Classification and entropy-based analysis of passes in italian men's football

Federico Casadei

03/09/2022

Introduction

For this project, event data of italian men's football referring to season 2017/2018 have been used to analyse ball possessions of the teams. Couples of passes have been employed to perform a classification procedure between two teams, *Juventus* and *Napoli*, using a neural network. Moreover, using the couples of passes, an indicator called *Entropy Ratio* has been defined with the aim of quantify the lack of 'memory' in the ball possessions of the italian teams in that season. Codes written in *Python* have been used for the manipulation of data and the simulations

1 The data used

The data used for this project are of the *event data* type: they consider all the most important actions on the ball during a football match, without considering the movements of the players but recording also the spatio-temporal coordinates of each event. They have been collected by *Wyscout* [1] and are part of a free set of data found on [2]. In this set are present different event data from several european competitions, and a description of them is given in [3].

The passes data of the men's italian football teams of the season 2017/2018 of *Serie A* have been used for this project. Each pass event consists in different informations, including the team and the player who performed it, the time at which it occurred and the starting and ending positions in the pitch[1]. For this study, the pitch has been discretized by dividing it into 20 zones, splitting the length in 5 parts and the witdh in 4 parts. The zones are represented in Figure 1.

The couples of passes have been selected taking, out of all the passes of a given team, all the couples of consecutive simple passes without any other event between them (i.e. no interruptions, fauls, ball out of the pitch, etc.). Everytime a chain of more than two consecutive passes was present, each pass was included both as the second pass of the precedent couple

[1] One must take into account, however, that all the positions on the pitch are reported in % of the length and width of the pitch itself, which are not fixed in football. For this analysis, a standard length of $105\,m$ and width of $68\,m$ have been chosen.

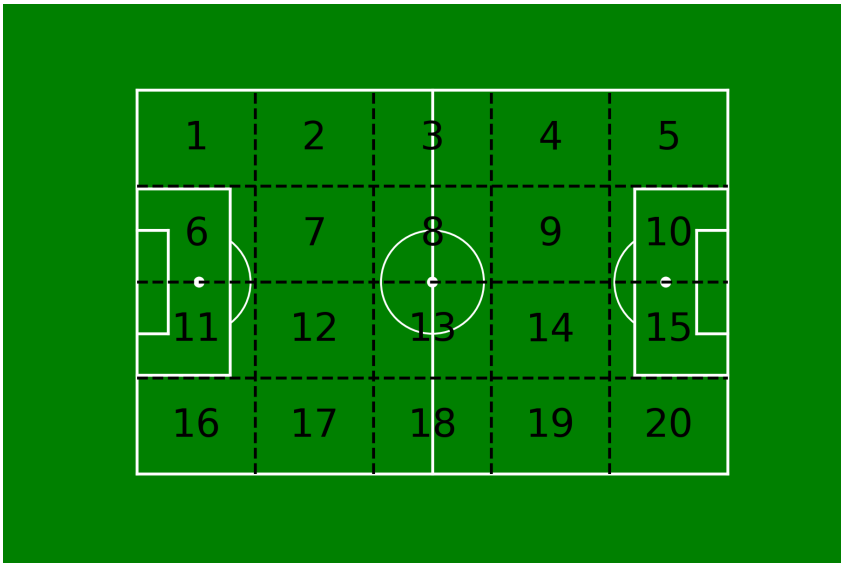

Figure 1: Football pitch with the 20 zones used for the analysis. The pitch has been drawn with the help of the code found in [4]. In these representations of the pitch, the reference team is always assumed to attack from left to right.

and as the first of the successive one (so, for instance, in a chain of three passes, two couples are extracted).

For each couple some parameters have been defined:

- *length*: the average length of the two passes;
- *time*: the time delay between the first and the second pass;
- *angle*: the angle between the directions of the two passes, between $0°$ and $180°$, where $0°$ means pass in the same direction and $180°$ pass in the opposite direction;
- *zones*: the three zones, among the 20 defined above, where, respectively, the first pass starts, the first pass ends and the second starts, the second pass ends[2].

A further cut was finally added, rejecting all the couples with a time delay bigger than $30\,s$ (they were assumed to be errors in the data, and that some sort of interruption was present betweeen the two passes).

[2]In football, of course, the ending position of a pass do not coincide in general with the starting position of the successive one. However, by defect of the data used, that was always the case. In absence of informations about what happens between two passes, this was just neglected in the treatment.

2 Classification of *Juventus* and *Napoli* passes

Using the dataset of the pass couples described in the previous section, a classification procedure has been performed, referring to all the passes[3] of the season 2017/2018 of the italian teams *Juventus* and *Napoli* in the league *Serie A*.

The former became the winning team of the *Serie A* that season, while the latter became the runner-up. In particular, according to several football experts, *Napoli* had a highly specific style of playing, characterized by a lot of short passes with a small time delay between them; on the other hand, *Juventus* seemed to play in a 'slower' way and with longer passes. This classification procedure have been used to confirm or deny these hypotheses.

The variables used for the classification were six, the ones defined in the previous section: the mean length, the time delay, the angle and the three positions on the pitch. The total dataset consisted in 35502 pass couples (15376 from *Juventus* and 20126 from *Napoli*).

The distributions of the lengths, the time delays and the angles are represented in Figures 2, 3 and 4.

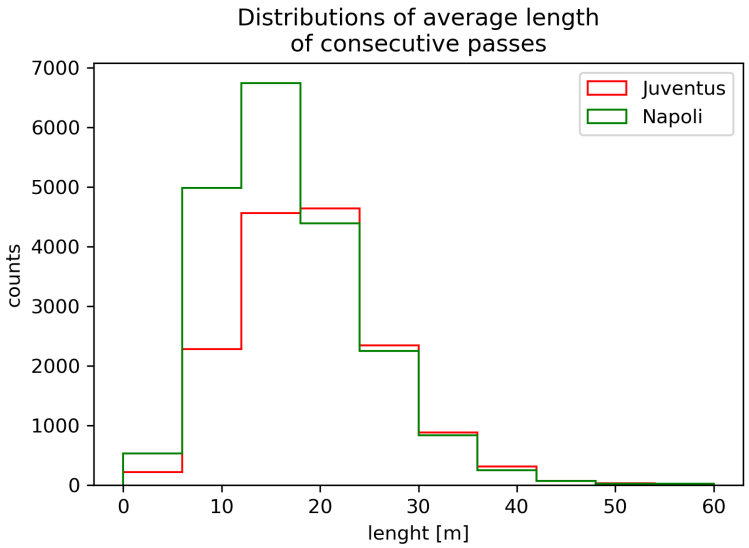

Figure 2: Pass lenght distributions of *Juventus* (red) and *Napoli* (green).

From the pictures is already visible a difference in the distributions, with the passes from *Napoli* being shorter and faster. A different behavior in the angle distributions is visible as well.

Training and testing

To perform the classification, a neural network has been trained, using the open-source library *Scikit-learn* [5]. After normalized all the variables to be between 0 and 1, a neural network with

[3]Not all the passes have been actually used, since the 'isolated' ones are not part of any couple of passes.

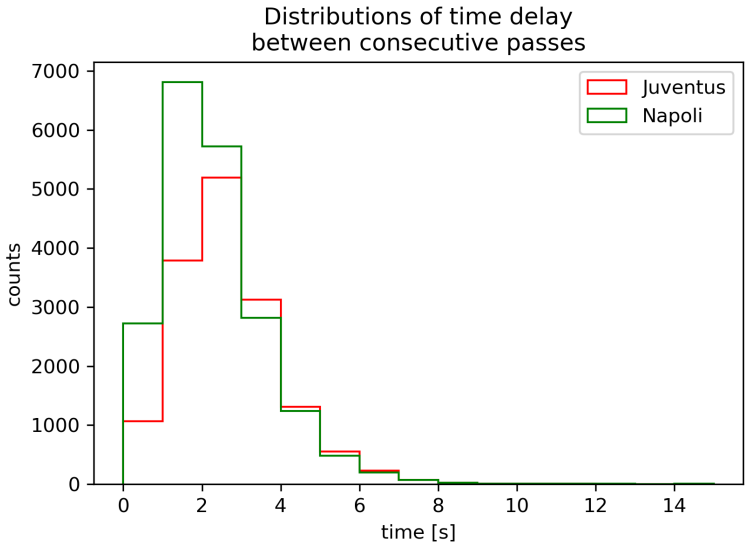

Figure 3: Pass time delay distributions of *Juventus* (red) and *Napoli* (green).

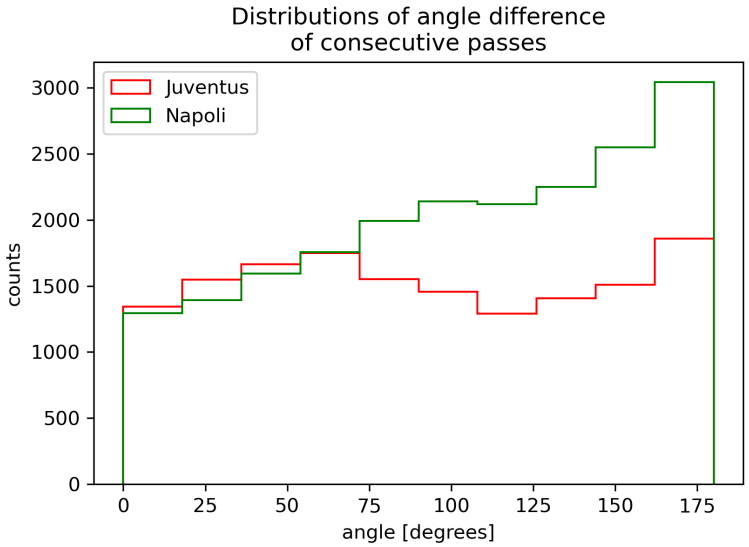

Figure 4: Pass angle distributions of *Juventus* (red) and *Napoli* (green).

3 hidden layers (each with 8 nodes) have been trained with 80% of the data. The remaining 20% was used for testing.

The ROC curve coming from the test sample is reported in Figure 5.

The area under the ROC curve is an indicator for the separation of the two sets of data: in this analysis the result was 0.64. Using a cut threshold of 0.5, the resulting confusion matrix is reported in Table 2.

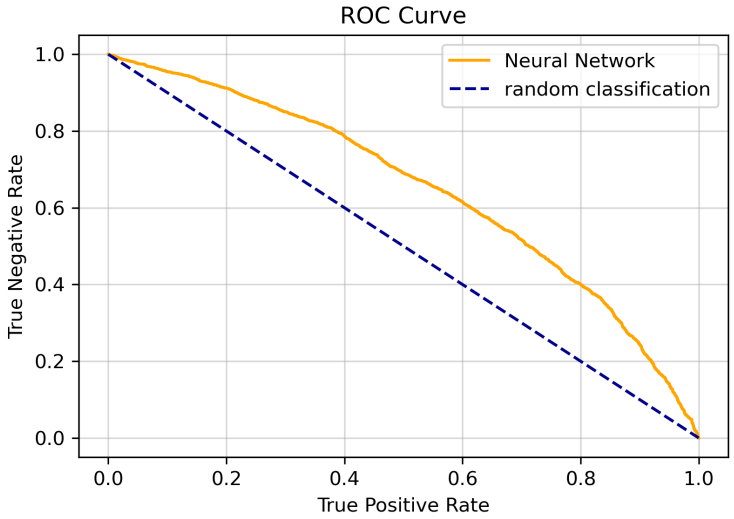

Figure 5: Receiver operating characteristic (ROC) curve for the classification made by the neural network. On the x axis, the true positive rate; on the y axis, the true negative rate.

	Assigned to *Napoli*	Assigned to *Juventus*
Passes from *Napoli*	2861	1126
Passes from *Juventus*	1607	1507

Table 1: Confusion matrix for the classification procedure, with the threshold value equal to 0.5.

3 Entropy and *Entropy Ratio*

Let's describe now a passing chain in football as a Markov process: each state is a zone in the pitch (out of the 20 defined above) and each pass is a transition from zone i to j. Let's consider, in addition, no time dependence. Each ball possession is so equivalent to a sequence of states.

It is useful to give the following definitions:

- p_i: probability to be in the state i;
- p_{ij}: probability, being in the state i, to have a transition to the state j;
- p_{ijk}: probability, after a transition $i \to j$, to have a transition $j \to k$.

It is clear that $\sum_{i=1}^{N_{states}} p_i = 1$, $\sum_{j=1}^{N_{states}} p_{ij} = 1$ and $\sum_{k=1}^{N_{states}} p_{ijk} = 1$. If one has N transitions and n_i of them start from zone i, n_{ij} times there is a transition $i \to j$ and n_{ijk} times

a chain $i \to j \to k$, the probabilities defined above can be estimated as:

$$\hat{p_i} = \frac{n_i}{N} \tag{1}$$

$$\hat{p_{ij}} = \frac{n_{ij}}{n_i} \tag{2}$$

$$\hat{p_{ijk}} = \frac{n_{ijk}}{n_{ij}} \,. \tag{3}$$

With the help of these definitions, it is possible to assign an entropy to each state:

$$S_i = -\sum_{j=1}^{N_{states}} p_{ij} \ln(p_{ij}) \tag{4}$$

and a total entropy:

$$S = \sum_{i=1}^{N_{states}} p_i S_i = -\sum_{i,j=1}^{N_{states}} p_i p_{ij} \ln(p_{ij}) \,. \tag{5}$$

It is well-known that these entropies can quantify the unpredictability of a transition: if, for some zone i, S_i is close to zero, that means that there are few favoured states where the system will be after a transition. On the other hand, if S_i is bigger[4], the probabilities of the successive state will be distributed in a less predictable way.

It is possible, furthermore, to define the entropy

$$S_{ij} = -\sum_{k=1}^{N_{zones}} p_{ijk} \ln(p_{ijk}) \,. \tag{6}$$

This entropy has the meaning to quantify the unpredictability of a transition from the state j, depending on the precendent transition $i \to j$.

The corrisponding estimators can be defined by substituting each probability with the corrisponding estimator defined in Equations (1), (2) and (3).

In a Markov process, by definition, the probability of a given transition only depends on the present state, without any 'memory' of the past history. Thus, we have

$$p^M_{ijk} = p_{jk} \,. \tag{7}$$

However, a passing chain in football is not, in general, a Markov process. In order to quantify the 'memory' in the passing chains of different teams, which is the purpose of this analysis, one can define the *Entropy Ratio*:

$$SR = \sum_{i,j=1}^{N_{zones}} p_i p_{ij} (SR)_{ij} \tag{8}$$

where

$$(SR)_{ij} = S_{ij}/S_j \,. \tag{9}$$

[4]In this case, the maximum value is $\ln(N_{states})$.

For a Markov process, due to Equation (7), $S_{ij} = S_j$ and so:

$$SR^M = 1\,. \tag{10}$$

Instead, for a fully deterministic process, where, given a transition $i \to j$, the successive transition $j \to k$ is automatically defined, S_{ij} is zero for each i and j: in that case the *Entropy Ratio* will be null. Each value between 0 and 1 for SR defines an 'intermediate' process.

Applying this treatment to the pass couples of italian football teams it is possible to define the lack of memory in the ball possessions of each of them.

Simulations

To verify the above statements, two different simulations have been carried out. Each of them made use of the *Numpy* [6] free library in Python for the pseudo-random number generation.

The first one aimed to prove the Equation (10). A Markov process has been simulated, using 20 states and different numbers of transitions between them. The estimator $\widehat{SR^M}$ for the *Entropy Ratio* has been computed and the results, reported in Figure 6, show that it asymptotically converges to 1 in the large sample limit.

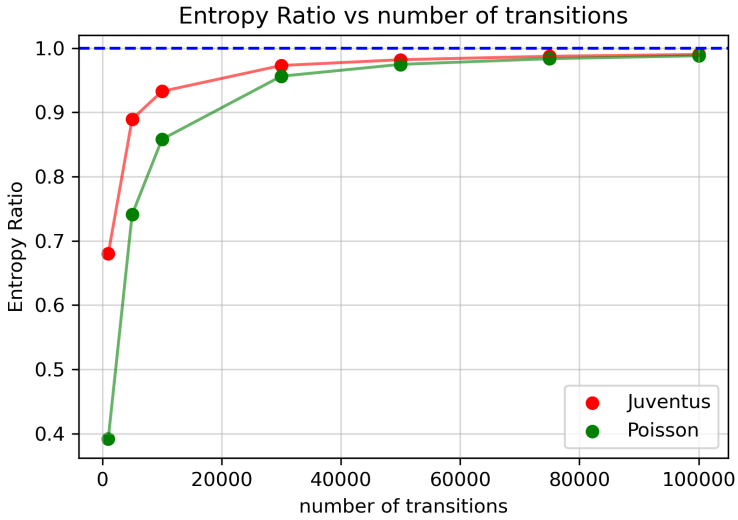

Figure 6: *Entropy Ratio* computed by the simulation of the Markov process as a function of the number of transitions. Green: transition probabilities defined by a Poisson distribution between states. Red: transition probabilities defined by the *Juventus* passes.

It is possible to see that between 10000 and 20000 passes, which is approximately the number of passes made by each analysed team, the asymptotic behavior is not already reached. This fact has to be taken into account to give a correct interpretation of the $\widehat{SR^M}$ of the different teams: in fact, a team that made more passes is in a zone where the asymptotic beahvior of the estimator is closer; but this estimator should give informations about the memory of the ball possession of each team independently on the number of passes.

Another correction that has not to be neglected is that a real football passing chain is not continuous as the Markov process considered is. In fact, the possession is continuously interrupted, and there are 'jumps' between states[5].

To fix this problem, another simulation has been implemented. This time, couples of transitions have been randomly generated: in this way the simulation was more similar to the datasets of the pass couples of the teams. After each couple of transitions, with a probability of 70% (close to a typical one for real passing chains of italian football teams) another couple begins, with the first transition corresponding to the second transition of the precedent couple. In the remaining 30% of the cases, a jump to another zone is made and a new couple is generated. The independence of each transition of the precedent one has been mantained also in this simulation.

The results of this simulation are shown in Figure 7. It is clear that this modification brings to a different limit value for the *Entropy Ratio*.

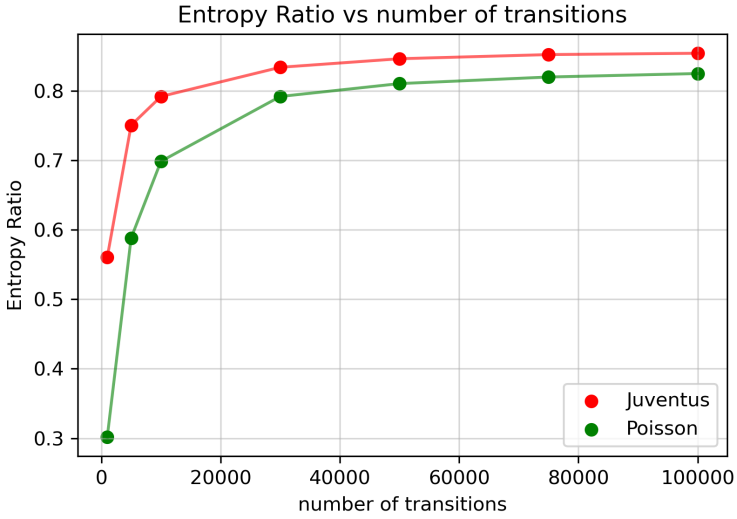

Figure 7: *Entropy Ratio* computed by the simulation of the Markov process with 'jumps' as a function of the number of transitions. Green: transition probabilities defined by a Poisson distribution between states. Red: transition probabilities defined by the *Juventus* passes. The asymptotic value for SR is smaller than 1 and dependent on the transition probabilities.

Entropy Ratio of the *Serie A* teams

For the reasons explained above, the *Entropy Ratio* computed with the Equation (8) and estimated with the Equations (1), (2) and (3) does not range between 0 and 1. It cannot by itself constitute an objective and sample-independent indicator of the lack of memory of ball possession of the teams. In Figure 8 are indeed plotted the $\widehat{SR}$ of each team, as a function of their total number of passes.

[5]A passing chain is interrupted somewhere and another one starts in another location.

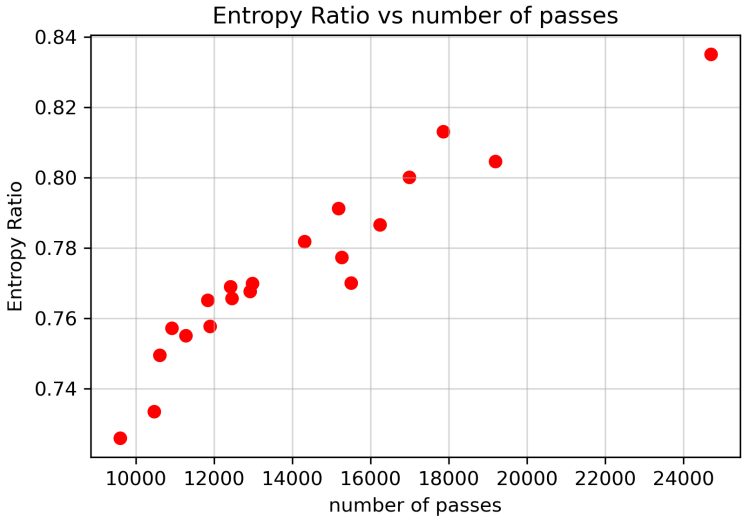

Figure 8: *Entropy Ratio* of the italian teams as a function of their number of passes. A clear dependence is visible.

It is very clear from the plot that there is a strong dependence of $\widehat{SR}$ on the number of passes.

To solve these problems, a simulation as the one done in the precedent section has been performed for each team: the number of transitions has been set to be equal to the number of passes of that team and the probability to do a jump to a new couple was estimated by the data. Moreover, the transition probabilities between zones have been defined to be equal to the real one, calculated by the data of that team. Then, a new variable has been defined as the ratio between the *Entropy Ratio* evaluated by the passes of a team and the one computed by the simulation:

$$\eta = \left. \frac{\widehat{SR}}{\widehat{SR^M}} \right|_n . \tag{11}$$

This new variable is a better estimator; it is included between 0 and 1 because it is 0, as before, when the process is totally deterministic, and it can reach values close to 1 even with jumps in the passing chains and far from the large sample limit. In fact, these effects are encounted in the denominator of the Equation (11).

Using η to quantify the lack of memory of the different teams, the dependence on the number of passes is removed, as shown in Figure 9.

The values of the indicator η for the different teams are reported in Table 2.

For all the teams, the value of η stays near to 0.95.

These values can give an indication on how much the ball possession of a specific team is similar to a Markov process. Anyway, it does not give relevant informations about the performance of a team: the correlation between η and points in the standings or goal scored by a team is not pronunced, and the number of passes is itself a better indicator.

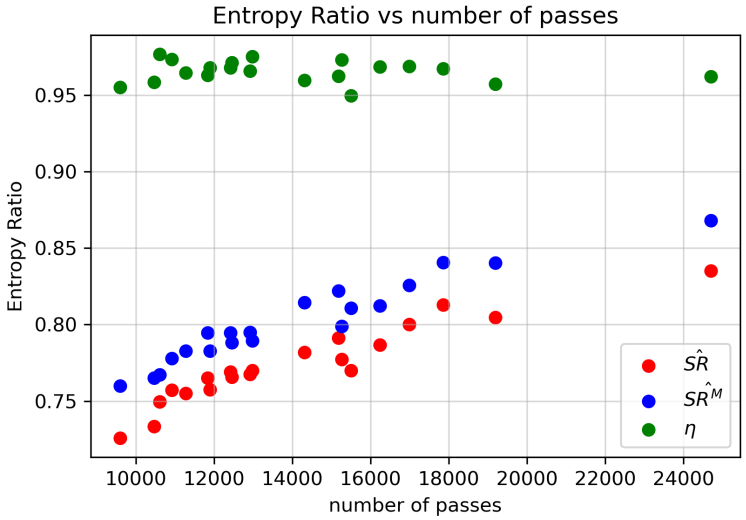

Figure 9: Red: *Entropy Ratio* of italian teams; blue: *Entropy Ratio* computed by a Markov simulation with the same number of transitions and jumps as the relative team; green: η, the ratio between the latter quantities.

Conclusions

For this analysis, event data from an entire season of italian men's football have been used, to build, in particular, passing chains for each team.

A classification procedure has been carried out, training a neural network to predict at what team, between *Juventus* and *Napoli*, belongs a given pass using variables got from data. This classification brought to a ROC curve with an area under the curve (AUC) equal to 0.64. The separation found between the style of playing of the two teams, together with the inspection of the distributions of the selected variables, gives a quantitative affirmation of the 'common feeling' about them. This classification could be repeated between different teams and even between different leagues, to look, for instance, if there are relevant differences in the style of playing in different countries.

By defining transition probabilities and describing the passing chains as Markov processes, an entropy-bases analysis has been performed, with the help of simulations, on the couples of passes of each team; in this mathematical background, the *Entropy Ratio* has been defined, in order to quantify the lack of 'memory' in the passing chains of each team. A further parameter η has been defined, to deal with some weaknesses in the former. This indicator, however, has not relevance in predicting or giving explanations to important performace indicators as points and scored goals.

References

[1] https://wyscout.com

Team	$\widehat{SR}$	$\widehat{SR^M}$	η
Atalanta	0.770	0.811	0.950
Benevento	0.770	0.789	0.975
Bologna	0.766	0.788	0.971
Cagliari	0.755	0.783	0.964
Chievo	0.758	0.783	0.968
Crotone	0.726	0.760	0.955
Fiorentina	0.782	0.815	0.960
Genoa	0.769	0.795	0.968
Hellas Verona	0.749	0.767	0.977
Internazionale	0.813	0.841	0.967
Juventus	0.805	0.840	0.957
Lazio	0.777	0.799	0.973
Milan	0.800	0.826	0.969
Napoli	0.835	0.868	0.962
Roma	0.787	0.812	0.968
Sampdoria	0.791	0.822	0.962
Sassuolo	0.733	0.765	0.959
SPAL	0.757	0.778	0.973
Torino	0.768	0.795	0.966
Udinese	0.765	0.795	0.963

Table 2: Values of *Entropy Ratio* for the italian teams in season 2017/2018, evaluated by data and by the simulation, and η.

[2] Pappalardo, Luca; Massucco, Emanuele (2019): Soccer match event dataset. figshare. Collection. https://doi.org/10.6084/m9.figshare.c.4415000

[3] https://www.nature.com/articles/s41597-019-0247-7

[4] https://fcpython.com/visualisation/drawing-pitchmap-adding-lines-circles-matplotlib

[5] https://scikit-learn.org/stable/index.html

[6] https://numpy.org