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

Table of Contents

[Theorical framework: 1](#_Toc165820199)

[Methodology 2](#_Toc165820200)

[Data selection and preparation 2](#_Toc165820201)

[Data Aggregation 3](#_Toc165820202)

[Data Cleaning 3](#_Toc165820203)

[Data Normalization 4](#_Toc165820204)

[Multivariate Analysis 4](#_Toc165820205)

[Effectivness Indicator 4](#_Toc165820206)

[Improvement Indicator 5](#_Toc165820207)

[Complexity Indicator: 5](#_Toc165820208)

[Popularity Indicator: 5](#_Toc165820209)

[Overall Index 5](#_Toc165820210)

[Findings 5](#_Toc165820211)

[The Results 5](#_Toc165820212)

[Link to other Indicators 6](#_Toc165820213)

[Clustering 7](#_Toc165820214)

# 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 is derived from an extensive database (<https://www.kaggle.com/datasets/alexandrelemercier/all-chess-openings>) that encompasses a vast array of recorded games, ranging from amateur matches to high-stakes grandmaster confrontations. 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 detailed move sequences.

By synthesizing this data, the index provides a nuanced view of the strategic value of different openings. It serves as a vital 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 under various competitive conditions.

# Methodology

The analysis is based on a public dataset that I have chosen carefully, with games from many years, This ensures the data is strong and varied, representing a broad range of player abilities and strategies, and with a full set of variables to ensure that the analysis will be relevant.

## Data selection and preparation

My 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 Aggregation

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

First of all, in the dataset I'm using, the openings are divided for each variation of it, so I had to aggregate the data by opening. So 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 add a new variable, ‘Number of variations’, by summing the number of variations for each opening, which I'll be able to use for the Complexity sub-index.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **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:

After 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 |

Data was aggregated to create a comprehensive profile for each type of chess opening. Each opening’s data points were compiled across various games to calculate average win percentages, draw rates, and usage frequency. Move sequences were analyzed using advanced pattern recognition algorithms to classify and quantify complexity levels.

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

### Data Normalization

To ensure that each variable contributed equally to the final index without bias from different scales or units (and be able to choose the contribution of variables with weights), I applied the Min-Max normalization technique:

This method was primarily used for win and draw percentages, and frequency of use, rescaling them to a 0-1 range.

Before applying the normalisation on the Number of games, I applied a Log function to make it more symmetric and handle skewed data.

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 |

This normalized data forms the basis for my multivariate analysis, enabling to construct a reliable and meaningful composite index of chess openings.

## Multivariate Analysis

I decided to split make a multivariate analysis for each of my sub-indicators and then on my overall index.

### Effectivness Indicator

* Player Win %
* Draw %
* Opponent Win %

=>Direct outcomes when the opening is used.

### Improvement Indicator

* perf : delta between player Rating and his performance rating

### 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.

### Popularity Indicator:

* Num Games: How frequently the opening is used.

### Overall Index

# 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** |  | **Complexity** | **Effectivness** | **Popularity** |
| **Indian Game** | 0.741 |  | 0.828 | 0.575 | 0.790 |
| **Queen Pawn Opening** | 0.726 |  | 1.0 | 0.808 | 0.279 |
| **Trompowsky Attack** | 0.724 |  | 0.851 | 0.650 | 0.627 |
| **Zukertort Opening** | 0.721 |  | 0.868 | 0.504 | 0.743 |
| **King's Pawn Opening** | 0.712 |  | 1.0 | 0.566 | 0.475 |
| **King's Indian Attack** | 0.711 |  | 0.780 | 0.627 | 0.702 |
| **Nimzo-Larsen Attack** | 0.709 |  | 0.957 | 0.517 | 0.569 |
| **Torre Attack** | 0.705 |  | 0.789 | 0.712 | 0.586 |
| **Queen Pawn Game** | 0.704 |  | 0.783 | 0.544 | 0.758 |
| **Bird Opening** | 0.691 |  | 0.891 | 0.544 | 0.570 |

## 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

## Clustering