PLOS ONE

RESEARCH ARTICLE

A computational study on outliers in world music

Maria Panteli*, Emmanouil Benetos, Simon Dixon

Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

* m.panteli@qmul.ac.uk

OPEN ACCESS

Citation: Panteli M, Benetos E, Dixon S (2017) A computational study on outliers in world music. PLoS ONE 12(12): e0189399. https://doi.org/10.1371/journal.pone.0189399

Editor: Chun-Hsi Huang, University of Connecticut, UNITED STATES

Received: May 17, 2017

Accepted: November 26, 2017

Published: December 18, 2017

Data Availability Statement: All code is available from the public repository github.com/mpanteli/music-outliers.

Funding: EB is supported by a RAEng Research Fellowship (RF/128) from the Royal Academy of Engineering (http://raeng.org.uk/). MP is supported by a Principal's research studentship from Queen Mary University of London (http://qmul.ac.uk). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The comparative analysis of world music cultures has been the focus of several ethnomusicological studies in the last century. With the advances of Music Information Retrieval and the increased accessibility of sound archives, large-scale analysis of world music with computational tools is today feasible. We investigate music similarity in a corpus of 8200 recordings of folk and traditional music from 137 countries around the world. In particular, we aim to identify music recordings that are most distinct compared to the rest of our corpus. We refer to these recordings as 'outliers'. We use signal processing tools to extract music information from audio recordings, data mining to quantify similarity and detect outliers, and spatial statistics to account for geographical correlation. Our findings suggest that Botswana is the country with the most distinct recordings in the corpus and China is the country with the most distinct recordings when considering spatial correlation. Our analysis includes a comparison of musical attributes and styles that contribute to the 'uniqueness' of the music of each country.

Introduction

With the increasing accessibility of large sound archives and advances in Music Information Retrieval (MIR) technologies [1] it is possible to automatically analyse vast amounts of sound recordings. This has been the target of several MIR studies, usually with a two-fold scope: first, the development of technology for the analysis of music audio, and second, the application of technology to study musical phenomena. While the development of MIR technologies has been advancing, few studies have attempted to apply it to the analysis of large corpora of folk and traditional music. We are interested in a large-scale comparison of world music with particular focus on music similarity and distinctiveness.

In the field of ethnomusicology, several studies have considered the comparison of world music cultures [2, 3]. Data collection and annotation for this type of research is usually done manually by ethnomusicologists, a process which limits the potential for large-scale results. In the field of MIR, large-scale comparative studies have focused mainly on Eurogenetic music [4, 5], where Eurogenetic defines music styles of mainly Western traditions for example classical and popular repertoires. The study of non-Eurogenetic music using computational tools

falls under the emerging field of Computational Ethnomusicology [6, 7]. While several research projects have focused on the development of MIR tools for world music analysis [8–12], no study, to the best of our knowledge, has applied such computational methods in the analysis of a large world music corpus.

Music similarity lies at the heart of most MIR applications, such as music classification, retrieval and recommendation [1]. In this study, we focus on music dissimilarity or musical distinctiveness. In particular we aim to detect music outliers. Outlier detection is a common pre-processing step in the analysis of big data collections [13]. In music, outlier detection can reveal recordings with outstanding musical characteristics. Tracing the geographic origin of these recordings could help identify areas of the world that have developed a unique musical character. Due to the long-lasting traditions of orally-transmitted repertoires and the lack of scores or consistent notation in world music, our music data is extracted solely from the audio. Music similarity/dissimilarity in this case is modelled by considering musical attributes captured in the audio signal.

In previous work we have explored the suitability of audio features for music similarity and content description [14]. Audio features for the purpose of studying world music need to be agnostic to style characteristics so that they can generalise to the diversity of music styles. We found rhythmic and melodic descriptors that are invariant to tempo and pitch transformations and are fairly robust to transformations of the recording quality. We used these features in combination with feature learning to assess music similarity in a relatively small world music corpus [15] as well as to detect and analyse music outliers in a preliminary study [16].

In this study we expand prior work to world music analysis using a larger corpus and evaluating additional methods. We use signal processing tools to process audio data from a collection of recorded world music. Machine learning and data embeddings are used to learn a feature space of music similarity. Data mining techniques are applied to detect outliers in this space. Results are evaluated quantitatively using metrics to assess classification accuracy and qualitatively via visualisation of the space and listening to audio examples. Our observations on music similarity comply with expected geographical and cultural links whereas outliers provide insights on the evolution of world music. This is the first study to investigate outliers in world music with such a large scale. Our developments contribute to defining concepts and methods from which future work in the study of large world music corpora can benefit.

This paper is organised as follows. The Related work section provides a literature review of related studies and methods. The Methodology section describes the materials and tools used in this study. It focuses on details of the music corpus under investigation, audio feature extraction and feature learning methods for music similarity, and data mining techniques to assess music similarity and distinctiveness as well as methods for modelling spatial relations. Results are presented in the Results section and limitations of the study as well as directions for future improvement are considered in the Discussion section. Findings are summarised in the Conclusion section.

Related work

Comparison of world music cultures

The comparison of world music cultures has been the topic of several ethnomusicological studies since the beginning of the 20th century [2, 3, 17, 18]. Alan Lomax, one of the major comparativists, made more than 4000 recordings from around the world and annotated their performance-style characteristics based on the system of 'Cantometrics' [2, 17]. Using a phylogenetic analysis, he formed the hypothesis that there are two music evolutionary roots, the eastern Asian and the Sub-Saharan African music cultures from which all other music styles

have possibly evolved [17]. In a similar manner, Savage et al. [3] analyse 304 recordings from the Garland Encyclopedia of Music [19] using the annotation system of 'Cantocore' [20] in addition to the Cantometrics descriptors. In this study, Savage et al. show that there are no 'absolute' music universals, i.e., music properties that are shared amongst all music of the world without exceptions, but rather 'statistical' universals, i.e., properties that occur with exceptions but are statistically consistent in music from around the world. This supports the hypothesis of the current study, that there are outliers, pieces outside the statistical norms shared by much of the world's music.

Applications of comparative musicology have also focused on contrasting music styles to genetic and language evolution [3, 18, 21–23]. The study of 220 traditional songs from 9 indigenous populations from Taiwan [18] showed that population structure for genetics exhibits stronger parallels to music than to language. The study of 700 recordings from 58 patrimonies of rural areas in Gabon [23] found that there is a predominant vertical transmission of musical characteristics such as metre, rhythm, and melody, where vertical transmission refers to the inheritance from ancestors in contrast to the horizontal exchange between neighbours.

Large-scale music corpus analysis

Computational approaches to music analysis enable the study of larger music corpora. Large-scale MIR studies have focused on the analysis of popular (mainly Eurogenetic) music [4, 5, 24]. For example, Serra et al. [4] analysed pitch, loudness and timbre characteristics in 464411 recordings of contemporary Western popular music between 1955–2010 and found that over the years music shows less variety in pitch transitions, consistent homogenisation of the timbral palette, and louder and potentially poorer volume dynamics. A related study of 24941 Western popular music recordings between 1922–2010 showed that the most influential songs were more innovative during the early 1970s and the mid 1990s [24]. Mauch et al. [5] analysed 17094 songs from the US Billboard Hot 100 between 1960–2010 and found that pop music evolved with particular rapidity during three stylistic 'revolutions'; around 1964, 1983 and 1991. Other corpus analysis studies have focused on the automatic classification of music by genre [25–27] via the combination of different audio features.

Fewer studies have considered the computational analysis of non-Western music corpora [12, 28]. Moelants et al. [12] analysed pitch distributions of 901 recordings from Central Africa and found that recent recordings exhibit Western-influenced scales. Gómez et al. [28] studied aspects of timbre, rhythm, and tonality in 5905 recordings from Western and non-Western music styles and showed that Western music is more equal-tempered than non-Western music. A comparison between music features and geographical latitude and longitude showed that latitude is mostly associated with tonal features whereas longitude with rhythmic ones. A number of studies have considered automatic classification of non-Western music styles. Liu et al. [29] classify 1300 music recordings into six cultural styles using timbre, rhythm, wavelet coefficients and musicology-based features. Kruspe et al. [30] study the automatic classification of 4400 recordings from non-Western music traditions into 9 geographical areas using features of timbre, rhythm and tonality. Zhou et al. [31] use a corpus of 1142 non-Western music tracks from 73 countries and predict the geographical location of each track via a regression method.

Computational approaches to music similarity

Music similarity is studied in several MIR application areas including automatic genre classification [32], cover song detection [33], structural segmentation [34], pattern recognition [35] and music recommendation [36]. In the Music Information Retrieval Evaluation eXchange (MIREX), the annual public evaluation of MIR systems and algorithms, there is a task on

Audio Music Similarity [37]. Since music is a multifaceted concept the study of music similarity is often divided into separate aspects [38]. For example, studies have focused on developing tools and datasets to investigate similarity in aspects of melody [39–41], rhythm [42–44], timbre [45–47], or harmony [48, 49].

The assessment of music similarity is subjective. Automatic systems built for music similarity tasks often need to be trained on a ground truth obtained from human listeners. Several approaches have used genre labels as a proxy for similarity [27]. In this case the assumption is made that songs from the same genre exhibit similar music characteristics. Other studies have focused on the creation of a ground truth set via the collection of similarity ratings from human listeners [50]. Given the scarcity of ground truth data, the evaluation of music similarity systems and the suitability to generalise to all music has been challenged [51, 52]. For example, music similarity systems that are evaluated based on the classification accuracy of genre labels are demonstrated to learn irrelevant music attributes [51]. On the other hand, music similarity systems evaluated with judgements from human listeners are limited by the inter-rater agreement [52]. In particular, due to the challenges in the definition of music similarity and the subjectivity of the task there is often a low inter-rater agreement. As computational models are not expected to outperform the level of human agreement there exists an upper bound beyond which the performance of the model cannot be further improved. Therefore the development and evaluation of a music similarity system still remains a challenge, especially in the yet unexplored space of world music.

Outliers in big data collections

Outlier detection is an essential step in the analysis of big data collections [53]. Outliers denote data points that deviate significantly from the distribution and often need to be filtered out or treated in a different manner. Applications of outlier detection include, amongst others, the identification of intrusions in computer networks [54], fraud in credit cards [55] and abnormal symptoms in disease diagnosis [56]. The study of outliers with respect to spatial relations, as assumed in this music research, adopts concepts of spatial statistics. A spatial outlier is usually viewed as a local anomaly whose non-spatial attribute values are extreme compared to its neighbours [57]. Spatial outlier detection can help locate extreme meteorological events [58], identify disease outbreaks [59], and predict crime hot spot areas [60].

The detection of outliers in music data is still a new area of research. Bountouridis et al. [61] investigate outlier detection in music data using multiple sequence alignment techniques. Lu et al. [62] compare outlier detection techniques applied on a music genre recognition dataset. Hansen et al. [63] apply outlier detection using probability density estimation methods to clean up large-scale datasets of mislabelled data. Livshin and Rodet [64] use outlier detection methods to identify badly recorded musical instrument samples. In the current study, outlier detection is used to identify geographical regions with distinct musical characteristics.

Methodology

The methodology is summarised as follows. For each audio recording in our dataset we extract music descriptors by a) filtering out speech segments as detected via a speech/music discriminator algorithm, b) extracting audio descriptors capturing aspects of music style, c) applying feature learning to reduce dimensionality and project the recordings into a similarity space. We optimise parameters and evaluate music similarity in the projected space by a classification task. The projected space is used to identify recordings that are outliers. We refer as 'outliers' to the recordings that stand out with respect to the whole set of recordings. Outliers are detected for different sets of features focusing on rhythm, melody, timbre, or harmony and a

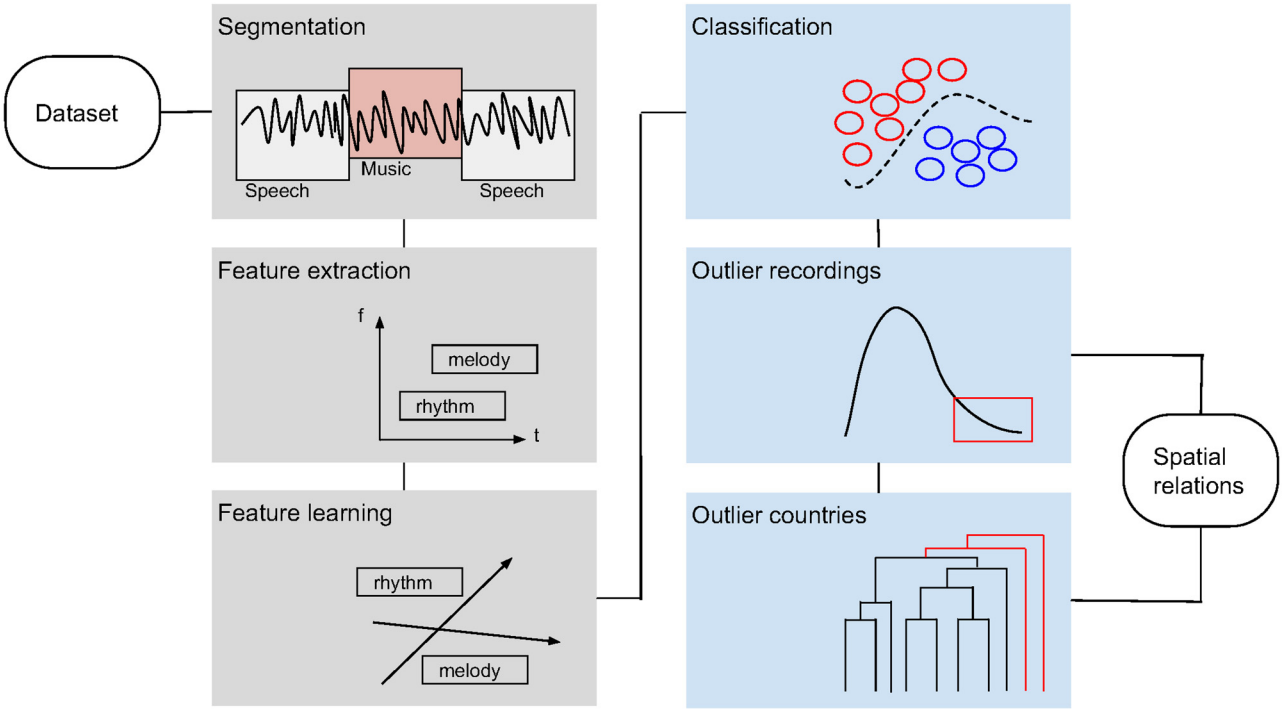

Fig 1. Overview of the methodology.

https://doi.org/10.1371/journal.pone.0189399.g001

combination of these. We take into account spatial relations to form geographical neighbourhoods and use these to detect spatial outliers, i.e., recordings that stand out with respect to their neighbours. Lastly, we extract a feature representation for each country by summarising information of its recordings. Hierarchical clustering is used to get an overview of similarity and dissimilarity between countries. The methodology is summarised in Fig 1 and explained in detail in the sections below.

In our analyses we use the country label of a recording as a proxy for music style. We assume that recordings originating from the same country have common musical characteristics and we use this as the ground truth to train our models. However, it is often the case that a music style is not unique to a single country. Music styles may be shared across many countries and a country may exhibit several music styles. The reason for choosing country as the unit of analysis in this study is two-fold: First, country label is the most consistent information available in our music metadata compared to, for example, music genre, language, or culture information (see also Data section). Second, several studies have considered larger geographical regions (e.g., continents or cultural areas) for the comparison of music styles [28, 30, 65]. Country boundaries work in a similar way but provide a more fine-grained unit for analysis. Alternative approaches are discussed further in the Discussion section.

Data

We aim to investigate music similarity in a world music corpus. The notion of world music is ambiguous often mixing folk, popular, and classical musics from around the world and from different eras [66]. In this study world music refers to recorded material from folk and traditional music styles from around the world. In particular we focus on field recordings collected by ethnomusicologists since the beginning of the 20*th* century. Our music dataset is drawn

from two large archives, the Smithsonian Folkways Recordings [67] and the World & Traditional music collection from the British Library Sound Archive [68]. Both archives include thousands of music recordings collected over decades of ethnomusicological research.

Even though access to large collections of world music recordings is now feasible, the creation of a representative world music corpus is still challenging. An ideal world music corpus would include samples from all inhabited geographical regions and provide information on the spatio-temporal and cultural origins of each music piece. The samples chosen would have to be sufficient to represent the diversity of styles within each music culture and the corpus as a whole should be a balanced collection of music cultures. Given the archives available today, the challenges in corpus creation involve addressing what defines a good sample, how to balance the diverse styles represented in the collection, how to avoid the Western-music bias and how to maximize the size of the corpus. These challenges have also been the main point of criticism for several music comparative studies [69–72]. Our effort to create a world music corpus from the currently available data is described below.

We use a subset of the Smithsonian Folkways Recordings collection which consists of more than 40000 audio recordings, including music as well as poetry. It has a large representation from North America (more than 21000 from the United States and around 1400 from Canada). It also includes around 7700 recordings from Eurasia (1700 from the United Kingdom, 800 from Russia, 800 from France), 4200 recordings from South America (Mexico 600, Trinidad and Tobago 400, Peru 400), 2300 from Asia (India 400, Indonesia 400, Philippines 200, China 200), 1900 from Africa (South Africa 200, Ghana 200, Kenya 100), and 400 from Oceania. Recording dates span from 1938 to 2014. We also use a subset of the World & Traditional music collection of the British Library Sound Archive as curated for the purposes of the Digital Music Lab project [8]. This subset consists of more than 29000 audio recordings with a large representation (17000) from the United Kingdom. It also includes around 7300 recordings from Africa (mostly from Uganda 3000), 2300 from Asia (mostly from Nepal 800 and Pakistan 700), and less than 1000 recordings from Oceania, North and South America. Recording dates span from 1898 to 2014. The metadata associated with each music recording include the country where the recording was made and the year it was recorded, the language and sometimes cultural background of the performers, the subject of the music or short description of its purpose, the title, album (if any), and information of the collector or collection it was accessed from.

In the above archives there is an unbalanced representation of music cultures, with the majority of recordings originating from Western-colonial areas. What is more, metadata for each recording is not always present or is inconsistent. To create a corpus we sample recordings based on the country information which in this case is more consistent than other culture-related metadata. In order to ensure geographical spread we require recordings from as many countries as possible. We set a minimum requirement of $N_{min} = 10$ recordings from each country and select a maximum of $N_{max} = 100$. Setting the minimum to 10 recordings is a trade-off between allowing under-represented areas to be included in the dataset and having a sufficient number of samples for each country. Although a sample of 10 recordings is too small to represent the diversity of music styles within a country, raising this minimum to e.g. 50 would exclude many of the countries we currently analyse and would limit the geographical scope of the study. Setting the maximum to 100 recordings prevents the over-represented areas from dominating the corpus. We sample at random N recordings from each country, where N is bounded by N_{min} and N_{max} as explained above.

Since the medium of analysis is digitised audio, most of our samples are dated since the 1950*s*, with the exception of some recordings from the British Library collection dated around 1900 which were digitised from wax cylinders. The duration of audio recordings from the

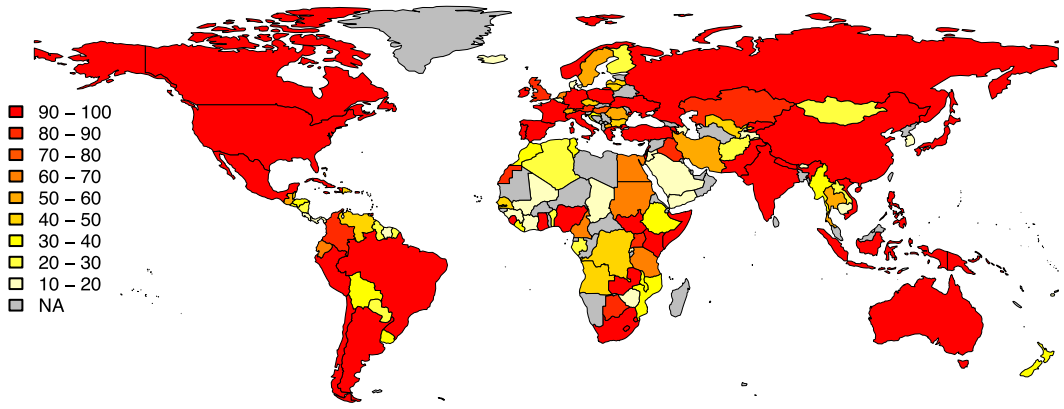

Fig 2. The distribution of countries in our dataset of 8200 world music recordings.

https://doi.org/10.1371/journal.pone.0189399.g002

Smithsonian Folkways Recordings collection is restricted to 30 seconds since we use the publicly available 30-second audio previews. For the British Library Sound Archive data we have access to complete recordings but we only sample the first music segments up to a total duration of 30 seconds for consistency with the short audio excerpts of the Smithsonian Folkways collection.

Given the above criteria, the final collection consists of a total of 8200 recordings, 6132 from the Smithsonian Folkways Recordings collection and 2068 from the British Library Sound Archive collection. The recordings originate from 137 countries with mean 59.9 and standard deviation 33.8 recordings per country (Fig 2). A total of 67 languages is represented by a minimum of 10 recordings, with a mean of 33.5 and standard deviation of 33.5 recordings per language (Fig 3). The recordings span the years between 1898–2014 with median year 1974 and standard deviation of 17.9 years (Fig 4).

Audio content analysis

Over the years several toolboxes have been developed for music content description [73–76]. Applications of these toolboxes include tasks of automatic classification and retrieval of mainly Eurogenetic music (Related work section). Audio content analysis of world music recordings has additional challenges. First, the audio material is recorded under a variety of recording conditions (live and field recordings), and is preserved to different degrees of fidelity (old and new recording media and equipment). Second, the music is very diverse and music descriptors designed primarily for Eurogenetic music might fail to capture particularities of world music styles. Our audio content analysis process includes a pre-processing step to remove speech segments from the dataset (Pre-processing section) and low-pass filtering to reflect limitations of old recording equipment (Features section). With respect to music descriptors, between specifically designing them as in other comparative music studies [28, 30, 31] and automatically deriving them from the spectrogram [77, 78] we choose a middle ground. We use expert knowledge to derive low-level music representations (Features section) and combine them with feature learning methods (Feature learning section) to adapt the representation to particularities of the music we analyse. Details for each step of the audio content analysis process are provided below.

Pre-processing. Our dataset consists of field recordings that sometimes mix speech and music segments. We are only interested in music segments but due to the lack of metadata speech segments cannot be filtered out a-priori. An essential pre-processing step is therefore

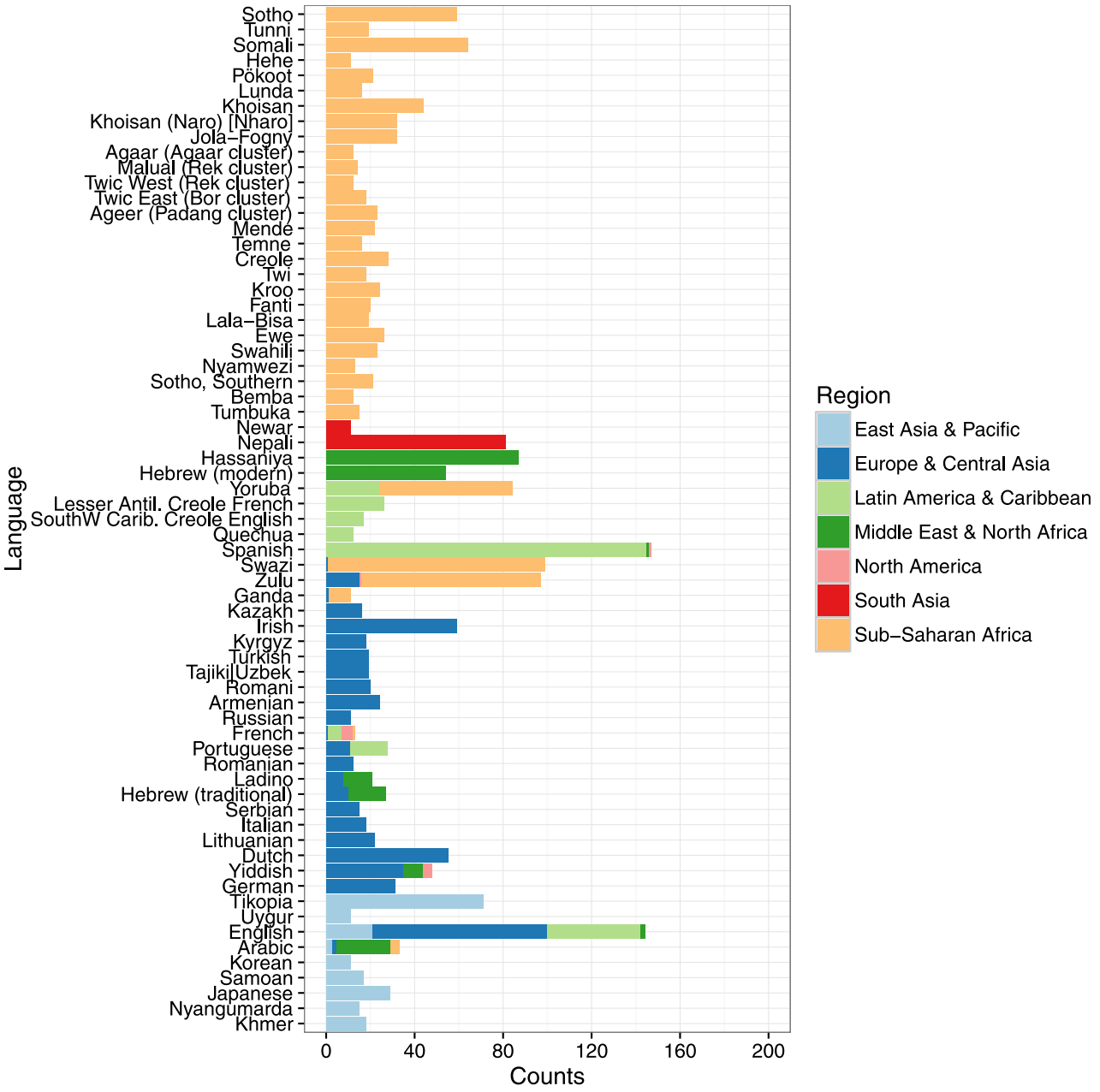

Fig 3. The languages in our world music corpus which are represented by a minimum of 10 recordings.

https://doi.org/10.1371/journal.pone.0189399.g003

the discrimination between speech and music segments. By speech/music segmentation we refer to the detection of segment boundaries and the classification of the segment as either speech or music. The task of speech/music segmentation has been the focus of several studies in the literature [79–81] and it was also identified as a challenge in the 2015 Music Information Retrieval Evaluation eXchange (MIREX) [82]. We select the best performing algorithm [83] from the MIREX 2015 evaluation. As part of the MIREX 2015 evaluation, the algorithm was tested on a non-overlapping set of British Library recordings which is very similar to the recording collection we use in this study and achieved a frame-based F-measure of 0.89. The

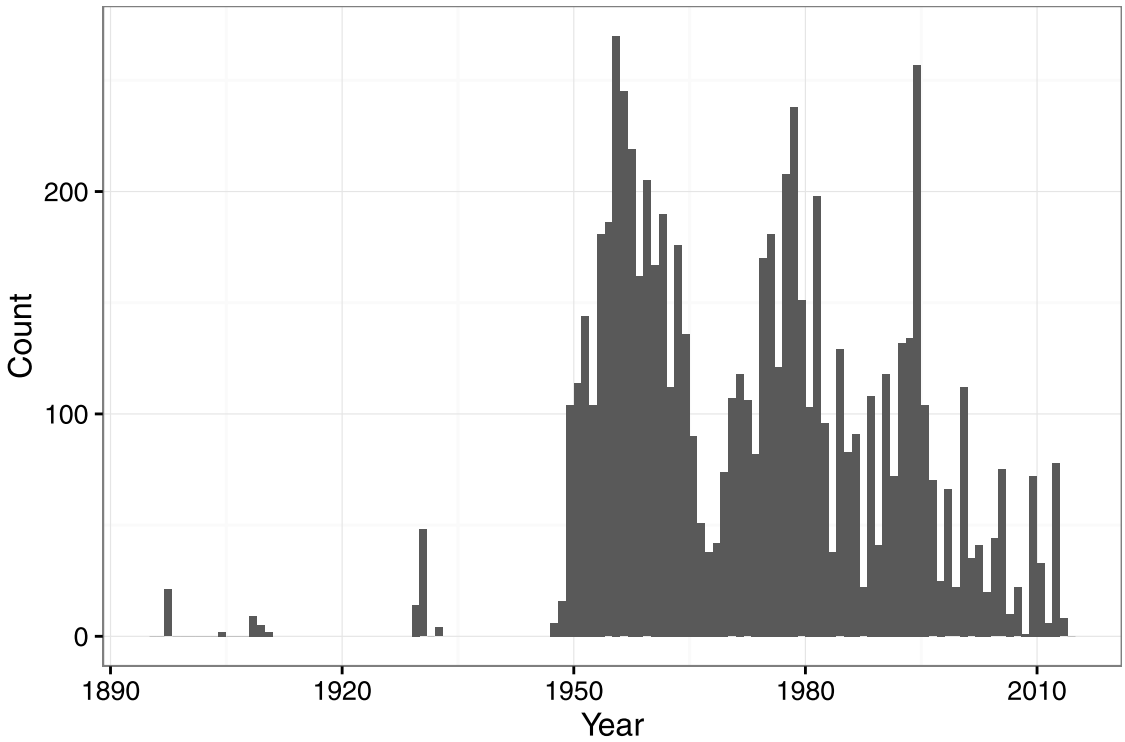

Fig 4. The time span of recordings in our world music corpus.

https://doi.org/10.1371/journal.pone.0189399.g004

algorithm is based on summary statistics of low-level features including Mel frequency cepstrum coefficients (MFCCs), spectral entropy, tonality, and 4 Hertz modulation, and is trained on folk music recordings [84]. We apply this algorithm to detect speech/music segments for all recordings in our dataset and use solely the music segments of each recording for further analysis. In case of long audio excerpts we only select the initial music segments up to a total duration of maximum 30 seconds (see also duration of recordings in Data section).

Features. We are interested in descriptors capturing aspects of world music style. We adopt the notion of music style by Sadie et al. [85], 'style can be recognized by characteristic uses of form, texture, harmony, melody, and rhythm'. The use of form is ignored in this study as most of our music collection is restricted to short audio excerpts rather than complete recordings. We focus on state of the art descriptors (and adaptations of them) that aim at capturing relevant rhythmic, timbral, melodic, and harmonic content. In particular, we extract onset patterns with the scale transform [86] for rhythm, pitch bi-histograms [87] for melody, average chromagrams [88] for harmony, and Mel frequency cepstrum coefficients (MFCCs) [89] for timbre content description. We choose these descriptors because they define low-level representations of the musical content, i.e., a less detailed representation but one that is more likely to be robust with respect to the diversity of the music styles we consider. In addition, these features achieved state-of-the-art performances in relevant classification and retrieval tasks [14], for example, onset patterns with the scale transform perform best in classifying Western and non-Western rhythms [90, 91] and pitch bi-histograms have been applied successfully in (melody-based) cover song recognition [87].

The audio features used in this study are computed with the following specifications. All recordings in our dataset have a sampling rate of 44100 Hz. For all features we compute the (first) frame decomposition using a window size of 40 ms and hop size of 5 ms. The output of

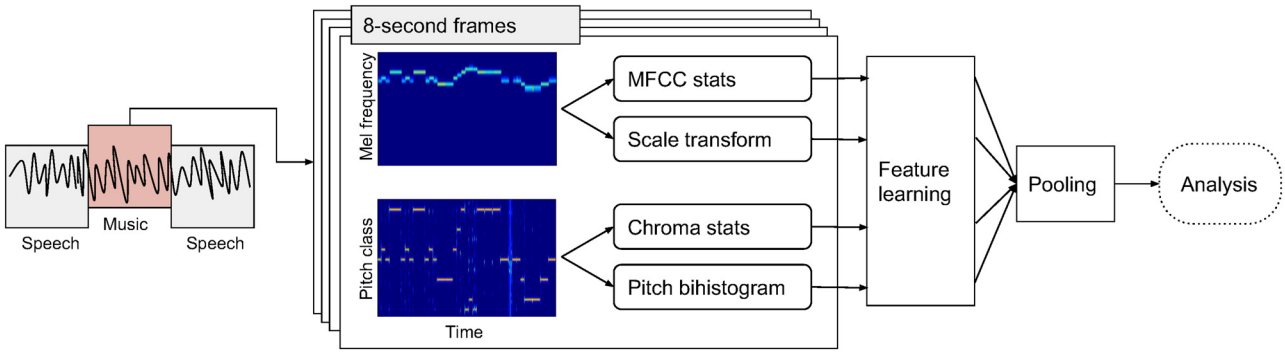

Fig 5. Overview of the audio content analysis process. Mel-spectrograms and chromagrams are processed in overlapping 8-second frames to extract rhythmic, timbral, harmonic, and melodic features. Feature learning is applied to the 8-second features and average pooling across time yields the representations for further analysis.

https://doi.org/10.1371/journal.pone.0189399.g005

the first frame decomposition is a Mel spectrogram and a chromagram. We use a second frame decomposition to extract descriptors over 8-second windows with 0.5-second hop size. This is particularly useful for rhythmic and melodic descriptors since rhythm and melody are perceived over longer time frames. Rhythmic and melodic descriptors considered in this study are derived from the second frame decomposition with overlapping 8-second windows. Timbral and harmonic descriptors are derived from the first frame decomposition with 0.04-second windows and for consistency with rhythmic and melodic features, they are summarised by their mean and standard deviation over the second frame decomposition with overlapping 8-second windows. The window of the second frame decomposition is hereby termed as 'texture window' [25]. The window size w of the texture window was set to 8 seconds after the parameter optimisation process described in the Parameter optimisation section. For all features we use a cutoff frequency at 8000 Hz since most of the older recordings do not contain higher frequencies than that. The audio content analysis process is summarised in Fig 5.

Rhythm and Timbre. For rhythm and timbre features we compute a Mel spectrogram with 40 Mel bands up to 8000 Hz using Librosa [76]. To describe rhythmic content we extract onset strength envelopes for each Mel band and compute rhythmic periodicities using a second Fourier transform with window size of 8 seconds and hop size of 0.5 seconds. We then apply the Mellin transform to achieve tempo invariance [90] and output rhythmic periodicities up to 960 beats per minute (bpm). The output is averaged across low and high frequency Mel bands with cutoff at 1758 Hz. The resulting rhythmic feature vector has length 400 values. Timbral aspects are characterised by 20 MFCCs and 20 first-order delta coefficients after removing the DC component [89]. We take the mean and standard deviation of these coefficients over 8-second windows with 0.5-second hop size. This results in a total of 80 feature values describing timbral aspects.

Harmony and Melody. To describe harmonic content we compute chromagrams using variable-Q transforms [92] up to 8000 Hz with 5 ms hop size and 20-cent pitch resolution to allow for microtonality. Chromagrams are aligned to the pitch class of the maximum magnitude per recording for key invariance. Harmonic content is described by the mean and standard deviation of chroma vectors using 8-second windows with 0.5-second hop size. The dimensionality of the harmonic feature vector results in a total of 120 values. To describe melodic content we extract pitch contours from polyphonic music signals using a method based on a time-pitch salience function [93]. The pitch contours are converted to 20-cent resolution binary chroma vectors with entries of 1, whenever a pitch estimate is active at a given

time, and 0 otherwise. Melodic aspects are captured via pitch bi-histograms which denote counts of transitions of pitch classes [87]. We use a window of $d = 0.5$ seconds to look for pitch class transitions in the binary chroma vectors. The resulting pitch bi-histogram matrix consists of $3600 = 60 \times 60$ values corresponding to pitch transitions with 20-cent pitch resolution. For efficient storage and processing, the matrix is decomposed using non-negative matrix factorisation [94]. We keep 2 basis vectors with their corresponding activations to represent melodic content. It was estimated that keeping only 2 bases was enough to provide sufficient reconstruction for most pitch bi-histogram matrices in our dataset (average reconstruction error $< 25\%$). Pitch bi-histograms are also computed over 8-second windows with 0.5-second hop size. This results in a total of 120 feature values describing melodic aspects.

Combining all features together results in a total of 840 descriptors for each recording in our dataset. A z-score standardisation of the 840 features is applied across all recordings before further processing.

Feature learning. For the low-level descriptors presented in the Features section we aim to learn high-level representations that best characterise music style similarity. Feature learning is also appropriate for reducing dimensionality, an essential step for the amount of data we analyse. We learn feature representations from the 8-second frame-based descriptors. In our analysis we consider the country label of a recording as a proxy for style and use this for supervised training and cross-validating our methods.

There are numerous feature learning techniques to choose from in the literature. Non-linear models such as neural networks usually require large training data sets [95]. We have a fairly limited number of audio recordings and our low-level descriptors partly incorporate expert knowledge of the music (section Features). In this case, simpler feature learning techniques are more suitable for the amount and type of data we have. We explore the applicability of 4 linear models trained in supervised and unsupervised fashions.

The audio features are standardised using z-scores and aggregated to a single feature vector for each 8-second frame of a recording. Feature representations are learned using Principal Component Analysis (PCA), Non-Negative Matrix Factorisation (NMF), Semi-Supervised Non-Negative Matrix Factorisation (SSNMF), and Linear Discriminant Analysis (LDA) methods [94]. PCA and NMF are unsupervised methods and try to extract components that account for the most variance in the data without any prior information on the data classes. LDA is a supervised method and tries to identify attributes that account for the most variance between classes (in this case country labels). SSNMF works similarly to NMF with the difference that ground truth labels are taken into account in addition to the data matrix in the optimisation step [96].

We split the 8200 recordings of our collection into training (60%), validation (20%), and testing (20%) sets. We train and test our models on the frame-based descriptors; this results in a dataset of 325435, 106632, and 107083 frames for training, validation, and testing, respectively. Frames used for training do not belong to the same recordings as frames used for testing or validation and vice versa. We use the training set to train the PCA, NMF, SSNMF, and LDA models and the validation set to optimise the parameters. In each experiment we retain components constituting to 99% of the variance. In the Results section we analyse the feature weights for the components of the best performing feature learning method.

A classification task is used to assess the quality of the learned space and optimise the parameters. An ideal music similarity space separates well data points belonging to different music classes and good classification results can be achieved with simple classifiers. We are not interested to build a powerful classifier since our primary aim is to assess the learned embeddings and not to optimise the classification task itself. We therefore focus on classifiers widely used in the machine learning community [97]. We train 4 classifiers, K-Nearest Neighbour

(KNN), Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and Random Forest (RF), to predict the country label of a recording. The purpose of the classification task is to optimise the window size w of the audio descriptors and assess the quality of the learned spaces in order to select the optimal feature learning method for our data. We use the classification F-score metric to compare the performance of the models. In the Results section we also analyse the coefficients of the best performing classifier.

In order to assess the contribution of different features to the classification task we consider 5 sets of features: a) scale transform (rhythmic) b) MFCCs (timbral), c) average chroma vectors (harmonic), d) pitch bi-histograms (melodic), and e) the combination of all the above. In each case, feature learning is applied on the selected feature set and frame-based projections are aggregated using the mean prior to classification. We also tested for aggregation using the mean and standard deviation of frame-based descriptors but this did not improve results; hence it was omitted. In the case of testing the combination of all features (e), we first reduce dimensionality for each feature set separately and then concatenate the components from all feature sets before mean aggregation and classification. Results for the feature learning optimisation and classification experiments are presented in the Results section.

Data mining

Outlier recordings. The feature learning and classification methods described above (Feature learning section) identify the optimal projection for the data. In the next step of the analysis we use the projected space to investigate music dissimilarity and identify outliers in the dataset. A recording is considered an outlier if it is distinct compared to the whole set of recordings. We detect outliers based on a method of squared Mahalanobis distances [13, 98]. Using Mahalanobis, a high-dimensional feature vector is expressed as the distance to the mean of the distribution in standard deviation units. Let $X \in \mathbb{R}^{I \times J}$ denote the set of observations for I recordings and J features. The Mahalanobis distance for observation $\mathbf{x_i} = (x_1, x_2, \ldots, x_J)^T$ for recording i from the set of observations X with mean $\mu = (\mu_1, \mu_2, \ldots, \mu_J)^T$ and covariance matrix S is denoted

$$D_M(\mathbf{x_i}) = \sqrt{(\mathbf{x_i} - \mu)^T S^{-1} (\mathbf{x_i} - \mu)}. \quad (1)$$

Data points that lie beyond a threshold, typically set to the $q = 97.5\%$ quantile of the chi-square distribution with J degrees of freedom [99], are considered outliers. This is denoted

$$O = \{i \in H | D_M(\mathbf{x}_i) > \sqrt{\chi^2_{J,q}}\} \quad (2)$$

where $H = \{1, 2, \ldots, I\}$ denotes the index of the observations.

Due to the high dimensionality of our feature vectors every data point can be considered far from the centre of the distribution [100]. To compensate for a possible large amount of outliers we consider a higher threshold based on the $q = 99.9\%$ quantile of the chi-square distribution.

To gain a better understanding of the type of outliers for each country we detect outliers using a) rhythmic, b) timbral, c) harmonic, and d) melodic features. For example, for J_R the dimensionality of the rhythmic feature vector and $X_R \in \mathbb{R}^{I \times J_R}$ the set of observations, the set of outlier recordings with respect to rhythmic characteristics is denoted

$$O_R = \{i \in H | D_M(\mathbf{x}_{R,i}) > \sqrt{\chi^2_{J_R,99.9}}\} \quad (3)$$

for observation $\mathbf{x}_{R,i} \in X_R$. We detect outliers with respect to rhythmic (O_R), timbral (O_T), melodic (O_M), and harmonic (O_H) characteristics.

Spatial neighbourhoods. In the previous section outliers were detected by comparing a recording to all other recordings in the dataset. Here we take into account spatial relations and compare recordings from a given country only to recordings of its neighbouring countries. In this way we are able to identify spatial outliers, i.e. recordings that are outliers compared to their spatial neighbours [57]. We construct spatial neighbourhoods based on contiguity and distance criteria: a) two countries are neighbours if they share a border (a vertex or an edge of their polygon shape), b) if a country doesn't border with any other country (e.g., the country is an island) its neighbours are defined by the 3 closest countries estimated via the Euclidean distance between the geographical coordinates (latitude and longitude) of the centre of each country.

Let N_i denote the set of neighbours for country i estimated via

$$N_i = \{j \in \{1, \ldots, R\} | \text{j is neighbour to i}\} \tag{4}$$

for R the number of countries. The spatial neighbourhood is represented as a weight matrix $W \in \mathbb{R}^{R \times R}$ where entry $w_{ij} \in W$ is non-zero whenever country j is neighbour to country i. This is denoted

$$w_{ij} = \begin{cases} \frac{1}{n_i}, & \text{if } j \in N_i \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $n_i = |N_i|$ denotes the total number of neighbours for country i. By definition, weight matrix W is row-standardized, $\sum_{j=1}^{R} w_{ij} = 1$.

Table in S1 Table provides the neighbours of each country as estimated via this approach. The geographical boundaries of each country are derived from spatial data available via the Natural Earth platform [101].

The set of recordings from a given country is appended with recordings from neighbouring countries as defined by the country's spatial neighbourhood (S1 Table). This set is used to detect outliers with the Mahalanobis distance as defined in Eq 2. Spatial outliers are detected in this manner for all countries in our dataset.

Outlier countries. The unit of analysis in the previous sections was the individual recordings. In this section we move one level up and place the focus at the country. We detect outlier countries in a similar manner as before where country features now summarise the information of the underlying recordings. The advantage of placing the focus at the country level is that the feature representations can now summarise the variety of styles that exist in the music of a country. Outliers are not judged by individual recordings but rather by the distribution of the whole set of recordings of each country.

We use K-means clustering to map recording representations to one of K clusters. The country representation is then derived from a histogram count of the K clusters of its recordings. Let $X \in \mathbb{R}^{I \times J}$ denote the set of observations for I recordings and J features. We compute K-means for X and map recordings to one of K clusters. We use a linear encoding function $f : \mathbb{R}^J \rightarrow \mathbb{R}^K$ so that each recording representation $\mathbf{x}_i \in \mathbb{R}^J$ for $i = 1, \ldots, I$ is mapped to a vector $\hat{\mathbf{x}}_i \in \mathbb{R}^K$ via the dot product between $\mathbf{x}_i$ and the cluster centroids $\mathbf{m}_k \in \mathbb{R}^J$ for $k = 1, \ldots, K$ clusters. The feature vector for a country $\mathbf{c}_r \in \mathbb{R}^K$ is the normalised histogram count of K clusters for recordings i from country r, denoted

$$\mathbf{c}'_r = \sum_i f(\mathbf{x}_i). \tag{6}$$

Each histogram is normalised to the unit norm, where $\mathbf{c}_r = \frac{c'_r}{\|c'_r\|}$. Let $C \in \mathbb{R}^{R \times K}$ denote the feature representations for R countries and K clusters derived as explained above. The optimal number K of clusters is decided based on the silhouette score [102] after evaluating K-means for K between 10 and 30 clusters.

We estimate similarity between countries via hierarchical clustering [103]. For consistency with the previous outlier detection method (section Outliers at the recording level), we use Mahalanobis distance to estimate pairwise similarity between countries. Pairwise Mahalanobis distance between countries is denoted

$$D_M(\mathbf{c}_i, \mathbf{c}_j) = \sqrt{(\mathbf{c}_i - \mathbf{c}_j)^T \bar{S}^{-1} (\mathbf{c}_i - \mathbf{c}_j)} \tag{7}$$

where $\bar{S}$ is the covariance matrix and $i, j \in \{1, 2, \ldots, R\}$. A hierarchy of countries is constructed using the average distance between sets of observations as the linkage criterion.

Results

Parameter optimisation

As mentioned in the Audio content analysis section, the window size w in the feature extraction process (Features section) was optimised based on a classification task. Given the feature transformed representations of each recording in the training set, we trained 4 classifiers (KNN, LDA, SVM, RF), to predict the country label of a recording. Parameter optimisation was based on the classification accuracy on the validation data. We used the weighted average of the F-measure of each class [104], referred to as F-score, to report classification performance in this case of unbalanced data classes. Fig 6 shows the classification F-score of the best performing classifier (LDA) for a range of window sizes w. Based on this evaluation the optimal window size was $w = 8$ seconds with highest F-score of 0.37 for the LDA classifier in combination with the LDA-transformed features.

The dimensions of the LDA-transformed features can be explained in the following way. LDA components for the rhythmic features give more weight to the periodicities of the high-frequency Mel bands (above 1758 Hz). Melodic features receive similar weights for both the bases and activations of the pitch bi-histogram. LDA components for the harmonic features assign more weight to relative pitch values (mean of chroma vectors) rather than pitch fluctuations (standard deviation of chroma vectors) over time. LDA components for timbral features focus on timbre fluctuation (mean and standard deviation of MFCC delta coefficients) over time. This is opposite to the behaviour of PCA transformation where components focus on absolute timbre qualities (mean and standard deviation of MFCC coefficients) over time. Fig 7 illustrates the difference between LDA and PCA components for the timbral features.

Classification

The classification results for the different classifiers in combination with the feature learning methods are presented in Table 1. Classification accuracy of the test set was assessed after fixing the window size of the feature extraction to $w = 8$ seconds as found optimal in section Parameter optimisation. Results suggest that the best classifier for our data when the combination of all features is considered is the LDA classifier with the LDA-transformed features (classification F-score of 0.321). Rhythmic, melodic, and harmonic features achieved best classification performance for the LDA-transformed features and the LDA classifier whereas timbral features achieved best classification performance for the LDA-transformed features and the SVM classifier. The first 10 components of the LDA classifier trained with the LDA-

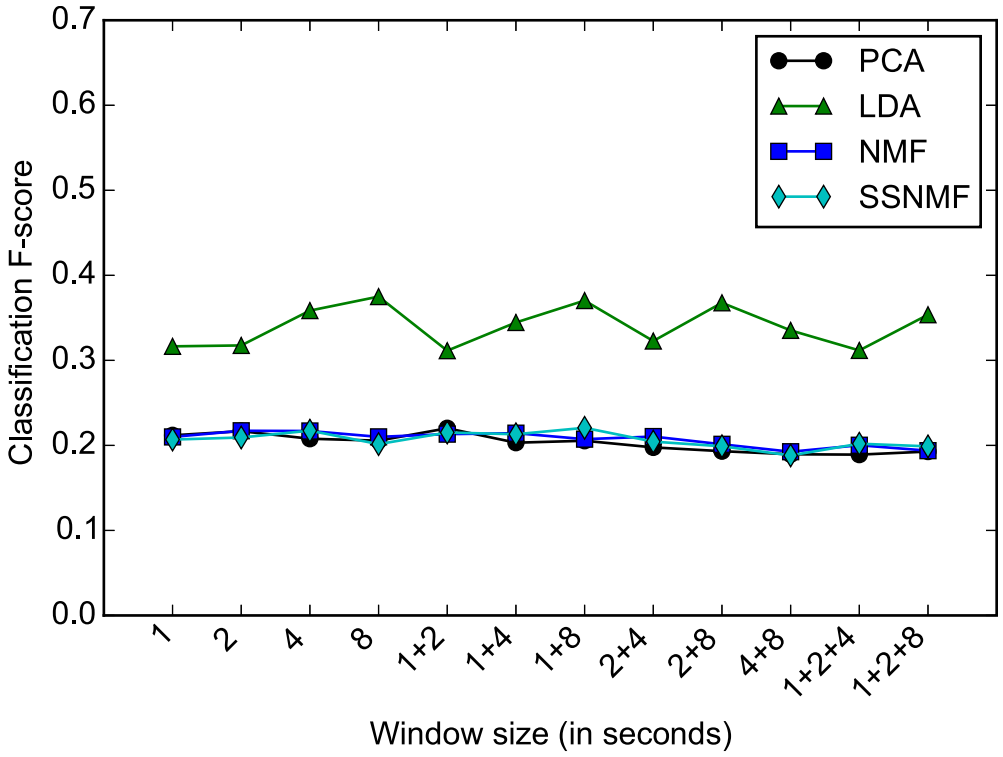

Fig 6. Classification F-score on the validation set for the best performing classifier (LDA) across different window sizes. Accuracies are compared for different feature learning methods (PCA, LDA, NMF, SSNMF). Combinations of window sizes are marked by '+' in (a), for example '4+8' represents the accuracy when combining features from the 4-second and the 8-second windows. Considering the performance of all feature learning methods, the optimal window size is 8 seconds.

https://doi.org/10.1371/journal.pone.0189399.g006

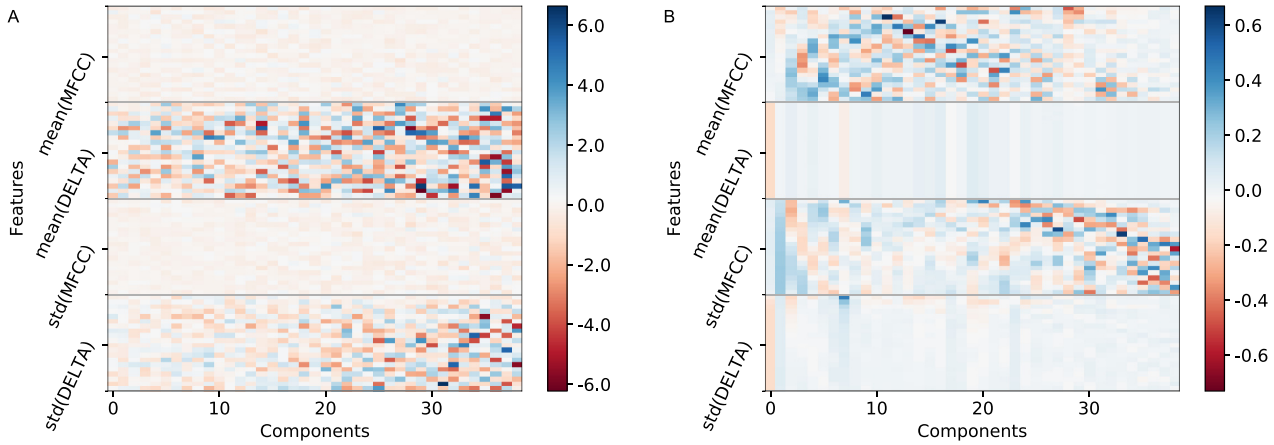

Fig 7. LDA and PCA components weigh timbral features in opposite ways. (A) LDA components focus on timbre fluctuation (mean and standard deviation of MFCC delta coefficients) over time. (B) PCA components focus on absolute timbre qualities (mean and standard deviation of MFCC coefficients) over time.

https://doi.org/10.1371/journal.pone.0189399.g007

Table 1. Classification F-scores of the test set for the country of recording (– denotes no transformation).

Transform	Classifier	F-score				
		All	Rhythm	Melody	Timbre	Harmony
LDA	LDA	0.321	0.150	0.070	0.199	0.107
SSNMF	LDA	0.183	0.053	0.039	0.165	0.082
NMF	LDA	0.178	0.059	0.046	0.166	0.086
–	LDA	0.177	0.060	0.038	0.191	0.084
PCA	LDA	0.175	0.055	0.046	0.162	0.084
LDA	KNN	0.152	0.055	0.023	0.282	0.086
SSNMF	KNN	0.143	0.043	0.015	0.227	0.072
PCA	KNN	0.141	0.053	0.027	0.221	0.081
–	KNN	0.140	0.052	0.027	0.222	0.082
NMF	KNN	0.114	0.043	0.029	0.178	0.080
–	RF	0.083	0.040	0.032	0.114	0.057
LDA	RF	0.071	0.031	0.017	0.150	0.051
NMF	RF	0.063	0.032	0.020	0.126	0.042
PCA	RF	0.046	0.026	0.019	0.140	0.045
SSNMF	RF	0.045	0.031	0.018	0.116	0.035
LDA	SVM	0.023	0.079	0.050	0.296	0.090
SSNMF	SVM	0.021	0.011	0.005	0.019	0.014
NMF	SVM	0.016	0.008	0.008	0.011	0.012
–	SVM	0.015	0.047	0.038	0.250	0.088
PCA	SVM	0.015	0.048	0.039	0.246	0.092

The window size of the features is 8 seconds as found optimal in section Parameter optimisation. Results are sorted by highest to lowest F-score of the combination of all features ('All').

https://doi.org/10.1371/journal.pone.0189399.t001

transformed features give more weight to the timbral and harmonic dimensions and explain 24% of the variance. The remaining components give more weight to the rhythmic and melodic dimensions. More information on the classification results and confusion matrices can be found in the published code repository (http://github.com/mpanteli/music-outliers).

Outliers at the recording level

We found the optimal feature learning method (LDA) that best approximates music similarity in our data as defined by the classification task (Classification section). We use the LDA-projected space to investigate music dissimilarity and identify outliers in the dataset.

From a total number of 8200 recordings we identify 1706 recordings as outliers. The distribution of outliers per country, normalised by the number of recordings per country in our dataset, is summarised in Fig 8. We observe that the country with the most outliers is Botswana with 61% (55 out of 90) of its recordings identified as outliers, followed by Ivory Coast (60%, 9 out of 15), Chad (55%, 6 out of 11), and Benin (54%, 14 out of 26). The percentage of outliers per country was not significantly correlated with the number of recordings sampled from that country (Pearson correlation coefficient $r = -0.01$ with p-value = 0.91).

Listening to some examples we summarise the following timbral characteristics for the outliers. Outlier recordings from Botswana include solo performances of the mouthbow and dance songs featuring group singing accompanied with handclapping or other percussion. Outlier recordings from Ivory Coast feature music from the Kroo ethnic group who originated in eastern Liberia and consist of polyphonic music with singing accompanied by woodwind

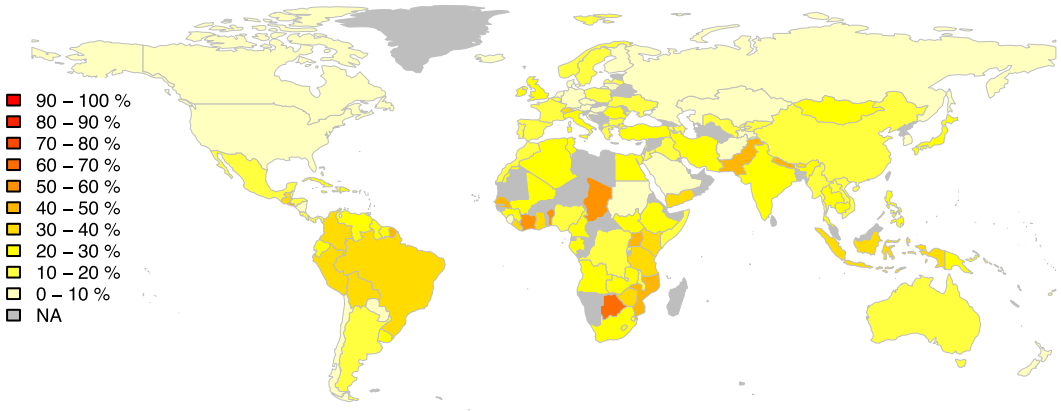

Fig 8. Distribution of outliers per country. The colour scale corresponds to the normalised number of outliers per country, where 0% indicates that none of the recordings of the country were identified as outliers and 100% indicates that all of the recordings of the country are outliers.

https://doi.org/10.1371/journal.pone.0189399.g008

and guitar instruments. Outlier recordings from Chad feature mainly dance music with emphasis on percussive and wind instruments as well as examples of the singing voice in solo and group performances. Outliers from French Guiana feature solo flute performances and singing with percussive accompaniment. Outlier recordings from Gambia include examples of group singing with percussive accompaniment of drums, jingles and wooden blocks, solo performances of the gong and flute. Outlier recordings from Benin include solo performances of the Yoruba drums and music from the Fon culture including examples of group singing with gong accompaniment.

To gain a deeper understanding of the type of outliers for each country we detect outliers using a) rhythmic, b) timbral, c) melodic, and d) harmonic features. Results are shown in Fig 9. With respect to rhythmic aspects the countries with the most outliers are Benin (50%, 13

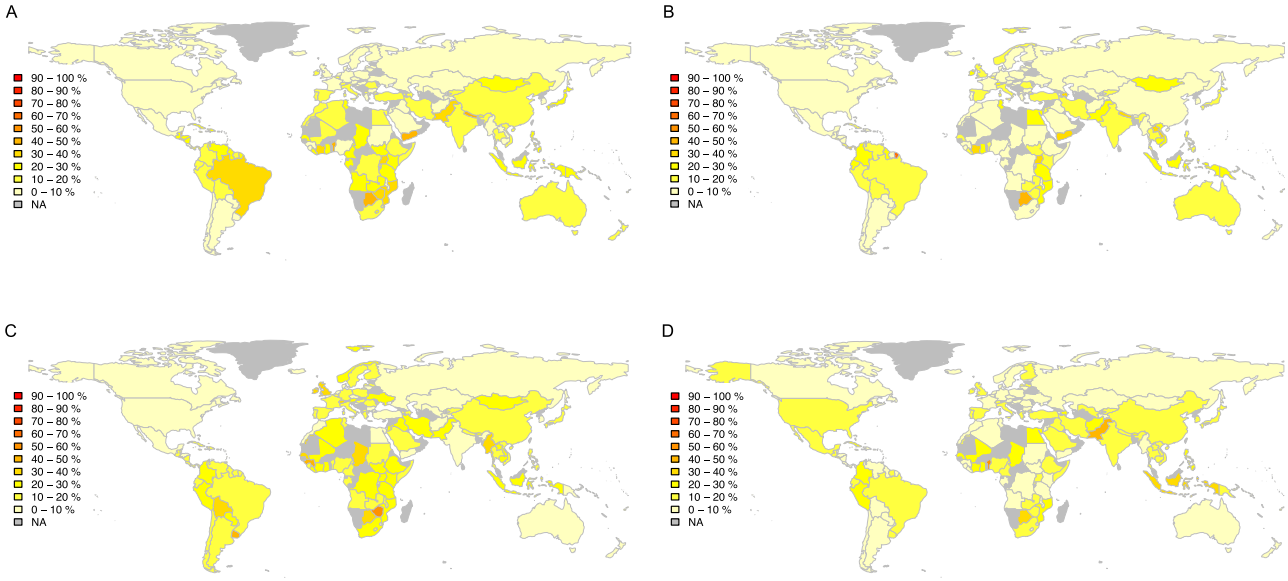

Fig 9. Distribution of outliers per country for each feature. Outliers detected for features of (A) rhythm, (B) timbre, (C) melody, and (D) harmony. The colour scale corresponds to the normalised number of outliers per country, from 0% of outliers (light colours) to 100% (dark colours).

https://doi.org/10.1371/journal.pone.0189399.g009

out of 26), Botswana (49%, 44 out of 90), and Nepal (42%, 40 out of 95). The countries with the most outliers with respect to timbral characteristics are French Guiana (78%, 19 out of 28), Botswana (48%, 43 out of 90), and Ivory Coast (40%, 5 out of 13). The countries with the most outliers with respect to melodic aspects are Zimbabwe (53%, 8 out of 15), Uruguay (48%, 15 out of 31), and Guinea (46%, 5 out of 11) and with respect to harmonic aspects Benin (54%, 14 out of 26), Pakistan (46%, 42 out of 91), and Gambia (36%, 18 out of 50).

Listening to some examples we summarise the following characteristics for the outliers. Rhythmic outliers include examples from African polyrhythms as well as examples with frequent transitions between binary and ternary subdivisions. The most prominent instruments in the rhythmic outliers are pitched and non-pitched percussion. Most rhythmic outliers tend to have a 'full' rhythm, i.e. there are many onsets within each bar duration. Outliers with respect to timbral characteristics include solo performances of xylophones and gongs for example recordings from Botswana, Indonesia, and Gamelan recordings from the Philippines. Another category of instruments that often gives rise to timbre outliers are wind instruments such as reedpipes and flutes. Outliers with respect to melodic characteristics include polyphonic melodies performed on the accordion (e.g. recordings from Uruguay) or the mbira (e.g. recordings from Zimbabwe). With respect to harmony, outliers exhibit microtonal scales and feature instruments with distinct tuning, for example solo sitar or surnai performances from Pakistan, xylophone and gong performances from Benin and Indonesia. Listening examples can be found at the online demo (see http://mpanteli.github.io/music-outliers/demo/outliers).

Spatial outliers. In the previous section we detected outliers by comparing a recording to all other recordings in the dataset. Here we take into account spatial relations and we compare recordings from a given country only to recordings of its neighbouring countries (section Spatial neighbourhoods). We summarise the distribution of spatial outliers, normalised by the total number of recordings in each spatial neighbourhood, in Fig 10. Results show that China is the country with the most spatial outliers (26%, 26 out of 100), followed by Brazil (24%, 24 out of 100), Colombia (21%, 19 out of 90), and Mozambique (21%, 7 out of 34).

China is the country with most spatial neighbours in our dataset, bordering with 12 other countries for which we have music data (S1 Table). Recordings from China feature the butterfly harp string instrument and singing examples from the Han cultural group, often with a bright sound and prominent singing in relatively high frequencies. These examples are

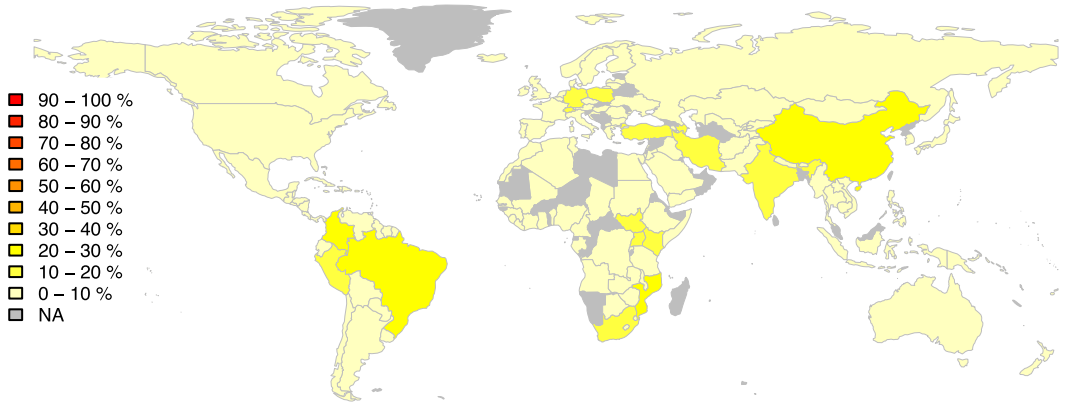

Fig 10. Distribution of outliers per country for the spatial neighbourhoods shown in S1 Table. The colour scale corresponds to the normalised number of outliers per country, from 0% of outliers (light colours) to 100% (dark colours).

https://doi.org/10.1371/journal.pone.0189399.g010

compared to various instruments and music styles from the neighbouring countries including lute performances from Kyrgyzstan, Mongolian jewish harp, Indian tala, Nepalese percussion and wind instrument performances, polyphonic singing from Vietnam and Laos, and instrumental pieces featuring the balalaika from Russia. Compared to the analysis of global outliers (Fig 8) we observe that recordings from China stand out only with respect to its spatial neighbourhoods but are not so distinct compared to the whole dataset of world music.

Outliers at the country level

In this section we consider the country instead of the individual recordings as the unit of analysis and detect outlier countries as described in section Outlier countries.

The silhouette score indicated an optimal number of $K = 10$ clusters. We refer to the country labels of each recording to give an overview of the music styles captured in each cluster. The 3 most frequent countries in each cluster are shown in Fig 11.

The similarity between countries was estimated via hierarchical clustering. Results are presented in a dendrogram in Fig 12. The countries with the most distinct feature representations are South Sudan, Botswana, Ghana, Austria and Switzerland (in order of most to least distinct). The aforementioned countries were found dissimilar (with respect to a threshold) to any other country in our dataset.

Recordings from South Sudan feature mostly examples of the singing voice in solo and group performances. The use of solely the singing voice is what we believe makes the feature representation of South Sudan so different from other countries. A similar observation holds for recordings from Austria and Switzerland featuring mostly dance songs with accordion accompaniment. This might not be a unique music style across our dataset but the consistent use of this style in the recordings from Austria and Switzerland is what we think makes them most distinct from other countries. Botswana and Ghana, also detected as outlier countries with the hierarchical clustering approach, exhibit the use of a variety of music styles. Botswana

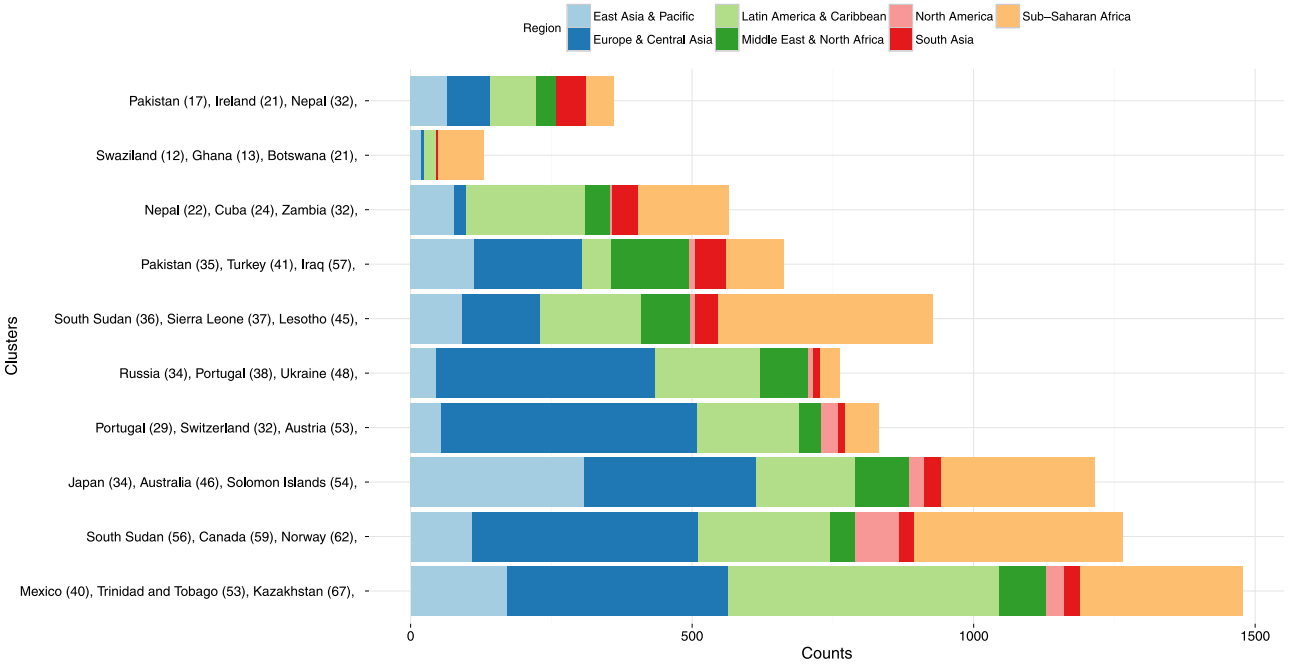

Fig 11. The top 3 countries for each of the 10 clusters.

https://doi.org/10.1371/journal.pone.0189399.g011

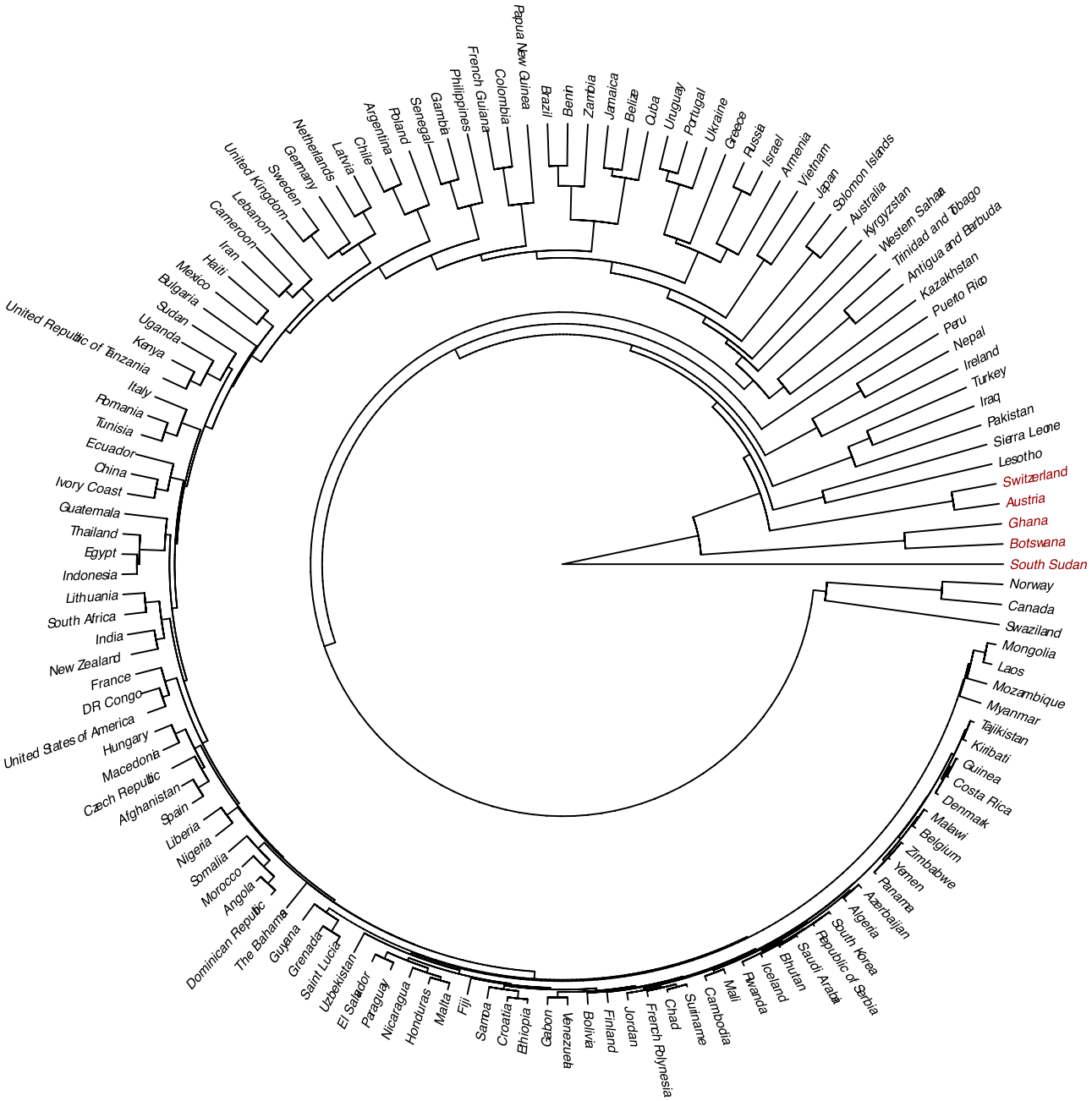

Fig 12. Hierarchical clustering of the 137 countries in our dataset. Each country was represented by the histogram of cluster mappings of its recordings (Outlier countries section). The most distinct countries are annotated with red colour.

https://doi.org/10.1371/journal.pone.0189399.g012

was also detected as the country with the most outlier recordings compared to the global dataset (section Outliers at the recording level). We note that Fig 12 also revealed some music similarity relationships between countries of geographical or cultural proximity. However, as the scope of this study is rather on music dissimilarity and outliers we leave the exploration of these relationships for future work.

Discussion

We combined world music recordings from two large archives and proposed a methodology to extract music features and detect outliers in the dataset. We developed signal processing methods to process music information from the audio signal taking into account the challenges imposed by noisy and musically diverse recordings. Our analyses explored differences and similarities of world music and revealed geographical patterns of music outliers.

We took into account several pre-processing steps to isolate relevant music information from the audio signal: speech segments were separated from music, frequencies above 8000 Hz were omitted for consistency with old recording equipment, and low-level music descriptors were combined with feature learning to give higher-level representations robust to diverse music characteristics. The size of the texture window was optimised and we found that longer windows (8 seconds) provide better representations for our music data than shorter ones (4,2,1 seconds). Feature learning was better in the supervised setting (LDA outperformed PCA and NMF) even though class labels (in this case countries) were not necessarily unique identifiers for the underlying musical content.

We proposed a method to detect outliers and explored several ways of understanding the musical differences. We listed the countries with the most outlier recordings and expanded the analysis to explain which music features are distinct in these outlier recordings. For example, Botswana was the country with most of its recordings detected as outliers and feature analysis showed that those outliers were mostly due to rhythmic and timbral features. With respect to rhythmic features, African countries indicated the largest amount of outliers with recordings often featuring the use of polyrhythms. Harmonic outliers originated mostly from Southeast Asian countries such as Pakistan and Indonesia, and African countries such as Benin and Gambia with recordings often featuring inharmonic instruments such as the gong and bell.

We ran a sensitivity experiment to check how stable our outlier findings are with respect to different datasets. We repeated the outlier analysis 10 times, each time selecting at random a stratified sample of 80% of the original dataset. The majority vote of outlier countries resulting in the top $K = 10$ positions of each experiment was used as the ground truth. Assessing the precision at $K = 10$ for each experiment assuming majority vote ground truth showed that the geographical patterns of outliers (Fig 8) were on average consistent across multiple random subsets of the original dataset (precision at K mean = 0.67, standard deviation = 0.06).

Incorporating spatial information we were able to compare recordings from neighbouring countries. This gave rise to music cultures that are not distinct compared to the global dataset but are still unique compared to their spatial neighbours. For example, music from China with bright timbres was found to be unique compared to its many spatial neighbours. Music from Brazil was also distinct compared to its spatial neighbours, an observation that could be attributed to cultural differences such as the use of different languages between Brazil and its neighbouring countries. Proving historical and cultural influence is not the aim of this study but we believe our findings could provide a good starting point for further investigation.

We also proposed a method to extract feature summaries for each country and estimated clusters for the whole set of recordings. We found 10 clusters to best represent the music styles in our dataset and observed recordings from geographically similar regions often clustered together. Hierarchical clustering at the country level representation revealed African countries such as South Sudan, Botswana, and Ghana as most distinct from others in the dataset.

Hubness

This research deals with high dimensional vectors and analysis of nearest neighbour relationships. High dimensional spaces are prone to produce data points that appear in the

neighbourhood of other points disproportionately often. We tested the effect of hubness in our data following the approach suggested by Schnitzer et al. [105]. We measured hubness as the skewness of the n-occurrence where n-occurrence defines the number of times track x occurs in the top n neighbours of other tracks. We used pairwise Mahalanobis distances and assessed the n nearest neighbours for each track in our dataset for $n = 60$, the average number of recordings per country. We observed a positively skewed distribution with hubness = 10.1. A total of 129 out of 8200 recordings occurred in the nearest neighbour lists of more than 1000 tracks (2% large hubs) and 3332 recordings had n-occurrence = 0 (41% orphans). Pairwise Mahalanobis distances in this study are only used for the computation of outlier countries (section Outlier countries). Future work could aim to reduce hubness via local scaling or mutual proximity [105].

Future work

There are several steps in the overall methodology that could be implemented differently and audio excerpts and features could be expanded and improved. Numerous audio features have been proposed in the literature for describing musical content in sound recordings for various applications. We selected a small set of features from the MIR domain based on their state-of-the-art performance and relevance for world music analysis. It is clear that any such set of features does not capture all aspects of a set of musical recordings. Future work could explore the suitability of feature sets proposed by ethnomusicologists [20] or embeddings learned from raw audio or spectrograms [106].

We used linear feature learning methods to learn higher-level representations from our low-level descriptors. Depending on the data and application, more powerful non-linear methods could be employed to learn meaningful feature representations [107]. What is more, our analysis relies on a bag-of-frames approach where temporal information of the entire music piece is lost by averaging short frames across time. Although this approach is in line with state of the art MIR research [87, 90] alternative methods capturing temporal relationships such as Hidden Markov Models [108] could be considered.

Like all studies of this nature our study is subject to sampling bias. Our observations on world music similarity are restricted to the dataset we analyse. It is difficult to gather representative samples of 'all' music of the world. We aimed to maximise geographical spread in the dataset by including as many countries as possible and representative samples from each country were drawn at random. This resulted in a total of 137 countries with a minimum of 10 recordings per country. Even though this is the largest and most diverse corpus of world music studied so far, there are many areas of the world and cultures that are not represented. The creation of a representative world music corpus will continue indefinitely as more music is recorded and the digitisation of archived recordings proceeds.

In this study country labels have been considered a proxy to music style and have been used to train models for music similarity and dissimilarity. While countries provide a broad notion of ethnic boundaries, music styles are not homogeneous within these boundaries. A country may exhibit several music styles and a music style may spread across many countries. The ambiguity of these boundaries provides an upper limit to the performance of our models. This ambiguity could be reduced by incorporating more information, for example the culture or language of the musicians, to better approximate the music style of a recording. Extracting culture or language information from the currently available metadata requires additional manual labour and this is a task left for future work.

Furthermore, a lot of information regarding the music style of a recording can be extracted from the date it was created. Music evolves over time, and two recordings from the same

location but recorded with a time difference of 50 years may vary in their style. In this study we ignored temporal information and considered our dataset as a static collection of world music. Country of origin and recording date could be used together to define the music style of a recording.

Our study focuses on the detection of outliers in music collections. The data we work with are numerical representations derived from a multi-step procedure of processing the audio signal. The suitability of the audio tools can be questioned with regard to their ability to capture and represent high-level musical concepts [70]. Likewise, the patterns we observe can sometimes be artifacts of the tools we use. We note that in this study the estimated outliers did not appear to be attributable to recording date differences or acoustic environments but quantitative and qualitative evaluation could be expanded [109].

Conclusion

The comparison of world music cultures has been traditionally studied with non-computational tools. We investigated similarity in a large corpus of world music using signal processing and data mining tools. We analysed thousands of recordings from folk and traditional music from around the world and quantified differences and similarities. Our findings identify regions that have possibly developed unique musical characteristics such as Botswana, as well as China, which is most distinct from its neighbours. We have also explored geographical patterns of music outliers for different sets of features and found that Benin has the most outlier recordings with respect to rhythm and harmony, French Guiana with respect to timbre, and Zimbabwe with respect to melody. A categorisation into world music styles identified 10 clusters with South Sudan and Botswana exhibiting the most distinct use of these clusters. This is the first study to consider the computational analysis of such a large world music corpus. There is a lot to be explored yet and we believe continuing on this line of research will help us understand better the music cultures of the world.

Supporting information

S1 Table. Spatial neighbours for each country in our dataset.
(PDF)

Author Contributions

Conceptualization: Maria Panteli, Emmanouil Benetos, Simon Dixon.

Data curation: Maria Panteli, Emmanouil Benetos.

Formal analysis: Maria Panteli.

Investigation: Maria Panteli.

Methodology: Maria Panteli, Emmanouil Benetos.

Software: Maria Panteli.

Supervision: Emmanouil Benetos, Simon Dixon.

Validation: Maria Panteli.

Visualization: Maria Panteli.

Writing – original draft: Maria Panteli.

Writing – review & editing: Emmanouil Benetos, Simon Dixon.

References

1. Schedl M, Gómez E, Urbano J. Music Information Retrieval: Recent Developments and Applications. Foundations and Trends® in Information Retrieval. 2014; 8(2-3):127–261. https://doi.org/10.1561/1500000042
2. Lomax A. Folk song style and culture. American Association for the Advancement of Science; 1968.
3. Savage PE, Brown S, Sakai E, Currie TE. Statistical universals reveal the structures and functions of human music. Proceedings of the National Academy of Sciences of the United States of America (PNAS). 2015; 112(29):8987–8992. https://doi.org/10.1073/pnas.1414495112
4. Serrà J, Corral Á, Boguñá M, Haro M, Arcos JL. Measuring the Evolution of Contemporary Western Popular Music. Scientific Reports. 2012; 2. https://doi.org/10.1038/srep00521 PMID: 22837813
5. Mauch M, MacCallum RM, Levy M, Leroi AM. The evolution of popular music: USA 1960-2010. Royal Society Open Science. 2015; 2(5). https://doi.org/10.1098/rsos.150081 PMID: 26064663
6. Tzanetakis G, Kapur A, Schloss Andrew W, Wright M. Computational Ethnomusicology. Journal of Interdisciplinary Music Studies. 2007; 1(2):1–24.
7. Gómez E, Herrera P, Gómez-Martin F. Computational Ethnomusicology: perspectives and challenges. Journal of New Music Research. 2013; 42(2):111–112. https://doi.org/10.1080/09298215.2013.818038
8. Abdallah S, Benetos E, Gold N, Hargreaves S, Weyde T, Wolff D. The Digital Music Lab: A Big Data Infrastructure for Digital Musicology. ACM Journal on Computing and Cultural Heritage. 2017; 10(1).
9. Serra X. A Multicultural Approach in Music Information Research. In: International Society for Music Information Retrieval Conference; 2011. p. 151–156.
10. Fillon T, Simonnot J, Mifune MF, Khoury S, Pellerin G, Le Coz M, et al. Telemeta: An open-source web framework for ethnomusicological audio archives management and automatic analysis. In: 1st International Digital Libraries for Musicology workshop (DLfM 2014); 2014. p. 1–8.
11. Kroher N, Díaz-Báñez JM, Mora J, Gómez E. Corpus COFLA: A Research Corpus for the Computational Study of Flamenco Music. Journal on Computing and Cultural Heritage. 2016; 9(2):10:1–10:21. https://doi.org/10.1145/2875428
12. Moelants D, Cornelis O, Leman M. Exploring African Tone Scales. In: Proceedings of the International Society for Music Information Retrieval Conference; 2009. p. 489–494.
13. Aggarwal CC, Yu PS. Outlier detection for high dimensional data. In: International Conference on Management of Data (ACM SIGMOD); 2001. p. 37–46.
14. Panteli M, Dixon S. On the evaluation of rhythmic and melodic descriptors for music similarity. In: Proceedings of the International Society for Music Information Retrieval Conference; 2016. p. 468–474.
15. Panteli M, Benetos E, Dixon S. Learning a feature space for similarity in world music. In: Proceedings of the International Society for Music Information Retrieval Conference; 2016. p. 538–544.
16. Panteli M, Benetos E, Dixon S. Automatic detection of outliers in world music collections. In: Analytical Approaches to World Music; 2016. p. 1–4.
17. Lomax A. Cantometrics: An Approach to the Anthropology of Music. Berkeley: Uneversity of California Extension Media Center; 1976.
18. Brown S, Savage PE, Ko AMS, Stoneking M, Ko YC, Loo JH, et al. Correlations in the population structure of music, genes and language. Proceedings of the Royal Society of London B: Biological Sciences. 2013; 281 (1774). https://doi.org/10.1098/rspb.2013.2072
19. Nettl B, Stone RM, Porter J, Rice T, editors. The Garland Encyclopedia of World Music. 1998th ed. New York: Garland Pub; 1998.
20. Savage PE, Merritt E, Rzeszutek T, Brown S. CantoCore: A new cross-cultural song classification scheme. Analytical Approaches to World Music. 2012; 2(1):87–137.
21. Rzeszutek T, Savage PE, Brown S. The structure of cross-cultural musical diversity. Proceedings of the Royal Society B-Biological Sciences. 2012; 279(1733):1606–1612. https://doi.org/10.1098/rspb.2011.1750
22. Savage PE, Brown S. Mapping Music: Cluster Analysis Of Song-Type Frequencies Within And Between Cultures. Ethnomusicology. 2014; 58(1):133–155. https://doi.org/10.5406/ethnomusicology.58.1.0133
23. Le Bomin S, Lecointre G, Heyer E. The evolution of musical diversity: The key role of vertical transmission. PLoS ONE. 2016; 11(3). https://doi.org/10.1371/journal.pone.0151570 PMID: 27027305
24. Shalit U, Weinshall D, Chechik G. Modeling Musical Influence with Topic Models. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13); 2013. p. 244–252.

25. Tzanetakis G, Cook P. Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing. 2002; 10(5):293–302. https://doi.org/10.1109/TSA.2002.800560
26. Pampalk E, Flexer A, Widmer G. Improvements of Audio-Based Music Similarity and Genre Classification. In: Proceedings of the International Symposium on Music Information Retrieval; 2005. p. 634–637.
27. Fu Z, Lu G, Ting KM, Zhang D. Music classification via the bag-of-features approach. Pattern Recognition Letters. 2011; 32(14):1768–1777. https://doi.org/10.1016/j.patrec.2011.06.026
28. Gómez E, Haro M, Herrera P. Music and geography: Content description of musical audio from different parts of the world. In: Proceedings of the International Society for Music Information Retrieval Conference; 2009. p. 753–758.
29. Liu Y, Xiang Q, Wang Y, Cai L. Cultural style based music classification of audio signals. In: IEEE International Conference on Acoustics, Speech, and Signal Processing; 2009. p. 57–60.
30. Kruspe A, Lukashevich H, Abeßer J, Großmann H, Dittmar C. Automatic Classification of Musical Pieces Into Global Cultural Areas. In: AES 42nd International Conference; 2011. p. 1–10.
31. Zhou F, Claire Q, King RD. Predicting the Geographical Origin of Music. In: IEEE International Conference on Data Mining; 2014. p. 1115–1120.
32. Fu Z, Lu G, Ting KM, Zhang D. A survey of audio-based music classification and annotation. IEEE Transactions on Multimedia. 2011; 13(2):303–319. https://doi.org/10.1109/TMM.2010.2098858
33. Serrà J, Gómez E, Herrera P. Audio cover song identification and similarity: background, approaches, evaluation, and beyond. In: Advances in Music Information Retrieval. Springer Berlin Heidelberg; 2010. p. 307–332.
34. Bello JP. Measuring structural similarity in music. IEEE Transactions on Audio, Speech and Language Processing. 2011; 19(7):2013–2025. https://doi.org/10.1109/TASL.2011.2108287
35. Collins T, Arzt A, Frostel H, Widmer G. Using Geometric Symbolic Fingerprinting to Discover Distinctive Patterns in Polyphonic Music Corpora. In: Meredith D, editor. Computational Music Analysis. Springer International Publishing; 2016. p. 445–474.
36. Celma Ò. Music Recommendation. In: Music Recommendation and Discovery. Springer Berlin Heidelberg; 2010. p. 43–85.
37. Downie JS, Ehmann AF, Bay M, Jones MC. The Music Information Retrieval Evaluation eXchange: Some Observations and Insights. Advances in Music Information Retrieval. 2010; 274:93–115.
38. Honingh A, Panteli M, Brockmeier T, López Mejía DI, Sadakata M. Perception of Timbre and Rhythm Similarity in Electronic Dance Music. Journal of New Music Research. 2015; 44(4):373–390. https://doi.org/10.1080/09298215.2015.1107102
39. Müllensiefen D, Frieler K. Melodic Similarity: Approaches and Applications. In: Proceedings of the 8th International Conference on Music Perception and Cognition; 2004. p. 1–7.
40. Typke R. Music Retrieval based on Melodic Similarity. Utrecht University; 2007.
41. Schmuckler MA. Melodic Contour Similarity Using Folk Melodies. Music Perception. 2010; 28(2): 169–194. https://doi.org/10.1525/mp.2010.28.2.169
42. Foote J, Cooper ML, Nam U. Audio Retrieval by Rhythmic Similarity. In: Proceedings of the International Society for Music Information Retrieval Conference; 2002. p. 265–266.
43. Dixon S, Gouyon F, Widmer G. Towards Characterisation of Music via Rhythmic Patterns. In: Proceedings of the International Symposium on Music Information Retrieval; 2004. p. 509–516.
44. Guastavino C, Gómez F, Toussaint G, Marandola F, Gómez E. Measuring Similarity between Flamenco Rhythmic Patterns. Journal of New Music Research. 2009; 38(2):129–138. https://doi.org/10.1080/09298210903229968
45. Toiviainen P, Tervaniemi M, Louhivuori J, Saher M, Huotilainen M, Näätänen R. Timbre Similarity: Convergence of Neural, Behavioral, and Computational Approaches. Music Perception. 1998; 16(2): 223–241. https://doi.org/10.2307/40285788
46. Pachet F, Aucouturier JJ. Improving timbre similarity: How high is the sky? Journal of negative results in speech and audio sciences. 2004; 1(1):1–13.
47. McAdams S. Musical Timbre Perception. In: The Psychology of Music. Elsevier Inc.; 2013. p. 35–67.
48. Haas WBD, Rohrmeier M, Wiering F. Modeling Harmonic Similarity using a Generative Grammar of Tonal Harmony. In: Proceedings of the International Society for Music Information Retrieval Conference; 2009. p. 549–554.
49. Müller M, Ewert S. Towards timbre-invariant audio features for harmony-based music. IEEE Transactions on Audio, Speech and Language Processing. 2010; 18(3):649–662. https://doi.org/10.1109/TASL.2010.2041394

50. Wolff D, Weyde T. Adapting Metrics for Music Similarity Using Comparative Ratings. In: Proceedings of the International Society for Music Information Retrieval Conference; 2011. p. 73–78.
51. Sturm BL. Classification accuracy is not enough. Journal of Intelligent Information Systems. 2013; 41(3):371–406. https://doi.org/10.1007/s10844-013-0250-y
52. Flexer A, Grill T. The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. Journal of New Music Research. 2016; 45(3):239–251. https://doi.org/10.1080/09298215.2016.1200631 PMID: 28190932
53. Ben-Gal I. Outlier Detection. In: Data Mining and Knowledge Discovery Handbook. New York: Springer-Verlag; 2005. p. 131–146.
54. Casas P, Mazel J, Owezarski P. Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge. Computer Communications. 2012; 35(7):772–783. https://doi.org/10.1016/j.comcom.2012.01.016
55. Bhattacharyya S, Jha S, Tharakunnel K, Westland JC. Data mining for credit card fraud: A comparative study. Decision Support Systems. 2011; 50(3):602–613. https://doi.org/10.1016/j.dss.2010.08.008
56. Podgorelec V, Heričko M, Rozman I. Improving mining of medical data by outliers prediction. In: Proceedings—IEEE Symposium on Computer-Based Medical Systems; 2005. p. 91–96.
57. Chen D, Lu CT, Kou Y, Chen F. On detecting spatial outliers. GeoInformatica. 2008; 12(4):455–475. https://doi.org/10.1007/s10707-007-0038-8
58. Lu CT, Kou Y, Zhao J, Chen L. Detecting and tracking regional outliers in meteorological data. Information Sciences. 2007; 177(7):1609–1632. https://doi.org/10.1016/j.ins.2006.09.013
59. WongWK, MooreA, CooperG, WagnerM. Rule-based anomaly pattern detection for detecting disease outbreaks. Proceedings of the Eighteenth National Conference on Artificial Intelligence. 2002; p. 217–223.
60. Grubesic TH. On the application of fuzzy clustering for crime hot spot detection. Journal of Quantitative Criminology. 2006; 22(1):77–105. https://doi.org/10.1007/s10940-005-9003-6
61. Bountouridis D, Koops HV, Wiering F, Veltkamp RC. Music Outlier Detection Using Multiple Sequence Alignment and Independent Ensembles. In: Similarity Search and Applications. vol. 9939. lecture no ed. Springer International Publishing; 2016. p. 286–300.
62. Lu Y, Wu C, Lu C, Lerch A. Automatic outlier detection in music genre datasets. International Society for Music Information Retrieval. 2016; p. 101–107.
63. Hansen LK, Lehn-Schi T, Petersen K. Learning and clean-up in a large scale music database. In: European Signal Processing Conference; 2007. p. 946–950.
64. Livshin A, Rodet X. Purging musical instrument sample databases using automatic musical instrument recognition methods. IEEE Transactions on Audio, Speech and Language Processing. 2009; 17(5): 1046–1051. https://doi.org/10.1109/TASL.2009.2018439
65. Titon JT, Cooley TJ, Locke D, McAllester DP, Rasmussen AK. Worlds of Music: An Introduction to the Music of the World's Peoples. Belmont: Schirmer Cengage Learning; 2009.
66. Bohlman PV. World Music: A Very Short Introduction. Oxford University Press; 2002.
67. Smithsonian Folkways Recordings. Smithsonian Institution;. Available from: http://www.folkways.si.edu/folkways-recordings/smithsonian.
68. World and Traditional Music. British Library Sounds;. Available from: http://sounds.bl.uk/World-and-traditional-music.
69. Nettl B. Review of Folk Song Style and Culture by Alan Lomax Source. American Anthropologist, New Series. 1970; 72(2):438–441. https://doi.org/10.1525/aa.1970.72.2.02a00600
70. Fink R. Big (Bad) Data; 2013. Available from: http://musicologynow.ams-net.org/2013/08/big-bad-data.html.
71. Clarke D. On Not Losing Heart: A Response to Savage and Brown's "Toward a New Comparative Musicology". Analytical Approaches to World Music. 2014; 3(2):1–14.
72. Trehub SE. Cross-cultural convergence of musical features. Proceedings of the National Academy of Sciences of the United States of America. 2015; 112(29):8809–8810. https://doi.org/10.1073/pnas.1510724112 PMID: 26157132
73. Tzanetakis G, Cook P. MARSYAS: a framework for audio analysis. Organised Sound. 2000; 4(3): 169–175. https://doi.org/10.1017/S1355771800003071
74. Peeters G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical Report IRCAM. 2004;.

75. Lartillot O, Toiviainen P. A Matlab Toolbox for Musical Feature Extraction From Audio. In: International Conference on Digital Audio Effects; 2007. p. 237–244.

76. McFee B, McVicar M, Raffel C, Liang D, Nieto O, Battenberg E, et al. librosa: 0.4.1; 2015. Available from: http://dx.doi.org/10.5281/zenodo.32193.

77. Hamel P, Eck D. Learning Features from Music Audio with Deep Belief Networks. In: International Society for Music Information Retrieval Conference. Ismir; 2010. p. 339–344.

78. Choi K, Fazekas G, Sandler M. Automatic tagging using deep convolutional neural networks. In: International Society for Music Information Retrieval Conference; 2016. p. 805–811.

79. Scheirer E, Slaney M. Construction and evaluation of a robust multifeature speech/music discriminator. In: IEEE International Conference on Acoustics, Speech and Signal Processing; 1997. p. 1331–1334.

80. El-Maleh K, Klein M, Petrucci G, Kabal P. Speech/music discrimination for multimedia applications. In: IEEE International Conference on Acoustics, Speech and Signal Processing; 2000. p. 2445–2448.

81. Panagiotakis C, Tziritas G. A speech/music discriminator based on RMS and zero-crossings. IEEE Transactions on Multimedia. 2005; 7(1):155–166. https://doi.org/10.1109/TMM.2004.840604

82. Downie JS. The Music Information Retrieval Evaluation Exchange (2005-2007): A window into music information retrieval research. Acoustical Science and Technology. 2008; 29(4):247–255. https://doi.org/10.1250/ast.29.247

83. Marolt M. Music/speech classification and detection submission for MIREX 2015. In: MIREX; 2015. p. 1.

84. Marolt M. Probabilistic Segmentation and Labeling of Ethnomusicological Field Recordings. In: International Society for Music Information Retrieval Conference; 2009. p. 75–80.

85. Sadie S, Tyrrell J, Levy M. The New Grove Dictionary of Music and Musicians. Oxford University Press; 2001.

86. Holzapfel A, Stylianou Y. Scale Transform in Rhythmic Similarity of Music. IEEE Transactions on Audio, Speech, and Language Processing. 2011; 19(1):176–185. https://doi.org/10.1109/TASL.2010.2045782

87. Van Balen J, Bountouridis D, Wiering F, Veltkamp R. Cognition-inspired Descriptors for Scalable Cover Song Retrieval. In: Proceedings of the International Society for Music Information Retrieval Conference; 2014. p. 379–384.

88. Bartsch MA, Wakefield GH. Audio thumbnailing of popular music using chroma-based representations. IEEE Transactions on Multimedia. 2005; 7(1):96–104. https://doi.org/10.1109/TMM.2004.840597

89. Aucouturier JJ, Pachet F, Sandler M. "The way it sounds": Timbre models for analysis and retrieval of music signals. IEEE Transactions on Multimedia. 2005; 7(6):1028–1035. https://doi.org/10.1109/TMM.2005.858380

90. Holzapfel A, Flexer A, Widmer G. Improving tempo-sensitive and tempo-robust descriptors for rhythmic similarity. In: Proceedings of the Sound and Music Computing Conference; 2011. p. 247–252.

91. Marchand U, Peeters G. The modulation scale spectrum and its application to rhythm-content description. In: International Conference on Difital Audio Effects; 2014. p. 167–172.

92. Schörkhuber C, Klapuri A, Holighaus N, Dörfler M. A Matlab Toolbox for Efficient Perfect Reconstruction Time-Frequency Transforms with Log-Frequency Resolution. In: AES 53rd Conference on Semantic Audio; 2014. p. 1–8.

93. Salamon J, Gómez E. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. IEEE Transactions on Audio, Speech, and Language Processing. 2012; 20(6):1759–1770. https://doi.org/10.1109/TASL.2012.2188515

94. Sun L, Ji S, Ye J. Multi-Label Dimensionality Reduction. CRC Press, Taylor & Francis Group; 2013.

95. Chen J, Sathe S, Aggarwal C, Turaga D. Outlier Detection with Autoencoder Ensembles. In: Proceedings of the 2017 SIAM International Conference on Data Mining.; 2017. p. 90–98.

96. Lee H, Yoo J, Choi S. Semi-Supervised Nonnegative Matrix Factorization. IEEE Signal Processing Letters. 2010; 17(1):4–7. https://doi.org/10.1109/LSP.2009.2027163

97. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.

98. Hodge V, Austin J. A Survey of Outlier Detection Methodologies. Artificial Intelligence Review. 2004; 22(2):85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

99. Filzmoser P. A Multivariate Outlier Detection Method. In: International Conference on Computer Data Analysis and Modeling; 2004. p. 18–22.

100. Filzmoser P, Maronna R, Werner M. Outlier identification in high dimensions. Computational Statistics and Data Analysis. 2008; 52(3):1694–1711. https://doi.org/10.1016/j.csda.2007.05.018

101. Kelso NV, Patterson T. Natural Earth;. Available from: http://www.naturalearthdata.com.

102. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987; 20(C):53–65. https://doi.org/10.1016/0377-0427(87)90125-7

103. Johnson SC. Hierarchical clustering schemes. Psychometrika. 1967; 32(3):241–254. https://doi.org/10.1007/BF02289588 PMID: 5234703

104. Powers DMW. Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies. 2011; 2(1):37–63.

105. Schnitzer D, Flexer A, Schedl M, Widmer G. Local and Global Scaling Reduce Hubs in Space. Journal of Machine Learning Research. 2012; 13:2871–2902.

106. Dieleman S, Schrauwen B. Multiscale Approaches To Music Audio Feature Learning. In: International Society for Music Information Retrieval Conference; 2013. p. 116–121.

107. Humphrey EJ, Glennon AP, Bello JP. Non-linear semantic embedding for organizing large instrument sample libraries. In: Proceedings of the International Conference on Machine Learning and Applications. vol. 2; 2011. p. 142–147.

108. Reed J, Lee CH. On the importance of modeling temporal information in music tag annotation. In: IEEE International Conference on Acoustics, Speech and Signal Processing; 2009. p. 1873–1876.

109. Sturm BL. Revisiting priorities: improving MIR evaluation practices. In: Proceedings of the International Society for Music Information Retrieval Conference; 2016. p. 488–494.