Computer Analysis of World Chess Championship Players

Oscar Romero
Universitat Politecnica de Valencia
Valencia, Spain
E-mail: -oromero@dcom.upv.es

Lorena Parra
Universitat Politecnica de Valencia
Valencia, Spain
E-mail: -loparbo@doctor.upv.es

Jose Fernando Cuenca
Chess24
Madrid, España
E-mail: -Jose.fernando.cuenca@gmail.com

Jaime Lloret
Universitat Politecnica de Valencia
Valencia, Spain
E-mail: -jlloret@dcom.upv.es

***Abstract*— In some sports, it is difficult to know who has been the best winner of the world championship. In athletism, it is not so difficult because world records clearly state which is the best mark. Nevertheless, in the case of chess, it is challenging to know who has been the best world chess championship player. Nowadays, it is well known that many chess engines can beat the best chess players in the world, so we can use it for comparison purposes. In this paper, we use one of the best chess engines, Stockfish 10, in order to know which world chess championship player is the best of all time. We have compared their moves during the world championship with the ones suggested by the chess engine in each game. Results show how good each one of them was, compared with Stockfish 10, which player obtained the greatest percentage of best moves during their games, how the quality of their moves evolved during the games and the average percentage of best moves throughout the games.**

Keywords-Chess; Computer Analysis; World Chess Championship.

I. INTRODUCTION

Chess is a strategy board game that involves two players. It is one of the most popular strategy games played across the globe. Modern chess is based on the rules adopted in Spain in the 15th century [1]. It is played on a checkered board with 64 squares and includes 6 different types of pieces. Each player starts with 16 pieces and the player who has the white pieces starts the game by moving first. The number of game states that can be reached through a legal play was estimated to be around 10^{46} [2]. Because of its complexity, chess has been used as a testbed for most of the Artificial Intelligence (AI) systems.

It 1947, Alan Turing designed a program to play chess for the first time in history. Since 1950, different programs have been developed to play chess. Different strategies have been applied to improve their results. While in 1960 chess programs only can beat amateur players, in 1990 those programs have become powerful and can win chess masters [3].

In the last decades, chess players have evolved and improved using chess programs to practice and learn. The Elo is a method used to calculate the skills of a chess player. The best players can foresee the development of a game 10 to 15 moves to decide the best strategy [4]. The current world champion (since 2013), Magnus Carlsen, has an Elo rating of 2882 [5]. On the other hand, the current versions of the best chess programs have more than 3400 [6].

In 2006, Guid and Bratko used CRAFTY, a chess program, to evaluate the quality of chess players regardless of the game score [7]. They evaluate players of the World Chess Championships (WCCs). CRAFTY calculates the best move for each given position and compare the move that did the players with the best move and assign an average error to each player. Their results were strongly criticized because some of the best players as Fischer were placed as weaker than players who only won the WCC one year. On the other hand, their results were disputed as the engine used to calculate the best moves was considered weaker than most of the analyzed players.

In this paper, we are going to analyze the performance of all chess players in the WCC, like the work presented by M. Guid and I. Bratko in 2006, but using a stronger engine. Our hypothesis is that the results obtained in the past were skewed by the used engine. Nowadays, computers are more powerful and probably the best move calculated in the past will be something different than the best move calculated in this paper. Therefore, the average error for each player may be different. We use all the games of the WCC, from 1886 to 2018, and the chess engine is configured with a depth of 28. A total of 20 computers were used to calculate the average error of each player and some other parameters. Then, the results of the average error or each player are compared. Finally, our results will be compared with the results obtained in the past to evaluate the ranking of best chess players according to their average error change.

The remaining of this paper is structured as follows. Section 2 describes the related work. Section 3 details the material and methods utilized. The results are shown in Section 4. Finally, Section 5 presents the conclusion and future work.

II. RELATED WORK

In this section, we are going to summarize the papers which deal with the topic of chess, algorithms and AI.

Guid and Bratko compared the quality of different chess players of the WCC [7]. They used the CRAFTY program. CRAFTY evaluated the individual move realized by each player. They used the games played in the WCC from 1886 to 2006. CRAFTY used a depth of 12 moves in the analyses. The parameters used to compare the players were the average error, % of blunders, complexity expected error, % of best moves and difference between best moves. Their results showed that Capablanca was the best player and Fischer was the one with the highest difference between best moves.

Ribeiro et al. [4] in 2013 used CRAFTY (Elo rating of 2950) to evaluate the white player advantage move-by-move. They used 73,444 high-level chess matches available in Portable Game Notation (PGN) Mentor. CRAFTY calculated de advantage in terms of the number of remaining pieces and its placement. A positive value indicated that the white player has an advantage, while the negative value indicated that is the black player who has the advantage. Their results included the advantage, mean of advantage and variance of advantage along with the game in each match. They compare the data from different periods, 1857-1918, 1919-1949 and 1950-2011 to evaluate the changes in the chess players. Their results suggested that the opening stage of a match is becoming longer and pointed out that this might be related to a collective learning process.

On the other hand, many authors used chess to test and train AI systems. One example is the work proposed by Vázquez-Fernández et al. in 2011 [8]. They proposed a method for tuning the weights of the evaluation function of chess based on evolutionary programming. They used 10 different players and 6 training games. They used as “theoretical” values: 100 (pawn (fixed value)), 300 (knight), 330 (bishop), 500 (rook) and 900 (queen). As mobility, a weight of 10, and bounds of [0,300], was used. After 50 generations, the weight changed to 100 (pawn), 310.89 (knight), 325.32 (bishop), 514.92 (rook), 841.61 (queen) and 5.62 (mobility). Finally, their engine was tested playing 10 games with a human player. It is important to note that, in this paper, the Elo of the used engine was 1463. The Elo of the human who played with the engine was 1737.

One year later, Vázquez-Fernández et al. [3] presented how their engine reaches an Elo ranking of 2425 after readjusting the weights. Moreover, they used the Hooke-Jeeves algorithm [9] in order to pursue the adjustment of the weights according to the best virtual players.

In 2014, Vecek et al. [10] presented a comparison between different evolutionary algorithms. They proposed to use Chess Rating System for Evolutionary Algorithms (CRS4EAs) instead of the typical Null Hypothesis Significance Testing (NHST). They claimed that NHST was often misused and misinterpreted. The CRS4EAs was planned as a tournament in which the algorithms are the players and the solutions of algorithms as the game outcome. A total of 15 evolutionary algorithms were tested. The best evolutionary algorithm according to CRS4EAs was the jDE/rand/1/bin (it was the second according to NHST). According to the positions, 9 out of 15 obtained the same position with both methods. Their results revealed that CRS4EAs is comparable with NHST, but it is easier to use and it is less sensitive to outliers.

In recent years, machine learning techniques have been applied to strategy games, and then, trying to play better than static algorithms. Alphago Zero was developed to play go, a simple game, with simple rules, but with many possible moves. It is an artificial intelligence, based on deep learning and neuronal networks. After a deep training, it was able to win the best human player on year 2016. In chess, several artificial intelligences have been developed. One of the most important is Leela Chess Zero (LCZero), also based on neuronal networks. After an intense training, it became the champion in Top Chess Engine Champion (TCEC) season 15, in May 2019, where Stockfish, one of the best static algorithms, obtained the second position. Recently, Stockfish has been improved and it is the current TCEC champion, season 16, celebrated on October 2019. This time, LCZero was in second position. Probably both of them will be improved again, and perhaps they will get an ELO above of 4000 very soon.

III. Test Bench and Method

In this work, we have evaluated all players in all world chess championships, from 1886 to 2018. Today, there are chess engines that play clearly better than the best human players. Thus, using one of those chess engines we can evaluate a human player. For this analysis, we get the score of each move and compare it with the best move obtained by the engine to get the human player error. From the information provided by the chess engine, we can extract additional information such as if the human selected the first, second, etc. best move.

A. Test Bench

The chess engine selected for this study is Stockfish 64 bits Version 10, one of the strongest engines in early 2019. We have created a program, using the Universal Chess Interface (UCI) protocol to communicate with the chess engine. The feature of the computers used to perform the test was intel i5, 8th Gen, at 2.8 GHz. One of the most important parameters to configure the engine is the depth. We used a depth of 28 in order to meet the time requirements (2 hours for 40 minutes plus 1 hour every new 20 moves, or 90 minutes plus 30 seconds every move) for each game taking into account the computer features. The chess engine and computer features provide us a good rating, and clearly, this engine beats the Crafty engine used in [7]. We tested other depths, like 29 and 30, but the results were similar, although they needed very much more calculation time. Then, with a depth of 28 and the used hardware, the average move evaluation time was 3 minutes per move or 5 hours per game. There was also the option to set up a fixed time per move or game, but obviously this would provide different results in different computers, or even in the same computer, because the actual time used by the engine would depend on the load of the computer, that may vary for different reasons (status of Operative System, running processes, RAM memory, etc.).

When our program is analyzing a game, it evaluates the human player move, and it compares with the best move given by the engine. Then, the difference will be the error made by the human, if ever. Also, the program will tell us if the human-made the best, the second-best, etc. move.

In total, near 1000 games were analyzed, employing almost 5000 hours by the 20 computers with exactly the same features, for 10 days.

For the analysis, some situations have been taking into account. The first moves have been widely studied, and chess players usually play opening books. The evaluation of these first moves would not add too much information to this study. Thus, we decide to start the analysis from move number 7, that is, we have not evaluated the first 12 moves (6 from the white and 6 from the black).

Another important decision is when to stop the analysis of a game. There may be a moment that the evaluation of the position is high (either positive por white pieces or negative por black pieces) before the end of the game. Under this situation, the white player may not play the best move, for example, because it may take a lot of time to find it, and it may be enough with a weaker move, fast to calculate, and good enough to keep the big advantage. We selected the limit of 2 white pieces (-2 for black pieces). On the other hand, if white player is losing, even for more than 2 (white position evaluation lower than -2) we will continue analysing the game, because the player will try to play the best move, trying not to lose the game (the same for black pieces with a score of 2 or more).

In other studies like [7], complexity had been analyzed. But in this paper, we are focused on the evaluation of each player, regardless of complexity, present material, or type of opening, for example. To get good results in all these situations is part of the skill of the player. In this way, blunders are also evaluated, when it may cause draw or lose a game. Not to make blunders is also part of the skill of the player.

IV. RESULTS

In this section, we have analyzed the average error for each player (compared with the move suggested by Stockfish version 10), the percentage of best moves performed during the game, the evolution of each player along the games (from past to present), for those that have played more than one world chess championship, and the average of best moves vs number of moves.

A. Average error

The average error is the difference between the evaluation of the human player and the best move provided by the chess engine. The formula used is

$$\text{Mean error} = \frac{\sum|\text{best move evaluation} - \text{player move evaluation}|}{\text{number of moves}} \qquad (1)$$

Figure 1 shows the average errors for all players in all world championships. As expected from the results in the last years, Magnus Carlsen is the top one chess player. Other present players like Caruana and Karjakin show good performance. Famous players like Kasparov, Karpov, and Fischer, although not in the top five, show good results, with an error average lower than 0.11.

Figure 2 shows the comparison of the results from this study and results from [6], thus, showing the world champions up to 2006. Results are similar, but the study of 2006 is giving in general higher average error, the mean average error of players was 0.143 according to the results of 2006 and 0.14 according to results of 2019.

It is interesting to discuss the evaluations for very important players like Kasparov, Fischer, where the difference between both studies is almost 30%, and also Karpov, with a difference of 16%. According to this study, that is more precise than the previous one, these three players have a good rating, as they have shown in their tournaments. On the other hand, players like Lasker and Smyslov had an over-evaluation in the past, and the new study shows that they have actually a lower play strength.

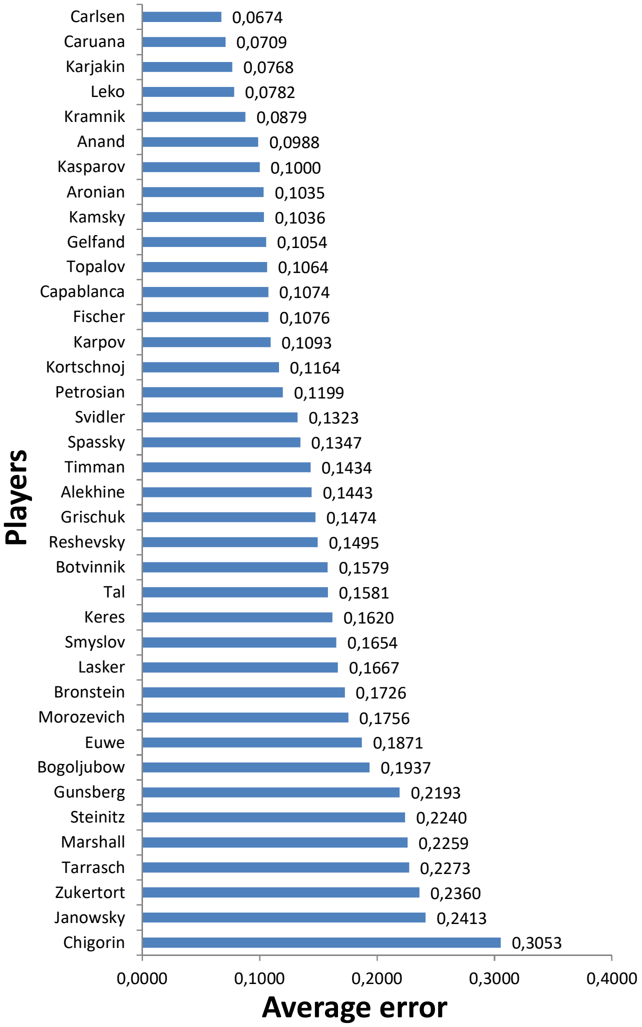

Figure 1. Comparison of players in terms of average error.

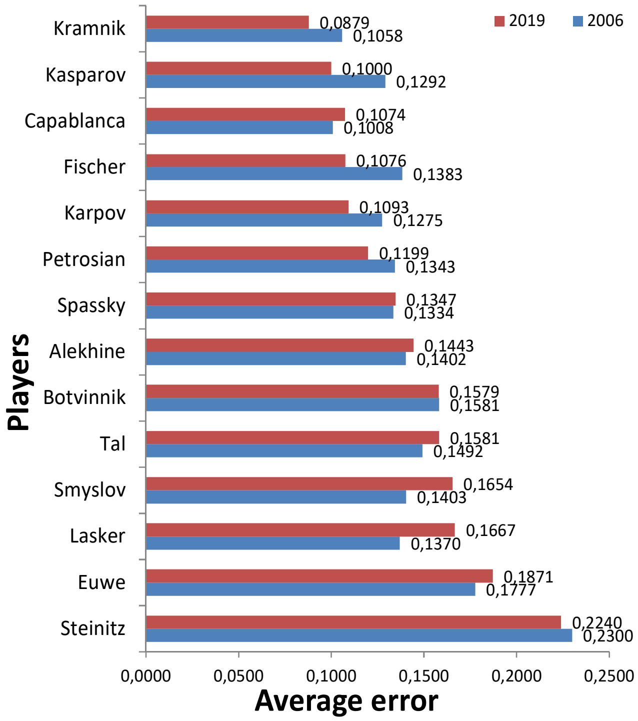

Figure 2. Comparison of average errors calculated in 2006 [7] and 2019.

B. *Player evolution along games over the time*

The quality of a player may change with time. In this subsection, we show the evolution of some players along the time, showing how the average error changes with the games played along the time. In this section, we have analyzed only those players who have played the world chess championship more than one time in order to have enough games for this analysis.

Figure 3 shows the average error evolution of some players along with their participation in world championships. The figure is divided into different periods, a) presents the data from WCC champions between 2000 and 2018 with the 3 champions of this period (Carlsen, Anand, and Kramnik). Note that even that the first played WCC by Anand was in 1996, he is included in this graphic. Carlsen, see Figure 3 a), although starting with a good average error in his first WCC, shows a general improvement along with his participation in four WCCs. Kramik and Anand had their best result in the second played WCC. The most relevant issue is that all of them have average errors lower than 0.105 between 2000 and 2018 (the average error of 0.12 of Anand corresponds to the WCC in 1996), and the variations in their average errors are minimum (lower than 0.025).

Figure 3 b) represents the average errors of players between 1980 to 1996 with the 4 champions of this period (Kasparov, Karpov, Spassky, and Botvínnik). There were other WCC champions in this period who only played one WCC and are not included. Kasparov had the best performance in his first WCC, Karpov, and Spassky had their best results in the second year and Botvinnik in ninth WCC. During this period, the average error of the players, 0.12, is higher than in the current period and their variations of the average error were higher than nowadays. Botvinnik is the player who had higher variations, his worst results were found in his fourth WCC.

The last period is represented in Figure 3 c) and corresponds to the campions of the oldest WCCs (1886 to 1946). During this period, the players had even higher variations than in the previous periods. The best results of Alekhine, Capablanca, Lasker, and Steintz were in their first, second seventh and fifth WCC respectively.

Apparently, there is no general trend that confirms that the more time a player plays in the WCC, the better he plays chess.

C. *Percentage of "best moves"*

Figure 4 shows the percentage of the best move selected by the players. Famous world champions like Kasparov, Carlsen, Karpov, and Fischer have similar rate choosing the best move, but none of them are in the top five. Other famous players from the past, like Capablanca, Lasker or Steinitz have 50% or less of rating for selecting the best move.

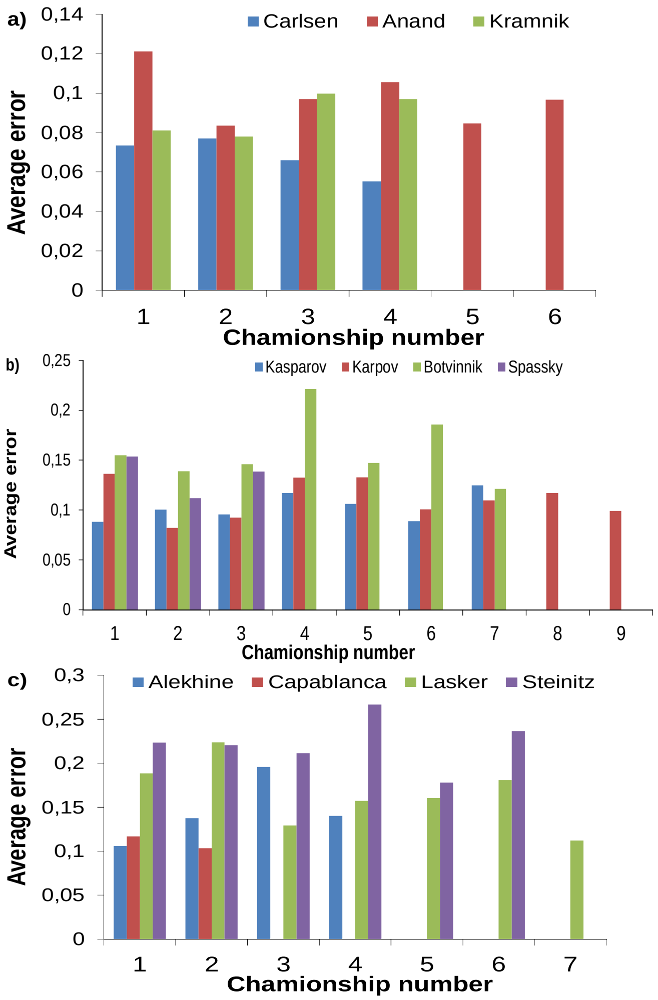

Figure 3. Player evolution by championships.

ISBN: 978-1-61208-752-8

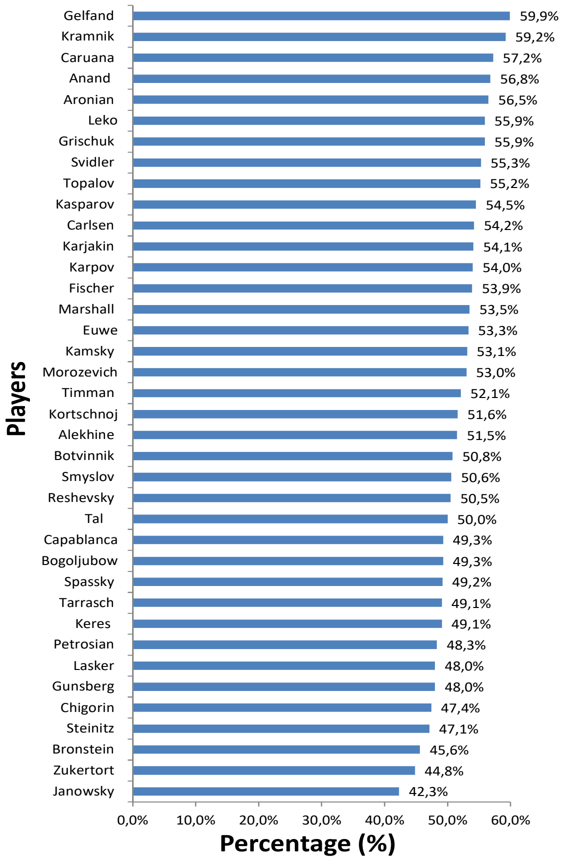

Figure 4. Percentage of best move of different players.

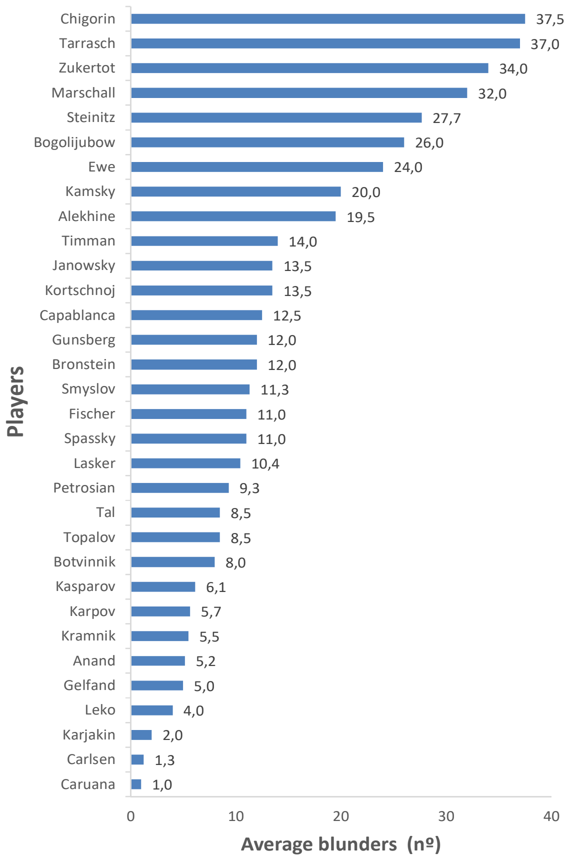

Figure 5. Comparison of average blunders played.

Carlsen, who is the current world champion, won 4 WCC and achieved the highest Elo in history. He has only 54.2% of best moves. He is placed in the 11th position behind some players who only played one WCC and never won the championship as Gelfand, Leko, or Aronian. Nonetheless, it is important to note that sometimes, the difference between the first and the second-best move are small and has almost no impact on the results of the game.

D. *Average of blunders*

The number of average blunders played by each one of the analyzed players is presented in Figure 4. For this analysis, we consider the WCC with only two players. A move is considered as a blunder when causes that white player has a score lower than 2 or black player has a score higher than -2.

While in the previous section, we saw how very famous players, such as Carlsen, Karpov or Kasparov, appeared in relative bad positions; when we analyze the data of blunders the results change, see Figure 5. The most recent WCC champions: Carlsen, Anand, and Kramnik, appear in the top seven positions. In their WCCs all of them played few blunders: 1.3, 5.1 and 5.5 blunders as average in all WCC respectively. It is relevant that Caruana is the player with the lowest average. Nonetheless, in the specific WCC that Caruana and Carlsen played, Carlsen had not played any blunder.

Some of the most famous champions, such as Kasparov and Karpov appears in the 9th and 10th position, with 5.5 and 6.1 average blunders per WCC. Meanwhile, other famous players from the past, like Alekhine, Capablanca, Lasker or Steinitz are in worse positions 24, 20, 12, and 28, with more than 10 average blunders in each WCC.

E. *Average of "best move" vs "number of move"*

Finally, we are going to evaluate the performance in terms of best move versus played move of different WCC players along with the game. Since in other section we pointed that the players were obtaining better performances in the last evaluated period of the history of WCC, 2000 to 2018, we are going to compare the players with best performances in this period: Kramik, Anand, Carlsen, and Caruana. As we may expect due to the time control at move 40, many players make mistakes in moves near move number 40, for example, in moves from 35 to 39. Then, along next moves, the quality of moves increases, but only for a short number of moves. Figure 6 shows how some players make mistakes again starting around move number 45. Then, we will compare the performance from move number 7 to move number 45, and from move 7 until the end.

The situation along a game, like beginning moves, ending, time left, may affect the capacity of the player to select the next move. Figure 6 shows if the players have selected the best move, second, etc, depending on the number of the move. Carlsen has a good rating, with an average around of second-best move, an average of 2.15 with standard deviation of 0.36 for moves 7 to 45. Only for very long games, he may make a bad selection move, such as in moves near number 60 or more than 70. Kramnik, also shows good rating, around second best move, average of 1.96 with standard deviation of 0.40 for moves 7 to 45 although not reaching long games. Anand has an average of 2.11 with standard deviation of 0.39 for moves 7 to 45. His results are similar to the observed results with Carlsen. On the other side, Caruana (who only played one WCC and did not win) is the one who has the highest variability in terms of best move versus played move, average of 2.09 with a standard deviation of 0.84 for moves 7 to 45. Thus, the player who played the best moves more times is Kramnik followed by Anand, Carlsen, and Caruana.

If we consider the whole game, moves 7 until the end, the results slightly change and the order of players who had the best average of best move versus played move is not maintained. The new order is Kramnik, Anand, Caruana, and Carlsen. The averages changed and, in most cases, the averages increased, which means that after long games the players selected worse moves. However, Kramnik had better average when we consider the entire game, 1.88. Thus, we can affirm that Kramnik had better performances in the endgame than the other players.

V. Conclusion and Future Work

In this paper, we have compared the performance of WCC players with is Stockfish 64 bits Version. Our study was based on the presented results in 2006 [7] and we have compared our findings with their outputs.

When we compared the average error of different players, the best player was Magnus Carlsen, the current WCC champion. Nonetheless, some famous champions such as Kasparov and Karpov were not in relevant positions attending to these parameters. Other parameters such as the percentage of best move and average of blunders were used to compare players. While percentage of best moves gave similar punctuations to players with different quality and Elo, the ranking by average of blunders offered more accurate results. Finally, we have evaluated the average of best move versus played move of best players from 2000 to today.

As future work, we will increase the complexity of our analysis including other factors as stages. We will also include statistical analysis in order to add more value to our results. Moreover, we will analyze the performance of the players based on their or her physical condition.

References

[1] J. L. Cazaux and R. Knowlton, "A World of Chess: Its Development and Variations Through Centuries and Civilizations", McFarland & Company, 2017.

[2] V. Janko and M. Guid, "A program for Progressive chess," Theoretical Computer Science, 644, 2016, pp. 76-91.

[3] E. Vázquez-Fernández, C. A. C. Coello, and F. D. S. Troncoso, "Assessing the positional values of chess pieces by tuning neural networks' weights with an evolutionary algorithm;" In 2012 World Automation Congress (WAC 2012), 24-28 June 2012, Puerto Vallarta, Mexico, pp. 1-6.

[4] H. V. Ribeiro, R. S. Mendes, E. K. Lenzi, M. del Castillo-Mussot, and L. A. Amaral, "Move-by-move dynamics of the advantage in chess matches reveals population-level learning of the game," PLoS One, 8(1), 2013, pp 1-7.

[5] Fide Ranking. Available at: https://ratings.fide.com/card.phtml?event=1503014. Last access 11/10/2019

[6] Ranking. of chess programs. Available at: https://ccrl.chessdom.com/ccrl/4040/. Last access 11/10/2019

[7] M. Guid and I. Bratko, "Computer analysis of world chess champions," ICGA Journal, 29(2), 2006, pp.65-73.

[8] E. Vázquez-Fernández, C. A. C. Coello, and F. D. S. Troncoso, "An evolutionary algorithm for tuning a chess evaluation function". IEEE Congress of Evolutionary Computation (CEC 2011), 5-8 June 2011, New Orleans, USA pp. 842-848.

[9] O. Tolga Altinoz, A. Egemen Yilmaz, Multiobjective Hooke–Jeeves algorithm with a stochastic Newton–Raphson-like step-size method, Expert Systems with Applications, Volume 117, 1 March 2019, pp. 166-175.

[10] N. Veček, M. Mernik, and M. Črepinšek. "A chess rating system for evolutionary algorithms: A new method for the comparison and ranking of evolutionary algorithms" Information Sciences, 277, 2014, pp. 656-679.

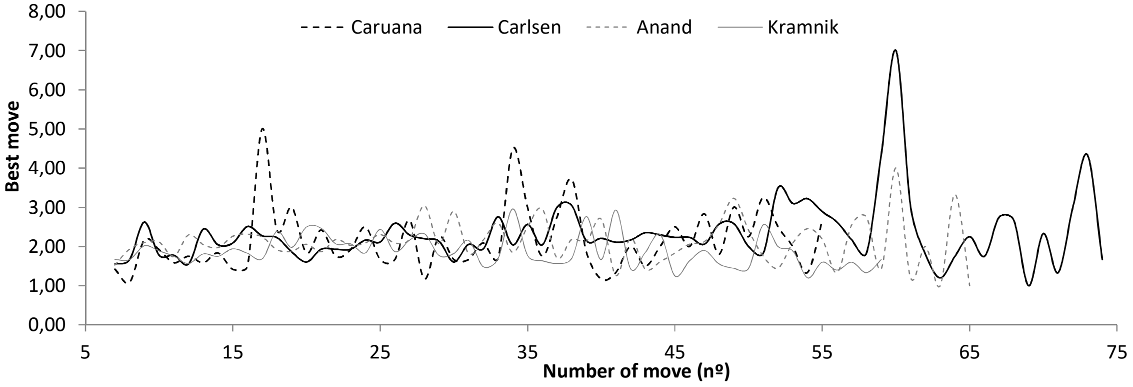

Figure 6. Comparison of the best move versus the played move.