UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Measuring the Eurovision Song Contest: A Living Dataset for Real-World MIR

Burgoyne, J.A.; Spijkervet, J.; Baker, D.J.

DOI
10.5281/zenodo.10265415

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 24th International Society for Music Information Retrieval Conference

License
CC BY

Link to publication

Citation for published version (APA):
Burgoyne, J. A., Spijkervet, J., & Baker, D. J. (2023). Measuring the Eurovision Song Contest: A Living Dataset for Real-World MIR. In A. Sarti, F. Antonacci, M. Sandler, P. Bestagini, S. Dixon, B. Liang, G. Richard, & J. Pauwels (Eds.), *Proceedings of the 24th International Society for Music Information Retrieval Conference: Milan, Italy, November 5-9, 2023* (pp. 817-823). ISMIR. https://doi.org/10.5281/zenodo.10265415

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
Download date:29 May 2024

MEASURING THE EUROVISION SONG CONTEST: A LIVING DATASET FOR REAL-WORLD MIR

John Ashley Burgoyne
University of Amsterdam
j.a.burgoyne@uva.nl

Janne Spijkervet
ByteDance
janne.spijkervet@gmail.com

David John Baker
University of Amsterdam
d.j.baker@uva.nl

ABSTRACT

Every year, several dozen, primarily European, countries, send performers to compete on live television at the Eurovision Song Contest, with the goal of entertaining an international audience of more than 150 million viewers. Each participating country is able to evaluate every other country's performance via a combination of rankings from professional jurors and telephone votes from viewers. Between fan sites and the official Song Contest organisation, a complete historical record of musical performances and country-to-country contest scores is available, back to the very first edition in 1956, and for the most recent contests, there is also information about each individual juror's rankings. In this paper, we introduce MIRoVision, a set of scripts which collates the data from these sources into a single, easy-to-use dataset, and a discrete-choice model to convert the raw contest scores into a stable, interval-scale measure of the competitiveness of Eurovision Song Contest entries across the years. We use this model to simulate contest outcomes from previous editions and compare the results to the implied win probabilities from bookmakers at various online betting markets. We also assess how successful content-based MIR could be at predicting Eurovision outcomes, using state-of-the-art music foundation models. Given its annual recurrence, emphasis on new music and lesser-known artists, and sophisticated voting structure, the Eurovision Song Contest is an outstanding testing ground for MIR algorithms, and we hope that this paper will inspire the community to use the contest as a regular assessment of the strength of modern MIR.

1. INTRODUCTION

The Eurovision Song Contest (ESC) is an annual event wherein several, primarily European, countries compete against one another by performing original, live songs during an internationally televised event. The contest began in 1956 and is typically held in the country of the previous year's winner in the spring.

© J. A. Burgoyne, J. Spijkervet, and D. J. Baker. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** J. A. Burgoyne, J. Spijkervet, and D. J. Baker, "Measuring the Eurovision Song Contest: A Living Dataset for Real-World MIR", in *Proc. of the 24th Int. Society for Music Information Retrieval Conf.,* Milan, Italy, 2023.

The content of the musical acts performed during the Eurovision Song Contest is always novel and notably diverse. Contestants are allowed to sing in whichever language they choose, often electing to sing in English to communicate the meaning of their song to a larger base, but some countries (notably France) have historically preferred to sing in their national language. According to the official Eurovision rules, all musical acts must perform an original song that is no more than three minutes in length, with the lead vocals performed live, and acts are limited to only six performers being on stage at any given moment during the performance [1].

Within these constraints, the musical acts of Eurovision are known for their ostentatious performances and camp aesthetics, which are often accompanied with visual spectacles from lightening to elaborate dance. As the contest is an international stage, the musical acts have also been a means in which countries are able to provide meta-political commentary on either national or global events [2, 3]. The contest has been noted as serving as an important platform for global LGBTQ+ visibility, which featured openly gay and transgender performers as early as the 1990s [4].

The winner of the contest is determined as a combination of both expert and panel voting, with no set criteria stated as to what should constitute a winning performance. A combination of the song's content, the visual performance, and the performer's ability to relate to the *zeitgeist* are all presumed to play an important role in determining the winner. Indeed, the Eurovision Song Contest can be and has been analysed from a variety of dimensions, summarised by Wolther as the media, the musical, the musical-economical, the political, the national-cultural, the national-economic, and the competitive [5].

We next detail the rules of the contest before introducing the MIRoVision data set, which contains a multi-faceted collection of historical data that could be used to predict the contest's winner and enable researchers to make deeper inquiries into the history and music of the contest.

1.1 Rules of Eurovision

In order to participate in the Eurovision Song Contest, participating countries work in coordination with the European Broadcasting Union. While each participating country – or more specifically the country's partnered national broadcaster – is allowed to decide for themselves which act to send to participate, the results of the Eurovision Song Contest are determined by voting over three events. These

three events referred to as the First Semi-Final, the Second Semi-Final, and the Grand Final. It is the Grand Final that typically receives the vast majority of the attention and viewership.

As described on the official Eurovision website [1] , all participating countries qualify for two semi-final shows in the week leading up to the Grand Final, which only a subset of the total countries will perform. France, Germany, Italy, Spain and the United Kingdom are automatically included in the Grand Final and are referred to as the 'Big Five'.

After a country has performed, each other country gives two sets of votes for the performance. The first set of votes comes from an expert panel of music industry professionals from within that country. Starting in 2016, the official Eurovision website has published the individual data of each juror from each participating country. The second set of votes comes from viewers from the of the performing country. The votes represent points that are added together and each country can use their set of points, $\{12, 10, 8, 7, 6, 5, 4, 3, 2, 1\}$ for one and only one country. No juror or television vote can be cast for one's own country. In the semi-finals, voting is limited to only countries participating in their respective show, whereas in the Grand Final, any country is allowed to vote. The Grand Final television show is also characterised by great fanfare surrounding each national jury's announcement of which country they chose to award '*douze points*'.

No explicit criteria are given as how any vote should be decided. Said another way, it should not be assumed that all participants attempt to vote for a measure of musical quality. Many factors have been discussed in academic literature on the topic, that suggest there are both geographic and political factors that can play into how countries decide to cast their votes [6–10].

2. MIROVISION DATASET

Data that comprises the MIRoVision dataset originates from three primary sources. The first is the official Eurovison website (https://eurovision.tv/), the second is the Eurovision World fan website (https://eurovisionworld.com), the third are audio features taken directly from the YouTube videos linked in the contestant metadata. The dataset contains five primary types of data: (1) contest meta-data; (2) contest results; (3) voting data; (4) audio features extracted from recorded performances of the musical acts and (5) betting office data. All data for each Eurovision Song Contest is available each year since the year 1956 until present day with the exception of 2020 when the contest was cancelled due to the global COVID-19 pandemic. As of 2016, the official Eurovision website has published data detailing how each of the five jurors from the expert panel have voted on all three nights of the contest. The current release of the data set contains the contestant metadata, contest ranking and voting data of 1719 entries. The dataset is hosted on a GitHub repository. [2]

[1] https://eurovision.tv/about/how-it-works

[2] https://github.com/Spijkervet/eurovision-dataset

In total, 56 countries are represented in the dataset, which includes countries that have been dissolved, renamed, or merged since the inception of the contest in 1956. Voting data for the contest is stored in three tables: (1) votes; (2) contestants; and (3) jurors.

The *votes* table contains data from the contest's beginning in 1956 and indicates how each country's aggregated jury and televoting points were distributed to each other participating country.

The *contestants* table contains all metadata regarding each song entry, such as the artist's name and song title, lyrics, composers and lyricists, the running order and the total points awarded by the jury and televoters in the Semi-Final and Final Rounds respectively. This table also includes links to YouTube videos of live performances from the televised Finals or Semi-Finals, as maintained by the Eurovision World team.

The *jurors* table contains data beginning from the year 2016 and indicates how the five anonymous jurors (designated with letter names A through E) voted for each other country and in which night of the contest. As noted above, countries are unable to vote for themselves, are only able to vote within the Semi-Final they are participating in, whereas all countries are able to vote in the Grand Final.

In addition to the voting tables, the *betting-offices* table provide tables of historical bookmakers' odds for the contest winners, as collected by Eurovision World. The Eurovision Song Contest is a popular target for online betting. Day-of-contest odds are available for 2016 and 2017, and daily odds up to six months prior to the contest are available from 2018 onward, for 10 to 20 betting offices.

3. A PREFERENCE MODEL FOR EUROVISION

The Eurovision Song Contest voting system is iconic, but because the number of contestants varies, it is not possible to use contest scores to make comparisons across years. Moreover, the contest scores do not operate on an interval level of measurement: even within a particular year, a difference of five or ten points may mean something quite different at the top end of the score range than it does at the bottom. With the rich data in the MIRoVision set, however, it is possible to fit statistical models with parameters that correspond monotonically to actual contest results but that *do* behave on an interval scale. Such an interval scale is not only interesting musicologically and sociologically, but also for machine-learning applications, as most common loss functions for training implicitly assume interval-scale outcomes. In short, we are looking for a true *measure* of competitiveness in the Eurovision Song Contest, and one that applies stably across years.

In order to achieve these desiderata, the contest results must be sufficient statistics for the model parameters of interest. If we make the stronger assumption that there be only a finite number of sufficient statistics beyond these, then by the Pitman–Koopman–Darmois theorem [11], the model must be a member of the exponential family. That leaves a surprisingly small class of plausible models.

The simplest model requires no sufficient statistics other

than the scores themselves. Under such a model, the probability of the set of scores from any particular country's jury or televoters

$$\Pr[\text{ranking}] \propto \exp(s_1\beta_1 + s_2\beta_2 + \cdots + s_N\beta_N)\,, \quad (1)$$

where the coefficients $s_n \in \{12, 10, 8, 7, 6, 5, 4, 3, 2, 1\}$ are the scores awarded from that jury or televoter group to contestant n and the β_n are the model's competitiveness parameters for contestant n. The normaliser $Z_0(\boldsymbol{\beta})$ for this distribution is the sum of these terms for any valid assignment of scores under the Eurovision system. After M juries and televote groups combine their scores independently to determine a winner, the combined probability

$$\Pr[\text{contest}] = \frac{\exp(s_1\beta_1 + s_2\beta_2 + \cdots + s_N\beta_N)}{Z_0(\boldsymbol{\beta})^M}\,, \quad (2)$$

where $s_1, s_2, \ldots, s_N$ now represent the *total* scores awarded to each contestant. The trouble with this model is that for a typical Eurovision show of 26 contestants, the normaliser contains ${}_{26}P_{10} \approx 19$ trillion terms. The model is thus infeasible in practice, despite its theoretical simplicity.

Most alternatives to this model lose their exponential-family properties. There is, however, an interesting alternative if we are willing to consider Eurovision contest scores from juries and televoters to be *ratings* instead of rankings. Specifically, assume that for each song, juries must award a scores in the set $\{12, 10, 8, 7, 6, 5, 4, 3, 2, 1, 0\}$, but that there is no restriction on how many times they can use each score. While the numerator of such a model remains the same as (1) and (2), its normaliser

$$Z(\boldsymbol{\beta}) = \prod_{n=1}^{N} \sum_{k \in \{12,10,8,7,6,5,4,3,2,1,0\}} \exp(k\beta_n)\,, \quad (3)$$

which can be computed easily. Although there are fundamental conceptual and mathematical differences between rankings and ratings [12], if we restrict the outcome space of rating model (3) to allow only outcomes that would also be valid in the ranking model (2), the models are equivalent [13]. Moreover, we can add an extra set of score-level parameters ξ_k to allow (3) to better approximate (2) without sacrificing equivalency on the restricted outcome space:

$$\Pr[\text{ratings}] = \frac{\exp\left(\sum_n s_n\beta_n\right) \cdot \exp\left(\sum_k \xi_k\right)}{\prod_n \sum_k \exp(k\beta_i + \xi_k)}\,, \quad (4)$$

where s_n are again the scores from a particular jury or televoter group and $k \in \{12, 10, 8, 7, 6, 5, 4, 3, 2, 1, 0\}$. This model is known in the psychometric literature as the *partial-credit model* [14] and is one of the standard mathematical tools used for assessing the reliability of rubrics, Likert scales, and educational test items with partial credit.

4. FITTING THE PREFERENCE MODEL

We fit the partial-credit model (4) to the MIRoVision data for all Song Contests since 1975, the year that the {12, 10, 8, 7, 6, 5, 4, 3, 2, 1, 0} scoring system was instituted. We considered every vote available as an individual observation: every country's jury, every country's televotes in years that those votes were counted separately from juries, and all votes from semi-final rounds when they occurred. We made the important but unavoidable assumption that the average competitiveness of a Eurovision entry has remained constant over time, as there are no cross-year comparisons that would make it possible to estimate the model otherwise.

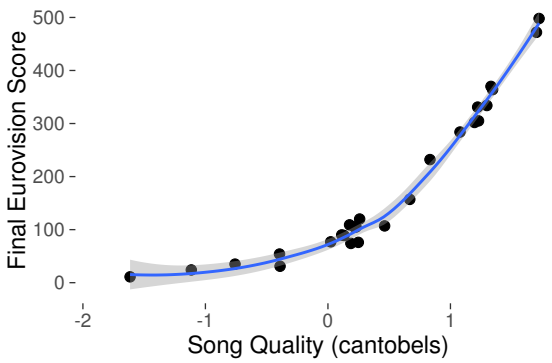

Figure 1. Correspondence between song competitiveness (in cantobels) and final Eurovision Song Contest scores in 2019. The pattern in this year is typical of all other years, with a relatively slow increase in points as competitiveness improves up to about 0.5 cantobels, followed by a rapid increase. Because of the semi-final rounds, the relationship between competitiveness and final score is not a strictly monotonic as in years without semi-finals, but it is still nearly monotonic.

We fit the joint probability model using the Bayesian probabilistic programming language Stan, with normal priors on average country competitiveness and song competitiveness and a multivariate normal prior on $\boldsymbol{\xi}$ for each contest. The complete model code is available in the supplemental material. For interpretive purposes, we fixed the mean of the song competitiveness parameters $\boldsymbol{\beta}$ to 0 and report them on a $10\log_{10}$ scale, analogous to the decibel. In honour of the singing at the contest, we deem this unit the *cantobel*. An increase of one cantobel in song competitiveness means that a song improves its chances of receiving one extra point from any given jury by $10^{\frac{1}{10}} \approx 1.26$. Like the decibel scale, an increase of 3 cantobels means that a song approximately doubles its chances of receiving one extra point.

Figure 1 illustrates the typical correspondence between competitiveness in cantobels and actual song contest results. After a slow increase, the slope rapidly increases for highly competitive entries. The Eurovision Song Contest scoring system compresses differences between relatively uncompetitive entries and dramatically exaggerates small differences at the top. While this surely contributes to the exciting television, cantobels are a better scale to use for scientific purposes.

Figures 2 and 3 reveal the heart of the model. The first shows the average song quality, as perceived by the Eurovision Song Contest juries and televoters, over the period from 1975 to 2022. Ukraine, Russia, Italy, and Sweden stand

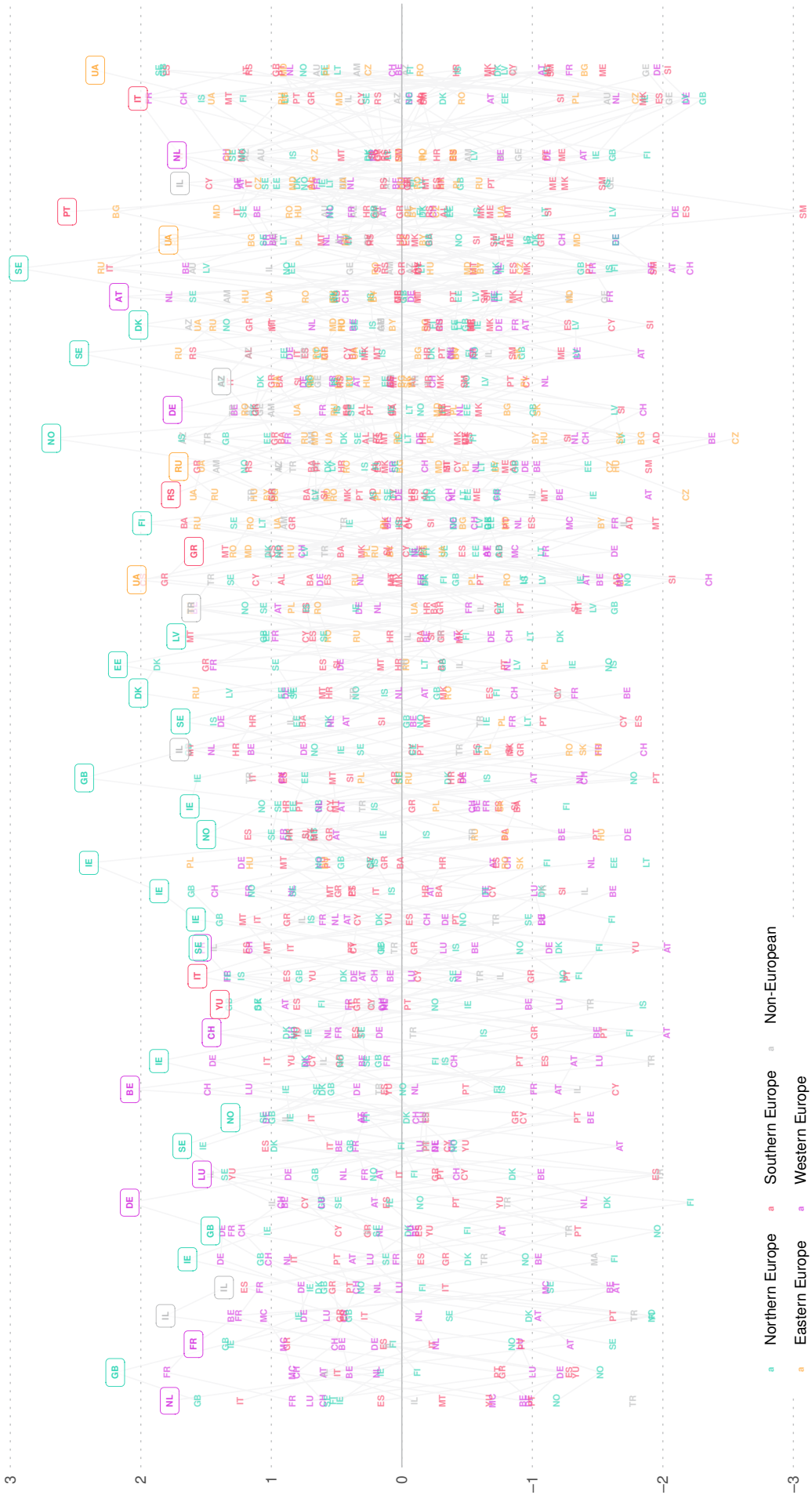

Figure 2. Historical competitiveness of Eurovision Song Contest entries (in cantobels). Countries are coloured by their geographic region as defined in the United Nations M49 standard. Winners are boxed. The standard error of estimates is roughly 0.5 cantobel in early years and roughly 0.3 after the institution of semi-final rounds in 2004; as such, difference of approximately 1.0 cantobels are likely statistically significant. After a period when Northern and Western Europe exchanged victories, there was a period of Northern European dominance; recent years have been characterised by a good geographic diversity of winners.

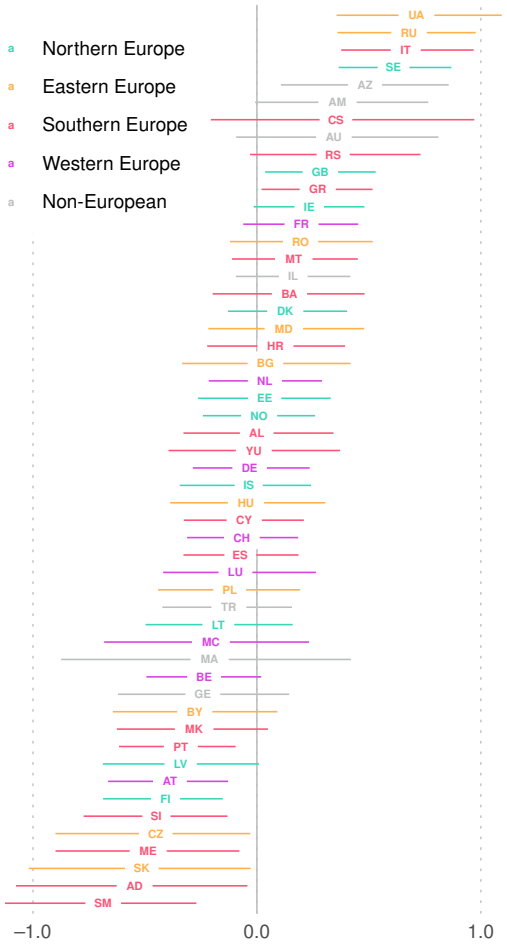

Figure 3. Median competitiveness of countries' Eurovision Song Contest entries, 1975–2022, in cantobels with 90% credible intervals. Countries are coloured by their geographic region as defined in the United Nations M49 standard. Ukraine, Russia, Italy, and Sweden stand out as having sent contestants of exceptional competitiveness, although Azerbaijan, the United Kingdom, and Greece's credible intervals are also strictly greater than zero.

out as having been particularly successful, even though they have suffered almost-wins instead of victories in many years. On average, songs from these countries have been a half cantobel above the average. But the first figure shows that there are dramatic swings from year to year underneath these averages. Even one the most convincing victories from one of the historically strongest countries – Måns Zelmerlöw's 'Heroes', Sweden's 2015 entry – was preceded and succeeded by much less appreciated acts.

4.1 Jury Model

Jury scores at the Eurovision Song Contest are determined by combining rankings from five independent jurors from each country, each of whom must make a complete ranking of contestants at a show, from best to worst. After averaging these ranks, they are converted to the better-known $\{12, 10, 8, 7, 6, 5, 4, 3, 2, 1, 0\}$ system that is reported on television. Since 2016, the European Broadcasting Union has made not only the final scores but also these individual rankings public. They have also publicised that they continue to experiment with the proper way to average the ranks across jurors, currently using exponential decay. [3]

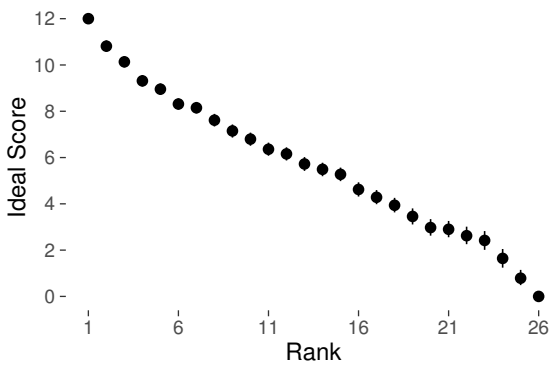

Figure 4. Ideal scores for averaging ranks within juries, according to a generalised partial-credit model, with 90% credible intervals. In recent years, the Eurovision Song Contest has used an exponential weighting scheme, but these results suggest that a linear scheme with a small bonus for the top-ranked entry would be sufficient.

The theory of partial-credit models offers an alternative, more empirical solution. Rather than taking the scoring rule in (4) as fixed, the *generalised partial-credit model* considers an optimal scoring rule that would lead the model to make the best predictions. Concretely, that would mean considering alternatives to the $\{12, 10, 8, 7, 6, 5, 4, 3, 2, 1, 0\}$ rule for the main contest, and by extension, to the simpler $\{1, 2, \ldots, N\}$ rule for jury members making a full ranking.

The MIRoVision dataset includes these jury scores, and we fit a generalised partial-credit model to them analogous to the model we fit for the contest overall. The code is available in the supplemental material. Figure 4 shows the results. Like the European Broadcasting Union's current rule, we arbitrarily fix the maximum score to 12. It seems that rather than replacing the former linear scheme with the current exponential one may be a more effective simply to give a small fixed bonus to each juror's top-ranked entry. Such a solution would also solve the core issue motivating the exponential weighting, namely that it it undesirable for one juror to have unilateral power to spoil the chances of some other juror's favourite.

5. PREDICTING WINNERS

The Eurovision Song Contest is also notorious for attracting online and offline bets on the outcome. Since 2015, the EurovisionWorld web site has been collecting the odds posted at a large number of online betting offices, for each day leading up to the contest. These odds can be converted into implicit probabilities of winning, and there is often much discussion in the weeks leading up to the contest about which acts the bookmakers are favouring.

[3] https://eurovision.tv/story/subtle-significant-ebu-changes-weight-individual-jury-rankings

Year	Country	Actual	Bookmakers
2018	Israel	.87	.24
2018	Cyprus	.12	.37
2018	Germany	.01	.09
2019	Netherlands	.53	.51
2019	Italy	.45	.09
2019	Switzerland	.01	.09
2019	Russia	.01	.02
2021	Italy	.63	.26
2021	France	.35	.22
2021	Switzerland	.02	.05
2022	Ukraine	.98	.62
2022	Sweden	.01	.14
2022	United Kingdom	.01	.06
2022	Spain	.01	.06

Table 1. Probability of winning the Eurovision Song Contest, 2018–2022, given the partial-credit model and perfect information about jurors' and televoters' preferences, compared to bookmakers' implied win probabilities immediately prior to the contest final.

We can use our model fits to compare the bookmakers' predictions to the actual probabilities countries had to win given jurors' and televoters' preferences and the assumptions of the partial-credit model. To compute these probabilities, we reshuffled the draws from our Bayesian samples independently for each country and tallied how often these would have been the highest, taking advantage of the fact that competitiveness in cantobels is a sufficient statistics for actual contest outcomes. Table 1 presents the results. Both 2019 and 2021 were rather close contests, whereas 2018 and 2022 had clearer frontrunners. The bookmakers markedly mis-called 2018, but have been more accurate since. If one had been able to place stakes at the online betting offices with perfect knowledge of the jurors' and televoters' preferences, one would have quadrupled one's stake on average (before paying out the bookmakers' sometimes shockingly high margins on Eurovision odds).

6. CONTENT-BASED CONTEST PREDICTIONS

Perfect information is of course never available, but perhaps deep learning and content-based MIR offer something? Self-supervised music representation learning has advanced considerably in recent years. It has successfully been applied to many downstream tasks, including music tagging [16], genre classification, key detection and emotion recognition [17, 18]. These foundation models are generally pre-trained in an unsupervised, end-to-end fashion on raw audio samples. By defining an auxiliary loss objective on large quantities of music and using data perturbations, models are able to learn effective and robust representations.

To evaluate whether a pre-trained foundation model is able to predict preferences, we extracted embeddings on all song entries using the TUNe+ [19] and MERT [18] models. On every window of 2 seconds, an embedding vector of 512 feature dimensions is computed for the TUNe+ model. The MERT model returns 25 representation layers, and 1024 feature dimensions on 5-second windows. For every song entry between 1975 and 2022, a single embedding vector is calculated by taking the arithmetic mean along the time dimension for TUNe+ and along the representation layers for MERT respectively. This results in 1 261 embeddings in total. For every song entry, we took 4 000 draws from the fitted model for song competitiveness (in cantobels) and treated these as our targets Y; using 4 000 draws instead of a single point estimate more accurately averages over our uncertainty about song competitiveness, given the inherently limited number of rankings available for any single edition of the contest. We freeze the pre-trained TUNe+ and MERT models and perform a linear probe using the mean-squared error between $(\hat{y}, y)$. We use 5-fold cross-validation and sample all song entries from two years within each decade between 1975 and 2023 as our validation set.

Model	L1	L2
TUNe+ [19]	0.828 (0.039)	1.063 (0.052)
MERT [18]	0.820 (0.019)	1.025 (0.027)

Table 2. L1 (MAE) and L2 (RMSE) losses and their standard deviations after training two state-of-the-art audio embeddings to predict the competitiveness of Eurovision Song Contest entries from 1975–2022, in cantobels.

Our results in Table 2 show that we can achieve RMSE of 1.025 cantobels by way of training a linear layer on embeddings extracted from a pre-trained foundation model. These models are not specifically trained or designed for our downstream task of preference prediction, e.g., features extracted by the different layers in MERT vary in their downstream task performance, and we leave further improvements to future work. But to contextualise the result, the overall standard deviation of our Eurovision competitiveness ratings is 1.064 cantobels, which means that state-of-the-art MIR audio embeddings are able to predict 7.2% of the variance in Eurovision Song Contest competitiveness.

7. CONCLUSION

We present MIRoVision, a collection of data and tools for studying the Eurovision Song Contest and applying music information retrieval to several types of data generated from the contest. One of our key results is a model for converting the highly non-linear contest scores into a well-behaved interval-scale measurement we dub the *cantobel*. Cantobels facilitate understanding of fluctuations in the contest over time and more accurately represent both the competitiveness and the uncertainty surrounding the competitiveness of Eurovision Song Contest entries. They also behave better with the standard loss functions used in machine learning systems, and allow us to predict a small but meaningful portion of variance in contest outcomes. We hope this result is sufficiently tantalising to encourage the community to try their own models – the Eurovision Song Contest offers a fresh set of contestants every year – and to find their own creative uses for this rich musicological data source.

8. REFERENCES

[1] "How the Eurovision Song Contest works," Jul 2022. [Online]. Available: https://eurovision.tv/about/how-it-works

[2] C. Baker, "Wild dances and dying wolves: Simulation, essentialization, and national identity at the Eurovision Song Contest," *Popular Communication*, vol. 6, no. 3, pp. 173–189, 2008.

[3] J. K. O'Connor, *The Eurovision Song Contest: The Official History*. Carlton, 2010.

[4] C. Baker, "The gay olympics? the Eurovision Song Contest and the politics of LGBT/European belonging," *European Journal of International Relations*, vol. 23, no. 1, pp. 97–121, 2017.

[5] I. Wolther, "More than just music: The seven dimensions of the Eurovision Song Contest," *Popular Music*, vol. 31, no. 1, pp. 165–171, 2012.

[6] G. Yair, "'Unite Unite Europe': The political and cultural structures of europe as reflected in the Eurovision Song Contest," *Social Networks*, vol. 17, no. 2, pp. 147–161, 1995.

[7] G. Yair and D. Maman, "The persistent structure of hegemony in the Eurovision Song Contest," *Acta Sociologica*, vol. 39, no. 3, pp. 309–325, 1996.

[8] D. Fenn, O. Suleman, J. Efstathiou, and N. F. Johnson, "How does Europe make its mind up? Connections, cliques, and compatibility between countries in the Eurovision Song Contest," *Physica A*, vol. 360, pp. 576–598, 2006.

[9] V. Ginsburgh and A. G. Noury, "The Eurovision Song Contest: Is voting political or cultural?" *European Journal of Political Economy*, vol. 24, no. 1, pp. 41–52, 2008.

[10] M. Blangiardo and G. Baio, "Evidence of bias in the Eurovision Song Contest: Modelling the votes using Bayesian hierarchical models," *Journal of Applied Statistics*, vol. 41, no. 10, pp. 2312–2322, 2014.

[11] B. O. Koopman, "On distributions admitting a sufficient statistic," *Transactions of the American Mathematical Society*, vol. 19, pp. 399–409, 1936.

[12] S. J. Brams and P. C. Fishburn, "Voting procedures," in *Handbook of Social Choice and Welfare*, K. J. Arrow, A. Sen, and K. Suzumura, Eds. Elsevier, 2002, vol. 1, pp. 173–236.

[13] D. Andrich, "Understanding the response structure and process in the polytomous Rasch model," in *Handbook of Polytomous Item Response Theory Models*, M. L. Nering and R. Ostini, Eds. New York: Routledge, 2010, pp. 123–152.

[14] G. N. Masters, "A Rasch model for partial credit scoring," *Psychometrika*, vol. 47, no. 2, pp. 149–174, 1982.

[15] United Nations Statistics Division, "Standard country area codes for statistical use (M49)," 2021. [Online]. Available: https://unstats.un.org/unsd/methodology/m49/

[16] J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," in *Proceedings of the 22nd Society for Music Information Retrieval Conference*, 2021.

[17] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," *Proceedings of the 22nd Society for Music Information Retrieval Conference*, 2021.

[18] Y. Li, R. Yuan, G. Zhang, Y. Ma, C. Lin, X. Chen, A. Ragni, H. Yin, Z. Hu, H. He *et al.*, "Large-scale pretrained model for self-supervised music audio representation learning," Presentation at the Digital Music Research Network, 2022.

[19] M. A. Vélez Vásquez and J. A. Burgoyne, "Tailed U-net: Multi-scale music representation learning," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, 2022.