Science of the Total Environment 505 (2015) 680–693

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

A general procedure to generate models for urban environmental-noise pollution using feature selection and machine learning methods

Antonio J. Torija [a,*], Diego P. Ruiz [b]

[a] Department of Electronic Technology, University of Malaga, Higher Technical School of Telecommunications Engineering, Campus de Teatinos, Malaga 29071, Spain
[b] Department of Applied Physics, University of Granada, Avda. Fuentenueva s/n, 18071 Granada, Spain

HIGHLIGHTS

- Machine-learning regression methods are implemented for L_{Aeq} prediction.
- Non-linear solvers outperform linear solver in estimating urban environmental noise.
- SMO and GPR algorithms achieve the best estimation of L_{Aeq}.
- CFS technique allows the greatest reduction in data-collection cost.
- Input variables chosen by WFS technique offers the best results in estimating L_{Aeq}.

ARTICLE INFO

Article history:
Received 23 May 2014
Received in revised form 4 August 2014
Accepted 19 August 2014
Available online xxxx

Editor: P. Kassomenos

Keywords:
Feature selection
Multiple linear regression
Multilayer perceptron
Sequential minimal optimisation
Gaussian processes for regression
Environmental-noise prediction

ABSTRACT

The prediction of environmental noise in urban environments requires the solution of a complex and non-linear problem, since there are complex relationships among the multitude of variables involved in the characterization and modelling of environmental noise and environmental-noise magnitudes. Moreover, the inclusion of the great spatial heterogeneity characteristic of urban environments seems to be essential in order to achieve an accurate environmental-noise prediction in cities. This problem is addressed in this paper, where a procedure based on feature-selection techniques and machine-learning regression methods is proposed and applied to this environmental problem. Three machine-learning regression methods, which are considered very robust in solving non-linear problems, are used to estimate the energy-equivalent sound-pressure level descriptor (L_{Aeq}). These three methods are: (i) multilayer perceptron (MLP), (ii) sequential minimal optimisation (SMO), and (iii) Gaussian processes for regression (GPR). In addition, because of the high number of input variables involved in environmental-noise modelling and estimation in urban environments, which make L_{Aeq} prediction models quite complex and costly in terms of time and resources for application to real situations, three different techniques are used to approach feature selection or data reduction. The feature-selection techniques used are: (i) correlation-based feature-subset selection (CFS), (ii) wrapper for feature-subset selection (WFS), and the data reduction technique is principal-component analysis (PCA). The subsequent analysis leads to a proposal of different schemes, depending on the needs regarding data collection and accuracy. The use of WFS as the feature-selection technique with the implementation of SMO or GPR as regression algorithm provides the best L_{Aeq} estimation ($R^2 = 0.94$ and mean absolute error (MAE) = 1.14–1.16 dB(A)).

1. Introduction

Exposure to environmental noise in urban areas is associated with adverse effects on human health and quality of life (Nega et al., 2012; Lucas de Souza and Giunta, 2011; Vlachokostas et al., 2012). Several studies have indicated that urban noise pollution can lead not only to psychological damage, such as annoyance, sleep disturbance, anxiety, etc., but also to physiological problems, such as cardiovascular risks (Belojevic et al., 1997, 2008; Hofman et al., 1995; Kurra et al., 1999; Shaw, 1996).

Environmental-noise-pollution models are important tools in planning more environmentally friendly urban sound spaces and assessing the impact of environmental noise on the exposed population (Pamanikabud and Tansatcha, 2003; Zhao et al., 2012). In addition, at design stages, environmental-noise-pollution models are needed for planning and designing new neighbourhoods as well as streets in urban environments so that impacted areas have comfortable sound conditions (Steele, 2001; Gündogdu et al., 2005). Consequently, noise models are extensively used in the monitoring of environmental-noise

* Corresponding author. Tel.: +34 95 21 32844.
E-mail addresses: ajtorija@ugr.es, ajtm19@gmail.com (A.J. Torija).

http://dx.doi.org/10.1016/j.scitotenv.2014.08.060

impact and in the management of practical solutions to existing noise problems (Givargis and Karimi, 2010).

Many scientific models have been developed focussing on road-traffic-noise prediction based on source emission and on empirical formulations for sound propagation (Garg and Maji, 2014). These models allow accurate road-traffic modelling. Although road traffic has been identified as the main environmental noise source in cities, it is widely known that one of the most salient characteristics of urban environments is their complexity (Tang and Wang, 2007; Torija et al., 2010, 2012), which is reflected in an accumulation, saturation, and diversity of sound sources, i.e. road traffic, industry, construction, commerce, and social as well as leisure activities. Therefore, the use of such emission–propagation empirical road-traffic-noise models can lead to a serious underestimation of sound levels in urban agglomerations, since the contribution of sound sources other than road traffic is not considered.

Also, built-up environments entail large spatial heterogeneity, with different types of locations (traffic street, square, urban park, etc.), varied geometry of locations, coexistence of diverse sound sources, as well as great temporal heterogeneity, depending on the time of day (day, evening or night) and the type of day (work day, weekend) (Torija et al., 2012). In environmental-noise modelling, this heterogeneity constitutes an essential aspect, since this situational diversity is a key factor to consider in order to develop a well-designed environmental model (Lucas de Souza and Giunta, 2011). For this, it is necessary to undertake a suitable selection of variables related to acoustic emission and propagation by which to characterize built-up environments appropriately (Torija et al., 2010).

These built-up environments, together with the large spatial, temporal, and spectral variability of environmental noise in urban spaces, make its modelling and prediction an extremely complex and non-linear problem (Torija et al., 2012). Because machine-learning techniques have a great ability to model non-linear relations, algorithms such as artificial neural networks (ANNs) (Givargis and Karimi, 2010; Lucas de Souza and Giunta, 2011; Torija and Ruiz, 2012; Torija et al., 2012; Verrelst et al., 2012; Yilmaz and Kaynar, 2011), support vector machines (SVM) (Chuang and Lee, 2011; Thissen et al., 2004; Vapnik, 1995; Verrelst et al., 2012; Xinjun, 2010; Torija et al., 2014), and Gaussian processes for regression (GPR) (Pasolli et al., 2010; Rasmussen and Williams, 2006; Verrelst et al., 2012; Wu et al., 2012), have been successfully employed in solving regression problems. Therefore, these machine-learning techniques for regression could be used to develop urban environmental-noise-pollution models.

The machine-learning-based models learn the relationship between a set of input variables and a dependent output, in this case energy-equivalent sound-pressure levels. Consequently, the performance of this approach depends heavily on an appropriate selection of input variables, but also on the collection of a representative sample of the diverse typology of scenarios where such models would be used. Thus, under unknown conditions for the models or new noise sources, a process for updating the database or the set of input variables is needed if reliable predictions are requested. On the other hand, the machine-learning models are flexible, adaptable and able to learn complex relationships, so that, a consequent improvement of the provided database might enable their application in a great range of scenarios. Also, other advantages of machine-learning-based models for estimating urban environmental noise that could be mentioned include: (i) easy implementation once trained and validated; (ii) ability to model non-conventional road-traffic sources and noise sources other than road traffic; (iii) possibility of implementation to deal with long-term but also short-time energy-equivalent sound levels. Thus, they can be used to identify impacts (or deviations from limit noise values) on a short-term scale, and hence offer faster action for noise management in noise action plans.

The objective of the present work is to develop a procedure for accurate environmental-noise-pollution modelling with a reduced computational and data collection cost in urban environments. Firstly, with the inclusion of a previously selected set of input variables (Torija et al., 2010), and with the use of machine-learning algorithms, a series of environmental-noise-pollution models to built-up environments are developed and tested. To develop machine learning for regression models, three approaches were used: multilayer perceptron neural networks (MLP), sequential minimal optimisation (SMO) for regression, and Gaussian process for regression (GPR). The Pearson VII function-based universal kernel (PUK) has been implemented to build SMO and GPR models. Furthermore, this it was hypothesised that the environmental-noise modelling in urban environments is a non-linear problem, so that the performance of the machine learning for regression models (non-linear solvers) is compared with the one of a classical linear solver (multiple linear-regression model, MLR). Secondly, a feature selection was undertaken to reduce the complexity of the models and thereby decrease both the computational and data-collection cost, thereby offering a more easily implemented model. Two different approaches were followed: correlation-based feature-subset selection and wrapper for feature-subset selection. The performance of these feature selection methods is compared to the one of a classical data reduction technique (principal components analysis). Thus, once the confirmation of the feasibility of machine learning for regression algorithms for developing urban environmental-noise-pollution models, different schemes for selecting the input variables set were proposed, depending on the needs regarding data-collection cost, computational cost and accuracy.

2. Methods

2.1. Methods used for urban environmental-noise-pollution

As stated above, this paper proposes a procedure for developing environmental-noise-pollution models for built-up environments. For this task, the performance of machine-learning algorithms such as MLP, SMO, and GRP was assessed and compared with that of a MLR. It should be noted that, the training and test conditions of the different models built are exactly the same for all the estimation techniques implemented.

Fig. 1 shows a diagram with information on the process used for model development.

2.1.1. Multiple-linear regression

MLR is employed to predict the variance in a dependent interval, based on linear combinations of interval, dichotomous, or dummy independent variables. The general purpose of MLR is to learn more about the relationship involving several independent or predictor variables and a dependent or criterion variable (Yilmaz and Kaynar, 2011). MLR is usually based on the least-squares method, so that the model is fit in such a way that the sum of squares of differences of observed and predicted values is minimized. Therefore, the predicted value is estimated as follows:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i1} + \hat{b}_2 x_{i2} + \cdots + \hat{b}_k x_{ik} \quad (1)$$

where, $\hat{b}_i$ are the estimations of the β_i parameters or standardized coefficients (the estimates resulting from the analysis made on the independent variable that have been standardized) and $\hat{y}_i$ is the predicted value (Agirre-Basurko et al., 2006).

In this study, a MLR is used to compare the performance of nonlinear techniques against a classical linear-estimation method for estimating environmental noise in urban environments.

2.1.2. Multilayer perceptron neural networks

The most common approach to develop nonparametric and nonlinear regression is based on ANN (Haykin, 1999). An ANN is a structure composed of artificial neurons (nodes) set in layers and connected