

# Project - Positional analysis of chess

Jake Moss - s46409665

April 27, 2021

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Project aims . . . . .	2
1.2	Obtaining data . . . . .	2
1.2.1	Sources . . . . .	2
1.2.2	PGN Format . . . . .	2
<b>2</b>	<b>Data Processing</b>	<b>3</b>
2.1	Tools used . . . . .	3
2.2	Meta Data Extraction . . . . .	4
2.3	Extracting Implicit Data . . . . .	4
<b>3</b>	<b>Data Analysis</b>	<b>4</b>
3.1	Local and Global Normalisation . . . . .	4
<b>4</b>	<b>Results</b>	<b>5</b>
4.1	Plot analysis . . . . .	5
4.1.1	ELO grid plots . . . . .	5
4.1.2	Date grid plots . . . . .	7
4.1.3	Individual piece plots . . . . .	8
4.1.4	Individual player plots . . . . .	9
4.2	Conclusion . . . . .	9
4.2.1	Areas for improvement . . . . .	9
<b>5</b>	<b>Reflection</b>	<b>9</b>
<b>6</b>	<b>Appendix</b>	<b>11</b>
6.1	ELO Grids (FISC) . . . . .	11
6.2	Date Grids (Tournament) . . . . .	15
6.3	DaenaliaEvandruile . . . . .	19
6.4	Individual piece plots . . . . .	22

# 1 Introduction

## 1.1 Project aims

The goal of this project was to provide a new and unique visualisations of chess games. To accomplish this two main plots were developed,

- Heat map of where pieces were captured
- Histogram of when pieces were captured

The purpose of this project is to show the positional differences between different levels of play and in different time periods. This is something not often visualised within the community.

## 1.2 Obtaining data

### 1.2.1 Sources

Chess games are found freely on the internet from many archives such as PGN Mentor and FICS Games Database. From here I am able to download millions of games at mass. However for large time periods most preserved games are of tournaments and high skill players this data doesn't not sure what happens in common games. To help mitigate this the FICS database was used to analyse a wide range of skill level games from a single year. This tool was also build with the capability to analyse a single players games given their username from Lichess and Chess.com. These games can be downloaded using their respective public API's.

To download games from Chess.com the bash command below can be used.

```
for g in $(curl -Ls https://api.chess.com/pub/player/$PLAYERNAME/games/archives | jq -rc  
↪ ".archives[]") ; do curl -Ls "$g" | jq -rc ".games[].pgn" ; done >> games.pgn
```

Lichess easily allows for downloading of an entire players archive at once with a simple curl.

```
curl https://lichess.org/games/export/$PLAYERNAME > games.pgn
```

DaenaliaEvandruile has given their permission for their games to be featured with this report to demonstrate this functionality.

### 1.2.2 PGN Format

The most common format used to store chess games is Portable Game Notation. It is a human readable plain text notation which stores individual moves rather than the board states. While this is considerably more efficient for both storage and transcribing it makes extracting data that isn't explicitly given difficult and costly.

The PGN standard defines a required seven tag roster and allows for optional tags such as WhiteElo and BlackElo. Each move is stored in Standard Algebraic Notation, which describes the change in board state such as captures, checks, and promotions.

SAN comes in a variety of formats short, long, or minimal. Each variant encodes moves with different levels of detail. Unfortunately no single variant is used in all PGN databases, although it is rare for the format to change from within a database.

A standard game might look something like this

```
[Event "WCh 2018"]
[Site "London ENG"]
[Date "2018.11.09"]
[White "Caruana, Fabiano"]
[Black "Carlsen, Magnus"]
[Result "1/2-1/2"]
[WhiteTitle "GM"]
[BlackTitle "GM"]
[WhiteElo "2832"]
[BlackElo "2835"]
[WhiteFideId "2020009"]
[BlackFideId "1503014"]
[EventDate "2018.11.09"]
```

1. e4 c5 2. Nf3 Nc6 3. Bb5 g6 ... 115. Ra8 1/2-1/2

Tags preface the game in [], often far more than required are provided. These tags are referred to as the headers. The moves themselves are listed in SAN with the result at the end.

## 2 Data Processing

### 2.1 Tools used

This project made use of the below tools,

- python-chess Used to handle the complex move validation and PGN parsing required to step through games.
- matplotlib Standard plotting library.
- seaborn A high level plotting library built on matplotlib.
- numpy Standard scientific computing library. Within this project it is used to for its pandas optimisations, and reshape function.
- pandas Data frames provide an efficient way to store and conditionally select games based on metadata.
- cPython Profiler Function call profiler
- snakeViz In browser cPython profiler visualiser.

## 2.2 Meta Data Extraction

To extract the metadata from a game the PGN file is read using standard python methods and the game object is initialised using the `aggragate` function from the `aggregation.py` module. From the game object the headers are extracted into a dictionary and a data frame is created consisting of the headers are column titles and rows as individual games.

	SAN string	Black	White	Result	BlackElo	WhiteElo	Date	...
1	string	string	string	string	int	int	DateTime	...
...								

## 2.3 Extracting Implicit Data

To extract the “implicit data”, such as piece captures, each game must be played out in its entirety with each move analysed as the format games are encoded in does not explicitly give this information.

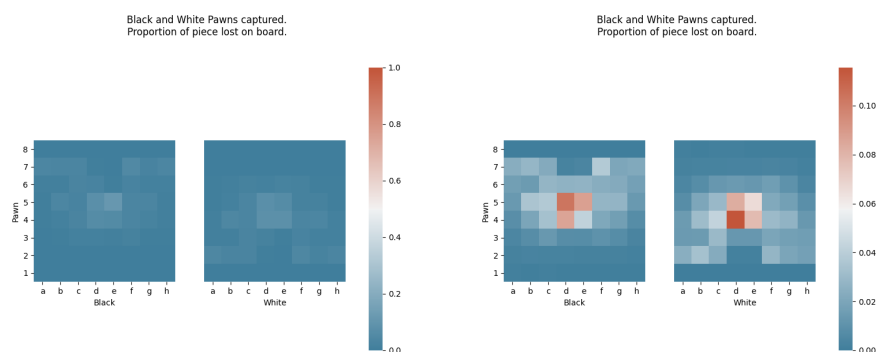
This data is then appended to the metadata data frame in another column. This allows for each conditioning of the data fame.

# 3 Data Analysis

## 3.1 Local and Global Normalisation

Within the heat map grid plots, each map shares the colour bar on the right. To accomplish this each square with each map is normalised to the sum of the local grid so that the map represents proportion instead of frequency. From this we now know that the maximum value is 1 and the colour bar can be calibrated to  $[0, 1]$ . However, as the proportion of lost pieces rarely gets close to 0.5 let alone 1, the colours become hard to differentiate.

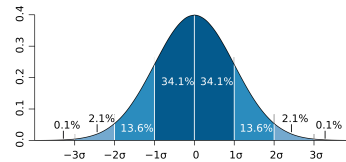
To remedy this I set the maximum of all local maximums to the max of the colour bar. This made it so that each plot is proportional to itself and the colouring is consistent between plots.



A similar strategy was employed to ensure the histograms and kernel density plots shared the same axis.

## 4 Results

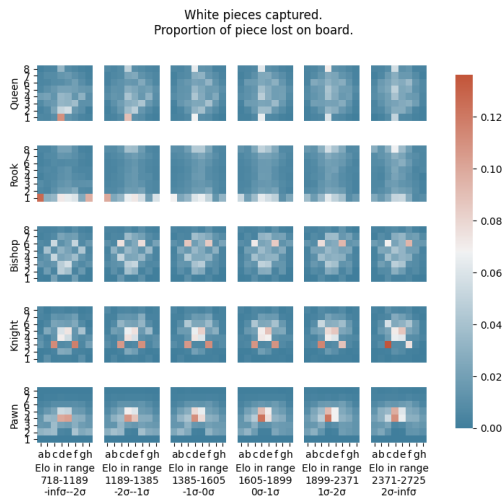
The skill level of players was assumed to be normally distributed. All ELO grid plots have their **x-axis** binned by percentiles from the normal dist. Sigma or  $\sigma$  here denotes one standard deviation [2]. Date grid plots are binned by quartiles.



### 4.1 Plot analysis

#### 4.1.1 ELO grid plots

This first visualisation shows positions where pieces were lost by a particular colour plotted against ELO ranking of the player. Each combination of ELO and piece is represented by an 8x8 heat map of the board from Whites perspective. The decision to make all plots from Whites POV was made in an attempt to improve comparability and consistency between plots.



The unique pattern for each piece shows the most common lines played by White.

Queens typically move through the C,D,E pawns which are the easiest lines out of the starting position.

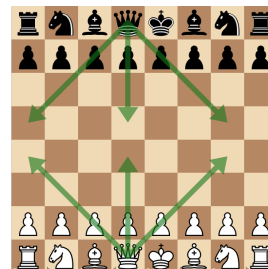


Figure 1: All captured White pieces binned by ELO

As ELO increases all pieces trend towards the centre as it is a highly advantageous position. The A1 and H1 Rooks (A8 and H8 for Black) losses almost disappear in higher ELO's, this is likely due to higher usage, and the utilisation of castling. The highest ELO players show a significant increase in spread of losses indicating that the Rooks are more active through a wider range of centre files. The decrease in losses on positions that are common from castling does not indicate a decrease in castling as the heat maps only show losses, as higher ELO players are more likely to play better positions, they are less likely to castle into a position where their Rook will be lost.

A similar spread increase can be seen in the Queens. Higher ELO players tend to lose their Queens in a much more even manner. This may be a symptom of sacrificing which is common in high level play.

Bishops, Knights, and Pawns do not display a pattern change between ELOs suggesting most players know common and effective positioning, likely from watching grand masters. Although there is negligible change in the patterns the frequency at which they are lost becomes more concentrated. This may indicate that higher ELO player are more likely to keep these pieces near positions where they are strongest.

A similar plot as **Figure 1** is attached in the Appendix which instead shows Black pieces. There is no not worthy pattern difference between the Black and White plot. All trends are present but with 180° rotation.

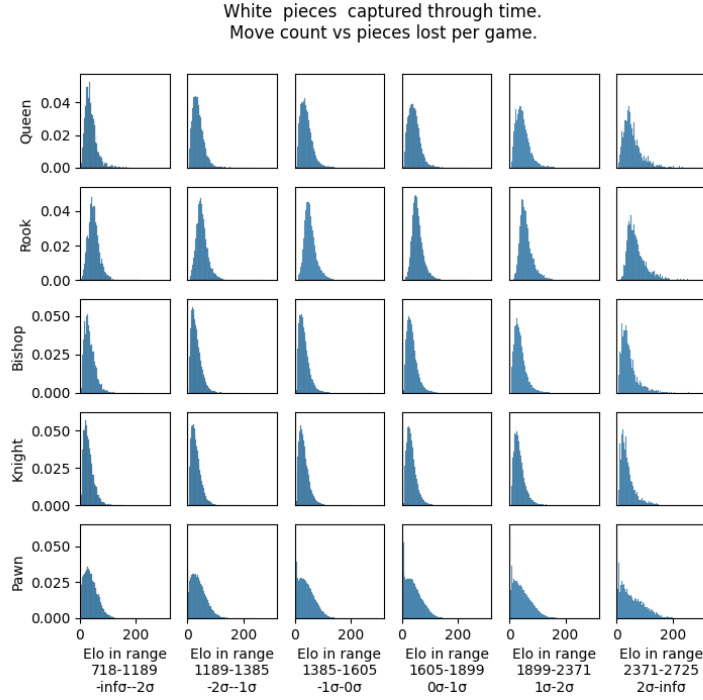


Figure 2: When White pieces were captured binned by ELO

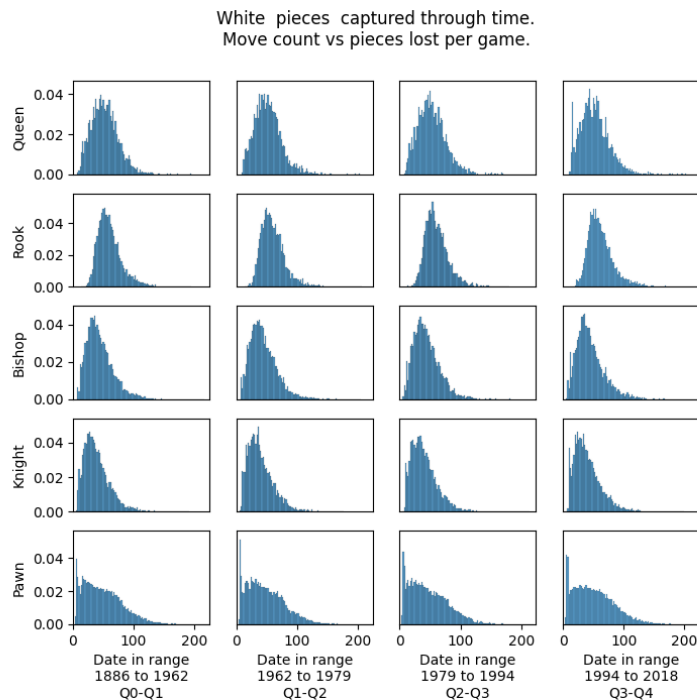
Here the number of captured pieces is divided by the number of games in the respective bin. This is to account for the different game counts due to the distribution of players based on ELO.

Although there is a lower number of games in the more extreme percentiles (col 1 and col 6) some conclusions can still be drawn. A common trend among most all pieces is the gradual spreading effect of pieces lost per game. This is probably due to higher skill players playing longer games. There is often a large spike in Pawn losses around move number 1-6 demonstrating the trading that is common at the beginning of games while competing for control of the centre of the board. This is corroborated by **Figure 1** where the Queen side Pawn is most commonly lost.

Further plots are included in the appendix.

### 4.1.2 Date grid plots

A different database was used for date binning as the FISC database only had games from 2000, instead a database comprising of **Tournaments**, **Candidates Interzonals**, and **World Championships** games were used. (These games were acquired from PGN mentor.)

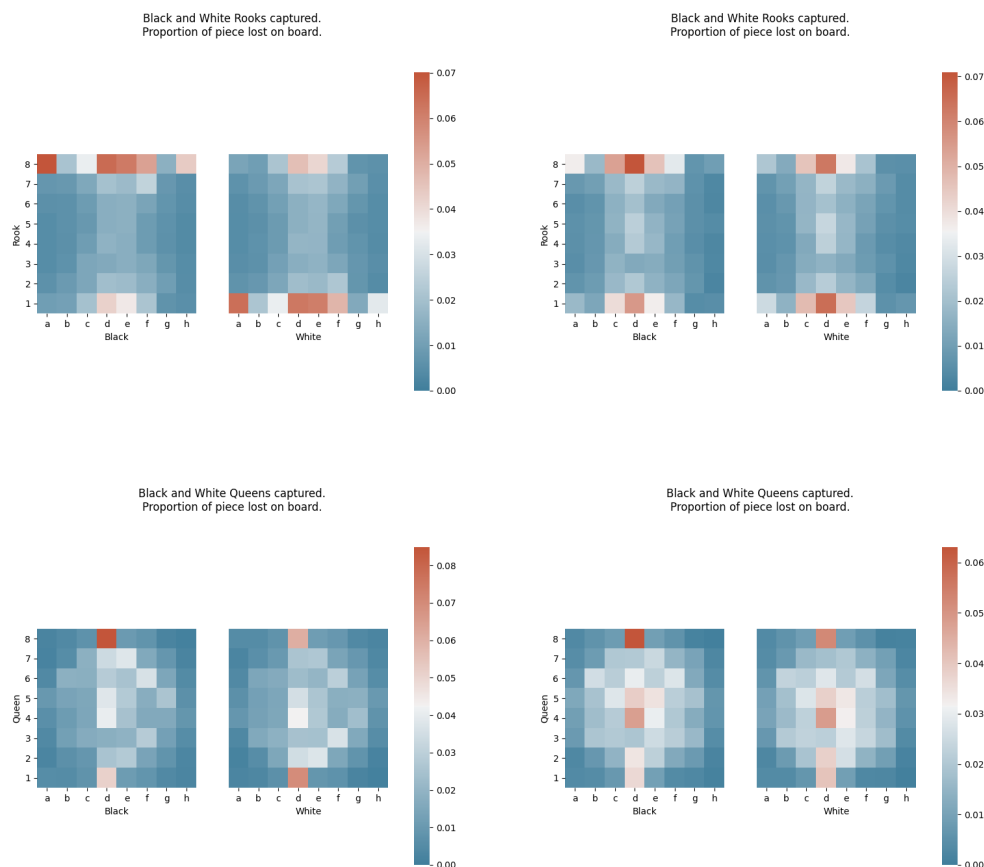


Unfortunately binning by date reveals no interesting patterns or trends. However it does show that chess does not change over the centuries.

Further plots are included in the appendix.

### 4.1.3 Individual piece plots

Left hand side plots are from the **FISC** database. Right hand side from the **Tournements** database.



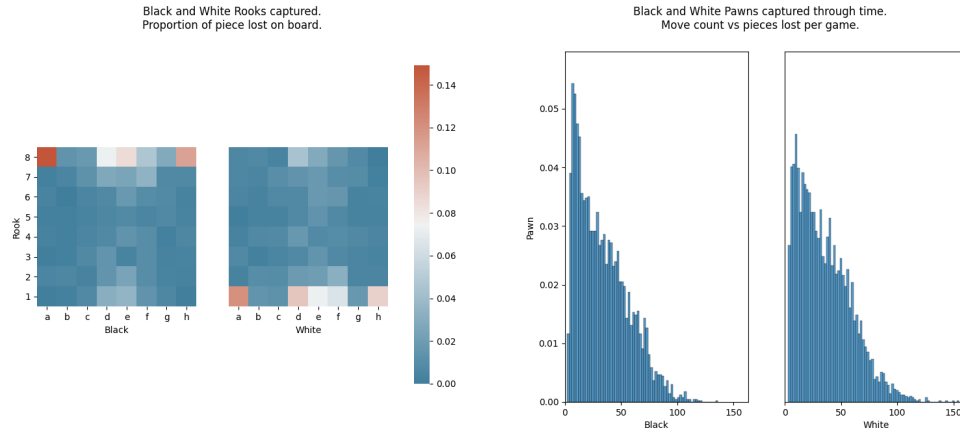
As the **FISC** database contains a varying skill level players the most prominent trends are those of average skill players. Similar to what was seen in **Figure 1**. The **Tournement** plots clearly exhibit patterns exactly the same as those found in high ELOs in **Figure 1** as expected.

These single piece plots help emphasise the individual lines. Interestingly high level players seem to lose their Queens in the same positions regardless of colour. This could be an consequence of first move advantage.



#### 4.1.4 Individual player plots

All previously shown plots are also capable of focusing on a single player. Here we have positions where **DaenaliaEvandruile**'s Rooks were lost and a histogram of when their Pawns were lost. This allows the user to compare their specific play patterns to common trends.



## 4.2 Conclusion

Overall this project has successfully visualised an uncommon statistic of chess which may provide insight into future trends or an individual players style.

All code and resources used are open source and can be found at this repository.

### 4.2.1 Areas for improvement

- Performance and memory requirements prevent me from analysing the huge databases I hoped to. A custom PGN parse and a rewrite in a lower level language such as Rust or Haskell would allow for this.
- More positional statistics such as positions that give and receive checks and checkmates may be interesting to visualise. This is very possible with the current implementation requiring only a few more functions due to the features of the `python-chess` library.

## 5 Reflection

Due to the complexity and design approach of this project I believe it is deserving of a 7. While the `Date` binning did not reveal and trends and patterns I initially hope it would it was successful in showing not much changes in chess over the centuries. As the program was design in with modularity and composability in mind it is highly general, and able to process any game database format and group based on any meta data field with a few changes. Binning based on location or opening would only require a few lines to group the games and the rest can remain unchanged.

No pre-processing is required by the user, all plots are not specific to a single database. Everything is self contained making it easy to work with. Analysing another database

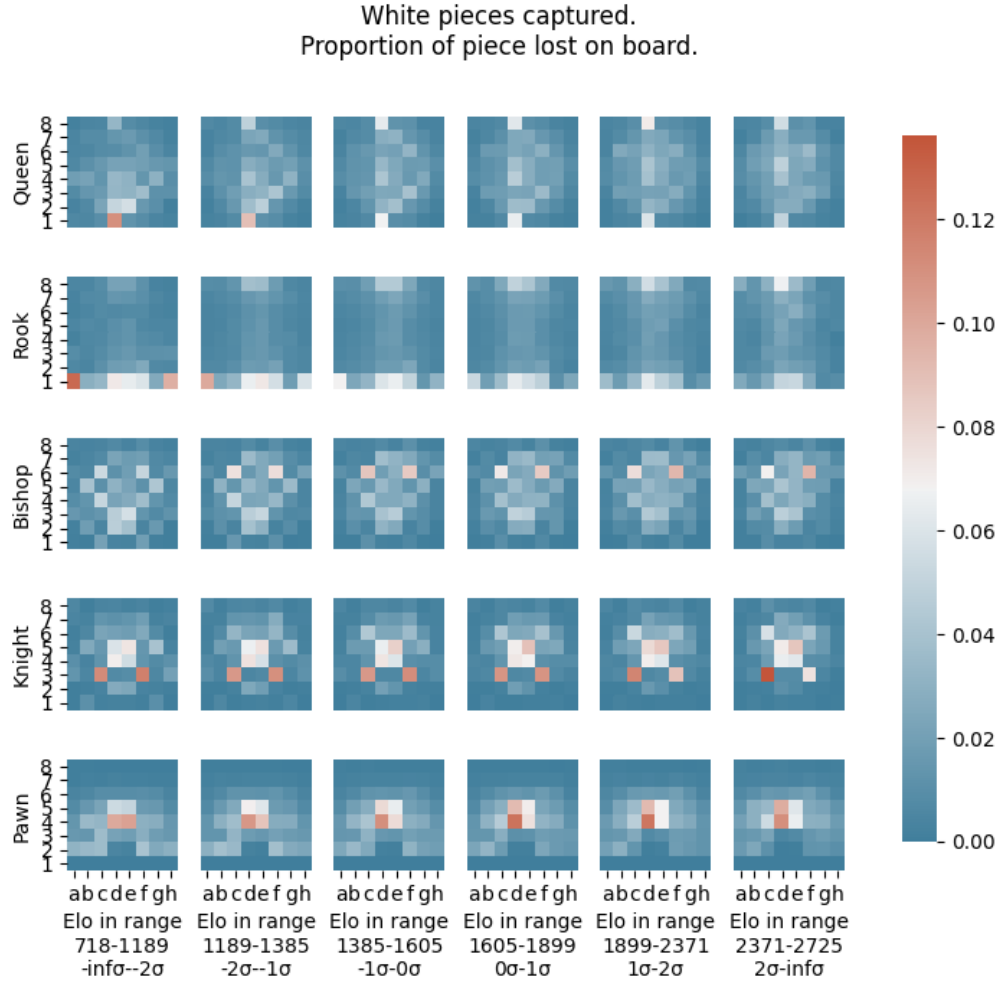
requires a single filename change. While writing this program my goal was not to visualise a single data set but rather to write a tool that achieved that same thing. This added a lot of complexity and time requirements but I believe this was for the best as it forced me to focus on the process itself rather than force a single data set to work for me.

## References

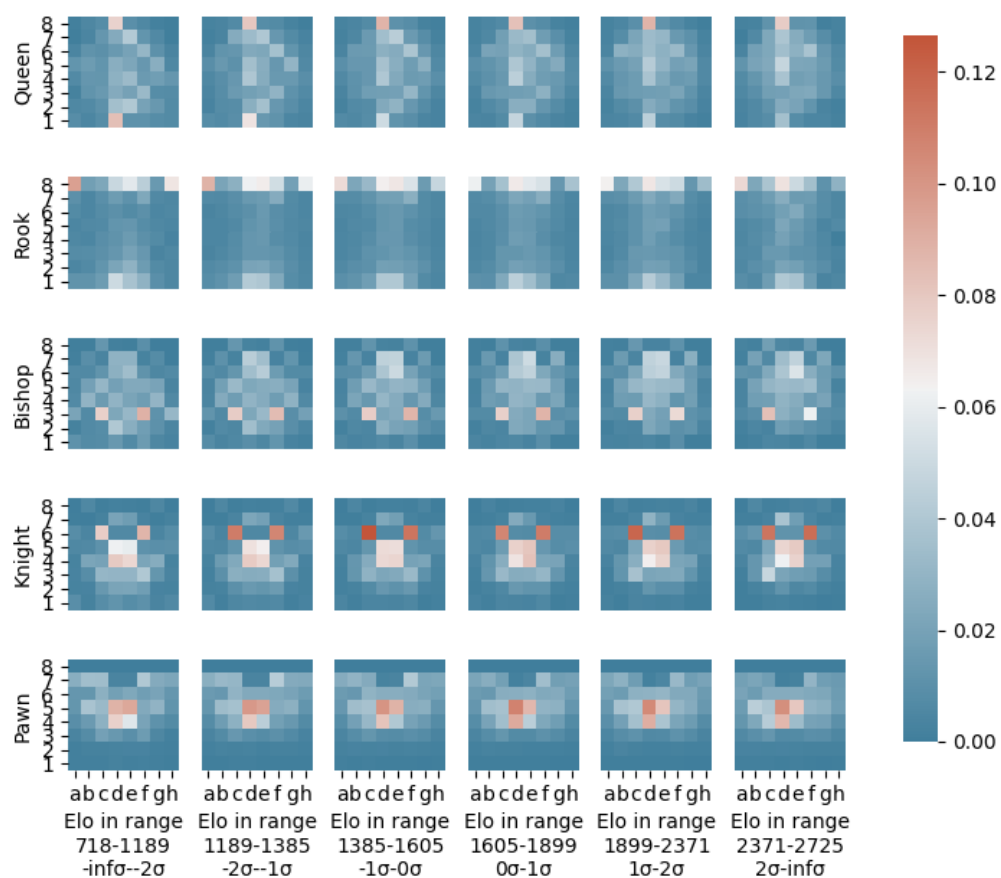
- [1] FIDE. Laws of chess, 2008.
- [2] M. W. Toews. Percentile - wikipedia, 2005.
- [3] Unknown. Soviet calendar - wikipedia, 2021.

## 6 Appendix

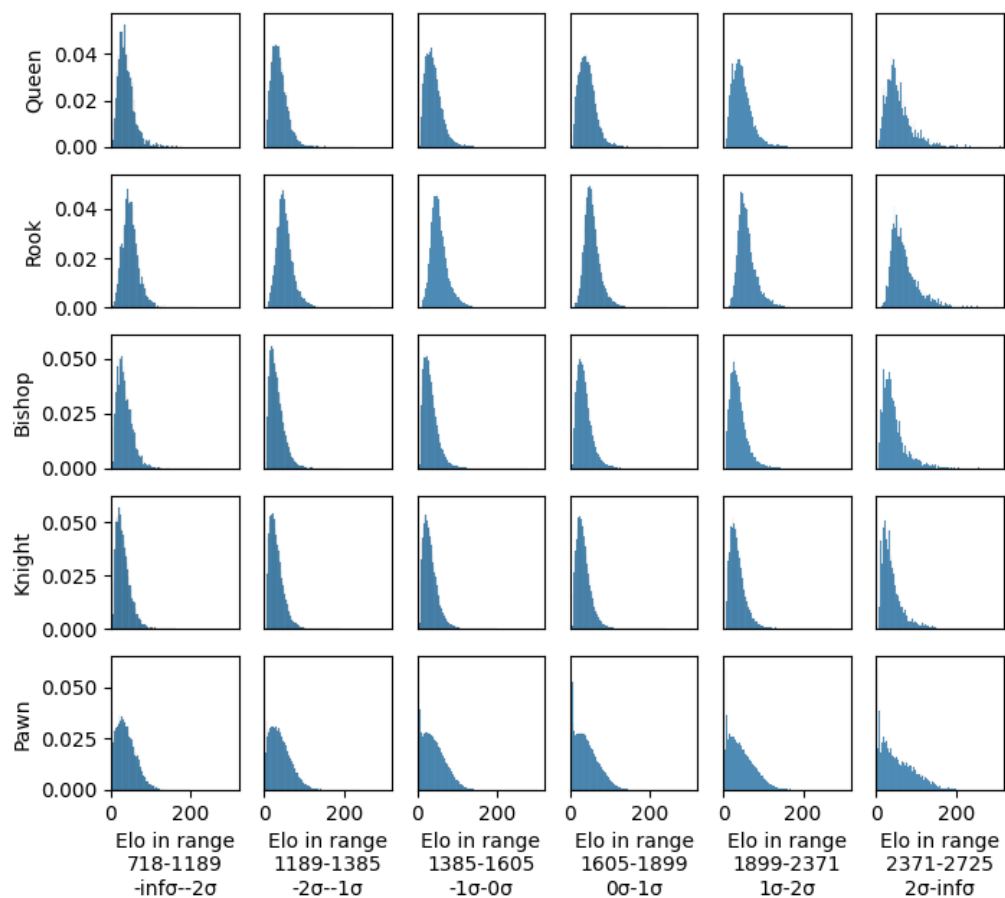
### 6.1 ELO Grids (FISC)



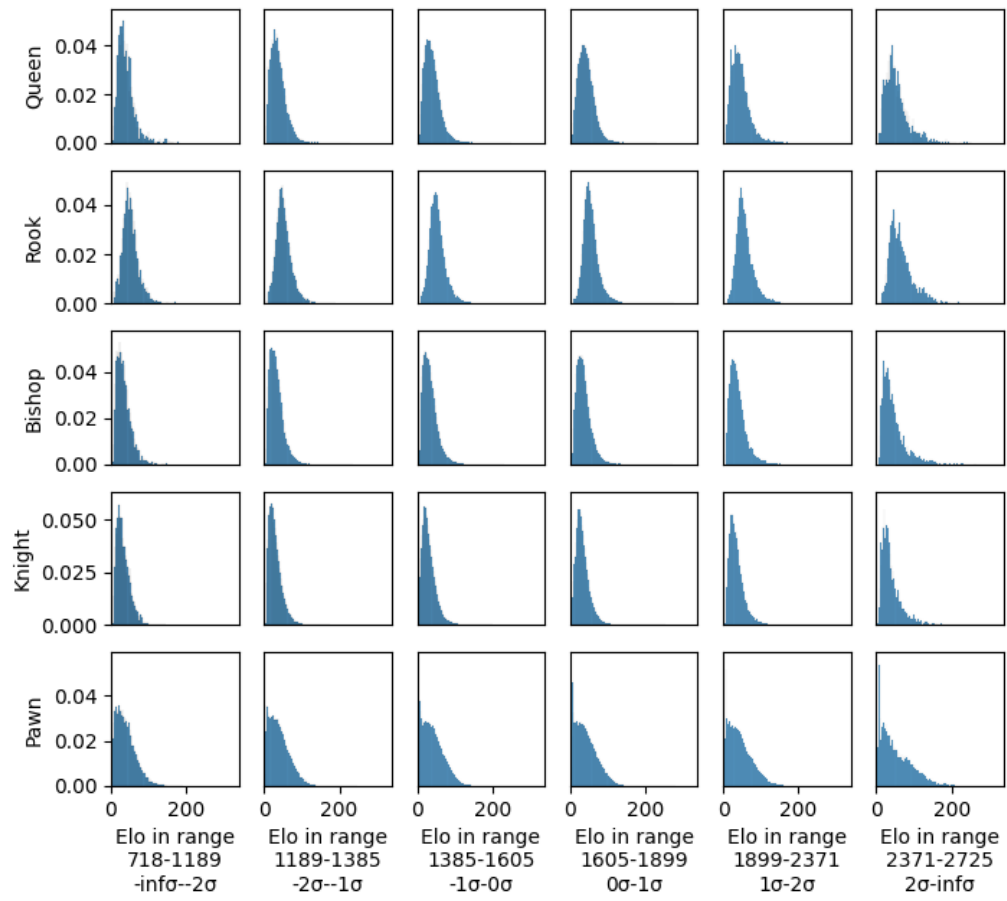
Black pieces captured.  
Proportion of piece lost on board.



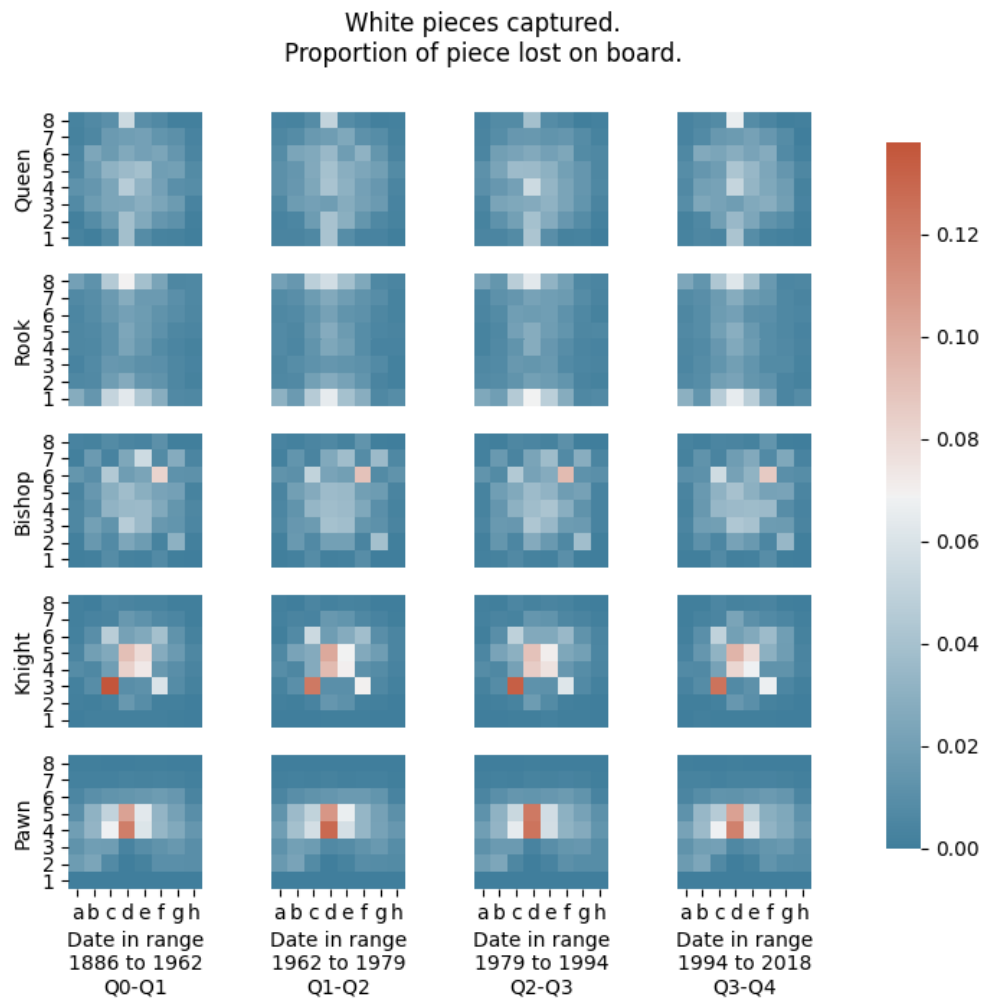
White pieces captured through time.  
Move count vs pieces lost per game.



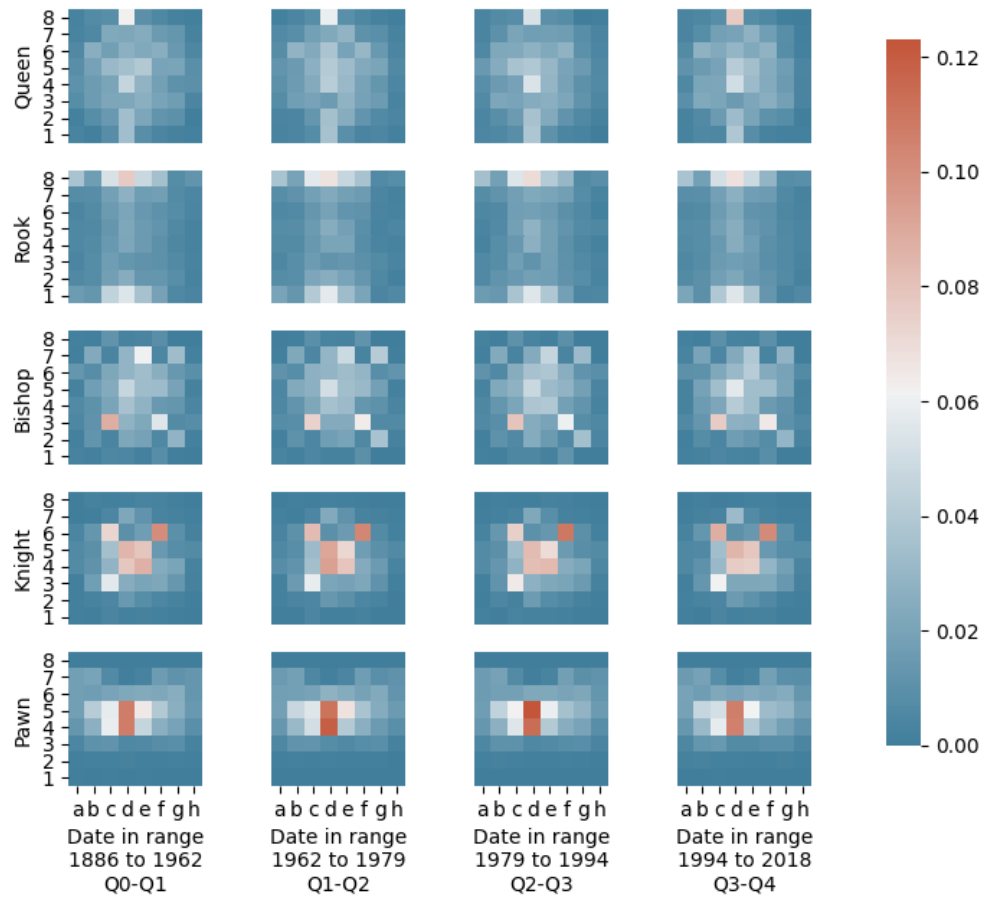
Black pieces captured through time.  
Move count vs pieces lost per game.



## 6.2 Date Grids (Tournament)

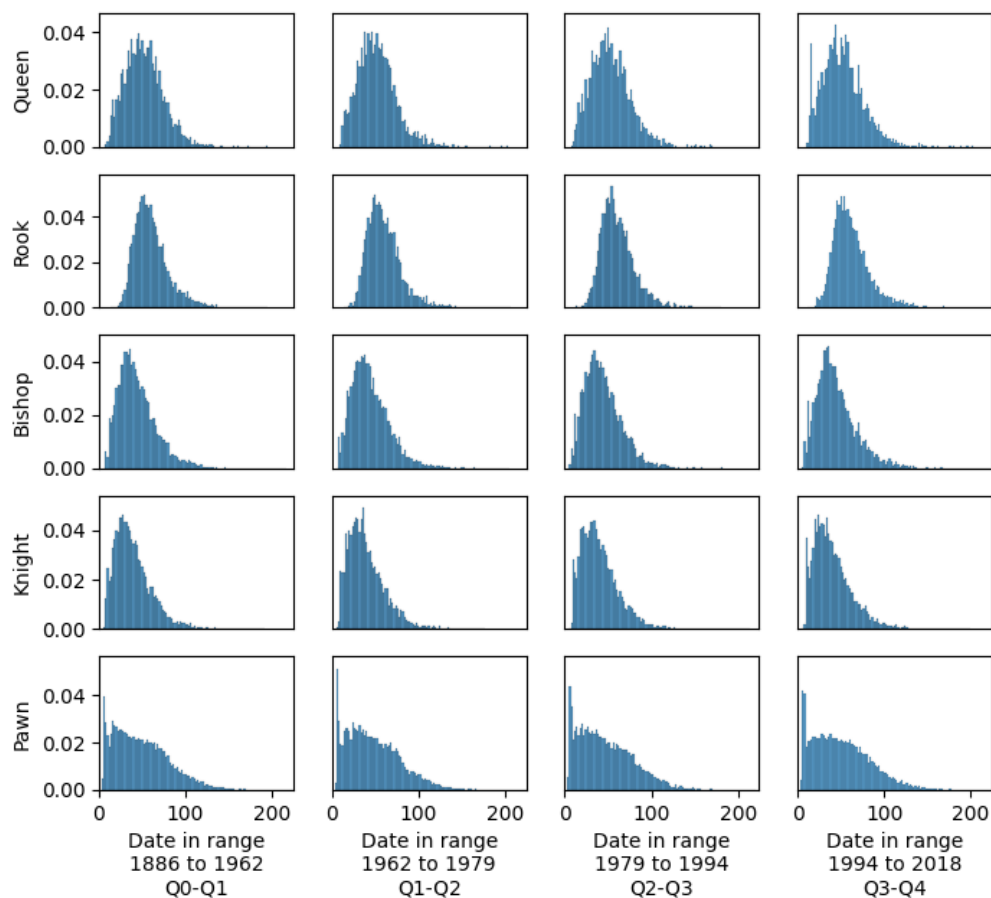


Black pieces captured.  
Proportion of piece lost on board.

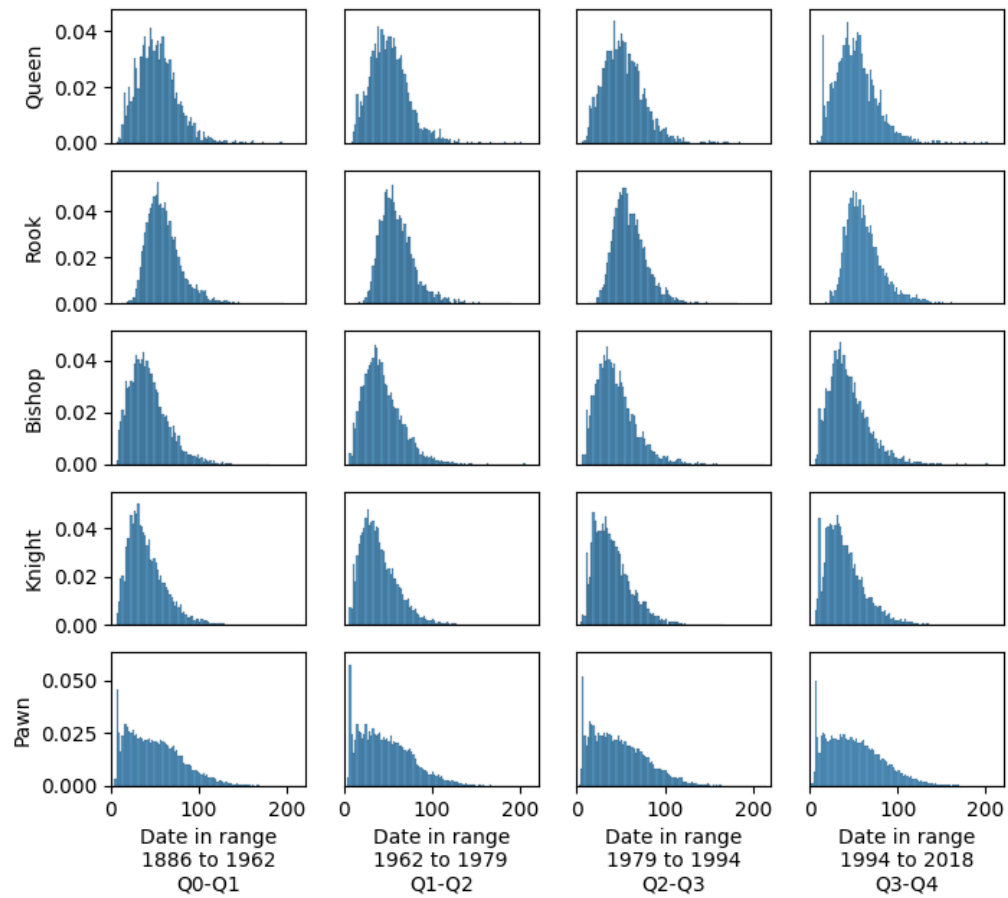




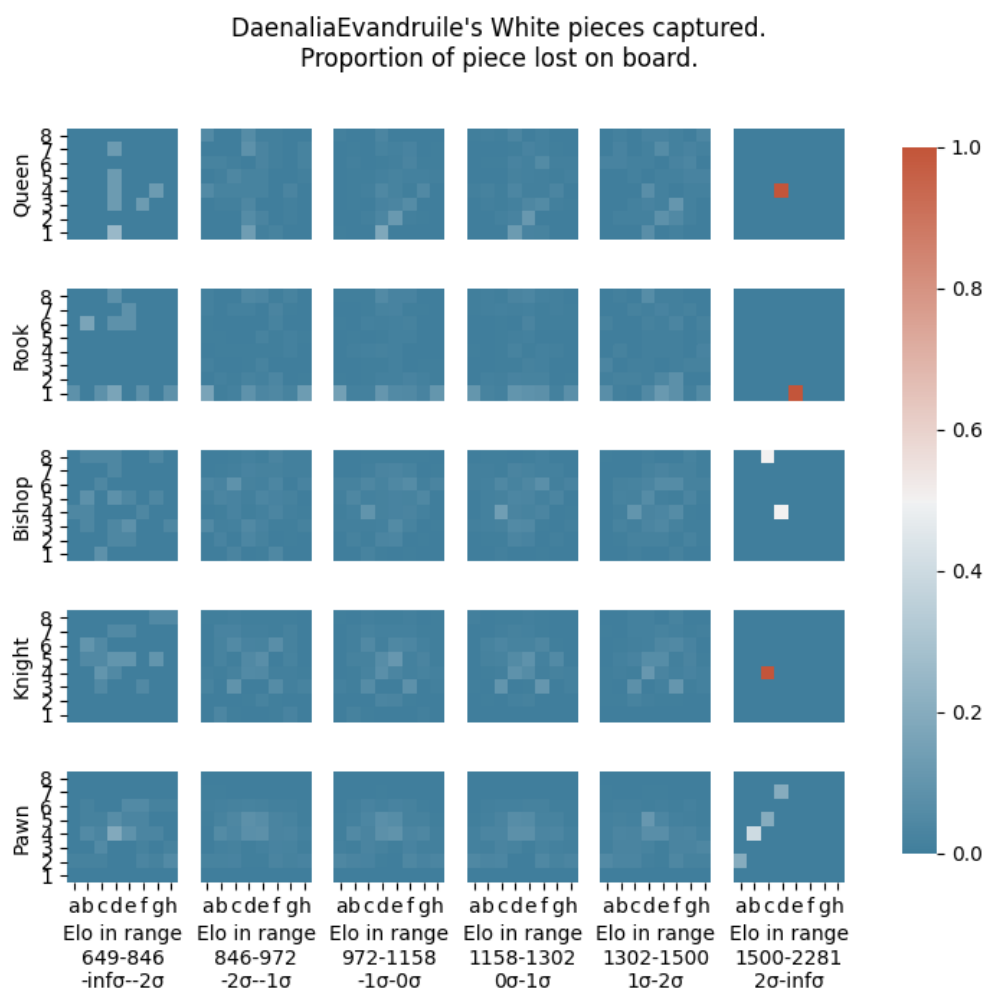
White pieces captured through time.  
Move count vs pieces lost per game.



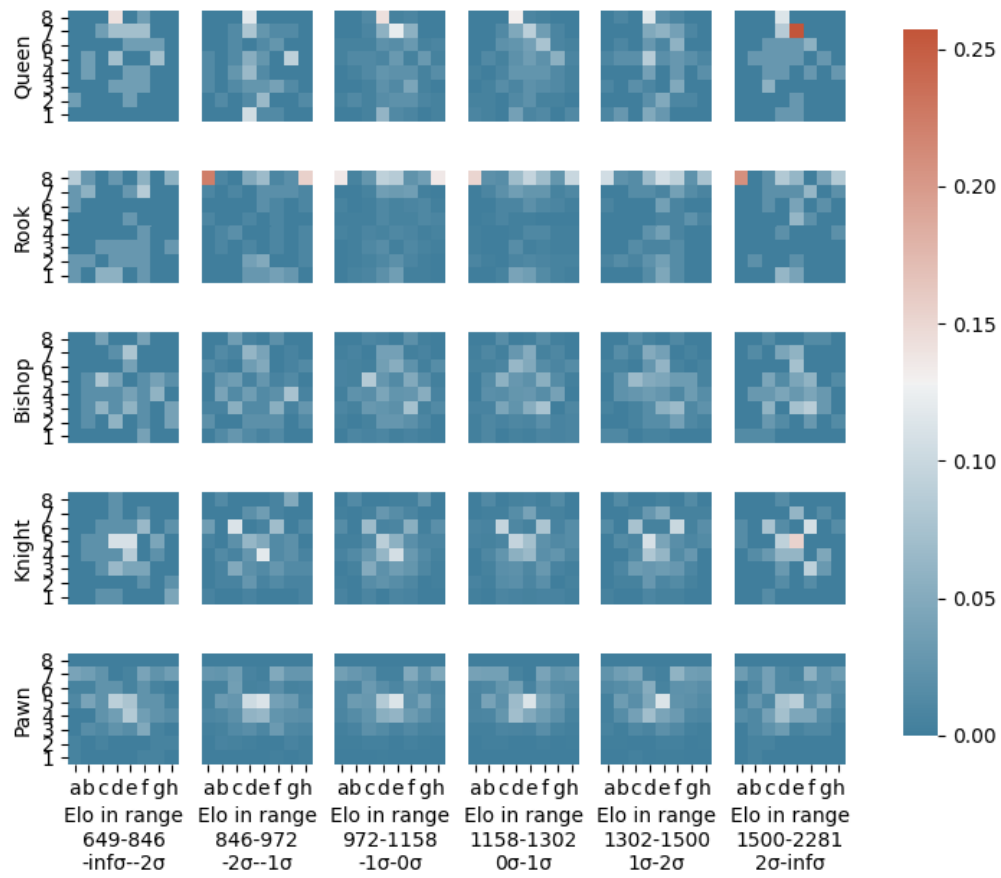
Black pieces captured through time.  
Move count vs pieces lost per game.



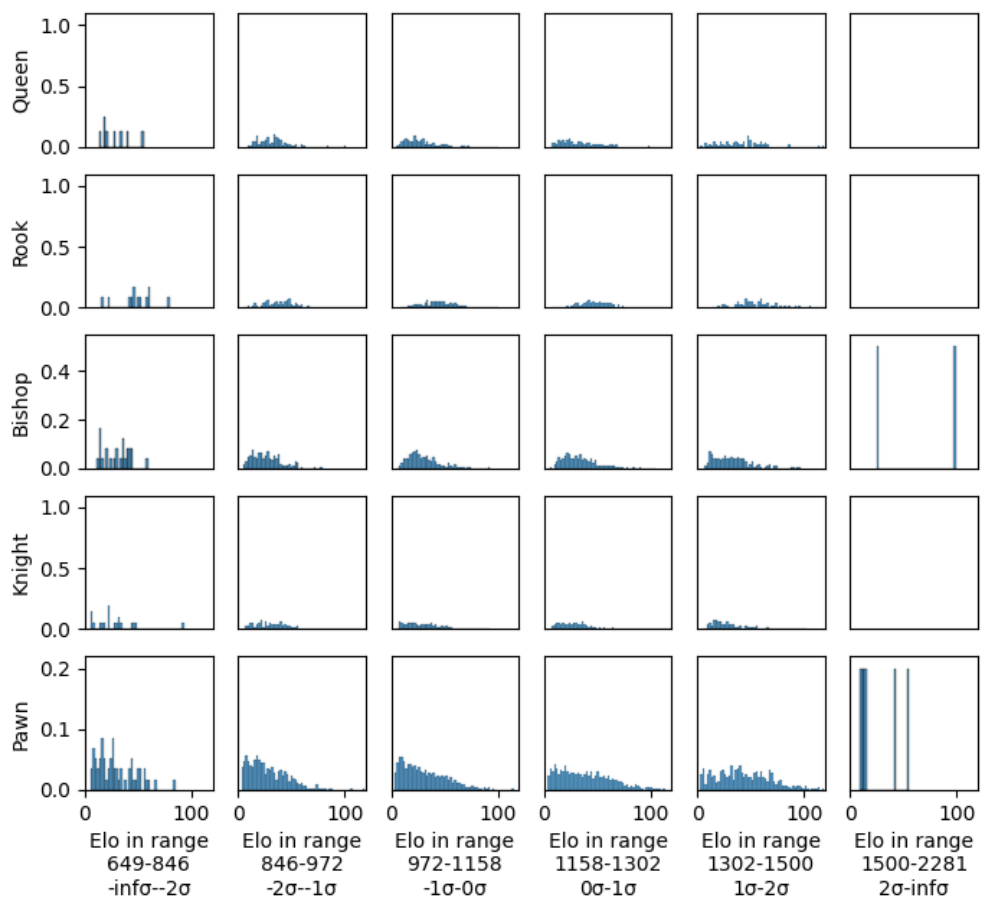
### 6.3 DaenaliaEvandruile



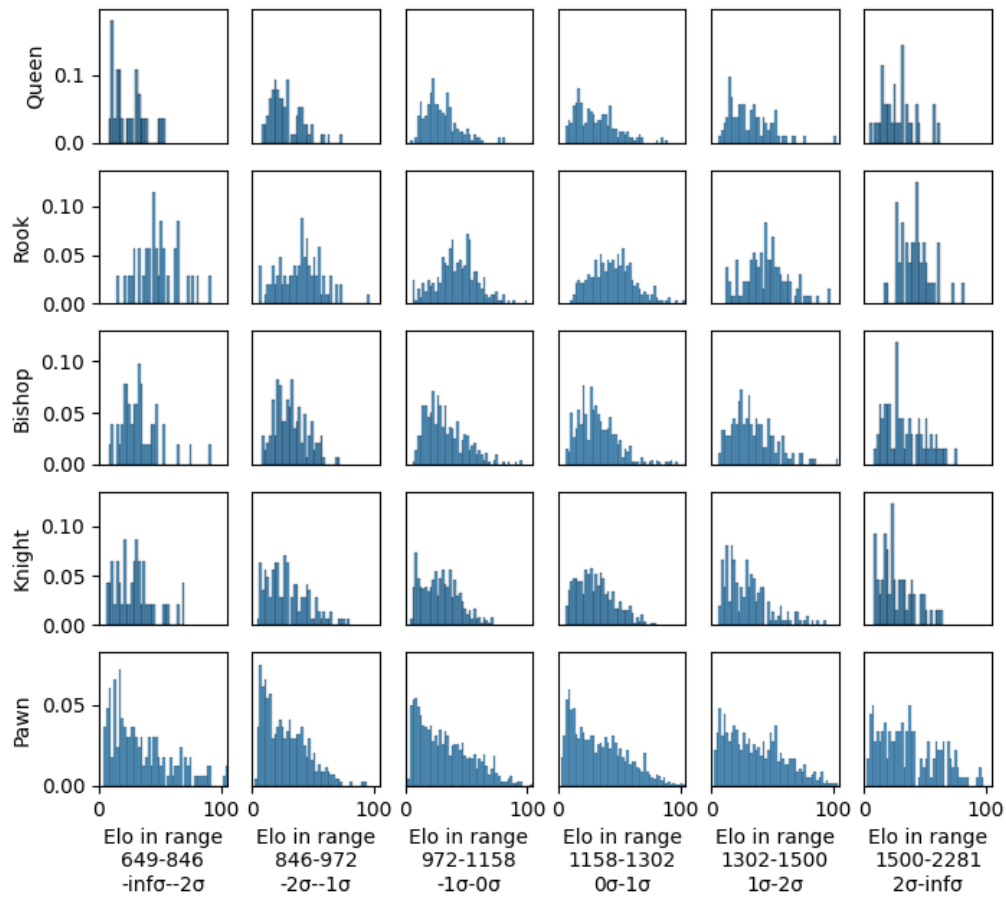
DaenaliaEvandruile's Black pieces captured.  
Proportion of piece lost on board.



White pieces captured through time.  
Move count vs pieces lost per game.



Black pieces captured through time.  
Move count vs pieces lost per game.



## 6.4 Individual piece plots

