Automatic Geocoding and Dating of Music Based on Audio Content

Mauro Domingues Teles
Instituto Superior Técnico
mauro.teles@tecnico.ulisboa.pt

***Abstract*—Throughout history, cultural interchanges have been a driving force for music to evolve. Previous studies have shown that these cultural traces are imprinted in music as information, accessible to be extracted and learned by automatic techniques. In this work, I address the tasks of predicting the geographical coordinates where a song was made and predicting the release year of songs. I specifically propose a deep neural network architecture with two main components. The first component is recurrent, and is capable of leveraging the sequential properties in music. The second component is neural attention, a mechanism that allows focusing on particular moments of the song, leveraging the context of a musical moment. In my experiments I use the Million Song Dataset, a collection of audio descriptors and meta-data for a million popular songs. The experimental evaluation shows that the proposed method performs better than baseline methods in the task of predicting the release year of a song, and has poor performance in most cases for the task of predicting the geographical location.**

1. Introduction

When discussing music it is not uncommon to use adjectives such as it being classical, African or oriental. However, it is not trivial to code these subjective definitions into a machine.

Recently, following the adoption of digital technology, computer scientists are working closely with musicologists in researching large scale music processing tools to support and enable their studies [1, 2, 3]. The work by French [4], where the author studies the geography of American rap music, is a practical example of a domain where musicologists might benefit from automated techniques to gather sociological insights.

Songs are multidimensional pieces which may be represented by properties such as rhythm, timbre, or pitch, and these are clues that can be used to distinguish between musical categories. Automatically classifying music into categories can be addressed using machine learning techniques [5]. These are computer algorithms that automatically learn to do a task from experience, in this case learning from audio or from features computed from the audio signal.

This work concerns with investigating whether some of the most promising machine learning algorithms can capture the high level semantics of geographical origin and year of creation for musical pieces. In previous studies, machine learning was already applied to similar problems. For instance, Zhou et al. [6] demonstrated the relationship between audio content and geographical coordinates, while Bertin-Mahieux et al. [7] addressed the task of predicting the release year of a song. However, given that these previous studies used simple algorithms or datasets with a few songs, there is an opportunity to study this task further.

I specifically use the Million Song Dataset, a collection of songs annotated with city of origin for the composing artist and release year. The audio descriptors contained in the MSD were calculated on a segment basis, that is, the features were extracted in short time intervals and each track is represented by a sequence of descriptors.

The experiments described in the present document include reproducing previous results with a baseline method, the K-Nearest Neighbors, an algorithm that leverages the overall song similarity, estimating similar labels for similar songs. Additionally, I propose a recurrent neural network architecture inspired on the success of previous works on similar tasks [5, 8]. The first part of the model consists of Gated Recurrent Units (GRU) [9], a neural unit capable of processing audio descriptors in order to capture sequential properties. The second part consists of a neural Attention mechanism [10], a trainable layer that allows the model to focus on the most relevant moments of the song. The rich representation learned by these units is passed to densely connected layers that output a single value for year regression or a class label for region classification.

This document reports on experiments with subsets of the Million Song Dataset with 396,070 examples for the task geo-coding and 208,503 examples for the task of predicting the release year. I measured the results in terms of the mean absolute error. The proposed neural network performed better than the baseline in both tasks, achieving a mean absolute error of 3203 km in the first task and 7.02 years in the second task.

The remainder of this document is structured as follows: Section 2 surveys previous related work on geocoding and dating music and also reviews relevant works on similar tasks. Section 3 presents the proposed approach, starting with fundamental concepts and then detailing the neural architecture. Section 4 presents the experimental setup. Section 5 presents the evaluation methodology and the obtained results. Finally, Section 6 summarizes the main conclusions and presents possible directions for future work.

2. Related Work

This section contextualizes the present work by reviewing two previous studies that address the task of predicting the geographical origin of music and another work that addresses the task of predicting the release year of songs.

Following that, I review other recent works that use state-of-the-art models and techniques for geocoding multimedia objects.

2.1. Geocoding and Dating Music Pieces

In Zhou et al. [6], the authors used supervised learning methods to address the task of predicting the geographical origin of a musical piece. This problem was modeled as a regression of latitude and longitude coordinates. The dataset used in the experiments contained 1,059 songs, computationally represented by timbral and chromatic features. In this study, the machine learning methods considered were the k-nearest neighbors method and random forest regression. In the case of random forest regression, three criteria were studied: variance, standard deviation, and absolute deviation. To evaluate the models, the authors considered the mean error distance, obtained by calculating the mean of the great circle distance between the predicted coordinates and the ground-truth locations. The main conclusions of the experiments were as follows: first, random forest regression using variance as partitioning criteria results in the most accurate mean prediction of 3,113 km. Second, random forest regression performed better than nearest neighbor interpolation for every partitioning criteria. Finally, the authors also observed that using chromatic features did not increase the performance of the model.

Following up on the aforementioned study, Schedl and Zhou [11] investigated how to improve results in the task of geocoding music. They proposed combining methods based on audio content with methods leveraging textual information automatically collected from the web. Regarding solely the audio-based prediction, they studied three sets of audio descriptors: block-level features, including spectral pattern, delta spectral pattern, variance delta spectral pattern, logarithmic fluctuation pattern, correlation pattern, and spectral contrast pattern, timbral features (as in Zhou et al. [6]), and timbral features with added chromatic features (as in Zhou et al. [6]). A k-nearest neighbor method is used to compare the listed sets of features. The estimated coordinates are obtained by calculating the geodesic midpoint of the K nearest examples, while the error distance is obtained through the great circle distance. The dataset and evaluation setup were the same as in Zhou et al. [6]. In this context, block-level features outperformed other features for every value of K, achieving the smallest mean error of $2,191$ km for $K = 1$. The second best obtained mean error was $3,410$ km when using only timbral features.

The Million Song Dataset[1] (MSD) is an open-source collection of features and meta-data for a million contemporary popular songs. This dataset provided researchers in the field of music information retrieval with a large scale music dataset, something that was not easily accessible up to that point. The MSD contains the following groups of descriptors:

- Time marks of beats, bars and tatums, which mark the rhythm of a song and are multiple of each other.
- Segment descriptors. A segment is defined as set of sound entities (typically under a second), each relatively uniform in timbre and harmony. Each segment is described by timbre, pitch, loudness and confidence.
- Estimated key, tempo, mode and time signature for each song.
- Miscellaneous information about each song such as title, duration and year of release.
- Miscellaneous information about each artist such as latitude, longitude and MusicBrainz[2] tags created by human users.

For additional information refer to the Echonest Analyzer Documentation.

In an introductory paper of the MSD, Bertin-Mahieux et al. [7] studied predicting the release year of songs. Overall, the subset used in the experiments contained 515,576 songs, from 28,223 artists. The features considered to represent the audio were the average and covariance of timbre vectors (12 audio texture features). The authors compared the performance of two different methods, the k-nearest neighbor approach, and a linear method, similar to the perceptron algorithm. The evaluation was performed by calculating the mean difference between the predicted years and the ground-truth years. Linear regression holds the best performance, with a mean difference of 6.14 years, followed by k-nearest neighbor method, which achieved 7.58 years in the same measure, for $K = 50$.

2.2. Deep Learning Methods for Geocoding or Dating Multimedia Objects

In the following Section, I review several previous works that use artificial neural networks to encode multimedia objects like music, text and images. Additionally, some of these works overlap with the task of geocoding, particularly relevant for the present work.

Neural networks are mathematical constructs that consume inputs and automatically tune a set of parameters in order to optimize a target function that is dependent on the output of the network and the expected values.

Choi et al. [5] reported good results on the task of predicting high-level tags (e.g., genre or mood) from music using an hybrid network consisting of a Convolutional Neural Network (CNN) feeding to a Recurrent Neural Network. The main idea behind this combined architecture is that the CNN works as a local feature extractor searching for patterns in the audio spectrum, while the RNN does temporal summarization. In this work, the *Convolutional Recurrent Neural Network* (CRNN) is compared with three other CNN architectures. These three networks are distinguished by the shape of the convolutional kernel (or filter) and the convolution dimension, one-dimensional or two-dimensional. In the three baseline CNN models, the resulting feature maps are fed to fully connected layers. In the case of the CRNN architecture, the last convolutional layer connects to Gated Recurrent Units [9]. The authors compared the four aforementioned architectures in the task of predicting the top-50 song tags from a subset of the MSD. The input of each network consisted of 96 log-amplitude Mel-spectrogram bins derived from raw audio. The experimental evaluation concluded that the CRNN model tends to perform better when controlling

1. https://labrosa.ee.columbia.edu/millionsong/

2. https://musicbrainz.org/

the number of parameters, arguably because of the ability of recurrent units to summarize features over time.

Weninger and Eyben [12] proposed using Long-Short Term Memories (LSTMs) [13], a type of recurrent neural network, to perform continuous-time regression of emotional valence and arousal in music. The features used to represent audio content were derived by applying statistics like the standard deviation and percentiles, to low level features such as chroma, MFCCs, and energy. The LSTM model was compared to Feed-Forward Neural Networks (FFNs) and Support Vector Regression (SVR) in the task of predicting the arousal and valence independently. In an experimental evaluation with one 1,000 songs, annotated by hand and balanced by genre, the authors evaluated three measures: the determination coefficient, the mean linear error, and the average Kendall's Tau per song. The LSTM model performed better than the SVR, obtaining an average improvement of 0.2 in the determination coefficient and a slightly lower mean linear error, for both arousal and valence.

There is an emerging body of work relying on neural networks to model multimedia artifacts for the task of geolocation [14, 15, 16, 17]. In the following paragraph, I briefly review some of these works, highlighting relevant mechanisms used for geographical estimation.

The first example is the work by Weyand et al. [14] where the objective is to estimate the place where a photo was taken. The two main aspects of the solution are: the representation of the earth's surface, and the Convolutional Neural Network (CNN) architecture. To discretize the earth's surface, the authors use a method that divides the world into quadrilateral cells. The size of cells is dictated by the number samples in the following way, the more images there are for a certain location, the smaller the cell surface. The CNN used in this work (PlaNet) was based on the Inception architecture [18], outputting the probability of an image belonging to a certain cell. The proposed model is able to localize 10.1% of the images at city-level accuracy, and 28.4% at country level. Additionally, the authors explored the idea that an album of photos, i.e., a sequence of photos, encodes common geographical information. In this manner, when the model is uncertain about where a certain picture was taken, it can use other pictures in the same sequence to improve the prediction. Practically, this was done by feeding the output vector of the CNN architecture into an LSTM unit. With regard to training the networks, both were trained separately, first the CNN and then the LSTM. The CNN/LSTM was compared to the CNN from the first experiment, using a dataset containing 29.7M albums and 616M images. In the end, the combined CNN/LSTM model is able to localize 45.6% of the images at city level, and 79.3% at country level.

Following [14], Vo et al. [15] revisited the problem of photo geolocation [14, 19]. In this work, the Earth's surface is also partitioned into regions using an adaptive mechanism, and a CNN is trained in the task of region classification. However, the authors use an image retrieval approach where they index a large set of image features. At test time, they calculate the nearest neighbors of the query image (in feature space). Combining adaptive discretization, deep learning and kernel estimation resulted in a relevant performance increase at street, city, region, and country levels.

Also very recently, Iso et al. [16] proposed a different approach to estimate the location of places named in tweets. The core concept consists of modeling the target as a probability distribution, casting the problem as a density estimation. The authors use a Convolutional Neural Network to process the text and generate an hidden representation, outputting the parameters of a Gaussian mixture model. The proposed Convolutional Mixture Density Network (CMDN) is flexible in terms of being able to represent more rich information (e.g., expressing ambiguity) in contrast to a single regression value. The evaluation was based on the mean and median error distances between the predicted location (the mode value of estimated density) and the ground-truth location. In a test with 4,633,478 Japanese tweets, the CMDN model had state-of-the-art performance, achieving the second lowest mean error of 159.4 km and the lowest median value of 10.7 km.

While some works were already dedicated to the task of predicting the geographical origin and dating of songs, they use small datasets or simple machine learning methods. As reviewed in this section, recent works with multimedia objects rely on neural networks to model the audio content. The present work builds on ideas surveyed in this section, namely recurrent neural networks for processing sequences, the attention mechanism and the adaptive discretization of the globe.

3. The Proposed Approach

In order to construct a classifier that leverages audio content, I need to define computational representations of sound that are meaningful to the classification objective. In section 3.1, I briefly review five categories of audio features, as described by Yang and Chen [20]. Next, in section 3.2 I present the theoretical foundations for a deep neural network capable of predicting the geographical origin and date of songs.

Taking inspiration on previous works with multimedia objects [14, 15, 16], and particularly music [12, 5], I propose a recurrent neural network for processing sequential data, complemented with an attention mechanism. Figure 1 presents the proposed deep learning model, which is detailed in section 3.3.

3.1. Features for Representing Music

Energy Features are correlated with the sensation of excitement in music. The *total loudness* [21] is an example of this class of features that tries to account the way humans perceive sound, calculated by aggregating the specific loudness sensation coefficients, derived from the bark scale, a non-linear psycho-acoustical scale.

Spectrum Features are a family of statistics which describe the shape of the power spectrum, obtained by taking the Short-Time Fourier Transform (STFT) of an audio signal, often associated with timbre qualities. Timbre can be understood as the underlying properties that allow humans to distinguish the sources of two different sounds with similar pitch. The *Mel-Frequency Cepstral Coefficients* (MFCCs) offer a compact representation of sound, mostly

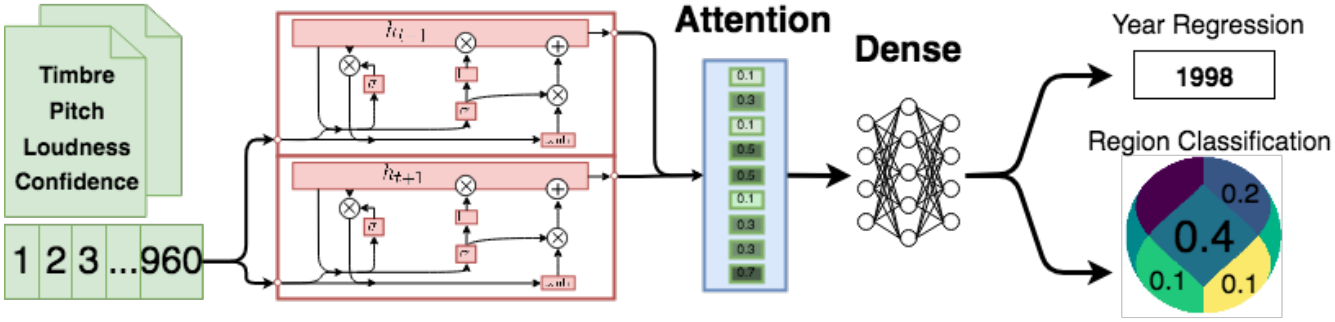


Figure 1: Representation of the proposed architecture.

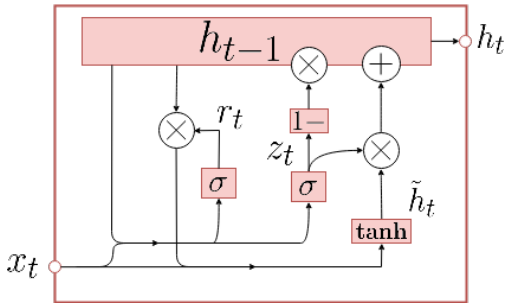
used in speech recognition [22]. The Mel scale is a non-linear scale of frequencies that takes the human perception of pitch into account. MFCCs can be computed by mapping frequencies to the mel scale, applying the logarithm to the power of each frequency, and then taking the discrete cosine transform, which outputs a set of coefficients. The first few dozens of coefficients are enough to represent voice/timbre characteristics, independent of speaker/source identity.

Finally, **Harmony Features** are related to pitch and to how different pitches blend together. Harmonic sounds are perceived as more pleasant than dissonant sounds. Such information can be analyzed by constructing a chromagram, by projecting the frequency spectrum onto twelve bins that correlate to the twelve semitones of the musical octave (i.e., the Western scale). The tonal centroid is an example of a feature belonging to this class and is calculated by projecting the chords along a chord progression scale (e.g., the circle of fifths).

3.2. Deep Neural Networks for Processing Sequential Data

The main computational unit of *Recurrent Neural Networks* (RNNs) is the recurrent neuron. The recurrent neuron is similar to the regular perceptron [23], complemented with a feed-back loop which causes the output of the neuron at time step $t+1$ to depend on the output at time step t. Recurrent Neural Networks are context sensitive and can be used to model input sequences of arbitrary length. Since the classic RNNs suffer from a problem of vanishing gradient, the *Long-Short Term Memory networks* [13] and *Gated Recurrent Units* [9] were proposed to model long sequences.

The main concept behind *Gated Recurrent Units* (GRU) is the gate. Gates are sets of trainable weights that can be combined with the input of time step t and the output of the previous time step $(t-1)$. A GRU has two gates, an update gate r, and a reset gate z. The update gate (z_t) defines which parts of the previous output will carry to the current state. The reset gate (r_t) can clear the context, so that the state is reset with the current input only. The output (h_t) is a combination on the mentioned gates.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right) \quad (1)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right) \quad (2)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right) \quad (3)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (4)$$

Figure 2: Representation of the Gated Recurrent Unit.

3.3. A Deep Neural Network for Geocoding or Dating Music Pieces

The proposed model uses two Gated Recurrent Units to generate an intermediate hidden representation of the audio descriptors. This unit allows the network to leverage the temporal structure of songs, being able to model rhythmic patterns or particular chord progressions.
RNNs are capable of leveraging information from previous time steps, however, information from forward steps may also be important to model a context, this can be achieved with a bidirectional [24] mechanism, a stack of RNNs, each processing the same input sequence in opposite directions. In my proposal the GRUs are bidirectional.
GRUs are connected to an Attention [10] layer. This mechanism allows a network to attend parts of an input sequence with more or less attention, effectively scoring what is most important for the prediction. A Recurrent Neural Network processes input vectors of length L, $x = (x_1, x_2, ..., x_{L-1}, x_L)$, into hidden representations $h = (h_1, h_2, ..., h_{L-1}, h_L)$. At each time step t, the output y of the attention layer, shown in equation 5, is generated by weighting the segments of the hidden state. The attention weights are trained similarly to a Multilayer Perceptron and

then normalized to sum to 1. See equations 6 and 7 where h is the input, w are the trainable weights and b is the bias parameter. At last, the result of the Attention layer is connected to dense layers, responsible for outputting the estimated target.

$$y_t = \sum_{j=1}^{L} h_{t,j} a_{t,j} \tag{5}$$

$$a_{t,j} = \frac{\exp(e_{t,j})}{\sum_{j=1}^{L} \exp(e_{t,j})} \tag{6}$$

$$e_{t,j} = \tanh\left(h \cdot w + b\right) \tag{7}$$

The proposed network is trained from end-to-end. In the case of estimating the geographical origin, the very last dense layer outputs two regression values, latitude and longitude, or the most likely cell, for region classification. In the first case (coordinate estimation), the network is trained with Vincenty's formulae [25], a differentiable function for calculating the distance between two points on the Earth, following the shortest path along the surface. In the case of region classification, the network is trained with categorical-crossentropy. In order to estimate the year of release, the only adaptation needed is to change the last dense layer to predict a single regression value and train the network with the loss function mean-squared-error.
I compare the proposed network to a shallow algorithm used in previous works [6, 11], the K-Nearest Neighbor (KNN), a method that I will use as baseline for comparison. The KNN can infer when and where a song was made based on the overall track similarity. In order to perform regression with the k-nearest neighbor method, we calculate the distance between the sample to predict and every other sample in the dataset. Then, considering the k closest, we aggregate their labels: in the case of predicting the year we simply calculate the mean year; in the case of coordinate regression, we can use the Geographical Midpoint Formula [26].

4. Experimental Setup

This section reports the experimental evaluation of the aforementioned machine learning algorithms in the tasks of predicting the release year and geographical origin of songs. The experiments performed are as follows:

- Predicting the release year of a song with:
  - The K-nearest neighbor method;
  - The proposed neural network model.
- Predicting the geographical origin of a song with:
  - The K-nearest neighbor method;
  - The proposed neural network model for latitude and longitude regression.
  - The proposed neural network model for region classification.

4.1. Features

From the set of audio descriptors available in the MSD, features extracted at segment level are particularly interesting given their sequential nature. For the rest of this document, the features considered in all experiments are the following:

- Segments timbre: 12 audio texture features (MFCCs), useful to distinguish between instruments.
- Segments pitches: 12 chromatic features, one value per semitone.
- Segments loudness maximum: maximum dB value, can be used to analyze how calm or busy a song is.
- Segments confidence: is a measure of how reliable the features extracted are.

For the experiments with neural networks, each track is trimmed to the mean segment length and padded with zeros, resulting in a sequence of 960 time steps, with 26 features each. In order to use sequential features with the k-nearest neighbor method, these features are aggregated in time, obtaining the mean and variance of each feature, resulting in a vector of 52 features per track.

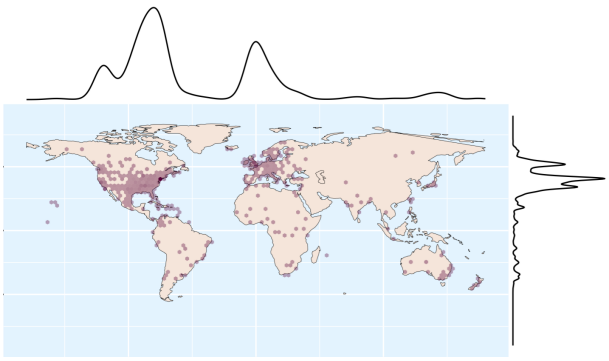


Figure 3: Distribution of tracks in the geographical subset.

4.2. Targets and Subsets

For the task of predicting the geographical origin of a song, I use the latitude and longitude of the location where the artist is associated. We should take into consideration that this target does not take into account if an artist lived in multiple countries. For the task of predicting the date of a song I simply use the release year.
Not every song in the MSD contains information about the release year, and not every artist is tagged with latitude and longitude, therefore, I created two different subsets, the Geographical Subset with 396,070 examples (Table 1 and Figure 3) and the Year Subset (Figure 4) with 208,503 examples.
For the tasks at hand, an ideal collection of music would contain many examples from multiple regions and eras. By observing the distribution of songs in the mentioned subsets it is noticeable that they are unbalanced, nevertheless, there are at least a thousand songs per continent. In this manner, we must keep in mind that we are dealing with a dataset that contains mostly Western contemporary popular music.
Drawing inspiration from previous works [14, 19, 15], I defined an adaptive discretization mechanism to convert geographical coordinates to bins. I use the HEALPix framework (Hierarchical Equal Area isoLatitude Pixelization) [27], which recursively partitions a sphere into geometrically

Table 1: Distribution of Songs per Continents in the Geographical Subset.

	Africa	Asia	Europe	North America	Oceania	South America
Track Count	2669	5565	105588	270114	6076	6058

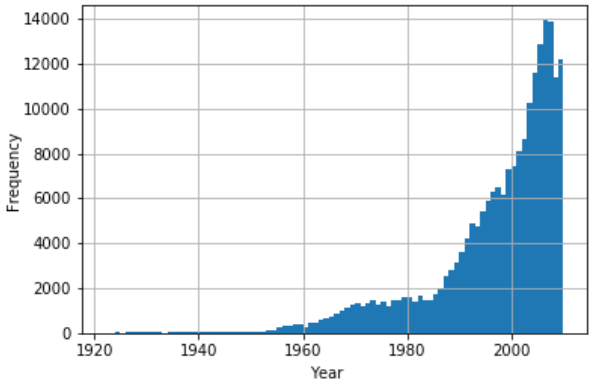


Figure 4: Frequency of songs per year in the year subset.

similar bins. The resolution of the grid can be controlled through a parameter N_{side}. Using this parameter I craft a mechanism to define smaller areas for regions with more samples.
The Geographical and Year subsets are divided into training and test splits with a ratio of 75%/25%, using stratified sampling based on the continent.

5. Obtained Results

For assessing the quality of the models I measure the mean and median absolute errors, in years and kilometers, according to the respective task. Besides that, I present the proportion of samples predicted within various ranges of error.

5.1. Predicting the Release Year

Table 2 shows the evaluation measures for both the K-Nearest Neighbors method and the proposed neural network architecture. While the KNN achieves a mean absolute error of 8.84 years for n = 50, the recurrent neural network augmented with an Attention mechanism achieves a mean absolute error of 7.02 years, outperforming the baseline.
Looking at Table 3, we can observe with more detail the prediction error per decade. The results are presented in the form of intervals of error. We can observe that songs from 1920 until 1960 are inaccurately predicted, while the model performs best for the music originated in the 80s, 90s and the first decade of 2000, predicting more than 50% of songs in a 10 year range (5 years before and after).

5.2. Predicting the Geographical Origin

The evaluation is performed by calculating the mean error distance between the ground truth coordinates and (i) the predicted coordinates, for regression, and (ii) the centroid of the estimated HealPix bin, for classification. Table 4 shows the results obtained.

The lowest mean error distance obtained with the KNN method is 3317 km for n = 80 neighbors. By comparing the original distribution of songs and the distribution of predicted coordinates, I find that the KNN method does not effectively learn to geo-locate music, the reason being that predictions are spread around a center point located at the highest frequency region (i.e., the east coast of the United States of America).
The mean error distance obtained with the proposed neural network architecture in the task of coordinate regression is 3810 km. With this setup, I also observe that the model tends to predict most songs around the highest frequency region, similarly to the results obtained with the KNN model.
Since most songs are originated from the East coast of the U.S.A and there is a large distance from that point to Europe or Asia, I use the same architecture in the task of region classification, circumventing the effect of predicting songs in the ocean.
The mean error distance obtained with the neural network architecture in the task of region classification is 3203 km. Table 5 allows a view of the classifier performance per continent. We can observe that the model is very inaccurate for regions outside North America, for example only 5% of songs from Europe are predicted within a 5,000 km radius. Table 5 also presents the error intervals for some representative countries, including some states from the United States of America. Southeastern American states like Louisiana and Mississippi have a considerable amount of correctly predicted songs (<500 km), the rest being confused with other American regions (estimated within a 2000 km radius). On the other hand, for Western states like California, most songs are predicted in a 5000 km radius, an indicator that the model is not able to separate music from this region.
Regarding other countries, Canadian songs are confused as being from the U.S.A. given the geographical proximity; European countries like Finland, Great Britain and Poland have around 10% of songs predicted in a reasonably close distance (<2,000 km); for countries other than the U.S.A, Mexico, Cuba, and particularly Jamaica present the best results, with a sizable portion of songs being placed within a 1000 km radius.

6. Conclusions and Future Work

In this document, I proposed a deep learning method for automatically estimating the year and geographical origin of songs. In the case of estimating the release year of a song, the results can be compared to the previous work by Bertin-Mahieux et al. [7], even though I use a different version of the MSD, with fewer songs. I obtained a mean absolute error of 7.02 year in a test with 208,503 samples while they achieved a mean absolute error of 6.14 years in a a test with 515,576 samples. In this manner I conclude that there is no particular advantage in using recurrent neural networks to leverage audio descriptors like timbre pitch and loudness.

Table 2: Performance Metrics in the task of Predicting the Release Year.

	Mean	Median	Median Confidence Interval at 95%
KNN	8.84 $\pm$ 0.6 years	7.11 years	[5.18, 9.14] years
Neural Network	7.02 $\pm$ 0.6 years	4.69 years	[3.22, 6.71] years

Table 3: Intervals of Error per Decade for the Neural Network Model.

	Absolute Year Difference per Interval				
Decade	<5 years	[5, 10] years	[10, 20] years	>20 years	Support
1920-1960	0%	0%	0%	100%	309
1960s	0%	0%	0.33%	0.67%	860
1970s	0.05%	0.17%	0.57%	0.21%	1671
1980s	0.53%	0.25%	0.21%	0.01%	2388
1990s	0.59%	0.33%	0.07%	0.01%	6602
2000s	0.67%	0.17%	0.15%	0.01%	12819
2010s	0.33%	0.44%	0.19%	0.04%	345

Table 4: Performance Metrics in the Task of Predicting the Geographical Origin.

Method	Mean	Median	Confidence Interval at 95 %
KNN	3317$\pm$ 28 km	2653 km	[1785, 3810] km
Neural Network Regression	3810 $\pm$ 32 km	2377 km	[1683, 4123] km
Neural Network Classification	3203 $\pm$ 19 km	2106 km	[1097, 3537] km

Table 5: Intervals of Error Distance per Country

	Error Distance in km						
Region	>500	[500, 1000]	[1000, 2000]	[2000, 5000]	[5000, 10000]	>10000	Support
Africa	0%	0.01%	0%	0.01%	45.6%	52.2%	662
Asia	0%	1.7%	0%	1.2%	19.3%	77.6%	1390
Europe	0.4%	1%	3%	0.6%	94.6%	0.1%	26405
North America	30.5%	20.6%	18.7%	28%	18%	0.1%	67510
Oceania	0%	0%	0%	0%	0.6%	99.3%	1519
South America	0%	0.1%	2.6%	22%	72.6%	1.9%	1529
Canada	29%	30%	12%	27%	2%	0%	3049
Jamaica	43%	0%	5%	51%	1%	0%	978
Mexico	18%	3%	3%	72%	2%	3%	616
Cuba	8%	2%	10%	77%	3%	0%	330
Finland	4%	6%	0%	1%	88%	0%	473
Great Britain	0%	0%	4%	0%	96%	1%	13165
Poland	1%	4%	7%	0%	87%	1%	204
U.S.A	31%	21%	20%	27%	2%	0%	62040
Alabama	13%	12%	74%	0%	1%	0%	649
Arkansas	5%	31%	62%	0%	1%	0%	766
California	0%	0%	0%	98%	1%	0%	10631
Louisiana	34%	0%	66%	0%	1%	0%	1513
Mississippi	45%	4%	51%	0%	0%	0%	1818

Regarding the problem of predicting the geographical origin of songs we can not directly compare the results to the previous works by Zhou et al. [6] and [11] since these works use much smaller collections. The performance of the proposed classifier is very limited in terms of accuracy, highlighting that Jamaican and Mexican music are easier to separate from the rest. The lacking performance can be due to one of two reasons: either the features do not serve for the task at hand or there is no significant relation between the audio content and the place where the song was released.
The main contributions of the present work include multiple experiments in the task of predicting the geographical origin of songs with the MSD. This work defines a baseline for future works to build upon.
For future work, following inspiration from recent works like Choi et al. [5], I envision using convolutional neural networks to extract features from raw audio instead of relying on pre-extracted features. Additionally, I propose formulating the problem as a density estimation, in the same manner as the work by Iso et al. [16]. Finally, instead of using only one location, we could automatically gather places associated to the artist, coming closer to model a geographical influence rather than a place where the artist was born.

References

[1] Samer Abdallah, Emmanouil Benetos, Nicolas Gold, Steven Hargreaves, Tillman Weyde, and Daniel Wolff. The digital music lab: A big data infrastructure for digital musicology. *Journal on Computing and Cultural Heritage*, 2016.

[2] Erik Duval, Marnix van Berchum, Anja Jentzsch, Gonzalo Alberto Parra Chico, and Andreas Drakos. Musicology of early music with europeana tools and services. *Proceedings of the Conference of the International Society for Music Information Retrieval*, 2015.

[3] Laurent Pugin. The challenge of data in digital musicology. *Frontiers in Digital Humanities*, 2, 2015.

[4] Kenneth French. Geography of American rap: rap diffusion and rap centers. *GeoJournal*, 82, 2017.

[5] Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional Recurrent Neural Networks for Music Classification. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[6] Fang Zhou, Q. Claire, and Ross D. King. Predicting the Geographical Origin of Music. *Proceedings of the IEEE International Conference on Data Mining*, 2014.

[7] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. *Proceedings of the Conference of the International Society for Music Information Retrieval*, 2012.

[8] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *Proceedings from the International Speech Communication Association Speech Synthesis Workshop*, 2016.

[9] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014.

[10] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the International Conference on Machine Learning*, 2015.

[11] Markus Schedl and Fang Zhou. Fusing web and audio predictors to localize the origin of music pieces for geospatial retrieval. *Proceedings of the European Conference on Information Retrieval*, 2016.

[12] Felix Weninger and Florian Eyben. On-line continuous-time music mood regression with deep recurrent neural networks machine intelligence. *Proceedings of the IEEE International Conference on Acoustic*, 2014.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9, 1997.

[14] Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet - Photo Geolocation with Convolutional Neural Networks. *Proceedings of the European Conference on Computer Vision*, 2016.

[15] Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[16] Hayate Iso, Shoko Wakamiya, and Eiji Aramaki. Density estimation for geolocation via convolutional mixture density network. *arXiv:1705.02750*, 2017.

[17] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-based localization using lstms for structured feature correlation. *IEEE International Conference on Computer Vision*, 2017.

[18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *Computer Vision and Pattern Recognition*, 2015.

[19] James Hays and Alexei A. Efros. im2gps: estimating geographic information from a single image. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[20] Yi-Hsuan Yang and Homer H. Chen. *Music Emotion Recognition*. CRC Press, Inc., 1st edition, 2011.

[21] Emmanouil Benetos, Margarita Kotti, and Constantine Kotropoulos. Large scale musical instrument identification. *Proceedings of the Sound and Music Computing Conferece*, 2007.

[22] Beth Logan. Mel Frequency Cepstral Coefficients for Music Modeling. *Proceedings of the International Symposium on Music Information Retrieval*, 2000.

[23] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 1958.

[24] Mike Schuster, Kuldip K. Paliwal, and A. General. Bidirectional recurrent neural networks. *Proceedings of the IEEE Transactions on Signal Processing*, 45, 1997.

[25] Thaddeus Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, 23(176), 2008.

[26] Jeff Jenness. Calculating areas and centroids on the sphere. *Proceedings of the International User Conference*, 2008.

[27] K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelman. HEALPix – a Framework for High Resolution Discretization, and Fast Analysis of Data Distributed on the Sphere. *The Astrophysical Journal*, 622, 2004.