UNIVERSITATEA BABEȘ-BOLYAI
BABEȘ-BOLYAI TUDOMÁNYEGYETEM
BABEȘ-BOLYAI UNIVERSITÄT
BABEȘ-BOLYAI UNIVERSITY

FACULTATEA DE FIZICĂ
Str. Mihail Kogălniceanu nr.1
Cluj-Napoca, RO-400084
Tel: +4(0)264-405300 | FAX: +4(0)264-591906
secretariat.phys@ubbcluj.ro
www.phys.ubbcluj.ro

UNIVERSITATEA "BABEȘ-BOLYAI" CLUJ NAPOCA
FACULTATEA DE FIZICĂ
SPECIALIZAREA FIZICĂ COMPUTAȚIONALĂ

LUCRARE DE DISERTAȚIE

Coordonator științific

Prof. dr. Néda Zoltán
Conf. dr. Járai-Szabó Ferenc
Lect. dr. Borbély Sándor

Absolvent

Bănică-Solymosi Írisz

2025

UNIVERSITATEA BABEȘ-BOLYAI
BABEȘ-BOLYAI TUDOMÁNYEGYETEM
BABEȘ-BOLYAI UNIVERSITÄT
BABEȘ-BOLYAI UNIVERSITY

FACULTATEA DE FIZICĂ
Str. Mihail Kogălniceanu nr.1
Cluj-Napoca, RO-400084
Tel: +4(0)264-405300 | FAX: +4(0)264-591906
secretariat.phys@ubbcluj.ro
www.phys.ubbcluj.ro

UNIVERSITATEA "BABEȘ-BOLYAI" CLUJ NAPOCA
FACULTATEA DE FIZICĂ
SPECIALIZAREA FIZICĂ COMPUTAȚIONALĂ

LUCRARE DE DISERTAȚIE

FOOTBALL MATCH SIMULATION BASED ON TRANSITION PROBABILITY MATRICES

Coordonator științific

Prof. dr. Néda Zoltán
Conf. dr. Járai-Szabó Ferenc
Lect. dr. Borbély Sándor

Absolvent

Bănică-Solymosi Írisz

2025

Table of Contents

Abstract

Analyzing data collected from sport events helped researchers to a better understanding of the game. In this work we intend to reproduce a football match with a simulation built as consecutive ball passes. To determine the path, we construct transition matrices to store, on one hand, the successful events of passing a ball and, on the other hand, those cases when there is an interception or a game interruption. Based on these probabilities the succession of the passes can be determined.

We display the results as a plot of the probability distribution of the ball's positions. We compare the obtained distribution with the one that is based on the original data, focusing on both the similarities and differences between them. We illustrate these distributions for data coming from successively simulated matches, too. Moreover, entropy values are also calculated for different number of consecutive game simulations.

Introduction

Curiosity is a vital part of human nature so it comes as no surprise that there is a vastness of researches in the domain that is as essential in everyday life as sport. Although these focus on all types of sporting activities, there is a predominance in those that are related to widespread sports as basketball or football ([1]). In the present work we aim to build a simulation of a football match which will be composed of successive ball passes, whose directions will be decided according to certain probabilities. These will be obtained from transition matrices that are computed from coordinates of the movement in the original data.

We start our paper with enumerating some previous studies that have been conducted in this field. These have a wide variety of themes, since some focus on, for example, temporal data ([2]), others also take into account spatial coordinates ([3]). In addition to these, possession analysis, trajectories, ball passes ([4]) are other factors that are examined to determine patterns, characterize dynamics and make predictions ([5], [6], [7]).

We then move our focus on the analyzed matches. After scaling the data to match the football field's sizes, we plot the probability distribution of the ball's position. Next, we filter the data, firstly eliminating those points when the ball was out of play, and after that reducing the size of the data set by keeping only the starting and end points of each pass. From the newly obtained dataset we calculate the transition matrices between the cells of the discretized football field. We store separately the events of successful passes and the lost balls (which can be due to an interception, a foul, the ball going out etc.). With the probabilities obtained from these matrices we build a match simulation, which is a series of consecutive ball passes.

In the next section our results are presented. We plot the probability distribution for the positions that we have from the simulation, and compare it to the distribution from the original, filtered data. We also plot distributions in case of data that is obtained from multiple match simulations. We calculate the relative difference between these probabilities (probabilities from the original, filtered distribution and from the case of the simulated match). Furthermore, using the Shannon-entropy formula, we compute the entropy values for both distributions.

1. Theoretical introduction

Sport plays a central role in both individual and societal contexts. Clearly, one advantage of performing a sporting activity is its positive effects on human health, yet we should also mention the benefits it brings to the social aspects of life. Not only those are involved in it who perform a given sport, but the supporters, people interested in following it as well. Thus, sport can move masses and connect people from all over the world.

However, when we take the competitive side of sport, achieving the best possible result is one of, if not, the most important goals of those who are engaged in it. This leads to a constant search for improvements, which can be achieved, on one hand, through previous experiences, and on the other hand, through experiments and researches. As technology advanced and new methods for data collection were utilized, more and more studies were conducted in the field of sport.

Sport science, especially physics, mathematics and informatics, aims to provide objective, quantitative ways to analyze games, to understand different situations, mechanisms. These are based on statistical methods, the laws of physics, and thus ensure a proper mean for studying sport events. If we consider, for example, the ball games, data can be collected regarding the players' and ball's position. Gudmundsson and Horton [1] review some research efforts which were based on spatio-temporal data. A significant part of research data is from basketball and football [1], which is unsurprising, since these two sports are among the most widespread around the globe.

Regarding football, the amount of data, which can be analyzed, is overwhelmingly high. Bialkowski et al. [5] focus on determining the formation of a team during matches and the role changes of players'. These can serve as a tool to characterize the style in which a team plays. Mendes et al. [2] concentrate on temporal sequences, times elapsed between ball touches to describe the dynamics of the games. Kijima et al. [8] analyze the time evolution for the players' and ball's positions to see whether it has self-similarities. Chacoma et al. [4] use a model composed of two teammates and one player from the opposing team. The former two can pass between them. Their aim with the simulations is to analyze the dynamics of ball possessions. With his paper, Peralta Alguacil [6] emphasizes the importance of the collective motion of the players. His model takes into account for example the probabilities of passes and pitch control. Stock et al. [7] implement a model in order to make predictions for the outcome of a tournament. This is based on the advantage that home soil has on teams when determining their potential. The enumeration is by no means exhaustive but it depicts the vastness of studies that were done in this field.

2. Methods

2.1. Analyzed matches

The match (referenced as Game), which we started to analyze, was played between two teams from the german first division, the Bundesliga. Data was collected with the resolution of 25 FPS (frames per second) and contains the following: the time frame, the x and y coordinates of the ball, which team has possession at each recorded time frame and the status of the ball (whether it is in play or out of play).

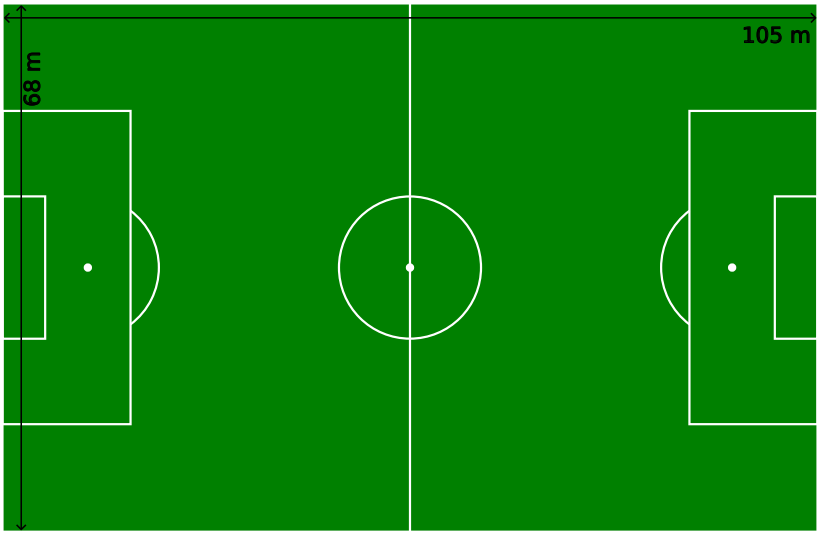

Figure 2.1: A standard football pitch with its respective line markings.

The playing area is of 105×68 metres (*Figure 2.1*), therefore the first step was to rescale the data according to these extents, since previously it corresponded to an 111×88 metres sized ground. After the transformation, on the x axis the data ranges from -52.5 to 52.5 and on the y axis it takes its values from the $[-34, 34]$ interval, with the $(0,0)$ origin point being in the middle of the field, at the kick-off point.

Initially we plotted the probability distribution of the ball's position in the pitch, using the whole dataset, as can be seen in *Figure 2.2 (a)*. Since those points that are out of play (meaning that the ball left the pitch boundaries or the game was stopped) may skew the distribution, we also plotted the probabilities for the in play coordinates (i.e when the ball was not out of play) (*Figure 2.2 (b)*). As we wanted to build a simulation based on passes between players, we restricted our analysis to these entries for the remainder of the study (when the ball is in play).

In addition to this football game, we analyzed two other matches, whose data is accessible on GitHub ([9]). The teams and players are anonymized. These datasets have a different format as opposed to the first match that we analyzed. Firstly, the coordinate values vary from 0 to 1 on both axes, with $(0,0)$ being in the top left corner and $(1,1)$ in the bottom right one. Secondly, there are events that are being recorded such as *pass, ball out, ball lost, shot, set piece* etc specifying their subtypes as well (for example *interception, throw in, on target-goal*).

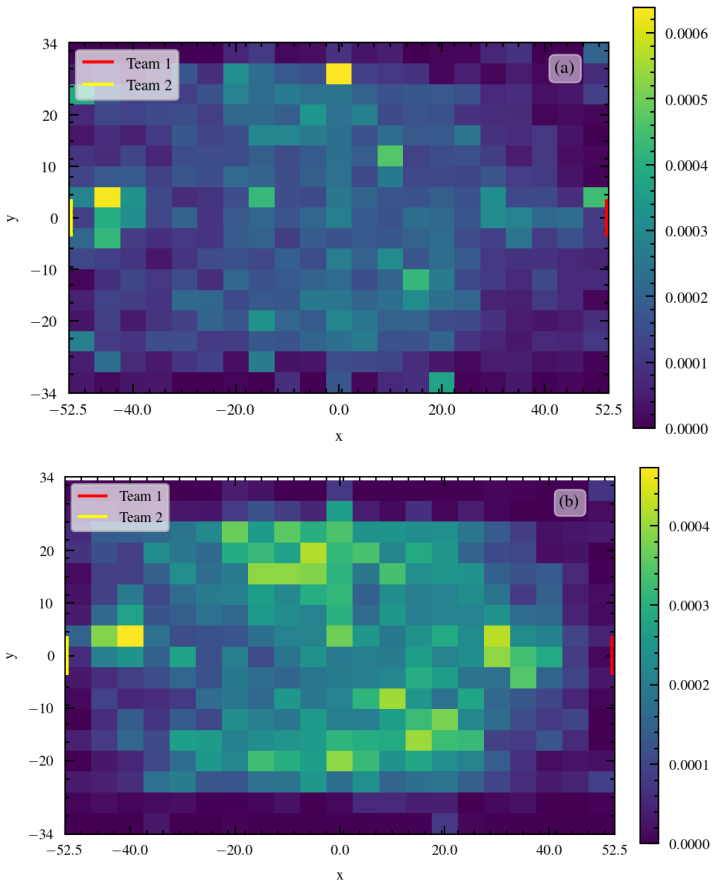

Figure 2.2: Probability distribution for the ball's position on the field. The bin size is 5×4 *m*. In (**a**) the whole dataset was used for the plot, whilst in (**b**) only those cases when the ball was in play.

The players involved in the respective events are listed, too. We will refer to these games as SampleGame1 and SampleGame2.

2.2. Data filtering

Our objective was to build a simulation of a match as a succession of ball passes, where the probability of each ball pass was extracted from the original experimental data set. In order to obtain these transition probabilities, in case of the Game, we aimed to extract discrete passing events by identifying directional discontinuities in the raw ball trajectory data. Our assumption was that an interception, whether it reached a teammate or an opponent, is usually indicated by a directional change in the ball's movement. Then, based on this angle difference between the initial and final direction of the ball, we would only keep those points that have a greater deviation than the fixed cut-off angle.

First we calculated how many events remain after the exclusion for a given cut-off angle. *Figure 2.3* shows the resulting plot for this step. There is a greater decrease in the number of events for the first part of the plot and then this decrease becomes less pronounced for larger angles. In addition to this, we plotted a segment from the match with three different cut-off angles to illustrate the influence of it, as shown in *Figure 2.4*. Upon their inspection, it can be

observed, that choosing a too large value will concatenate consecutive ball passes. After this we decided that we would continue to work with the $i = 15^\circ$ angle and filtered the data according to this, separately for both teams.

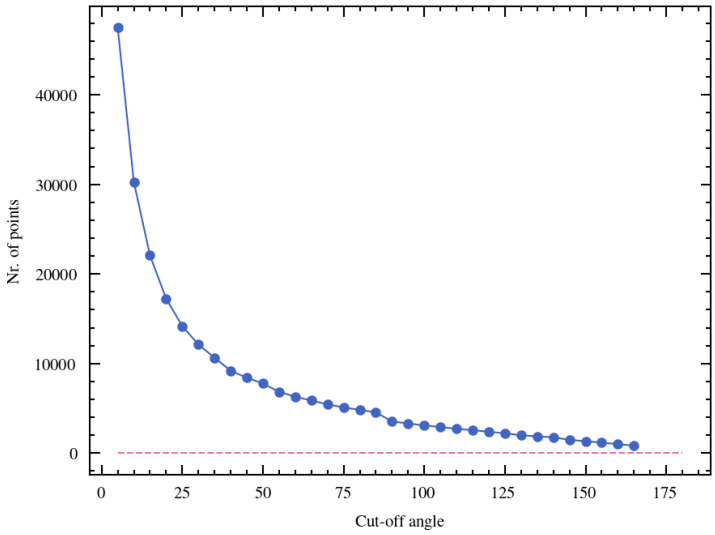

Figure 2.3: Number of remaining points after the data filtering for the different cut-off angles.

2.3. Transition matrix and match simulation

To build the simulation, we required the transition matrices for the successful passes and for the lost balls. These were needed for the two teams separately. A pass is considered successful if it goes from one teammate to another, a ball is considered lost if a player from the opposing team intercepts it or the game is paused (ex. foul, the ball goes out).

We divided the pitch into 30 equal sized rectangles of 17.5×13.6 metres (*Figure 2.5*). There is a matrix for every rectangle where we count the passes that started from this area and ended at another rectangle. The same is true for the case of lost balls. After each event (pass or loss of possession) is accounted for, we normalized each matrix with the sum of the two cases (passes + lost balls) for that respective rectangle. Since storing these two transition probabilities, for a given domain, in one array would help in the building of the simulation, these matrices were converted, for every rectangle, into a one-dimensional list. After this, they were concatenated (the first part of the new list from the passes and the second part from the lost balls). The resulting arrays then are held in a single two-dimensional matrix, where every row corresponds to a rectangle. Moreover, for every rectangle the cumulative sum of that given row is stored. As a consequence each team has their own two-dimensional array which holds the probabilities of passes and ball losses.

To decide which team in which direction is attacking, we plotted the passes that are originating from the rectangles that enclose the goals. We executed this for both teams separately (*Figure 2.6*). This was needed, first of all, for the analysis of the simulation results. On the other hand, because there are an even number of rectangles on the horizontal axis, when we started the simulation we needed to know from which rectangle to start the play, depending on which

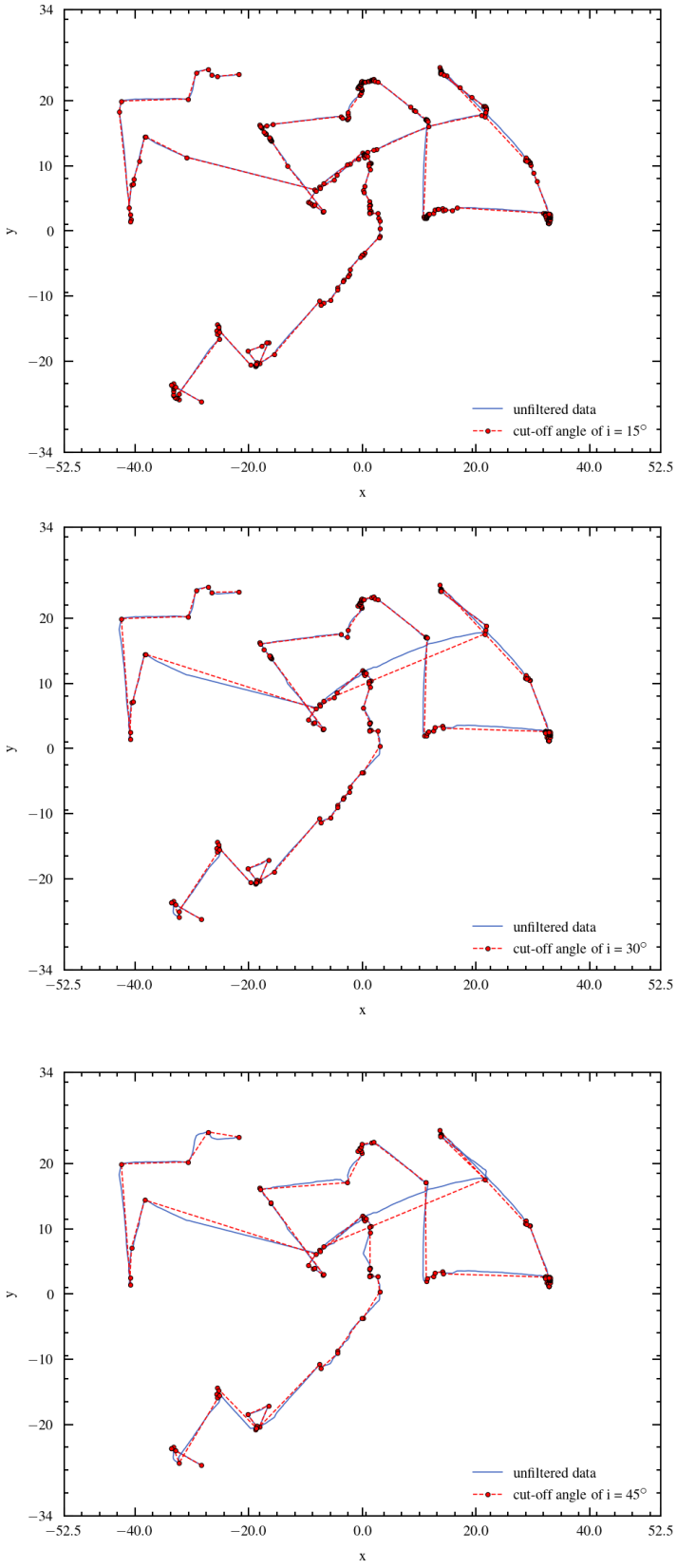

Figure 2.4: A segment of the ball's trajectory and the sorted data points for three different cut-off angles ($i = 15°, 30°, 45°$). It can be seen that for the cases of $i = 30°$ (**middle**) and $i = 45°$ (**bottom**) cut-off angles the newly obtained passes are not entirely accurate since some consecutive passes from the original data are now omitted.

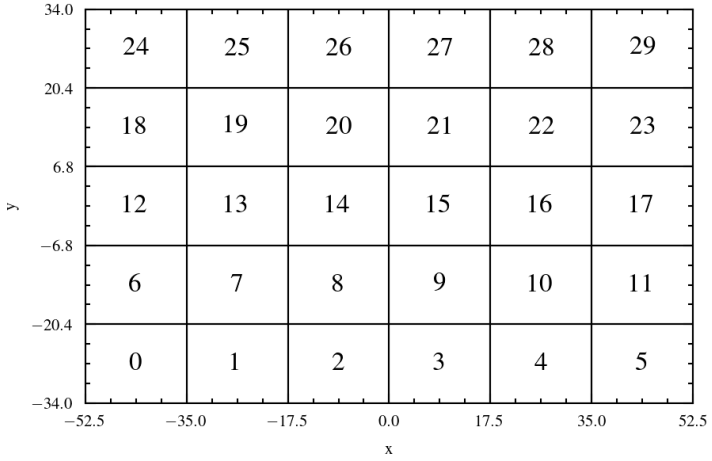

Figure 2.5: The football field is divided into 30 equal-sized rectangles. Each of these rectangle has an index number as illustrated here.

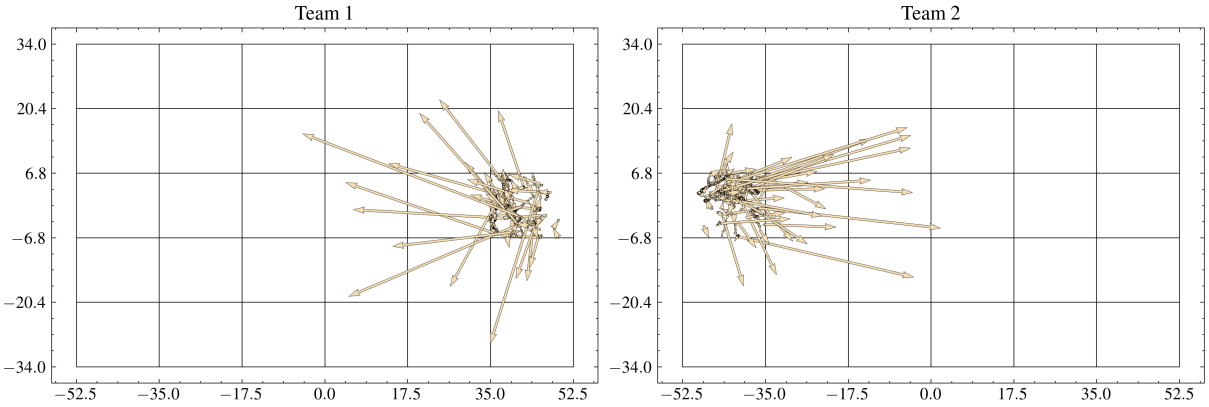

Figure 2.6: Determining the attacking directions for the two teams, separately. In the **left** part it can be seen that the team labeled with the number '1' is moving the ball away from this rectangles (index 17). This implies that the corresponding area is likely their own goal zone. On the **right**-hand side of the image we can see the same for Team 2 in the case of the rectangle with index number 12.

team has possession of the ball. In this case, since Team 1 is attacking from right to left, they started the match from the rectangle with index number 15.

The simulation is initialized from one of the central cells, then we determine where the next pass goes. For this a uniform random number is generated from the interval [0,1) and the index of that value, from the cumulative-summed array, for which this random number is smaller is then determined. For a given rectangle the transition probability list is constructed as: indeces 0-29 refer to successful passing probabilities, while indeces 30-59 refer to the case of lost balls. So if the index, whose corresponding value was greater than the random value, is less then 30, it means that the pass was successful. Then we move to the rectangle corresponding to the index identified above and we continue with the next pass. However, if this index is greater or equal with 30, it means that there was an interception or the game was stopped. This event is treated as a change in possession and only after this is the process of passing resumed (from the rectangle, whose index can be calculated as the array-index minus 30).

The model assumes that all interruptions result in a turnover, thus excluding cases like fouls

where possession is retained. Here after every lost ball there is a change in possession. In addition to this, there is no goal detection built into the simulation. So there is no such case where there is a goal and after that the game is resumed from the kick-off point.

In the case of the other two games from the GitHub repository, the raw data is composed of the events recorded during these plays. Therefore, the transition matrices were constructed using these and the simulation steps are the same as described previously. Since there were goals scored in these matches, from these event coordinates we could deduce the teams' attacking directions.

3. Results

3.1. Distribution

We plotted the probability distribution for the ball's position for the initial data and for the simulation results to see if there is a similarity between them.

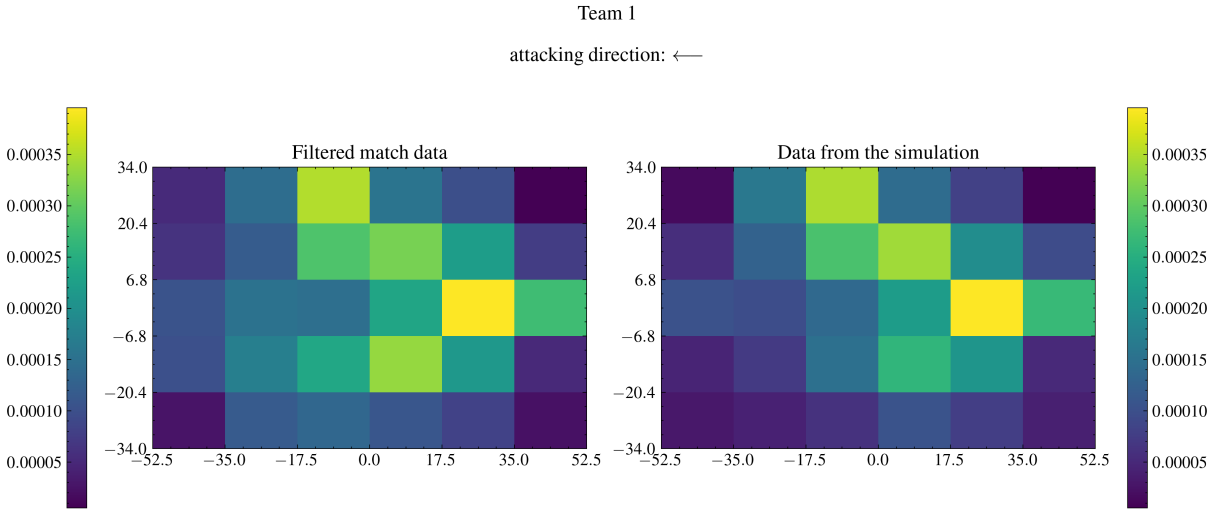

Figure 3.1: Probability distribution for Team 1 in case of one single simulated match.

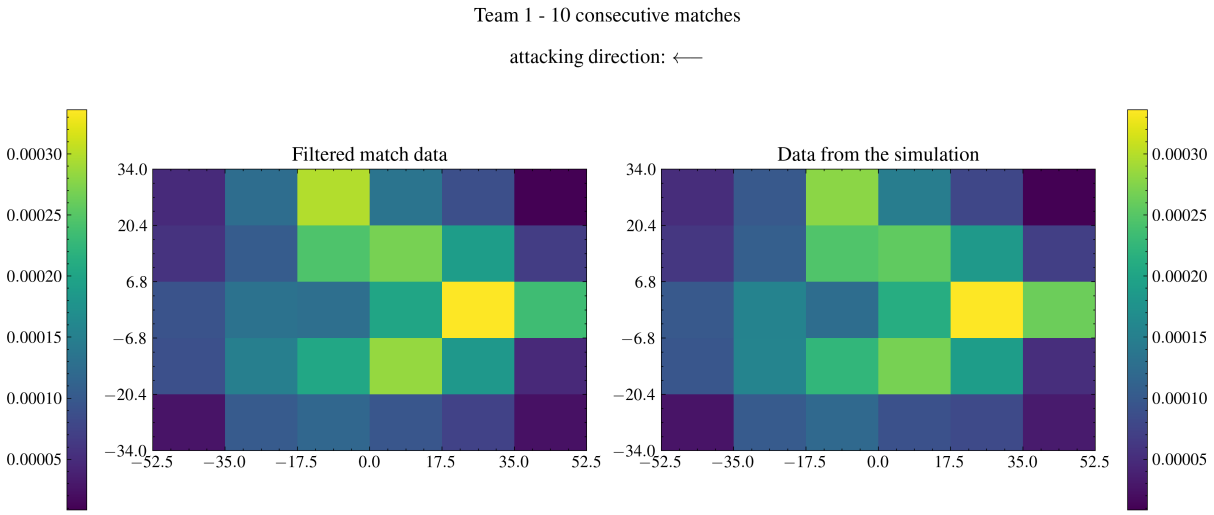

Figure 3.2: Probability distribution for Team 1 in case of 10 consecutively simulated matches.

As shown in *Figure 3.1*, after one simulated match there is a similarity in the distribution. *Figure 3.2* illustrates the case, when we calculated the probabilities for data from 10 match simulations. This resulted in an even better agreement with the distribution from the filtered data. We can see that Team 1 was predominantly playing on the right side of the pitch, leading their attacks through this part.

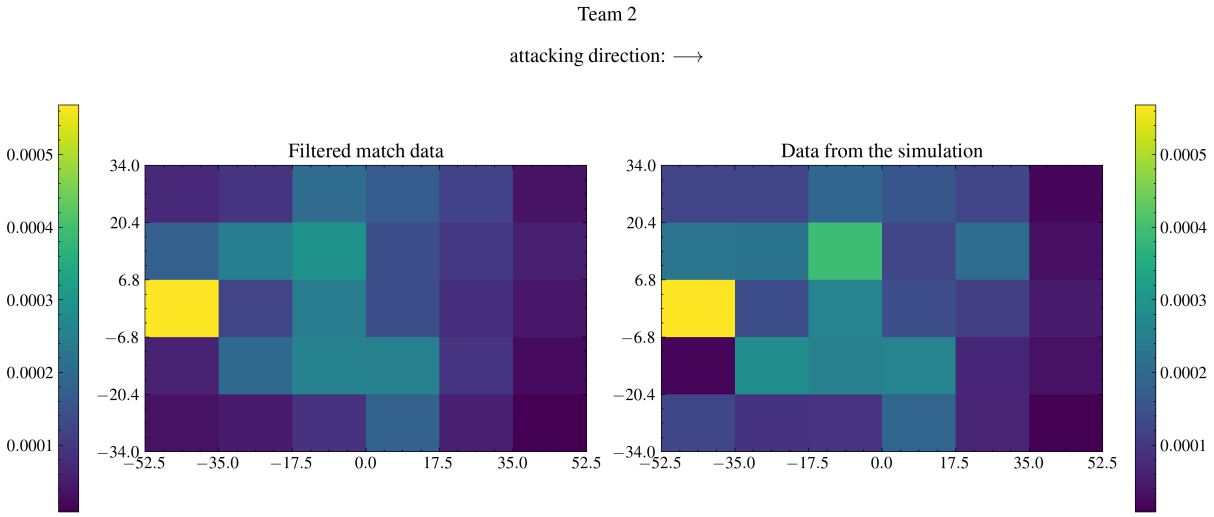

Figure 3.3: Probability distribution for Team 2 in case of one single simulated match.

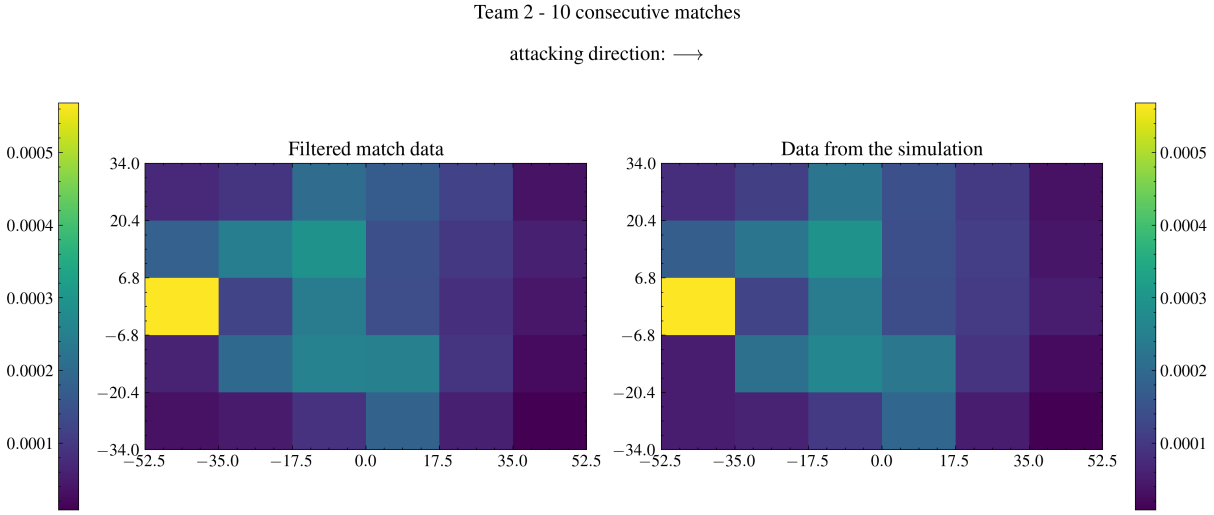

Figure 3.4: Probability distribution for Team 2 in case of 10 consecutively simulated matches.

We performed the same analysis for the other team, the results of which are shown in *Figure 3.3* and *Figure 3.4*. Here we can see that Team 2 was rather in a defensive position, forced to be in front of their own goal.

To illustrate the difference between the probability values of the football field cells we graphed bar plots shown in *Figure 3.5* and *Figure 3.6*. These plots contain the probability values from the original, filtered distribution and from the match simulation, in the case of 50 successive games.

To further examine the resemblance between the two distributions we calculated the relative error from the difference of the probability values for the simulation and the original data as specified in equation (1):

$$\Delta = \frac{1}{N} \sum_i \frac{|p_i - p_{0i}|}{p_{0i}}, \tag{1}$$

where p_i is the probability that the ball is in the i-th rectangle in case of the simulation, p_{0i} for the original data and N is the number of bins.

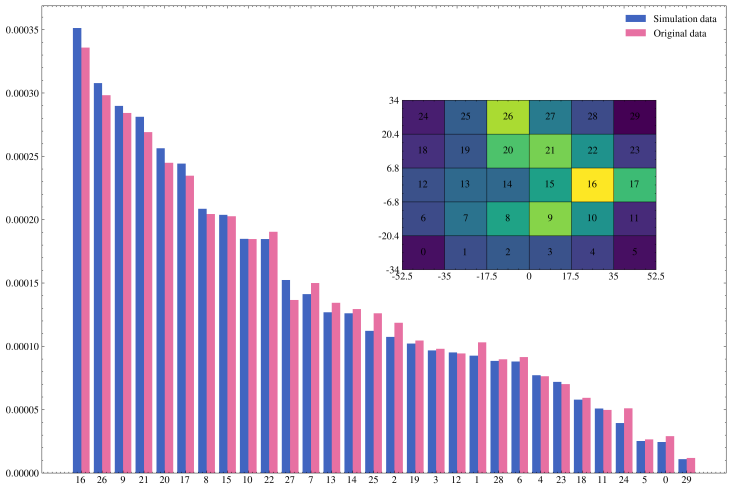

Figure 3.5: The comparison of the probability values, obtained from the original and simulated data, for every rectangle in case of Team 1. The simulation data is from 50 consecutive matches.

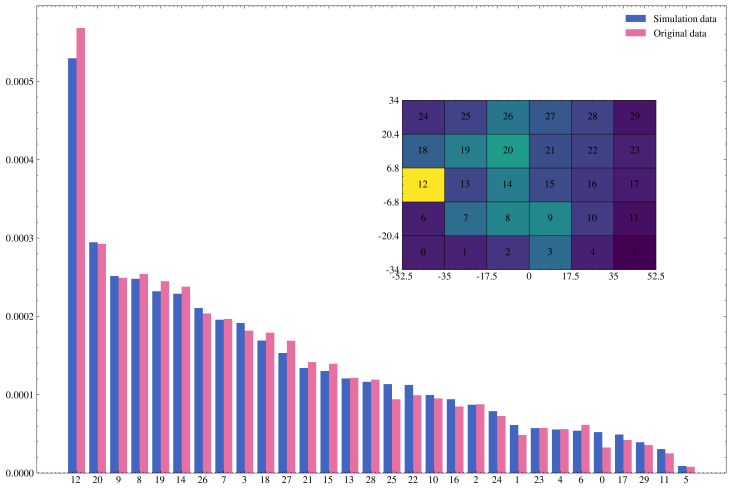

Figure 3.6: The comparison of the probability values, obtained from the original and simulated data, for every rectangle in case of Team 2. The simulation data is from 50 consecutive matches.

These results are shown in *Figure 3.7* and *Figure 3.8*. As we increase the number of consecutively simulated matches, for which the probability values are calculated, this difference tends to approach a small value, which is not equal with 0. This indicates that important positional or temporal correlations not captured by the model may affect the outcome.

For the other two matches the distributions showed less similarity in case of a single simulation. For 10 consecutive matches the resemblance improved, though. For SampleGame1 *Figures 3.9, 3.10*, for SampleGame2 *Figures 3.11, 3.12* illustrate these results for the 10-match cases. For SampleGame 1, we can observe that both teams lead their attacks on parts of the pitch that correspond to their team's right side. In the case of SampleGame 2 one segment of the field is frequented heavier than the other by both teams.

Nevertheless, these do not have the same level of similarity as the Game, which might suggest that the approach we used for the first game may not work properly for the other two,

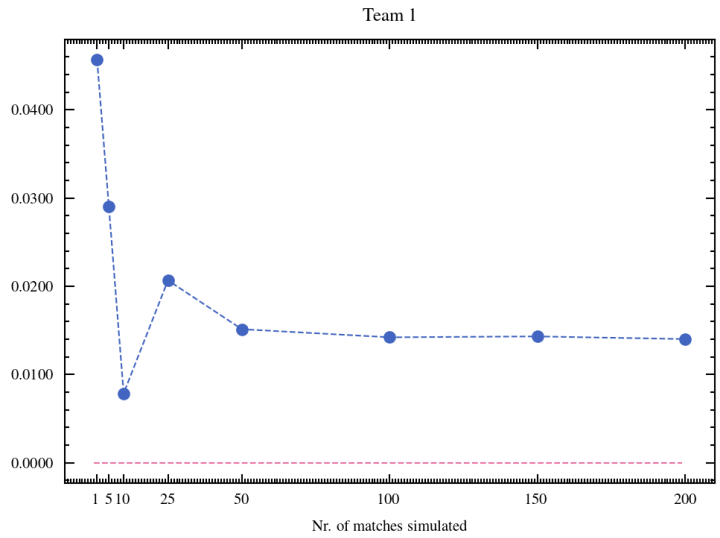

Figure 3.7: The relative difference between the probability values for the original data and the simulation data in case of Team 1.

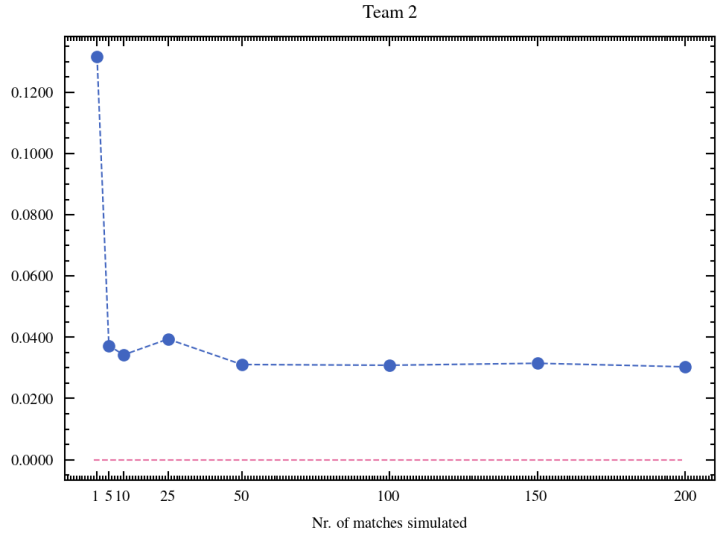

Figure 3.8: The relative difference between the probability values for the original data and the simulation data in case of Team 2.

since they have a different type of data recorded. The relative differences obtained in case of SampleGame2 are depicted in *Figure 3.13* and *Figure 3.14*.

3.2. Entropy

For a given probability distribution we calculated the entropy with the

$$H(X) = -\sum_i p_i(x) \log p_i(x) \tag{2}$$

formula, where p_i is the probability that the ball is in the i-th rectangle. We calculated this value for multiple distributions resulting from different simulations. In *Figures 3.15* and *3.16* the average values are illustrated, with their standard deviations as error bars. Entropy values

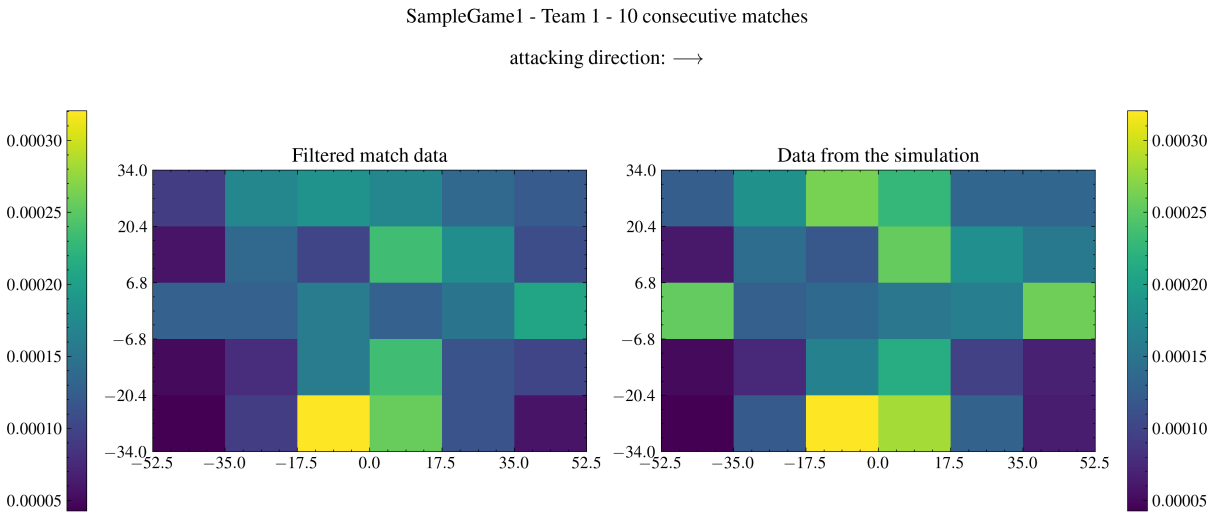

Figure 3.9: Probability distribution for Team 1 in case of 10 consecutively simulated matches for SampleGame 1.

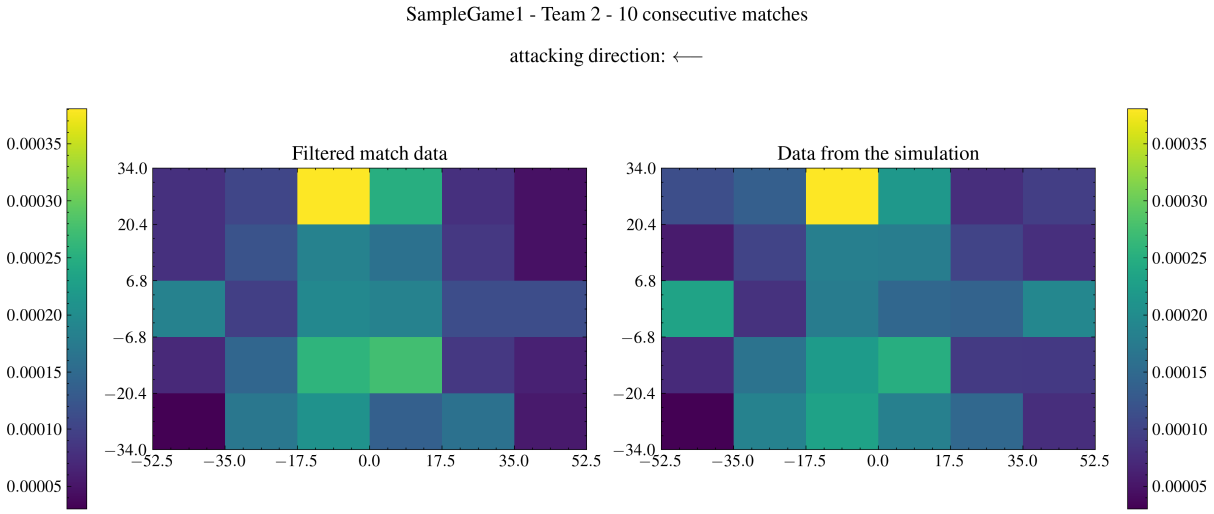

Figure 3.10: Probability distribution for Team 2 in case of 10 consecutively simulated matches for SampleGame 1.

exhibit a convergence trend similar to that observed in the relative error analysis, meaning that they also approach a certain value which is close but not equal with the one obtained from the distribution of the original, filtered data. This phenomenon is present in all three matches that we have analyzed. In addition to the aforementioned images, *Figure 3.17* and *Figure 3.18* show the results for SampleGame2.

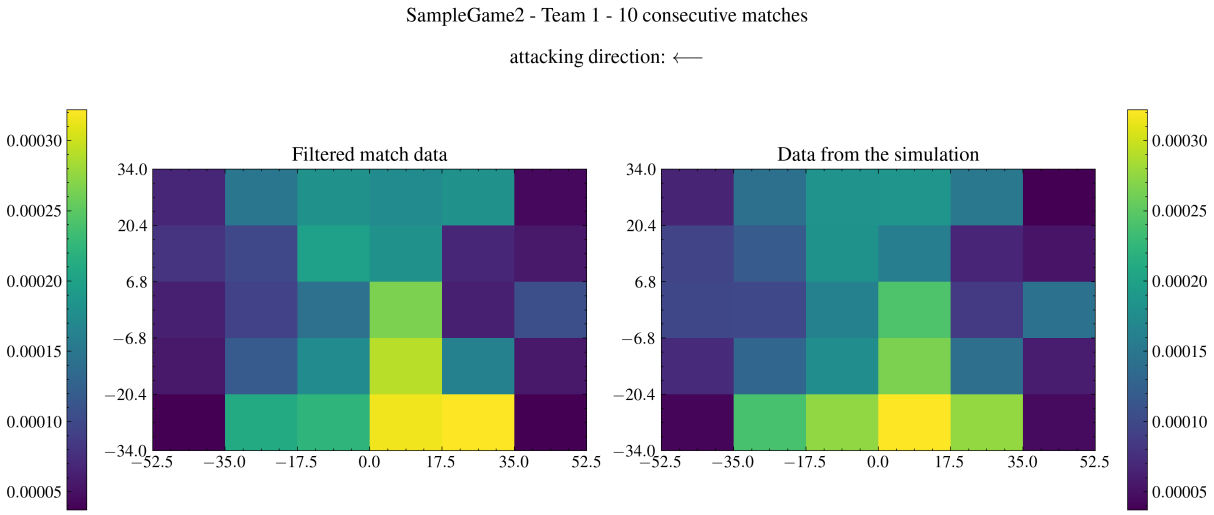

Figure 3.11: Probability distribution for Team 1 in case of 10 consecutively simulated matches for SampleGame 2.

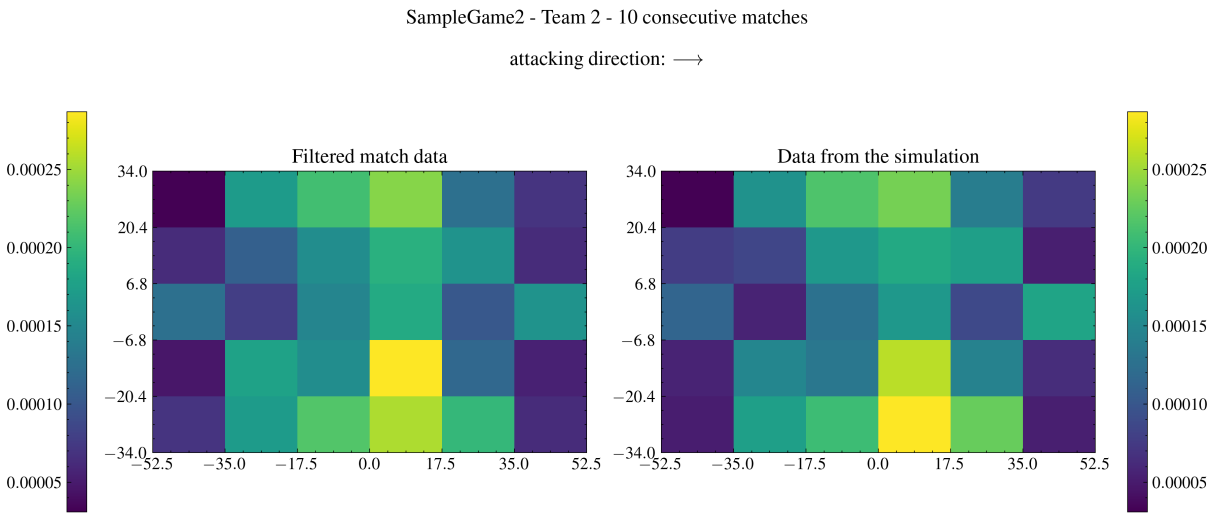

Figure 3.12: Probability distribution for Team 2 in case of 10 consecutively simulated matches for SampleGame 2.

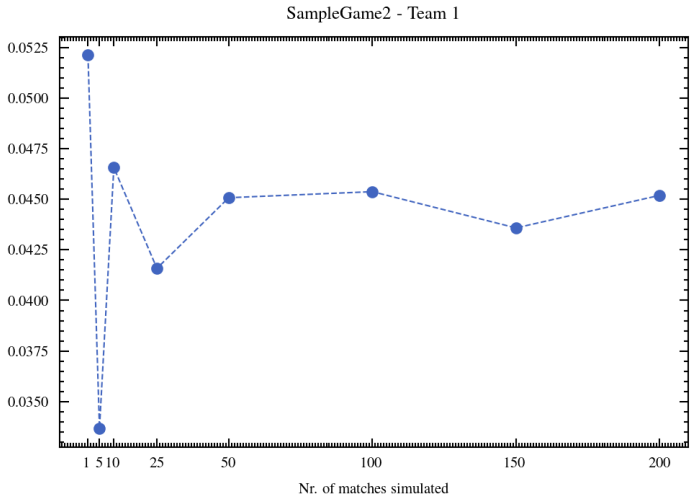

Figure 3.13: The relative difference between the probability values for the original data and the simulation data in case of Team 1 for SampleGame2.

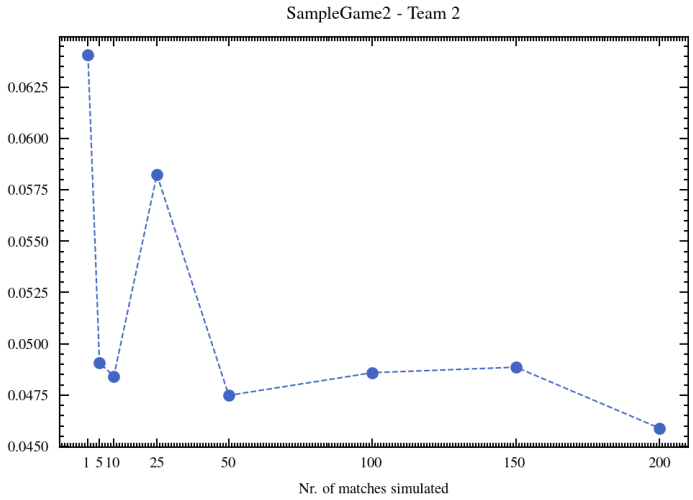

Figure 3.14: The relative difference between the probability values for the original data and the simulation data in case of Team 2 for SampleGame2.

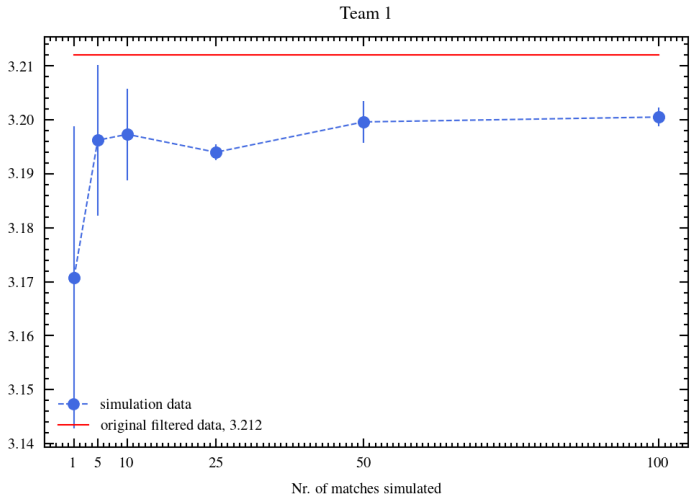

Figure 3.15: Entropy values for different number of match simulations for Team 1.

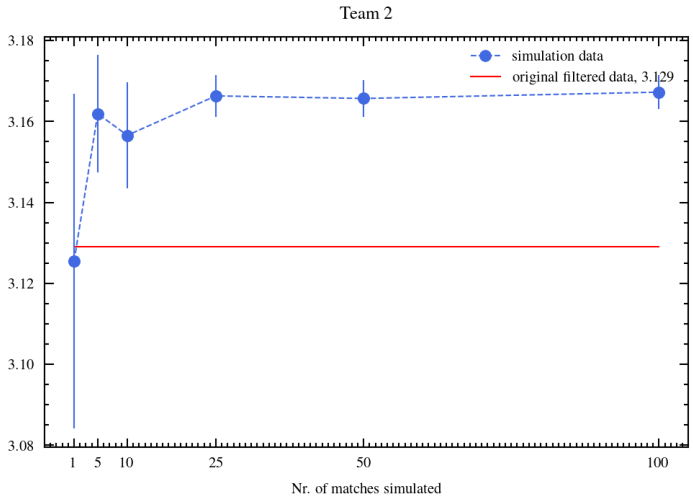

Figure 3.16: Entropy values for different number of match simulations for Team 2.

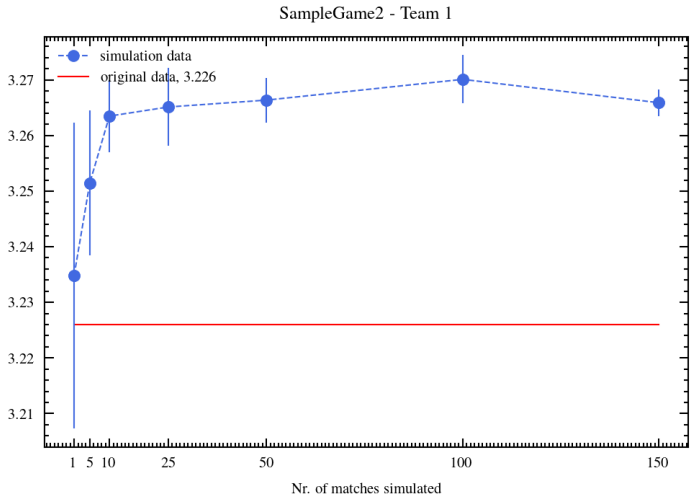

Figure 3.17: Entropy values for different number of match simulations for Team 1 for SampleGame2.

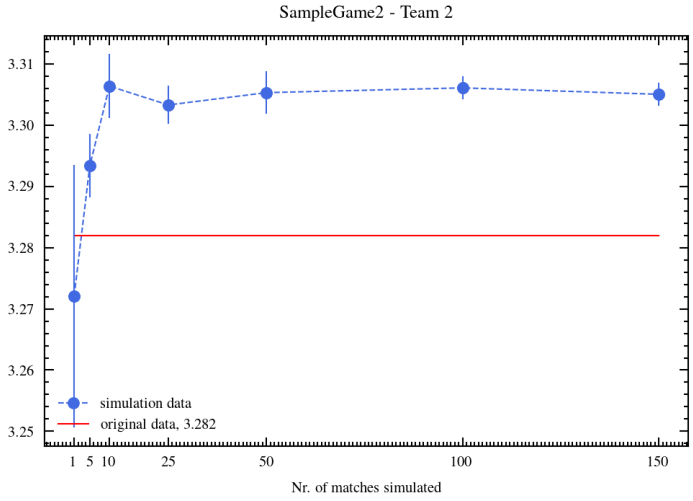

Figure 3.18: Entropy values for different number of match simulations for Team 2 for SampleGame2.

Conclusion

To summarize it, with data from multiple football matches, we built a simulation of consecutive passes. The probability distribution of the ball's position obtained from the simulation data reproduces the original distribution to a fair extent.

We have observed that in the case of data from one single simulated match the differences between the distributions could be easily identified. Nevertheless, if we calculated the probability values based on information gathered from several matches, the resemblance improved significantly.

Since the relative difference between the two distributions (for the original, filtered data and for the ones from the simulation) did not converge to zero, further improvements are needed to be implemented into the simulation. One idea that might enhance the performance is related to the changes in possession. Currently we shift the ball between teams whenever there is either a lost ball or a game interruption. While for the former case this is the correct procedure, in the second case a break in the play can also be caused by a fault, not only by a ball that is out. Most often that player is fouled whose team is in possession, so after the interruption they should be those who start passing again.

Another aspect that was not yet implemented is the event of a goal. We would like to detect if a team scores a goal, since then we could have a final result for the match. Furthermore, the event of a goal is ought to be know, because then the game resumes from the kick-off point.

Acknowledgement

We would like to express our gratitude to Máté Nagy (ELTE, Department of Biological Physics) and Rui Marcelino Maciel Oliveira for providing the dataset in the case of the first match that we have analyzed. In addition to this, we also thank the possibility of using the datasets from the Metrica Sports Sample Data accessible on GitHub ([9]).

Bibliography

[1] Joachim Gudmundsson and Michael Horton. “Spatio-temporal analysis of team sports”. In: *ACM Computing Surveys (CSUR)* 50.2 (2017), pp. 1–34.

[2] RS Mendes, LC Malacarne, and C Anteneodo. “Statistics of football dynamics”. In: *The European Physical Journal B* 57 (2007), pp. 357–363.

[3] A Chacoma et al. “Stochastic model for football’s collective dynamics”. In: *Physical Review E* 104.2 (2021), p. 024110.

[4] A Chacoma et al. “Modeling ball possession dynamics in the game of football”. In: *Physical Review E* 102.4 (2020), p. 042120.

[5] Alina Bialkowski et al. “Discovering team structures in soccer from spatiotemporal data”. In: *IEEE Transactions on Knowledge and Data Engineering* 28.10 (2016), pp. 2596–2605.

[6] Francisco José Peralta Alguacil. *Modelling the Collective Movement of Football Players*. 2019.

[7] Eduardo Velasco Stock, Roberto da Silva, and Henrique A Fernandes. “A physics-based algorithm to perform predictions in football leagues”. In: *Physica A: Statistical Mechanics and its Applications* 600 (2022), p. 127532.

[8] Akifumi Kijima et al. “Emergence of self-similarity in football dynamics”. In: *The European Physical Journal B* 87 (2014), pp. 1–6.

[9] *Metrica Sports Sample Data*. URL: `https://github.com/metrica-sports/sample-data`.

Anexa 2

DECLARAȚIE PE PROPRIE RĂSPUNDERE

Subsemnatul, BĂNICĂ-SOLYMOSI IRISZ, declar că Lucrarea de absolvire/Lucrarea de licență/Proiectul de diplomă/Lucrarea de disertație pe care o voi prezenta în cadrul examenului de finalizare a studiilor la Facultatea de FIZICĂ, din cadrul Universității Babeș-Bolyai, în sesiunea IULIE 2025, sub îndrumarea Prof. dr. NÉDA ZOLTÁN, conf. dr. FARAI-SZABÓ FERENC, lect. dr. BORBÉLY SÁNDOR .....reprezintă o operă personală. Menționez că nu am plagiat o altă lucrare publicată, prezentată public sau un fișier postat pe Internet. Pentru realizarea lucrării am folosit exclusiv bibliografia prezentată și nu am ascuns nici o altă sursă bibliografică sau fișier electronic pe care să le fi folosit la redactarea lucrării.

Prezenta declarație este parte a lucrării și se anexează la aceasta.

Data,

25.06.2025.

Nume,

BĂNICĂ-SOLYMOSI IRISZ

Semnătură

Bănică-Solymosi