CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

| | |
|---|---|
| Title | Soundscape segregation based on visual analysis and discriminating features |
| Authors | Dias, Fábio Felix;Pedrini, Helio;Minghim, Rosane |
| Publication date | 2021-11-04 |
| Original Citation | Dias, F. F., Pedrini, H. and Minghim, R. (2021) 'Soundscape segregation based on visual analysis and discriminating features', Ecological Informatics, 61, 101184 (13 pp). doi: 10.1016/j.ecoinf.2020.101184 |
| Type of publication | Article (peer-reviewed) |
| Link to publisher's version | https://www.sciencedirect.com/science/article/pii/S1574954120301345 - 10.1016/j.ecoinf.2020.101184 |
| Rights | © 2020 Elsevier B.V. All rights reserved. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/ - http://creativecommons.org/licenses/by-nc-nd/4.0/ |
| Download date | 2026-11-15 01:48:37 |
| Item downloaded from | https://hdl.handle.net/10468/11042 |

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

# Soundscape segregation based on visual analysis and discriminating features

Fábio Felix Dias[a,*], Helio Pedrini[b], Rosane Minghim[a,c]

[a]*Instituto de Ciências Matemáticas e de Computação, University of São Paulo*
*Av. Trabalhador São-carlense, 400, São Carlos, SP, Brazil, 13566-590*
[b]*Institute of Computing, University of Campinas*
*Av. Albert Einstein, 1251, Campinas, SP, Brazil, 13083-852*
[c]*School of Computer Science and Information Technology, University College Cork, Ireland*

**Abstract**

The distinction of landscapes based on their sound patterns is useful for several analyses. For instance, comparisons of audio files from different periods enable the detection of changes over time in a particular habitat, signaling events of importance, such as modifications in the balance between species and presence of new ones. The handling of a large number of different sound recordings in wild environments also reduces the set of sounds to be examined. However, the current efforts towards soundscape interpretation do not provide enough elements for researchers to automatically split soundscape datasets with degrees of similarity, thus requiring users' feedback for the grouping of highly related recordings. This work introduces a strategy for the exploration and analysis of soundscapes that highlights data characteristics related to differences and similarities among distinct soundscapes. It is based on a visual and numerical evaluation of feature spaces and was applied to three feature sets, namely acoustic indices and measurements, images from audio spectrograms depicted by classic features, and the same images depicted by features automatically generated by Deep Learning techniques. The results indicate that certain combinations of acoustic indices and measurements perform well for the discrimination task, although other feature sets have not been discarded. In addition, visual techniques were able to assist this type of analysis.

*Keywords:* Acoustic features, Spectrogram image, Image descriptors, Deep learning, Information visualization

## 1. Introduction

Soundscape ecology is the study of the relationship between a landscape and its sounds [1]; it refers to a collection of sounds that emanate from a landscape, defining its spatio-temporal patterns [2]. Such sounds are categorized as biophonic, geophonic and anthrophonic. Krause [3] defined biophony and geophony as the collection of biological and non-biological sounds (wind, thunder, river, and so on), respectively, whereas Pijanowski et al. [2] described anthrophony as sounds created directly or indirectly by human beings.

By analyzing recordings, soundscape studies have attempted to evaluate diversity, as well as understand landscape changes and the way certain sounds produced by elements, such as airplanes, vessels, and new species affect the environment [4]. Examples of results in that field include analyses of influence of urbanization on animal life [5], estimation of birds diversity in certain forest areas [6], measurement of animal diversity in woodlands [7], quantification of biodiversity in marine areas [8], and description of acoustic activity in a natural reserve [9].

The large number of recordings prompted by advances in technology is a great challenge for specialists (e.g. the amount of data in several specialized laboratories has easily reached many terabytes in only a few years [4, 10, 11, 12, 13]). As a consequence, specialists need to rely on proper strategies (e.g. feature extraction, machine learning, and data visualization techniques) to efficiently analyze such data and improve knowledge acquisition.

*Corresponding author
*Email address:* f_diasfabio@usp.br (Fábio Felix Dias)

Among the approaches designed for summarizing, representation, visualization, and analyses of soundscape data, the use of acoustic indices, which are mathematical functions or algorithms that evaluate sound dynamics and biodiversity aspects from sound signals, has been recurrently applied [14]. However, some authors have claimed that it is imprecisely formulated [15], is sensitive to noise [8], provides limited representation power [16], and requires extensive testing prior to its application to distinct environments [17, 18]. Regardless of such criticisms, different indices can capture certain aspects that may work towards explaining and discriminating environments (e.g. [19, 20, 21]), in the same manner that image features are employed in visual categorization. Other researches, however, have applied different signal features [22, 23] and features extracted from other domains, such as spectrogram images [10, 24, 25, 26].

The main questions the present study aims to tackle refer to (i) whether soundscapes can be segregated based on features from audio recorded in natural environments, and (ii) whether a suitable approach to visual layout and interaction can support data exploration and identification of groups of soundscapes for describing similar environments. The following aspects were, therefore, investigated: (i) a set of characteristics that better describe a specific soundscape (acoustic indices and measurements, and spectrogram image descriptors), and (ii) application of numerical coefficients and multidimensional projection techniques to support quantitative and qualitative evaluation and exploration of such data spaces.

The use of visual approaches in soundscape investigation is justified by the lack of full knowledge of features that differentiate soundscapes and the power of such techniques to communicate data patterns. An example of a visual exploration of soundscape features is the tool presented in Reis et al. [27], which identifies relevant acoustic features and their relationships for representing specific events or general trends in soundscapes. Therefore, distinct soundscapes can be characterized, described, and compared. The authors based their technique on the clustering of highly correlated features and exploration of feature spaces with visual feedback. Phillips et al. [11] applied clustering techniques to summarize acoustic features without loss of ecological meaning and explored audio sets with visualization techniques, such as *dial plots* and *polar histograms*, etc. These approaches facilitate the navigation of long-duration audio recordings and reveal relevant ecological content, such as frequency of events through time, relation among events, and similarities among distinct places. Znidersic et al. [28] created a strategy to monitor a bird species based on visual and machine learning techniques. To visually confirm the presence of calling patterns, the authors employed a false-color spectrogram [13] and, to quantify the number of bird calls, they used a Random Forest regressor. With the visual approach, Znidersic et al. [28] analyzed recordings from both continuous and non-continuous periods, and determined the presence of the species of interest. Moreover, details of the false-color spectrogram showed cryptic species that were not the focus of the researchers.

Advances in visual and algorithmic techniques have therefore enhanced the exploration of acoustic data and supported decision-making on serious ecological problems as identified by the works reported above. Moreover, they can answer more specific questions associated with the level of anthrophony over a landscape, effects of climate events and pollution, species migration or invasion, and other events that may change the acoustic signature of a particular area.

We have also found a few soundscape ecology researches that employ projection techniques (e.g., [11]), which have been central in several applications with different data scenarios [29, 30]. Each of these techniques can yield distinct layouts due to their different objectives, formulations, and capabilities (see [31]). That approach can provide interesting insight for soundscape researchers in several tasks, such as segregation of large groups, identification of local patterns, and even qualification of environments.

This work is organized as follows. Section 2 briefly reviews the relevant concepts employed in this research. Section 3 presents the proposed methods and materials used in our experiments. Section 4 reports the experimental results obtained with the proposed methodology. Section 5 discusses the experimental results. Finally, Section 6 provides the conclusions and directions for future work.

## 2. Basic concepts

As depicted in Figure 1, although acoustic indices and measurements can be calculated both from original signals and spectrum of signal frequencies, many tools extract relevant features from other sources related to an audio signal. An example is the set of features extracted from the spectrogram image, which graphically represents the relation of sound power and time-frequency domain [10]. Features generated from time-frequency information or their images

can represent, at some level, patterns contained in audio signals, such as those highlighted in the figure. Beyond acoustic features, this section describes other tools used in our investigations on soundscape feature extraction from various perspectives, and the numerical and visual tools that evaluated and explored the feature space representation.

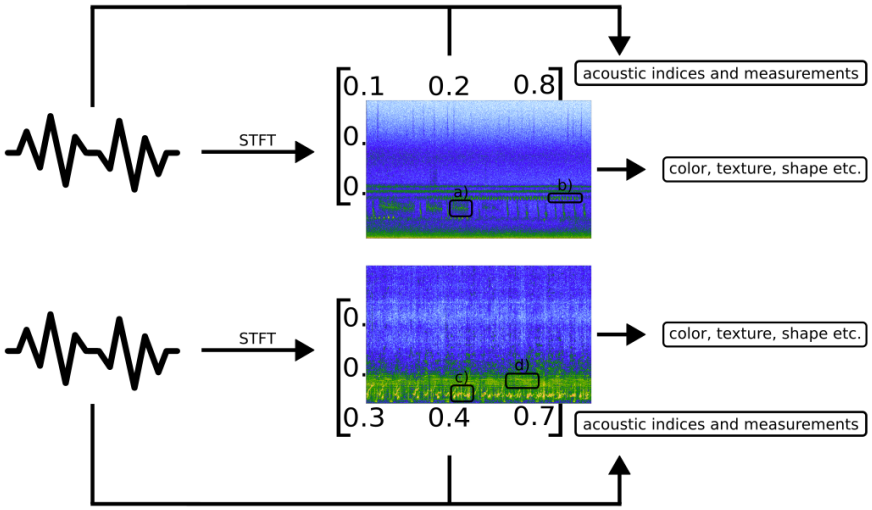



Figure 1: Feature spaces of recordings from terrestrial (top) and underwater areas (bottom). Some features can be generated from time domain, frequency domain (spectrum or spectrogram), a combination of time/frequency domains, and image domain, and represent audio patterns, such as those highlighted in the spectrogram images (a to d).

### *2.1. Image representation and description*

Image descriptors are algorithms that compute feature vectors from images. They can be categorized by the type of information they extract, such as color, texture or shape[32]. The next subsections describe those employed in this study for the extraction of features from spectrograms.

#### *2.1.1. Color descriptors*

Color is an important feature for image identification [32] and is frequently employed to represent a picture. This research employs color descriptors, such as Global Color Histogram (GCH) [33], a common and simple descriptor that creates a histogram of quantized image colors.

Auto Color Correlation (ACC) [34] attempts to measure the probability of two pixels of the same color at a certain distance from each other being found. Color Coherence Vector (CCV) [35] creates a histogram and classifies pixels into coherent and not coherent, based on a previously defined color threshold. Border/Interior Pixel Classification (BIC) [36] defines a histogram that separates colors in the border and interior regions.

#### *2.1.2. Texture descriptors*

Texture is a well-known image characteristic of difficult description. Generally speaking, it is a collection of intensity variations that create patterns and repetitions [32]. Our study employed texture descriptors, such as Gray Level Co-occurrence Matrix (GLCM) [37], which is a common statistical method that quantifies intensity relations among image pixels in different directions.

Spectral Descriptors [38] employ Fourier Transform to represent texture patterns and can be interpreted as the spectrum of high energy areas. Local Binary Patterns (LBP) [39] extract information on intensity variation (this method is invariant to scaling, rotation and illumination conditions).

### 2.2. Autoencoder

Machine Learning (ML) approaches are widely used to improve or automate tasks, such as web search, speech recognition, autonomous driving and data analysis, and are characterized as an automatic process that extracts significant patterns from data collections [40].

ML is permeated with tools and techniques, such as autoencoder, a recent and particularly successful one. This technique is a Deep Learning approach, which is a collection of methods that employ Neural Networks (NN) with a large number of layers (tens or hundreds) [41].

An autoencoder aims to create an approximate representation of a dataset [42], based solely on the structure of data without considering its labels (unsupervised approach). This network has two blocks of layers: (i) one (encoder) that creates compact data representation and (ii) the other (decoder) that approximately reconstructs the incoming data from the representation created by the encoder. After a training process, the representation produced by the encoder can be employed as a well-suited feature space of the data.

The modeling of an autoencoder is similar to the process of representing any neural network. The number of layers and neurons (kernels or filters) must be initially defined, as well as the size of the inputs, type of layers to be applied, activation functions, loss functions and optimization method. Additionally, some parameters, such as batch size, number of epochs, size of datasets (training, validation and testing), must be defined.

### 2.3. Data visualization techniques

The field of Information Visualization (InfoVis) applies visual representation to explore attribute values and relationships among data, in order to acquire knowledge about them [43, 44].

The representation for data types with no intrinsic physical attributes is abstract and must encode patterns that may appear in the data [44]. For datasets with many attributes, such as the ones from genetic sequencing, social networks, text, image and audio, it is currently understood that embedding them in two dimensions can help find groups of interest, as well as common characteristics in datasets. In the following subsections, we describe some point-based techniques employed in our visual pipeline, which represents data items as some graphic entities (e.g., a circle or rectangle).

#### 2.3.1. Multidimensional projections

Multidimensional Projections (MDP) map a dataset from a high-dimensional space (with tens, hundreds or thousands of features) to a lower one with 2 or 3 dimensions (2D or 3D) through the application of a mathematical transformation. The numerical representation provides a visual representation, such as the scatter plot on the right side of Figure 2. The layout is a visual representation that allows the identification of patterns of similarity based on proximity.

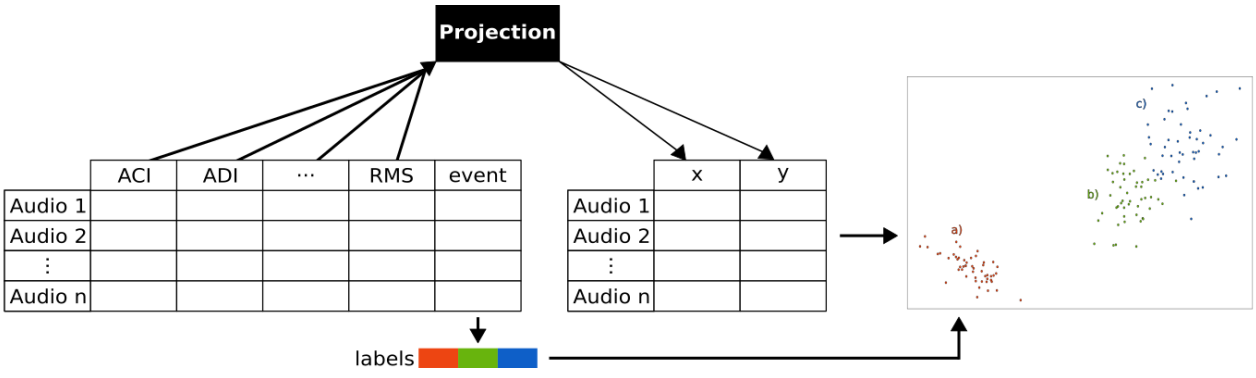



Figure 2: The projection converts a large number of acoustic features (left) into a two-dimensional space (middle) that can be graphically represented in a scatter plot (right), where each point represents an audio file and the distances among them represent the similarity of audio content. Event types can be normalized for color space and associated with points to facilitate visual identification of recordings with the same events. Events (b) and (c) are similar, but different from event (a).

Techniques for MDP aim to preserve a certain property of the original feature spaces, such as neighborhood relationships, grouping of similar data points, or distances. Figure 2 shows group relations among data points, labeled as (a) red, (b) green and (c) blue. Although projections do not always reach the goal of preservation, due to the size, dimension and typical sparsity of the data [45], they can aid data exploration, revealing structures and relationships.

Several dimensionality reduction and projection approaches have been designed to place points in visual space and preserve relevant properties of the original one, for example, Principal Component Analysis (PCA) [46, 47], Multidimensional Scaling (MDS) [48], Force Scheme [49], t-Stochastic Neighbor Embedding (t-SNE) [50], Least Square Projection (LSP) [51], Local Affine Multidimensional Projection (LAMP) [52].

### *2.4. Evaluation of projection results and feature space*

*Silhouette coefficient* [53] has been originally applied to validate results of clustering algorithms through measurements of cohesion (Equation 2) and separation (Equation 3) of data groups. Nevertheless, it has been considered to evaluate quality of projections and feature spaces [30, 54]. Data groups or labels are required for the calculation of coefficient of trial projections and data spaces.

A silhouette coefficient is generated by Equation 1 for each data point $\mathbf{x}_i$ and a value of the complete dataset is obtained from the average of all coefficient values. These values vary between -1 and 1, where the best cohesion/segregation measures are represented by values closer to 1.

$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}} \tag{1}$$

$$a(\mathbf{x}_i) = \frac{1}{N_A - 1} \sum_{\substack{i \neq j \\ x_j \in A}} d(\mathbf{x}_i, \mathbf{x}_j), \text{ where } \begin{cases} A & \text{is the } \mathbf{x}_i \text{ group,} \\ N_A & \text{is the quantities of items in } A, \\ d & \text{is a similarity function} \end{cases} \tag{2}$$

$$b(\mathbf{x}_i) = \min_{\forall C \neq A} \{D(\mathbf{x}_i, C)\} \tag{3}$$

$$D(\mathbf{x}_i, C) = \frac{1}{N_C} \sum_{x_j \in C} d(\mathbf{x}_i, \mathbf{x}_j) \tag{4}$$

In Equation 3, $A$ is the $\mathbf{x}_i$ group and $C$ is any other data group. In Equation 4, $N_C$ is the quantities of items in the group $C$ and $d$ is the same similarity function used in Equation 2.

A silhouette value denotes how close each point is to its data group and how far each point is to other groups, which enables evaluations of how well a specific feature space represents data characteristics based on grouping and segregation.

## 3. Material and methods

This section presents the main steps for the exploration and analysis of soundscape, as depicted in Figure 3. There are format conversion steps applied to recordings, and the raw files are used to extract features. The first step refers to feature extraction, which uses, in addition to acoustic indices and measurements, descriptors (standard and learned) for spectrogram images for the creation of a multidimensional data space. The second and third steps employ visual and numerical methods to detect the feature set that best represents the analyzed soundscapes. Visual techniques also reveal audio content and characteristics related to sound similarities.

### *3.1. Feature extraction*

The next subsections describe techniques for the extraction of audio and image features applied in this study.

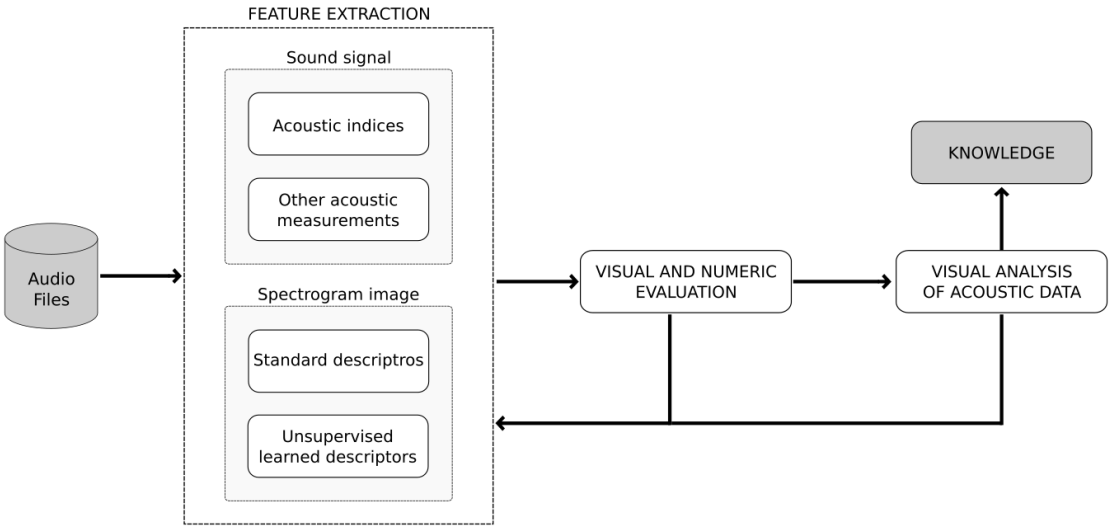



Figure 3: Main steps of the proposed method.

*3.1.1. Acoustic indices and other acoustic measurements*

As addressed in Section 1, acoustic indices are extensively employed for analyses of soundscapes. The present study applied Bioacoustic Index (Bio) [55], a set composed of Temporal Entropy ($H_t$), Frequency Entropy ($H_f$) and Acoustic Entropy Index (H) [56], Acoustic Complexity Index (ACI) [6], Acoustic Evenness Index (AEI) [57], tuple $M$ index and Acoustic Richness (AR) [7], Normalized Difference Soundscape Index (NDSI) [58] and Acoustic Diversity Index (ADI) [59].

Other classic measurements were calculated for evaluation of audio signal, such as Sound Pressure Level (SPL) [23], Number of Peaks (NP), Root Mean Square (RMS) and functions that describe signal variations, such as Roughness [60] and Rugosity [61]. Mel-frequency Cepstrum Coefficients (MFCC), applied to audio analysis [22, 62], were also generated.

*3.1.2. Standard and unsupervised learned image descriptors*

Apart from acoustic indices and measurements, this research employed features extracted from spectrogram images as an alternative soundscape description. The use of images, which constrains the analysis space and provides a user visual interaction, has been successfully applied in other researches, as stated by Harvey [10] and Xie et al. [26].

Our study adopted two approaches for the description of images. In the first, standard features, such as GLCM, LBP, Fourier for texture, GCH, ACC, CCV, and BIC (see Section 2.1), were extracted and an autoencoder (see Section 2.2) was built to learn and automatically extract features that better represent image patterns generated by sound structure. Deep learning techniques have been attained suitable results in many soundscape researches, as stated by Lostanlen et al. [63], Thomas et al. [64], and Kirsebom et al. [65].

### *3.2. Feature set evaluation*

Instead of exploring attributes for evincing specific and internal questions of a region, this study analyzed the power of extracted characteristics for distinguishing environments, as highlighted in the introduction. Recordings from five collection areas were used in the evaluation of the feature set (see Section 3.3), being two terrestrial areas and the other three underwater areas. Differences between the terrestrial and underwater soundscapes were clear, however, there were weak assumptions about differences between the two terrestrial areas and among the three underwater areas. Silhouette coefficient values were calculated for each feature set after extraction. Collection areas were considered labels, and the silhouette values enabled measurements of the capability of sets and subsets of features to discriminate data.

Techniques, such as Force Scheme, LAMP, LSP, MDC, PCA and t-SNE (see Section 2.3.1), were employed for visual inspections of datasets and enabled estimations of the quality of features. If a projection cannot represent separations of distinct areas, the features chosen are probably not the best to describe differences in supposedly distinct data. The collection areas were considered as colors and associated with points to facilitate the visual distinction of areas and to identify recordings with the same patterns.

The definition of the best MDP for the description of data can be visual, however, this is subjective and complex task, since different projections can depict similar layouts and due to the cluttering from the lack of visual screen space. As a result, an actual group separation becomes a challenge, and, towards a proper MDP evaluation, visual inspection was combined with the silhouette coefficient calculated from the projected space. The values were compared with the silhouette values of the original data space and the level of disturbance in projections was verified.

A *min-max* normalization of feature values was performed, mapping feature values into $[0, 1]$ range, in order to verify its impact on the results. Silhouette coefficients and MDPs were computed again, and the new results were compared with the previous ones (without normalization).

Finally, Automatic Feature Selection techniques can identify more coherent features for a given task, such as segregation or classification. Therefore, some techniques were applied for the selection of proper sets of features and their MDP and silhouette results were compared with those obtained in this study. This research applied Correlation-based, Information Gain and Relief-F methods via Weka data mining program[1] described by Jović et al. [66].

### *3.3. Datasets*

Our experiments employed a set of 4,340 audio files (≈485 hours) from natural landscapes of the following different environments:

1. **Terrestrial:** the data was provided by professor Bryan C. Pijanowski from Purdue University, Indiana, USA. The data was collected in two areas at the *La Selva* Biological Station, Costa Rica. The first (CostaRica1) is an old-growth forest near *Sarapiquí* river, whereas the second (CostaRica2) is a secondary forest farther from the same river, but in the same biological station[2]. Four files were recorded for each hour of the day (the first with 10 minutes and the others with 1-minute audio). There are 3,061 files divided into 1,246 (4.054 min.) from CostaRica1 and 1,815 (5.883 min.) from CostaRica2. The recording periods were from March 6th to 19th, 2015, for CostaRica1, and from March 6th to 20th and from April 15th to 20th, 2015, for CostaRica2. All audio files are stereo in Free Lossless Audio Codec (FLAC) format, recorded at a sampling rate of 44,100 Hz, 16 bit-depth and Pulse Code Modulation (PCM). They include sounds of insects, amphibians, rain, engines, and other commonly sounds present in natural environments in the proximity of human activity;
2. **Underwater:** the data was provided by professor Linilson R. Padovese from Polytechnic School of the University of São Paulo, Brazil. The audio files were recorded in the following two areas:
    - *Ilhéus*, southern coast of Bahia State, Brazil. There are 480 files with 15 minutes each. The recording period was from September 3rd to 4th, and from September 18th to 22nd, 2014. The dataset was initially divided into two parts: one with 200 (3.000 min.) audio files (Ilheus1) and the other (Ilheus2) with 280 (4.200 min.) files. All audio files are mono in WAVEform (WAV) audio format, recorded at sampling rate of 11,025 Hz, 16 bit-depth and PCM modulation. Sounds of humpback whales and fish choruses are predominant in these files;
    - *Laje de Santos* Marine State Park, southern coast of São Paulo State, Brazil. There are 799 (11.985 min.) audio files with 15 minutes each. The recording period was from March 17th to 18th, and from March 27th to April 3rd, 2015. All audio files are mono in WAV audio format, recorded at sampling rate of 11,025 Hz, 16 bit-depth and PCM modulation. This dataset contains sounds of fish, crustaceans and vessels.

[1] `https://www.cs.waikato.ac.nz/ml/weka/`

[2] GPS coordinates from the sensor in the old-growth forest: 10.43167528 -84.02136972. GPS sensor in secondary forest: 10.42254278 -84.01599944

### 3.4. Setup

We implemented and used routines from distinct programming languages. For a better comprehension, Figure 4 illustrates the input/output of each of them. R packages Seewave[3], Soundecology[4], and tuneR[5] were used in the tests for the generation of spectrograms and extraction of acoustic features, mp[6] package ran MDPs, and silhouette coefficient was calculated with standard R libraries.

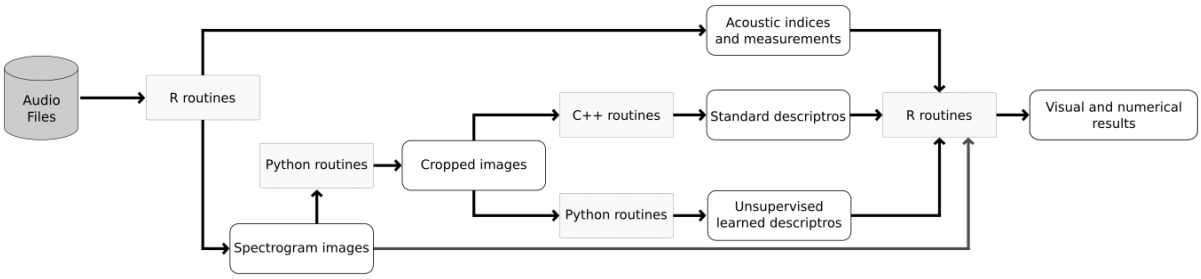



Figure 4: Input/output of routines used to achieve the steps of Figure 3.

The parameters for the extraction of features and generation of projections were the default values of the packages. Stereo recordings were converted to mono and all FLAC audio files were converted to WAV format, since the packages work well with this format to extract features. Each feature value (or set of values) represents the entire file content, independently of the recording duration (e.g., a 10-minute or a 1-minute file generates one value of an index, such as ACI). The MFCC result is a matrix with coefficients (columns) and their components (rows). Consequently, column means were calculated for representing coefficients, and twelve of them were considered in the tests. A table with rows representing each recording and columns representing the respective feature values was maintained as a CSV file.

Image spectrograms were generated for each audio file with Seewave and standard R routines for saving PNG files with 1366×768 pixels. The images were used for (i) analyzing the content of sounds by inspecting images with visual data tools, so it is important to have information about frequencies, time and sound levels in the image captions, and (ii) evaluating the feature space generated by image descriptors extracted from image spectrograms. No caption information for the task is required, so it can be discarded.

The best spectrogram resolution was achieved with Hanning window, with a length of 1024 and an overlap of 75%. We configured Seewave function to generate gray-scale spectrograms to reduce the complexity of extracting descriptors and the autoencoder input size, avoiding learning a large number of parameters. The more parameters, the more difficult it is to train a model. Figure 7 shows some images generated with colors (standard palette of Seewave package for visual presentation only) and captions. These captions were eliminated by applying cropping functions[7] available in the Pillow Python package, maintaining the time-frequency region (1110×680 pixels) for extracting image descriptors and providing inputs to the autoencoder. We could generate two images, one with captions and another without them, however, creating a complete image and discarding its caption, when necessary, requiring less processing time.

The OpenCV[8] library with C++ programming language was employed for the implementation and extraction of descriptors from gray-scale image (1110×680 pixels). The LBP descriptor was configured to utilize 8-neighborhood. The GLCM descriptor was obtained by the calculation of contrast, angular second moment, correlation and entropy for each co-occurrence matrix (0°, 45°, 90°, and 135°directions). The averages of each measure are the descriptor values. The Fourier descriptor used values generated by the radius variation (10 different radius values), and values outside the last radius were summed and employed as a last descriptor value. The ACC descriptor applied distance

[3] http://rug.mnhn.fr/seewave/
[4] http://ljvillanueva.github.io/soundecology/
[5] https://cran.r-project.org/web/packages/tuneR/index.html
[6] https://cran.r-project.org/web/packages/mp/index.html
[7] https://pillow.readthedocs.io/en/stable/
[8] https://opencv.org/

values defined in the original study (1, 3, 5, 7) [34]. The CCV threshold was assigned to 100, whereas the BIC used 8-neighborhood.

The Python programming language with Keras[9] and TensorFlow[10] libraries were used to build the autoencoder, which was implemented with the architecture shown in Figure 5. Each gray-scale image (1110×680 pixels) was split into 20 sub-images (222×170 pixels) to ensure that the autoencoder represented specific patterns (example in the supplementary material). The *pooling* layers in the encoder were configured to reduce the dimensions of entering data by half, whereas the *upsampling* layers of the decoder doubled in size. As a result, sub-images were re-scaled from 222×170 to 192×192, to ensure correct size reduction and reconstruction. This resizing process did not compromise the image resolution.

*Mean squared logarithmic error* was employed as the loss function and *Stochastic gradient descent* was the optimization method, with default learning rate. The batch size was assigned to 160, whereas the number of epochs was initialized as 30. *Selu*[11] function was applied as activation function to all convolutional layers, except the last decoder layer, which used *Sigmoid*[11] function.

We generated 4,340 images (same number of recordings) and this dataset was divided into 3 subsets: training (3,334 images), validation (833 images) and testing (173 images). The images in each subset were randomly selected, but maintaining the original proportion of labels. Approximately, each subset is comprised of 71% images from *Costa Rica*, 11% images from *Ilheus* and 18% from *Laje de Santos*.

As mentioned previously, each image was divided into 20 sub-images. Consequently, the number of images in each subset turned into 66,680 images for training, 16,660 for validation and 3,460 for testing.

Figure 5 shows that the model can extract 1,152 features from each sub-image,resulting in 23,040 features for the complete spectrogram image. Features from a sub-image can be summarized by 8 statistical measures: minimum and maximum values, mean, standard deviation, 1°, 2°and 3°quartile, and interquartile range. Consequently, a complete spectrogram can be represented by 160 summarized values.

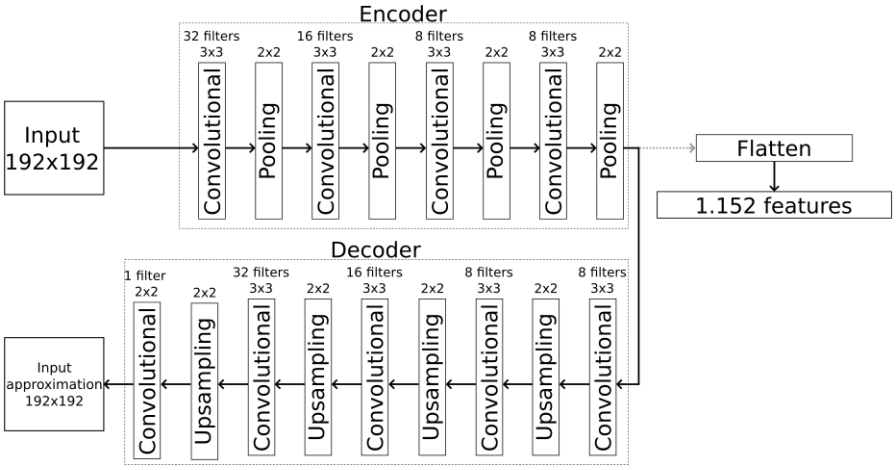



Figure 5: The autoencoder architecture.

Finally, the audio and image features were extracted by an Intel Core i7-2600, 3.40GHz, 8 cores, and 16GB RAM. The autoencoder training and testing were performed employing an NVidia Titan Xp video card.

## 4. Experimental results

This section reports the experimental results of each step of the proposed pipeline, namely extraction, visualization and evaluation of soundscape feature spaces. The datasets described in Section 3.3 were labeled as CostaRica1,

[9] https://keras.io/
[10] https://www.tensorflow.org/
[11] https://keras.io/activations/

CostaRica2, Ilheus1, Ilheus2, and Laje (5 labels), and the features addressed in Section 3.1 were generated for all datasets and divided into four groups: acoustic indices and other acoustic measurements, standard image descriptors, complete autoencoder (23,040 features per spectrogram) and summarized autoencoder (160 summarized features per spectrogram). Additionally, six MDP techniques listed in Section 3.2 were tested for the choice of the one that was best suited for supporting the task at hand. In all cases, features were normalized by the *min-max* technique, which yielded better results than those obtained with the original values and by other normalization techniques. Additional results, including an evaluation of other normalization procedures, can be found in the supplementary material.

To fairly compare the results of features and projections, we first considered the complete dataset (all data combined) including autoencoder features. However, in actual situations, a model is trained in a subset and applied to another one, as the results presented in Section 4.3. Table 1 reports the silhouette values for the complete dataset and its columns display silhouettes for original and projected spaces. First and most importantly, the feature space generated by acoustic indices and measurements consistently yielded the best silhouette, thus indicating this feature space can produce the best segregation results between the distinct environments tested.

Table 1 also shows the choice of the most discriminating MDP based on silhouette values according to the feature space. In most cases, the difference is not significant, and prompts a visual examination of the projection, highlighting group and label segregation during exploration. Figure 6 shows visual results for the same combinations of features and MDPs listed in Table 1. The examination of this figure revealed that t-SNE is visually more supportive for global discrimination, although it did not provide the best silhouette values. This is probably due to some distortion in the final distances in the projected space within particular groups performed by t-SNE. Although we have chosen t-SNE to present the remaining visual results, each MDP can highlight different aspects of the original dataset and be combined for exploration tasks.

| | **Original Normalized** | **Force** | **LAMP** | **LSP** | **MDS** | **PCA** | **t-SNE** |
|---|---|---|---|---|---|---|---|
| *Indices and Measurements* | 0.12 | 0.0545 | 0.08 | 0.1234 | **0.1313** | **0.1313** | 0.1214 |
| *Image Descriptors* | 0.09 | 0.0866 | 0.0808 | **0.111** | 0.0965 | 0.0965 | 0.0451 |
| *Complete Autoencoder* | 0.02 | 0.0679 | 0.0615 | **0.0697** | 0.0691 | 0.0691 | 0.0192 |
| *Summarized Autoencoder* | 0.01 | **0.0764** | 0.0749 | 0.0746 | 0.0577 | 0.0577 | -0.0172 |

Table 1: Silhouette coefficient values considering *Ilheus1* and *Ilheus2* as separated sets. Values are obtained from original normalized feature sets and corresponding results of different MDP. The best value in each row is highlighted. The *complete* autoencoder has all features (23,040 attributes) generated by the autoencoder technique, while *summarized* has a statistical summary of them with 160 values.

In the figure, the areas of *Costa Rica*, *Laje de Santos* and *Ilheus* are mostly separable in at least some of the feature sets and projections. CostaRica1 and CostaRica2 are less separable, nonetheless, they can be segregated for various degrees of precision, particularly employing indices and measurements as feature space and t-SNE as the visual aid for exploration.

The first tests and explorations have shown that Ilheus1 and Ilheus2 areas were not distinguishable, either numerically or visually (see Figure 6). The researchers that provided these datasets were then contacted and confirmed that no distinction was expected since Ilheus recordings were simply divided into two collections acquired by the same type of equipment not far apart from each other. Therefore, we again generated silhouettes and MDPs, making Ilheus1 and Ilheus2 a single label. Results for the modified four label dataset are shown in Table 2. The silhouette improved considerably, so that the analysis was more coherent with the contents of datasets, and the numerical analysis of the projections changed slightly.

Bellow are the results and observations for each of the feature sets tested to represent the soundscapes.

| | Original Normalized | Force | LAMP | LSP | MDS | PCA | t-SNE |
|---|---|---|---|---|---|---|---|
| *Indices and Measurements* | 0.20 | 0.1064 | 0.1571 | 0.2011 | **0.2181** | **0.2181** | 0.1953 |
| *Image Descriptors* | 0.10 | 0.1117 | 0.1341 | **0.1405** | 0.1305 | 0.1305 | 0.1103 |
| *Complete Autoencoder* | 0.08 | **0.0985** | 0.0895 | 0.0979 | 0.0965 | 0.0965 | 0.0631 |
| *Summarized Autoencoder* | 0.08 | **0.0998** | 0.0983 | 0.0982 | 0.0815 | 0.0815 | 0.0259 |

Table 2: Silhouette coefficient values considering Ilheus as **one** class. Values are obtained from original normalized feature sets and corresponding results of different MDP. The best value in each row is highlighted. The *complete* autoencoder has all features (23,040 attributes) generated by the autoencoder technique, while *summarized* has a statistical summary of them with 160 values.

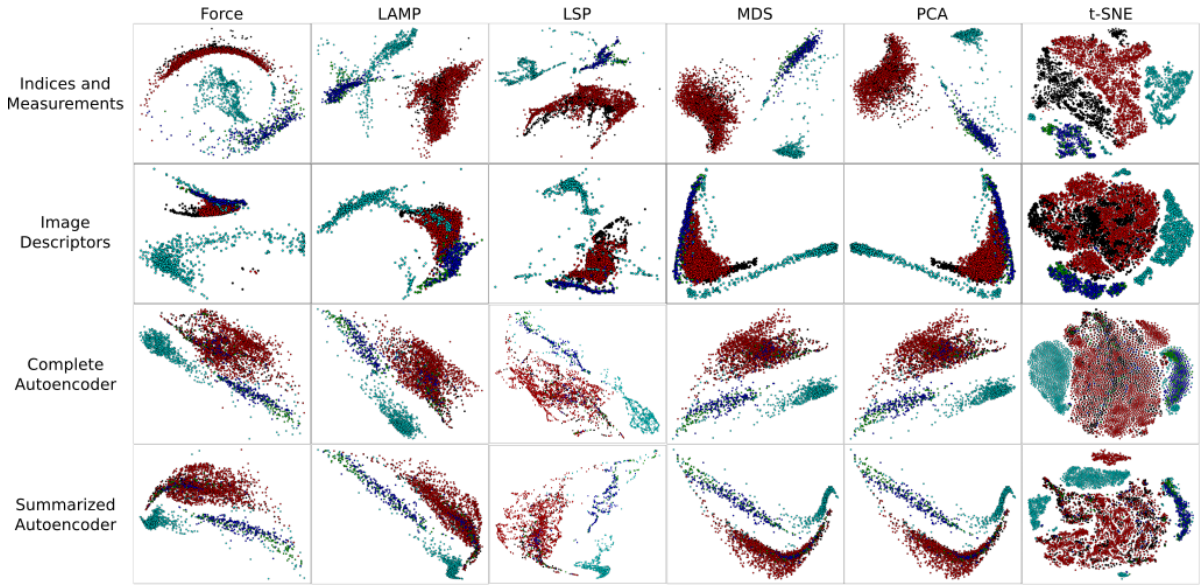



Figure 6: Visual presentation from normalized feature sets associated with MDP of all datasets described by all feature space definitions. Colors red and black represent CostaRica1 and CostaRica2, dark blue and dark green represent Ilheus1 and Ilheus2, and light blue represents Laje.

### *4.1. Acoustic indices and other acoustic measurements*

This section provides results for the feature space formed by 27 acoustic indices and measurements cited in Section 3.1.1. Each audio file minute demanded 1-minute to generate acoustic features, e.g., the process of CostaRica2 area (1,815 audio files, approximately 5,883 minutes) required 4 days.

Figure 7 shows all datasets described by this feature space, projected by t-SNE, and presents the distinction among CostaRica1, CostaRica2, Ilheus, and Laje. As aforementioned, Ilheus1 and Ilheus2 are not clearly separated. An analyst can naturally identify more than 4 groups, which is important for an exploration. However, the purpose of this step was to observe whether features could segregate previously known groups, as discussed in Section 3, for testing and justifying our approach.

Figure 7 also shows representative spectrograms in several regions of the projection. On the bottom left, the spectrogram represents a group containing audio files with low-intensity sound, whereas, on the left, it highlights cicada sound patterns, which are recurrent in most surrounding spectrograms. The sample spectrogram shown at the top indicates recordings containing fish choruses, a repeated pattern in that region, and the spectrogram on the bottom

right refers to recordings containing humpback whales and fish choruses, repeated patterns in those data.

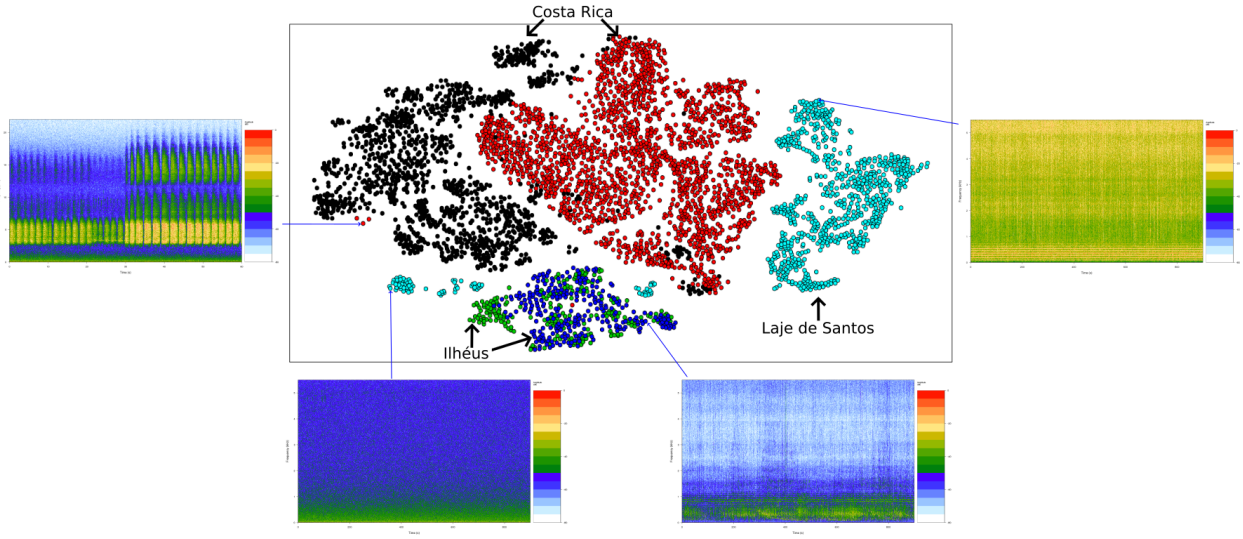



Figure 7: t-SNE projection of 4,340 audio files described by 27 acoustic indices and measurements. The spectrograms show the main content of some groups presented. Left: cicada sound pattern; top: fish chorus pattern; bottom left: low-intensity sound; bottom right: humpback whale and fish chorus sound patterns.

Figure 8 displays data for each natural area. CostaRica1 and CostaRica2 were segregated again, while two Ilheus areas were not. Laje, which represents a single label, tends to create sub-groups of files with certain patterns. The highlighted group shows a set of files with low-intensity sounds, or no sound, represented by two random samples in Figure 9.

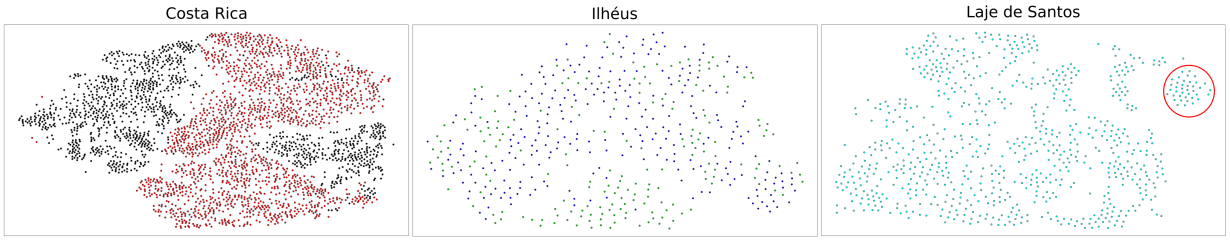



Figure 8: t-SNE projection of individual areas described by 27 acoustic indices and measurements, showing segregation between two sub-areas in *Costa Rica*, as well as *Laje* and whole *Ilheus*. The recordings of the group highlighted for *Laje de Santos* show low-intensity sound or no sound.

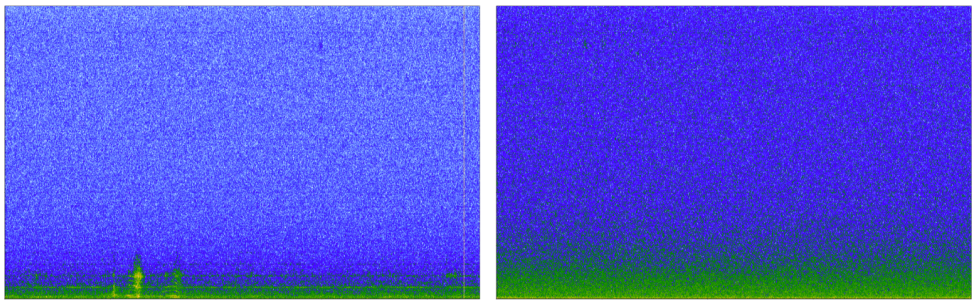


Figure 9: Examples of audio spectrograms from the *Laje de Santos* group highlighted in Figures 8, 11, 13 and 15.

An experiment identified whether sub-groups of features could produce a similar type of segregation pattern. Figures 10 and 11 show tests that used features divided into three sets: acoustic indices (ACI, ADI, AEI, BIO, Hf, Ht, H, NDSI, M, AR), MFCC (12 coefficients), and other acoustic measurements (SPL, RMS, Roughness, Rugosity, Number of Peaks).

The application of MFCC produced reasonable results regarding the separation of CostaRica1 and CostaRica2, whereas Ilheus and Laje datasets were intermingled when *other acoustic measurements* were used. As shown in Figure 11, Laje continued forming groups also formed in other feature sets, and whose recordings showed low-intensity sound or no sound (see Figure 9).

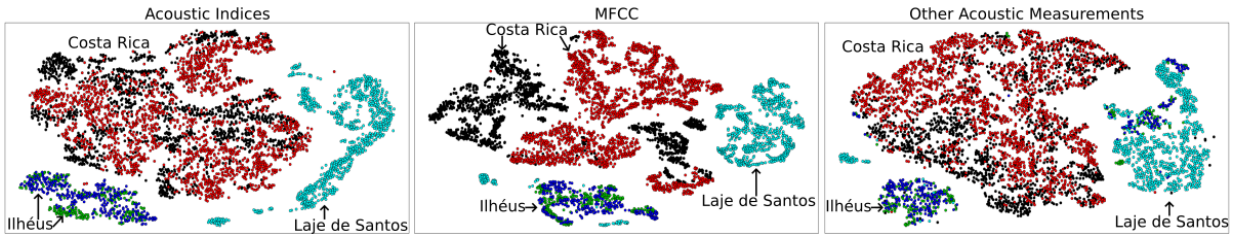



Figure 10: t-SNE projection with the use of different feature sets.

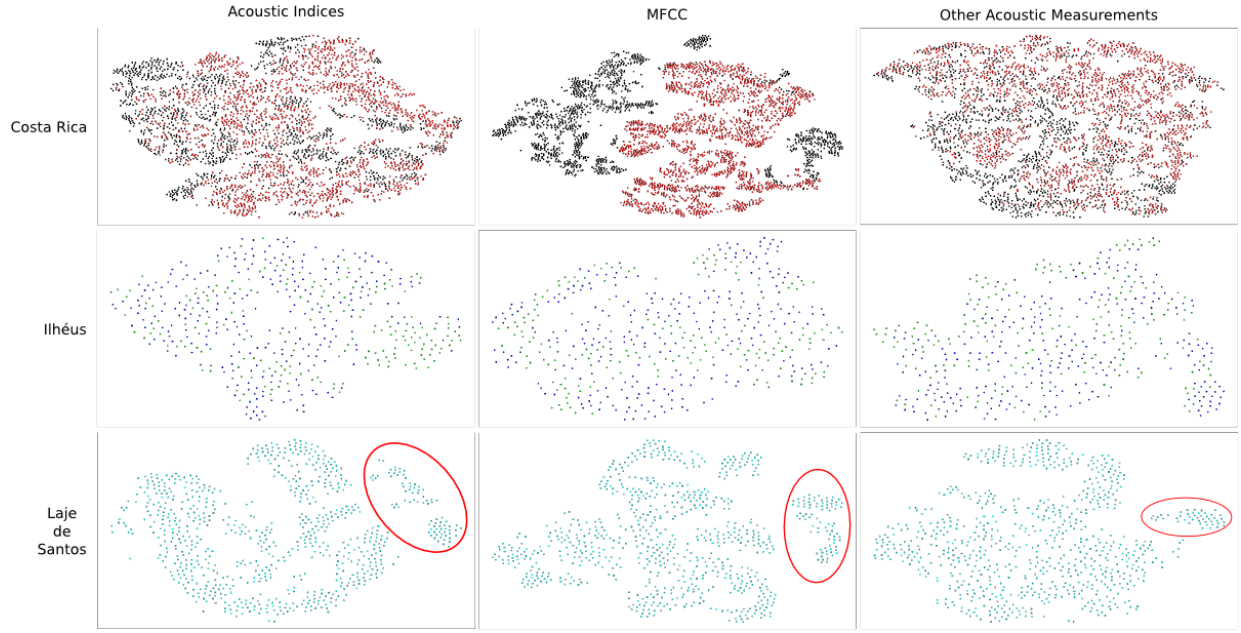



Figure 11: t-SNE projection of separated datasets with the use of different feature sets. The recordings of the group highlighted for *Laje de Santos* show low-intensity sound or no sound.

### *4.2. Standard image descriptors*

This section reports results of the same analyses, however, with the application of a feature space built upon the standard image descriptors presented in Section 2.1. Approximately 30 seconds were necessary for the generation of a spectrogram for each audio file. GLCM, LBP, Fourier, GCH, ACC, CCV, BIC descriptors were employed and generated a feature vector with 600 values. Approximately 2 seconds of processing for each image were necessary for the extraction of the descriptors, e.g., the process of CostaRica2 area (1,815 images) required around 1 hour. Figure 12 shows the result of a data projection that used image descriptors and the separation among *Costa Rica* areas, *Ilheus*

and *Laje*. However, the segregation between CostaRica1 and CostaRica2 is not as clear as the one achieved with acoustic features.

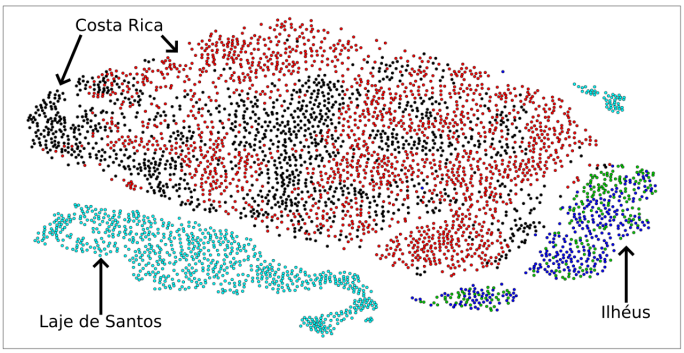



Figure 12: t-SNE projection of 4,340 audio files described by 600 image descriptor values.

Figure 13 displays collection areas separately. CostaRica1 and CostaRica2 do not segregate in general, only in small groups within some areas, and Ilheus1 and Ilheus2 do not segregate again. Laje continues to show a group with files containing low-intensity sound (see Figure 9).

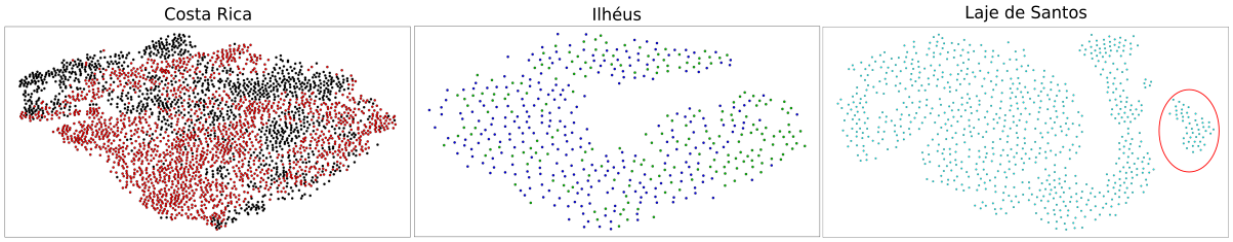



Figure 13: t-SNE projection of each sound dataset described by 600 image descriptor values. The recordings of the group highlighted for *Laje de Santos* show low-intensity sound or no sound.

Figures 14 and 15 show tests that used characteristics divided into two groups: texture (GLCM, LBP, Fourier) and color (GCH, ACC, CCV, BIC) descriptors. Color descriptors can distinguish CostaRica1 from CostaRica2 better than texture descriptors, however, Ilheus1 and Ilheus2 are not segregated by any features group. Low-intensity sounds are still segregated in the Laje dataset.

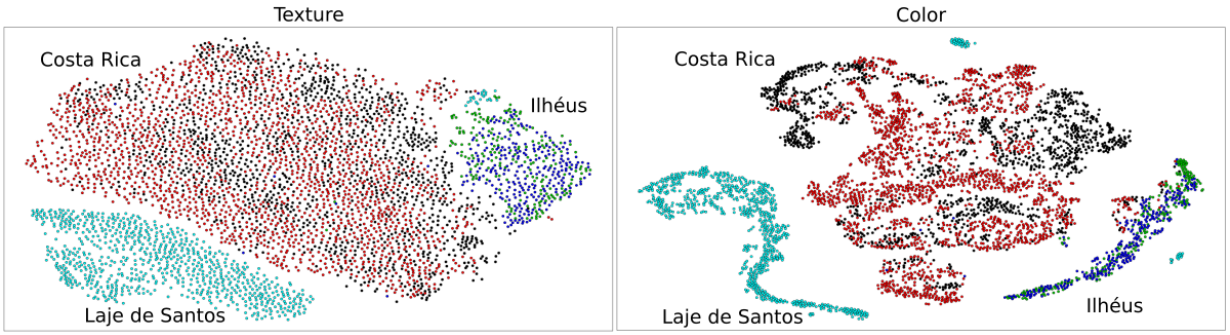



Figure 14: t-SNE projection with the use of different feature sets.

### *4.3. Unsupervised learned descriptors*

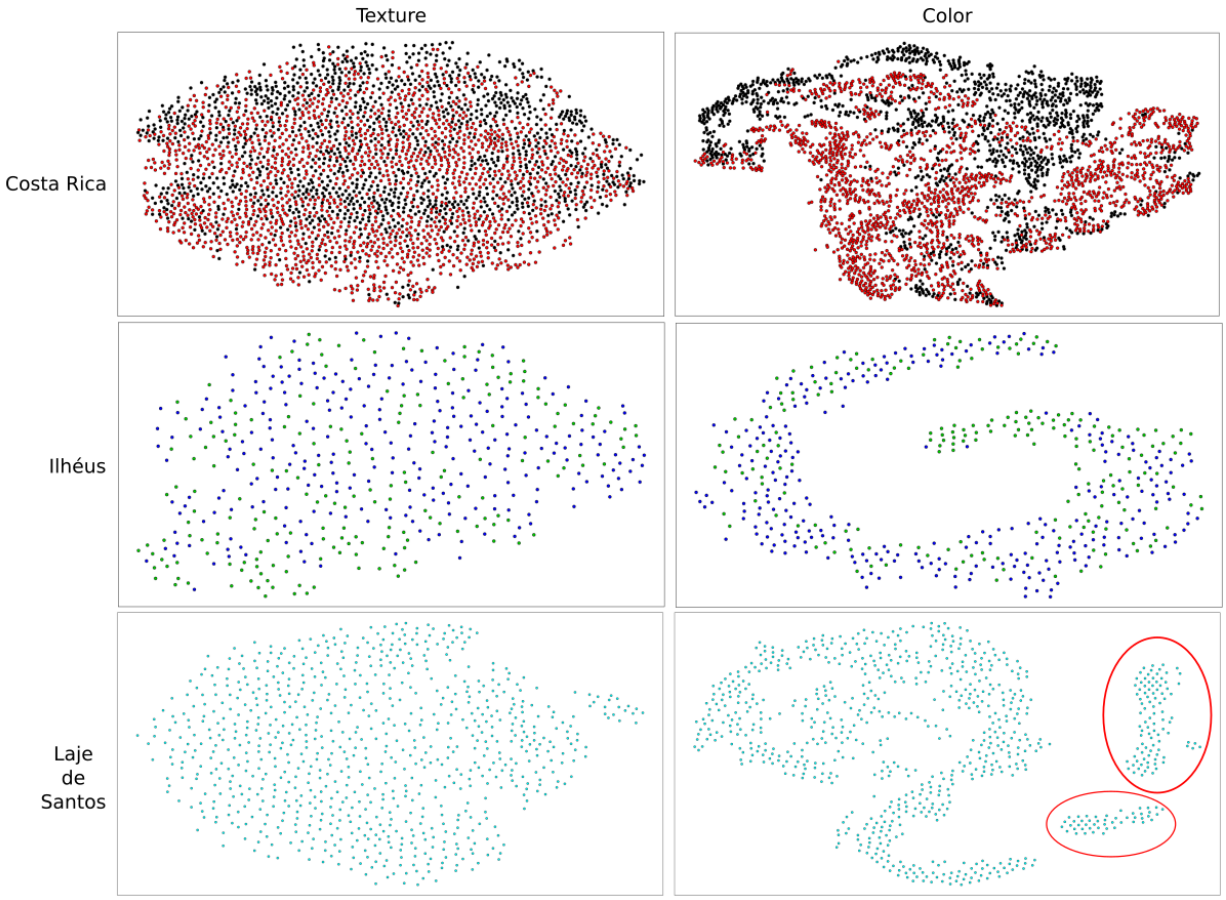



Figure 15: t-SNE projection of separated datasets using different feature sets. The recordings of the group highlighted for *Laje de Santos* show low-intensity sound or no sound.

This section presents and analyzes the results for the feature space learned with the autoencoder model described in Section 3.4. It was tested as an alternative technique for the "manual" definition and extraction of significant features from spectrogram images. The loss function values, both for training and validation, started near 0.01 and tended to 0 throughout the training process, which required 1 hour and a half.

The trained model was applied for the extraction of features from the test images (173 spectrogram images, 3,460 sub-images), and required approximately 6 seconds. The same process was applied to all images (training, validation and testing) to present results in Figure 6 and Tables 1 and 2, that is, 86,800 sub-images were processed and it took 3 minutes for the feature extraction.

Figure 16 depicts a projection that used features generated by the autoencoder for test images. A proper segregation was achieved for *Laje* points, but *Costa Rica* and *Ilheus* were not clearly separated. On the other hand, if features from all datasets are considered, as in Figure 6, the projection with the use of autoencoder features is similar to that of image descriptors, i.e., an adequate segregation of terrestrial and underwater data.

Silhouette values of the test dataset (173 spectrogram images), described by previous feature sets, were generated for comparisons with autoencoder features. Table 3 shows values calculated for Ilheus1 and Ilheus2 as separate labels (column 5 classes) and a single label (column 4 classes). The results are consistent with those from tests for the full dataset, i.e., a significant difference was observed among the results of different feature spaces, regarding segregation capacity. The best silhouette values were also reached with color descriptors and in a higher degree with MFCC.

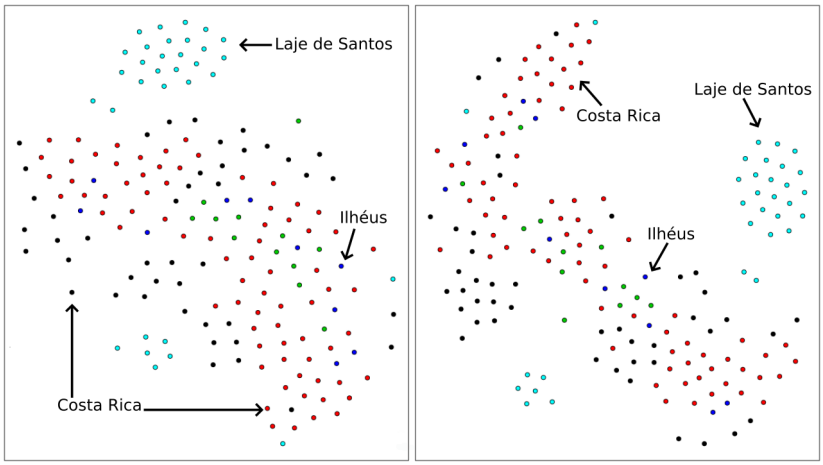



Figure 16: t-SNE projection of test subset with 173 spectrograms. The images show projection using 23,040 features (left) and their summarization (right), both normalized.

| **Features** | **5 classes** | **4 classes** |
|---|---|---|
| *Indices and Measurements* | 0.10 | **0.17** |
| *Image Descriptors* | 0.06 | 0.07 |
| *Complete Autoencoder* | -0.06 | 0.00 |
| *Summarized Autoencoder* | -0.08 | -0.02 |
| *MFCC* | 0.12 | **0.21** |
| *Color descriptors* | 0.06 | 0.07 |

Table 3: Silhouette coefficient values with distinct features sets, generated with the test sub-set (173 samples) used in the autoencoder tests. The feature values were normalized between [0, 1]. Values considered Ilheus1 and Ilheus2 separate labels (column 5 classes) and a single label (column 4 classes). The best silhouette values are highlighted.

## 5. Discussion

In this work we have applied visual and numerical approaches to evaluate feature sets composed of acoustic features and image descriptors, manually and automatically extracted from spectrogram images, with the goals of segregation of soundscapes from different areas as well as exploration of sub-sets of audio patterns. Different patterns emerged from the recordings employed such that the features achieved success in visually and numerically segregating underwater from terrestrial recordings.

From the different sets of features employed, acoustic features performed best and were also sensitive to different types of environments in contiguous terrestrial collection areas (primary and secondary forest vegetation in CostaRica1 and Costarica2). The other features attained relatively less success in such a refined segregation.

According to the experiments, MFCC features best segregated terrestrial areas (CostaRica1 and CostaRica2). Although they are most frequently applied for speech recognition and classification of specific natural sounds, our tests revealed their good performance in differentiating soundscape environments. Color descriptors for spectrogram images also segregated *Costa Rica* data, but at a lower degree of efficacy. Our setting for autoencoder did not discriminate terrestrial regions well on either global or local scales. In fact, the separation of those terrestrial areas is a good indicator of what recordings from heterogeneous regions a feature space can handle in a single framework.

Additionally features and projections indicated similarities between *Ilheus* and *Laje* soundscapes, due to the presence of fish sounds. Our processes also reflected area differences, due to a large presence of whale sounds in *Ilheus*, and crustaceans and vessels in *Laje*. All feature sets segregated these areas and separated them from terrestrial areas.

Regarding the features generated by the autoencoder, despite the unsatisfactory results in the current setup, the architecture can be improved to achieve a better feature space. Our future research will explore other inputs (spectro-

gram matrix, spectrum, or even raw audio), type of layers (recurrent layers for analysis of time context information), layer layout (sequential or parallel) and other network parameters.

Acoustic indices required longer processing time in comparison to all other features (days were necessary for a sequential processing). Image descriptors took approximately 1 hour to process, and the autoencoder required 2 hours to build the model and some minutes to extract features. This panorama suggests the applications can use two different approaches regarding feature spaces. The first relies on faster processes and compromises precision to some degree, and the second takes advantage of data independence, in the case of acoustic indices and measurements, and calculates large datasets in parallel. Therefore, each sound file can be processed independently, which would speed up feature extraction. Indeed, differences between machines that processed autoencoder and the other feature sets, as well as the type of implementations performed must be considered. Moreover, a parallel extraction of acoustic features can facilitate the use of recording equipment with processing power that calculate indices as data are collected.

Regardless of the discussion on whether indices are a proper tool for analysis of soundscapes (see [8, 15, 17, 18, 67]) or pose significant problems (see [23, 68, 69]), they performed very well when used together in our scenario, which reinforces previous research (e.g., Brown et al. [70] and Phillips et al. [11]).

Many of our analyses, including some reported here, have revealed that the best feature space in segregating areas can also group similar patterns in each area, i.e., recordings with similar sound patterns are mapped near each other. As a result, samples can be collected from a same area in different periods and labeled with period, and a set of suitable features can be extracted from them. The projection of these data can show segregation among samples, which highlights differences among distinct periods and enables visual explorations by experts for finding further patterns. This type of analysis will be performed in our future studies.

It is worth reminding that the acoustic features and spectrogram images were generated with recordings with different duration (see Section 3.4) and standardizing file duration is a better approach. Nonetheless, even with distinct file lengths, all features could represent audio file similarities at different levels and achieve reasonable results. Standardizing duration probably generates better representations and can improve the results of all feature spaces, due to the higher resolution of the acoustic pattern description. As a future work, we intend to carry out tests with this pre-processing step.

The examples presented here confirmed that InfoVis (mainly projections) techniques can support the exploration and analysis of soundscapes, as stated by Phillips et al. [11], Reis et al. [27] and others. Our visual results were confirmed with numerical values, and visualizations based on MDP revealed several patterns within soundscapes. Visualizations can efficiently explore grouping on larger and smaller scales, and on complete or sampled datasets for the choice of feature spaces.

In order to choose a good feature set, additional tests were conducted through an automatic feature selection with three known algorithms (*Correlation-based*, *Information Gain* and *Relief-F*) to compare their results with those of this study (see supplementary material). Although the silhouette results were slightly better in some cases, the difference was not significant and the use of automatically reduced feature sets did not offer any advantage in the MDP results. However, this is a feasible solution for data reduction in the processing, exploration and visualization of larger datasets.

The set of techniques employed in this work can open up several possibilities for the exploration of soundscapes. In addition to comparisons of distinct areas, temporal changes in acoustic patterns can appear in a global analysis if a representation is split into new partitions of the dataset. These partitions can provide additional information about various events, such as biodiversity loss or gain, increase in human-related sounds and detection of invasive species. Our future studies on computational methods in natural acoustic environments aim to adapt the current approach to those and other pressing problems.

## 6. Conclusions and future work

This paper addressed a novel exploration and evaluation process that applies the MDP as a visual aid (with a numerical silhouette coefficient support) to validate and select relevant features that represent acoustic data. The process allows analyses and comparisons of feature sets regarding the grouping of similar audio files that describe an environment, and showed evidence of global and local segregation, that is, groups of similar recordings within distinct soundscapes. The methods adopted here are feasible for assessment and correlation studies of acoustic domains and

scenarios in terrestrial and underwater environments. To the best of our knowledge, this is one of the first studies that apply MDP to analyze feature spaces in the soundscape context. Its applicability ranges from distinguishing historical data or monitoring changes in the environment to finding features for identifying specific events of scenarios in audio sets.

Acoustic indices and measurements performed better than a manual and automatic image descriptor for the setups presented. However, some image features also attained satisfactory results and are faster for calculations in most computer configurations. Finally, the MFCC were more sensitive to segregation tasks.

This research imposed the following limitations: (i) the results that compare different feature spaces must be confirmed by additional datasets (from distinct and more diverse regions) and other data mining tasks, such as clustering and classification, regarding the advantages of feature space analyses for future automatic interpretations. Therefore, specific questions, for instance, about groups to be found, or more specific goals for the analyses must be established, and (ii) the processing times for the values calculated in our best feature space are prohibitive for very large datasets and real-time applications, since they were calculated sequentially. However, the same features can be calculated independently for each sound file, which suggests a setup for large-scale processing based on parallel architectures or more processing power at the collection stations. Associated with these improvements, audio files will be pre-processed to standardize their duration and provide a higher resolution of patterns.

In addition, more in-depth investigation will be performed to verify the suitability of the proposed approach to highlight temporal changes in soundscape patterns. Further studies on Deep Learning strategies for the same data types are also necessary. Although we tried some different setups for the approach, they were not as successful as the other alternatives. We believe that, with another architecture, or in a more specific scenario, this technology can offer advantages over other types of extraction. Other specific studies are, therefore, required, since the proposed method has been applied for the categorization of acoustic or other data, such as image and video, achieving satisfactory results.

## Acknowledgements

This study was partially financed by the Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001, National Council for Scientific and Technological Development (CNPq) and São Paulo Research Foundation (FAPESP). The authors would like to thank professors Linilson R. Padovese, from Polytechnic School of the University of São Paulo, Brazil, and Bryan C. Pijanowski, from Purdue University, Indiana, USA, for their data and useful feedback, and Angela C. P. Giampedro, from the University of São Paulo, for her invaluable help with English language review.

## References

[1] B. C. Pijanowski, A. Farina, S. H. Gage, S. L. Dumyahn, B. L. Krause, What is soundscape ecology? An introduction and overview of an emerging new science, Landscape Ecology 26 (2011) 1213–1232.

[2] B. C. Pijanowski, L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, S. H. Gage, N. Pieretti, Soundscape ecology: the science of sound in the landscape, BioScience 61 (2011) 23–216.

[3] B. Krause, Bioacoustics, habitat ambience in ecological balance, Whole Earth Review 57 (1987) 14–18.

[4] K. Servick, Eavesdropping on Ecosystems, Science (New York, N.Y.) 343 (2014) 834–837.

[5] W. Joo, S. H. Gage, E. P. Kasten, Analysis and interpretation of variability in soundscapes along an urban-rural gradient, Landscape and Urban Planning 103 (2011) 259–276.

[6] N. Pieretti, A. Farina, D. Morri, A new methodology to infer the singing activity of an avian community: The acoustic complexity index (aci), Ecological Indicators 11 (2011) 868–873.

[7] M. Depraetere, S. Pavoine, F. Jiguet, A. Gasc, S. Duvail, J. Sueur, Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland, Ecological Indicators 13 (2012) 46–54.

[8] S. E. Parks, J. L. Miksis-Olds, S. L. Denes, Assessing marine ecosystem acoustic diversity across ocean basins, Ecological Informatics 21 (2014) 81–88.

[9] R. Righini, G. Pavan, A soundscape assessment of the sasso fratino integral nature reserve in the central apennines, italy, Biodiversity 21 (2020) 4–14.

[10] M. Harvey, Acoustic Detection of Humpback Whales Using a Convolutional Neural Network, 2018. URL: `https://ai.googleblog.com/2018/10/acoustic-detection-of-humpback-whales.html`.

[11] Y. F. Phillips, M. Towsey, P. Roe, Revealing the ecological content of long-duration audio-recordings of the environment through clustering and visualisation, PloS One 13 (2018) e0193345.

[12] M. Sankupellay, M. Towsey, A. Truskinger, P. Roe, Visual fingerprints of the acoustic environment: The use of acoustic indices to characterise natural habitats, in: Big Data Visual Analytics (BDVA), 2015, IEEE, 2015, pp. 1–8.
[13] M. Towsey, L. Zhang, M. Cottman-Fields, J. Wimmer, J. Zhang, P. Roe, Visualization of long-duration acoustic recordings of the environment, Procedia Computer Science 29 (2014) 703–712.
[14] J. Sueur, A. Farina, A. Gasc, N. Pieretti, S. Pavoine, Acoustic indices for biodiversity assessment and landscape investigation, Acta Acustica United with Acustica 100 (2014) 772–781.
[15] L. Jost, Entropy and diversity, Oikos 113 (2006) 363–375.
[16] A. Eldridge, M. Casey, P. Moscoso, M. Peck, A new method for ecoacoustics? Toward the extraction and evaluation of ecologically-meaningful soundscape components using sparse coding methods, PeerJ 4 (2016) e2108.
[17] A. Gasc, J. Sueur, F. Jiguet, V. Devictor, P. Grandcolas, C. Burrow, M. Depraetere, S. Pavoine, Assessing biodiversity with sound: Do acoustic diversity indices reflect phylogenetic and functional diversities of bird communities?, Ecological Indicators 25 (2013) 279–287.
[18] S. Fuller, A. C. Axel, D. Tucker, S. H. Gage, Connecting soundscape to landscape: Which acoustic index best describes landscape configuration?, Ecological Indicators 58 (2015) 207–215.
[19] T. Bradfer-Lawrence, N. Gardner, L. Bunnefeld, N. Bunnefeld, S. G. Willis, D. H. Dent, Guidelines for the use of acoustic indices in environmental research, Methods in Ecology and Evolution 10 (2019) 1796–1807.
[20] S. Dröge, D. A. Martin, R. Andriafanomezantsoa, Z. Burivalova, T. R. Fulgence, K. Osen, E. Rakotomalala, D. Schwab, A. Wurz, T. Richter, et al., Listening to a changing landscape: Acoustic indices reflect bird species richness and plot-scale vegetation structure across different land-use types in north-eastern madagascar, Ecological Indicators 120 (2021) 106929.
[21] S. L. Mitchell, J. E. Bicknell, D. P. Edwards, N. J. Deere, H. Bernard, Z. G. Davies, M. J. Struebig, Spatial replication and habitat context matters for assessments of tropical biodiversity using acoustic indices, Ecological Indicators 119 (2020) 106717.
[22] D. Stowell, M. D. Plumbley, Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning, PeerJ 2 (2014) e488.
[23] I. Sánchez-Gendriz, L. Padovese, Underwater soundscape of marine protected areas in the south Brazilian coast, Marine Pollution Bulletin 105 (2016) 65–72.
[24] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, M. G. Betts, Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach., The Journal of the Acoustical Society of America 131 (2012) 4640–4650.
[25] X. Dong, M. Towsey, J. Zhang, P. Roe, Compact features for birdcall retrieval from environmental acoustic recordings, in: Proceedings of the 2015 IEEE 15th International Conference on Data Mining Workshops, IEEE Computer Society, 2015, pp. 1–6.
[26] J. Xie, M. Towsey, J. Zhang, X. Dong, P. Roe, Application of image processing techniques for frog call classification, 2015 IEEE International Conference on Image Processing (ICIP) (2015) 4190–4194.
[27] C. D. G. Reis, T. N. Santos, et al., A visualization framework for feature investigation in soundscape recordings, in: 2018 22nd International Conference Information Visualisation (IV), IEEE, 2018, pp. 490–497.
[28] E. Znidersic, M. Towsey, W. K. Roy, S. E. Darling, A. Truskinger, P. Roe, D. M. Watson, Using visualization and machine learning methods to monitor low detectability species—the least bittern as a case study, Ecological Informatics 55 (2020) 101014.
[29] F. V. Paulovich, Mapeamento de dados multi-dimensionais-integrando mineração e visualização, Ph.D. thesis, Universidade de São Paulo, 2008.
[30] D. B. Coimbra, Multidimensional projections for the visual exploration of multimedia data, Ph.D. thesis, Universidade de São Paulo, 2016.
[31] L. G. Nonato, M. Aupetit, Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment, IEEE Transactions on Visualization and Computer Graphics 25 (2018) 2650–2673.
[32] O. A. B. Penatti, et al., Estudo comparativo de descritores para recuperação de imagens por conteudo na web, Master's thesis, Universidade Estadual de Campinas - Unicamp, 2009.
[33] M. J. Swain, D. H. Ballard, Color indexing, International Journal of Computer Vision 7 (1991) 11–32.
[34] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, R. Zabih, Image indexing using color correlograms, in: Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, IEEE, 1997, pp. 762–768.
[35] G. Pass, R. Zabih, J. Miller, Comparing images using color coherence vectors, in: Proceedings of the fourth ACM International Conference on Multimedia, ACM, 1997, pp. 65–73.
[36] R. O. Stehling, M. A. Nascimento, A. X. Falcão, A compact and efficient image retrieval approach based on border/interior pixel classification, in: Proceedings of the eleventh international conference on Information and knowledge management, ACM, 2002, pp. 102–109.
[37] R. M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man, and Cybernetics 3 (1973) 610–621.
[38] R. C. Gonzalez, R. E. Woods, Digital Image Processing, 3rd ed., Pearson Education, 2011.
[39] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 971–987.
[40] S. Shalev-Shwartz, S. Ben-David, Understanding machine learning: From theory to algorithms, Cambridge University Press, 2014.
[41] A. Gulli, S. Pal, Deep Learning with Keras, Packt Publishing, 2017.
[42] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, volume 1, MIT Press Cambridge, 2016.
[43] E. H. Chi, A framework for Visualization Information, Springer, 2002.
[44] A. C. Telea, Data visualization: principles and practice, 2 ed., CRC Press, 2014.
[45] M. O. Ward, G. Grinstein, D. Keim, Interactive data visualizaton: foundations, techniques, and applications, 2 ed., CRC Press, 2015.
[46] K. Pearson, On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (1901) 559–572.
[47] H. Hotelling, Analysis of a complex of statistical variables into principal components., Journal of Educational Psychology 24 (1933) 417.
[48] J. B. Kruskal, M. Wish, Multidimensional scaling, volume 11, Sage, 1978.
[49] E. Tejada, R. Minghim, L. G. Nonato, On improved projection techniques to support visual exploration of multi-dimensional data sets, Information Visualization 2 (2003) 218–231.

[50] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605.
[51] F. V. Paulovich, L. G. Nonato, R. Minghim, H. Levkowitz, Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping, IEEE Transactions on Visualization and Computer Graphics 14 (2008) 564–575.
[52] P. Joia, D. Coimbra, J. A. Cuminato, F. V. Paulovich, L. G. Nonato, Local affine multidimensional projection, IEEE Transactions on Visualization and Computer Graphics 17 (2011) 2563–2571.
[53] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to data mining. 1st, 2005.
[54] W. E. Marcilio, D. M. Eler, R. E. Garcia, An approach to perform local analysis on multidimensional projection, in: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, 2017, pp. 351–358.
[55] N. T. Boelman, G. P. Asner, P. J. Hart, R. E. Martin, Multi-trophic invasion resistance in hawaii: bioacoustics, field surveys, and airborne remote sensing, Ecological Applications 17 (2007) 2137–2144.
[56] J. Sueur, S. Pavoine, O. Hamerlynck, S. Duvail, Rapid acoustic survey for biodiversity appraisal, PloS One 3 (2008) e4065.
[57] L. Villanueva-Rivera, B. Pijanowski, j. Doucette, B. Pekin, A primer of acoustic analysis for landscape ecologists, Landscape Ecology 26 (2011) 1233–1246.
[58] E. P. Kasten, S. H. Gage, J. Fox, W. Joo, The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology, Ecological Informatics 12 (2012) 50–67.
[59] B. Pekin, J. Jung, L. Villanueva-Rivera, B. Pijanowski, J. Ahumada, Modeling acoustic diversity using soundscape recordings and lidar-derived metrics of vertical forest structure in aneotropical rainforest, Landscape Ecology 27 (2012) 1513–1522.
[60] J. O. Ramsay, Functional data analysis, Wiley Online Library, 2006.
[61] D. A. Mezquida, J. L. Martínez, Platform for bee-hives monitoring based on sound analysis. a perpetual warehouse for swarm apos; s daily activity, Spanish Journal of Agricultural Research 7 (2009) 824–828.
[62] M. Lutter, Mel-Frequency Cepstral Coefficients, 2014. URL: `http://recognize-speech.com/feature-extraction/mfcc`.
[63] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, J. P. Bello, Robust sound event detection in bioacoustic sensor networks, PloS one 14 (2019) e0214168.
[64] M. Thomas, B. Martin, K. Kowarski, B. Gaudet, S. Matwin, Marine mammal species classification using convolutional neural networks and a novel acoustic representation, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 290–305.
[65] O. S. Kirsebom, F. Frazao, Y. Simard, N. Roy, S. Matwin, S. Giard, Performance of a deep neural network at detecting north atlantic right whale upcalls, The Journal of the Acoustical Society of America 147 (2020) 2636–2646.
[66] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, in: Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on, IEEE, 2015, pp. 1200–1205.
[67] L. Lellouch, S. Pavoine, F. Jiguet, H. Glotin, J. Sueur, Monitoring temporal change of bird communities with dissimilarity acoustic indices, Methods in Ecology and Evolution 5 (2014) 495–505.
[68] S. E. Freeman, F. L. Rohwer, G. L. D'Spain, A. M. Friedlander, A. K. Gregg, S. a. Sandin, M. J. Buckingham, The origins of ambient biological sound from coral reef ecosystems in the Line Islands archipelago, The Journal of the Acoustical Society of America 135 (2014) 1775–1788.
[69] I. Sánchez-Gendriz, L. R. Padovese, A methodology for analyzing biological choruses from long-term passive acoustic monitoring in natural areas, Ecological Informatics 41 (2017) 1–10.
[70] A. Brown, S. Garg, J. Montgomery, Automatic rain and cicada chorus filtering of bird acoustic data, Applied Soft Computing 81 (2019) 105501.