Estimating the Maximal Speed of Soccer Players on Scale

Laszlo Gyarmati and Mohamed Hefeeda

Qatar Computing Research Institute, HBKU
{lgyarmati,mhefeeda}@qf.org.qa

Abstract. Excellent physical performance of soccer players is inevitable for the success of a team. Despite of this, a large-scale, quantitative analysis of the maximal speed of the players is missing due to the sensitive nature of trajectory datasets. We propose a novel method to derive the in-game speed profile of soccer players from event-based datasets, which are widely accessible. We show that eight games are enough to derive an accurate speed profile. We also reveal team level discrepancies: to estimate the maximal speed of the players of some teams 50% more games may be necessary. The speed characteristics of the players provide valuable insights for domains such as player scouting.

1 Introduction

Quantitative performance analysis in sports has become mainstream in the last decade. The focus of the analyses is shifting towards more sport-specific metrics due to novel technologies. These systems measure the movements of the players and the events happening during trainings and games. This allows for a more detailed evaluation of the professional athletes with implications on areas such as opponent scouting, planning of training sessions, or player scouting.

Previous works that analyze soccer-related logs focus on the game-related performance of the players and teams. Vast majority of these methodologies concentrate on descriptive statistics that capture some part of the strategy of the players. For example, in case of soccer, the average number of shots, goals, fouls, passes are derived both for the teams and the players [1, 2]. Other works identify and analyze the outcome of the strategies that teams apply [10, 8, 6, 4]. However, the physical performance of the players has not received detailed attention from the research community.

It is a challenging task to get access to metrics related to the physical performance of soccer players. The teams consider such information highly confidential, especially if it covers in-game performance. Despite the fact that numerous teams deployed player tracking systems in their stadiums, datasets of this nature are not available for the research or public domain. It is nearly impossible to have quantitative information on the physical performance of all the teams in a competition. Hence, most of the analysis and evaluation of the players' performance do not contain too much information on the physical aspect of the game.

We address this issue by proposing a methodology that is able to derive the in-game speed profile of soccer players, *i.e.*, how much time a player needs to

cover a certain distance in the best case scenario. In other words, we determine the relation between the maximal speed of a player for a given range. In addition, we are able to do this on scale: our method is able to analyze the physical performance of the players across multiple seasons and competitions without any major investment. It is not required to have an expensive, dedicated player tracking system deployed in the stadium. Instead, if the game is broadcasted, our methodology can be used. As a consequence, our technique does not require the consent of the involved teams yet it provides insights on the physical performance of the players of both teams. Soccer data companies are covering 50+ leagues providing the potential to analyze the speed profile of tens of thousands of players. The main contribution of our work is threefold:

1. we propose a methodology to extract the maximal speed characteristics of the players,
2. we determine the minimal number of games necessary to determine the physical capabilities of a player,
3. and we show that the playing style of a team has a significant impact on the accuracy of the speed estimation.

2 Methodology

In this section we introduce our methodology used to extract the movements of the players and then to estimate their maximal speed. Our final goal is to derive a regression model between the distance of the movement and the minimal time necessary for it. We use an event-based dataset throughout our analyses that we describe next.

Dataset. We use an event-based dataset generated by Opta [9] covering the 2012/13 season of La Liga (*i.e.*, the first division soccer league of Spain). The dataset contains all the major events of a soccer game including passes, shots, dribbles, tackles, *etc.*. For example, the dataset has more than 300,000 passes and nearly 10,000 shots. The feature of the dataset we explore is that it contains the time and the location of these events as well along with the identity of the involved players. Hence, it is possible to derive a coarse grain time-series of the movements of the players. We note that the precision of the time annotation is one second. The procedure uses all the (x, y) positions a player has during a game and creates a movement vector using a consecutive pair of (x, y) coordinates and timestamps to create a movement vector. We illustrate the derived movements of a player in Figure 1 given a single game. This is the first step of our methodology: extracting the movement vectors of the players. The event-based dataset we use is sparse in terms of the position of the players, *i.e.*, the physical location of a player is only recorded when the player was involved in some ball-related event[1]. As such, the elapsed time between two events of a player can be as low as couple of seconds but it can reach several minutes too. This introduces significant noise to the data that we have to handle in the regression model.

[1] This is a consequence of the data acquisition process: the games are annotated based on the television broadcast that focuses on the ball all the time.

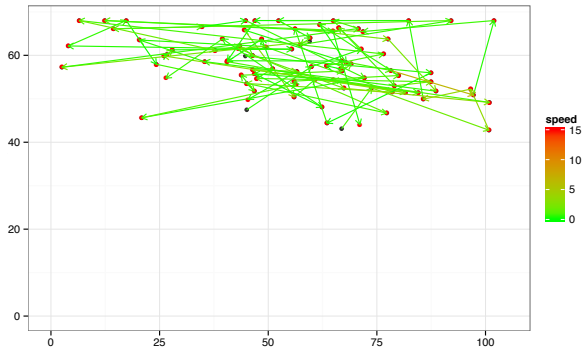


Fig. 1. Movement vectors of a player derived from an event-based dataset. Not only the location of the end points are present in the data but the speed of the movement too.

Handling passes. It is straightforward to determine the timestamp and the position of the players in case of single-player events (*i.e.*, all the events except passes). In terms of passes, we have a complete datapoint for the initiator of the pass (*i.e.*, timestamp and location), however, at the receiving end, the dataset does not contain a timestamp. To overcome this issues, and to increase the wealth of the extracted time-series, we apply four methods to estimate the time when a pass was received. The four options are:

- *0. Neglect.* The event of receiving a pass is neglected, *i.e.*, we do not use this partial information.
- *1. Previous event.* The timestamp of the previous event is used, *i.e.*, the initiation of the pass. This is a lower-bound estimation of the time of reception.
- *2. Next event.* The timestamp of the next event is applied. This timestamp is an upper-bound on the reception of the pass.
- *3. Regression.* Two passes may follow each other immediately in soccer, *i.e.*, when a player receives a pass, handles the ball, and passes the ball forward with a single touch. We can select these consecutive passes from the dataset, *i.e.*, in this case the (x, y) coordinates and the identity of the player are the same (the receiver of the first pass and the initiator of the next one). Therefore, in case of the first passes we know the timestamp of both the initiation and the reception. Therefore, based on these accurate ball movements, we build a linear regression model between the range and the elapsed time of the passes. We apply a 10-fold cross validation of the model; the accuracy score is 33.26%, while 73.2% of the times we are able to estimate the time duration of the pass with an error of at most one second. Using this regression model, we estimate the speed of the passes and as such the time of the pass reception to increase the instances where the position of the players are known.

At the end of the data extraction step, for each game and each player we have a list movements done during the game. Such a tuple contains the start and end (x, y) coordinates of the player along with the appropriate timestamps.

Diverse field sizes. An interesting property of the rules of soccer is that the sizes of the field are not fixed, there is some room to design a soccer pitch even in case of international matches. According to the first law of the game, the length of the pitch shall be between 100 and 110 meters, while the width between 64 and 75 meters [3]. There is an ongoing standardization effort, most of the newly constructed stadiums have a pitch with a size of 105x68m[11]. Spain is not an exception to this extent, where the dimension of Elche's stadium is 108x70m while the same is 100x65m in case of Rayo Vallecano [7]. The dataset we apply uses relative coordinates, *i.e.*, both sides of the pitch are measured between 0 and 100 unit. We transform these relative units into the metric system using the sizes of the stadiums. At the end of this transformation, the end points of the movement vectors are measured in meters.

Filtering. Before building the regression model of the maximal speed, we apply a data cleaning step. As we mentioned above, the derived movement dataset contains a lot of noise. On the one hand, it is owed to the methodology we use to derive the movement vectors, while on the other hand the time is annotated in seconds. As a sanity check, we apply two filters to remove the obvious flaws from the dataset. We filter out all the movement vectors that span more than 20 seconds. Our choice of this constraint is based on the fact that professional sprinters are able to run 100 meters in less than 10 seconds. Thus, it is reasonable to assume that the maximal speed of soccer players is above 50% of the sprinters. The second filter is based on the speed of the movement: we remove those movements where the speed of the player is larger than 15m/s.

Quantile regression. We use the filtered movement vectors to build a regression model that estimates the maximal speed of the players depending on the distance they cover. Our goal is to determine the minimal time a player needs to cover a certain distance. For this purpose, we apply the techniques of quantile regression where the regression model estimates a specific quantile of the dataset (instead of the mean in case of the linear regression) [5]. We show the speed of all the movement vectors of a player throughout a whole season in Figure 2 along with the 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5 quantile regression lines. We note that the 0.5 quantile regression model equals the regular linear model. Due to the lack of accessibility of ground truth, it is challenging to evaluate quantitatively which quantile is the best estimator of the players' maximal speed. Based on extensive qualitative analysis we decided to use the 0.05 quantile regression model for the speed estimation (annotated by red solid line in the figure).

We evaluate the accuracy of the derived regression models based on their consistency, *i.e.*, how stable the parameters of the regression model are. If the parameters of the regression model—namely, the intercept and the slope—are similar irrespective to which subset of the dataset we use, the model can be considered sound.

3 Evaluation

The evaluation of the proposed methodology is twofold. First we focus on the overall performance of the speed estimators and then we analyze the scalability of the methods. To investigate the accuracy of the regression models (*i.e.*, the four

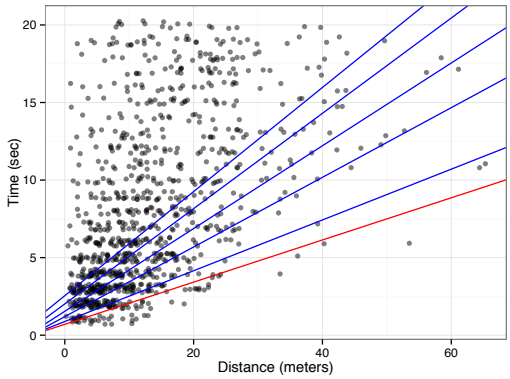


Fig. 2. The speed of the movement vectors of a player throughout a season. The solid lines show the 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5 quantile regression models. The red line is the 0.05 quantile we decided to use as the estimator of the players' maximal speed.

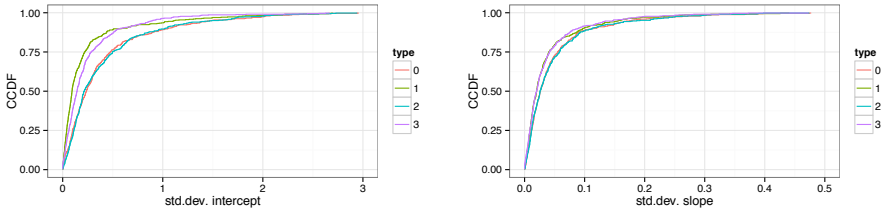


Fig. 3. The accuracy of the methods estimating the maximal in-game speed of soccer players. The cumulative distribution functions of the parameters of the quantile regression models reveal that the *previous event* method provides the best speed estimation.

variants how we handle the passes), we derive the quantile regression model of all the players in the dataset using all the movements the players had throughout the season. In case of each player, we divide the movement vectors into two and then compute the parameters of the quantile regression line. Afterwards, we determine the standard deviation of the parameters in case of all the players separately. In Figure 3 we show the cumulative distribution function of the parameters in case of the four methods. In case of both parameters, the *previous event* (#1) provides the best accuracy, *i.e.*, it has the lowest deviation in the parameters given the random subsets of the sample. Not only the precision of speed estimation is the highest in case of the previous event method but it enables us to investigate the maximal speed of more players compared to the *neglect* version (539 vs. 529 players). The *next event* method (#2) does not enhance the accuracy of the speed estimation as the results reveal.

We next focus on the following question: how many games do we need to accurately estimate the maximal speed of the players? We answer this question by analyzing the standard deviation of the parameters of the quantile regression models given different subset of the games a player was involved. Specifically,

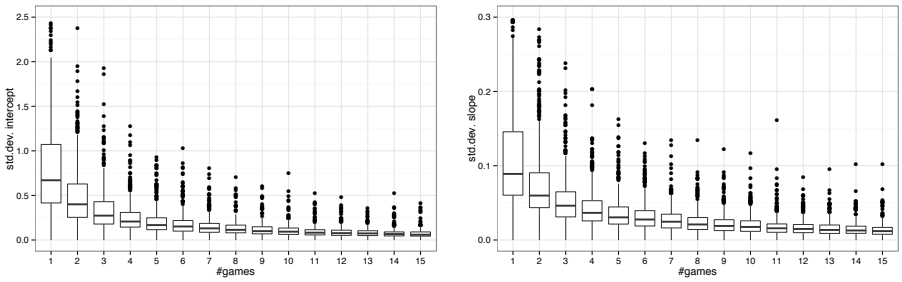
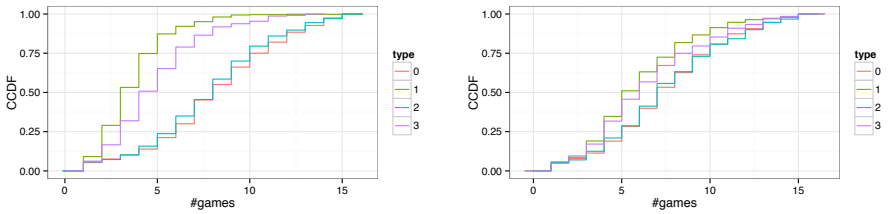


Fig. 4. The accuracy of the *previous event* method given the number of games from which we derive movement information. Having data from eight games provides a quite accurate estimation of the speed of the players.

we randomly select $n = 1, 2, \ldots$ games from the ones the player participated in and derive the regression model; we repeat this ten times for each n and for each player. We show the deviation of the parameters in Figure 4, where we focus on the best estimator we have seen above. As the results reveal, the accuracy of the *previous event* method is stable if we have data from at least 8 games. This is a fascinating result that implies we are able to characterize the in-game maximal speed of a player based on one quarter of a season—which has 38 games.

We analyze the accuracy and the information need of the different methods in Figure 5. Here we apply thresholds for the deviation of the parameters. For each player we determine the minimal number of games that enable us to estimate the maximal speed of the players with the given accuracy. Specifically, the thresholds are 0.25 and 0.025 in case of the intercept and the slope, respectively. In case of 50 percent of the players, it is enough to have data for five games to have an accurate enough estimation of their maximal speed (in case of the *previous event* method). There are large discrepancies among the methods, *e.g.*, the *neglect* and *next event* methods need twice as much games to provide accurate speed estimation for 80 percent of the players compared to the *previous event* method. Based on the results we can draw the following conclusion: one should use the *previous event* or the *regression* methods.

There are team specific discrepancies in case of the information need of the methods. Table 1 presents the mean number of games required to estimate the speed of the players of a given team accurately. In general, we need the fewest number of games in case of the players of FC Barcelona. This is inline with the fact that FC Barcelona dominates the ball possession in its games and such its players have numerous ball related events, and as such, movement vectors. However, in case of the *previous event* method,we need only 2.6 games to estimate the speed of the players of Celta de Vigo too. In some cases, the discrepancy of the required number of games is significant, *e.g.*, we need 50% more games in case of Espanyol and Valencia using the *neglect* methodology compared to FC Barcelona. These differences have a crucial impact on one of the application domain of the methodology: player scouting (*i.e.*, one has to analyze more games if the player is part of a specific team).



(a) Intercept (b) Slope

Fig. 5. The accuracy of the speed estimation methods given the used amount of data (*i.e.*, the number of games). The cumulative distribution functions show the minimal number of games required for a good enough speed estimation. The *previous event* method provides the best accuracy based on a given set of games.

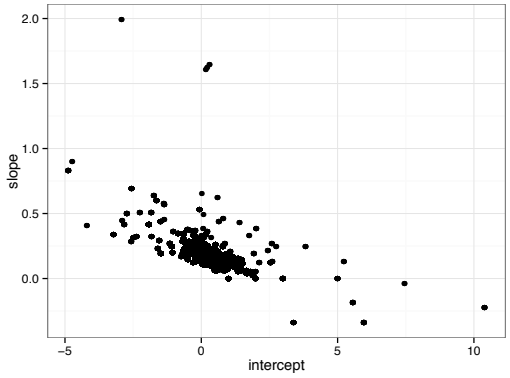


Fig. 6. The speed profile of players in the Spanish first division (the intercept and the slope of their quantile regression model). Players have diverse, in-game physical activities that reveals a novel aspect of their performance.

The proposed methodology indeed can be used for player scouting. As Figure 6 shows, the maximal in-game speed characteristics of the players are diverse, hence, it provides an additional facet for performance evaluation. One can identify suitable candidate to sign who has the physical capabilities necessary for the playing style of a given team.

4 Conclusions

We proposed a new technique to estimate the maximal speed of soccer players. Using event-based datasets of eight games we are able to accurately determine the speed profile of the players. The investigations revealed that teams require diverse size of datasets for a precise speed estimations. As a future work, we plan to analyze the discrepancies of the estimations across players and leagues. Our method provides a new way to evaluate the performance of soccer players,

	intercept				slope			
Team	#0	#1	#2	#3	#0	#1	#2	#3
Athletic Bilbao	8.1	3.1	7.6	4.0	7.9	5.7	7.2	5.5
Atletico Madrid	7.7	3.2	6.8	4.1	7.3	5.5	6.4	5.0
Barcelona	6.1	2.8	7.0	3.6	6.5	5.2	7.7	5.5
Celta de Vigo	7.4	2.6	7.9	4.2	7.0	6.0	6.8	6.2
Deportivo La Coruna	8.6	3.5	8.1	4.2	7.6	5.7	7.8	6.4
Espanyol	10.1	4.4	9.8	5.8	8.0	6.1	8.7	8.3
Getafe	8.0	4.6	8.4	5.8	7.3	8.0	8.5	7.9
Granada CF	9.3	3.7	9.2	5.5	8.3	6.5	8.3	7.0
Levante	8.8	4.4	8.3	6.3	8.2	5.6	7.8	6.9
Mallorca	9.2	4.1	7.9	6.1	9.2	6.4	8.2	7.5
Malaga	8.1	3.3	8.4	4.6	6.9	5.4	7.5	6.7
Osasuna	8.0	3.6	7.3	5.3	8.0	5.8	8.0	6.5
Rayo Vallecano	7.2	4.4	6.9	4.5	6.6	7.3	6.5	6.0
Real Betis	8.5	3.4	7.8	4.8	8.0	5.2	6.8	5.5
Real Madrid	7.3	3.2	7.3	4.2	6.8	5.3	7.1	6.1
Real Sociedad	9.2	2.5	8.4	4.1	7.7	5.6	7.1	6.3
Real Valladolid	9.0	3.6	8.4	5.3	7.4	6.2	7.5	7.6
Real Zaragoza	6.4	4.2	6.3	4.9	5.6	6.4	6.1	6.1
Sevilla	7.8	4.3	7.5	4.4	7.3	6.2	8.2	5.9
Valencia	8.8	3.6	8.0	5.7	9.0	5.7	8.3	7.3

Table 1. The mean number of games needed by the methods to estimate the speed of the players of a given team. There are significant discrepancies among the teams, *e.g.*, 50% more games may be needed for an accurate estimation in case of Espanyol and Valencia.

particularly, from a physical performance point of view. Such insights can be used as competitive advantage for opponent and player scouting.

References

1. Anderson, C., Sally, D.: The Numbers Game: Why Everything You Know about Football is Wrong (2013)
2. Duch, J., Waitzman, J.S., Amaral, L.A.N.: Quantifying the performance of individual players in a team activity. PloS one 5(6), e10937 (2010)
3. FIFA: Laws of the Game (2014)
4. Gyarmati, L., Kwak, H., Rodriguez, P.: Searching for a unique style in soccer. In: Proc. 2014 KDD Workshop on Large-Scale Sports Analytics (2014)
5. Koenker, R., Hallock, K.: Quantile regression: An introduction. Journal of Economic Perspectives 15(4), 43–56 (2001)
6. Lucey, P., Oliver, D., Carr, P., Roth, J., Matthews, I.: Assessing team strategy using spatiotemporal data. In: Proc. 19th ACM SIGKDD. ACM (2013)
7. Marca: Cual es el campo mas grande de la Liga? (2014)
8. Narizuka, T., Yamamoto, K., Yamazaki, Y.: Statistical properties of position-dependent ball-passing networks in football games. arXiv:1311.0641 (2013)
9. OptaPro: `http://optasportspro.com` (2014)
10. Peña, J.L., Touchette, H.: A network theory analysis of football strategies. arXiv preprint arXiv:1206.6904 (2012)
11. UEFA: Guide to Quality Stadiums (2014)