cont_cluster

2024-12-07

At first glance, there doesn’t seem to be any obvious grouping by geography. Group 1 seems like it is the teams that did better in the tournament. Let’s evaluate more carefully.

Let’s split the countries into 5 groups based on geography. Europe and Africa can be their own groups. Forming the other 3 groups is more interesting. Let’s try a group of countries from the Americas, not including Canada and the USA. Let’s make another group for Asian countries. The last group can be Canada, USA, and Australia. The reasoning here is that Canada, USA, and Australia seem more culturally similar to each other and Mexico, central America, and South America seem more culturally similar to each other. Objectively, this is at least true in terms of language.

```
library(mclust)
```

```
## Package ’mclust’ version 6.1.1
## Type ’citation("mclust")’ for citing this R package in publications.
```

```
coords_df <- read.csv("avg_team_data_kmeans.csv")
continent_g1 <- c("latam", "can_us_aus", "eur", "latam", "afr", "can_us_aus",
                  "latam", "eur", "eur", "latam", "eur", "eur", "eur", "afr",
                  "asia", "asia", "asia", "latam", "afr", "eur", "eur", "eur",
                  "asia", "asia", "afr", "eur", "eur", "eur", "afr", "can_us_aus",
                  "latam", "eur")

coords_df$cont_g1 <- continent_g1

#Confusion matrix
cont_conf <- table(coords_df$cluster5, coords_df$cont_g1)
write.csv(cont_conf, "continent_confusion.csv")
```

Does not look like there is any intelligent clustering by geography.

Let’s check the adjusted Rand index to be sure of this.

```
# adjusted Rand index
adjusted_rand <- adjustedRandIndex(coords_df$cluster5, coords_df$cont_g1)
print(adjusted_rand)
```

```
## [1] -0.03497156
```

The value close to zero indicates random assignment!