Sean Duggan and Thomas Stephens
Professor Pekelis
MCS 100/Stats 50
December 9th, 2014

Modeling El Clásico:

Real Madrid's Passing Trends and Corresponding Insights

As the name suggests, soccer aficionados and casual fans alike hold El Clásico (translation: "The Classic") to the highest of soccer standard. Famed rivals FC Barcelona and Real Madrid CF face off in the world's top league, Spain's La Liga. The star studded affair pits World Players of the Year, international superstars, and legendary mangers against one another for what's always to be an instant classic. Naturally, El Clásico offers an ideal setting for soccer purists who covet it as "the beautiful game," and not one to defined by numbers. Viewpoints such as this have traditionally left statistics with a sigma of being more of a distraction as opposed to a focal point. However, the sports statistics wave has not missed the world's most popular game. Numerous companies and websites now develop complete team strategies, percentages, and player zones. While emerging recognition of importance has created a market for broad team statistics, we feel as though they do not tell the entire story.

Anyone who has watched a soccer game knows that passing is the lifeblood of the sport, yet—save assists and completions—passing is as underreported as any category in the sport. Hundreds of passes per game statistically validate questions such as: "which passing interaction is most effective in leading to a goal?" or "which player is most a least effective in converting passing to goals?" With questions like these in mind, we chose Real Madrid in El Clásico to

analyze for two reasons: 1) Real is the most prolific scoring machine in soccer, 2) FC Barcelona doesn't allow Real to hold anything back.

While looking at this game, careful attention was paid to the players' passing interaction. This information would allow for a truly in-depth, player-to-player description of team passing trends. A soccer ball's movement lends itself nicely to a Markov chain model. To better aim our study, our Markov chain breaks with convention in that "actions," not a unit of time, control its movement. Markov properties allow us to probabilistically predict the location of the ball *x* steps from the current state given the transition probabilities. By doing so, we can show how successful a possession—particularly one starting in friendly territory—can be by the probability of a goal based on the initial state of the ball. Because this analysis focuses only on possessions and actions by Real Madrid, the state space can be described by 17 separate states: 14 states for each Real player who took the pitch, 1 state for turnovers (i.e. FC Barcelona has possession), 1 state for a shot, and 1 state for a goal.

To build the transition matrix of this Markov chain, we watched game film repeatedly to determine the frequencies movements to and from these 17 states. Possessions were tracked in strings, which not only tracked which players were involved in each string, but also how many "actions" a possession lasted. Every possession string starts from a turnover (with the exception of the kickoff), and one might look something like: *Casillas→Pepe→James→Ronoldo→Shot*, meaning that the frequencies of each of those transitions get a tick in their column. Players can also recur to themselves (i.e. *Ronaldo→Ronaldo*) by making deliberate moves with the ball. Please refer to appendix Exhibit A for a full state space diagram. To form this frequency matrix (Exhibit B) into a proper Markov transition matrix (Exhibit C), the rows were then normalized. Due to the low number of players who score per game in soccer, we altered the *P(shot→goal)*

frequency in the matrix. Our modification makes our Markov chain pseudo-power-2 chain, as the *Prob(shot→goal)* depends on the player who transitioned the ball to "shot" in the first place. Throughout our model, where *i* is the current step:

$$P_i(shot \rightarrow goal) = [state_{i-1}\ row\ vector] \times [column\ vector\ of\ Players'\ goals/shot\ ratio] \quad (1)$$

Using interactions between each player and finding the mean number of actions per possession, we can calculate the best player to initiate a counter attack and the player with the most productive passing tendencies. It is important to note that we do not account for some of the less common goal types, such as goals off of penalty kicks or own-goals scored by the opposing team.

After our pseudo-power-2 transition matrix was up and running, we were then able to begin iterating through he chain. With 635 actions on 83 possessions, there was an average of 7.65 actions/possession. Accordingly, we used 8 actions to step though our Markov chain (as steps need to be integers). For each of the 14 players who played, we found the probability of a goal being scored on each action. We then summed these probabilities for each state *i* to yield the probability of a goal given a possession starting with that player *j*. In an equation:

$$Prob(Goal\ |Player\ j\ starts\ ) = \sum_{i}^{8} P(Goal\ on\ action\ i|Player\ j\ start)\ \forall j1, \dots, 14 \quad (2)$$

The same calculations were also performed with the goal/shot ratios adjusted for the Super batter/Super pitcher scenario:

$$\frac{G*}{1-G*} = \frac{G}{1-G}\frac{1-G_L}{G_L}\frac{G_G}{1-G_G} \quad (3)$$

Where G* = result ratio, G = player's current ratio, G_L = league average, and G_G = goalie's ratio.

* * *

Upon completion of the frequency matrix normalization and before a single calculation was made, the Markov transition matrix immediately gave us insight into the game. We noted that Dani Carvajal passes to James Rodriquez 41% of the time. Additionally, Cristiano Ronaldo—one of the world's very best players—lived up to his ball-hog reputation and kept the ball for himself 22% of the time. He also turned the ball over 22% of his possessions—a statistic that doesn't exactly scream "FIFA World Player of the Year." Karim Benzema took shots on 16% of his possessions. And while this analysis predominately focus' on offense, it is important to note that Carvajal pulls in 23% of the team's turnovers.

Without the SB/SP adjustment, findings from the state iteration had sensible outputs. Off of each possession's second action (consider it takes two actions to go from a players foot to the goal) in each possession, James, Benzema, and Ronaldo were the most effective, from most to least, respectively. James was 62.4% more effective than Ronaldo and 5.4% more effective than Benzema on 40% and 16.7% more shots, respectively. These statistics coupled with a 22% turnover rate show that either Ronaldo had am exceptionally bad game, or he is markedly overrated. Carvajal and Toni Kroos distinguish themselves as players who inspire peak probabilities of scoring on the third action of scorings they started. Carvajal's success in this category is undoubtedly linked to his high frequency of passing to James, while Kroos favors Benzema. After four actions, the chain begins to achieve its Markovian steady-state probability of $E[goal/action] = 0.0043$. When the 8 actions were summed, James, Benzema, and Ronaldo were unsurprisingly in the lead again. Interestingly though, of the three substitutes that went in the game, all were amongst the bottom five players in *E[Goals/possession*]. Full graphs for these outputs can be seen in Exhibit D and Exhibit E in the appendix. This seems to validate why they

are substitutes. Finally, with this analysis we found that *E[goals/game] = 2.37* with the *E[goals/possession] = 0.02854*.

After making the SB/SP adjustment to shooting ratios, the findings more-or-less mirrored those without the adjustment. Claudio Bravo, the FC Barcelona keeper, is without a doubt a super goalie, holding down a .056 goals/shot ratio. And with the league average sitting at .1136 goals/shot, Real Madrid is a team of super shooters. Again, James, Benzema, and Ronaldo were the heroes. However, with the adjustment James was 69.2% and 10.2% more effective than Ronaldo and Benzema, respectively. The adjustment essentially reinforced all of the insights from the unadjusted model. Interestingly though, the adjusted model had larger extremes than the unadjusted one. This brings to mind Wayne Gretzky's famous "You miss 100% of the shots you never take," quote. While obviously taking the same number of shots as the unadjusted scenario, James' newly-adjusted goals/shot ratio fell by 49% while Ronaldo's newly-adjusted goals/shot ratio only fell by 45%. Yet, James' effectiveness over Ronaldo's grew by 7.0ppt. This leads us to speculate that the SB/SP adjustment favors quantity (chances) to quality. Full graphs for these outputs can be seen in Exhibit F and Exhibit G in the appendix. Finally, with this analysis we found that *E[goals/game] = 1.30* with the *E[goals/possession] = 0.01574*. For a game that ended with 3 (albeit on of which came in a penalty kick) scores for Real Madrid, these figures imply that Claudio Bravo and a very subpar day.

In closing, the analysis offers insights that both reaffirm subjective belief as well as opens new questions. Teams watch hours on end of film to understand the tendencies of their opponent. With a frequency analysis at his disposal, an opposing Valencia midfielder could watch hours on end of Carvajal tape *or* he could just know that roughly half the time Carvajal passes to James. While the frequency analysis was very useful, the usefulness of the Markov data depreciated

quickly as the process moved to a steady state only 4 actions into a possession. Higher power Markov chains could potentially impede this quick steady state formation. They could expose even more doors in this type of analysis, as teams naturally have certain executions (clearance patterns, etc.) unknowingly hardwired into their play. However, a power-1 Markov chain might work more effectively in a game with fewer actions/possession, such as basketball (especially for an analysis to decide who to inbounds the ball to for the last shot). As for soccer though, we found Markov chains to be a powerful framework to express the movements and events of the time. Despite its limitations, the framework produced interesting that were not quite shocking, but not necessarily obvious either (i.e. James was the most lethal player on the field). As desired, it exposed who the most and least valuable offensive players on the field.

Due to lack of resources this study was not exhaustive, yet the passing frequencies themselves still serve for identifying strengths and weaknesses in strategy. We believe that this type of analysis needs to be done on a larger level to have more meaning, as a game is but a snapshot of a player's career. Although outside the scope of this work, we suspect managers of national teams could utilize this type of data when selecting squads to optimize how a team would sync, or agencies could use the method to rate player game performance. Although a large data sample of many games would be necessary, we believe this study to gained an initial insight into the potential of this framework.

References

- YouTube (footage of El Clásico)
- WhoScored.com (Soccer Statistics)
- FoxSports.com (Soccer Statistics)
- Wikipedia.com (Markov Reference)

Appendix

Exhibit A:

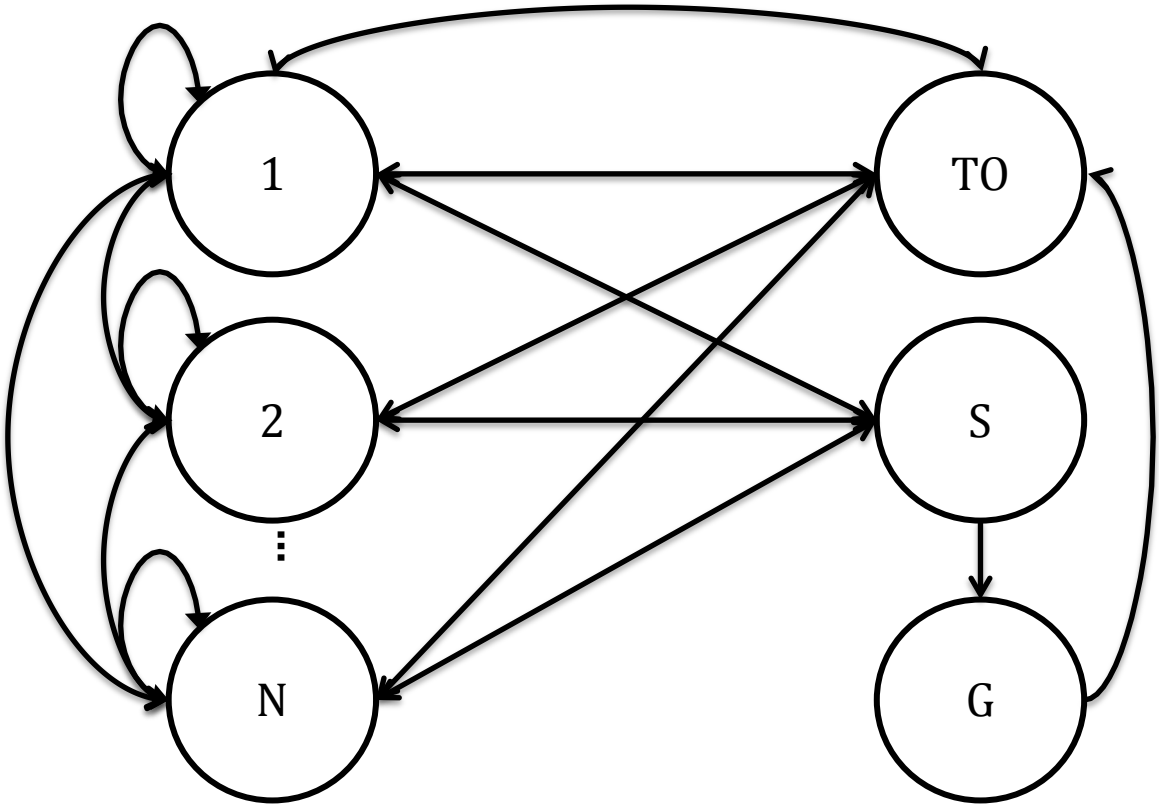

Exhibit B:

	I.C 1	Pepe 3	S.R 4	D.C 15	M.V 12	L.M 19	A.A 17	T.K 8	Isco 23	A.I 24	J.R 10	K.B 9	S.K 6	C.R 7	T.O.	Shots	Goals
Casillas 1	0	2	1	1	2	0	1	1	2	0	0	0	0	2	3	0	0
Pepe 3	3	1	4	3	3	2	0	7	1	2	1	1	1	5	4	1	0
Ramos 4	0	3	0	1	7	3	0	5	3	1	1	1	0	0	0	0	0
Carvajal 15	1	10	1	2	1	7	0	3	0	0	22	0	1	2	4	0	0
Marcelo 12	1	0	4	1	5	0	0	12	15	1	4	6	0	11	12	0	0
Modric 19	0	4	1	7	6	0	0	9	3	2	7	1	0	3	1	0	0
Arbeloa 17	0	0	0	0	2	0	0	1	0	0	0	0	0	0	1	0	0
Kroos 8	0	4	3	5	8	9	1	0	3	1	7	11	2	4	2	1	0
Isco 23	0	0	2	1	10	2	0	5	4	0	2	3	0	4	11	0	0
Illarramendi 24	0	1	2	0	0	0	0	1	0	0	0	0	2	1	1	0	0
James 10	0	0	1	11	4	5	1	1	4	0	5	6	0	7	12	7	0
Benzema 9	0	0	0	1	7	4	0	4	1	0	4	2	0	5	3	6	0
Khedira 6	0	2	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0
Ronaldo 7	0	0	0	1	4	4	1	3	5	0	6	3	0	13	13	5	0
Turnover	11	12	6	19	14	7	0	7	3	0	3	2	0	0	0	0	0
Shots	0	0	0	0	0	1	0	2	0	0	1	1	0	1	11	0	3
Goals	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0

Exhibit C:

	I.C 1	Pepe 3	S.R 4	D.C 15	M.V. 12	L.M. 19	A.A. 17	T.K 8	Isco 23	A.I. 24	J.R. 10	K.B. 9	S.K. 6	C.R. 7	T.O	S	G
Casillas 1	0.00	0.13	0.07	0.07	0.13	0.00	0.07	0.07	0.13	0.00	0.00	0.00	0.00	0.13	0.20	0.00	0.00
Pepe 3	0.08	0.03	0.10	0.08	0.08	0.05	0.00	0.18	0.03	0.05	0.03	0.03	0.03	0.13	0.10	0.03	0.00
Ramos 4	0.00	0.12	0.00	0.04	0.28	0.12	0.00	0.20	0.12	0.04	0.04	0.04	0.00	0.00	0.00	0.00	0.00
Carvajal 15	0.02	0.19	0.02	0.04	0.02	0.13	0.00	0.06	0.00	0.00	0.41	0.00	0.02	0.04	0.07	0.00	0.00
Marcelo 12	0.01	0.00	0.06	0.01	0.07	0.00	0.00	0.17	0.21	0.01	0.06	0.08	0.00	0.15	0.17	0.00	0.00
Modric 19	0.00	0.09	0.02	0.16	0.14	0.00	0.00	0.20	0.07	0.05	0.16	0.02	0.00	0.07	0.02	0.00	0.00
Arbeloa 17	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00
Kroos 8	0.00	0.07	0.05	0.08	0.13	0.15	0.02	0.00	0.05	0.02	0.11	0.18	0.03	0.07	0.03	0.02	0.00
Isco 23	0.00	0.00	0.05	0.02	0.23	0.05	0.00	0.11	0.09	0.00	0.05	0.07	0.00	0.09	0.25	0.00	0.00
Illarramendi 24	0.00	0.13	0.25	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.25	0.13	0.13	0.00	0.00
James 10	0.00	0.00	0.02	0.17	0.06	0.08	0.02	0.02	0.06	0.00	0.08	0.09	0.00	0.11	0.19	0.11	0.00
Benzema 9	0.00	0.00	0.00	0.03	0.19	0.11	0.00	0.11	0.03	0.00	0.11	0.05	0.00	0.14	0.08	0.16	0.00
Khedira 6	0.00	0.40	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.20	0.00	0.00	0.00	0.00	0.00	0.00
Ronaldo 7	0.00	0.00	0.00	0.02	0.07	0.07	0.02	0.05	0.09	0.00	0.10	0.05	0.00	0.22	0.22	0.09	0.00
Turnovers	0.13	0.14	0.07	0.23	0.17	0.08	0.00	0.08	0.04	0.00	0.04	0.02	0.00	0.00	0.00	0.00	0.00
Shots	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.10	0.00	0.00	0.05	0.05	0.00	0.05	0.55	0.00	0.15
Goals	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00

Exhibit D:

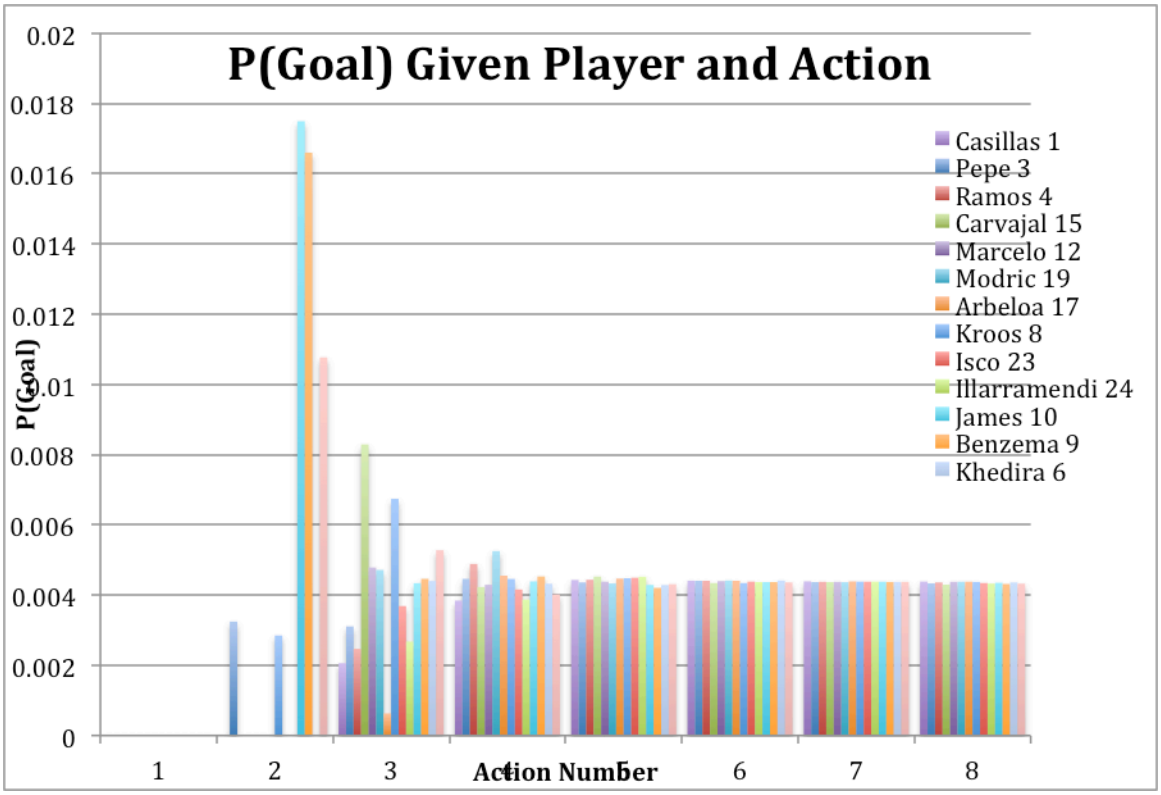

Exhibit E:

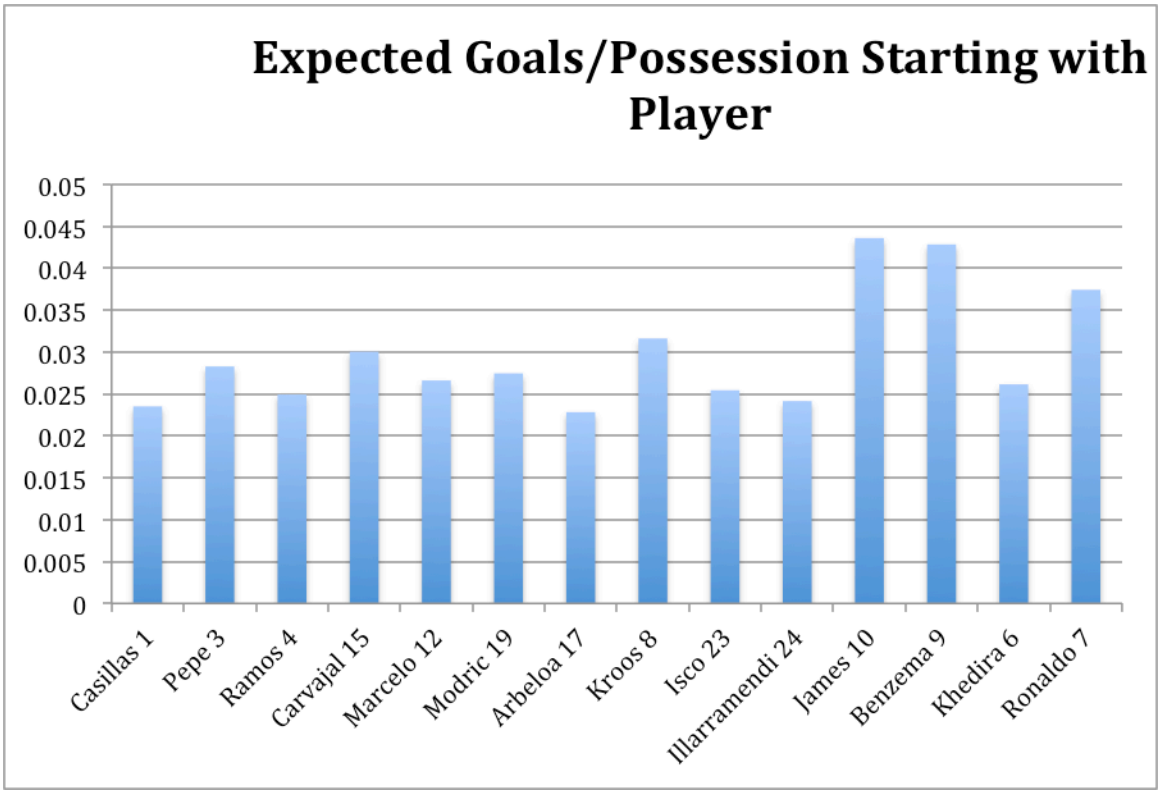

Exhibit F:

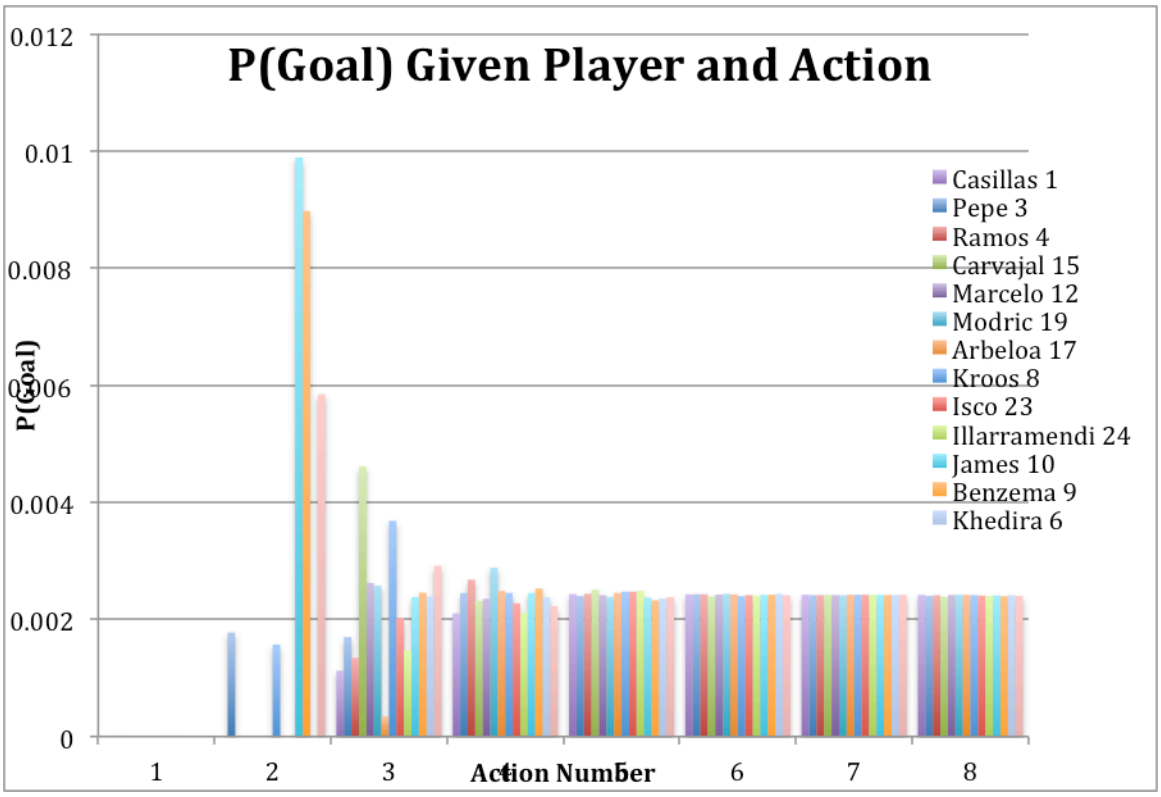

Exhibit G: