|  |  |
| --- | --- |
| Chess best opening index | Data analysis and visualization - CA1 Specification Index Generation and Visualization  **Alexandre Desbos** |

Table of Contents

[Theorical framework: 1](#_Toc165831239)

[Methodology 2](#_Toc165831240)

[Data selection and preparation 2](#_Toc165831241)

[Data Distribution 2](#_Toc165831242)

[Data Aggregation and Modifications 4](#_Toc165831243)

[Data Cleaning 5](#_Toc165831244)

[Data Normalization 6](#_Toc165831245)

[Multivariate Analysis 7](#_Toc165831246)

[Effectiveness Indicator 7](#_Toc165831247)

[Complexity Indicator 7](#_Toc165831248)

[Popularity Indicator 8](#_Toc165831249)

[Improvement Indicator 8](#_Toc165831250)

[Overall Index 9](#_Toc165831251)

[Findings 11](#_Toc165831252)

[The Results 11](#_Toc165831253)

[Link to other Indicators 12](#_Toc165831254)

# 

# Theorical framework:

In chess, the selection of an opening often plays a pivotal role in determining a player's success. To address this, I have developed the "Best Chess Opening Composite Index", designed to quantify the multifaceted nature of chess openings. This index integrates data across three concept: effectiveness, popularity, and complexity, offering a comprehensive resource for players at all skill levels to take well-informed decisions about their opening strategies. By adjusting the weighting of indicators, I also generate 2 supplement indexes, one for beginner and one for experiment players to allow a choice of opening levels.

The data underpinning this index come from an extensive database (<https://www.kaggle.com/datasets/alexandrelemercier/all-chess-openings>) that encompasses a vast array of recorded games. This diverse dataset ensures that the index is robust and reflective of strategies employed across the entire spectrum of the chess-playing community. The variables integrated into the index include quantifiable measures such as win and draw percentages, frequency of opening utilization, and number of moves of an opening.

By using this data, the index provides a nuanced view of the strategic value of different openings. It serves as a tool for strategic preparation and decision-making, enabling players to choose openings that not only align with their personal style and strengths but also enhance their chances of winning or securing a draw.

# 

# Methodology

The analysis is based on a public dataset on chess opening, with games from many years and a big variety of variables, This ensures the data is strong and varied, representing a broad range of player abilities and strategies to make a relevant analysis. Here is the methodology I have used to create the index.

## Data selection and preparation

The dataset contains a variety of variables, so the first step was to analyse them to select the variables I can use to build my 4 sub-indicators, which will then enable me to create my index.

Data select:

* **Number of game** : The total number of games played with this opening.
* **Perf Rating**: The average performance rating of players who have played this opening.
* **Player Rating**: The overall average rating of players in the dataset.
* **Player Win %:** The win rate for players using the opening.
* **Draw %**: The percentage of games that ended in a draw.
* **Opponent Win %:** The win rate against players using the opening.
* **Moves List**: A comprehensive list of all moves made in the opening sequence.

Data create using the dataset:

* **Number of variations:** The number of variations for each opening
* **Delta Perf:** Difference between player Rating and his performance rating

## Data Distribution

Plot of variable distributions to assess the modifications and aggregations to be made.

A graph of different colored bars

Description automatically generated

A graph of number of num variations

Description automatically generated

A graph of a number of players

Description automatically generated

### 

### Data Aggregation and Modifications

In order to have more relevant data, I had to make aggregation and modification of variables.

First of all, I change the variables “*moves list”* to “*number of moves”* by changing it to an int of the length of the list.

Then, in the dataset I'm using, the openings are divided for each variation of it, so I had to aggregate the data by opening. For each variation of an opening I added up the number of games and averaged the other variables (perf rating, player rating, player win%, draw%, opponent win%, number of moves).

This also allowed me to create a new variable, ‘Number of variations’, by adding up the number of variations for each opening, which I'll be able to use for the Complexity sub-index.

Finally, I have calculated and add a Delta Perf variable by calculating the difference between the average rating and the performance rating. This variable is useful for determining whether a player can get a better score for a game with a specific opening than for all his games.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Opening** | **Num Games** | **Perf Rating** | **Avg Player** | **Player Win %** | **Draw %** | **Opponent Win %** | **moves\_list** |
| **0** | Alekhine Defense, Balogh Variation | 692 | 2247 | 2225 | 40.8 | 24.3 | 35.0 | ['1.e4', 'Nf6', '2.e5', 'Nd5', '3.d4', 'd6', '4.Bc4'] |
| **1** | Alekhine Defense, Brooklyn Variation | 228 | 2145 | 2193 | 29.8 | 22.4 | 47.8 | ['1.e4', 'Nf6', '2.e5', 'Ng8'] |
| **2** | Alekhine Defense, Exchange Variation | 6485 | 2244 | 2194 | 40.8 | 27.7 | 31.5 | ['1.e4', 'Nf6', '2.e5', 'Nd5', '3.d4', 'd6', '4.c4', 'Nb6', '5.exd6'] |

Before aggregation and modifications:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Opening Name** | **Num Games** | **Perf Rating** | **Avg Player** | **Player Win %** | **Draw %** | **Opponent Win %** | **Avg Num Moves** | **Num Variations** | **DeltaPerf** |
| **0** | **Alekhine Defense** | 34710 | 2207.925925925930 | 2208.44 | 36.133 | 26.78 | 37.08 | 7.62 | 27 | -0.51 |
| **1** | **Anderssen Opening** | 1308 | 2124.0 | 2126.0 | 35.7 | 25.6 | 38.7 | 1.0 | 1 | -2.0 |
| **2** | **Benko Gambit** | 24543 | 2245.0588235294100 | 2229.29 | 40.13 | 25.17 | 34.68 | 10.58 | 17 | 15.76 |

After aggregation and modifications:

### Data Cleaning

I ensured the consistency of the data by cleaning it up. This process involved:

1. **Removing Duplicates**: No duplicates were found in the dataset

A black and white background with white text

Description automatically generated

1. **Handling Missing Values**: I didn't need to impute any data because there were no missing values.

A screen shot of a computer

Description automatically generated

### Data Normalization

To ensure that each variable contributed equally to the final index without bias from different scales or units, I applied the Min-Max normalization technique. This method was used for each variable and enabled them to be scaled within a range of 0 to 1.

Before applying the normalisation on the “Number of games”, I applied a Log function to make it more symmetric because it was skewed data that gave me not relevant results.

Data after normalization:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Opening Name** | **Num Games** | **Perf Rating** | **Avg Player** | **Player Win %** | **Draw %** | **Opponent Win %** | **Avg Num Moves** | **Num Variations** | **DeltaPerf** | **Log Num Games** |
| **Alekhine Defense** | 34710 | 0.67 | 0.63 | 0.55 | 0.45 | 0.56 | 0.52 | 0.90 | 0.55 | 0.65 |
| **Anderssen Opening** | 1308 | 0.51 | 0.470 | 0.54 | 0.41 | 0.51 | 1.0 | 1.0 | 0.54 | 0.28 |
| **Benko Gambit** | 24543 | 0.75 | 0.68 | 0.71 | 0.39 | 0.63 | 0.31 | 0.94 | 0.65 | 0.61 |

These standardised data form the basis of my multivariate analysis, making it possible to construct a reliable and significant composite index of chess openings and to easily choose the contribution of variables with weightings.

A screenshot of a graph

Description automatically generated

## 

## Multivariate Analysis

I made a multivariate analysis for each of my sub-indicators and then on my overall index to make sure my indexes are consistent.

### Effectiveness Indicator

* Player Win %
* Draw % =>Direct outcomes when the opening is used.
* Opponent Win %

A graph with a red line and blue dots

Description automatically generatedA screen shot of a graph

Description automatically generated

A graph with a red line

Description automatically generated

A screenshot of a computer

Description automatically generated

We can see that the model have a good R-squared and adjust R-square, the standar errors is very low (good).

### Complexity Indicator

To build the complexity indicator, I wanted to use two variables, so I performed a scatterplot analysis to spot and avoid potential multicollinearity.

Variables for complexity indicator:

* Number of moves: The length and complexity of the opening moves can indicate strategic depth.
* Number of variations: The number of possible move sequences can reflect the complexity of the opening.

Results of the scatterplot analysis and correlation:

A graph with blue dots and a red line

Description automatically generated

*PearsonRResult(statistic=0.5471878958641806, p-value=2.8813028407624773e-08)*

We can see the person correlation coefficient value is 0.54, that represent a moderate relationship between the two variables and the p-value is way smaller than 0.05 so it’s very significant.

With this analysis, I decided to keep both of my variables because a correlation of 0.55 is generally not high enough to cause concerns on about multicollinearity and I believe both variables contribute unique information to the index and reflect different aspects of the complexity of chess opening.

### Popularity Indicator

The popularity indicator uses only the number of games played, so no analysis has been carried out for this indicator.

* Num Games: Represent the number of times the opening is used.

### Improvement Indicator

The improvement indicator also uses only on variable, the Delta perf to represent if in average, playing an opening can have a better performance rating than player rating. It’s better to use Delta perf than perf rating and Avg rating because they are highly correlated.

A graph with blue dots and a red line

Description automatically generated

The Avg player and Perf Rating are highly correlated (0.9636) and this is very significant (pvalue < 0.05)

* Delta perf : Difference between player Rating and his performance rating

### Overall Index

**Cluster Analysis**

A graph with a dotted line

Description automatically generated

A collage of graphs

Description automatically generated

A screenshot of a graph

Description automatically generated

A graph with blue dots and a red line

Description automatically generated

The correaltion Matrix of the indicators show that Effectiveness and Improvement look highly correlated and this is confirm by the scatterplot, the p-value is < 0.05 so the corelation is very significant.

# Findings

## The Results

*The complete results from the index, including the overall result of 10 best opening as well as the individual rankings within the 3 categories. (complete result in the file final\_data.csv)*

|  |  |  |
| --- | --- | --- |
| **Opening Name** | **Overall** |  |
| **Indian Game** | 0.741 |  |
| **Queen Pawn Opening** | 0.726 |  |
| **Trompowsky Attack** | 0.724 |  |
| **Zukertort Opening** | 0.721 |  |
| **King's Pawn Opening** | 0.712 |  |
| **King's Indian Attack** | 0.711 |  |
| **Nimzo-Larsen Attack** | 0.709 |  |
| **Torre Attack** | 0.705 |  |
| **Queen Pawn Game** | 0.704 |  |
| **Bird Opening** | 0.691 |  |

## Link to other Indicators

https://thechessworld.com/articles/openings/chess-statistics-top-10-best-openings-for-white-and-black/

A table with numbers and a number of points

Description automatically generated

https://www.chess.com/forum/view/chess-openings/100-most-popular-openingsA screen shot of a computer

Description automatically generated