Towards Automated GIS-based Analysis of Scoring Attempt Patterns in Association Football

Gilbert Kotzbek
University of Vienna
Department of Geography and Regional Research
Universitätsstraße 7
Vienna, Austria
gilbert.kotzbek@univie.ac.at

Wolfgang Kainz
University of Vienna
Department of Geography and Regional Research
Universitätsstraße 7
Vienna, Austria
wolfgang.kainz@univie.ac.at

Abstract

This article is about GIS-based football game analysis in general and focusses on an approach towards an automated spatiotemporal analysis of scoring attempt patterns based on football-specific geo data, which were kindly provided by *Prozone Sports*. The main objective of this paper is to present our developed *Analyse Scoring Attempts* tool, which corresponds to a *Python* script but is designated for the execution as a custom tool for *ArcGIS for Desktop 10.x*. In this context we attempt to portray the functionality of that tool in as much detail as possible. We also demonstrate that by the application of that tool it is possible to gain the following information automatically: First, the number of players involved in a scoring attempt sequence. Second, the length of that sequence. Furthermore, the duration between the capture of the ball and the final shot. Moreover, both the initial as well as the finishing position of each sequence are intersected with special pitch zones. In addition to that the developed *Spatial Scoring Probability Index* (*SSPI*) is determined for the shooting position. This pure theoretical index indicates the probability of scoring a goal and is composed of weighted values for distance and angle of the shooting position in relation to the goal. Besides, the tool's textual and graphical outcome are presented. Finally, this paper is concluded with a brief discussion about upcoming challenges.

Keywords: GIS-based Football Game Analysis, geographical information systems, sports analytics, scoring attempts, Prozone Sports

1 Prologue and Introduction

This paper was written in the course of the PhD project *GIS-based Football Game Analysis* at the University of Vienna, Austria. In this context, three surveys have already been published. Whereas [5] and [7] provide an adequate overview of the project as well as its main objectives, [6] focusses primarily on the project's data base and how it can be automatically prepared for GIS-applications utilising *ArcGIS for Desktop* developed by *ESRI*.

Today the importance of game analysis in sports can be regarded as undisputed. Since football is a union of space and time [5] we are aware of the potential of GIS in this context which is far from being exhausted. Therefore, we assume that GIS-based applications can and finally will play a key role in the course of professional football game analysis in the future. Our project can be regarded as one step towards this direction.

This study comprises just one particular aspect of our project and can be considered as our first approach to analyse scoring attempts in football games. In this context a tool was developed which has to be considered as a prototype at the time of this paper's submission. We attempted to focus on the spatiotemporal context of goal attempts and their immediately preceding passing sequences. In this context goal attempts are considered as shots on target and shots not on target. Especially by counting the players who are directly involved in individual attacks and further by calculating the time between the ball capture and the shot, assertions about the team's gameplay or style can be derived [4]. Furthermore, the length of the passing sequences which led to shots are also calculated. Moreover, both the positions where the ball was captured and shot are recorded and spatially intersected with special zones of the pitch. In addition, a specific parameter called *Spatial Scoring Probability Index* (SSPI) was developed. It is based on the players' shooting positions and their distance as well as angle to the goal to determine the shooters' scoring probability for each goal attempt.

Although the existing data which represents the gameplay of football games fulfills all requirements of GIS applications [5, 6, 7] corresponding publications out of the field of cartography and GI science are still rare. Therefore, we have to look beyond our discipline's edge which gets increasingly fuzzy because of the interdisciplinary character of our topic. In the following, a brief overview is given about contributions which can be regarded as far related to this study. In [1] passing sequences of the entire game are analysed by counting all passes per player, whereby the pass lengths and the duration of the players' ball possession are taken into account. Furthermore, a potential function was developed which can be used for many purposes. A similar approach to [1] is described in [2] and in [3], whereby the focus lies on network analysis. In contrast to [2], in [3] a social network analysis is conducted in order to evaluate the individual player's performance. Whereas a certain degree of consideration of space and time can be observed in the yet cited studies, [4] just analyse passes, shots and goals in a purely statistical manner. Amongst all not clear GIS-based contributions, the papers of P. Lucey et al. have to be highlighted because of their emphasis on the spatiotemporal character of football. According to this paper, [8, 9, 10] are the most relevant.

The rest of the paper is organised as follows: First, the applied data base is briefly characterised. Then the developed tool for analysing scoring attempts is outlined. Afterwards the tool's preliminary measures are described in detail, such as (1) the detection of the game direction, (2) the preparation of the output data, the detection of ball possession changes (3) and scoring attempts (4) as well as (5) the determination of scoring attempt sequences. Finally, the tools' analysing process is portrayed. Following this, the tool's outcome is presented and the paper will be concluded in its last section where further adaptions are discussed.

2 The Data

As already mentioned, the project's data base is described in more detail in [6]. For this study matching and anonymised tracking and event dataset were utilised. In general, *"…tracking data in football corresponds to tracks gathered as consecutive point data…"* [6], whereas *"…event data represents connected dynamic interactions between the players…"* [6], such as passes and shots, for instance. The data was kindly provided by *Prozone Sports* and represent an entire football game. Applying the developed *Match Data Preparation* tool, which is also described in [6], the data were automatically prepared in a GIS appropriate manner and stored as point feature classes (FC) within a suitable file geodatabase (GDB). In addition, the developed tools *Add Time Index* and *Add Ball Possession Index* have to be executed so that the input data fulfills all requirements of the *Analyse Scoring Attempts* tool which was created for this study.

Although several game datasets are available it is purposeless to utilise all of them in this specific case as it is unknown if for instance *Team A* of a particular game is identical to *Team A* of any other game. Therefore, performance analyses over a couple of games, as conducted in [3], are impossible. However, in the course of our project, we explicitly requested anonymous data as analysing them we are not tempted to subjective interpretations, which are probably inevitable if the players' names are known. Furthermore, at this stage of the project it is more important that the tool is working accurately because this is the main precondition for real applications within the analysis sections of professional football clubs.

3 The Tool and its Input Parameters

To conduct the proposed analyses a *Python* script, which can be run as a custom tool in *ArcGIS*, was written. This is a common practice within our project as we cannot expect that all users are familiar with executing *Python* scripts or *ArcGIS* models. Hence, we strived to keep the analyst's requirement profile at a minimum. Moreover, we can provide all tools comprehensively as a geoprocessing package.

The *Analyse Score Attempts* tool requires five input parameters. First, an input GDB, which corresponds to the *Match Data Preparation* tool's output GDB wherein the original data is stored. Furthermore, the user is able to choose if the tool should be executed for either both teams or just one of them. Moreover, it is possible to choose between a specific half or the entire game. These options have technically been realised with Boolean queries.

The tool's performance depends on several factors such as the applied computer system and software package as well as the total amount of goal attempts during the selected period. In our case, we used a computer based on Windows 7 64 Bit OS with 8GB RAM and an Intel® Core™ i5-3470 CPU with 3.20 GHz. Applying ArcGIS 10.3 for Desktop – Background Geoprocessing (64 Bit) and 32 goal attempts, the tool's execution required 4 minutes and 19 seconds. The stated amount of scoring attempts occurred for the period of an entire game, whereby both teams were considered.

4 Preliminary Measures

Hereafter the tool's specific parts are described. However, no script fragments are provided due to the limitation of space. However, on request the script will be gladly provided.

As an entire GDB is required as an input several measures have to be conducted in advance. First, an output folder for the textual outcome, which shares the same path as the GDB, is created. Then specific feature datasets (FDS) within the GDB, which are separated by half times as well as tracking and event data, have to be located in order to find the necessary FCs. The tool's execution requires the tracking FC of the ball and the event FC where all events are contained. Those FCs could be select manually also of course. However, the indication of a GDB means less effort and guarantees that the correct FCs are selected. Afterwards an output GDB containing two FDS for each half time is created.

Since the teams' game directions are unknown in advance its detection is crucial to ensure a correct analysis. In this particular context the X-coordinate of the home teams' goalkeeper at the very first frame of the first half is taken into account. Similar to the detection of the ball FC, the goalkeeper can be found easily as the players' tactical positions are part of the created FC names. In dependence upon the algebraic sign of the X-value not only the game direction of the home team can be determined for the first half but also for the second half and the away team respectively. It has to be mentioned that the coordinates systems' origin corresponds to the pitch's centre.

Subsequently the ball FC is copied to the specific output FDS and joined attributively with the corresponding event FC via their *Frame* field. This FC is equal to the further processed output FC. First, all features which occurred during an interruption are deleted. Those features can be identified easily as the *Add Time Index* tool append that information to every feature based upon the game events. For example, the event *Out for throw-in* indicates a game interruption until the ball is throw in. Afterwards a new field named *Attack* is appended to the output FC.

In accordance to [9] the segmentation of the ball's trajectory into team specific possession parts is considered as inevitable. Since the *Add Ball Possession Index* tool determines if the ball is either possessed by the home team or the away team for every frame possession changes can be detected without difficulty.

While iterating through the FC twice applying an *UpdateCursor* the *Attack* field is populated with a certain text where possession changes occurred.

Once possession changes are detected, scoring attempts are identified based on the event names *Shot on target* and *Shot not on target*. Again, an *UpdateCursor* is applied.

In the course of the two previous detection processes *Python* lists were created containing the frames of all possession changes and scoring attempts. With that information it is possible to update a *Python* dictionary while iterating over two *for* loops. The dictionary's *key* corresponds to the goal attempt's frame, whereas the *value* matches the start frame of the sequences. Based on the dictionary, all features which do not relate to a scoring attempt sequence are deleted. Moreover, each sequence is numbered, whereby this information is stored in a new field named *Attack_ID*. For both processes an *UpdateCursor* is applied.

5 Analysis Procedure

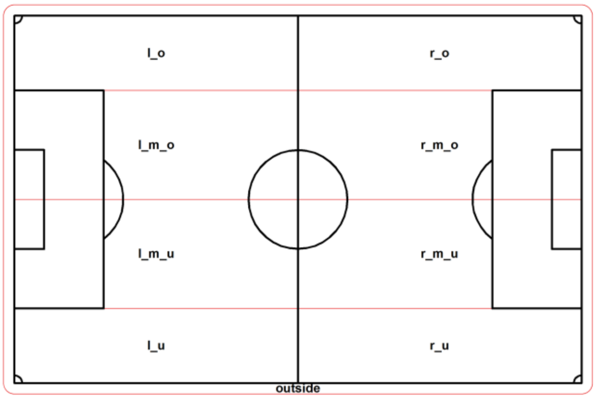
At this time of the tool's execution the output FC is ready for the analysis itself, whose process is described in this section of the paper. Since all relevant events are connected to a player, the number of players involved in a sequence can be counted. For this a *SearchCursor* is applied. A constant total amount of players in such sequences gives some indication of the team's tactical orientation [4]. For instance, if the team favours a fast play in the course of counter attacks or a controlled build-up via several open men. Besides, in [2] it is suggested that the number of consecutive passes indicates a team's elaborateness.

In order to measure the length of the ball's trajectory during a sequence the balls' tracking points have to be linked. For this *ESRI*'s *Points to Line* tool is applied and the fields *Frame* and *Attack_ID* serve as *sort field* and *line field* respectively. The tool's conduction implies the creation of a new line FC which is also stored within the output FDS. The new file contains the length of each sequence and is particularly important for coaches who favour pressing and short attacks.

Since the duration between a ball capture event and the final goal attempt can be limited in a club's game philosophy it is reasonable to gather that information too. In this context the sequences first *Frame* value is subtracted from the last one and finally converted into seconds. This is possible because the data is recorded constantly with 10 fps, which means that just a simple division by 10 is necessary to gain the duration time. In accordance to [1] the measurement of time each player holds the ball during the sequence is also conceivable for future tool revisions.

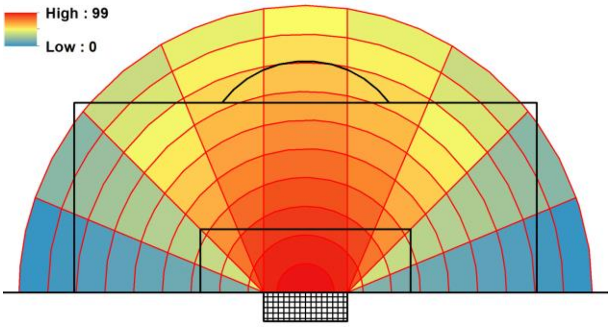
The analysis of scoring attempt patterns demands the identification of the zone where the ball, which was eventually shot, was captured originally. For this particular case the pitch was divided into eight zones as illustrated in Figure 1. Naturally, other zoning alternatives can be utilised such as in [2, 8, 9]. However, a distinction between the halves and within that four another vertical zones seems to be appropriate for that special kind of analysis. This zoning FC represents just one of several other zoning versions that can be optionally created by executing the *Match Data Preparation* tool. This is why it can be automatically taken into account for an intersection with the sequence's initial position.

Figure 1: Applied zoning of the pitch.



The next analysis process does not differ from the last one with the exception of the intersection input. In this particular case a zoning including the developed *SSPI* is applied, as illustrated in Figure 2.

Figure 2: Zoning based on the SSPI.



The *SSPI* indicates the pure theoretical probability of scoring a goal in dependence upon the distance and angle of the shooting position, as outlined in (1).

$$x_d = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$
$$x_a = \{0, 11.25, 33.5, 56.25, 78.75\}$$
$$SSPI = (x_d * 10) * 0{,}33 + (\cos x_a * 100) * 0{,}66 \qquad (1)$$

In this case the shooters' skills cannot be considered. In general, the *SSPI* follows the logic that the closer the distance and the more obtuse the angle the more likely is it to score a goal. In this context the zoning area corresponds to a semicircle inside the pitch with a radius of 25 m starting at the centre of the goal line. The comprising area is further divided by 10 other semicircles with a similar origin but different radii beginning from 2.5 m in equal distances of 2.5 m each. The distribution of distance values follows a linear function, whereby the smallest and largest semicircles have already categorised values of 10 and 100 respectively. In regard of the angles the zones are divided by a rectangle whose long sides are on the levels of the goal posts. From each post the areas on the left and the right side of the goal are further consistently divided into intervals of 22.5°. Unlike the value distribution of the distances the values of the angles follow a cosine function, whereby its x_a values correspond to the sectors' average angles. Considering angle to be more crucial than distance, both ranges are further weighted. To be precise the angle and distance values are multiplied by the factors 0.66 and 0.33 respectively. Finally, the values for each zone are summed.

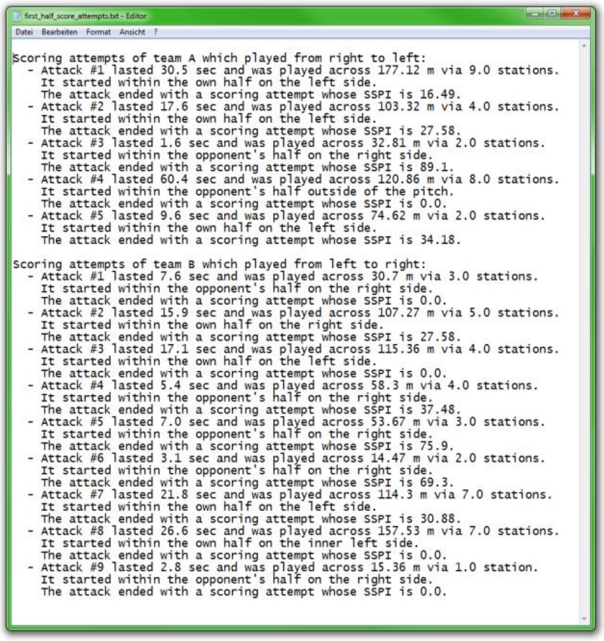
In contrast to that, in [11] scoring probability contours were developed based on a regression model. Admittedly, this report is quite interesting, the validity of its results is questionably though. Since this article was published in 1997 its results might not fit the reality of today's football.

6 Analysis Outcome

Among others, the tool's outcome consists of a textual output, as illustrated in Figure 3. The storage of the statistical output within a text file might be arguable in terms of its practicability. However, it demonstrates that the gained information can be extracted from the input data, which actually is the crucial point.

On the other hand a graphical output is provided in form of the created point and line FCs. This is illustrated in Figure 4, where the scoring attempt sequences of both teams are depicted for the first half time.

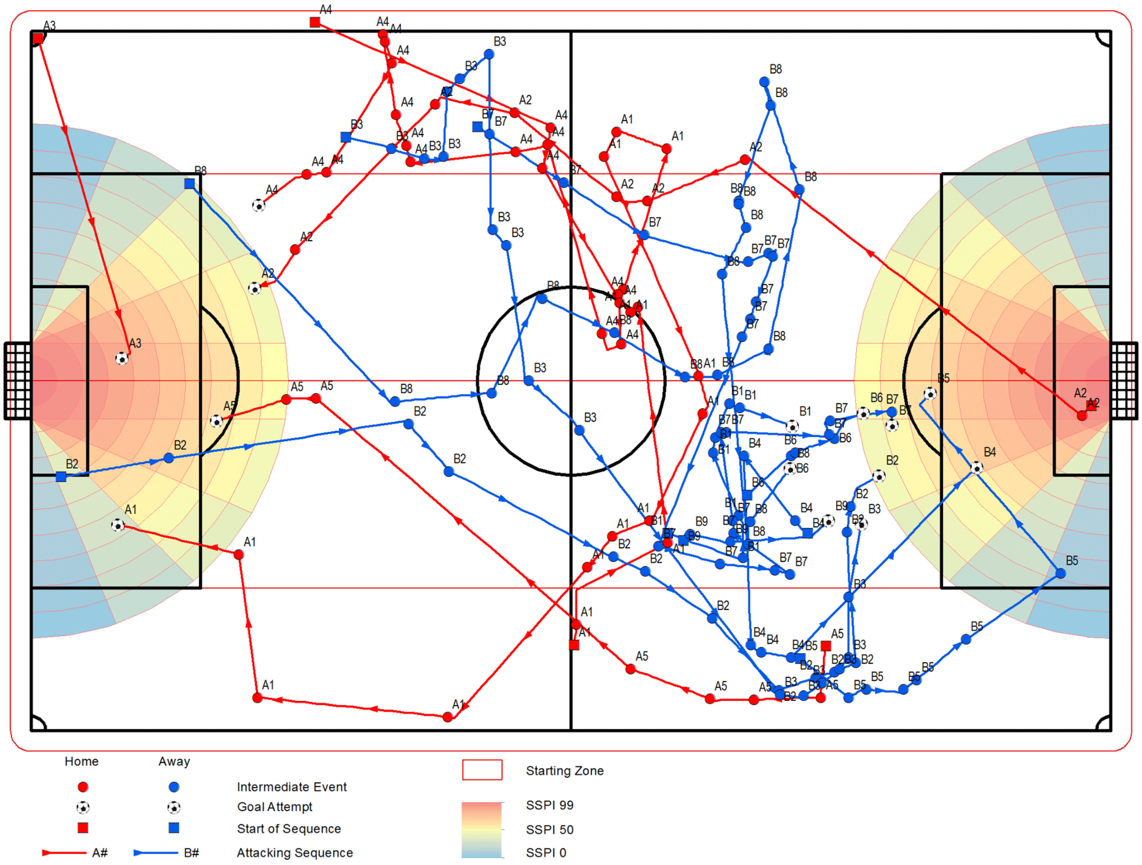
Figure 3: Textual outcome of the *Analyse Scoring Attempts* tool, which corresponds to a statistical summary.



```
Scoring attempts of team A which played from right to left:
  - Attack #1 lasted 30.5 sec and was played across 177.12 m via 9.0 stations.
    It started within the own half on the left side.
    The attack ended with a scoring attempt whose SSPI is 16.49.
  - Attack #2 lasted 17.6 sec and was played across 103.32 m via 4.0 stations.
    It started within the own half on the left side.
    The attack ended with a scoring attempt whose SSPI is 27.58.
  - Attack #3 lasted 1.6 sec and was played across 32.81 m via 2.0 stations.
    It started within the opponent's half on the right side.
    The attack ended with a scoring attempt whose SSPI is 89.1.
  - Attack #4 lasted 60.4 sec and was played across 120.86 m via 8.0 stations.
    It started within the opponent's half outside of the pitch.
    The attack ended with a scoring attempt whose SSPI is 0.0.
  - Attack #5 lasted 9.6 sec and was played across 74.62 m via 2.0 stations.
    It started within the own half on the left side.
    The attack ended with a scoring attempt whose SSPI is 34.18.

Scoring attempts of team B which played from left to right:
  - Attack #1 lasted 7.6 sec and was played across 30.7 m via 3.0 stations.
    It started within the opponent's half on the right side.
    The attack ended with a scoring attempt whose SSPI is 0.0.
  - Attack #2 lasted 15.9 sec and was played across 107.27 m via 5.0 stations.
    It started within the own half on the right side.
    The attack ended with a scoring attempt whose SSPI is 27.58.
  - Attack #3 lasted 17.1 sec and was played across 115.36 m via 4.0 stations.
    It started within the own half on the left side.
    The attack ended with a scoring attempt whose SSPI is 0.0.
  - Attack #4 lasted 5.4 sec and was played across 58.3 m via 4.0 stations.
    It started within the opponent's half on the right side.
    The attack ended with a scoring attempt whose SSPI is 37.48.
  - Attack #5 lasted 7.0 sec and was played across 53.67 m via 3.0 stations.
    It started within the opponent's half on the right side.
    The attack ended with a scoring attempt whose SSPI is 75.9.
  - Attack #6 lasted 3.1 sec and was played across 14.47 m via 2.0 stations.
    It started within the opponent's half on the right side.
    The attack ended with a scoring attempt whose SSPI is 69.3.
  - Attack #7 lasted 21.8 sec and was played across 114.3 m via 7.0 stations.
    It started within the own half on the left side.
    The attack ended with a scoring attempt whose SSPI is 30.88.
  - Attack #8 lasted 26.6 sec and was played across 157.53 m via 7.0 stations.
    It started within the own half on the inner left side.
    The attack ended with a scoring attempt whose SSPI is 0.0.
  - Attack #9 lasted 2.8 sec and was played across 15.36 m via 1.0 station.
    It started within the opponent's half on the right side.
    The attack ended with a scoring attempt whose SSPI is 0.0.
```

Figure 4: Graphical outcome of the *Analyse Scoring Attempts* tool depicting all scoring attempt sequences of one half time



7 Discussion and Conclusion

As football is an amalgamation of spatiotemporal components and phenomena it can be analysed appropriately by taking the spatiotemporal character of the game sufficiently into account. In this particular context the utilisation of GI-technology is worthwhile because of its applicability in the case of football-specific geo data.

This paper demonstrates that GIS can be applied for any kind of spatiotemporal football game analyses. Nevertheless, we have to evaluate the results critically. The presented contents can be regarded as a first approach towards automated GIS-based analysis of scoring attempt patterns. Nevertheless there are three crucial questions left. First: Are the gained information relevant for game analysts? Second: How can we derive more information from the input data? And finally: How can we connect these results to the outcomes of other analyses? Although some progress has been made so far there is still a lot of work ahead.

The answer to the first two questions definitely requires a close cooperation with both sports scientists and professional game analysts. However, at the time of the paper's submission an adaptation of the tool is in progress. Thereby the focus is on the players' position within the shooting field which corresponds to a triangle between the shooter's position and the goal posts. By taking those potential human barriers into account the *SSPI* decreases and the shot quality can be determined. By contrast, the third question is more a technical one, primarily a concern for GI experts.

Due to those unresolved issues we have strived to modify our tools based on external expert knowledge, which is why certain input parameters such as more small scaling zoning options can be altered or exchanged easily. Furthermore, the textual output can be provided in a different form of course, e.g. as a table, a diagram, etc. In regard of the graphical output different forms of static and even dynamic visualisations are definitely conceivable.

As already stated, there is still a lot of research to do. In any form whatsoever, the key issue of the future is the cooperation not only within our own scientific community but also beyond that.

References

[1] D. R. Brillinger. A Potential Function Approach to the Flow of Play in Soccer. *In Journal of Quantitative Analysis in Sports*, Vol. 3, Issue 1, Article 3, 2007

[2] C. Cotta and A. M. Mora, J. J. Merelo, C. Merelo-Molina. A Network Analysis of the 2010 FIFA World Cup Champion Team Play. In *Journal of Systems Science and Complexity*, Issue 26, pages 21-42, 2013

[3] J. Duch and J. S. Waitzman, L. A. Nunes Amaral. Quantifying the Performance of Individual Players in a Team Activity. In *PLoS ONE*, Vol. 5, Issue 6, 2010

[4] M. Hughes and I. Franks. Analysis of Passing Sequences, Shots and Goals in Soccer. In *Journal of Sports Sciences*, Issue 23(5), pages 509-514, 2005

[5] G. Kotzbek and W. Kainz. Football Game Analysis: A New Application Area for Cartographers and GI-Scientists? In *Proceedings*, *Vol.1 and Vol.2 of the 5th International Conference on Cartography and GIS*, Riviera, pages 299-306, 2014

[6] G. Kotzbek and W. Kainz. GIS-based Football Game Analysis – A Brief Introduction to the Applied Data Base and a Guideline on How to Utilise It. *In Proceedings of the 27th International Cartographic Conference*, Rio de Janeiro, 2015

[7] G. Kotzbek and W. Kainz. Das Runde muss ins GIS – Neue Wege im Bereich der Fußball-Spielanalyse. In *gis.Science – Die Zeitschrift für Geoinformatik*, Issue 03/2015, pages 117-124, 2015

[8] P. Lucey and A. Bialkowski, P. Carr, E. Foote, I. Matthews. Characterizing Multi-Agent Team Behavior from Partial Team Tracing: Evidence from the English Premier League. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 1387-1393, Toronto, 2012

[9] P. Lucey and D. Oliver, P. Carr, J. Roth, I. Matthews. Assessing Team Strategy using Spatiotemporal Data. Accepted paper to the *ACM SIGKDD Conference on Knowledge, Discovery and Data Mining (KDD)*, Chicago, 2013

[10] P. Lucey and A. Bialkowski, M. Monfort, P. Carr, I. Matthews. "Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data. Accepted paper to the *MIT Sloan Sports Analytics Conference (MITSSAC)*, Boston, 2015

[11] R. Pollard and C. Reep. Measuring the effectiveness of playing strategies at soccer. In *The Statistician* 46, No. 4, pages 541-550, 1997