Olympics Data Exploration and analysis

Analyze trends in data and make predictions for future Olympics.

Chinmayi P S (01FB15ECS078)
Data Analytics, Dept. of Computer Science and Engineering
PES University
Bangalore,India
chinmayi2398@gmail.com

Devika Mishra (01FB15ECS093)
Data Analytics, Dept. of Computer Science and Engineering
PES University
Bangalore,India
mishra.devika23@gmail.com

***Abstract*— A country's success in sport can be evaluated relative to a small number of demographic and economic characteristics, such as population size and GDP per capita.**

In this project we try to evaluate the effects of these major factors on medal count:
1) Population size
2) GDP per capita.
3) Whether a country is the host of the Olympics.
4) Performance in previous Olympics .

Also analyzing the health conditions (BMI) of the population and its correlation to the performance of players.

Visualizing the trends in performance of players/countries and predicting what can be the results in the next coming Olympic events.

Keywords— POPULATION, GDP, CCR MODEL, REFERCENCE SET COUNT, BMI.

I. INTRODUCTION TO THE CONTEXT OF THE PROBLEM

The modern Olympic Games, held quadrennially in different countries, are considered as the most important sport events, and maybe among the most influential activities in the world. For almost all of the top athletes from all over the world, the Olympic Games are the best arena to show their excellence. So the achievement in the Olympic Games for one country is regarded as one of the most important success, although the games are to seek more friendship and intercommunion than competition.

A country's performance in hosting events and performing at it is an assertion of it's ability to compete at the global stage and this brand is very essential for business.
Data analysis has played an important role in providing that extra edge required to excel in various fields, not just sports. Data will help inform decision-making. Our aim here is to analyze the trends in data, make predictions that can benefit player, coaches and sports representatives of countries.

Spotting patterns in the data is very important. By looking for patterns across different datasets and combining that knowledge to improve performance, this applies to sports teams and businesses alike.

II. WHAT HAVE OTHERS DONE TO SOLVE IT?

A) Assumptions Made

1. Talented athletes were randomly distributed in the world population (WHO WINS THE OLYMPIC GAMES: ECONOMIC RESOURCES AND MEDAL TOTALS).

2. Wealthier countries are more likely to have individuals, organizations, governments willing to make an investment into sports.(WHO WINS THE OLYMPIC GAMES: ECONOMIC RESOURCES AND MEDAL TOTALS).

3. Factors that are used to determine the levels of success in sports for developed countries are not necessarily the same as for developing countries.

4. Studies make the assumption that sporting talent is equally distributed throughout the world, and every nation has equal opportunity of producing successful athletes.

5. Olympics affects the Electric power supply of countries like Brazil which in turn drastically affects the living conditions of the remaining population.

6. Weight restrictions are imposed on two parameters - GDP AND POPULATION in order to increase the validity of the results.

7. Gold , silver , bronze i.e. each type of medal is given a different weight .Example : 3 points for a gold , 2 points for a silver , 1 point for a bronze.

B) Approach used - a summary

1. The likelihood-ratio test is used for the prediction. Both population and GDP factors are considered. As these factors have there net value to be very large the values have been log transformed. Uses dummy variables to capture the effect of certain factors like the countries ability to mobilizing resources , etc.

2. Multiple linear regression has been used. WtdOlympic = b0 + b1 Population + b2 GDP + b3 Climate + b4 Elite + e

3. The actual relationship predicted by the talent distribution is that the expected the medal share accruing to a country should be equal to its share of the total population of countries participating in the Olympics: E(medalshare) = $medals_{it}$ / $medals_{jt}$ = $population_{it}$ / $population_{jt}$ = $popshare_{it}$

4. Reference set count, was used to rank efficient countries, that is, the efficient units are ranked by simply counting the number of times they appear in the reference sets of inefficient units.

5. Use of the "gravity" model of international trade to determine Olympic Effect on Trade

C) Summary of the results reported

1. GDP is indeed a very important factor . Besides GDP , population size and elite facilities are also important.

2. Climate does not play a major role in sporting success. Furthermore , olympic medal count also depends on the level of investment in education and health.

3. The similarity of the coefficients on log population and log GDP per capita suggests that log GDP is the relevant determinant of country medal shares. A likelihood ratio test rejects the equality of these coefficients at the 5% level.

4. Using the DEA's CCR model, a total of seven countries are considered to be efficient in Athens 2004, namely Australia, Bahamas, China, Cuba, Ethiopia, Russia and Ukraine. Since this model fails to differentiate among efficient countries, Utilizing the reference set count to rank the efficient countries, and Cuba (36 times) appears more frequently as benchmarks, followed by Australia (25 times), Bahamas (20times), Russia (16 times), Ukraine (4 times), Ethiopia (3 times) and China (1 time).

D) Limitations Reported

1. The work already been done specifically looks only into the GDP criteria, independent of the other probable factors like population, host country, ratio of the number of participants to medals won etc., which could have a direct after on the performance of a country as a whole. Moreover the work done till now has been very specific to countries like China, USA, Spain etc. Or focuses on a sub-group of countries that are relatively poor (developing countries).

2. The earlier works have taken into consideration a very limited period of time only a few Olympics have been considered and not all.

3. The DEA's CCR model used fails to differentiate among efficient countries. As the parameters increase, the size of the output multipliers feasible region in DEA model gets smaller and, therefore, the efficiency in DEA's CCR model generally decreases.

4. Although traditional CCR model provides benchmarks for inefficient countries, it has certain limitations. The main issue is that an inefficient country and its benchmarks may not be inherently similar in their practices.

E) Any lacuna in their approach/ evaluation that you inferred?

The methodology in previous studies are all limited to only one technique.

For example, the AN ECONOMIC ANALYSIS OF SPORTS PERFORMANCE IN AFRICA has only used multiple linear regression.

Also the paper which analyzed the overall performances of all countries used the CCR model of DEA which had the limitations mentioned above. In order to overcome these limitations we would cluster the countries based on "significant" parameters and try and overcome the limitations.

We will try and use a few different techniques and combine the results as effectively as possible.

III. PROPOSED PROBLEM STATEMENT WITH SPECIFIC ISSUE

Analysis of Sports data from the economical and social aspects.

In this project we try to take into account the different factors like population, BMI, GDP , Number of athletes and possible trends in the performance of athletes individually and their contribution to the performance of a country as a whole.
We try to answer the question: factors that matter for Olympic performance. Comparing GDP and Cost per athlete, and total cost (does a country's GDP affect the performance of individual and overall performance of all athletes?)

We also will try and predict the performance of different countries in the upcoming 2020 Olympics by training the model for data available till 2008 and testing on 2012 and 2016 data.

We will also try to check if any correlation exists between Olympic success of the athletes of a country and the BMI in the general population of that country.

IV. HOW IS OUR APPROACH DIFFERENT?

There are several visualizations made to understand the data. There are graphs for medals by discipline , athlete maps , total medals won by different countries , medals won by men and women in each year etc.

We try to answer the question: factors that matter for Olympic performance.
We also will try and predict the performance of different countries in the upcoming 2020 Olympics by training the model for data available till 2008 and testing on 2012 and 2016 data.

We will take into consideration all the Olympic Games that have been held till date.

Most Olympic medal predictions assess athletic talent by sport and predict winners in each event. We will follow a more generalized approach, and consider the major factors like population, GDP etc., and not look at specific athletes or their sporting talents.

We will use multiple techniques to predict performance in upcoming Olympics and combine, compare the results to get the best possible outcome. Using multiple linear regression, deep learning and a few other methods.

We will also perform a correlation test between BMI of population and the medals won. For this we consider weighted count, Gold: 3 points, Silver : 2 points, Bronze : 1 point . This is done to give more importance to the type of medal, rather than just the number.

V. REFERENCES

[1] Andrew B. Bemard and Meghan R. Busse, "WHO WINS THE OLYMPIC GAMES: ECONOMIC RESOURCES AND MEDAL TOTALS".

[2] Paulo Gomes, Antonio Felipe Aquino, Sumara Duarte Ticom ,Bernardo Fernandes, and James Feltes, "HOW BRAZIL AIMS FOR GOLD IN RELIABILITY".

[3] John Manuel Luiz , Riyas Fadal , "AN ECONOMIC ANALYSIS OF SPORTS PERFORMANCE IN AFRICA".

[4] Jie Wu a, Liang Liang a and Feng Yang a, "ACHIEVEMENT AND BENCHMARKING OF COUNTRIES AT THE SUMMER OLYMPICS USING CROSS EFFICIENCY EVALUATION METHOD".