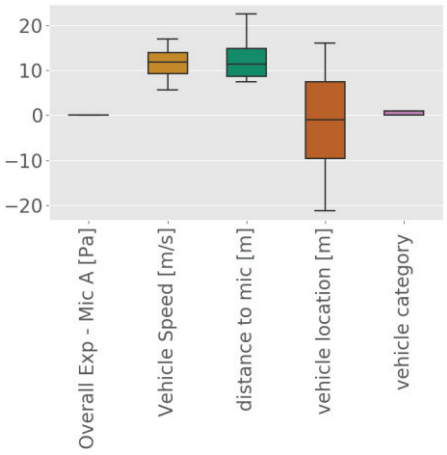

Figure 3. Boxplot of Dataset5

Index	Overall Exp-Mic A[Pa]	Vehicle Speed [m/s]	distance to mic [m]	vehicle location [m]	vehicle category
count	684	684	684	684	684
mean	0.040656	11.6242	12.065	-1.20436	0.545322
std	0.0131771	3.07468	3.79213	10.118	0.498306
min	0.0102728	5.68902	7.5	-21.2	0
25%	0.03048	9.23574	8.63937	-9.5972	0
50%	0.0441843	11.8779	11.3731	-1.06894	1
75%	0.05225	13.9364	14.8478	7.46553	1
max	0.0570808	16.9788	22.4876	16.005	1

Table 2. Dataset5 description

Fig. 3 and Tab. 2 show the descriptive statistics of Dataset5, there are overall 684 samples with one target variable and 4 feature variables. We can see the structure more intuitively in the boxplot of Fig. 3, which shows the input data in quartiles, containing minimum, lower quartile (25th Percentile), median (50th Percentile), upper quartile (75th Percentile) and maximum. In Tab. 2, it shows that the variance/standard deviation of each feature variable (except vehicle category) is quite comparable, feature scaling is not necessary to be applied here. Feature scaling can be a very important step in some cases, as the feature with big variance might dominate the cost function and the selection of the estimator [18], especially in some learning algorithms like SVM (support vector machine) and neural networks. With feature scaling, the learning process can be stabilized [19].

2.3 Data Splitting

The overall data is first split into training set (80%) and test set (20%). The test set is held out aside, without any pre-processing, only used for final model evaluation. The training data is split again to training and validation part with KFold cross validation. Cross validation is a very common method in machine learning to have lower bias. It is a resampling procedure, especially useful and advantageous when the overall sample size is small, so that we can use all our data as both training and validation dataset. With cross validation, the training data is divided into K parts and the training happens K times. Each time, (K-1) parts of the training data are used for training the model and the remaining Kth part as validation set is used for calculating the error and adjusting the parameters. The model with the lowest error is chosen. The test set is at the end used for checking the performance of the final model and the calculated estimation skill can be also used for different model comparison. KFold cross validation can be used also on the entire data set without the held-out test set when no parameter optimization is needed [18]. In this paper K is set to 20. 20Fold cross validation as shown in Fig. 4 is applied to each dataset.

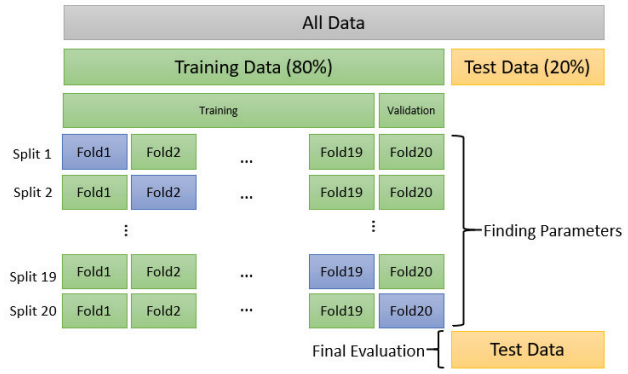

Figure 4. Data split and 20Fold cross validation, figure adapted after [18]

3. MACHINE LEARNING MODEL TRAINING

Predicting the sound pressure with labelled data is a supervised learning regression task. In this paper, linear regression, polynomial regression, decision trees and support vector regression have been tested with all five datasets respectively.

3.1 Multiple Linear Regression (LR)

According to [20], 'the linear model has been a mainstay of statistics for the past 30 years and remains one of our most important tools.' As the most popular method, least squares estimate has been used to fit the data to the linear model, with the aim to minimize the mean squared error.

$$min \sum_{i=1}^{m}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{m}\left(y_i - (\omega \cdot x_i + b)\right)^2 \quad (1)$$

$$\hat{\omega} = (X^T X)^{-1} X^T Y \quad (2)$$

Eqn. (1) is the cost function, where y_i is the true value, $\hat{y}_i$ is the predicted value, $\hat{\omega}$ is the estimation of the parameter. With least square method, $\hat{\omega}$ in Eqn. (2) will be finally obtained.

3.2 Polynomial Regression (PO)

Although polynomial regression fits a nonlinear model to the data, from estimation point of view, the target is the same as in Eqn. (2), the estimated parameter $\hat{\omega}$ is in the linear form, therefore polynomial regression is considered as linear model.

$$y = \omega_0 + \omega_1 \cdot x + \omega_2 \cdot x^2 + \cdots + \omega_p \cdot x^p + \epsilon \quad (3)$$

$$y = \omega_0 + \omega_1 \cdot x_1 + \omega_2 \cdot x_2 + \cdots + \omega_p \cdot x_p + \epsilon \quad (4)$$

$$x_i = x^i \quad (5)$$

Although the variables are raised to higher power in Eqn. (3), the problem can be converted to Eqn. (4) of multivariate linear regression with replacing the variables with Eqn. (5). Then the estimation of parameters can be done by least squares method, same as in multiple linear regression mentioned in Eqn. (2).

3.3 Decision Trees (DT)

Decision tree regression breaks down a dataset into smaller subsets to develop a tree structure. Each time, it picks a value to split the data into two subsets as depicted in Fig. 5.

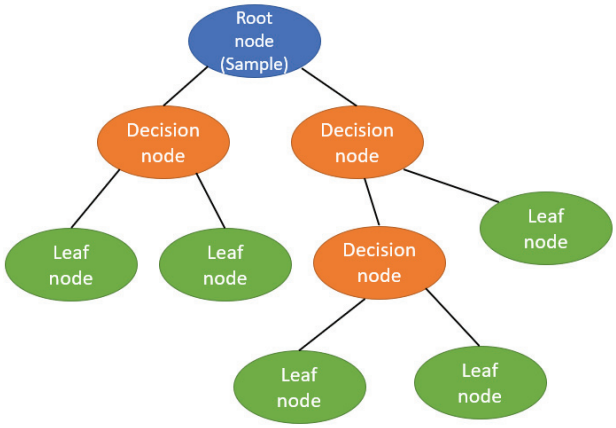

Figure 5. Structure of decision trees

This splitting value selection is based on if it results in smallest mean squared error in each region and will stop growing when a user-defined stopping criterion is met. In

regression the predicted value in each leaf node is the mean value of observations in that region.

The decision trees algorithm is not algebraic, it has a lot of advantages comparing with other algebraic algorithms. Decision trees is well known to be immune to outliers, able to handle both numeric and categorical data, without need of scaling or other transformations of variables. The model is quite interpretable [20]. On the other hand, it is very data sensitive, a small change in the data can cause big change of the tree structure. It can take long time to train the model as it is based on top-down greedy approach.

3.4 Support Vector Regression (SVR)

Different from linear regression, in SVR the error function is ϵ-insensitive as shown below in Fig. 6. That means the model depends only on part of the training data and doesn't care about training points outside of the margin. Inside the margin, the error is tolerated. SVR allows also nonlinear fitting problems (Fig. 6), depending on the kernel function, Eqn. (6), which transforms the data into higher dimensional feature space.

$$K(x_i, x_j) = \Phi^T(x_i)\Phi(x_j) \tag{6}$$

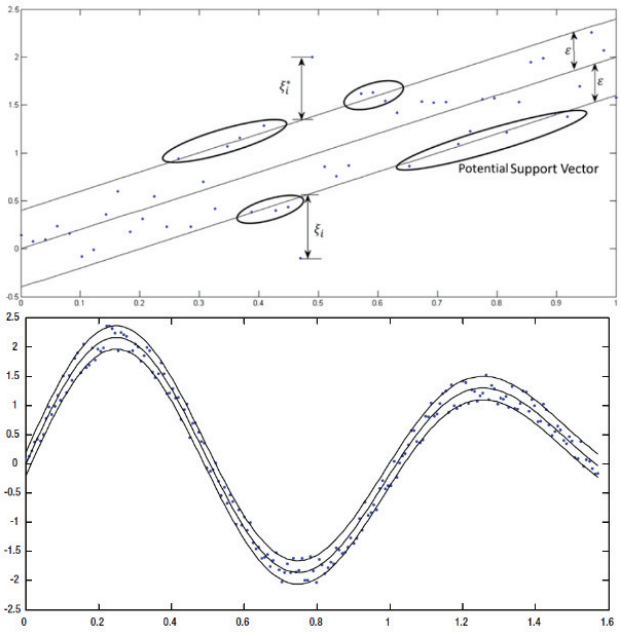

Figure 6. Linear SVR regression (up) and non-linear SVR regression (down) [21]

The SVR first defines a convex ϵ-insensitive loss function and aims to minimize the loss function by finding a flattest tube containing most of the training points. The optimization problem is constrained to:

Minimize:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{7}$$

Subject to:

$$y_i - \langle w, x_i\rangle - b \leq \epsilon + \xi_i \tag{8}$$

$$\langle w, x_i\rangle + b - y_i \leq \epsilon + \xi_i^* \tag{9}$$

$$\epsilon, \xi_i, \xi_i^* \geq 0 \tag{10}$$

The final optimization issue is converted to a constrained problem as shown in Eqn. (7-10), where ξ_i and ξ_i^* are slack variables, which allow regression errors up to the value of ξ_i, ξ_i^*. C is the penalty term, when C =∞, meaning a hard-margin. The strength of C means how much penalty for the error. This full term $C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$ can also be helpful to avoid overfitting.

4. RESULTS AND DISCUSSIONS

4.1 Evaluation methods

In regression task, R-squared is used as a statistical measure of how close the data are to the fitted regression line. R-squared is the coefficient of determination, it is the percentage of variance of dependent variable that has been explained by the independent variables in the model. The value is usually between 0 and 1. If the value is 1, it means the model could explain 100% of the variance, it is an indication of a perfect fit. A constant model that always predicts the expected value of y regardless of the input features will get a R squared score of 0 [18]. R squared can be mathematically expressed as below, Eqn. (11-14):

$$R^2(y_i, \hat{y}) = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \tag{11}$$

$$RSS = \sum(y_i - \hat{y})^2 \tag{12}$$

$$ESS = \sum(\hat{y} - \bar{y})^2 \tag{13}$$

$$TSS = \sum(y_i - \bar{y})^2 \tag{14}$$

where $\hat{y}$ is the predicted value and y_i is the true value, $\bar{y}$ is the average of the samples of the true value y_i. TSS (total sum of squares) is the summation of RSS (residual sum of squares) and ESS (estimation sum of squares). R-squared is the proportion of ESS and TSS.

Besides, in regression task mean squared error and root mean squared error are also often used for model evaluation, which can be calculated through Eqn. (15-16):

$$MSE = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2 \tag{15}$$

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{n}} \tag{16}$$

4.2 Model evaluation

With the held-out test data from each dataset, Fig. 7 and Fig. 8 show the predicted sound pressure vs. the real sound pressure from test data.

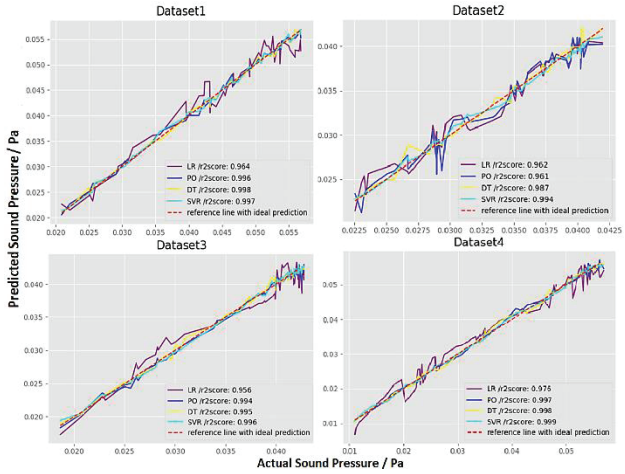

Figure 7. Prediction performance of different models based on Dataset1-4 test set

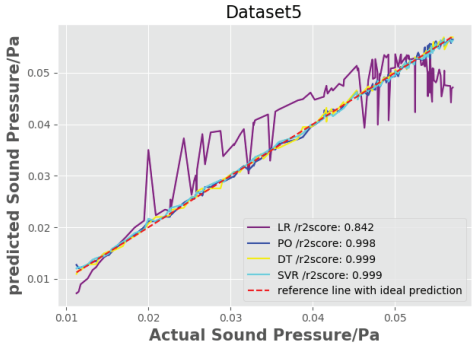

Figure 8. Prediction performance of different models based on Dataset5 test set

These results show that machine learning performs well in predicting pass-by noise of a single vehicle, regardless of driving mode or vehicle type. Decision trees and support vector regression have almost always best prediction with highest coefficient of determination (R2 score). Comparing results from Dataset1 (WOT) and Dataset2 (CRS), the result from Dataset2 is slightly worse as in the CRS driving mode the speed is nearly constant around 50 km/h with a very small variance. Usually a variable with very similar observations is dropped during data pre-processing, because it barely affects the target. Comparing results between Dataset1 (WOT, Gear 3) and Dataset3 (WOT, Gear 4), the prediction performance is almost the same, i.e. machine learning performance is not much affected by gear selection. The same is found comparing Dataset1 and Dataset4 (LCV). In Dataset5, with vehicle category as feature variable, linear regression behaves poorer in comparison with the other three algorithms, while polynomial regression, decision trees and support vector regression algorithms still achieve the best performance with R squared score reaching 0.99.

Algorithms (**Dataset1**)	R squared	MSE	RMSE
Multiple linear Regression	0.964359	3.70889E-06	0.00192585
Polynomial Regression	0.996398	3.74863E-07	0.00061226
Decision Trees	0.997874	2.21194E-07	0.000470312
SVR	0.997235	2.87745E-07	0.000536419
Algorithms (**Dataset2**)	R squared	MSE	RMSE
Multiple linear Regression	0.962213	1.25374E-06	0.00111971
Polynomial Regression	0.960527	1.30966E-06	0.0011444
Decision Trees	0.987217	4.24111E-07	0.000651238
SVR	0.994019	1.98459E-07	0.000445488
Algorithms (**Dataset3**)	R squared	MSE	RMSE
Multiple linear Regression	0.956488	2.10892E-06	0.00145221
Polynomial Regression	0.994061	2.87856E-07	0.000536522
Decision Trees	0.995487	2.18737E-07	0.000467694
SVR	0.995779	2.04562E-07	0.000452285
Algorithms (**Dataset4**)	R squared	MSE	RMSE
Multiple linear Regression	0.975546	5.30286E-06	0.00230279
Polynomial Regression	0.997019	6.46486E-07	0.000804043
Decision Trees	0.99811	4.09815E-07	0.000640168
SVR	0.998568	3.10533E-07	0.000557254
Algorithms (**Dataset5**)	R squared	MSE	RMSE
Multiple linear Regression	0.842048	2.93732E-05	0.0054197
Polynomial Regression	0.997721	4.23855E-07	0.000651042
Decision Trees	0.998805	2.22164E-07	0.000471343
SVR	0.998624	2.55919E-07	0.000505884

Table 3. Evaluation comparison based on test data from Dataset1 to 5

Tab. 3 lists both R squared, MSE and RMSE with the test data from each dataset, applying to each model. All algorithms are performing well for Datset1 to 4 while for Dataset5 polynomial regression, decision trees and support vector regression are performing better with small MSE and high R squared score in comparison with multiple linear regression.

Furthermore, the models of Dataset 5, PO, DT, SVR, are compared pair-wise through statistical significance test with the distribution of error samples obtained through test dataset.

4.3 Statistical testing

Model selection is a crucial part of applied machine learning. The best model is not only decided by regarding the best estimated skill (indicated by R squared score or RMSE) but also by proving that the best estimated skill is not caused by statistical chance.

To solve this, apply the test dataset to different models and get the corresponding error distribution of the test dataset. Then statistical hypothesis test can be applied to the error samples, given the assumption that the samples have the same distribution (null hypothesis). The result of a statistical test such as p-value can be interpreted to quantify the level of confidence or significance in the difference between models, which is compared with the significance level to decide if to reject or fail to reject the null hypothesis. Significance level is often selected as 0.05 or much lower, depending on the condition of the real scenario. Here is 0.05 selected.

There are quite a few different common statistical significance tests. The determination of the corresponding hypothesis test is based on the distribution type of the samples and other factors such as type of variable or the number of models being compared, e.g. Student's t-test assumes a normal distribution of the samples, Chi-squared- test is used to compare categorical variables,

ANOVA (analysis of variance) test is used to compare multiple (three or more) models with a single test.

In this paper, the Wilcoxon signed-rank test is applied pair-wise, because the error samples from the compared models from Dataset5 (PO, DT, SVR) do not all have normal distribution. Tab. 4 lists the result of Wilcoxon signed-rank test, where H0 is the null hypothesis that the error samples have the same distribution:

Wilcoxon signed-rank test	Result
Test 1	Polynomial regression model and decision tree model are NOT significantly different (fail to reject H0)
Test 2	Polynomial regression model and SVR model are NOT significantly different (fail to reject H0)
Test 3	SVR model and decision tree model are NOT significantly different (fail to reject H0)

Table 4. Pair-wise statistical significance test with Wilcoxon signed-rank

We can conclude that the decision trees model, support vector regression model as well as polynomial regression model are all not statistical different with each other in the performance of modelling Dataset5.
Considering the computing time, the simplest well-performed model decision trees model is selected for further model explainability and feature importance study.

4.4 Model Interpretation, Feature Importance

Feature importance check is part of model interpretation, which is necessary for knowing which features play a more important role in determining the prediction. Identifying which variables are important may help in early detection. The more accurate the model is, the more we can trust the result of the feature importance measures. With decision trees and other tree related algorithms, feature importance can be calculated as the normalized total reduction of the Gini impurity [21] brought by that feature.

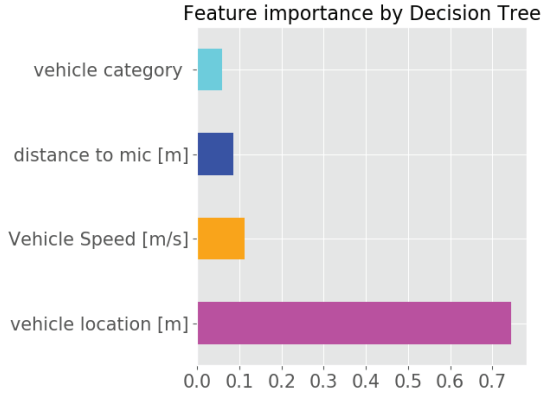

Figure 9. Feature importance from decision trees model

Fig. 9 shows that 'vehicle location' is the most important variable for predicting the sound pressure. 'Vehicle speed' and 'distance to mic' are of similar importance, while 'vehicle category' is slightly less important. 'Sound pressure' versus 'vehicle location' is plotted as below in Fig. 10.

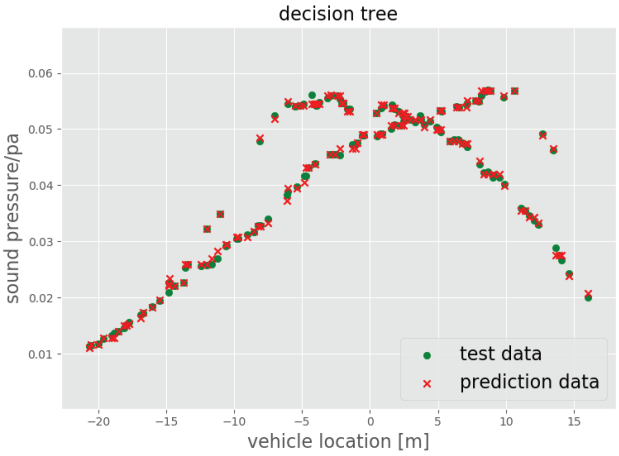

Figure 10. The relation between vehicle location and sound pressure with test data and prediction data

The green plot is the real data from the test and the red plot is the prediction result by the decision trees estimator. Prediction and test data highly overlap, which proves again the good performance of the decision trees model. Besides, it is notable, that the overall data separates in two distinct curves in the figure that correspond to each vehicle category.

5. CONCLUSIONS

This paper investigates the application of machine learning algorithms in single vehicle pass-by noise prediction during the standard pass-by noise testing. Machine learning methods show promising performance for noise prediction. Decision trees, support vector regression and polynomial regression models have very stable performance over all tested datasets with high coefficient of determination and small mean square error. 'Vehicle location' has been shown to have the biggest impact on 'sound pressure'.

In the next step, the method will be applied to a real traffic scenario through a specific case study with more complex traffic situation to evaluate the effectiveness of machine learning algorithms in real-life traffic noise modelling.

6. ACKNOWLEDGEMENT

The author is sincerely thankful to Applus IDIADA for providing the standard pass-by noise test data. The author gratefully acknowledges the European Commission for its support of the Marie Sklodowska Curie program through the ETN PBNv2 project (GA 721615).

7. REFERENCES

[1] M. Yap and M. Munizaga, "Workshop 8 report: Big data in the digital age and how it can benefit public transport users," *Res. Transp. Econ.*, vol.

69, pp. 615–620, 2018.

[2] C. Ding, D. Wang, X. Ma, and H. Li, "Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees," *Sustain.*, vol. 8, no. 11, pp. 1–16, 2016.

[3] F. Toque, M. Khouadjia, E. Come, M. Trepanier, and L. Oukhellou, "Short & long term forecasting of multimodal transport passenger flows with machine learning methods," *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, vol. 2018-March, pp. 560–566, 2018.

[4] X. Ma, Y. J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transp. Res. Part C Emerg. Technol.*, vol. 36, pp. 1–12, 2013.

[5] C. H. Lee and C. H. Wu, "Collecting and Mining Big Data for Electric Vehicle Systems Using Battery Modeling Data," *Proc. - 12th Int. Conf. Inf. Technol. New Gener. ITNG 2015*, pp. 626–631, 2015.

[6] H. P. Lu, Z. Y. Sun, and W. C. Qu, "Big Data-Driven Based Real-Time Traffic Flow State Identification and Prediction," *Discret. Dyn. Nat. Soc.*, vol. 2015, 2015.

[7] M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using machine learning paradigms," *Inform.*, vol. 29, no. 1, pp. 89–98, 2005.

[8] B. Li, M. C. Kisacikoglu, C. Liu, N. Singh, and M. Erol-Kantarci, "Big Data Analytics for Electric Vehicle Integration in Green Smart Cities," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 19–25, Nov. 2017.

[9] V. K. Bolly, J. Springer, and E. Dietz, "Using open source NoSQL technologies in designing systems for delivering electric vehicle data analytics," *ASEE Annu. Conf. Expo. Conf. Proc.*, 2014.

[10] P. Jochem, S. Babrowski, and W. Fichtner, "Assessing CO2 emissions of electric vehicles in Germany in 2030," *Transp. Res. Part A Policy Pract.*, vol. 78, no. 2015, pp. 68–83, 2015.

[11] M. De Gennaro, E. Paffumi, and G. Martini, "Big Data for Supporting Low-Carbon Road Transport Policies in Europe: Applications, Challenges and Opportunities," *Big Data Res.*, vol. 6, pp. 11–25, 2016.

[12] J. Zawieska and J. Pieriegud, "Smart city as a tool for sustainable mobility and transport decarbonisation," *Transp. Policy*, vol. 63, no. June 2017, pp. 39–50, 2018.

[13] C. Steele, "Critical review of some traffic noise prediction models," *Appl. Acoust.*, vol. 62, no. 3, pp. 271–287, 2001.

[14] N. Garg and S. Maji, "A critical review of principal traffic noise models: Strategies and implications," *Environ. Impact Assess. Rev.*, vol. 46, pp. 68–81, 2014.

[15] S. Kephalopoulos, E. Commission, M. Paviotti, E. Commission, and F. A. Ledee, *Common noise assessment methods in Europe (CNOSSOS-EU)*, no. September 2015. 2012.

[16] M. G. Dittrich and X. Zhang, "The Harmonoise/IMAGINE model for traction noise of powered railway vehicles," *J. Sound Vib.*, vol. 293, no. 3–5, pp. 986–994, 2006.

[17] S. Kephalopoulos, M. Paviotti, F. Anfosso-Lédée, D. Van Maercke, S. Shilton, and N. Jones, "Advances in the development of common noise assessment methods in Europe: The CNOSSOS-EU framework for strategic environmental noise mapping," *Sci. Total Environ.*, vol. 482–483, no. 1, pp. 400–410, 2014.

[18] G. M. R. Garreta, *Learning scikit-learn: machine learning in python*. Birmingham: Packt Publishing Ltd, 2013.

[19] K. A. Toh, *Training a reciprocal-sigmoid classifier by feature scaling-space*, vol. 65, no. 1. 2006.

[20] R. T. J. Friedman, T. Hastie, *The elements of statistical learning*. New York: Springer series in statistics, 2001.

[21] K. R. Awad M., *Support Vector Regression. In: Efficient Learning Machines.* Berkeley, 2015.