# Missing values reconstruction in sound level monitoring station by means of intelligent computing

*Leszek* Radziszewski[1], *Michał* Kekez[1,*], *Alžbeta* Sapietová[2]

[1]Kielce University of Technology, Faculty of Mechatronics and Mechanical Engineering, Aleja Tysiąclecia Państwa Polskiego 7, 25314 Kielce, Poland
[2]University of Žilina, Faculty of Mechanical Engineering, Univerzitná 8215/1, 01026 Žilina, Slovakia

**Abstract.** The aim of the paper was to reconstruct the missing data by applying the model which describes variability of sound level in the whole period from 2013 to 2016. To build the model, the computational intelligence methods, like fuzzy systems, or regression trees can be used. The latter approach was applied and we built the model with Cubist regression tree software, using equivalent sound levels recorded in 2013. For the reconstruction of sound level data in short period of time (several days), time series values and day_of_week values together should be used in the training dataset. For the reconstruction of sound level data in long period of time (several months) day_of_week values should be used in the training dataset.

**Keywords:** reconstruction of missing sound level data, random forest, regression trees

## 1 Introduction

The work presents the analysis of equivalent sound level [1] data recorded during measurements in monitoring stations located at Krakowska Street in Kielce, Poland. Krakowska street is a dual-carriageway road, which connects the south-western part of the city with the expressway towards Cracow. The monitoring station consists of sound level meter and weather station. The microphone for the acoustic pressure measurements is mounted at a distance of 4 m from the edge of the road and 4 m above ground level. Acoustic measurements were carried out [2] by using SVAN 958A, digital, four-channel, class-1, vibration and sound meter. In the research, the ½” prepolarized free field condenser microphone MIKROTECH GEFELL MK 250, which has sensitivity of 50 mV/Pa, was used together with SV 12L preamplifier. The range of frequencies was 3.5 Hz to 20 kHz, and dynamic range was 15–146 dB. The resolution of the signal RMS detector is 0.1 dB. The measurements were carried out 24 hours a day. The RMS values of the A sound level were registered in the buffer every 1 s and the results were recorded every 60 seconds. Based on the measurements that were conducted 24 hours a day, the equivalent sound levels were calculated for three time periods: day (6 a.m. to 6 p.m.), evening (6 p.m. to 10

[*] Corresponding author: mkekez@tu.kielce.pl

Reviewers: *Bohuš* Leitner, *Josef* Soukup

p.m.), and night (10 p.m. to 6 a.m.) [3]. The noise pollution (nuisance) due to long-term exposure to noise is very often measured by the equivalent sound level ($L_{Aeq,T}$), expressed in (dB), defined as [4]

$$L_{Aeq,T} = 10 \cdot \log\left[\frac{1}{T}\int_0^T \left(\frac{p_A(t)}{p_0}\right)^2 dt\right] , \tag{1}$$

where $p_0$ is the standardized reference acoustic pressure of 20 µPa. According to the ISO standard, this parameter can be determined from [5]

$$L_{eq} = 10 \log\left(\frac{1}{N}\sum_{i=1}^{i=N} 10^{0.1 L_{A,i}}\right) , \tag{2}$$

where $L_{A,i}$ is the A-weighted acoustic pressure level measured in the measurement interval $i$. The average sound level can be determined as expected value from [6]

$$\bar{L} = \frac{1}{N}\sum_{i=1}^{i=N} L_{A,i} . \tag{3}$$

## 2 Measurements results

The data collected at Krakowska Street in 2013 consists of 905 records describing equivalent sound level for a particular night (Fig. 1a), evening (Fig. 1b), or day (Fig. 1c). In the year 2013 there are missing values [7] for beginning of January, second half of January, first half of February, and most part of December.

When the missing values for a particular night, evening, and day are replaced by 0 dB, the calculated $L_{Aeq,T}$ values for the whole year are: 63.98 dB (nights), 68.13 dB (evenings), and 69.50 dB (days). When records which originally contained missing values are omitted, the equivalent sound levels are: 64.80 dB, 68.95, and 70.35 dB respectively. The correct calculation of $L_{Aeq,T}$ values requires the reconstruction of the missing data records.

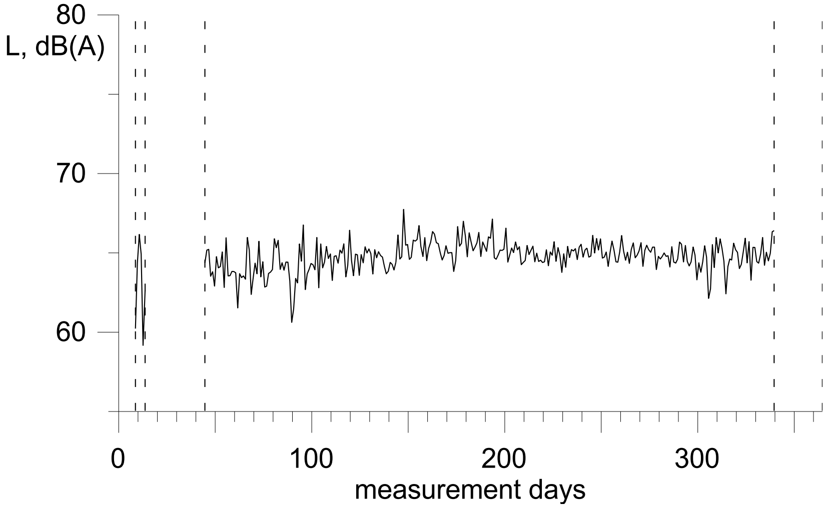

**Fig. 1a.** Krakowska Street, 2013, equivalent sound levels for nights (time_of_day=0)

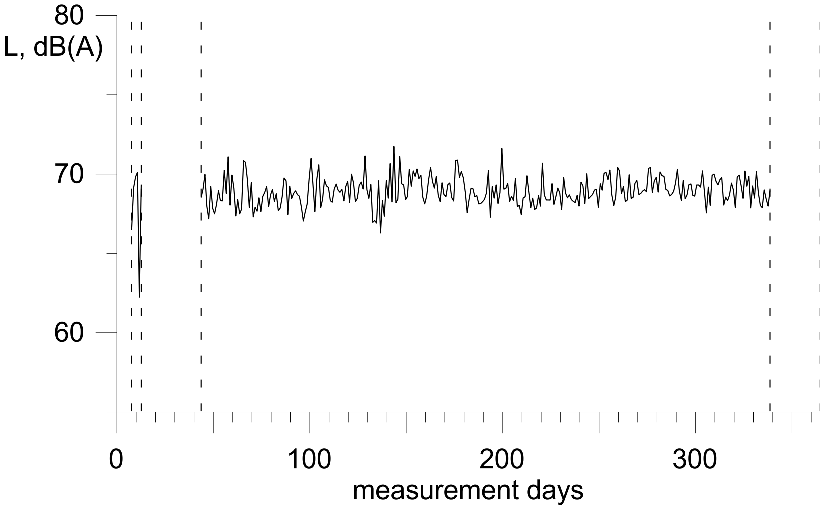

**Fig. 1b.** Krakowska Street, 2013, equivalent sound levels for evenings (time_of_day=1)

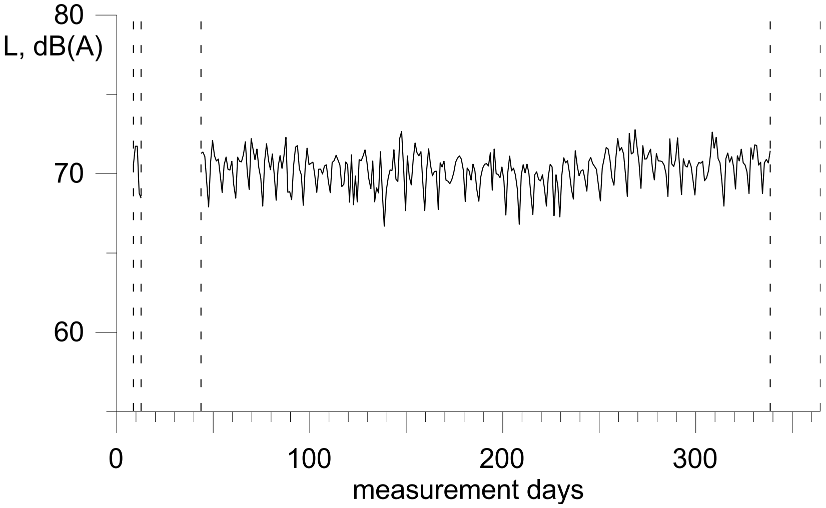

**Fig. 1c.** Krakowska Street, 2013, equivalent sound levels for days (time_of_day=2)

## 3 The proposed models for reconstruction of data for longer periods of time

The aim of this paper is to reconstruct the missing data by applying the model which describes variability of sound level in the whole year 2013. To build the model, the computational intelligence methods, like fuzzy systems [8], regression trees or time series analysis [9] can be used. The second approach was applied and we built the model with Cubist regression tree software [10], using equivalent sound levels recorded in 2013. The training data for Cubist consisted of records, each containing values of one output attribute, *dB_A* (sound level) and 2 input attributes: *day_of_the_week* (from 1 – Monday to 7 – Sunday), and *time_of_day* (0 – night, 1 – evening, 2 – day). The training dataset consisted of records collected in the first half of the year, while the test dataset contained records coming from the second half of the year.

The data was dived into 3 separate sets, one for each time of day (night, day, and evening). Cubist produced 3 rulesets (one for each dataset). Later, they were merged, and the obtained regression tree model had 5 rules:

IF day_of_the_week>6 AND time_of_day=0 THEN db_A = 63.48
IF day_of_the_week≤6 AND time_of_day=0 THEN db_A = 64.03 + 0.15 day_of_the_week
IF time_of_day=1 THEN db_A = 68.81
IF day_of_the_week>5 AND time_of_day=2 THEN db_A = 76.098 – 1.078 day_of_the_week
IF day_of_the_week≤5 AND time_of_day=2 THEN db_A = 70.74

The accuracy of the model on the data regarding the whole year 2013 is quite good (Tab. 1) in terms of mean absolute error (MAE, eq. 4) and root mean square error (RMSE, eq. 5)

$$MAE = \frac{\sum_{i=1}^{n}|\overline{y_i} - y_i|}{n}, \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\overline{y_i} - y_i)^2}{n}}, \tag{5}$$

where $y_i$ denotes real $db\_A$ value, $\overline{y_i}$ denotes $db\_A$ value calculated by the model, and $n$ is the number of records. For every day of week, and every time of day, MAE lies within the range of 0.47 dB (Saturday, day) to 0.98 dB (Sunday, night). The values of RMSE for every day of week and every time of day lies within the range of 0.59 dB (Saturday, day) to 1.28 dB (Sunday, night). The values of MAE for all days of week together are: 0.83 and 0.65 dB (nights), 0.87 and 0.61 dB (evenings), 0.58 and 0.59 dB (days), on the training and test data respectively.

**Table 1.** Accuracy of the Cubist regression tree model on the data from the whole year 2013

| | **Monday** | **Wednesday** | **Friday** | **Sunday** |
|---|---|---|---|---|
| RMSE [dB] | | | | |
| Night | 1.02 | 1.05 | 0.74 | 1.28 |
| Day | 0.93 | 0.72 | 0.79 | 0.75 |
| Evening | 0.92 | 0.92 | 1.10 | 0.75 |
| MAE [dB] | | | | |
| Night | 0.85 | 0.74 | 0.59 | 0.98 |
| Day | 0.67 | 0.54 | 0.60 | 0.59 |
| Evening | 0.77 | 0.71 | 0.81 | 0.62 |

Ten-fold cross validation (repeated several times) shows that MAE (mean absolute error) of the model for the first half of the year 2013 is 0.84–0.87 dB (nights), 0.87–0.88 dB (evenings), 0.58–0.60 dB (days). The scatter plots for prediction of sound level, for the test data (second half of the year 2013) are shown in Fig. 2.

When the missing values of sound level for a particular night, evening, and day are replaced by the values obtained by the Cubist model, the $L_{Aeq,T}$ values calculated for the whole year are: 64.74 (nights), 68.93 dB (evenings), and 70.35 (days).

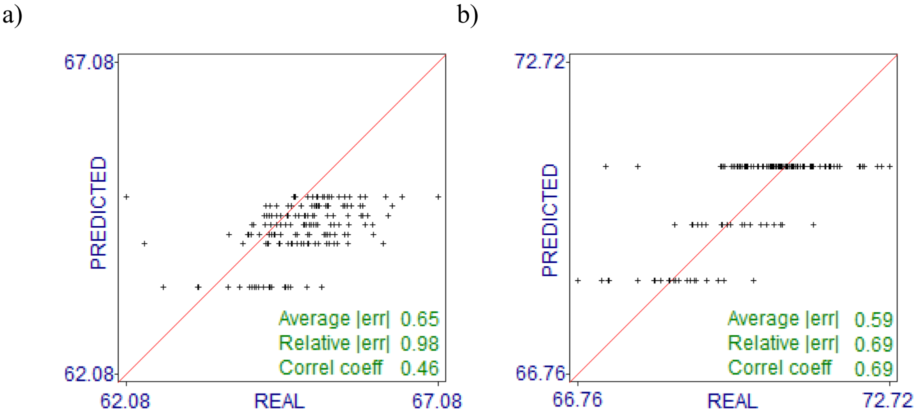

**Fig. 2.** Scatter plots for prediction of equivalent sound level by Cubist regression tree model in second half of the year: a) for nights (time_of_day=0), b) for days (time_of_day=2)

Later, the committee of models was built. At first, the Cubist system was run with the option „committee of 5 members". The transparency of the obtained model was low (5 regression trees in one model), and the accuracy was exactly the same as for single regression tree. Later, Cubist software was run with the option „committee of 100 members". In this case, the transparency of the models was very low (100 regression trees), and the accuracy was almost exactly the same as for the single regression tree.

Later, RandomForest algorithm from WEKA software package was used to build random forest. Because the number of features (i.e. input attributes) is only 2, the random feature selection had almost „nothing to do". The accuracy of the obtained model for evenings is better than that of the single regression tree built by Cubist. Mean average error (MAE) was: 0.84 dB and 0.65 dB for nights, 0.82 dB and 0.59 dB for evenings, 0.59 dB and 0.61 dB for days, on training and test data respectively.

Next, the neural network model (multilayer perceptron) was built by Multilayer Perceptron algorithm from WEKA software. The accuracy of the obtained model is worse than accuracy of regression tree built by Cubist. Comprehensibility of the model is very low, because of the form of the models (weights and thresholds compared to if-then rules in the previously presented models).

## 4 The proposed models for reconstruction of data for short periods of time

The next aim of this paper is to reconstruct the missing data in short periods of time (several consecutive days only) with high accuracy by applying the model which describes variability of sound level in the whole year 2013. In this case, the training data contains time series data *t*_1*, t*_2,.. ,*t*_6. The dataset consisted of records, each containing values of one output attribute, *dB_A* (sound level), and 8 input attributes: *day_of_the_week* (from 1 – Monday to 7 – Sunday), *time_of_day* (0 – night, 1 – evening, 2 – day), *t*_1 (sound level value at the same time of day, but one day earlier), *t*_2 (similarly, but two days earlier), *t*_3, *t*_4*, t*_5, and *t*_6. The training dataset consisted of records collected in the first half of the year, while the test dataset contained records registered in the second half of the year. Again, the data was dived into 3 separate sets, one for each time of day (night, day, and evening). Again, Cubist produced 3 rulesets (one for each dataset). Later, they were merged, and the obtained regression tree model had 5 rules:

IF day_week <= 6 AND time_of_day=0 THEN db_A = 31.598 + 0.343 t_1 + 0.169 t_2
IF day_week > 6 AND time_of_day=0 THEN db_A = 2.999 + 0.656 t_1 + 0.255 t_5 - 0.019 day_week + 0.023 t_2
IF time_of_day=1 THEN db_A = 56.69 + 0.176 t_1
IF day_week > 5 AND time_of_day=2 THEN db_A = 76.346 - 1.116 day_week
IF day_week <= 5 AND time_of_day=2 THEN db_A = 48.393 + 0.25 t_1 - 0.113 day_week + 0.141 t_4 - 0.068 t_5

The transparency of the model is moderate (5 complicated rules). The accuracy is the best, but the model can be applied only for short periods of missing data (up to one week). For longer periods, most input values for the model *(t*_6, *t*_5, ..., *t*_1) must be produced by the model itself, which would lead to lower accuracy. When *t*_1 to *t*_6 are taken from measurement data, the accuracy is 0.65 and 0.53 dB (nights), 0.79 and 0.60 dB (evenings), 0.56 and 0.56 dB (days), on training and test data respectively.

When the missing values of sound level for a particular night, evening, and day are replaced by the values obtained by this model, the $L_{Aeq,T}$ values calculated for the whole year are: 64.74 dB (nights), 68.93 dB (evenings), and 70.35 (days).

When the training dataset did not contain *day_of_the_week* and *time_of_day* input attributes (only *dB_A* and *t*_1, *t*_2, ..., *t*_6 time series attributes were present), the Cubist software produced regression tree, which had much lower accuracy for nights and days than the previous model.

The comparison of results (Tab. 2) shows that the highest accuracy is obtained by the Cubist regression tree built for *day_of_week* and time series (*t*_1, …, *t*_6) training data, and this accuracy is better than that of the classifier presented in the previous section.

**Table 2.** Accuracy of models for equivalent sound level at Krakowska Street monitoring station

| **Mean Average Error, dB(A)** | **Night** | | **Evening** | | **Day** | |
|---|---|---|---|---|---|---|
| | train data | test data | train data | test data | train data | test data |
| Cubist regression tree | 0.83 | 0.65 | 0.87 | 0.61 | 0.58 | 0.59 |
| Committee of 5 regression trees, Cubist software | 0.83 | 0.65 | 0.87 | 0.61 | 0.58 | 0.59 |
| Committee of 100 regression trees, Cubist software | 0.83 | 0.65 | 0.87 | 0.61 | 0.58 | 0.59 |
| Random forest, WEKA software | 0.84 | 0.65 | 0.82 | 0.59 | 0.59 | 0.60 |
| Neural network, WEKA software | 0.93 | 0.61 | 0.87 | 0.62 | 0.63 | 0.61 |
| Cubist regression tree for day_of_week and time series data | 0.65 | 0.53 | 0.79 | 0.60 | 0.56 | 0.56 |
| Cubist regression tree for time series data only | 0.76 | 0.57 | 0.79 | 0.60 | 0.65 | 0.85 |

## Conclusions

The comprehensibility of the classifier produced by Cubist regression tree for Krakowska Street monitoring station is highest among all obtained models. The comparison of results (Tab. 2) shows that for short periods of missing data, the highest accuracy is obtained by the Cubist regression tree built for *day_of_week* and time series (*t*_1, …, *t*_6) training data – 0.53 dB, 0.60 dB, 0.56 dB on test data for nights, evenings, and days respectively. For longer periods of missing data, the best accuracy is achieved by two models, obtained from training dataset which does not contain time series data, namely Cubist regression tree (for night and day equivalent sound levels – 0.83 dB and 0.58 dB respectively, on training data), and WEKA implementation of random forest (for evening equivalent sound levels – 0.82

dB on training data). The time necessary to build most of the presented models does not exceed 0.2 s.

The obtained models allow the calculation of the equivalent sound levels for the whole year: 64.74 dB (for nights), 68.93 dB (for evenings), 70.35 (for days). All the obtained results can be used for verification of sound propagation models during elaboration of city acoustic map.

## References

1. W. Batko, B. Stępień, *Type A Standard Uncertainty of Long-Term Noise Indicators*. Arch. Acoust. **39** (1), 25-36 (2014)
2. A. Bąkowski, L. Radziszewski, Z. Skrobacki, *Assessment of uncertainty in urban traffic noise measurement*. Procedia Engineering **177**, 281–288 (2017)
3. K. Marciniuk, M. Szczodrak, B. Kostek, *Performance of Noise Map Service Working in Cloud Computing Environment*. Arch. Acoust. **41** (2), 297–302 (2016) DOI: 10.1515/aoa-2016-0029
4. R. Makarewicz, R. Gołębiewski, *The Uncertainty of Noise Composed of Separate Sound Events*. Arch. Acoust. **41** (1), 133–138 (2016) DOI: 10.1515/aoa-2016-0013
5. *ISO 9612 :2009, Acoustics – Determination of occupational noise exposure – Engineering method*.
6. D. I. Mihajlov, M. R. Prascevic, *Permanent and Semi-permanent Road Traffic Noise Monitoring in the City of Nis (Serbia)*. J. of Low Frequency Noise, Vibration and Active Control **34** (3), 251–268 (2015)
7. M. Spławińska, *The Problem Of Imputation Of The Missing Data From The Continuous Counts Of Road Traffic*. Archives of Civil Engineering **61** (1), 131–145 (2015) DOI:10.1515/ace-2015-0009
8. M. Kekez, L. Radziszewski, A. Sapietova, *Fuel type recognition by classifiers developed with computational intelligence methods using combustion pressure data and the crankshaft angle at which heat release reaches its maximum*. Procedia Engineering **136**, 353–358 (2016)
9. E. Solazzo, S. Galmarini, *Comparing apples with apples: Using spatially distributed time series of monitoring data for model evaluation*. Atmospheric Environment **112**, 234–245 (2015)
10. https://www.rulequest.com/cubist-info.html