Machine Learning as an Objective Approach to Understanding Musical Origin

Claire Q[1] and Ross King[2]

[1] Aberystwyth University, UK, ceq08@aber.ac.uk
[2] Manchester University, UK, ross.king@manchester.ac.uk

Abstract. Traditional research into the arts has almost always been based around the subjective judgment of human critics. We propose an alternative approach based on the use of objective machine learning programs. To illustrate this approach we investigated the distribution of music from around the world: geographical ethnomusicology. To ensure that our understanding of geographical ethnomusicology is objective and operational, we cast the problem as training a machine learning program to predict the geographical origin of pieces of music. We collected 1,142 pieces of music from 73 countries, and described them using 2 different sets of standard audio descriptors using MARSYAS. To predict the location of origin of the music we developed several methods designed to deal with the spherical surface topology based upon a modified k-nearest-neighbour. We investigate the utility of *a priori* geographical knowledge in the predictions: a land and sea mask, and a population distribution overlay. Our best-performing algorithm so far achieved a median land distance error of 1,506km, with comparable random trials having mean of medians 3,190km - this is significant at $P < 0.001$.

1 Introduction

1.1 An Objective Approach to Understanding Art

We hold the strong philosophical position that we do not fully understand a phenomenon unless we can make a machine that reproduces it: "What I cannot create, I do not understand" (written on Richard Feynman's blackboard at the time of his death). The advantage of this approach to understanding is that it is wholly objective and fully operational. This type of approach to understanding a phenomenon is taken by the AI community to be working towards understanding intelligence - they aim to develop intelligent machines. If they succeed in doing so it will be possible to objectively determine their success. This approach to understanding contrasts strikingly with that that of the traditional research into understanding art, which has almost always been based around the subjective judgment of human critics. Often great insight is gained by this subjective approach, but it also has to be granted that there are limitations to the results relying upon the peculiarities of the listener. We propose to extend the objective approach to understanding phenomena to art – in this case, music. Specifically

we will use the success of predictive machine learning programs as a measure of objective success in understanding a phenomenon. To illustrate and demonstrate our proposed objective approach to art we investigated the distribution of music from around the world: geographical ethnomusicology. To ensure that any understanding is objective the problem is cast as a machine learning one. This removes personal opinions and expectations in the listener because all decisions are made by machine.

1.2 Geographical Ethnomusicology

The world contains a vast variety of types of music. This music arose as the result of complex geographical, historical, and prehistorical processes. One way to better understand these processes is to analyse the current geographical distribution of music. The study of this distribution is termed Geographical Ethnomusicology. The problem of determining the geographical origin of a piece of music is complicated. Musical forms are rarely pure. Over time they have influenced each other, and many forms of music have travelled far from their point of origin. Western music's influence is nearly ubiquitous. The influence of other forms of music are also widely distributed, for example: Arabic musical influence spread all around the Indian Ocean, across North Africa, to Spain, to central Asia, etc.; more recently reggae has spread from Jamaica, to the UK, Brazil, Mauritius, etc. The question we wish to answer is given these complications, how well can a computer predict the geographical origin of a piece of music? It could be argued that unsupervised spatial clustering methods such as Kohonen nets [1] would be best suited to such a task. However, the problem with such clustering methods is that there is generally no objective measure of success. We could find groups of similar music in terms of the audio, but it would not extract those features most suited to predicting a location. This contrasts with supervised methods, where the labels on known examples (classes or numbers) enable the objective measuring of whether a method is working or not - does it predict well or badly? As we know the geographical location of origin of the music (to some degree) in our corpus we should exploit this information. We therefore cast the problem as that of training a machine learning program to be able to predict the geographical origin of pieces of music, i.e. the computer learns a functional relationship between the audio content and its geographic origin on the globe.This predictive task is possible to some extent by human musicologists.

1.3 Related Work

A large amount of research has been recently done on the development of audio features (attributes) for the computational analysis of music, e.g. [2]. These attributes have typically been used to perform automatic classification and clustering to identify similar pieces of music (for recommendation systems), e.g. to identify mood, genre, emotive content, and various other purposes for which it would be impossible to provide an exhaustive list [3]. In addition to audio attributes other meta-data have been utilised via, for example, web searches and

social tags, but also MIDI, score reading and lyric mining [4–7]. Various machine learning and statistical methods have been used upon these attributes: Support Vector Machines, k-Nearest-Neigbour, Neural Nets etc. with good success for certain applications [8]. Despite these advances little computational work has been done on computational ethnomusicology. Some few examples have been offered, such as Liu *et al.* which demonstrated the applicability of music analysis techniques to non-western music [9]. Gomez *et al.* have applied these techniques to classifying music as western or non-western with success, and also found some important features relating to the latitude and longitude of origin of a piece [10,11]. Tzanetakis' work on computational ethnomusicology [12] seeks to illustrate the potential application of music information retrieval (MIR) to ethnomusicology. This allows the analysis of large corpuses of music to obtain automatically features that would take copious time to transcribe by hand. Though the features are from signal processing they relate closely to how humans perceive music. One example is the spectral centroid, which is mathematically simple (it is the weighted mean of the frequencies in the signal) and yet is strongly correlated with human perception of 'brightness' in sound [13, 14].

Spatial statistics [15] consider the topology of the data as part or all of that which is to be measured, implying a connection between values and their spatial location. The most common applications are in statistical geography but its use in epidemiology is also well-known. One common example in statistics textbooks is the early work of John Snow who in 1855 proved that cholera was waterborne via statistical means and a geographical dot map showing the occurrences of cholera were clustered around a particular water pump. On a larger scale, geographic information systems including global positioning systems in recent years have created the discipline geospatial information studies wherein large databases of geographic information are analysed using geospatial relationships such as adjacency, containment and distance. Our work can be considered part of the latter category as a distance measure is the eventual output which measures success. Recent work in Self Organising Maps refers to the formation of geospatial shapes to avoid the edge problem, one example being GeoSOM [16]. Though we have already mentioned why clustering is unsuited to our prediction task, the difference in topology between a flat map of latitude and longitude and a spherical representation which wraps around is an important distinction.

2 Method

2.1 Music Collection

Our corpus was built from a personal collection of 1,142 tracks covering 73 countries.[3] The music used is traditional, ethnic or 'world' only, as classified by the publishers of the product on which it appears. We have not included any Western music as is naturally hard to place since its influence is global

[3] The music used is subject to copyright, but the processed data is not, and the data is to be made publicly available.

– what we seek are the aspects of music that most influence location. Thus, being able to specify a location with strong influence to the music is paramount. This will form the target function for the learning algorithm. To determine the geographical location of origin we manually collected the information from the CD sleeve notes, and when this information was inadequate we searched other information sources. There are most certainly other options as demonstrated by Govaerts *et. al.* but these have varying levels of accuracy and indeed their ground truth for the experiment was 'personal knowledge' or 'by looking up the origin' [17]. We did not wish to confound the ability of the predictor with incorrect location information. The location data is limited in precision to the country of origin - we did not have time to try to find out more about each track. In many cases the level of detailed precision possible for musical origin is arguably not much smaller than perhaps a region of a country, except in certain cases where a community has been extremely isolated.

The country of origin was determined by the artist's or artists' main country of residence. Of course, many artists live in different places throughout their lives, but our aim was to determine the major influence. For example, if a Malian writing Mali music lives in retirement in Paris, we consider the music Malian. We recognise that some music is less linked to countries than cultures, but countries at least have true geographical locations that are measurable - by virtue of having defined borders - allowing our machine learning approach to give objective output. We have taken the position of each country's capital city by latitude and longitude as the absolute point of origin in the beginning. The assumption here is that the political capital is also the cultural capital of the country. This assumption also utilises coarse a priori knowledge about population - most, but not all, countries have a highly populous capital city. Using the capital takes into account country-level population distribution in a simple way without resorting to time-consuming investigations into the exact place each artist spent most time. In the population distribution task we altered this to the centre of population, or population centroid, of each country, which is a fairer measure that takes into account skewed population distribution and capitals with low populations. Countries are linked to artists, not tracks.

It is clear that the country of *production* may have no bearing on the origin, since many world music CDs are made in Germany or the US despite the music coming from e.g. Kenya. The artist in question is usually the composer where known, or else the performer where the music is traditional to their home country. If several artists have made a contribution to the music, all substantial (not merely 'featuring') contributions are taken into account when deciding if it should be included. Any track that had ambiguous origin – whether because the artist's own origin was ambiguous, or many artists from different countries collaborated, or the track is a deliberate fusion of styles – was removed from the dataset. There are no "right answers" in such cases. For example, Bhangra music is a fusion of Indian Punjab culture, UK culture and hip-hop. Therefore to try to determine a single geographical location for a Bhangra track would be nonsensical as it has multiple sites. One could suggest that the geographical midpoint

of all the influences is the right answer, but this does not fully encapsulate the data and would result in a position for Bhangra music near Volgagrad!
Figure 1 shows the distribution of tracks per country, for the 20 best represented

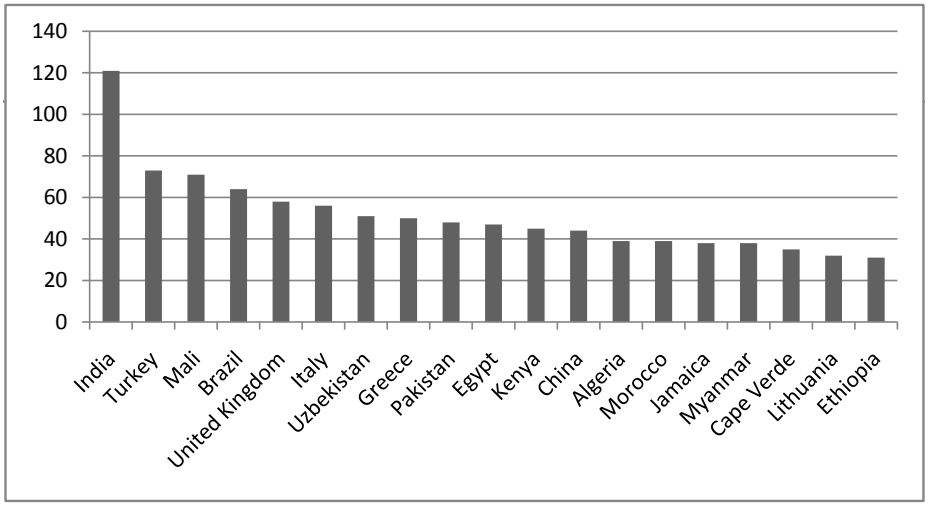

Fig. 1. Partial sample of music distribution by country.

countries. It can be seen that some countries are much better represented than others, which will affect the performance - more examples from a country gives more data about that country's music and thus better predictions.

2.2 Audio Features

The program MARSYAS [18] was used to extract audio features from the wave files. We first used the default MARSYAS settings in single vector format (68 features) to estimate the performance with basic timbral information covering the entire length of each track. Each of the default features is an indicator of timbre, which is one of the main ways (another being attack-decay-sustain-release models) [13] to distinguish musical instruments. Since instrumentation is also a major difference between cultural music traditions, these are appropriate to the task. No feature weighting or pre-filtering was applied. All numerical features (that is, all features) were transformed to have a mean of 0, and a standard deviation of 1. We also investigated the utility of adding chromatic attributes. These describe the notes of the scale being used. This is especially important as a distinguishing feature in geographical ethnomusicology – unlike Western music which largely conforms to one tuning system. The chromatic features provided by MARSYAS are 12 per octave – Western tuning, but it

may be possible to tell something from how similar to or different from Western tuning the music is.

2.3 Geographic Representation

The problem of predicting a point on the surface of a sphere is made more complicated as the standard coordinate system (latitudes and longitudes), which are the natural targets for regression, have a complicated relationship to surface area where the predicted point will be. This is illustrated in the standard Mercator projection of the globe where countries near the poles are unnaturally large. Area is not preserved equally on such projections. In general there is no perfect flat projection - a compromise is made for some parameter or parameters in favour of some desired quality, perhaps straight lines of latitude and longitude, perhaps equal area, even such complicated solutions as the butterfly map by Cahill later re-imagined by Waterman. None of these addresses the true mathematical problem fully - the Earth is not flat, and should therefore ideally be treated as close as possible to its true topology. In considering the sparsity of our data, we choose a sphere as an approximate representation for the globe, though with more precision still it is an oblate spheroid with certain peaks and troughs across the surface.

2.4 Spherical k-Nearest Neighbour Prediction Method

We decided to cast the problem as a regression problem (predicting a point) rather than a classification problem (predicting a country) because the large number of countries, and low number of examples per country, would complicate classification. Most regression methods assume either that either only one real number is to be predicted, or if multiple real numbers are to be predicted that they are independent. Perhaps the simplest approach to running regression with spherical coordinates is to side-step this difficulty and use a k-Nearest Neighbour method to predict points [1]. A Euclidean (in attribute space) k-NN algorithm was run using the musical features as axes. For each track the nearest k neighbours were found, the geodesic mean of their locations was taken, and the result compared to the true origin. To adopt this method to predict geographical location we used spherical geometry and took the average positions on an idealised sphere of Earth radius, using standard geodesic distance calculations. The results can be measured in terms of its great-circle distance from the true location (capital city) of the piece under consideration.

Finding the geodesic midpoint: with λ as latitude and ϕ as longitude (both in radians), convert to cartesian coordinates on a unit sphere:

$$x = cos(\lambda)cos(\phi) \tag{1}$$

$$y = cos(\lambda)sin(\phi) \tag{2}$$

$$z = sin(\lambda) \tag{3}$$

Take means per dimension $\bar{x}$,$\bar{y}$,$\bar{z}$. Find the longitude $\bar{\phi}$ of the midpoint:

$$\bar{\phi} = arctan(\bar{y}, \bar{x}) \tag{4}$$

Find the latitude $\bar{\lambda}$ of the midpoint

$$\bar{\lambda} = arctan(\bar{z}, \sqrt{\bar{x}^2 + \bar{y}^2}) \tag{5}$$

2.5 Utilising *a priori* Background Knowledge

We investigated the utility of using *a priori* knowledge to improve the predictions.

Land and Sea

The first piece of knowledge used is that music is produced on land. To utilise this we applied the NASA LandMask projected onto the idealised sphere of the Earth. This gives the terrain type for each square *degree* latitude by longitude. This varies in true size from about 110km wide to exactly 0 at each pole for longitude, whereas the latitude separation at 1 degree remains roughly 69km apart, excepting differences for the oblate spheroidal shape of the Earth. Because the landmask is given in latitude-longitude squares, a line-drawing algorithm must be employed to traverse a spherical distance across the earth, ensuring that the great-circle calculation is performed between each step, so that the correct next square is chosen. The total land contained in the path of error is weighted against the total water coverage. Different weightings were investigated but we settled with the simplest option: land 1:0 water such that only the land part counts. The reasoning behind this is that though human migration is slow across land, it is very fast across water, comparatively. For example, Brazilian music is close musically to Portuguese music because of migration patterns despite the size of the Atlantic Ocean. Different weightings will be tested in future.

Population Centroids and Population Density

The first change is to recenter each country to its population centroid, which were collated from the GPWv3 per administrative area data [20] scaled up to per country via the method detailed by Greg Hamerly (http://cs.ecs.baylor.edu/~hamerly/). The same dataset also provides population density grids at several resolutions. We chose to use the coarsest – per square degree – since it matches what we have for the land mask and is thus more easily comparable. Population density (as opposed to count) is chosen because it avoids the problem of varying size of square degree on the earth's surface owing to the diminishing distance between longitudinal degrees as either pole is approached, as explained above.

The algorithm for applying this mask is as follows:

1. The k nearest neighbours are determined based upon musical features.
2. For the group of nearest neighbours, we find the geodesic midpoint
3. The geodesic distance to the furthest neighbour from that midpoint gives us a radius of a geodesic circle which contains all the neighbours.

4. We calculate a new midpoint for this circle which is weighted by the population density in each of the points, at a resolution of one square degree.

The weighted midpoint is found as in equations 1-5 but each dimensional mean is calculated from weighted coordinates such that $\bar{x} = \sum \frac{x_i \rho}{N}$ where ρ is the population density for the square degree at the point (λ, ϕ).

3 Results

3.1 kNN performance

Using the default MARSYAS features the best predictive performance we achieved was a 2,827km median distance, and a 2,125km median land distance from the true position. This was achieved with k=10. This like all above results has a P-value < 0.001 since it was tested against a random distribution of 1,000 medians, and its value falls outside the entire distribution. We selected to use the median rather than the mean as this is more robust to outliers, but the means were also outside the relevant random distributions each time.

Table 1. Median distance from true location per k, featureset and algorithm

k	Feature	Mapping	Median
10	68	Spherical	**2827**
5	68	Spherical	2940
3	68	Spherical	2987
2	68	Spherical	3039
10	68	Land	2125
5	68	Land	2024
3	68	Land	1966
2	68	Land	**1850**
10	116	Spherical	**3013**
5	116	Spherical	3025
3	116	Spherical	3048
2	116	Spherical	3222
10	116	Land	**1506**
5	116	Land	1550
3	116	Land	2087
2	116	Land	1996

Table 1 shows the breakdown of results varying k (nearest neighbours), the number of features (116 includes chromatic features where 68 is the default set without them), mapping is whether the land measure is used to find the closest landmass (spherical means a pure distance measure using spherical geometry) and median is the median distance from the true answer for each combination.

Whilst demonstrably better than random this result could be improved upon. What we show here is that even with one of the simplest possible algorithms the information contained in the features is enough to indicate some geographic information. The land proportion is, of course always smaller than the total distance, but we argued that this is a fairer measure of error. This indicates that some measure of population distribution would be useful information to have *a priori*. It is interesting that for land measures the k=2 method performed best before addition of chromatic features, but the k=10 version performed best with the extra features. We hypothesise that for very similar music the chromatic features add no new information, yet when more dissimilar music is encountered, these provide a coarser measure of similarity with the music of other countries. Since some countries were underrepresented to the point of having fewer than k member tracks, this follows. However, the simple spherical method medians show no such inversion.

3.2 kNN with Population Distribution

Early indications are that this has not performed better than the above results. One reason for this may be that the larger problem is the feature sets rather than the prediction location, so taking population into consideration – which matters more for larger countries – would only help if those larger countries were being predicted well. The trend seems to be for smaller isolated countries such as Taiwan to be better predicted.

3.3 Statistical Significance

The determination of whether a music geographical prediction method is performing better than random is difficult using traditional statistical methods, because of the complicated, non-uniform geographical distribution of the music. Therefore we initially used a computationally intensive statistical method based on resampling. We programmed 1,000 random trials for each k and method combination, and measure our distance means against the distribution of random means to ensure the statistical significance of our results.

3.4 Performance by Country

The algorithm performed much better on some countries than others – even with the same number of tracks available, suggesting that some countries are more musically diverse than others. An additional problem is the relative size of countries – the level of precision required for e.g. Russian music is much less strict than that for e.g. Croatian music. Figure 2 (courtesy of Google Maps) shows half of the estimates (for clarity and to avoid overloading of positions) for Greek music, taken evenly from the distribution for Greece. From this it is clear that the prediction range for a country can be quite tightly distributed – the furthest estimate is 3,652km but there is a close cluster central to the image that reflects the more general skew in the distribution of estimates for all countries.

Fig. 2. Greek music: predictions of location

4 Discussion And Future Work

The work here is preliminary and there is much scope for further research and improvement in prediction performance.

More data is required: With a larger corpus with both more tracks from each country, and more countries represented, the prediction results will inevitably improve. If one is only interested in predicting location (as opposed to understanding the historical/pre-historical reasons for musical distribution) the problem is in many ways similar to statistical analysis of text, where organisations such as Google have now indexed so many pieces of text that they can solve many problems that were once thought to require solving deep problems in computational linguistics.

More geographical information could be utilised. It would be better to have access to the exact location of the origin of the music, rather than just the capital or population centroid, as most countries have strong regional variations in style. Some cultures change drastically over small areas, some are unchanged over large expanses, and this needs to be learnt by the prediction method.

Better representations. The music could be better represented for computational analysis. It is a truism within machine learning that the hard part is getting the features correct, and with the correct features almost any learning algorithm will work. For example, extra features such as the fine chromatic feature used by Gomez *et al.* could be applied. It would also be useful to explore the possibility of filtering and pre-selection of descriptors.

Many other forms of machine learning could be applied: neural-networks, support vector machines, decision trees, etc. It is also generally possible to improve the performance of individual methods by combining them together to form consensus predictions [1].

It is difficult to know how good our prediction results are as there are no previously published related comparisons. It would therefore be very interesting to compare the results of the machine learning programs with that of human performance in predicting musical origin.

The motivation for this work is to better understand the diversity of world music. To do this we have to go beyond just the prediction of location, but to analyse what features of the music are responsible for these predictions. This is now the main focus of our research.

References

1. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis, Wiley, New York, 1973.
2. Grachten, M., Schedl, M., Pohle, T., Widmer, G.: The ISMIR Cloud: a Decade of ISMIR Conferences at Your Fingertips. In: International Symposium of Music Information Retrieval, pp. 63–68, 2009.
3. Laurier, C., Grivolla, J., Herrera, P.: Multimodal Music Mood Classification Using Audio and Lyrics. In: Seventh International Conference on Machine Learning and Applications, pp. 688–693, 2008.

4. Widmer, G., Dixon, S., Knees, S., Pampalk E.,Pohle, T.: From sound to sense via feature extraction and machine learning: Deriving high-level descriptors for characterising music In: Sound to Sense: Sense to Sound: A State-of-the-Art. S2S2 Consortium, Florence, 2005.
5. Turnbull, D.R., Barrington, L., Lanckriet, G., Yazdani, M.: Combining audio content and social context for semantic music discovery.In: SIGIR 09, ACM, New York, NY, USA, pp. 387–394, 2009.
6. Knees, P., Pohle, T., Schedl, M., Widmer, G.: A music search engine built upon audio-based and web-based similarity measures. In Proceedings of SIGIR 07, ACM, New York, NY, USA, pp. 447–454, 2007.
7. Mckay, C., Fujinaga, I.: Automatic genre classification using large high-level musical feature sets.In Int. Conf. on Music Information Retrieval, pp. 525–530, 2004.
8. Mckinney, M., Breebaart, J.: Features for Audio and Music Classification, In: International Symposium on Music Information Retrieval, pp. 151–158, 2003.
9. 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, pp. 57–60, 2009.
10. Gomez, E., Herrera, P.: Comparative Analysis of Music Recordings from Western and Non-Western traditions by Automatic Tonal Feature Extraction. In: Empirical Musicology Review Vol. 3, No. 3, pp. 140–156, 2008.
11. Gomez, E., Haro, M., Herrera, P.: Music and Geography: Content Description of Musical Audio from Different Parts of the World. In: International Symposium on Music Information Retrieval, pp. 753–758, 2009.
12. Tzanetakis, G., Kapur, A., Schloss, A., Wright, M.: Computational Ethnomusicology. In: Journal of Interdisciplinary Music Studies, Vol. 1, No. 2, pp 1–24, 2007.
13. Lichte, W.H.: Attributes of Complex Tones. In: Journal of Experimental Psychology, Vol. 28, pp. 455–480, 1941.
14. Grey, J.M.: Multidimensional perceptual scaling of musical timbres. In: Journal of the Acoustic Society of America, Vol. 61, No. 5, pp. 1270–1277, 1977.
15. Ripley, B.D.: Spatial Statistics. Wiley, New York, 2004.
16. Wu, Y., Takatsuka, M.: The Geodesic Self-Organizing Map and its Error Analysis. In: Twenty-eighth Australasian Conference on Computer Science, Vol. 38, pp. 343–351, 2005
17. Govaerts, S., Duval, E.: A Web-based Approach to Determine the Origin of an Artist.In: International Symposium on Music Information Retrieval, pp. 261–266, 2009.
18. Tzanetakis, G., Cook, P.: MARSYAS: a framework for audio analysis. In: Organised Sound, Vol. 4, No. 3, pp 169–175, 2000.
19. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. In: IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 5, pp 293–302, 2002.
20. CIESIN, Columbia University; and CIAT 2005. Gridded Population of the World, Version 3 (GPWv3). Palisades, NY: SEDAC, Columbia University. http://sedac.ciesin.columbia.edu/gpw