Ukrainian Catholic University

Faculty of Applied Sciences

IT & Business Analytics Programme

Research on Eurovision Song Contest voting results based on ethnicity groups in each country, and geopolitical and cultural factors

Econometrics Project

Authors:
Anastasiia Havryliv
Alina Voronina

01 June 2022

1   Introduction and problem setting

Voting systems are extremely sensitive to different factors, and thus are easy to manipulate, implicitly or explicitly. The Eurovision Song Contest (ESC) is an example of such voting ambiguousness. It is upon people to decide which artists will win, but the results of the contest do not depend only on their opinions. The locations of main diasporas, national minorities' distribution, and political presumptions are the main factors that might affect the authenticity of the results. For instance, being in solidarity with Ukraine, neighbouring countries may give it more points, and vice versa. Likewise, countries with large diasporas of ukrainians will more likely vote for their native country.

To reduce the effect of the demographic distribution and politics on the contest results, Eurovision authorities introduced different types of voting systems, them being jury voting, televoting, or both ones combined.

This project has a goal to analyse the results of Eurovision contests, and identify the trends they may have. Therefore, the audience of this project contains people connected to the area of international competitions, them being contest holders, jurys, and prognosticators.

The aim of this project is to identify whether peoples' votes are correlated with political and demographic situations in the countries, based on the previous Eurovision contest results, data about ethnicities' distribution over participant countries, and geopolitical and cultural factors.

2   Literature review

There exist many studies dedicated to different factors' influence on the outcomes of the competition. One of them, made by Ginsburgh and Noury [1], investigated the effect of cultural voting in the ESC. Their research was based on dataset starting from 1956, and was focused on determining the features that allow competitors to succeed. They used a multiple regression model to test the significance of the variables referring to linguistic and cultural factors [2]. Moreover, they suggested that migration patterns also determined the outcomes of the voting results: people from other nation, living in the participating country, were more likely to vote for their home country. Furthermore, migrants were more motivated to participate in voting than the country's nationals, as they could ameliorate the chances of their homeland victory. Therefore, the results of the competition were affected by the migration situation and ethnicity groups on the participant countries.

3   Data description and analysis

Although data for ECS is available from 1956, the year when the contest started, not all the years were included in the research. The manipulations were done on the data starting from 1990, with the aim to logically connect datasets on voting results and migration information, as the latter is available only from 1990.

A Main data: ECS Votes and Migration

Primary data used in the research comes from three datasets: *Eurovision data 1956-2020* [3], *Eurovision data 2021* [4], and *Population division dataset* [5]. All data sets were cleared and brought to a convenient view. Moreover, special functions to get information about the specific country were implemented. They allow to obtain a dictionary, where each key represents a voter, and each value represents number of points given from the voter to the particular country.

From the two *Eurovision datasets* the following dictionary is obtained:

$$countryA = \{countryB: pointsB, \ countryC: pointsC\},$$

where each country in the dictionary is the country giving points to countryA in the specified year.

Regarding *Population division dataset*, it has similar function, which allows to get information for each country of origin. The key represents a country where this nation migrates to, and the value depicts the number of the migrants in specific country, respectively.

$$countryA = \{countryB: migrantsB, \ countryC: migrantsC\},$$

where each country in the dictionary is the country to which people from countryA migrated, and the number of migrants is represented in the specified year.

The two dictionaries are the main data collections that the work is performed on.

B Auxiliary data: Borders and Language

The *borders.csv* dataset was made specially for the research. It is depicted as a matrix, where row and column names are country names, and their intersection includes a binary number: 0 — if countries do not share a physical border, 1 — if countries share a physical border. Moreover, this file contains information about countries' official languages (if there are few of them, the one that is used the most is stated in the column).

The *eurovision_winners.csv* dataset contains information on every winner of the ESC, and the language their song was performed.

C Dependent and independent variables

Taking the above data views into consideration, one can determine the dependent variable as a number of points to country A from country B in particular year. The independent variables are the metrics of the country B that have relation to the dependent one.

In the first model, they are a number of migrants in the specified year, the presence

of physical border between countries A and B, and the language of the song country A performs.

The overall regression formula looks as follows:

$$Y = \beta_0 + \sum_{i=1990}^{c}(\beta_i * \text{number of migrants from A to B in year i}) + \beta_1 * \text{border A with B} + \beta_2 * \text{language of the song A} \,,$$

where Y = points to A from B in year c; language = 1 if country A performs a song in a language, that is native to country B, 0 otherwise.

In the second model, the votes to country A from country B in the previous years instead of migrants data are considered as independent variables, which are included in the model.

The overall regression formula looks as follows:

$$Y = \beta_0 + \sum_{i=1990}^{c-1}(\beta_i * \text{previous i-th year points to A from B}) + \beta_1 * \text{border A with B} + \beta_2 * \text{language of the song A} \,,$$

where Y = points to A from B in year c; language = 1 if country A performs a song in a language, that is native to country B, 0 otherwise.

4 Methodology

The null hypothesis states that people vote for the neighbouring countries, or counties they have relations to. This covers the demographic notions of diasporas and migrant groups, as well as political aspects of nations' solidarity and/or historical relations.

To implement the model fitting, there were taken a few preparation steps:

- **Choosing the country for analysis**. The country can be chosen randomly, from the list of winners of the ESC in the time frame 1990-2021. However, in the examples of the OLS regression and data preparations specific country was chosen to make observations more clear and easier.

- **Taking the needed data and combining dataframes**. This step is needed in purpose to take only relevant observations and variables for the analysis and prepare the final dataframe for the OLS regression.

- **Defining dependent and independent variables for the model**. (More detailed explanation about the variables in the previous section)

- **Fitting the OLS regression model for collected and processed data**.

In this research the **Ordinary Least Squares regression** was used as the main method for the modeling. This technique allows to estimate the relationship between the dependent variable and one or more independent variables.

Intuitively speaking, OLS method aims to minimize the sum of squared errors, which is basically the prediction error between the real and the predicted values. And to do that, it provides a single regression equation to represent the process of modeling the prediction.

From the results of the two fitted models such a conclusion can be drawn: The null hypothesis should be rejected, because of the very small R-squared in the first model (the one using migrant data). From the R-squared score can be taken the information that the migrants have some influence on the number of votes, but it is note as big as one might think.
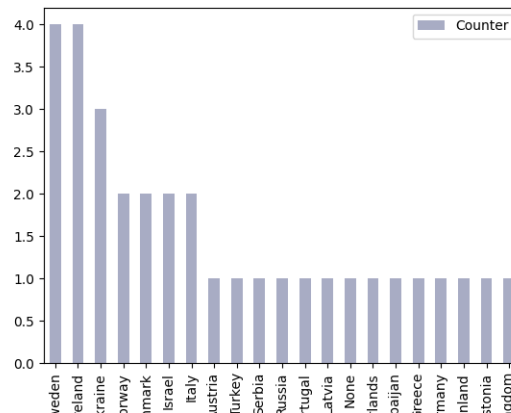
5   Results

The results of the project are important to identify the relevance of the contest voting systems, in particular the ESC one. The precise analysis also gave insights into the possibility of anticipating contest results long before the voting polls are even open.

As a result of the work, an analysis of exploring the ESC voting results was done. After analysing different geopolitical, ethnical, and cultural features, the authors were able to explain these factors' influence on the overall ESC voting results and competitors' victories throughout the years, starting from 1990.

It turned out that the model, based on the migration data, has a very low R-squared score of 13.7%, meaning it is not reliable in predicting the voting results. On the contrary, the model, which takes into consideration voting results from the previous years, is more credible, showing the R-squared score of 93%. The p-values to the variables in the migration-based model are not low, thus the null-hypothesis of those variables being significant is rejected. Regarding the borders and language variables, they tend to influence the overall voting results at some point.

The following plots demonstrate the winners of the ESC by the number of victories, and the languages that most winning songs were performed in.

Figure 1. *Countries' victories by number of total wins*
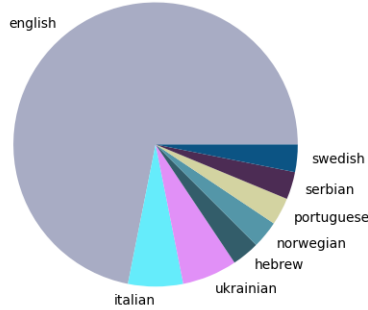
Figure 2. *Languages of songs by number of total wins*

As can be seen from above, the language of the song is not influencing the results as much, although it is still an important variable. Both artists performing in their native language and the ones performing in English have chances to win. The sample of the songs in English language that won the ESC is bigger, though.

6 Conclusions

The two models were obtained as a result of the project, with considering migration and previous voting results as main factors. The former's performance is quite low, meanwhile the latter shows much better results. The *borders* and *song language* variables are considered as significant, although they are not the main factors which influence the voting outcomes for this or that country.

Although the results of the newest 2022 ESC are available, they were not taken into account due to peculiarities with political situation. Because of russo-ukrainian war, a lot of ukrainians migrate abroad, thus causing major changes in the migrant data, compared to the previous year. Furthermore, there is no structured, or even unstructured but credible data on the Internet on the movement of migrants from Ukraine. Different sources suggest information on single countries, but the problem is they are all of different dates, which makes the research irrelevant. Moreover, the sources themselves are quite random too.

The results of both models can be ameliorated by including more significant features into the regression, and by exploring the correlations between each variable more presicely.

7 Data and code

All re-executable Python code, datasets, and visualizations can be found on the following GitHub repository:

https://github.com/linvieson/eurovision-voting

References

[1] Victor Ginsburgh, Abdul Noury, *Cultural Voting The Eurovision Song Contest*, November 2004

[2] Oisín Leonard, *An Investigation into Factors Which Explain the Scores Which Explain the Scores and Voting Patterns of the Eurovision Song Contest*, 25 January 2018

[3] www.kaggle.com/datasets/orianao/eurovision-finals-voting-results

[4] www.kaggle.com/datasets/prajittr/eurovision-2021-song-competition-voting-patterns

[5] www.google.com/url?q=https://www.un.org/en/development