Soccer Analytics with Two Pieces of Paper and a Pencil

Michael A. Rutter, Ph.D.

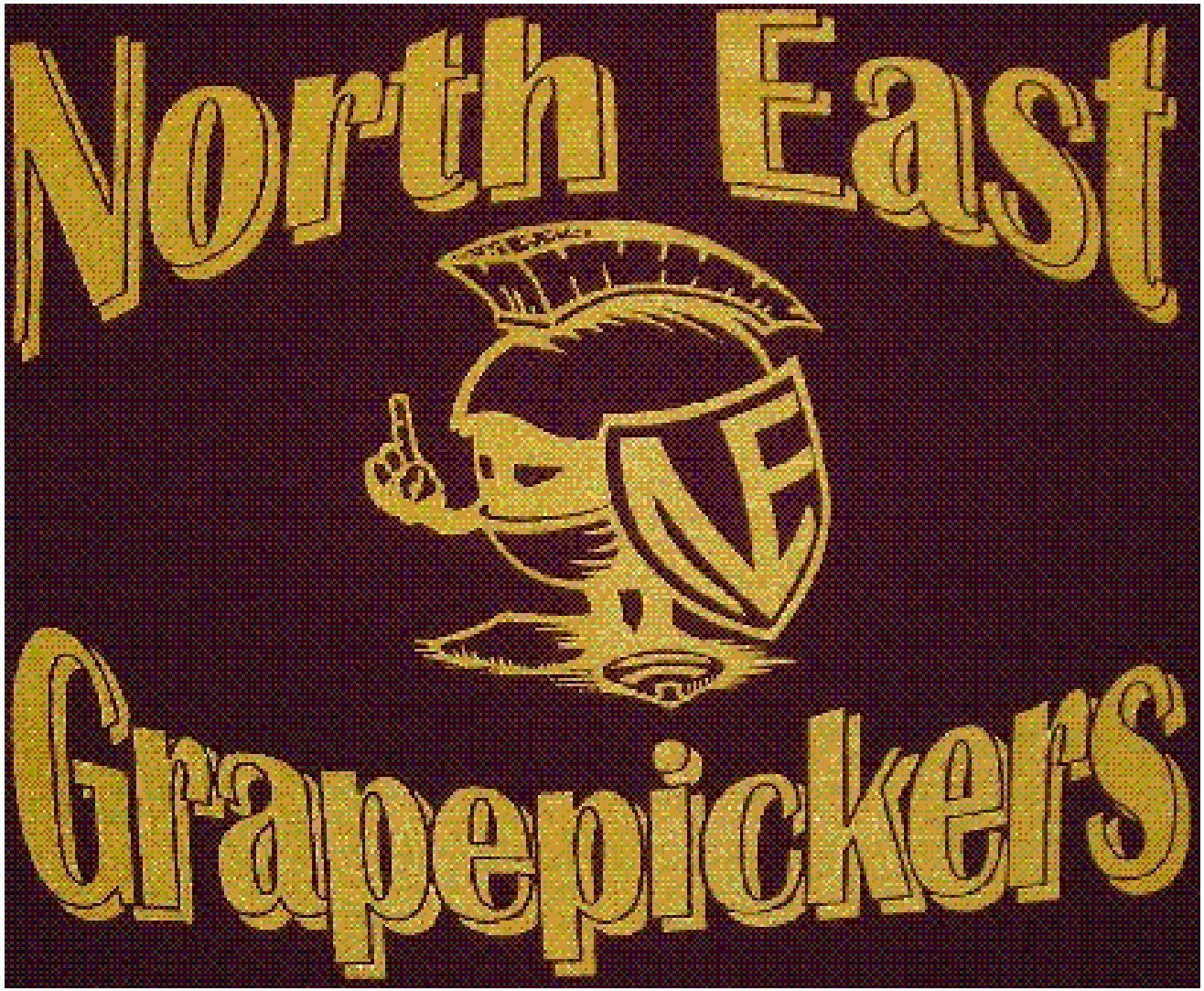
Associate Professor of Statistics
Associate Director, School of Science
Penn State Behrend

JSM 2019 (7/31/19)

Alternative Titles

- “What Happens When a Statistician is Asked to do Statistics for a Varsity Soccer Team”
- “Please Stop Shooting From There”

North East Grapepickers



North East Grapepickers

- Located in NE Erie County in NW Pennsylvania
- 10 wins, 8 losses, and 2 ties in 2018
- Region 4 co-champions, Lost in District 10 2A semi-finals
- Budget for statistics: $0

North East Grapepickers

- Located in NE Erie County in NW Pennsylvania
- 10 wins, 8 losses, and 2 ties in 2018
- Region 4 co-champions, Lost in District 10 2A semi-finals
- Budget for statistics: $0

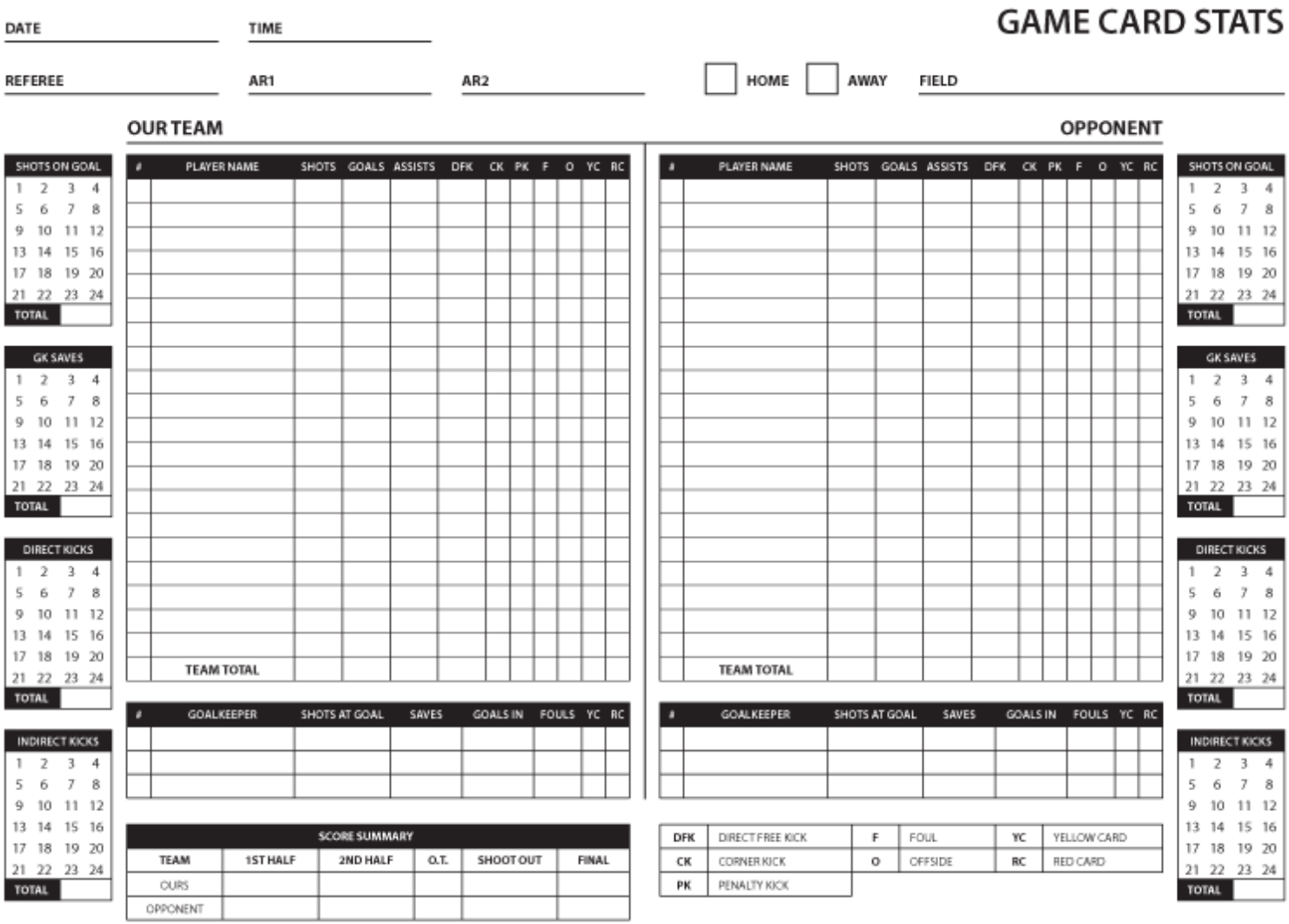
North East Grapepickers

- Located in NE Erie County in NW Pennsylvania
- 10 wins, 8 losses, and 2 ties in 2018
- Region 4 co-champions, Lost in District 10 2A semi-finals
- Budget for statistics: $0

North East Grapepickers

- Located in NE Erie County in NW Pennsylvania
- 10 wins, 8 losses, and 2 ties in 2018
- Region 4 co-champions, Lost in District 10 2A semi-finals
- Budget for statistics: $0

Typical Soccer Statistics



GAME CARD STATS

DATE _____ TIME _____

REFEREE _____ AR1 _____ AR2 _____ ☐ HOME ☐ AWAY FIELD _____

	OUR TEAM											OPPONENT										
#	PLAYER NAME	SHOTS	GOALS	ASSISTS	DFK	CK	PK	F	O	YC	RC	#	PLAYER NAME	SHOTS	GOALS	ASSISTS	DFK	CK	PK	F	O	YC	RC
	TEAM TOTAL												TEAM TOTAL										

SHOTS ON GOAL				
1	2	3	4	
5	6	7	8	
9	10	11	12	
13	14	15	16	
17	18	19	20	
21	22	23	24	
TOTAL				

GK SAVES				
1	2	3	4	
5	6	7	8	
9	10	11	12	
13	14	15	16	
17	18	19	20	
21	22	23	24	
TOTAL				

DIRECT KICKS				
1	2	3	4	
5	6	7	8	
9	10	11	12	
13	14	15	16	
17	18	19	20	
21	22	23	24	
TOTAL				

INDIRECT KICKS				
1	2	3	4	
5	6	7	8	
9	10	11	12	
13	14	15	16	
17	18	19	20	
21	22	23	24	
TOTAL				

#	GOALKEEPER	SHOTS AT GOAL	SAVES	GOALS IN	FOULS	YC	RC

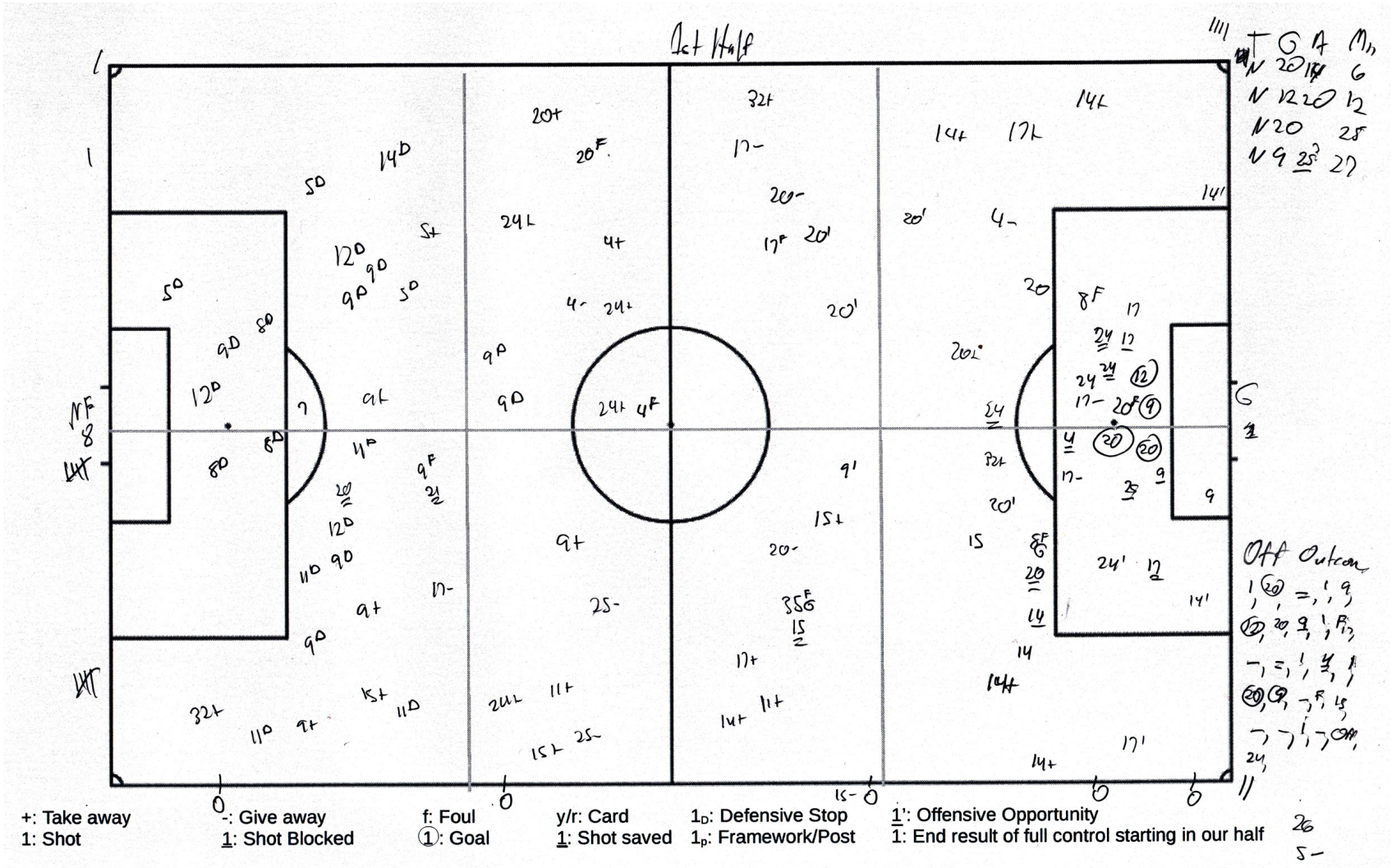
#	GOALKEEPER	SHOTS AT GOAL	SAVES	GOALS IN	FOULS	YC	RC

SCORE SUMMARY					
TEAM	1ST HALF	2ND HALF	O.T.	SHOOT OUT	FINAL
OURS					
OPPONENT					

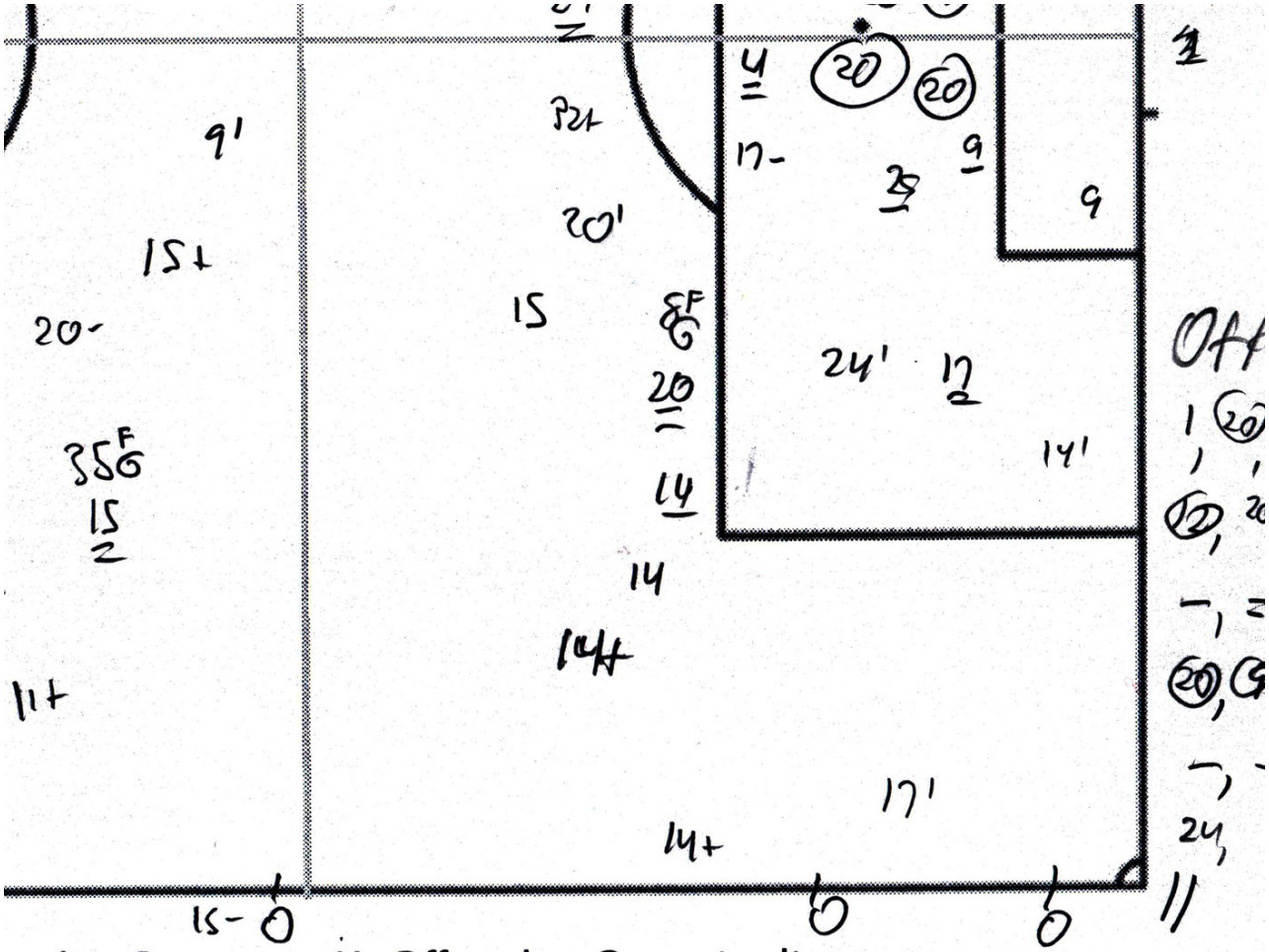
DFK	DIRECT FREE KICK	F	FOUL	YC	YELLOW CARD
CK	CORNER KICK	O	OFFSIDE	RC	RED CARD
PK	PENALTY KICK				

https://www.brantwojack.com

My Version


1st Half
+: Take away
1: Shot
-: Give away
1: Shot Blocked
f: Foul
①: Goal
y/r: Card
1: Shot saved
1D: Defensive Stop
1P: Framework/Post
1': Offensive Opportunity
1: End result of full control starting in our half

My Version


ensive Stop
1'': Offensive Opportunity
nework/Post
1: End result of full control starting in our half

Obtaining the Data

- Digitized each shot recorded for season
- Labeled “Miss” or “Goal”, team, and game number
- Did not distinguish type of shot (header, etc.)
- Used “WebPlotDigitizer” (`https://automeris.io/WebPlotDigitizer`)
- Manipulated data using R (github: marutter/NEsoccer)

Obtaining the Data

- Digitized each shot recorded for season
- Labeled “Miss” or “Goal”, team, and game number
- Did not distinguish type of shot (header, etc.)
- Used “WebPlotDigitizer” (`https://automeris.io/WebPlotDigitizer`)
- Manipulated data using R (github: marutter/NEsoccer)

Obtaining the Data

- Digitized each shot recorded for season
- Labeled “Miss” or “Goal”, team, and game number
- Did not distinguish type of shot (header, etc.)
- Used “WebPlotDigitizer” (https://automeris.io/WebPlotDigitizer)
- Manipulated data using R (github: marutter/NEsoccer)

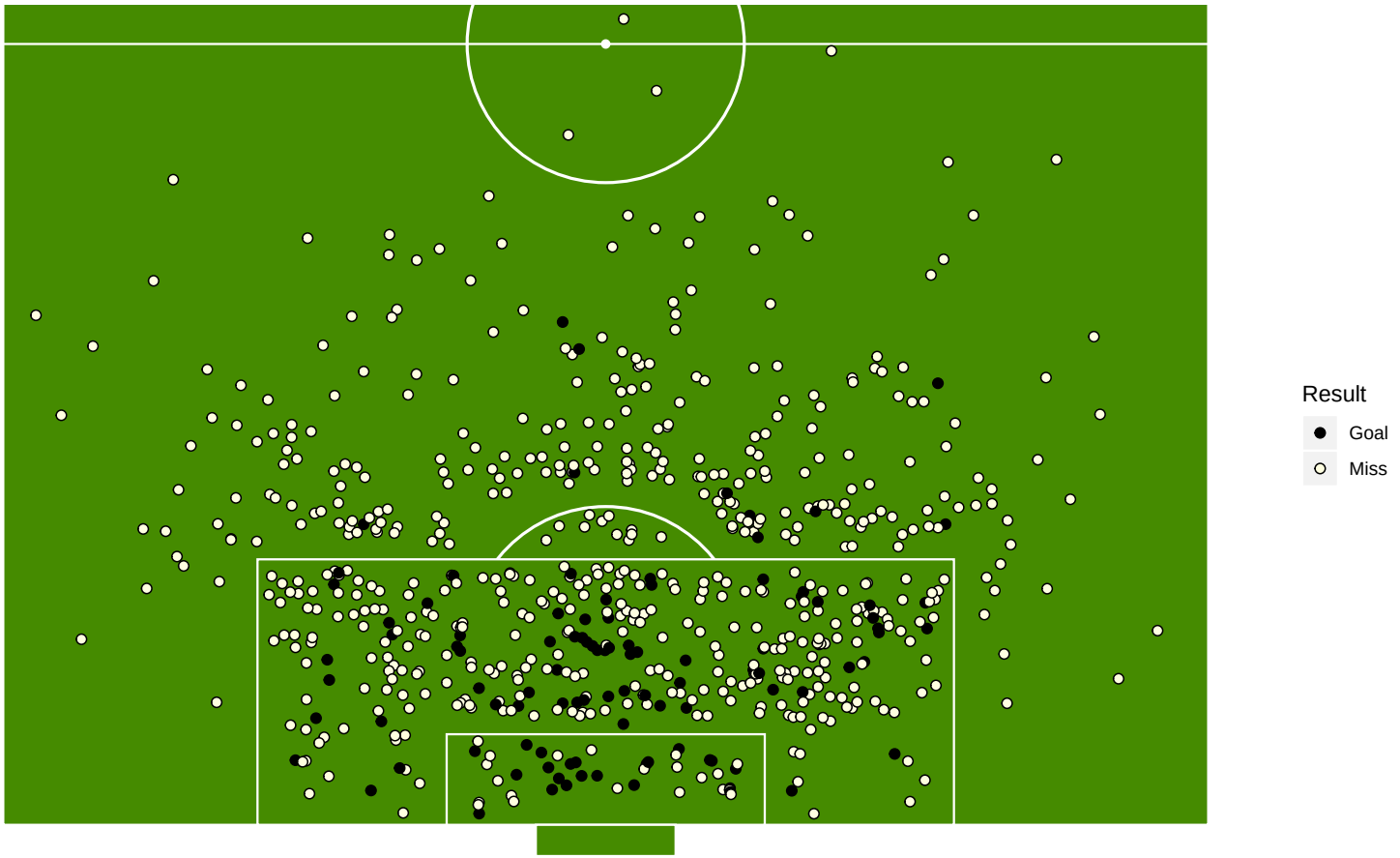
Obtaining the Data

- Digitized each shot recorded for season
- Labeled “Miss” or “Goal”, team, and game number
- Did not distinguish type of shot (header, etc.)
- Used “WebPlotDigitizer” (`https://automeris.io/WebPlotDigitizer`)
- Manipulated data using R (github: marutter/NEsoccer)

Obtaining the Data

- Digitized each shot recorded for season
- Labeled “Miss” or “Goal”, team, and game number
- Did not distinguish type of shot (header, etc.)
- Used “WebPlotDigitizer” (`https://automeris.io/WebPlotDigitizer`)
- Manipulated data using R (github: marutter/NEsoccer)

All Shots Recorded



Initial Results

- Of 680 shots, 101 were goals (14.9% success rate)
- This ignores location
- Divide the pitch into six zones
- First introduced by Jacob Beckett (@jacobbeckett22) on the “American Soccer Analysis” blog

Initial Results

- Of 680 shots, 101 were goals (14.9% success rate)
- This ignores location
- Divide the pitch into six zones
- First introduced by Jacob Beckett (@jacobbeckett22) on the “American Soccer Analysis” blog

Initial Results

- Of 680 shots, 101 were goals (14.9% success rate)
- This ignores location
- Divide the pitch into six zones
- First introduced by Jacob Beckett (@jacobbeckett22) on the “American Soccer Analysis” blog

Initial Results

- Of 680 shots, 101 were goals (14.9% success rate)
- This ignores location
- Divide the pitch into six zones
- First introduced by Jacob Beckett (@jacobbeckett22) on the “American Soccer Analysis” blog

Scoring Zones

Percentages by Scoring Area

Zone	Goals	Misses	Goal %	MLS Goal %
1	22	21	51.2%	31.1%
2	42	128	24.7%	17.7%
3	27	154	14.9%	7.1%
4	7	136	4.9%	5.3%
5	3	95	3.1%	2.3%
6	0	45	0.0%	3.5%

Lack of Independence

- This data set is much smaller than the MLS data set (680 vs. 8335 shots)
- Results are correlated based on the team
- Estimated the probability of a goal using a mixed logistic model
- All shots by the same team in the same game were assumed correlated

Lack of Independence

- This data set is much smaller than the MLS data set (680 vs. 8335 shots)
- Results are correlated based on the team
- Estimated the probability of a goal using a mixed logistic model
- All shots by the same team in the same game were assumed correlated

Lack of Independence

- This data set is much smaller than the MLS data set (680 vs. 8335 shots)
- Results are correlated based on the team
- Estimated the probability of a goal using a mixed logistic model
- All shots by the same team in the same game were assumed correlated

Lack of Independence

- This data set is much smaller than the MLS data set (680 vs. 8335 shots)
- Results are correlated based on the team
- Estimated the probability of a goal using a mixed logistic model
- All shots by the same team in the same game were assumed correlated

Updated Percentages

Zone	Goals	Misses	Goal %	Mixed Goal %
1	22	21	51.2%	54.0%
2	42	128	24.7%	24.9%
3	27	154	14.9%	15.1%
4	7	136	4.9%	5.0%
5	3	95	3.1%	3.0%
6	0	45	0.0%	0.0%

Expected Goals

- Especially for high school soccer teams, shots attempted may not be a good metric of the quality of offensive play
- Ten shots from zones five and six are not the same as ten shots from zones one and two
- Given the number of shots per zone, the *expected goals* can be calculated
- $E_g = \sum_{z=1}^{6} p_z n_z$

Expected Goals

- Especially for high school soccer teams, shots attempted may not be a good metric of the quality of offensive play
- Ten shots from zones five and six are not the same as ten shots from zones one and two
- Given the number of shots per zone, the *expected goals* can be calculated
- $E_g = \sum_{z=1}^{6} p_z n_z$

Expected Goals

- Especially for high school soccer teams, shots attempted may not be a good metric of the quality of offensive play
- Ten shots from zones five and six are not the same as ten shots from zones one and two
- Given the number of shots per zone, the *expected goals* can be calculated
- $E_g = \sum_{z=1}^{6} p_z n_z$

Expected Goals

- Especially for high school soccer teams, shots attempted may not be a good metric of the quality of offensive play
- Ten shots from zones five and six are not the same as ten shots from zones one and two
- Given the number of shots per zone, the *expected goals* can be calculated
- $E_g = \sum_{z=1}^{6} p_z n_z$

Some Specific Results

Game vs. Conneaut, Ohio

Zone	1	2	3	4	5	6
Shots	2	6	9	9	8	2

- $E_G = 4.5$ (from 36 shots)
- Actual score: 2-2 (Conneaut had 6 shots)

Some Specific Results

Game vs. Conneaut, Ohio

Zone	1	2	3	4	5	6
Shots	2	6	9	9	8	2

- $E_G = 4.5$ (from 36 shots)
- Actual score: 2-2 (Conneaut had 6 shots)

Some Specific Results

Game vs. Titusville

Zone	1	2	3	4	5	6
Shots	3	6	7	1	1	0

- $E_G = 4.2$ (from 18 shots)
- Actual score: 4-3 (Tittusville had $E_G = 1.7$)

Some Specific Results

Game vs. Titusville

Zone	1	2	3	4	5	6
Shots	3	6	7	1	1	0

- E_G = 4.2 (from 18 shots)
- Actual score: 4-3 (Tittusville had E_G = 1.7)

Conclusions

- As in professional soccer, all shots in high school soccer are not the same quality
- High school players suffer from “selection bias” in terms of long distance shots due to highlight packages
- Expected goals based on zones useful data for both strategy discussions and post-game analysis
- Data set is small, but this shows usefulness of data collected by hand
- Questions?

Conclusions

- As in professional soccer, all shots in high school soccer are not the same quality
- High school players suffer from “selection bias” in terms of long distance shots due to highlight packages
- Expected goals based on zones useful data for both strategy discussions and post-game analysis
- Data set is small, but this shows usefulness of data collected by hand
- Questions?

Conclusions

- As in professional soccer, all shots in high school soccer are not the same quality
- High school players suffer from “selection bias” in terms of long distance shots due to highlight packages
- Expected goals based on zones useful data for both strategy discussions and post-game analysis
- Data set is small, but this shows usefulness of data collected by hand
- Questions?

Conclusions

- As in professional soccer, all shots in high school soccer are not the same quality
- High school players suffer from “selection bias” in terms of long distance shots due to highlight packages
- Expected goals based on zones useful data for both strategy discussions and post-game analysis
- Data set is small, but this shows usefulness of data collected by hand
- Questions?

Conclusions

- As in professional soccer, all shots in high school soccer are not the same quality
- High school players suffer from “selection bias” in terms of long distance shots due to highlight packages
- Expected goals based on zones useful data for both strategy discussions and post-game analysis
- Data set is small, but this shows usefulness of data collected by hand
- Questions?