IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 June 2024, accepted 27 June 2024, date of publication 3 July 2024, date of current version 12 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3422832

RESEARCH ARTICLE

A Variable Selection Analysis for Soundscape Emotion Modeling Using Decision Tree Regression and Modern Information Criteria

ROBERTO SAN MILLÁN-CASTILLO, LUCA MARTINO, AND EDUARDO MORGADO
Departamento de Teoría de la Señal y Comunicaciones y Sistemas Telemáticos y Computaciòn, Universidad Rey Juan Carlos (URJC), 28942 Fuenlabrada, Spain

Corresponding author: Roberto San Millán-Castillo (roberto.sanmillan@urjc.es)

This work was supported in part by Agencia Estatal de Investigación (AEI) through the project SPGRAPH under Grant PID2019-105032GB-I00; in part by MCIN/AEI/10.13039/501100011033 under Grant PID2022-136887NB-I00 (POLIGRAPH); in part by the Comunidad de Madrid and Universidad Rey Juan Carlos (Proyecto I+D Jóvenes Doctores, AUTO-BA-GRAPH) under Grant F861; and in part by the Programa de Excelencia-Convenio Plurianual entre Comunidad de Madrid y la URJC under Grant Y158/DF007003/30-06-2020 and Grant F840.

ABSTRACT During the last decade, soundscape research has become one of the most active topics in Acoustics. This work provides a nonlinear variable selection analysis over the well-known dataset 'emo-soundscapes'. Namely, we provide a selection of the soundscape indicators using a nonlinear and nonparametric model as a regression tree method. Modern techniques (proposed in the literature) have been used, first for ranking the variables and then for choosing the effective number of features. We have also compared and discussed our results with those provided previously in the literature. This study, based on modern techniques in selecting the effective number of variables, confirms the result presented in previous recent work (but based on a linear model) that very parsimonious models should be considered (in the case of a nonlinear model, it is based on very few variables, from 2 to 4, depending on the output). All the results are obtained by analyzing a single dataset. As future research works, we plan to extend our study by also considering alternative datasets.

INDEX TERMS Soundscape emotion recognition, decision tree regression, ranking methods, variable selection.

I. INTRODUCTION

In recent years, Soundscape research has shifted the paradigm in environmental acoustics, from a traditional focus on noise levels to a more comprehensive approach that includes the sound perception of community and individuals, as well as acoustic environments and contexts. Southworth originally coined the term soundscape [1] and was popularised by Schafer [2]. This topic is relevant in acoustics nowadays due to the extensive range of new and innovative applications for real and practical problems. Music and speech-elicited emotions are less subtle and more noticed than those related to soundscapes. The subjective evaluation of soundscapes relies on many factors including acoustical, visual, personal, physical, psychological, social, and even cultural estimators and their intricate interplay. Up to now, research continues to push the limits of knowledge in this field even with recent attempts of standardization [3] regarding: soundscape attributes and the different languages (e.g., in Spanish [4] but also in many others), probabilistic approaches to soundscape perception [5], the influence of well-being and demographic facts [6], or even changing the paradigms of affective data collection [7], to name a few. These challenges suggest that soundscape emotion recognition (SER) requires further research to support perception descriptors [8]. SER is a general definition from a broad signal processing point

The associate editor coordinating the review of this manuscript and approving it for publication was Li Zhang.

of view [9], [10], [11], which is identified as soundscape descriptors [12] in particular topics. Several applications, including urban sound planning [13], [14], [15], noise control [16], [17], acoustic monitoring [18], [19], sound design in films and digital games [20], [21], or sonification [22] are increasingly employing soundscapes-elicited emotions to provide advanced criteria on sound assessment.

Modelling the perception of soundscapes helps forecast human responses to different acoustic circumstances with few resources. Hence, design and planning processes can be cost-effective within practical frameworks, without the need for constant reliance on specific experiments. Moreover, modelling may provide a seamless workflow which helps the soundscape approach be a valuable tool. SER is a challenging task that requires further research to better understand the dynamics of soundscape perception, as described in [23]. Russell's circumplex model has been applied for scaling the perception of soundscapes [11], [24], [25], [26]. Russell's affect representation can be modelled with two key factors: arousal and valence, which respectively represent the eventfulness and the pleasantness of a soundscape [26]. While arousal and valence have been accepted as the principal and sufficient affective descriptors in some research [26], [27], there are various proposals to enhance the SER with additional or alternative descriptors [12], [28], [29], and even to include emotion appraisal in standardised procedures of soundscape data collection and analysis [8].

Up to now, there are no standardised estimators or features for SER. Thus, researchers have been working on an extensive range of soundscape descriptors (i.e., outputs) and soundscape indicators (i.e., variables/features). In terms of modelling algorithms, nonlinear models seem to outperform linear models, which are preferred due to their simplicity in development [30]. Most of these studies lack a commom comparison framework. In this study, we employ the Emo-soundscapes database (EMO) [11], which has been a reference in SER studies by different authors and methods, including modelling and variable selection [9], [22], [31], [32], [33], [34], [35]. Recently, other promising databases focusing on urban soundscapes are available: Athens Urban Soundscape (ATHUS) [36], in the city of Athens (Greece), from 2019; International Soundscape Database (ISD) [37] in London and Venice, from 2022; Affective Responses to Augmented Urban Soundscapes (ARAUS) [38] is a data-augmented database combining previous databases and adding maskers, from 2023. All databases show advantages and limitations. EMO provides a wider comparison framework currently, a ready-to-use dimensional problem (with up to 122 varied features), and a balanced range of soundscapes with all Schafer's categories annotated from 74 countries. The authors consider these reasons valuable to make progress in the EMO framework and our research in [31].

This article studies a variable selection approach in the SER application. This is a central task in several scientific fields such as signal processing, statistics, and machine learning, to name a few. Generally, variable selection focuses on selecting the relevant variables (i.e., features) that improve the performance of a regression or classification model. An appropiate and efficient strategy on variable selection may benefit models in several ways : (a) a decrease in features leads to less complex models; (b) the simpler the model, the lower the computational load; (c) a reduced complexity also yields more interpretable models. We divide the variable selection problem into two theoretically separate stages. Namely, it consists of two main parts: (a) firstly, ranking variables and (b) secondly, selecting the effective number of variables.

In the literature, most ranking methods are classified into 3 different classes: *filter* methods (e.g., a variance threshold in [11]), *wrapper* methods (e.g., step-wise regressions used in [31]), and finally, *intrinsic/embedded* methods. DTR provides an *embedded* method for ranking variables based on the so-called Gini importance index. Furthermore, several researchers have already performed different variable selection techniques based on EMO for regression problems. Some research chose features heuristically like in [33] with 23 features, and in [32], with 54 features. In [11], the technique was a heuristic threshold for the variance to select 39 variables out of the 122, which are available. Principal components analysis (PCA) and univariate linear regression were combined to select the most relevant variables in the best models in [9], resulting in 26 and 29 variables for valence and arousal respectively. Recursive feature elimination and hyperparameter grid search for random forests in [22] provided the best models, including 15 and 14 features for arousal and valence, respectively. Finally, recent research on classification with EMO, has presented different variable selection approaches (i.e., PCA, recursive feature elimination, forward feature elimination, random forests) within five machine learning classifiers, which generally improve their performance. However, the number of variables is not reported for each experiment [35]. Research on the variable selection process with other databases is available in the literature. In [39], the authors performed a backwards-step feature selection in an urban traffic noise dataset. This same technique was included in [40] applying also the Akaike information criterion (AIC) as the performance criterion within ISD dataset over 11 original variables.

In this work, our goal is to conduct a variable selection study with recent modern techniques - comparing with classical ones, as cross-validation (CV) - while considering a nonlinear, nonparametric flexible model. Note that, in this sense, the use of a *decision tree regressor* (DTR) method seems a perfect fit: it is a nonlinear and nonparametric model that also provides an intrinsically embedded method for ranking the features, which can be used for comparison with other ranking methods. In previous work on the EMO database [31], variable selection was conducted by considering a linear model (for arousal and valence). Unlike in [31], this research we use an *intrinsic*

(called also *embedded*) ranking approach based on the Gini importance index provided directly by DTR. Other classes of ranking methods are generally used in the literature, such as the *filter* and *wrapper* ranking methods [9], [11], [32]. For model selection purposes (i.e., for finding the effective number of variables), we employ recent techniques such as the Spectral Information Criterion (SIC) and the Universal Automatic Elbow Detector (UAED) [41], [42]. Both can be applied in more general scenarios than other information criteria in the literature, as the well-known BIC [43] and AIC [44] (see Table 9). Moreover, it has been proved that SIC and UAED exhibit more robust behaviour when tested in various scenarios and applications [41], [42]. See also Appendix for further details. Generally, resampling methods such as cross-validation are often employed for variable selection methods, but the split of data into training and test sets significantly influences performance. Alternatively, information criteria make use of the entire dataset (e.g., BIC and AIC) but require probabilistic assumptions. Modern techniques based on elbow detection (e.g., UAED and SIC) are equivalent to information criteria but present evidence of being more general and robust.

As the final outcome of our analysis in this work, we have determined that if the regression model is sufficiently flexible (nonlinear and nonparametric) we suggest the use of only 2 variables for the output ''arousal'' and only 4 variables for the output ''valence''. Therefore, considering only the EMO database, we have confirmed the conjecture shown in [31] with a linear model: regarding the soundscape emotion recognition (SER) more parsimonious models can be considered, mainly based on psychoacoustic features, such as ''Loudness Mean'' and ''Loudness Standard Deviation''.

The highlights of the work are the following:

- Evaluation of the performance of DTR models in SER as a competitive alternative in terms of parsimony and interpretability to linear regression, and nonlinear methods used in the literature such as Random Forests, Support Vector Machines and Artificial Neural Network strategies.
- Analysis of a variable selection framework based on DTR, using the Gini Importance index as a ranking strategy.
- Analysis of the influence of cross-validation settings on the performance, the complexity of the model and the selection of an optimal subset of features.
- Use of two innovative elbow detection methods to decide the optimal variable subset in SER, the Spectral Information Criterion (SIC) and the Universal Automatic Elbow Detection (UAED).

The rest of the work is organised as follows. Section II describes the database that was employed in this study. Section III presents some background material on DTR and describes a procedure for tuning a crucial hyperparameter of DTR. Section IV is devoted to the description of the employed ranking method. Then, Section V introduces different techniques for choosing the effective number of variables. Section VI provides a detailed discussion about the obtained results and we compare them with other results given in the literature. Finally, Section VII draws some conclusions.

II. DATABASE

This study works with EMO, a large, extensively benchmarked, and publicly available database of soundscapes with annotations of emotion labels [11], which is extremely convenient for variable selection studies due to the dimensional problem that it provides (with 122 assorted and meaningful features). EMO consists of 1213 audio files under a Creative Commons license, which were created by material included in the Freesound collaborative platform [45]. The EMO's files are classified into six categories according to Schafer's taxonomy, which identifies both the sound source and the listening context [2]: human sounds (e.g. laugh), sounds and society (e.g. party), mechanical sounds (e.g. engine), natural sounds (e.g. birds), quiet and silence (e.g. quiet park), and sounds as indicators (e.g. church bells). Schafer's classification also has fewer categories than Brown's taxonomy [16]. Given the above, Schafer's taxonomy seems to provide a general and straightforward approach for practical cases [16]. EMO contains 100 audio clips per category in a first subset (i.e. 600 audio clips) and, additionally, 613 manually mixed audio clips of two or three categories from the first subset. A crowd-sourcing procedure provides the perceived soundscape emotions (i.e., arousal and valence) by 1182 trusted annotators, who performed a ranking-based questionnaire of two clips pairwise comparison with adequate inter-subject reliability

EMO's files are monophonic audio clips with a 44100 Hz sample rate. According to [46], monophonic audio seems to be adequate for assessing eventfulness (i.e., arousal) and pleasantness (i.e., valence) of acoustic environments, among many other soundscapes descriptors. EMO provides up to 122 normalised variables for each output, that are extracted from every audio file, with a 50% overlapping Hanning window (23 ms wide), through general use tools such as YAAFE [47] and MIRToolbox [48].

EMO's variables of the audio signals can be classified into three main categories as follows:

- Time-domain variables: These represent the signal dynamics, such as classical estimators based on samples of the audio signal (i.e., energy, entropy of energy, root mean square (RMS), or zero-crossing). They are indexed from 1 to 7, and 22, 23, 52, 115, 116. The index of every feature represents their order number in EMO. Hence, the features may also be referred to in a condensed way.
- Psychoacoustic variables: These represent perceptual (i.e., subjective) attributes of sounds such as level (i.e., loudness for overall level and MFCC for band-limited levels). They are indexed 4, from 24 to 49, and 113, 114, 117, 118, 119.

- Frequency-domain variables: These represent the shape of the spectrum and the harmonic structure of sounds such as the proportion of frequencies that are not multiple of the fundamental frequency (i.e., inharmonicity). They are represented by the remainder indexes, i.e., the rest of the variables.

EMO's complete data and further information, are publicly available at https://metacreation.net/emo-soundscapes/.

Additional Considerations: It is important to remark that the EMO database suffers from some drawbacks, as all databases do. In literature, there are different approaches to consider representative soundscape excerpts, to name some of them, from a few seconds [10], [11], [49], to the general trend with 30 s [25], [40], [50], and also longer duration such as 3 m in the ISO standard [3], [46], [51], which seems to be more reliable than shorter clips. Even, some authors claim that around 40 min of exposure to soundscape can influence the perception of some soundscape dimensions [52]. Nevertheless, the procedure of soundscape data collection should be a trade-off between the representative audio file and the global survey. Too long surveys may lead to a deterioration of response quality [53], [54], [55]. Breaking points lay between 10 min and 20 min, which means that soundscape clips should be a few and/or short. When it comes to the soundscape predictive framework, data quality and quantity are paramount. Hence, the data collection procedure needs to balance these critical variables depending on the application and the database.

Perhaps the most adequate data is available via on-site data collection, but these surveys are resource-intensive and require a detailed design, which results in a limited range of data(i.e., locations, contexts, sources, and participants). Audio signal acquisition equipment should be unnoticeable to avoid any distortion of the acoustic context. Hence, the exposure level of participants is the closest to calibrating further processing stages for feature engineering (e.g., ISD dataset). Conversely, laboratory experiments (e.g., ARAUS) and online surveys (e.g., EMO) provide vast data with limited resources.

Calibration is mostly convenient to provide robust validity on some signal features although sound level adjustment of soundscape reproductions produces more ecologically valid results in certain conditions than on-site actual level according to [56].

Other subsequent soundscape datasets have attempted to address these drawbacks, but they also show their limitations. ISD [37] offers less than 1000 valid in situ recordings and questionnaires collected in 13 urban spaces only in London and Venice. ISD provides the eight dimensions of the Perceived Affective Quality (PAQ) model [25] which is interesting for multi-output models. ARAUS [38] provides also PAQ outputs for more participants, up to 127 different urban soundscape clips, split into two halves and augmented with 280 monophonic maskers, but in a laboratory experiment. Both of them provided a limited range of features, soundscape types and cultural origins of participants in comparison with EMO.

III. DECISION TREE REGRESSION (DTR) MODEL

DTR is a nonlinear and nonparametric supervised learning method, that works as follows. Let us denote as $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ the N pairs of data, where $\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,M}]^\top \in \mathbb{R}^M$ represents an input vector and y_i denotes a scalar output. The idea in DTR is to divide the input space $\mathbb{R}^M$ using a partition formed by hyper-rectangular regions $\mathcal{R}_j$ with $j = 1, \ldots, J$. The value of J is related to the hyper-parameter "MaxDepth" of DTR (i.e., the depth of the tree). Note that the regions $\mathcal{R}_j$, since they are hyper-rectangular, can be easily defined by inequalities over the components of the input $x_{t,j}$. In each region $\mathcal{R}_j$, the regression value $\widehat{y}_j$ is given by the arithmetic mean of the outputs of the samples *inside* $\mathcal{R}_j$. Denoting as $\mathcal{K}_t$ the set of indices of the samples contained inside $\mathcal{R}_j$,

$$\mathcal{K}_t = \{t \in \{1, \ldots, N\} : \text{ such that } \quad \mathbf{x}_t \in \mathcal{R}_j\},$$

for all $j = 1, \ldots, J$, then we have

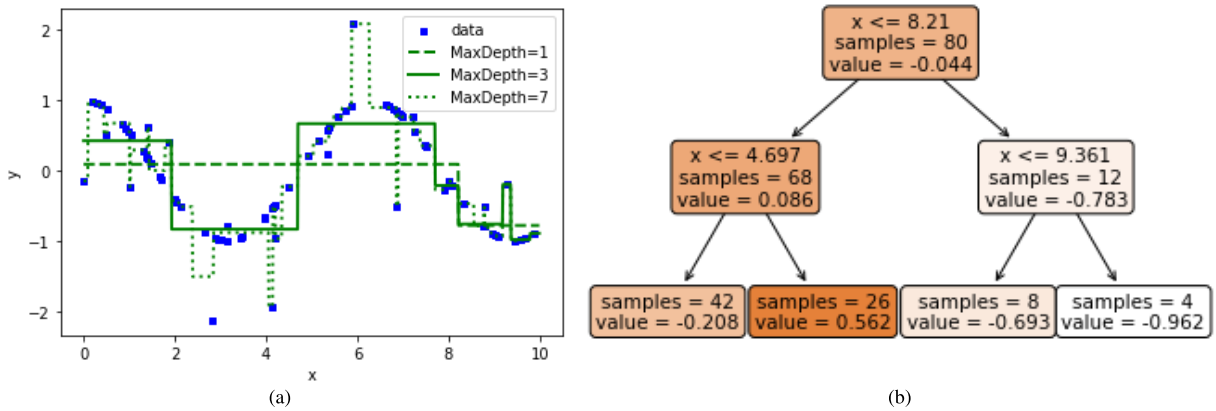
$$\widehat{y}_j = \frac{1}{n_j} \sum_{t \in \mathcal{K}_t} y_k, \tag{1}$$

where n_j are the number of samples inside $\mathcal{R}_j$, i.e., $n_j = |\mathcal{R}_j|$ and $\sum_{j=1}^{J} n_j = N$. Due to the partition of the space and that in each region the regression function takes a constant value $\widehat{y}_j$, the resulting regression functions based on DTR are piecewise constant functions. Graphical examples are given in Figure 1 (for a uni-dimensional input space, i.e., $M = 1$). The goal of DTR is to search the optimal partition regions $\mathcal{R}_j$ that minimizes the Mean Square Error (MSE) considering the overall dataset.

DTR presents well-known advantages: no assumptions are needed for the underlying distribution of the data; graphical and easy interpretation is available (see Figure 1); it is robust to noise, outliers and missing data, thus, affordable data pre-processing. Another relevant benefit of DTR is also that it provides the Gini importance index (GI), as the variable importance criterion, which calculates the times a variable is used to split a tree node (weighted by the number of samples of the node) for each component $x_{t,j}$ of the input vector that measures the importance of the specific feature [57], [58]. Our analysis is based on the DTR model. More specifically, to obtain reliable results, we have employed the code provided by scikit-learn library in Python programming language [59]. On the other hand, DTR needs a good tuning of hyperparameters to control the overfitting and increase the ability to provide good results in other (but related) datasets.

A. SETTING SOME PARAMETER OF THE TREES: CHOICE OF MAXDEPTH

DTR deals with a wide range of hyperparameters for tuning the models' predictions. For the sake of simplicity, this work only varies one of the most relevant hyperparameters: the



(a) (b)

FIGURE 1. **(a) Examples of regression tree functions. The pairs of data $\{x_i, y_i\}$ are shown in blue squares and green lines. In this example, the inputs are scalar values, i.e., $x_i \in \mathbb{R}$ ($M = 1$). We depict 3 different regression tree curves (that are piecewise constant functions, as expected), each one corresponding to a different hyperparameter "MaxDepth" value. We can observe that bigger values of "MaxDepth" foster the overfitting. (b) The tree corresponds to the solid line in figure (a). With "samples", we denote the number of samples n_j in each interval/region, and with "value", we denote the mean value of the outputs $\hat{y}_j$ in that region.**

maximum number of splits for a sample (*MaxDepth*). To limit model complexity and adequate interpretability of outcomes, *MaxDepth* is the key hyperparameter within DTR.[1]

We apply a CV procedure with $80\% - 20\%$ (training-test) and average the results in terms of MSE ($CV - MSE$) over 10^4 independent runs. Figure 2 depicts the performance of the DTR model with a classical cross-validation approach ($CV - MSE$), versus the MaxDepth for both outputs, arousal and valence. Both curves reach a minimum value of MaxDepth. The arousal DTR model with CV provides a minimum CV MSE of 0.053, and the valence DTR model presents its best CV MSE of 0.151 (both values considering all the 122 variables), with a MaxDepth value of 5 and 4 respectively. Therefore, for the sake of simplicity, in most of the rest of the work, we decided to use $MaxDepth = 4$. More precisely, we employ $MaxDepth = 4$ in all the sections except for this section, where we average the results obtained with different *MaxDepth*. Figure 3 presents a flow chart of the variable selection framework proposed in this work. The choice of the principal DTR hyperparameter is the first step. Then, the features are ranked according to their Gini importance index. After that, we perform DTR models adding one variable on each stage, according to the ranking, and starting with the more important feature. Finally, we conduct variable selection with various elbow detection methods (UAED and SIC), and cross-validation (with different training-test data sizes). Details on each part will follow in the next sections.

IV. RANKING OF THE VARIABLES BY GINI IMPORTANCE INDEX

In this section, we rank the 122 variables/features using the Gini importance index (GI) provided within the DTR model. We provide a brief explanation of GI below:

[1]The remainder of the DTR hyperparameters are: the best split considers all the input variables, the minimum number of required samples to split a node is two, and leaf nodes are unlimited and are correct with only one sample.

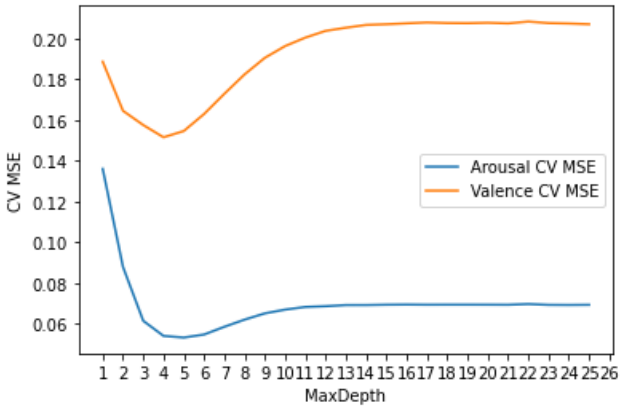


FIGURE 2. ***CV − MSE*** **of a 122-variables DTR model (Arousal + Valence) vs MaxDepth.**

Gini Importance Index: Regression tree algorithms deploy the method of the Gini importance index to originate binary splits in the constructed tree. In each split in the tree, the GI measures how the corresponding feature affects the prediction error (in case of regression). The GI calculates the times a variable is used to split a tree node (weighted by the number of samples of the node) for each component $x_{t,j}$ of the input vector, which measures the importance of the specific feature. For further technical details, see [57] and [58].

Here, we consider all the data together (without using any CV) and consider different values of $MaxDepth \in \{1, 2 \ldots, 25\}$. The results are then averaged over 10^4 independent runs.

Table 1 and Table 2 show the first 8 most important variables (sorted in decreasing order of GI) as a result of this average. Table 1 refers to the output "Arousal" whereas Table 2 refers to the output "Valence". Note that *perception features* based on "Loudness" become the most significant by far both for arousal and valence. This is in agreement with the reported analysis with a linear regression model in [31],

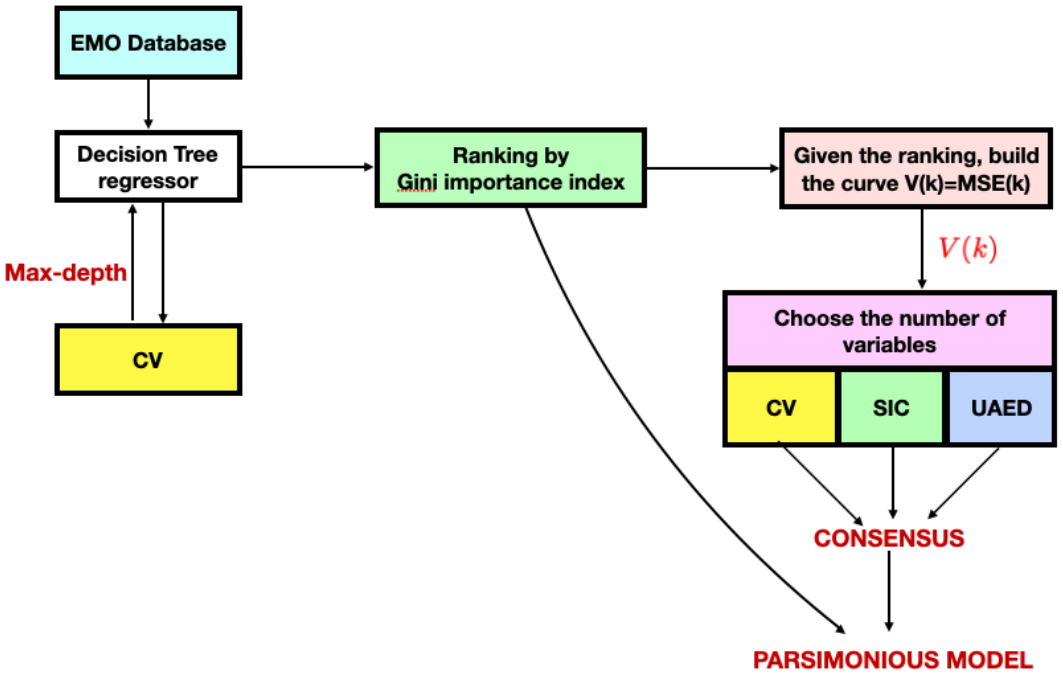


FIGURE 3. General flow chart of the variable selection framework that we propose in this work.

TABLE 1. GI of the 8 most relevant variables for the training set of a 122-variables DTR model for arousal. Identification of the variables by the name and the position in EMO. M: Mean, Std: Standard Deviation.

Variable	Loudness M (113)	Fluct. max (4)	RMS M (1)	Energy M (115)	Zero-cross. M (6)	MFCC M 8 (31)	RMS Std (2)	MFCC Std 9 (45)
GI	0.783	0.052	0.030	0.023	0.009	0.004	0.004	0.004

TABLE 2. GI of the 8 most relevant variables for the training set of a 122-variables DTR model for valence. Identification of the variables by their name and their position in EMO. M: Mean, Std: Standard Deviation.

Variable	Loudness Std (114)	Chromag. Std 2 (102)	Entropy Std (23)	Chromag. M 12 (68)	MFCC M 5 (28)	Decrease M (3)	Loudness M (113)	Kurtosis Std (19)
GI	0.620	0.030	0.028	0.020	0.017	0.013	0.011	0.0

the arousal DTR model seems to require fewer variables than the valence one according to the GI of the 8 most relevant variables. The summation of the GI of 5 first variables in the arousal model results in 0.98, while in the valence model, the first 10 variables only reach 0.71. This suggests the need to use more features for modelling valence for arousal, as also reported by other works in the literature [9], [31].

V. CHOOSING THE NUMBER OF VARIABLES

A. ELBOW DETECTION METHODS

In this section, we apply an elbow detection approach that allows us to avoid applying again CV. However, later we use a CV approach, in order to compare and check the results. First of all, we need to construct a non-increasing MSE curve (considering all the data - without CV), adding progressively variables (in decreasing order of importance) considering the ranking obtained by the GI above. This curve is denoted as

$$V(k) = \mathrm{MSE}(k), \quad k = 0, \ldots, 122,$$

where $V(0)$ is the power of the signal. We build the non-increasing MSE curve $V(k)$ for arousal and another one for valence. They are shown in Figure 4. Thus, we apply the spectral information criterion (SIC) method [41] and the universal automatic elbow detector (UAED) technique [42]. See Appendix for more information about SIC and UAED. The SIC method includes the main information criteria in the literature as special cases [43], [44]. Both SIC and UAED, can be applied to any non-increasing function $V(k)$ (unlike BIC and AIC). See Table 9 for further details. Moreover, SIC and UAED have obtained very good results in several applications and numerical experiments [41], [42].

We apply SIC with $2 \cdot 10^6$ particles for the internal Monte Carlo approximation. SIC returns the set of all possible models (number of variables),

$$\mathcal{E}_{\texttt{arousal}} = \{1, 2, 3, 6, 9\},$$
$$\mathcal{E}_{\texttt{valence}} = \{1, 2, 3, 4, 6, 7, 10, 11, 16\},$$

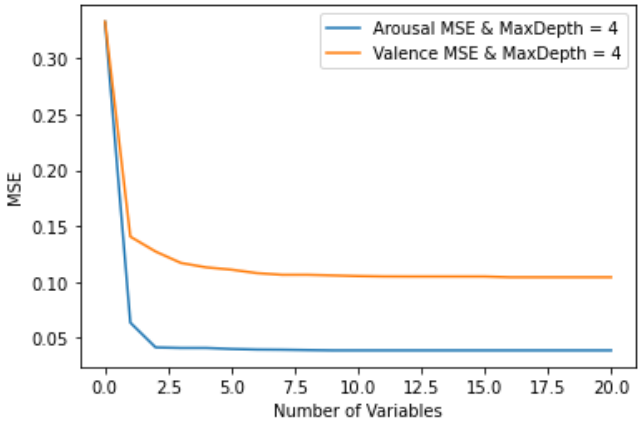


FIGURE 4. The non-increasing curves $V(k) = \text{MSE}(k)$ versus number of variables k with $MaxDepth = 4$, for arousal and valence.

i.e., only 5 models are possible for arousal and only 9 models are possible for valence, following SIC. The final model suggestions of SIC are:

(considering the 95% and 100% of probability - see [41] and the Appendix in this work)

For arousal:

SIC-95 $\longrightarrow$ 2 variables, SIC-100 $\longrightarrow$ 9 variables.

For valence:

SIC-95 $\longrightarrow$ 3 variables, SIC-100$\longrightarrow$ 16 variables.

Moreover, we apply the UAED in [42] to the $V(k)$ curves in Figure 4, and this method suggests 2 variables for both, arousal and valence. Therefore, there seems to be an almost perfect agreement between the SIC-95 and UAED suggestions. Note also that the derivative of the decrease of the MSE is similar to the derivative of the decrease of the GI.

B. SELECTING THE NUMBER OF VARIABLES WITH CV

Many authors also use a CV procedure for selecting the number of variables in a variable selection problem and, more generally, for DTR [60], [61], [62]. Thus, we also apply a CV approach to compare with the results obtained in the previous section. However, in advance, we can assert, since SIC and UAED do not require any split of the data (and hence no averaging) they are both much less costly than the CV procedure.

Clearly, the proportion of data for training and testing affects the results. We modified CV rates to observe any changes both in performance and variable selection terms. The experiments considered test sizes (TS) of 50%, 30%, 20%, and 10% of the available data. We average the results over $T_{iter} = 10^4$ independent runs (we keep MaxDepth $= 4$, as explained in Section III-A).

We employ the Gini ranking described in Section IV, in order to add new variables progressively, exactly as done with the elbow detectors in Section V-A above and $MaxDepth = 4$. Figures 5(a), 5(b), depict the curves MSE(k) for the two outputs, arousal and valence. That, in this case, has a minimum due to the use of the CV. These curves confirm that the results change depending on the CV rate. They are given in Tables 3-4.

TABLE 3. Choice of the number of variables searching for the minima of the CV MSE curves in Figure 5(a), with different CV procedure (using the Gini ranking), for arousal.

CV-50/50: 50% training — 50% test	2 variables
CV-70/30: 70% training — 30% test	2 variables
CV-80/20: 80% training — 20% test	2 variables
CV-90/10: 90% training — 10% test	9 variables

TABLE 4. Choice of the number of variables searching for the minima of the CV MSE curves in Figure 5(b), with different CV procedure (using the Gini ranking), for valence.

CV-50/50: 50% training — 50% test	4 variables
CV-70/30: 70% training — 30% test	4 variables
CV-80/20: 80% training — 20% test	5 variables
CV-90/10: 90% training — 10% test	5 variables

VI. DISCUSSION AND FINAL CHOICES

First of all, we can observe that SIC and the CV procedures suggest using more variables for valence (than for arousal), confirming the results in other papers that consider other models [9], [22], [31]. Below, we discuss in detail, and separately, the cases of arousal and valence.

A. MAIN VARIABLES FOR AROUSAL

For arousal, it seems there is an agreement among SIC-95, UAED and the CV procedures (CV-50/50, CV-70/30 and CV-80/20) in order to choose only 2 variables. Only CV-90/10 suggests 9 variables which is also the suggestion of SIC-100. The 6 most important variables following the Gini ranking are (in decreasing order of importance by the Gini index),

$$113, 4, 1, 115, 6, 31,$$

We can observe an agreement with 3 of the six variables selected in [31] with the linear model (in decreasing order of importance according to [31]),

$$113, 114, 14, 115, 4, 8, 56,$$

which are

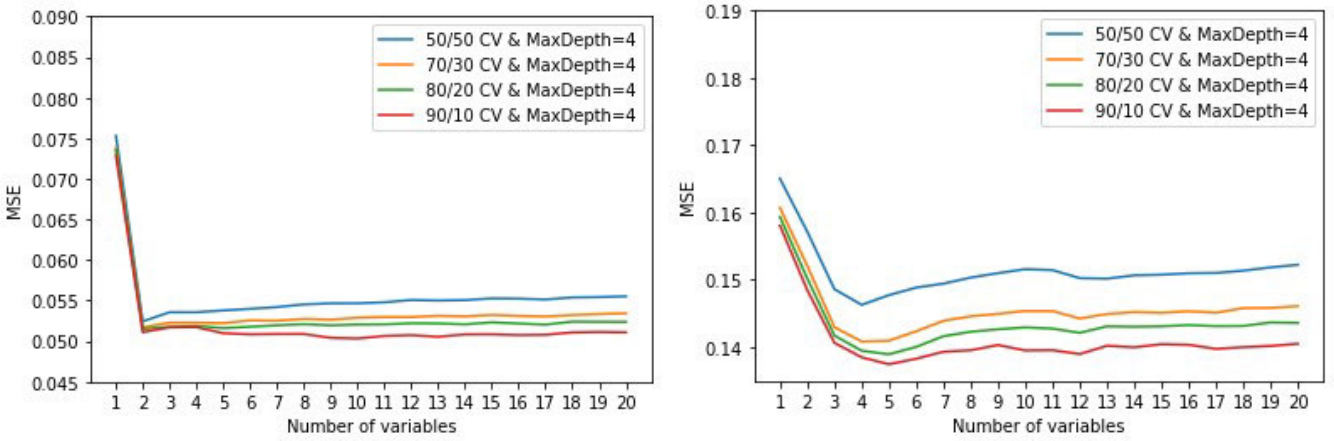
- Loudness mean (113),
- Fluctuation max (4), and
- Energy mean (115).

Here, since the agreement among SIC-95, UAED and CV-50/50, CV-70/30 and CV-80/20, we suggest the use of a DTR model with only two variables: Loudness mean (113) and Fluctuation max (4). Clearly, a nonlinear nonparametric model as a regression tree requires fewer variables than a linear model (e.g., the 6 variables in [31]). We have included the DTR model model for arousal in *supplementary material.*

B. MAIN VARIABLES FOR VALENCE

For valence, the summary of the suggestions is given below:

- UAED: 2 variables,
- SIC-95: 3 variables,



(a) Arousal (b) Valence

FIGURE 5. CV MSE versus the number of variables for different CV rates with *MaxDepth* = 4, (a) for arousal; (b) for valence.

- CV-50/50 and CV-70/30: 4 variables,
- CV-80/20 and CV-90/10: 5 variables,
- SIC-100: 16 variables,

The study in [31], which employs a linear model, suggests the use of 16 variables (note that is the same number suggested by SIC-100 in this work), divided into in 6 *very relevant* variables and 10 *relevant* variables. Comparing 8 variables in the Gini ranking in Table 2, and the 16 variables in [31], we have an agreement with 4 variables

- Loudness mean (114),
- Loudness standard deviation (113),
- Decrease Slope mean (3), and
- Energy mean (115).

According to the list of suggestions, here we consider using a DTR model with only 4 variables (the first of the Gini ranking above). The application of a nonparametric nonlinear model allows the use of more parsimonious models to a linear one [31].

C. FINAL COMMENTS

The suggestions in this work are in line with the results in [31] where a simple linear model was considered. Since the DTR model is nonlinear, this justifies the use of a smaller number of features with respect to [31]. The need for more variables for valence is also shown in [9] and [22]. Tables 5 and 6, below, summarize the results in other works considering different methods and show the DTR model proposal developed in this research, highlighted in bold. Table 5 presents the prediction model, the number of variables for the two outputs, and the more relevant variables, reported in the literature. Table 6 shows the applied methods for ranking. Furthermore, the number and names of the chosen variables with the corresponding research are given in Table 5. Note that, in some previous works, the names of selected variables are not provided.

We can observe in Table 5 that we suggest much more parsimonious models compared to the other works in the literature (but in line with [31]). Our suggestion is supported by the small increase in the MSE that we obtain (considering all the sample data, *MaxDepth* = 4, see Table 7). Namely, our suggested very parsimonious models increase the MSE by just +6% in arousal and +4.6% in valence regarding the final model with only 2 and 4 features, respectively. The figures in Table 7 show that the training stage performs reasonably with few variables in comparison with *all variables* models. This indicates that these variables are relevant and confirms the parsimonious choices of UAED, SIC and CV procedures. From Table 5, we can observe that psychoacoustic variables become relevant in previous work with EMO, such as MFCCs, Roughness, Fluctuation Strength, and Loudness. Also, predictive models based on recent and powerful soundscape databases underline the key role of psychoacoustic features. ARAUS showed that the maximum Loudness was the most important variable for ISO Pleasantness (i.e., the equivalent to valence in this work) [38], and then frequency-domain variables. On the other hand, ISD proposed a model for ISO Pleasantness with Loudness (fifth percentile) as an extremely important variable, along with time-domain and frequency-domain variables (i.e., based on L_{Aeq}, L_{Ceq}, and sound pressure level percentiles). Moreover, ISO Eventfulness (i.e., the equivalent to arousal in this work), was modelled considering other psychoacoustic variables (e.g., Fluctuation strength, Roughness). All these ideas are in line with the conclusions in a recent review on soundscape modelling [30]. The soundscape models presented in this study result in a similar, and parsimonious, set of features that have been used in the literature with other methods and datasets. Our proposal highlights psychoacoustic variables as central (i.e., Loudness), and then, a fine-tuning with other frequency-domain and time-domain features.

TABLE 5. Summary of current methods and models employed in the literature with EMO, along with the number of variables and the main ones. A: Arousal, V: Valence; M: Mean, Std: Standard Deviation.

Work	Name of method	Number of variables		Main variable/s (name)	
		A	V	A	V
[11]	Support Vector Regression	39	39	—	—
[32]	Several Deep Learning methods	54	54	—	—
[9]	Random Forest and other methods	26	29	RMS Mean, RMS Std, Fluctuation Max	
[22]	Random Forest	15	14	Roughness M, Flux M	Roughness M, RMS M
[33]	Convolutional Neuronal Networks	23	23	∼ MFCCs	
[31]	Multiple Linear Regression	7	16	Loudness M, Loudness Std	Loudness Std, Loudness M
Here	**Decision Tree Regression**	**2**	**4**	**Loudness M, Fluctuation Max**	**Loudness Std, Chromagram**

TABLE 6. Summary of current ranking and variable selection methods employed in the literature with EMO.

Work	Method for ranking	Choice of number of variables
[11]	Variance	Heuristic Threshold
[32]	No Ranking available	Heuristic
[9]	Principal Components Analysis (PCA)	Univariate Linear Regression Test (KBest)
[22]	Recursive Feature Elimination	Grid Search (wrapper method)
[33]	No Ranking	Heuristic
[31]	Mainly Forward Selection, but also other ones	Heuristic choice driven by Gibbs analysis
Here	**Gini Importance**	**SIC – UAED - CV procedure**

TABLE 7. MSE in smoothing for arousal and valence (all data).

Output	**All variables (122)**	**2 Variables**	**3 variables**	**8 variables**
Arousal	0.0387	0.0410 +6.0%	0.0406 +4.9%	0.0388 +0.25%
Output	**All variables (122)**	**3 Variables**	**4 variables**	**16 variables**
Valence	0.1035	0.1170 +13.0%	0.1083 +4.6%	0.1042 +0.6%

TABLE 8. CV MSE summary of DTR models for arousal and valence, with CV-80/20).

Output	**All variables (122)**	**2 Variables**	**3 variables**	**8 variables**
Arousal	0.0530	**0.0515** -2.83%	0.0517 -2.45%	0.0520 -1.88%
Output	**All variables (122)**	**3 Variables**	**4 variables**	**16 variables**
Valence	0.1510	0.1473 -2.45%	**0.1394** -7.6%	0.14330 -5.09%

It is also interesting to observe that the MSEs obtained with the linear models suggested in [31] were 0.0432 for arousal (using 7 variables) and 0.1182 for valence (using 16 variables). Hence, we can observe that linear models with much fewer variables (than 122) are not so far away from 0.0387 and 0.1035 obtained with 122 variables and a nonlinear, nonparametric method as a DTR model.

Important Consideration: The difference between the results in this work with respect to the other results in the literature, depends on three factors (used in the analysis):

- choice of regression model,
- choice of ranking method,
- choice of "elbow detection" method, i.e., a procedure for choosing the number of variables (or a stopping rule).

In our opinion, in this work, we have used more adequate methodologies: first of all, a non-parametric nonlinear regression model, an embedded method given by the regression model for the ranking, and modern procedures (SIC and UAEC) for choosing the number of variables. Furthermore, we also test the results given by these modern procedures using the classical and robust cross-validation (CV) obtaining very similar results in terms of number of variables (2 features for arousal, and 4 or 5 features for valence). Therefore, the classical CV procedure supports the results given by SIC and UAED. Namely, we have applied SIC, UAED and CV arriving at a consensus. The other "elbow detection" methods (or stopping rules) employed in the literature, that are based on p-values, or other information criteria (such as AIC and/or BIc) often tend to provide more complex models (i.e., to overfit, in our opinion).

Table 8 summarizes the performance of different DTR models within a general approach of CV-80/20. Other test size results can be observed in Figure 5. The Gini ranking and the applied variable selection criteria point out the more convenient models. Hence, the 2-feature model and the 4-feature model for arousal and valence, respectively, present competitive CV MSEs. Linear models proposed in [31] also decrease their performance in a CV framework resulting in MSEs of 0.450 (using 7 variables)

and 0.1233 (using 16 variables), for arousal and valence respectively. CV MSEs are worse than training MSEs as expected due to the overfitting trend of DTR [60].

Finally, some research on Random Forest in [22] shows CV MSE of 0.055 (training MSE = 0.0106) for arousal and 0.1367 (training MSE = 0.0283) for valence, using the six Schafer's categories of EMO. The training MSEs may indicate an overfitting model. These results were achieved by 50 estimators (i.e. number of trees in the forest) and MaxDepth values were 20 for arousal (15 features) and 10 (14 features) for valence. Therefore, the reported Random Forest presents more complexity (i.e. less interpretability) and higher computational load than our parsimonious DTR models, with similar performance metrics. Other authors using even more complex nonlinear models, such as convolutional neural networks with windowing augmented data, obtain a CV MSE of 0.035 for arousal, and a CV MSE of 0.078 for valence (using 54 variables) [32]. Again, these values, 0.035 and 0.078, are competitive to 0.051 (with just 2 features, for arousal) and 0.139 (with just 4 features, for valence) considering a trade-off between features and performance. The rest of literature proposals (summarised in Table 5) present similar metrics to our DTR models but also more features, ranging from 14 to 54. All these considerations support the idea that just a few variables are relevant in this soundscape emotion dataset.

VII. CONCLUSION

In the last decade, soundscape research has become one of the most active topics in Acoustics. The related research grows exponentially in this field, requiring intensive and time-consuming resources, such as considerable locations and surveys.

The use of parsimonious models, designed by applying variable selection techniques, can drive simpler and clearer strategies to employ soundscape benefits in practical problems. Providing parsimonious models helps in a range of soundscape applications. In urban planning, reducing the overall cost of the decided public-political actions needs straight decision rules, providing just fewer variables makes it easier. Noise monitoring and sonification production may also improve their performance with the implementation of simpler SER models, which may efficiently shorten delays. Generally, the selection of key features lets further studies and research focus on a more particular direction, and hence, saving research time and resources. From a technical point of view, variable/feature selection also allows to work in low dimensional spaces. This helps understand and decrease the computational cost of future research analyses on the field. Hence, this simplification may help efficiently develop IoT or Tiny ML systems in real time and save energy and storage resources. All these points may lead to a faster progression in research.

By the analysis in this work, we have confirmed that much more parsimonious models can be considered regarding the SER, as also confirmed in [31]. In this study, we suggest the use of just 2 variables for the output "arousal" and just 4 variables for the output "valence". In order to support these suggestions, we have considered a ranking based on the Gini Importance index provided directly (and embedded) by the DTR model, and we have applied both (a) classical techniques as CV and (b) modern techniques as SIC and UAED methods for deciding the final number of variables. These results are in line with previous studies with EMO (in terms of selected features) and some works with other datasets and different methods. Hence, psychoacoustic features seem to become crucial in the presented models for arousal and valence, like in other works. Here we suggest employing a much reduced set of features, but obtaining a competitive model's performance (*MSE*), in comparison to previous works. Including some time-domain and frequency-domain variables led to more accurate models.

Another interesting aspect to remark is that the optimal MaxDepth parameter (employing all the 122 input variables) that we have found is 4, which is quite small. Hence, this shows a quite simple regression functions (nonlinear, but relatively simple regressors). This observation also confirms the idea of the possible use of parsimonious models. It is important to remark that all the obtained results in this work depend on the choice of the EMO database. Although it has been employed in many different works, it presents some limitations that can affect to the generalisation of our conclusion.

APPENDIX
SPECTRAL INFORMATION CRITERION (SIC) AND UNIVERSAL AUTOMATIC ELBOW DETECTOR (UAED)

Let us consider an error function $V(k)$ as a function of the dimension/complexity of the model, for $k = 0, \ldots, K$. For instance, two examples are $V(k) = \text{MSE}(k)$ or when likelihood function is available, $V(k) = -2\log \ell_{\max}(k)$ where $\ell_{\max}(k)$ is maximum likelihood value.

Generally, $V(k)$ is a non-increasing or decreasing function. Hence, to force to have a minimum value, a penalization of the model complexity is required. The idea used in the so-called *information criteria* is to employ a linear penalization on k [63], [64], [65]. Therefore, a cost function is constructed as follows:

$$C(k,\lambda) = \underbrace{V(k)}_{\text{fitting}} + \underbrace{\lambda k}_{\text{penalization}}, \qquad k = 0, \ldots, K, \quad (2)$$

where $\lambda > 0$. The first term is a fitting one, whereas the second term is a linear penalty for the model complexity. Many information criteria set $V(k) = -2\log \ell_{\max}(k)$ and differs for the choice of the penalty slope λ [63], [64]. A more recent information criterion, the universal automatic elbow detector (UAED) [42], based on geometric considerations, can be used with any generic function $V(k)$. Moreover, the spectral information criterion (SIC) encompasses all the previous ones (in some sense that we explain later on)

TABLE 9. Summary of different information criteria. Many of them are contained as special cases in SIC, with the proper choice of $V(k) = -2\log \ell_{\max}(k)$; N represents the total number of data points.

Information criterion	Choice of λ	$V(k)$
Bayesian information criterion (BIC) [43]	$\log N$	$-2\log \ell_{\max}(k)$
Akaike information criterion (AIC) [44]	2	$-2\log \ell_{\max}(k)$
Hannan-Quinn information criterion (HQIC) [66]	$\log(\log(N))$	$-2\log \ell_{\max}(k)$
Universal automatic elbow detector (UAED) [67]	$\frac{V(0)}{\min[\arg\min V(k)]}$	any
Spectral information criterion (SIC) [41]	any $\lambda \in [0, \lambda_{\max}]$	any

[41]. Table 9) summarizes different information criteria in literature [41], [63], [64].

Spectral Information Criterion (SIC): In SIC [41], the fitting term $V(k)$ can be any non-increasing function. Moreover, SIC obtains the distribution of minima of the cost function $C(k, \lambda)$ as λ varies in the interval $[0, \lambda_{\max}]$, where $\lambda_{\max}$ is defined as $\lambda_{\max} = \{\min \lambda : \arg\min_k C(k, \lambda) = 0\}$, i.e., is the minimum value of λ which provides the strongest possible model penalization (choosing a model with zero variables all the features are irrelevant for predicting the output y). The value of $\lambda_{\max}$ can be analytically obtained as

$$\lambda_{\max} = \max_k \left[\frac{V(0) - V(k)}{k}\right], \quad \text{for } k = 1, \ldots, K. \tag{3}$$

Since above we consider$k = 1, \ldots, K$, we can perform an exhaustive search and obtain $\lambda_{\max}$ from Eq. (3). The SIC approach is inspired by the idea of "integrating out" λ, i.e., to remove the dependence of λ and hence avoid picking a specific value of λ. To each value of k', SIC associates an interval of λ values, $\mathcal{S}_{k'} \subset [0, \lambda_{\max}]$, such that for each $\lambda^* \in \mathcal{S}_{k'}$, then $\arg\min_k C(k, \lambda^*) = k'$. These intervals, for $k = 1, \ldots, d$, form a partition of $[0, \lambda_{\max}]$, i.e.,

$$\mathcal{S}_1 \cup \mathcal{S}_2 \ldots \cup \mathcal{S}_d = [0, \lambda_{\max}], \tag{4}$$

and $\mathcal{S}_k \cap \mathcal{S}_j = 0$, for all $k \neq j$. Observe that, by construction, $\mathcal{S}_0 = \emptyset$ due to the definition of $\lambda_{\max}$. Furthermore, we can use the information provided by the measures $|\mathcal{S}_k|$, defining the weights $\bar{w}_k \propto |\mathcal{S}_k|$, i.e.,

$$\bar{w}_k = \frac{|\mathcal{S}_k|}{\sum_{j=0}^{d} |\mathcal{S}_j|} = \frac{|\mathcal{S}_k|}{\sum_{j=1}^{d} |\mathcal{S}_j|}, \tag{5}$$

where $|\mathcal{S}_0| = 0$. Note that $\bar{w}_k$, for $k = 1, \ldots, d$, defines a probability mass function, $\sum_{k=1}^{K} \bar{w}_k = 1$. These weights can be computed by the Monte Carlo method [41].

A first output of SIC is the set $\mathcal{E}$ of indices such that the corresponding weights are non-zero:

$$\mathcal{E} = \{\text{all } k : \bar{w}_k > 0\} = \{k_E^{(1)}, k_E^{(2)}, \ldots, k_E^{(J)}\}. \tag{6}$$

They can be interpreted as a possible "elbow" of the curve $V(k)$, i.e., any possible selected model is represented by one index $k_E^{(j)}$.

Remark: If $V(k) = -2\log \ell_{\max}$, then the set $\mathcal{E}$ virtually contains all the solutions of the rest of the information criteria in the literature, such as BIC and AIC. See Table 9 for further details.

Note that we have denoted $J = |\mathcal{E}|$ with $J \leq K$ and, in some cases, $J \ll K$. This is because some value $k' \neq 0$ could never be a minimum so that $|\mathcal{S}_{k'}| = 0$. Therefore, we can have a sensible reduction of the number of possible models to choose from. To select just one model, the more conservative solution is $k_E = \max k_E^{(j)}$ choosing the more complex model. However, the suggestion in [41], is to define the cumulative sum of the first m weights i.e., $W_m = \sum_{i=1}^{m} \bar{w}_i$, with $1 < m \leq K$ and choose as "elbow" the index defined as

$$k_E = \min\{k : W_k \geq \ell\}, \tag{7}$$

where $\ell \in (0, 1]$ is a confidence level, for instance, $\ell = 0.9$ or $\ell = 0.95$ (denote as SIC-95, in this work). A more conservative choice (i.e., selecting a more complex model within $\mathcal{E}$) can be obtained by setting $\ell = 0.99$ or even $\ell = 1$ (denote as SIC-100, in this work).

ACKNOWLEDGMENT

This work was supported in part by Agencia Estatal de Investigación-AEI (project SPGRAPH, ref. num. PID2019-105032GBI00, and Grant PID2022-136887NB-I00 (POLI-GRAPH) funded by MCIN/AEI/10.13039/501100011033), in part by Comunidad de Madrid and Universidad Rey Juan Carlos (Proyecto I+D Jóvenes Doctores, AUTO-BA-GRAPH, ref. num. F861), and in part by Programa de Excelencia-Convenio Plurianual entre Comunidad de Madrid y la Universidad Rey Juan Carlos (ref. num. Y158/DF007003/30-06-2020, ref. num F840).

REFERENCES

[1] M. F. Southworth, "The sonic environment of cities," Ph.D. dissertation, Dept. City Regional Planning, Massachusetts Inst. Technol., Cambridge, MA, USA, 1967.
[2] R. M. Schafer, *The Soundscape: Our Sonic Environment and the Tuning of the World*. New York, NY, USA: Simon and Schuster, 1993.
[3] *Part 1 Definition and Conceptual Framework*, ISO Acoustics—Soundscape, International Standards Organization, Geneva, Switzerland, 2014, vol. 12913, no. 1.
[4] J. Vida, J. A. Almagro, R. García-Quesada, F. Aletta, T. Oberman, A. Mitchell, and J. Kang, "Soundscape attributes in Spanish: A comparison with the English version of the protocol proposed in method a of the ISO/TS 12913–2," *Appl. Acoust.*, vol. 211, Aug. 2023, Art. no. 109516.
[5] A. Mitchell, F. Aletta, and J. Kang, "How to analyse and represent quantitative soundscape data," *JASA Exp. Lett.*, vol. 2, no. 3, pp. 1–9, Mar. 2022.
[6] M. Erfanian, A. Mitchell, F. Aletta, and J. Kang, "Psychological well-being and demographic factors can mediate soundscape pleasantness and eventfulness: A large sample study," *J. Environ. Psychol.*, vol. 77, Oct. 2021, Art. no. 101660.

[7] M. Lionello, F. Aletta, A. Mitchell, and J. Kang, "Introducing a method for intervals correction on multiple Likert scales: A case study on an urban soundscape data collection instrument," *Frontiers Psychol.*, vol. 11, Jan. 2021, Art. no. 602831.

[8] A. Fiebig, P. Jordan, and C. C. Moshona, "Assessments of acoustic environments by emotions—The application of emotion theory in soundscape," *Frontiers Psychol.*, vol. 11, p. 3261, Nov. 2020.

[9] F. Abri, L. F. Gutiérrez, A. S. Namin, D. R. W. Sears, and K. S. Jones, "Predicting emotions perceived from sounds," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 2057–2064.

[10] J. Fan, M. Thorogood, and P. Pasquier, "Automatic soundscape affect recognition using a dimensional approach," *J. Audio Eng. Soc.*, vol. 64, no. 9, pp. 646–653, Sep. 2016.

[11] J. Fan, M. Thorogood, and P. Pasquier, "Emo-soundscapes: A dataset for soundscape emotion recognition," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 196–201.

[12] F. Aletta, J. Kang, and Ö. Axelsson, "Soundscape descriptors and a conceptual framework for developing predictive soundscape models," *Landscape Urban Planning*, vol. 149, pp. 65–74, May 2016.

[13] K. Herranz-Pascual, I. García, I. Aspuru, I. Díez, and Á. Santander, "Progress in the understanding of soundscape: Objective variables and objectifiable criteria that predict acoustic comfort in urban places," *Noise Mapping*, vol. 3, no. 1, pp. 247–263, Sep. 2016.

[14] T. Giannakopoulos, G. Siantikos, S. Perantonis, N.-E. Votsi, and J. Pantis, "Automatic soundscape quality estimation using audio analysis," in *Proc. 8th ACM Int. Conf. Pervasive Technol. Rel. Assistive Environ.*, Jul. 2015, pp. 1–9.

[15] T. Van Renterghem, L. Dekoninck, and D. Botteldooren, "Multi-stage sound planning methodology for urban redevelopment," *Sustain. Cities Soc.*, vol. 62, Nov. 2020, Art. no. 102362.

[16] A. L. Brown, "Soundscapes and environmental noise management," *Noise Control Eng. J.*, vol. 58, no. 5, p. 493, 2010.

[17] M. Adams, T. Cox, G. Moore, B. Croxford, M. Refaee, and S. Sharples, "Sustainable soundscapes: Noise policy and the urban experience," *Urban Stud.*, vol. 43, no. 13, pp. 2385–2398, Dec. 2006.

[18] J. Segura-Garcia, J. M. A. Calero, A. Pastor-Aparicio, R. Marco-Alaez, S. Felici-Castell, and Q. Wang, "5G IoT system for real-time psychoacoustic soundscape monitoring in smart cities with dynamic computational offloading to the edge," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 12467–12475, Aug. 2021.

[19] Y. Zhao, S. Sheppard, Z. Sun, Z. Hao, J. Jin, Z. Bai, Q. Bian, and C. Wang, "Soundscapes of urban parks: An innovative approach for ecosystem monitoring and adaptive management," *Urban Forestry Urban Greening*, vol. 71, May 2022, Art. no. 127555.

[20] P. Lopes, A. Liapis, and G. N. Yannakakis, "Modelling affect for horror soundscapes," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 209–222, Apr. 2019.

[21] T. Görne, "The emotional impact of sound: A short theory of film sound design," *EPiC Ser. Technol.*, vol. 1, pp. 17–30, Jan. 2019.

[22] F. Abri, L. F. Gutiérrez, P. Datta, D. R. W. Sears, A. S. Namin, and K. S. Jones, "A comparative analysis of modeling and predicting perceived and induced emotions in sonification," *Electronics*, vol. 10, no. 20, p. 2519, Oct. 2021.

[23] A. Mitchell, F. Aletta, T. Oberman, M. Erfanian, and J. Kang, "A conceptual framework for the practical use of predictive models and soundscape indices: Goals, constraints, and applications," in *Proc. INTER-NOISE NOISE-CON Congr. Conf.*, Wakefield, MA, USA: Institute of Noise Control Engineering, 2023, pp. 2108–2118.

[24] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, Dec. 1980.

[25] Ö. Axelsson, M. E. Nilsson, and B. Berglund, "A principal components model of soundscape perception," *J. Acoust. Soc. Amer.*, vol. 128, no. 5, pp. 2836–2846, Nov. 2010.

[26] W. J. Davies, N. S. Bruce, and J. E. Murphy, "Soundscape reproduction and synthesis," *Acta Acustica United With Acustica*, vol. 100, no. 2, pp. 285–292, Mar. 2014.

[27] D. Västfjäll, M. Kleiner, and T. Gärling, "Affective reactions to interior aircraft sounds," *Acta Acustica United With Acustica*, vol. 89, no. 4, pp. 693–701, 2003.

[28] B. T. Lawrence, J. Hornberg, T. Haselhoff, R. Sutcliffe, S. Ahmed, S. Moebus, and D. Gruehn, "A widened array of metrics (WAM) approach to characterize the urban acoustic environment; a case comparison of urban mixed-use and forest," *Appl. Acoust.*, vol. 185, Jan. 2022, Art. no. 108387.

[29] F. Aletta, Ö. Axelsson, and J. Kang, "Dimensions underlying the perceived similarity of acoustic environments," *Frontiers Psychol.*, vol. 8, p. 1162, Jul. 2017.

[30] M. Lionello, F. Aletta, and J. Kang, "A systematic review of prediction models for the experience of urban soundscapes," *Appl. Acoust.*, vol. 170, Dec. 2020, Art. no. 107479.

[31] R. S. Millán-Castillo, L. Martino, E. Morgado, and F. Llorente, "An exhaustive variable selection study for linear models of soundscape emotions: Rankings and Gibbs analysis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 2460–2474, Jul. 2022.

[32] J. Fan, F. Tung, W. Li, and P. Pasquier, "Soundscape emotion recognition via deep learning," in *Proc. Sound Music Comput.*, Jul. 2018, pp. 1–6.

[33] S. Ntalampiras, "Emotional quantification of soundscapes by learning between samples," *Multimedia Tools Appl.*, vol. 79, nos. 41–42, pp. 30387–30395, Nov. 2020.

[34] R. San Millán-Castillo, L. Martino, and E. Morgado, "Variable selection analysis for decision tree regression models of soundscapes emotions," in *Proc. 53th Congreso Español de Acústica XII Congreso Ibérico de Acústica*, 2022, pp. 930–938.

[35] P. Krishan and F. Abri, "Classifying perceived emotions based on polarity of arousal and valence from sound events," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2022, pp. 2849–2856.

[36] T. Giannakopoulos, M. Orfanidi, and S. Perantonis, "Athens urban soundscape (ATHUS): A dataset for urban soundscape quality recognition," in *Proc. Int. Conf. Multimedia Modeling*, Thessaloniki, Greece. Cham, Switzerland: Springer, Jan. 2019, pp. 338–348.

[37] A. Mitchell, T. Oberman, F. Aletta, M. Erfanian, M. Kachlicka, M. Lionello, and J. Kang, "The international soundscape database: An integrated multimedia database of urban soundscape surveys—Questionnaires with acoustical and contextual information (0.2.4) [data set]," *Data set*, vol. 74, no. 2, pp. 248–254, 2022.

[38] K. Ooi, Z.-T. Ong, K. N. Watcharasupat, B. Lam, J. Y. Hong, and W.-S. Gan, "ARAUS: A large-scale dataset and baseline models of affective responses to augmented urban soundscapes," *IEEE Trans. Affect. Comput.*, vol. 15, no. 1, pp. 1–17, Jan./Mar. 2024.

[39] F. Orga, A. Mitchell, M. Freixes, F. Aletta, R. M. Alsina-Pagès, and M. Foraster, "Multilevel annoyance modelling of short environmental sound recordings," *Sustainability*, vol. 13, no. 11, p. 5779, May 2021.

[40] A. Mitchell, T. Oberman, F. Aletta, M. Kachlicka, M. Lionello, M. Erfanian, and J. Kang, "Investigating urban soundscapes of the COVID-19 lockdown: A predictive soundscape modeling approach," *J. Acoust. Soc. Amer.*, vol. 150, no. 6, pp. 4474–4488, Dec. 2021.

[41] L. Martino, R. San Millán-Castillo, and E. Morgado, "Spectral information criterion for automatic elbow detection," *Expert Syst. Appl.*, vol. 231, Nov. 2023, Art. no. 120705.

[42] E. Morgado, L. Martino, and R. S. Millán-Castillo, "Universal and automatic elbow detection for learning the effective number of components in model selection problems," *Digit. Signal Process.*, vol. 140, Aug. 2023, Art. no. 104103.

[43] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[44] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 64, no. 4, pp. 583–639, Oct. 2002.

[45] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: A platform for the creation of open audio datasets," in *Proc. 18th ISMIR Conf.*, X. Hu, S. J. Cunningham, D. Turnbull, and Z. Duan, Eds., Suzhou, China: International Society for Music Information Retrieval (ISMIR), Oct. 2017, pp. 93–486.

[46] C. Xu and J. Kang, "Soundscape evaluation: Binaural or monaural?" *J. Acoust. Soc. Amer.*, vol. 145, no. 5, pp. 3208–3217, May 2019.

[47] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software," in *Proc. ISMIR*. Princeton, NJ, USA: Citeseer, 2010, pp. 441–446.

[48] O. Lartillot, P. Toiviainen, and T. Eerola, "A MATLAB toolbox for music information retrieval," in *Data Analysis, Machine Learning and Applications*. Cham, Switzerland: Springer, 2008, pp. 261–268.

[49] R. Cain, P. Jennings, and J. Poxon, "The development and application of the emotional dimensions of a soundscape," *Appl. Acoust.*, vol. 74, no. 2, pp. 232–239, Feb. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X1100301X

[50] D. de la Prida, A. Pedrero, M. Á. Navacerrada, and C. Díaz, "Relationship between the geometric profile of the city and the subjective perception of urban soundscapes," *Appl. Acoust.*, vol. 149, pp. 74–84, Jun. 2019.
[51] A. Fiebig, "Reliability of in-situ measurements of acoustic environments," in *Proc. DAGA*, 2016, pp. 1335–1338.
[52] S. R. Payne and C. Guastavino, "Exploring the validity of the perceived restorativeness soundscape scale: A psycholinguistic approach," *Frontiers Psychol.*, vol. 9, p. 2224, Nov. 2018.
[53] D. A. Bligh, *What's the Use of Lectures: First U.S. Edition of the Classic Work on Lecturing*. Hoboken, NJ, USA: Wiley, 2000.
[54] M. Revilla and C. Ochoa, "Ideal and maximum length for a web survey," *Int. J. Market Res.*, vol. 59, no. 5, pp. 557–565, 2017.
[55] R. San Millán-Castillo, I. López-Peñalver, L. Martínez-Cano, E. Morgado, and L. Martino, "Análisis de técnicas de grabación asequibles para paisajes sonoros," in *Proc. 54th Congreso Español de Acústica XIII Congreso Ibérico de Acústica*, Nov. 2023, pp. 139–144.
[56] A. S. Sudarsono, Y. W. Lam, and W. J. Davies, "The effect of sound level on perception of reproduced soundscapes," *Appl. Acoust.*, vol. 110, pp. 53–60, Sep. 2016.
[57] L. Breiman, *Classification Regression Trees*. Evanston, IL, USA: Routledge, 2017.
[58] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.
[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[60] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. Cham, Switzerland: Springer, 2009.
[61] F. Questier, R. Put, D. Coomans, B. Walczak, and Y. V. Heyden, "The use of CART and multivariate regression trees for supervised and unsupervised feature selection," *Chemometric Intell. Lab. Syst.*, vol. 76, no. 1, pp. 45–54, Mar. 2005.
[62] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2225–2236, Oct. 2010.
[63] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.
[64] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, "Marginal likelihood computation for model selection and hypothesis testing: An extensive review," 2020, *arXiv:2005.08334*.
[65] F. Llorente, L. Martino, E. Curbelo, J. López-Santiago, and D. Delgado, "On the safe use of prior densities for Bayesian model selection," *WIREs Comput. Statist.*, vol. 15, no. 1, Jan. 2023, Art. no. e1595.
[66] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 41, no. 2, pp. 190–195, Jan. 1979.
[67] E. Morgado, L. Martino, and R. S. Millán-Castillo, "Universal and automatic elbow detection for learning the effective number of components in model selection problems," 2022, *arXiv:2308.09102*.

ROBERTO SAN MILLÁN-CASTILLO received the B.Sc. degree in telecommunications engineering (electrical engineering), with a minor in sound and image, from Universidad Politécnica de Madrid (UPM), in 2000, the M.Sc. degree in project management from Universidad Antonio de Nebrija, in 2012, the M.Sc. degree in acoustic engineering (research training) from UPM, in 2013, and the Ph.D. degree in multimedia and communications from UC3M+URJC, in 2020. In 2011, he joined Universidad Rey Juan Carlos (URJC), Spain, as a Lecturer. Since 1999, he had hands-on experience in the industry as an acoustics, noise control, audio, and instrumentation consultant in different positions, such as engineering, project management, and sales. All at once, he taught an endless number of training courses at companies, universities, professional associations, and government institutions. His main research interests include signal processing and machine learning techniques applied to practical problems with acoustical and audio signals.

LUCA MARTINO received the Ph.D. degree in statistical signal processing from Universidad Carlos III de Madrid, Spain, in 2011. He was an Assistant Professor with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid. In August 2013, he joined with the Department of Mathematics and Statistics, University of Helsinki. He was a Postdoctoral Researcher with Universidade de São Paulo (USP) and Universitat de València, València, Spain. He is currently an Associate Professor with Universidad Rey Juan Carlos (URJC). His research interests include Bayesian inference, Monte Carlo methods, and exact methods for generating random variables.

EDUARDO MORGADO received the degree in telecommunication engineering from the University Carlos III of Madrid, Leganés, Spain, in 2004, and the Ph.D. degree in telecommunications engineering from Universidad Rey Juan Carlos (URJC), Fuenlabrada, Spain, in 2009. He is currently an Associate Professor with the Department of Signal Theory and Communications, URJC. His research interests include signal processing for wireless communications with applications to ad hoc and sensor networks.