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

Table of Contents

[Introduction: 1](#_Toc165809821)

[Theorical framework: 1](#_Toc165809822)

[Methodology 2](#_Toc165809823)

[Data Selection 2](#_Toc165809824)

[Data Aggregation 2](#_Toc165809825)

[Data Cleaning 3](#_Toc165809826)

[Data Normalization 3](#_Toc165809827)

[The Results 4](#_Toc165809828)

[Sub-Indicators 4](#_Toc165809829)

[Effectivness 4](#_Toc165809830)

[Popularity 4](#_Toc165809831)

[Complexity 4](#_Toc165809832)

## 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. To perform this analysis and index creation, the first step was to

### Data Selection

From this extensive database, specific data points were chosen for analysis based on their relevance to the study's objectives. The primary variables selected include:

Player Win Percentage: Reflects the effectiveness of openings in securing wins.

Draw Percentage: Indicates the stability and defensive strength of openings.

Frequency of Use: Measures the popularity of each opening among players of all levels.

Move Sequences: Provides insight into the complexity and depth of each opening.

These variables were identified as critical in understanding the multifaceted nature of chess openings and their impact on game outcomes.

### Data Aggregation

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

The aggregated data underwent a thorough cleaning process to ensure accuracy and consistency. This process involved:

Removing Duplicates: Ensuring that each game's data is only counted once.

Handling Missing Values: Imputing missing data where appropriate, or removing data points with insufficient information, based on the nature of the missing data.

Error Checking: Correcting any discrepancies in data entry, particularly in player ratings and game outcomes.

### Data Normalization

To ensure that each variable contributed equally to the final index without bias from different scales or units, we applied normalization techniques:

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

Z-Score Normalization: Applied to the complexity scores derived from move sequences to standardize them around the mean, reducing the impact of outliers.

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

**Data select:**

* Number of game
* Perf Rating
* Player Rating
* Player Win %
* Draw %
* Opponent Win %
* Number of moves
* Number of variations

1. Popularity indicator:

* Num Games: How frequently the opening is used.

1. Effectivness indicator:

* Player Win %
* Draw %
* Opponent Win %

=>Direct outcomes when the opening is used.

1. Improvement indicator:

* perf : delta between player Rating and his performance rating

1. 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.

## The Results

## Sub-Indicators

### Effectivness

### Popularity

### Complexity