Project 1: Impact Factor
STAT 497 – Sports Analytics
Dr. Joshua Wyatt Smith
Nikolas Argiropoulos 40044358
Taniel Migdesyan 40060644
Concordia University

1. Introduction

Over the past decade, data driven decisions have extended beyond the medical and financial world, and into sports (Parra et al., 2022). The use of data and mathematical modeling to determine the best possible outcome of a match or strategy is fairly novel. Sports, like hockey and American football (NFL), have adapt various machine learning techniques to revolutionize their game planning where teams that are not data driven are at a great disadvantage. For example, NFL teams greatly rely on probability theory to determine the outcome of success of going on 4^{th} down or kicking a field-goal/punt, evidently increasing, or decreasing the probability of a win (Williams et al., 2022). Due to the stagnate nature of the NFL, probabilities regarding the success of the next play or strategy could be directed to the coach or players prior to the play. In other words, set plays with the highest level of success could be determined seconds before the play begins. This unique approach cannot be done in soccer due to the fluidity of the game. Many strategies and set plays in soccer must be studied and practiced prior to a game, this includes their associated probabilities of success. There are many metrics in which soccer analyst uses to determine which set plays to use, what to focus on in future practices and games, and the weakness of the offense/defence of the opposing team. There are two main metrics sports analysts gather; tracking data such as distance and speed covered or event data like passes or shots (Smith, 2022). For this project, we will be focusing on passes, more specifically the impact factor of said passes.

Impact factor (*IF*) is defined as the increase or decrease of the significance of a pass to a teammate. In general, impact factor can be operationalized as the summation of opposing player the ball outplays. For example, if player P_1 passes the ball to player P_2 in the positive x-direction and the ball outplays play M_1-M_4 then the impact factor would be 4 (Smith, 2022). The aim of the project was to determine the advantages and disadvantages of the *IF* definition and propose a new definition.

2. Methods

2.1. *Advantages of the current Impact Factor*

The current definition of *IF* encompasses the over goal of soccer; to out-pass your opponent in an efficient manner to maximize the number of goals scored. By summing the amount of out-pass that occur, you can approximate the player and team *IF* to generate scoring chances.

2.2. *Disadvantages of the current Impact Factor*

2.2.1. *Simple Definition*

The simplicity of the current definition of *IF* is too minimal for the complexities of soccer. Soccer is a game with high amount of passing, low amounts of quality shots and even fewer goals. To use the current definition oversimplifies the importance of a pass to create a high quality chance at goal.

2.2.2. *Importance of proximity to the net*

To attribute an impact based on the amount of opposing player a ball out-plays can be generalized in the offensive and defensive zones. For example, a pass that out-plays 4 players in the box that results in a goal would be equivalent to a pass that out-plays 4 players in center field. Both scenarios result in an *IF* of 4, however the former resulted a goal whereas the latter results in a prevalent pass. Thus, the current definition does not account for where the pass occurred.

2.2.3. *Back passing*

Passing back to the keeper or the initial player is quite common in soccer. In fact, the majority passes occur within a short proximity of one another. Due to the current definition not accounting for back-passing or passing within a short proximity, summation of *IF* could stack quite quickly, thus inflating the players and teams total *IF*.

2.3. New Impact Factor

2.3.1. *Defining and dimensions of the pitch and Zones*

The dimensions of the field follow the Metrica dataset #2 provided by Dr. Smith. The official coordinates for the pitch are (0,0) is the top left, (1,1) is the bottom right, and (0.5,0.5) is center field (Metrica, 2022). See figure 1 for Zone A-F and their respective *initial IF*'s. Zones were determined based on the amount of goals scored and the probability of scoring within their respected area (Figure 1) (Simiyu, 2013).

2.3.2. *New Impact Factor Definition*

We defined *Grand IF* as the sum of pass (*initial IF*) or progression towards high value zones within the X proximity of the goal, the result of a shot (*potential assist*) and the total amount of interceptions. In other words, if player P_1 passes the ball to player P_2 in the positive x-direction within Zone A, the *initial IF* would be considered a 1 for that pass. If a successive pass remains within the same Zone, no increase or decrease in *IF* will occur. However, if a back-pass occur into a lower valued Zone, the *initial IF* will be calculated as the subtraction of the Zone where the pass ended by the original Zone. For example, if player P_1 is located in Zone B and passes the ball to player P_2 in Zone A the resulting *initial IF* for that pass would be $IF_{A-B} = Zone\ A - Zone\ B = 0 - 2 = -0$. A negative *IF* should not be interpreted as uninfluential or non-important pass, but rather the negative indicates a direction, and the numerical value indicates a pass to a lower goal scoring area. A *potential assist* was operationalized as +2 *IF* and a pass that resulted in an *interception passes* was coded as -2 *IF*. Mathematically, individual *Grand IF* is calculated as $IF_i = \sum_{x=A}^{G}\sum_{k=1}^{n} Zone\ x_k + \sum_{m=1}^{n} PA_m - \sum_{j=i}^{n} interceptions_j$, where i: player number, X: Zone, AA: potential assist (passes that result in a shot). The subscripts k, m and j indicate the number of passes. *Team IF* by period is calculated as $IF_{p,t} = \sum_{i}^{10} IF_i, where\ p\ is\ the\ period\ and\ t\ is\ the\ team$ and the *total team IF* is calculated as $IF_{Team} = \sum_{t}^{2} IF_{p,t}\ \ where\ p\ is\ the\ period\ and\ t\ is\ the\ team.$

2.3.3. *Advantages New Impact Factor Definition*

2.3.3.1. *Accounting for back-passing*

Under the previous definition of *IF*, back-passing was not incorporated into the scoring system. The new *IF* scoring system scores back-passing as negative values. This method allows analysts to infer which players pass back more often than passes forward or pass to a lower scoring area.

2.3.3.2. *Zonal Passing*

The old *IF* metric did not account for passing within Zones. Under the new definition, any additional pass that occurs within the receiving Zone will not receive any addition *initial IF* points (i.e. if a pass is made into Zone B, that pass will receive an *initial IF* = 1, however if the proceeding pass remains within Zone B no addition *initial IF* points will be given). The logic behind this is to eliminate rewarding parallel passing that does not lead to significant forward passing. In addition, analyst could view if the *initial IF* increased within a certain time frame. In other words, if the *initial IF* remains consistent for an extended period of time, this suggests parallel passing and no significant forward movement of the ball.

2.3.3.3. *Rewarding proximity to the net*

Lastly, the current definition incorporates proximity to the net. As pass occur closer to the net, the probability of a goal increase. Thus, passes closest to the net are scored higher than passes further from the net.

2.3.4. *Limitations New Impact Factor Definition*

2.3.4.1. *Opposing Player proximity*

A limitation of this definition is it does not integrate the proximity of opposing players. A future project could incorporate this metric as the difficulty of passing increases as players converge to the net due to the close distance between opposing teams and the goal.

2.3.4.2. *Back-passes*

Although we include back-passing as an advantage in our definition, back-passing can be efficient to set up open players in advantageous zones. For example, if player P_1 pass the ball back to player P_2, player P_2 can then pass it to player P_3 who is located near the goal. This sequence of events is not accounted for in the model.

2.3.4.3. *Amount/ locations of Zones*

Our definition only include 7 zones, however, future research could look into if more zone and non-uniform locations with varying *IF*'s could be more accurate in determining individual and team *IF*'s.

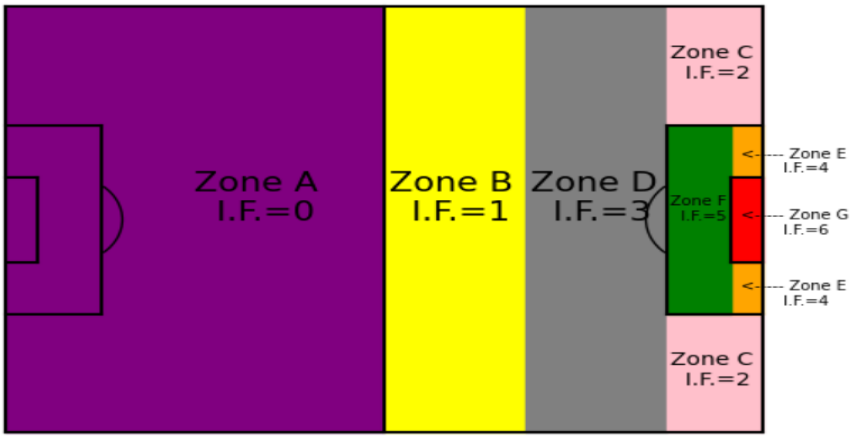

Figure 1. *Description of Zones and Impact Factors*

	Period	From	Initial I.F.	Intercepted Passes	Potential Assists	Grand I.F.
0	1	Player1	6.0	1	1	6.0
0	1	Player2	3.0	2	0	-1.0
0	1	Player3	4.0	3	0	-2.0
0	1	Player4	20.0	1	0	18.0
0	1	Player5	3.0	4	0	-5.0
0	1	Player6	0.0	1	0	-2.0
0	1	Player7	-2.0	2	0	-6.0
0	1	Player8	1.0	4	1	-5.0
0	1	Player9	-3.0	4	1	-9.0
0	1	Player10	-2.0	5	1	-10.0
0	2	Player1	1.0	4	2	-3.0
0	2	Player2	3.0	0	0	3.0
0	2	Player3	1.0	1	0	-1.0
0	2	Player4	8.0	4	0	0.0
0	2	Player5	-1.0	6	1	-11.0
0	2	Player6	11.0	4	0	3.0
0	2	Player7	-1.0	4	0	-9.0
0	2	Player8	1.0	5	1	-7.0
0	2	Player9	-2.0	4	3	-4.0
0	2	Player10	-3.0	1	0	-5.0

Figure 2. *Team 1 Data*

	Period	From	Initial I.F.	Intercepted Passes	Potential Assists	Grand I.F.
0	1	Player15	1.0	5	0	-9.0
0	1	Player16	3.0	5	0	-7.0
0	1	Player17	4.0	1	0	2.0
0	1	Player18	6.0	5	1	-2.0
0	1	Player19	-1.0	3	0	-7.0
0	1	Player20	14.0	3	0	8.0
0	1	Player21	7.0	3	0	1.0
0	1	Player22	-5.0	3	1	-9.0
0	1	Player23	-3.0	2	1	-5.0
0	1	Player24	0.0	4	0	-8.0
0	2	Player15	-2.0	4	0	-10.0
0	2	Player16	7.0	3	0	1.0
0	2	Player17	6.0	5	0	-4.0
0	2	Player18	3.0	1	1	3.0
0	2	Player19	-2.0	2	0	-6.0
0	2	Player20	7.0	4	0	-1.0
0	2	Player21	1.0	2	0	-3.0
0	2	Player22	1.0	0	1	3.0
0	2	Player23	3.0	0	1	5.0
0	2	Player24	-5.0	3	1	-9.0

Figure 3. *Team 2 Data*

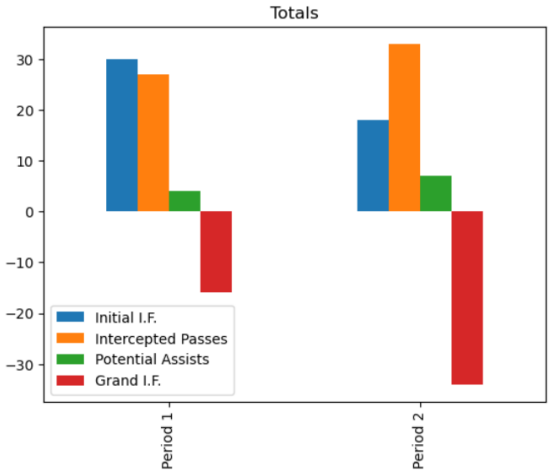

Figure 4. *Team 1 Total Score*

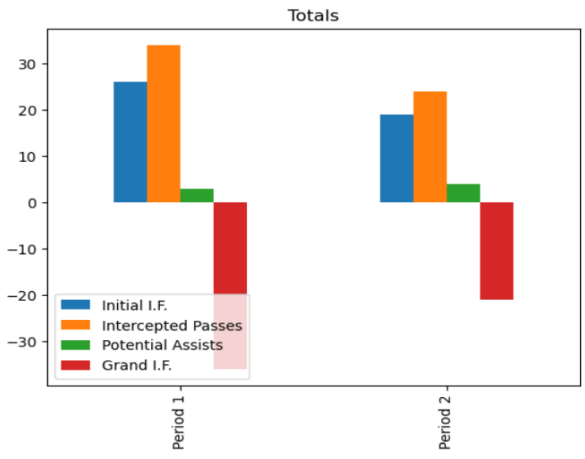

Figure 5. *Team 2 Total Score*

Results/Discussion

Total and means (see code for means) for each player and team were calculated, and partitioned by period and to compare *initial IF*, *interceptions*, *potential assists* and *grand IF*. When comparing the metrics total, we notice that Team 1 and Team 2 do not differ drastically. There is a maximum difference of 7 *IF* scores. Partitioning by period we see that Team 1 has less total interceptions in period 1 compared to Team 2 (27vs.34). when comparing Grand *IF* we see that Team 1 has -16 compared to Team 2's -36. These values suggest that Team 1 was passing more forward than Team 2 in the first period. However, these metrics change in period 2 as Team 2 gave up less interception (24) and had a larger Grand *IF* (-21) than Team 1 (Inter=33, Grand *IF*=-34). One reason for this is Team 2 was behind by a goal in period 2, thus they were trying to pass the ball further into high valued areas. Furthermore, by our definition, we see that Team 1 has a Grand *IF* of -50 and Team 2 has a Grand *IF* of -57. These results suggest that Team 2 had less total forward passing than Team 1, which in turn led to less *potential assists*. Team 1 had 4 more *potential assists* and 3 more *initial IF* score compared to Team 2 (T1: AA=11 *initial IF*=48 , T2: AA=7 *initial IF*=45). This suggests that Team 1 has attempted to pass and progress towards high valued zones thus leading to greater chances of scoring a goal. The results of the match further suggest the accuracy of the new definition. The match concluded with Team 1 winning 3-2 over Team 2. Due to the close final score and Team 1's slight increase in *potential assists*, *initial IF* and *Grand IF* may suggest that the new *IF* definition can describe a possible link between the final outcome and *IF* metrics. However, more research and model refining is needed.

References

The first seventeen "In[1]-In[17]" on Jupyter notebook were taking from https://github.com/wyatt-ai/Project1/blob/master/example_notebook.ipynb

Drawing a pitchmap - adding lines & circles in Matplotlib. FC Python. (2020, December 18). Retrieved October 6, 2022, from https://fcpython.com/visualisation/drawing-pitchmap-adding-lines-circles-matplotlib

Metrica. (2022). About the data. Github.

Parra, X., Tort-Martorell, X., Alvarez-Gomez, F., & Ruiz-Viñals, C. (2022). Chronological evolution of the information-driven decision-making process (1950–2020). *Journal of the Knowledge Economy*. https://doi.org/10.1007/s13132-022-00917-y

Simiyu, W. N. (2013). Analysis of goals scored in the 2010 world cup soccer tournament held in South Africa. *Journal of Physical Education and Sport*, *13*(1), 6–13. https://doi.org/10.7752/jpes.2013.01002

Smith, W. J. (2022). In-class Notes.

Williams, B., Palmquist, W., & Elmore, R. (2022). Simulation-based decision making in the NFL using nflsimulator. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04524-7