Traffic noise prediction using machine learning and monte carlo data augmentation: a case study on the Patiala city in India

Daljeet Singh1,a, Priyal Kaler2, Ishita Lyall2, Aekamjot Singh2 and H S Pannu2

1 Mechanical Engineering Department, Thapar Institute of Engineering and Technology, Patiala-147004, India
2 Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala-147004, India

a E-mail: daljeet.singh@thapar.edu

Abstract. Traffic noise pollution is a serious problem in the modern urban areas especially to design new architecture of smart cities, highways, hospitals, schools for an efficient and healthy environment. To analyse this aspect, we have proposed a machine learning based prediction of sound pressure level on an original dataset collected in Patiala city in India. Vehicular traffic and sound pressure level data was collected on different sites in the city. A total of 502 data samples on the identified sites were obtained for the study. Further this data is augmented using Monte Carlo simulation to 10 times of its initial size and the Artificial Neural Networks (ANN) have been trained and compared with other Machine learning methods for the vehicular traffic noise prediction. The input parameters in the model are traffic volume Q, percentage of heavy vehicles P and the average speed of vehicles V and the output parameter is the equivalent continuous sound pressure level, $L_{eq}$ dB(A). The experimental results show ANN which is trained on the augmented data using Monte Carlo simulations outperforms other advanced methods making it an effective measure for vehicular traffic noise prediction to develop a healthy environment which is free of noise pollution.

1. Introduction
The increased vehicular traffic on the roads is a cause of serious concern as it leads to environmental noise pollution, annoyance, sleep disturbance [1], [2], anxiety to drivers and passengers and various other related health problems [3]. The scientific community and the regulatory agencies the world over are working to address this issue. One aspect related to this problem is the assessment of the present noise levels using different measurement techniques and equipment and prediction of the future traffic noise levels to take adequate preventive measures. Figure 1 depicts the proportion of various noise sources causing annoyance in which the road traffic noise pollution contributes to about 80% of overall noise annoyance according to [4].

Some of the existing traffic noise prediction models to address the issue of traffic noise assessment and plan mitigation measures include FHWA (USA), RLS-90 (Germany), CoRTN (UK), ASJ (Japan) [5] which are country specific and may not be applicable everywhere. There are some methods which are based on machine learning techniques like ANN, GA etc. [6]. The parameters generally considered include traffic volume, percentage of heavy vehicles, average speed of vehicles etc.
Figure 1. Distribution of noise annoyance created by human beings through various sources [4]. Road traffic noise is responsible for almost 80% of the overall noise annoyance.

A comparison of some of these methods on the criteria of MSE, correlation coefficient and accuracy has been done by Singh et al. [7]–[9]. But the methods considered are limited in number. Also some other aspects like the augmentation of the data sets considered and the effect of the size of the data set needs to be studied. Some of these aspects have been addressed in the present study. A Monte Carlo simulation has been attempted to augment the dataset and see the effects on the results, i.e. the mean square error and the prediction accuracy, obtained using different Machine Learning techniques.

The material and methods have been presented in section 2, results and discussion in section 3 and finally the conclusions drawn from the study are presented in section 4.

2. Materials and Methods
The dataset collected for the study, the measurement locations, equipment and methods used are presented in this section.

2.1 Dataset for the study
The study locations selected were on the highway roads that bring the traffic into Patiala city. The traffic flow observed is in the free-flow condition as there is no congestion, traffic lights, roundabouts etc. on these locations. The input parameters include the traffic volume (Q), percentage of heavy vehicles (P) and the average speed of vehicles (V). The values of these traffic parameters were measured manually using videography. The output parameter considered is the equivalent continuous sound pressure level, $L_{eq}$. The values of $L_{eq}$ were measured using a class 1 integrating type sound level meter (SLM) confirming with the IEC specifications [10]. The SLM was mounted on a tripod to avoid the human impedance effects and a windscreen was used to cover the microphone. The SLM was kept at a distance of 1m from the edge of the road. A total of 502 datasets were obtained on the three measurement locations during different times of the day and different months throughout the year to account for the variation in the same. A sample dataset is presented in table 1 with a sample of five values for the considered parameters and also the mean and the standard deviation of the entire dataset of 5000 values.
Table 1. Sample dataset where log Q, P and log V are input variables and $L_{eq}$ is the predicted variable in dB(A) units.

| S.No. | $L_{eq}$, dB(A) | log Q | P     | log V  |
|-------|----------------|-------|-------|-------|
| 1     | 74.51          | 3.21  | 19.49 | 1.68  |
| 2     | 76.32          | 3.17  | 15.18 | 1.64  |
| 3     | 76.35          | 3.06  | 39.12 | 1.76  |
| 4     | 77.30          | 3.18  | 35.62 | 1.66  |
| 5     | 73.03          | 3.12  | 22.50 | 1.54  |
| ...   | ...            | ...   | ...   | ...   |
| 502   | 75.2           | 3.1   | 32.6  | 1.7   |
| Mean  | 76.45          | 3.14  | 32.74 | 1.69  |
| S.D.  | 1.68           | 0.04  | 11.68 | 0.05  |

In order to achieve the objective of accurately predicting the values of the output parameter of $L_{eq}$ based on the set of input parameters, the approach of machine learning combined with Monte Carlo simulation was used and is presented in the next section.

2.2 Monte Carlo simulation
The concept of Monte Carlo simulation is used to increase the dataset with random samples rather than using a fixed set of input values for the prediction [11]. The number of experiments using the Monte Carlo method can be increased thousands of times to see more possible outcomes and therefore increase the prediction accuracy as shown in figure 2. It is also observed from figure 2 that as the number of samples increases, the Central Limit Theorem (CLT) states that the distribution becomes normal which is illustrated using the bell curve. The main steps involve studying the probability distribution of the values of the independent parameters or input parameters, generating a range of values for these parameters (using the concept of random sampling) and running the prediction model (machine learning based in the present case) hundreds or thousands of times to get the output results.
Figure 2. Effect of Monte Carlo simulation as the number of repeated samples increases. The distribution tends to be normal as more samples are extracted.

Monte Carlo simulation is a computational algorithm which randomly extracts samples based upon a given data distribution such as a specified mean and standard deviation as given by the following formulae,

$$\mu_X = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$  \hspace{1cm} (1)

$$\sigma_X = \sqrt{\frac{1}{n} \left( \sum_{i=1}^{n} X_i^2 \right) - \frac{1}{n} \left( \sum_{i=1}^{n} X_i \right)^2}$$  \hspace{1cm} (2)

Usually real world problems are deterministic in principle in which random selection is imperative and the original data set is not sufficiently representative of enough population. Monte Carlo repeated random sampling is often used in obtaining a probability distribution, numerical integration and optimization [11]. The simulations are useful where degrees of freedom of input variables are coupled in a complex fashion to generate the output, such as the current problem of noise volume prediction $L_{eq}$, using traffic volume $Q$, percentage of heavy vehicles $P$ and their average speed $V$. Monte Carlo simulation is quite popular in financial modeling, risk management, and complex decision boundary conditions. Central limit theorem (CLT) states that if the sample size is larger than 30 then it approximates to the normal distribution, regardless of the population distribution [12]. Therefore the original data collected was of the size 502 samples can be assumed as an approximation of normal distribution. This relatively smaller dataset of 502 samples has been augmented to 5,000 samples using Monte Carlo simulation. Monte Carlo has been implemented in Excel using NORMINV(.) which takes three inputs (a) random number generator function within a specified range (b) mean and (c)
standard deviation of the underlying distribution of smaller dataset. Every time it is used, it creates a new data entry, so we have used this function 5,000 times with the drag function in Excel.

Table 2. Formulas of performance metrics

| S.No. | Performance Metric                        | Formula                                      |
|-------|-------------------------------------------|----------------------------------------------|
| 1     | Root Mean Square Error (RMSE)             | \[ \sqrt{\frac{\sum(x - y)^2}{n}} \]           |
| 2     | Coefficient of determination R²           | \[ \frac{SS}{TSOS} = 1 - \frac{ESS}{TSOS} \]  |
| 3     | How good is the fit within the given tolerance e (Accuracy) | \[ \frac{\sum f(|x - y| \leq e)}{n} \]          |

2.3 Results and Discussion
The results obtained by applying the methodology discussed in the previous sections are presented below. Neural Network fitting application in MATLAB R2020a has been used to train a 2-layer feed-forward network with linear output neurons and sigmoid hidden neurons. The network is trained using Bayesian Regularization algorithm option with train:validate:test ratio of 70:15:15. Figure 3 shows the outcome of Neural network results for 118 epochs which implies the convergence validation and other parameters using MATLAB fitting application.

Mu refers to the adaption parameter in the underlying optimisation used by ANN during the parameter update calculations. It is defined as momentum update and its value lies in \([0, 1]\). It affects the convergence error and is a control parameter used by ANN during training. It helps to avoid problems such as network getting stuck in the local minima. The details of mu are given in the Levenberg-Marquardt algorithm. ‘Gradient’ stands for minimum performance gradient. It has a default value of 1e-7. ‘gamk’ shows the number of parameters used for training, ‘ssX’ means sum of squares of the parameters, validation checks (val fail) graph at the bottom is about monitoring the possible degradation of the validation failures.

Figure 4 shows the regression plots by the proposed method (Monte Carlo simulation and Neural Networks regression) which is almost 1 for the value of R. Figure 5 shows the frequency histogram of the errors for training and testing target variables (L_eq). Figure 6 shows that the mean square error was minimum at 108 epochs.

Various state-of-art machine learning algorithms have been considered for performance comparison such as, GLM: Generalised Linear Model; DT: Decision Tree; RF: Random Forest; ANN: Artificial Neural Network. Finally the proposed regression model using Monte Carlo data augmentation has been considered which has superior performance. Figure 7 and 8 show the performance comparison histograms for R, RMSE [13], [14] & accuracies within 1 dBA and 0.5 dBA. Tables 3-4 show the values of R, RMSE and accuracies in detail corresponding to the figures 7-8. Formulas of all the performance metrics have been described in table 2.
Figure 3. Neural network results for 118 epochs show the convergence validation and other parameters using MATLAB fitting application.

Figure 4. Regression plots for training, testing and all data samples.
Figure 5. Error histogram for training and testing for 20 bins show the normal distribution and values are negligible due to high accuracy of the trained model.

Figure 6. Graph between epochs and mean squared error. Training is best at 108 epochs.
Figure 7. R and RMSE values for training of GLM, Decision Tree, Random Forest, ANN, and proposed method

Figure 8. Accuracy values for training and testing of GLM, Decision Tree, Random Forest, ANN, and proposed method

Table 3. Performance comparison during the training phase by the underlying techniques

| S.No. | Method           | R    | RMSE | Acc (+1 dBA) | Acc (+0.5 dBA) |
|-------|------------------|------|------|--------------|----------------|
| 1     | GLM              | 0.86 | 0.63 | 81           | 56             |
| 2     | DT               | 0.89 | 0.56 | 88           | 57             |
| 3     | RF               | 0.97 | 0.28 | 97           | 84             |
| 4     | ANN              | 0.9  | 0.49 | 87           | 69             |
| 5     | Monte Carlo+ANN  | 0.99 | 0.18 | 100          | 100            |
Table 4. Performance comparison during the testing phase by the underlying techniques

| S.No. | Method           | R   | RMSE | Acc (±1 dBA) | Acc (±0.5 dBA) |
|-------|------------------|-----|------|--------------|---------------|
| 1     | GLM              | 0.89| 0.58 | 88           | 74            |
| 2     | DT               | 0.89| 0.52 | 86           | 62            |
| 3     | RF               | 0.92| 0.43 | 88           | 73            |
| 4     | ANN              | 0.91| 0.45 | 88           | 68            |
| 5     | Monte Carlo+ANN  | 0.99| 0.22 | 100          | 100           |

3. Conclusions

In the modern world especially in the urban areas, noise pollution is a serious problem which causes health and environmental issues especially around the schools, hospitals, highways and industries. To analyze this problem, we have studied the original dataset collected in the Patiala city including traffic volume, percentage of heavy vehicles, average speed of vehicles and equivalent sound pressure level in dBA. In this paper, vehicular traffic noise has been predicted using three variables: (a) traffic volume Q, (b) percentage of heavy vehicles P and (c) the average speed of vehicles V. Then the output parameter is predicted which is equivalent continuous sound pressure level, Leq dBA. The original data of 502 samples were collected from Patiala city in India and then Monte Carlo simulation has been used to augment the data to increase the size to 10 times i.e., 5,000 samples. It has been observed that machine learning algorithms work well for larger dataset as compared to traditional machine learning algorithms with smaller dataset. Thus, Monte Carlo simulation has been observed to justify its advantage for data augmentation. The machine learning algorithms perform better on a larger training dataset since numerous internal parameters (such as neuron weights in case of ANN) needs to be setup accurately for better generalization and accuracy. This dataset enlargement is provided by Monte Carlo simulation in our study. The assumption machine learning algorithms is parameter tuning and data distribution which must be analyzed carefully. Moreover, machine learning depends upon the data distribution, so a collection of techniques must be explored for every data analysis problem. Our future plan is to collect data specific to places and vehicle types while including more input variables for accurate prediction.

References

[1] Öhrström E, Hadzibajramovic E, Holmes M, and Svensson H 2006 “Effects of road traffic noise on sleep: studies on children and adults,” *J. Environ. Psychol.* 26 (2) 116–126.

[2] Graziuso G, Mancini S, Francavilla A B, Grimaldi M and Guarnaccia C 2021 “Geocrowdsourced Sound Level Data in Support of the Community Facilities Planning. A Methodological Proposal,” *Sustainability* 13(10) 5486.

[3] Babisch W 2003 “Stress hormones in the research on cardiovascular effects of noise,” *Noise Heal.* 5(18) 1.

[4] Amundsen A H and Klæboe R 2005 *A Nordic perspective on noise reduction at the source*, no. 806/2005. Transportøkonomisk institutt.

[5] Steele C, 2001 “A critical review of some traffic noise prediction models,” *Appl. Acoust.* 62(3) 271–287 doi: 10.1016/S0003-682X(00)00030-X.

[6] Kumar P, Nigam S P, and Kumar N 2014 “Vehicular traffic noise modeling using artificial neural network approach,” *Transp. Res. Part C Emerg. Technol.* 40 111-122 doi: 10.1016/j.trc.2014.01.006.

[7] Singh, D, Nigam, S P, Agrawal V P, and Kumar M 2016 “Vehicular traffic noise prediction
using soft computing approach,” *J. Environ. Manage.* **183**(1) 59–66 doi: 10.1016/j.jenvman.2016.08.053.

[8] Singh D, Upadhyay R, Pannu H S, and Leray D 2021 “Development of an adaptive neuro fuzzy inference system based vehicular traffic noise prediction model,” *J. Ambient Intell. Humaniz. Comput.* **12**(2) 2685–2701.

[9] Singh D, Francavilla A B, Mancini S and Guarnaccia C 2021 “Application of machine learning to include honking effect in vehicular traffic noise prediction,” *Appl. Sci.* **11**(13) 6030.

[10] “IEC 61672-1: 2002, International Electrotechnical Commission”.

[11] Raychaudhuri S 2008 “Introduction to monte carlo simulation” *Winter simulation conference* 91–100.

[12] Mulekar M S 2006 “Visualizing The Central Limit Theorem,” *Alabama J. Math.* 25-36.

[13] Kyong-Ii K, Song-Chol K, Song-Chon L and Ryu Chol C 2019 “An Improved Thresholding Method for Wavelet Denoising of Acoustic Signal” *WSEAS Trans. Signal Process.* **15** 65–76.

[14] Vagaská A, Michal P, Bukovský I, Gombár M and Kmec J 2019 “The Application of Neural Networks to Control Technological Process” *WSEAS Trans. Circuits Syst.* **18** 147–153.