

A Clustering-based Analysis of Geopolitical bias in Eurovision Voting

Keyu

Jan. 7, 2025

1 Introduction

This is just an informal draft, simply explaining the clustering method.

2 Method

Our motivation for clustering is to explore whether countries classified into the same category based on geographical location have similar voting performances. We designed two strategies. The first one clusters based on countries, and the second one clusters based on country-country pairs.

There are a total of n countries ($n=51$, currently only considering the countries that appear in `distances.csv`). Based on the distances between capitals and the number of votes for each other, two $n \times n$ matrices can be constructed, the distance matrix D and the voting matrix V . I have processed the order of rows and columns of D and V , and the i -th row (i -th column) of D and V corresponds to the i -th country in `country_info_aligned.csv`. Both D and V are normalized.

For the first strategy, we use k -means clustering on each row of D . Similar rows in D (i.e., countries with close/similar geographical locations) will be classified into the same category. We obtain the classification labels (country: label) and use these labels to visualize D and V (for convenience, both visualizations are reduced to 2 dimensions using PCA). If the points of the same category in V are scattered, it indicates that countries grouped together due to similar geographical locations have significant differences in their voting preferences. The silhouette score is calculated to measure the degree of dispersion of points of the same category in V (the closer to 1, the more aggregated the points of the same category are; the closer to -1, the more dispersed they are), and it is compared with random clustering.

The second strategy is similar in method, but differs in that it clusters the $n \times n$ elements within the matrix. For example, (Germany, Morocco) and (Germany, Iceland) would be clustered into one category (the distances from Berlin to Reykjavík and Rabat are 2382 and 2598, respectively), while (Germany, Lithuania) and (Germany, Latvia) would be clustered into another category

(the distances from Berlin to Vilnius and Riga are 822 and 844, respectively). After obtaining the clustering labels, we use a similar method to visualize D and V.

3 Results & Analysis