Statistical Analysis of Defensive Production in Major League Baseball

A Major Qualifying Project Report:

submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

by

_____

Allison M. Haskins

_____

Alyssa M. Lopes

_____

Christopher M. Teixeira

Date: March 13, 2006

Approved:

_____

Professor Jayson D. Wilbur, Advisor

1. Statistics
2. Baseball
3. Factor Analysis

Abstract

While many statistical measures have been devised to measure team and player performance in professional baseball, relatively little work has been done to study and improve measures of defensive performance. This project's goals were to model defensive contribution to team performance and develop related measures of individual player defensive production. Factor analysis was employed to extract positional factors and combine them with pitching and offensive measures into a final logistic regression model to determine effects of defense on team performance.

Executive Summary

Many statistics are used to evaluate both team and player performance in baseball. However, most of these measure pitching and offensive production, while relatively few measure defensive ability. In response, this project aimed to model the defensive contributions made by individual positions as well as a team's complete fielding performance to its overall success.

The project group began by creating a list of qualities and abilities for each position that encompassed the responsibilities of the position. Using a combination of commonly used statistics and their knowledge of the game, members of the project group recognized 3-8 variables for each position, for a total of 43 position-specific traits. Each of these was paired with a measured statistic(s) that best described that characteristic. The data was obtained from two sources: the Retrosheet database and the Baseball Archive database. The project team decided to use data from the 2000-2004 seasons for several reasons, including that it was most complete, available in both databases, and most recent. The defensive statistics from the 2000-2004 seasons were extracted for every player and summed according to position, team, and season; a new database was created with this data in which each entry was for a given position, team, and year.

Next, the project group continued by using factor analysis methods to determine several factors comprised of the position-specific variables to model defensive performance on a position-by-position basis. There were 6 different methods of factor analysis proposed for each position including analysis using the correlation and covariance matrices, in addition to transformations of the data using the log function, ranking, and normalizations by the coefficient of variation and coefficient of variation

squared; analyses of the transformed data sets utilized the covariance matrix. Each of the 6 methods was run using a varied number of factors. The number of factors that explained either 90 or 95 percent of the variation in the data as well as the number of factors that had eigenvalues greater than one (or greater than the average of the eigenvalues using the covariance matrix) were considered.

Selection of the best factor analysis model for each position was based on three criteria: interpretability of factors, model consistency amongst the 5 seasons, and AIC,

- Pitcher – Correlation with 3 factors
- Catcher – Rank transformation with 3 factors
- First Base – Normalized coefficient of variation with 2 factors
- Second Base – Rank transformation with 2 factors
- Third Base – Correlation with 3 factors
- Shortstop – Rank transformation with 4 factors
- Outfield – Log transformation with 2 factors

Each generated factor was given a name that reflected the quality or ability that was best described by those variables with high loadings within the factor.

As a means of verifying that those factors determined for each position should be used in the final model, cluster analysis was conducted that compared the plotted factor scores between Gold Glove recipients and non Gold Glove recipients. If the factors were valid, then the results from the cluster analysis were expected to show not only differences between those Gold Glovers and the rest of the players, but a trend where the Gold Glovers would receive the most desired scores.

The cluster analysis results were varied, with some positions exemplifying the desired trend of Gold Glovers performing best, some lacking in any noticeable distinction between the two groups, and others making no logical relation between factors. Regarding the degree of variation found between the Gold Glovers and non-Gold Glovers

after entering the players' statistics into the validation plots, the project group determined that this was of not great consequence. The selection of non-Gold Glove recipients was restricted to only those players who had represented their team for the greatest number of innings for their particular position, meaning that they were skilled everyday players.

After the validation process was completed, the 19 defensive factors were combined with measures for pitching (earned runs allowed), hitting (runs scored), and division in a final logistic regression model. Statistically, it was determined that none of the defensive factors were significant in the final model, however the pitching and offense factors were; thus nearly all of the variation in the data could be explained in terms of a team's offensive production and pitching performance. While this did not bode well for the project group's interest in developing an innovative defensive measure at the team level, it did help to support the argument that perhaps defensive skills are not that important to consider when composing a team

Even though a significant team-based defensive model was not found, the results of the exploration on individual position defensive qualities were noteworthy. For each position, the selected factors not only made sense intuitively, but also they appeared to cover the spectrum of a player's responsibilities that were suggested by the project group at the start of the project. Additionally, even though not every validation analysis resulted in exactly what the group members anticipated, the majority of the data plots confirmed the validity of the factors selected to represent each position. Therefore while the project group was unable to model defensive contribution made by a team, the results from the individual position factor analyses satisfied the group's goal of determining important defensive characteristics in terms of position.

Acknowledgements

Our group would like to thank the following individuals and organizations for their time, effort, support, cooperation, and guidance throughout the project:

Jayson D. Wilbur
Andrew W. Swift
Carlos Morales
Balgobin Nandram
Debra Dexter
Christine Drew
Ashley Moraski
Pat Malloy
Ethan Thompson
Jim Albert
Retrosheet
The Baseball Archive

Table of Contents

Table of Tables

Table of Figures

1. Introduction

Baseball is more than just a sport in the United States; it is a defining characteristic of the nation's culture, a successful and profitable industry, and a passion shared by millions of Americans. There are many individuals involved in the baseball industry with a substantial interest in the success of their team during a season. Each year, a baseball organization strives to put together the best team possible in order to have a chance at being the top team at season's end. As baseball's popularity and profitability are continually increasing, the demand for more sophisticated ways of determining the best players and organizations continues to rise as well (Spalding, 1992).

There currently exist many statistical measures and models to evaluate pitching and offensive performance, yet relatively few measures exist to evaluate defensive production. This project aimed to model defensive contribution by both individual positions, as well as overall team fielding performance. The main goals were to identify those variables most important to a team's defensive production, explore defense characteristics in terms of individual positions, develop a way to effectively model defensive contribution to a team's overall success, and to validate the resulting model through cluster analysis and other statistical procedures.

The project group studied numerous statistical topics with the intention of determining the best way to approach the problem. Topics studied included Bayesian statistical inference, generalized linear methods, analysis of observational studies, Markov chains, factor analysis, and cluster analysis. After considering these topics, the project group decided to use factor analysis methods as the primary basis of its investigation as it was deemed most appropriate for the project goals as well as the type

of data available. Factor analysis procedures were applied to fielding data from the 2000-2004 major league baseball seasons, with the eventual outcome of the development of a final model which combined the factors from individual position models with measures that accounted for pitching, hitting, and division factors. The resulting logistic regression model, along with the observations and conclusions made during the entire procedure leading up to its completion, provided great insight on the role that defense has in the success of a major league baseball team

The main goals were to identify those variables most important to a team's defensive production, explore defense characteristics in terms of individual positions, develop a way to effectively model defensive contribution to a team's overall success, and to validate the resulting model through cluster analysis and other statistical procedures.

2. Background

Whether playing or watching, baseball is one of the most popular forms of entertainment in the United States. American professional baseball began with the establishment of the National League (NL) in 1876, closely followed by the founding of the American League (AL) in 1900 (Albert & Bennett, 2001). Today, these two leagues combine to form the major leagues. Currently, a total of 30 teams, divided among three divisions in each league, compete annually in a 162-game season. After each team completes 162 games, the six teams with the best win-loss records in each division are eligible to compete in the playoffs. Since 1995 the team with the best record from each league that does not win its division also has also participated in the postseason. Initially, four best-of-five game divisional series are played, with each of the eight teams facing another from their league. Whichever teams win three of the five games then play in a best-of-seven game league championship series with the other three-game winner from their league. Finally, the two teams that win four games in their respective series, the American and National League champions, participate in the seven-game World Series. At its conclusion, the team that wins four games is given the title of World Champions (MLB.com, 2005).

2.1 Sabermetrics

Since early in baseball history, statistics, to some extent, have been recorded, describing both team and individual performance. With the advent of computers and the Internet, the amount of data available to the public has increased tremendously and thus public interest has grown as well. Fans of the game use statistics to discuss the

performance of their favorite player, to speculate the chances of a team winning a particular game, and perhaps to guess the eventual outcome of the regular season.

Emerging from this popularity, the ideology of "sabermetrics" was developed in the 1970s by SABR, the Society for American Baseball Research. "Sabermetricians" seek to find objective answers to questions in baseball using statistical measures (Grabiner, n.d.). Sabermetrics consists of several tasks, primarily evaluating certain statistics for validity and accuracy. New statistics have also been formulated in order to better assess players' abilities and performance, which can then be applied to either examine past performances or predict possible future occurrences (Grabiner, n.d.). Therefore, these statistics have become much more than a form of recreation. Each major league team uses statistics to aid in managerial decisions during games as well as regarding adjustments to their roster before, during, and after the regular season.

2.2 Measures of Hitting Performance

Undoubtedly a team's ability to produce runs is a significant factor in its overall success. Presently, several different statistics are used to measure a player's hitting ability. However one must question whether these measures are evaluating all of the relevant information. An effective hitter not only has the ability to produce singles, doubles, triples, and homeruns, but also performs well in a situation such as "*sacrificing*" (italicized words are listed in the Glossary) with either a *bunt* or deep fly ball in order to advance runners and earn runs (Albert, 2001b). A player's overall ability to help his team score runs should be evaluated and not just the hitting percentage, or the proportion of total number of *hits* to the number of batting attempts, not including *walks* and *hit by pitches*, of that player.

Currently, the most commonly used offensive measure in baseball is the batting average (Bennett, 1983). This is simply computed as the number of hits earned divided by the number of *at bats*.

$$BA = \frac{H}{AB} \tag{2.1}$$

Many sabermetricians argue that this figure is not representative of a player's ability to score runs, and at most it may lead one to infer that the more hits a player makes, the more runs he is able to score (Bennett, 1983). However, the batting average does not take into account the number of times a batter is advanced to first base due to a walk, hit by pitch play, or a fielder's *error*. The batting average can also be misleading due to its lack of differentiation between hits. In other words, since this statistic is unweighted, it does not view a single any differently than a homerun (Albert, 2001a).

Another popular measure of hitting performance is slugging percentage, which is used to measure the advancement of runners (Albert, 2001b). This is the total base accumulation from hits divided by the number of at bats.

$$SLG = \frac{TB}{AB} = \frac{(1B) + 2(2B) + 3(3B) + 4(HR)}{AB} \tag{2.2}$$

Like batting average, the slugging percentage also does not include walks or a hit by pitch play and thus is not completely accurate (Bennett, 1983). Aiming to remedy this discrepancy, the on base percentage (OBP) measures the rate of getting on base by hits, walks, and hit by pitch plays (Albert, 2001b).

The three aforementioned measures of offense are the most popularly used amongst fans. Even though these may not prove to most accurately portray a player's ability to score runs, they tend to be the easiest to understand and hence are widely

accepted in the baseball community. However, a baseball statistician has much different views and goals when evaluating the game than the average fan, resulting in many alternative, more complex measures.

A statistic that seems to be growing in popularity is OPS, which is the sum of the on base and slugging percentages.

$$OPS = (OBP + SLG) \tag{2.3}$$

Bill James, a prominent "sabermetrician" and perhaps one of the first to use statistics to answer objective questions about the game (Roney, 2003), proposed a measure called Runs Created (RC) which aims to incorporate both the amount of times on base and the advancement of runners by multiplying them together and dividing by the total number of plate appearances (Albert, 2001b).

$$RC = \frac{(H + BB) * TB}{(AB + BB)} \tag{2.4}$$

In order to further evaluate the relevance of the type of hit produced, a weighing system was developed by Thorn and Palmer (1989) called Linear Weights. In this model, the number of singles, doubles, triples, homeruns, and walks a batter produces are each multiplied by some constant weight, where a homerun would have a greater weight than a single. Palmer determined these constant weight values by running a computer simulation of all the major league games played in about a 75 year span. The frequencies of offensive events were then tabulated and assigned advancement values which corresponded to the expected average run value for each hitting event. The values were determined for c1 through c5 as follows: 0.46, 0.80, 1.02, 1.40, and 0.33. These values are then summed for the final measure (Albert, 2001b).

$$LW = c1(1B) + c2(2B) + c3(3B) + c4(HR) + c5(BB + HBP) \tag{2.5}$$

Other statisticians have researched whether a player's offensive ability is influenced by the current situation during an at bat. Albert (2001a) evaluates player ability to score runs for his team in different situations involving the number of outs and number and position of players on base. For example, it seems logical to conclude that it is more likely a player will drive in a run if the bases are loaded with no outs than if no one is on base and there are two outs. Resulting, an important question is whether or not a batter is affected by the situation and either performs better or worse in "clutch" situations, or situations where the result is very important to the outcome of the game (Albert, 2001a). Specifically, Albert addresses the issue by using a measure proposed by Lindsey (1963) known as Value Added or Value at Plate Appearance.

	Runners							
Outs	None on	1st	2nd	3rd	1st, 2nd	1st, 3rd	2nd, 3rd	Bases Loaded
0	0.49	0.85	1.11	1.3	1.39	1.62	1.76	2.15
1	0.27	0.51	0.68	0.94	0.86	1.11	1.32	1.39
2	0.10	0.23	0.31	0.38	0.42	0.48	0.52	0.65

Table 1: Run Potential Probabilities

Using a table of run potential probabilities (Table 1) (Albert, 2001a) determined by the amount of players on base and number of outs, Value Added is calculated by subtracting the expected runs probability before the player bats from the expected runs probability after the player bats, and this total is added to the actual number of runs resulting from the play. A positive value yields a successful plate appearance, whereas a negative value indicates an unsuccessful plate appearance (Albert, 2001a).

$$VA = (ExpR\_A) - (ExpR\_B) + R \qquad (2.6)$$

Even though many players have been coined as "clutch hitters," many argue that this is not an ability, but rather a random variation of hitting. Several studies have been performed in order to analyze this claim. One of the first was by Cramer (1977) who compared players' hitting in clutch situations for two consecutive seasons. He found that a player who performed well under clutch situations in one season did not necessarily perform well in the next season. Therefore, he concluded that clutch hitting was actually just a random occurrence (Neyer, 1999). Conversely, some argue that the reason why these studies do not detect a substantial difference in situational hitting is because the commonly used hitting statistics are not valid measures of this ability, often omitting sacrifice bunts or fly balls used to score runs (Hakes, 2004). This issue of clutch hitting is an ongoing debate amongst statisticians and fans alike.

2.3 Measures of Pitching Performance

The quality of a team's pitching staff is essential to the success of a major league baseball organization. Both *starting pitchers* and *relief pitchers* comprise a sizeable portion of each team's roster. Starting pitchers are those pitchers who begin pitching in the first inning and are expected to play for most of the game while relief pitchers enter the game in later innings to replace, or "relieve", another pitcher. Relief pitchers are generally not expected to remain in the game for many innings and may be called upon to pitch for a single batter under certain situations. Effective pitchers switch pitch types, change velocities, and make effective use of the *strike zone* to increase their success against opposing batters. Since little emphasis is placed on a pitcher's batting or fielding ability, the evaluation of a pitcher is almost entirely based on his performance on the mound. To gauge just how successful a pitcher is, a wide variety of statistical measures

have been developed by coaches, statisticians, and baseball enthusiasts alike to evaluate and make comparisons between major league pitchers.

The most commonly used measure in evaluating pitching is the earned run average, (ERA). A player's ERA is defined as the average number of runs allowed during the course of nine innings, ignoring those runs that result from some defensive error or misplay by another player.

$$\text{ERA} = \frac{\text{ER}}{(\text{IP} / 9)} \qquad (2.7)$$

Hence when using solely this quantity, one pitcher is "better" than another if his ERA is lower. Another way of evaluating a pitcher is by simply considering the number of wins and losses accumulated over the course of a season. For starting pitchers, the mark that is currently accepted as defining an exceptional pitcher is twenty wins in a season (Albert & Cochran, 2005). In addition to the above measures, the recorded number of *strikeouts* and walks are also considered in pitcher assessment. Often these measures are transformed into ratios depicting the number of either strikeouts or walks per nine innings. Accordingly, this is the total number of strikeouts or walks in a given period of time divided by the total number of innings pitched. Table 2 lists and describes some other commonly used pitcher statistics.

Statistic	Abbreviation	Description
Earned Runs	ER	Number of earned runs allowed
Earned Run Average	ERA	Average number of earned runs allowed per nine innings
Walks*	W	Number of walks allowed
Strikeouts	SO	Number of strikeouts
Hits	H	Number of hits allowed
Runs	R	Number of runs, both earned and unearned
Home Runs	HR	Number of home runs allowed
Opponent Batting Average	OBA	Combined batting average of all batters against a particular pitcher
Hit Batsman	HP	Number of batters hit
Wild Pitch	WP	Number of pitches determined to be too erratic for the catcher to handle and result in the advancement of a base runner
Walks and Hits Per Innings Pitched	WHIP	Number of hits and walks (totaled) per inning
Saves	S	Number of times a relief pitcher enters the game with his team leading by less than 3 runs and is still pitching at game's end with his team victorious

Table 2: Commonly Used Pitching Statistics

*Many of the above statistics are more commonly found as ratios of the total number of walks, hits, strikeouts, etc per inning or per nine innings

While the preceding statistics have been used for many years to assess pitching ability, there are various problems associated with their usage. First, these statistics do not consider the strength of the pitcher's or the opposing team. An average pitcher playing for a team with an exceptional offense may win as many games as a pitcher with greater ability playing for a below average team. Alternatively, an above average pitcher may lose more games than expected if he plays for a team that produces few runs and/or commits many errors in the field.

A second factor to consider when using these measures is the difference in applying them to both starters and relievers. As starting pitchers have completely

different roles then relief pitchers and face many more batters in a season, it is not always sensible to use the same statistics or conclude that a starter and reliever with the same value of a particular measure are equal. As an example, consider the ERA of a pitcher. If a pitcher records all 27 outs in a game and no *earned runs* are recorded, then his ERA for the game is zero. But if a relief pitcher enters the game, strikes out one batter, and then exits the game immediately after, his ERA for the game is zero as well. It is obvious that these two performances are not equivalent, yet if one simply looks at the ERA for each player, then no difference can be discerned. Figure 1 illustrates the differences between the ERA of starters and relievers and alludes to the difficulties that arise when using this measure to compare pitching performances.

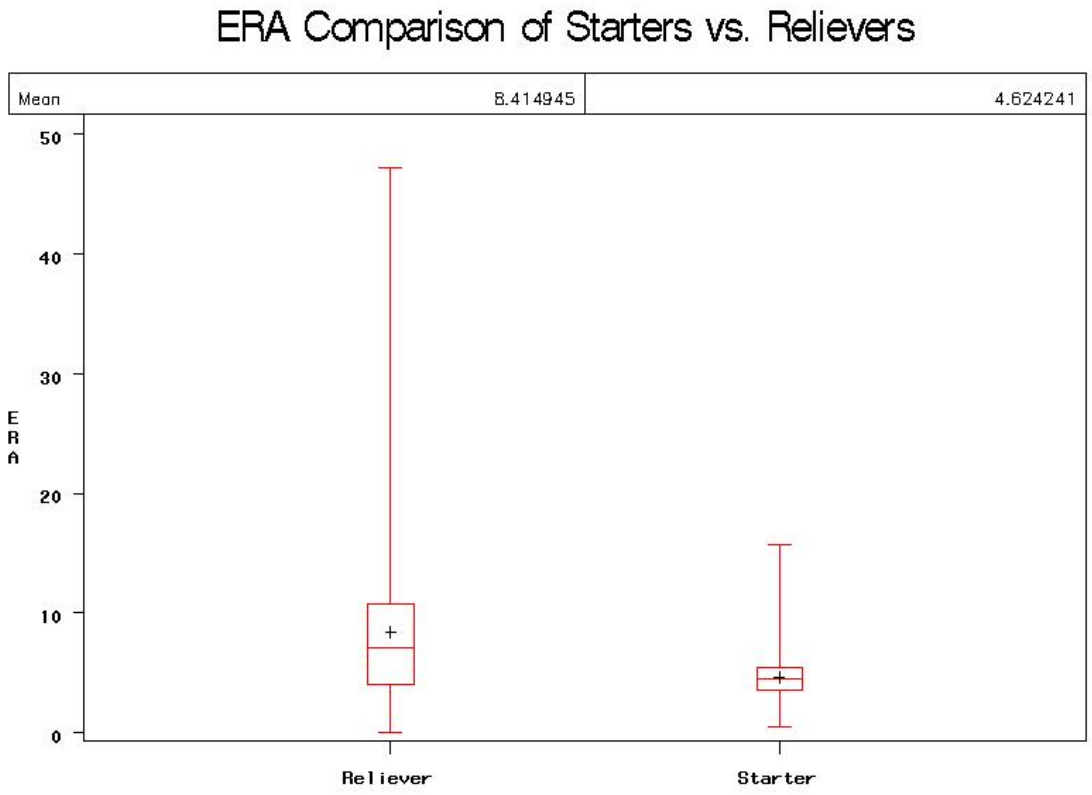


Figure 1: ERA Comparison of Starters vs. Relievers

Alternative methods and measures have been proposed to eliminate these problems. Typically, when evaluating a pitcher across a season, statisticians will tend to avoid using those measures that are confounded with other variables beyond the pitcher's control. Thus certain statistics, including number of wins and losses, are not as accepted as measures like ERA. Although ERA is generally regarded as a good measure of pitching ability, an alternative measure to both ERA and win/loss records is proposed by Thorn and Palmer (1993). The two argue that ERA fails to demonstrate the efficiency of a given pitcher for his team during a particular season while wins and losses are not determined solely by the pitcher's ability. Instead of ERA, they suggest using Pitching Runs, which takes the number of innings pitched multiplied by the difference between the league ERA (LERA) divided by nine, and the number of earned runs allowed.

$$PR = (IP/9) \times LERA - ER \qquad (2.8)$$

Here, a value greater than zero characterizes an above average pitcher while anything less than zero refers to a below average pitcher (Albert & Cochran, 2005).

To eliminate problems in using measures that compare pitchers with a sizeable difference in number of innings pitched, it is often preferable to use statistics that are calculated as ratios, rates or proportions. So for this particular problem, taking ratios between some observed measure (number of strikeouts, *home runs* against, walks, etc) and the number of innings pitched is an improvement over the number of occurrences themselves (Albert & Bennett, 2001). Similarly, when looking to create models based on pitcher performances, either the use of derived baseball statistics or the exclusion of all data from pitchers who played for less than a specified number of innings is preferred.

2.4 Measure of Defensive Performance

In addition to its pitching and offensive production, a team's defensive ability is a contributing factor to its success. The defensive makeup of a team consists of nine players that take the field, each of which assume a different role. There are three outfielders, the left fielder, center fielder, and right fielder, who are responsible for any ball that leaves the infield. A good outfielder is skilled at catching fly balls, backing up infield plays when necessary, and making assists. This is more difficult than it appears since the dimensions of every ballpark differ, forcing the outfielders to make frequent adjustments. The dimensions may vary in terms of the position of the outfield wall and even the amount of foul territory that exists. Examples of some unique obstacles are the center field hill at Minute Maid Park in Houston, referred to as Tal's Hill, and the high left field wall, often referred to as the "Green Monster," at Fenway Park (Sandalow, 2004). Another problem presented with outfield data is trying to record exactly where plays are made. Measurement error exits as different people, with unavoidably varied measuring techniques, are responsible for recording the hit location of every play.

The remaining six players are positioned in the infield. Behind the pitcher are four players which primarily cover the area between the outfield and the pitcher's mound. The left side of the infield is split between the third baseman and the shortstop while the right side is split between the first and second basemen. Common tasks for infielders include fielding *ground balls*, catching pop flies, and throwing out base runners. The middle infielders, the second baseman and shortstop, must be particularly skilled at executing relay throws and turning double plays, or plays that result in two outs. The third baseman is often responsible for preventing sacrifice bunt attempts. Due to its

proximity to home plate, this position requires very good reflexes and quick reaction time. Also positioned just ninety feet from home plate, the first baseman shares many of the same responsibilities of a third baseman. In addition, the first baseman is often at the receiving end of many errant throws, and therefore must be able to react quickly to prevent errors.

Two of the most important defensive positions are the pitcher and the catcher because they are involved in every play that occurs. The catcher is accountable for the area surrounding home plate. Constant communication between the pitcher and catcher must take place to prevent base runners from advancing around the bases. When a base runner is on first base, the pitcher must take care not to let the runner take too large of a lead or else he may easily steal second base. To prevent a runner from attempting to steal second base, pitchers routinely throw to first base to keep the runner's lead to a minimal. If a steal attempt is initiated as a pitch is thrown, it becomes the catcher's responsibility to throw out the runner at second base. Similar scenarios occur when base runners are situated at other bases. Passed balls and *wild pitches* are a concern for catchers after every pitch. Both of these can result in the advancement of a runner or a run scored. However, wild pitches are charged to the pitcher while passed balls are charged to the catcher (Tangotiger, 2004).

Many defensive statistics are maintained for both team and individual performance. Common statistics among all positions include putouts (PO) and assists (A). Putouts refer to the first out made during a play while assists are exactly the number of outs recorded on the play minus the putout. For example, if an outfielder catches a fly ball, the batter is out and a putout is recorded. If the outfielder in turn throws out a base

runner after making the original play, then he is credited with an assist. Another statistic, Range Factor, combines both putouts and assists to create a way of measuring players against each other, taking into consideration the total number of innings played (INN) (Nutting, 2004). Range Factor is calculated using the following formula:

$$RF = \frac{(PO + A) * 9}{INN} \tag{2.9}$$

Fielding percentage, which is the total number of successful plays divided by the number of chances, is by far the most commonly used defensive statistic. One of the bigger problems for this statistic is the perception of an error; one individual scorekeeper may call a particular play an error while another may call it a failed attempt on what would have been an incredible play, and hence not an error. The location of the stadium where the play takes place would also have an effect on the call. Also, it is important to consider the effect of a player's ability to reach balls put into play. Players with a greater range can cover more ground and get to the more difficult to reach balls. If these players make a miscue on the play, they could be charged with an error. Meanwhile, a player with a smaller range who could not reach the ball will not be charged with an error for essentially the same play.

A catcher's involvement in various types of plays throughout a game can lead to questions about his effectiveness. Several factors come into play for measuring the quality of a catcher. Consequently, specific measures were created in order to cover the different aspects of being an efficient catcher. Statistics like fielding percentage, base runners caught stealing, number of stolen base attempts, and passed balls are all taken into account (Rosciam, 2004). Each one plays an important role in evaluating a catcher. Major League Baseball has recently started recording data like innings caught and

opponents caught stealing. These have been used to create some new measures like Stolen Base Attempts per Game (SBA/G). Weigand (2004) modeled SBA/G for a single catcher to compare him to the league's value of SBA/G. Rosciam's model (Rosciam, 2004) is more complicated and includes more statistics like passed balls and a catcher's game calling ability. He combines six measures to create an intricate model that produces a ranking for catchers. Basic statistics are used to compute stamina, good glove, good arm, ball handling, effective game play, and game calling. These measures are multiplied together to create his model called catcher defensive rating.

There are certain plays in baseball that are remarkable or worthy of notice. Starting with the pitcher, picking off a base runner is a rare feat. It is easy to make the throw over to a base and attempt to pick off a runner. However, pitchers must be very accurate in their throws as considerations of the risk of either *balking* or throwing the ball away exist. Throwing out a base runner trying to steal a base is also a substantial feat. The catcher must be able to catch the pitch and then throw out the runner with the batter in his face.

Other great plays in the infield include plays at home and traditional double plays. Plays at home plate tend to be some of the most exciting plays in baseball. Runners try to slide in around the catcher or even try to take out the catcher in order to score a run for their team. From the defensive side of the ball, double plays can turn around an inning instantly. Most double plays occur in the infield where a ground ball is hit and the infielders can create two outs easily. Other double plays consist of an outfielder catching a *pop fly* and then throwing out a runner that tried to advance or a pitcher striking out a batter and the catcher throwing out someone trying to steal a base.

2.5 Databases

During the past fifteen to twenty years, sophisticated quantitative baseball analysis has experienced a tremendous growth. Several organizations, including Retrosheet and the Baseball Archive, have created comprehensive databases containing player and team statistics to aid in statistical analysis.

2.5.1 Retrosheet

As a consequence of the demand for more accessible data to perform these sophisticated analyses, the non-profit organization, Retrosheet, developed an organized database of baseball figures beginning in 1989. The database contains scoresheets, collected by Project Scoresheet and other similar organizations that are taken from official major league baseball scorers, game announcers, and fans watching the games at home. These scoresheets are then taken and translated into a common notational system.

Retrosheet compiles all of the data into a simple system available through their website. The event files, detailing the 1959-2004 seasons, contain play-by-play data for each game during that year. They list box scores, league leaders, and every player that played a certain position during a particular year. Retrosheet developed its own programs that are able to search through the event files for specific data. Additionally, Retrosheet keeps a listing of changes to team rosters, including results of the annual draft as well as trades between teams that take place during a season.

Because these data are donated to the organization, researchers can easily access it without having to be charged a fee. The organization's founders encourages use of its event files for research and allows people to publish papers through its website as long as their research incorporates use of the Retrosheet database.

Unfortunately, one of the largest challenges Retrosheet faces is obtaining complete and accurate data. The data from past years that is just now being submitted is comparatively less desirable as the data taken at more recent games. Common examples of missing data include pitch counts or even complete games. With the abundance of computers today, statistics can be recorded and stored so that more data can be easily retained. Retrosheet maintains the long-term goal of constructing a complete and accurate record of all games played in the modern era.

2.5.2 Baseball Archive's Database

An alternative source of free and comprehensive baseball data is located on the Baseball Archive's website. Steven Lahman, a columnist for the Baseball Archive and author of *Reversing the (Other) Curse*, created the database in 1994 to satisfy the need for an easily searchable source of baseball statistics. The database contains season totals of many of the most common baseball statistics, including those of pitching, fielding, and hitting. It also contains newer statistics, such as intentional base on balls and zone rating. Today, the database contains data from the 1871 through 2004 seasons, and is assisted by organizations like SABR and Project Scoresheet in its data collection.

2.6 Data Sources

The development of any measure or model is dependent on the data available on which the calculations are based. For this project, there were two main sources of data that were not only comprehensive enough to support the problem at hand, but also were accessible to the project group, namely, databases created by the Baseball Archive and

Retrosheet. Each is unique in its presentation and in the data that it contains. However both include ample defensive data.

The Baseball Archive database contains yearly totals of individual player statistics in many categories, including fielding data. In each of the years from 1871 through 2004, the database organizes the statistics by player, with separate entries for a given player who either plays for a different team over the course of one season, or plays more than one position. For example, a player who splits his time at second base and shortstop during one season will be listed twice, with his information specific to his performance at each position kept separate. The defensive data available for all years in the database includes number of games played, assists, putouts, errors, double plays turned, and passed balls. A total of 123,944 entries contain this information. Of these, 11,359 entries from the 2000 through 2004 seasons contain zone ratings for each player which measure the percentage of balls that a player fields in his conventional defensive area. Because approximately 91 percent of the entries do not have a recorded zone rating, it is uncommon to see it used by statisticians. The Baseball Archive database is clearly organized and made available in several formats, including Microsoft Access and Microsoft Excel.

The Retrosheet database contains a large amount of data on the fielding performances of individual players. However, it is limited to the 1974 through 1992 seasons. The main difference between Retrosheet and Baseball Archives is that Retrosheet's database contains play-by-play data from 1987 to 1992, and also from 2000 to 2004. The play-by-play data describe what happens every time an out is recorded or a runner advances to another base. In certain instances, each pitch of the at-bats is

recorded as well. The play-by-play data is organized by each play in the course of a game, and lists the inning, the player at bat, the current defensive field arrangement, and exactly what happened and which defensive positions took part in the play. Using other data tables in the database, it is possible to determine not only which positions took part in a play, but also the names of the actual players involved. Unique to the Retrosheet, the database also describes where each play occurred using a system that breaks down the playing field into numerous labeled regions. The complete list of the 97 fields and their descriptions are found in Appendix A.

When the Baseball Archive database was created, it included the player identifications from the Retrosheet database in its master table. Therefore, the two databases can be cross-referenced and information about a particular player can be compiled. In this manner, one can work with all data available between the two databases and be able to construct better models and measures by incorporating more data. The reliability of the consistency between the data found in both databases was a concern. However, after the project group randomly selected statistics for various positions and compared the results from the Retrosheet and Baseball Archive databases, no relevant differences were found.

3. Statistical methods

The project group employed several statistical methods throughout the project. Techniques included factor analysis, cluster analysis, generalized linear models, and logistic regression.

3.1 Factor Analysis

Factor analysis is used to determine whether relationships exist amongst numerous variables. The goal of factor analysis is to explain the variation and interrelationships among a large number of related variables by a smaller number of uncorrelated latent factors which are not directly observed. Each factor explains a certain aspect of the variability of the response variable and the factor's meaning is then interpreted by the statistician. One of the problems commonly associated with factor analysis is the interpretation of the factors produced. Since these factors are linear combinations of the original variables, they cannot always be interpreted.

3.1.1 Factor Analysis Model

Factor analysis starts by trying to solve a factor model. A typical form of the factor model is

$$\mathbf{y} - \mu = \lambda \times f + \varepsilon \qquad (3.1)$$

where **y** is a vector of the random sample taken from a homogenous population, μ is the vector of mean values corresponding to the random samples, λ is the vector of factor loadings, f is the vector of factors used in the models, and ε is a vector of the random errors associated with each random sample. The covariance of the ε vector is defined as

$$\text{cov}(\varepsilon) = \Psi = \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \Psi_p \end{bmatrix} \tag{3.2}$$

Further analyses of the data can be performed once the factor model has been determined.

Each factor in the model is associated with an eigenvalue which measures the variance in the original variables that the factor explains. In other words, the higher the value of the eigenvalue is for a factor, the more variance is explained by it. Therefore, factors with low eigenvalues are usually disregarded since they do not explain much of the total variance.

3.1.2 Principal Component Method

The principal component method produces uncorrelated factors which explain as much of the variance in the original variables as possible with a small number of factors. The estimated covariance matrix *S* from the sample of data is used along with the spectral composition to form the following equation,

$$S = CDC^T \tag{3.3}$$

where *C* is an orthogonal matrix comprised of the normalized eigenvectors of *S* and *D* is a diagonal matrix with the eigenvalues of *S* on its diagonal. Considering that $D = D^{\frac{1}{2}} D^{\frac{1}{2}}$, one obtains the following:

$$S = \Lambda\Lambda^T \text{ where } \Lambda = CD^{\frac{1}{2}} \tag{3.4}$$

where Λ is a *p* x *m* matrix that is approximately equal to the first *m* columns of $CD^{\frac{1}{2}}$, a *p* x *p* matrix. In other words, one chooses the largest *m* eigenvalues to create matrices C_1

and D_1 such that $\widehat{\Lambda} = C_1 D_1^{\frac{1}{2}}$ and S is a p x p matrix. Here p is the number of variables and m is the number of factors (Rencher, 2002).

A matrix $\widehat{\psi}$ is created using the diagonals of the difference between S and $\widehat{\Lambda}\widehat{\Lambda}^T$. The diagonal elements of $\widehat{\Lambda}$ are called the loadings for each factor. Note that if one begins by first standardizing the data, then the covariance matrix can be replaced by the correlation matrix. Additionally one can choose to transform the data by taking the logarithm, ranking the data, or normalizing, such as by the coefficient of variation (Rencher, 2002).

3.1.3 Rotation Method

The rotation method is often used in order to make interpretation of the factors easier by distinguishing those variables in the first factor that have the highest loadings from those with medium loadings and decreases the complexity of the factor. Complexity of a factor is measured by the number of high loadings within the factor. The most common rotation method is an orthogonal rotation, in which the variables that make up each factor are only highly associated with one factor. The Varimax rotation maximizes the variance of the vectors of the squared loadings in $\widehat{\Lambda}$ causing changes in the magnitude of the loadings. This method of orthogonal rotation is only used with models that contain more than two factors (Rencher, 2002). In SAS, this option is available through ROTATION=VARIMAX in PROC FACTOR.

3.1.4 Choosing the Number of Factors

The appropriate number of factors can be determined using a number of different methods. Perhaps the simplest is through a scree diagram, which plots the eigenvalues against the number of factors. The number of factors is determined by examining the slope of the graph. In general the graph will be steep on the left and then decrease in slopes as it continues to the. The furthest point to the right which is still considered to be a part of the steep slope corresponds to the number of appropriate factors.

Another method to determine the amount of factors to be used is to select a desired amount of common variance to be explained by the factors. More factors are needed to explain a higher proportion of variance. When selecting factors, statisticians commonly choose those that explain either 90 or 95 percent of the variance in the data. Also, a minimum eigenvalue can be specified in order to restrict the number of factors. The minimum eigenvalue should be the average of the eigenvalues, which is 1 if a correlation matrix is used. In SAS, a scree diagram, the proportion of variance, and the minimum eigenvalue can be output by using the SCREE, PROPORTION, and, MINEIGEN options, respectively, in PROC FACTOR.

Additionally, the Root Mean Square (RMS) test can be performed to determine the number of factors. First, a matrix of residuals is obtained and then the sum of the off-diagonal entries is computed. Using SAS, the residual matrix is determined using the RESIDUALS option in PROC FACTOR. The least number of factors which produce a RMS score of less than or equal to some desired value, often 0.05, is the desired number of factors. Once the number of factors is established using one of these methods, the option of NFACT can be used in SAS to specify the exact number of factors to be used.

Factor analysis assumes the data has very few outliers. Outliers impact the correlations between variables and thus can lead to an inaccurate factor analysis. Also factor analysis presumes the data has a linear relationship amongst the variables. With these assumptions met, the factor relationships should be evaluated accurately and meaningfully.

3.2 Cluster Analysis

Cluster analysis, looks for natural groupings of observations and searches for patterns in the data, so that the observations in a particular cluster are highly associated with each other, but not with the observations in other clusters. The simplest method of clustering items is to measure the distance, or similarity, using the joining or tree clustering method. First, the distances between observations are calculated; the Euclidean distance is the most commonly used measure of distance:

$$\text{distance}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2} \tag{3.5}$$

The observations are linked according to the least amount of distance amongst the other observations. After the initial clusters have been formed, a linkage rule must be established in order to determine how these groups can be further clustered, or linked. The single linkage or nearest neighbor rule links two clusters by considering the distance between the two closest objects in the different clusters. Another method is the complete linkage or furthest neighbor rule, which uses the greatest distance between two observations in two different clusters to link the groups. This procedure of linking clusters can be continued until all groups are linked into one final cluster (Johnson, 2002).

The use of factor scores can be easily applied to cluster analysis procedures. Once the factor scores are determined for each of the observations, they can be plotted to ascertain the relation between the factors. That is to say, one can identify patterns that dictate whether increasing one factor will result in an increase or a decrease in another (Rencher, 2002).

3.3 Generalized Linear Models

Generalized linear models are one kind of regression model that involves fitting a model of the form:

$$\phi(E[Y]) = Z'\beta \tag{3.6}$$

where *Y* is assumed to be a random variable from an exponential family, $Z'\beta$ is the linear predictor, and ϕ is the link function (Montgomery, 1997). This method is different from a typical linear regression model because generalized linear models do not assume normal data. Generalized linear models require the response variable to be from a distribution in the exponential family. Using generalized linear models does not require the data to be normal like the least-squares regression models.

SAS incorporates generalized linear models through the procedure GENMOD. The model also uses LINK option to link the response mean to the linear predictor; this option specifies the distribution of the data. The variables listed under the CLASS option are treated as classification variables. This adds each value of the classification variables to the model. Like the logistic procedure, GENMOD uses the chi-squared test statistic to test for significance of each variable. (Johnson, 2002)

3.4 Logistic Regression

Logistic regression is an example of a type of generalized linear model. This method is very similar to the simple least-squares regression as far as results are concerned, but both the mathematics used to get the results and the assumptions of the response variable are different. The response variable of the logistic regression is assumed to be binomial whereas the least squares linear model is assumed to be a normal response variable. Since the logistic regression uses a linear function to fit the data, the response is transformed using the logit function to accommodate the range of the linear function. The logit function is given as

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \text{ where } \pi \in (0,1) \tag{3.7}$$

and used to fit $\text{logit}(\pi) = \alpha + \beta x$ where α and β are parameter vectors of the model. The response for the model is the logarithm of the odds ratio for the mean of the observations π, where π is the expected value for a random variable *Y* that follows a binomial distribution. Note that because the binomial distribution is an exponential family, then logit(π) becomes the link function of a generalized linear model (Allison, 1999).

SAS has a specific procedure for logistic regression called LOGISTIC. The CLASS option allows certain variables to be defined as classification variables. This allows for the values of these variables to be added to the model as individual variables. The LOGISTIC procedure uses the chi-squared test statistic to test the null hypothesis associated with logistic regression (Khattree, 2000).

3.5 Log Likelihood and AIC

A model can be evaluated by the amount of the response variable that is explained by its parameters. The log likelihood procedure measures this by taking the likelihood of a base model and then comparing it to a model with additional parameters. In SAS, the log likelihood can be determined in PROC FACTOR using OUTEST to output the results. The criterion's magnitude increases with the addition of more parameters, so one may be tempted to include as many as possible. However this should be done cautiously since the interpretability of the model may be sacrificed if arbitrary parameters are present. By introducing the Akaike Information Criterion (AIC), the likelihood is penalized by the number of parameters, F, in the model.

$$AIC = -2\log(L) - 2F \tag{3.8}$$

4. Problem Selection

The project group considered several different topics of study before making a final selection. Three distinct problems were defined and the positive and negative aspects of each considered. From there, the group selected the topic that was deemed most substantial, interesting, and potentially innovative to baseball statistics.

4.1 Evaluation of ERA as a Pitching Measure

The first problem considered was to assess the value of the pitching measure, earned run average or, ERA. ERA is the most commonly used statistic used to evaluate major league pitchers. However, it has its limitations, in particular, that it is able to assume the values of zero and infinity in certain situations. In these two cases, a pitcher has either not surrendered an earned run or has not recorded an out for the duration of the season respectively. In either instance, ERA is not a good measure of a pitcher's ability and creates problems when ERA is used in models. ERA is also not necessarily a good measure to use when comparing starting pitchers to relief pitchers. Since each type of pitcher has a different role to fill on the team, causing the length of appearance in a game and the number of appearances to be varied, its usage is further complicated. Ultimately, the goal of this investigation would be not only to evaluate ERA, but also to propose an alternative gauge of pitcher performance.

The ERA problem was a promising choice of topic in that the necessary data would be easily extracted from the databases available to the project group. Also, the data kept for pitching have been kept for many years and seem to be more accurate than data from other areas, such as defense. One drawback to executing this project idea was

that much research has already been done on pitching and the development of viable pitching measures, so whatever this project produced would not necessarily be innovative or important to the statistical and baseball communities. In addition, there was a concern that the amount of work required for this project would not be substantial.

4.2 Paired Comparison Model

A second considered topic involved the creation of a model that would show the likelihood of a win for a team given the team's and its opponent's strengths. The team's current number of wins and losses, home field advantage, and measures for pitching, hitting and defense would all potentially be included in the model, where each of the significant factors would be given a calculated weight.

Since the model would not take into account the information obtained from play-by-play data, the use of the Baseball Archive database would suffice for the development of the model. Hence the data would be readily available and exist in a format that would be easy to work with. Similar to the ERA topic idea, similar projects have already been undertaken by statisticians, and so there existed concern that the project would not be innovative.

4.3 Defensive Model

Although the problems discussed in both of the previous section are reasonable, a third considered topic was selected involving the development of a statistical model to determine the specific defensive factors that are significant to a team's success. Very little has been done with fielding statistics, at least nothing comparable to that which has been done with pitching and hitting statistics. One reason has been the lack of

comprehensive or even accurate data on defense. Another factor in the lack of production of defensive models and measures is the complexities involved in such calculations. Each defensive position demands different responsibilities from a player, and so each, with perhaps the exception of the three outfield positions, requires a different measure; one would not compare the number of double plays turned by a particular shortstop with that of a third baseman. Additionally, recorded defensive data has some level of subjectivity to it, further deterring statisticians from analyzing it.

While it was recognized that such difficulties existed and had the potential to hinder progress, the project group decided that the creation of a defensive model would not only be an exciting challenge but also if successful, an innovative and substantial contribution in baseball analysis. The ultimate goal was to establish a model that would take into account the significant defensive characteristics that are crucial to a team's success and then verify that it was indeed a good model. To accomplish this, it was first necessary to identify all possible qualities or current measures particular to each position that were significant to a team's success as a defensive unit. Each statistic for a different position was treated as a different variable, with the exception of the three outfield positions, which were considered together as one. For example, the number of assists recorded by a first baseman and the number of assists recorded by an outfielder were treated as two different variables with the potential of having different levels of importance to the final model.

Once all possible variables were identified, factor analysis procedures were used to select a determined number of factors for each position that would contribute to the overall final model. Data from the 2000-2004 seasons were used to fit the model. With

the defensive factors identified, they, along with pitching, hitting, and divisional factors, were then used to generate a model that determines the winning percentage of a team for a given season. More importantly, this model describes the relative importance of each individual defensive factor to a team's overall success.

5. Procedure

Before investigating the defensive statistics, the project group developed a rough formulation of a logistic regression model for winning percentage in terms of offensive, defensive, and pitching components. This comprehensive model would be used to rate or rank all teams across a season. The group selected total runs scored as the hitting component and number of earned runs to measure pitching. Basic defensive statistics, including errors, assists, and putouts were added as a temporary stand-in for defense. One major concern that the group had with this model was the lack of a factor to account for the unequal distribution of the number of games that a particular team plays against another team; a team will play the most games against teams in its own division and the least against teams outside of their league. Consequently, the group added a division factor to the model that displays which of the six divisions a team belongs to. Throughout the length of this project, SAS programs were created in order to complete the tasks outlined previously (Appendices F, G, H).

5.1 Determination of Position Qualities

For each of the seven positions, the project group came up with a list of responsibilities and qualities that a player should posses and that could be used to judge his defensive ability (Appendix B). Then it was necessary to determine appropriate statistical measures that could be used to measure those responsibilities and qualities. However, with limited statistical information, namely, that information found in either the Baseball Archive or Retrosheet databases, the group had to eliminate some of the qualities that it had wished to measure as there was no way to extract the necessary data

from the sources. In the end, there ranged between three and eight measures to consider for each position, forty-three in total. Eventually these position qualities were used in a linear combination to create factors to be used in the final model.

5.2 Collection of Team Statistics

Since all of the data stored in the Baseball Archive and Retrosheet databases is sorted by player, the group had to construct a comprehensive database that combined the statistics according to team rather than by player (Appendix G). Team totals for each of the 2000-2004 seasons were recorded for all thirty major league teams. Additionally, the database contains the combined statistics for the five seasons; this was used later in the project.

5.3 Model Selection by Position

In order to determine the best factor analysis for each position, 6 different methods were proposed. The correlation matrix and covariance matrices were used to analyze the original data set. The data was then transformed in four manners: log function, ranking, normalization by the coefficient of variation, and normalization by the coefficient of variation squared. A factor analysis was run for each of these transformed data sets using the covariance matrix.

Additionally, each position and method had a varied number of factors included. The number of included factors was determined using two methods. First, the number of factors was chosen according to the minimum eigenvalue rule. The set eigenvalue was 1 when the correlation matrix was used and the average of the eigenvalues when the covariance matrix was used. The factors selected were those which had corresponding

eigenvalues greater than 1 and the average, respectively. Second, the number of factors that explained either 90 or 95 percent of the variation in the model was selected. This procedure would generally limit the number of factors to 1-2 for each of the 6 methods.

The group considered 3 criteria, consistency, AIC, and interpretability, to select the best method for each position. For consistency, the factors' loadings were compared from year to year and were then ranked on a scale of 1 to 5, where 1 corresponded to all the years having the same factor loadings and 5 to all of the years being different. The AIC measured the contribution to the model with a penalty for number of factors and was later ranked. Each of the methods was evaluated for interpretability that considered the number of factors and the type of information that was explained by the factor loadings. After the 3 rankings were computed, their sum for each method provided an overall ranking. The group selected the method for each position using this overall ranking (Appendix D). For further details on how the project group made its selections please refer to section 6.1.

5.4 Data Adjustment

Throughout the various analyses and models, the group encountered problems with the pitching results. Year to year, the results lacked consistency and interpretability. In an attempt to alleviate the problem, the group eliminated one pitching variable (number of steal attempts against) and combined two others (number of line drives fielded and number of ground balls fielded). The first variable was eliminated because the group determined that the characteristic that it was trying to measure, namely a pitcher's ability to prevent stolen bases, was already described by the pickoff factor. As for the combining of the number of ground balls fielded and line drives caught, both

measure the pitcher's reaction time and ability to cleanly field a ball, and thus the combining of the two would simply help eliminate to reduce the large number of variables being used to describe the pitcher. Once the pitching variables were adjusted, the group ran the factor analyses again.

5.5 Naming the Factors

After the factors were determined for each position, decisions were made in order to name each factor. For example, the first factor for outfield had a factor loading which was most concentrated in assists. Knowing that outfield assists require a strong, accurate, and timely through, it was thus decided that this was a measure of the fielder's arm ability, and hence given the factor name, "arm." A full listing of factor names can be found in Appendix K. Sometimes, even though the factor loadings were similar from position to position, different names were designated due to the nature of the position. For example, a catcher's agility can be described as his ability to make outs on bunts; however, this same characteristic of making outs on bunts to a first baseman more accurately describes his reaction time.

5.6 Model Validation with Cluster Analysis

The project group wanted to verify that the factors selected for each position were associated with available assessments of relative defensive performance. *Gold Glove* winners, while whether they are the absolute best defensive player for their position in a season is debatable, are no doubt some of the most capable players in the field. Therefore, the group decided to compare Gold Glove recipients with a random selection of position players receiving a significant amount of time in the field to see if the models

would first, indicate that the Gold Glovers are good defensive players, and second, that they would generally be rated higher than non-recipients.

For each position, a total of 25 players were selected from the 2000-2004 seasons (Appendix C). Ten were the Gold Glove recipients, with outfield being an exception. For outfield, 30 Gold Glovers were awarded from 2000 to 2004, of which the group randomly selected 10 for database inclusion. Fifteen additional players not receiving a Gold Glove for a particular season were selected as well. These 15 came from a population of each player who played the greatest number of innings for his team in a given season. A random sampling was made in order to avoid selection bias. Thus, there were 150 (450 for outfield) players from which 15 were randomly chosen to be used in the validation procedure.

Each player's stats were tabulated, and then normalized by the time played in the field expressed as outs. These normalized values were then used in a cluster analysis which plotted each factor for a given position against its other factors (Appendix I). For each model, 2-D plots were constructed (1 for 2-factor models, 3 for 3-factor models, and 6 for 4-factor models), and also, where appropriate, 3-D plots. Then any patterns or separation between gold glove players and non-gold glove players on each scatter plot was noted. Also, the project group looked to verify that gold glove players were located near the ends of the plot that would best indicate a strong defensive player.

5.7 Final Logistic Model

With the defensive component complete, the group combined it, along with the pitching, hitting, and division factors, to create the overall final model using logistic

regression. The group then noted the sign of each parameter's coefficient for consistency.

6. Results

The project group found the results from several statistical procedures. First deciding upon the best factor method for each position was determined by ranking three different criteria. After the factors were established, a logistic regression model was developed and validated utilizing cluster analysis.

6.1 Ranking Factor Methods

Perhaps the most important decision made during the project was the determination of the factor analysis methods for each position. For this reason, three criteria were used to make the selections: interpretability of factors, consistency across seasons, and importance to the model (measured by AIC). Each position was considered individually in order to obtain the optimal representation of defensive contribution (Appendix D).

The best way to explain the selection process is to go through the step-by-step procedure and results for one position; for this example second base will be used. Table 3 shows the description and rankings of the three criteria for several methods considered for second base. The consistency was determined by comparing each year's factor loadings. If the variables with the greatest magnitudes matched in all of the factor loadings from one year to another, it was considered to be consistent. So the best consistency description would be 5/5, representing all 5 years having the same pattern of factor loadings. For second base, the consistency description ranged from 2/5 to 5/5, with 3/5 describing the consistency of the selected method.

Method	*Number of Factors*	*Reason for Selected Number of Factors*	*Consistency Description*	*Interpretability Description*	*AIC*	*Consistency Rank*	*Interpretability Rank*	*Log Likelihood/ AIC Rank*	*Final Ranking*
cov	2	Mineigen	5/5	assists; Putouts	0.199	1	3	4	8
cov	3	95%	4/5	doubleplays; putouts; Assists	0.193	2	2	3	7
corr	3	Mineigen	2/5	errors/err on gb; doubleplays; relaythrows	0.133	4	5	1	10
log	2	mineigen/90%	5/5	errors/err on gb; relaythrows	0.37	1	4	10	15
log	3	95%	3/5	errors/err on gb; relaythrows; --	0.339	3	9	8	20
rank	**2**	**Mineigen**	**3/5**	**errors/err on gb; assists/doubleplays**	**0.174**	**3**	**1**	**2**	**6**
rank	4	90%	2/5	errors/err on gb; assists/doubleplays; putouts; relaythrows	0.412	4	7	11	22
rank	5	95%	5/5	errors/err on gb; assists; putouts; relaythrows; doubleplays	0.483	1	8	12	21
norm_v	2	mineigen/90%	5/5	errors/err on gb; relaythrows	0.292	1	4	6	11
norm_v	3	95%	2/5	errors/err on gb; relaythrows; --	0.277	4	9	5	18
norm_cv	2	Mineigen	5/5	errors/err on gb; relaythrows	0.302	1	4	7	12
norm_cv	4	90%	2/5	errors/err on gb; assists/doubleplays; relaythrows; Putouts	0.347	4	6	9	19

Table 3: Ranking Factor Methods for Second Base

Each variable in a factor loading with a magnitude of 0.75 or greater was noted for the interpretability description; all other variables with magnitudes less than 0.75 were not considered important to the overall factor and hence ignored. The ranking of interpretability was subjectively done by the project group, however a few standard rules were established for judgment. Low rankings were given to methods which had a high

number of factors, a factor which did not have any variables with high magnitudes in its loading, and illogical combinations of factor loadings. A good ranking resulted from a factor loading which seemed to explain a valued characteristic for that position. For second base, the two factor rank transformation method was chosen to have the best interpretability. When evaluating these combinations of variables in each factor loading, one could see that the variables were interrelated. Additionally the factors were easily interpretable, since the first factor may be thought of as a second baseman's mistakes on the field and the second factor as his teamwork skills. The factor loadings were consecutively ranked according to their level of interpretability with 1 being the highest ranking. Additionally, when different methods had the same factor loadings, these methods were given an equal ranking.

When ranking the various factor methods, the AIC was also considered, which measured the contribution the factors for each method made to the final model. A smaller AIC value measured a higher contribution and thus received a higher ranking. After the three rankings were completed for consistency, interpretability, and AIC, their sum was determined for the final ranking. Often the decision was made to use the method with the lowest final ranking. A few exceptions were made when the final ranking for a particular method was the lowest but had a very poor score in one of the three criteria. The methods used for each position are shown in Table 4.

Position	***Method***	***Number of Factors***
Pitchers	Correlation	3
Catchers	Rank	3
First Base	Norm_CV	2
Second Base	Rank	2
Third Base	Correlation	3
Shortstop	Rank	4
Outfielders	Log	2

Table 4: Final Factor Methods

6.2 Logistic Regression Model

Along with pitching, offense, and divisional elements, the logistic regression was formulated using the aforementioned factors as the defensive portion of the model. Whether or not an intercept would be appropriate for the model was debated. While making this decision, the following scenario was considered: if no information about a team was available and no intercept was included in the model, a team's winning percentage according to the model would be 0.500. Given this situation, the model seems to produce an inaccurate winning percentage. Since a team will always have statistics associated with it, however, this situation is very unrealistic. An intercept would also attempt to explain some of the noise or variation not explained by the parameters in the model. Seeing as so much of winning percentage is already explained through offense and pitching, the project group felt that an intercept would just hinder the amount of defensive contribution included in the model. For this reason, an intercept is not included in the model.

$$\begin{aligned}\text{Team Winning Percentage} = {} & -0.00266\text{SumOfER} + 0.00245\text{SumOfR} \\ & - 0.0113\text{divisionAC} + 0.00589\text{divisionAE} - 0.0149\text{divisionAW} - 0.0183\text{divisionNC} + 0.0206\text{divisionNE} \\ & - 0.00234P_{\text{Errors}} + 0.0160P_{\text{Reaction Time}} + 0.00660P_{\text{Putouts}} \\ & - 0.00642C_{\text{Teamwork}} + 0.0205C_{\text{Errors}} - 0.0133C_{\text{Agility}} \\ & + 0.00348FB_{\text{Errors}} - 0.0286FB_{\text{Reaction Time}} \\ & - 0.0165SB_{\text{Errors}} + 0.0214SB_{\text{Teamwork}} \\ & - 0.0113TB_{\text{Errors}} - 0.00102TB_{\text{Range}} + 0.0231TB_{\text{Reaction Time}} \\ & + 0.0206SS_{\text{Errors}} - 0.00038SS_{\text{Teamwork}} - 0.00681SS_{\text{Putouts}} - 0.00599SS_{\text{Arm}} \\ & + 0.00764OF_{\text{Arm}} + 0.00885OF_{\text{Errors}}\end{aligned} \tag{6.1}$$

At first glance, the coefficients of the factors seem to conflict with what instinctually one expects. Not all of the error factors had negative coefficients and not all of the putouts and teamwork factors were positive; however, there were several possible reasons for this outcome. Since an orthogonal rotation was used for factor analysis, all of the factors within a position were uncorrelated with each other. This does not guarantee that a factor from one position is uncorrelated with a factor from another position. Seeing as the model measures the defensive contribution from all of the positions, the coefficients may vary due to the factors' dependence. With this in mind, a correlation matrix was constructed. One such instance can be seen with C_Errors and OF_Assists, where a relatively high negative correlation is present. It seems reasonable to assume that a team with outfielders adept at making assists, many of which correspond to outs made at home plate, would have catchers with somewhat lower error totals as the throws from these skilled outfielders would likely not be error-causing. Meanwhile, if few outfield assists are recorded for a team, it is possible that instead of a simple lack of opportunity, that the low total is a result of many throws resulting in errors, a portion of which would be credited to the catcher (Appendix E).

	P_Errors	P_Reaction_Time	P_Putouts	C_Teamwork	C_Errors	C_Agility	FB_Errors	FB_Reaction_Time	SB_Errors	SB_Teamwork	TB_Errors	TB_Range	TB_Reaction_Time	SS_Errors	SS_Teamwork	SS_Putouts	SS_Arm	OF_Arm	OF_Errors
P_Errors	1.00																		
P_Reaction_Time	0.00	1.00																	
P_Putouts	0.00	0.00	1.00																
C_Teamwork	-0.04	0.07	-0.13	1.00															
C_Errors	-0.23	0.07	-0.05	0.00	1.00														
C_Agility	-0.06	0.39	0.04	0.00	0.00	1.00													
FB_Errors	0.14	-0.01	-0.11	0.06	-0.16	0.14	1.00												
FB_Reaction_Time	-0.06	0.27	0.05	0.21	0.17	0.37	0.00	1.00											
SB_Errors	-0.08	0.00	0.05	-0.06	-0.05	0.02	0.18	-0.06	1.00										
SB_Teamwork	-0.04	-0.19	-0.02	0.19	0.26	-0.20	-0.13	0.00	0.00	1.00									
TB_Errors	0.18	0.05	-0.02	-0.07	-0.17	0.03	0.10	-0.15	0.12	-0.04	1.00								
TB_Range	0.06	-0.16	0.19	-0.20	0.10	-0.19	-0.13	-0.22	-0.07	0.14	0.00	1.00							
TB_Reaction_Time	-0.06	0.28	-0.20	-0.01	0.17	0.34	-0.08	0.27	-0.08	-0.15	0.00	0.00	1.00						
SS_Errors	-0.01	-0.07	0.20	-0.05	0.06	-0.04	-0.15	0.01	-0.13	-0.05	-0.11	0.07	-0.07	1.00					
SS_Teamwork	0.06	0.14	-0.01	-0.23	-0.19	0.25	0.25	0.00	-0.02	-0.61	0.14	-0.22	0.10	0.00	1.00				
SS_Putouts	0.12	-0.24	0.11	-0.21	-0.03	-0.21	0.00	-0.20	0.16	-0.11	-0.07	0.14	-0.34	0.00	0.00	1.00			
SS_Arm	0.00	-0.10	-0.08	-0.10	-0.04	0.04	0.10	-0.01	0.08	-0.06	0.10	0.00	-0.06	0.00	0.00	0.00	1.00		
OF_Arm	0.13	-0.09	0.04	-0.13	-0.32	0.13	0.05	-0.06	0.03	-0.10	0.01	-0.05	-0.12	-0.17	0.09	0.12	0.11	1.00	
OF_Errors	0.06	0.15	0.09	-0.03	-0.21	0.24	0.04	-0.04	-0.01	-0.25	0.19	-0.02	0.14	0.08	0.16	-0.04	-0.18	0.00	1.00

Table 5: Factors Correlation Matrix

Even though defense does play a role in baseball, the contribution may not be able to be measured through a statistical model. Unfortunately, there may not have been enough variation present that was not explained through hitting and pitching alone in the studied seasons. Dividing the remaining variation amongst several players' defensive abilities could be ineffective, which may explain why many of the factors in the model have large p-values. Moreover, the variation that the project group aimed to explain using defense may simply just be attributed to noise.

6.3 Player Validation

The results of the cluster analysis were mixed (Appendix I). About a third of the plots were intuitively what was expected, showing a player's good qualities as being

positively related with each other and negatively related to errors. For instance, a shortstop's ability to produce putouts and relay throws are positively associated as shown in Figure 2, so that a player who makes many relay throws has a tendency to produce many putouts as well.

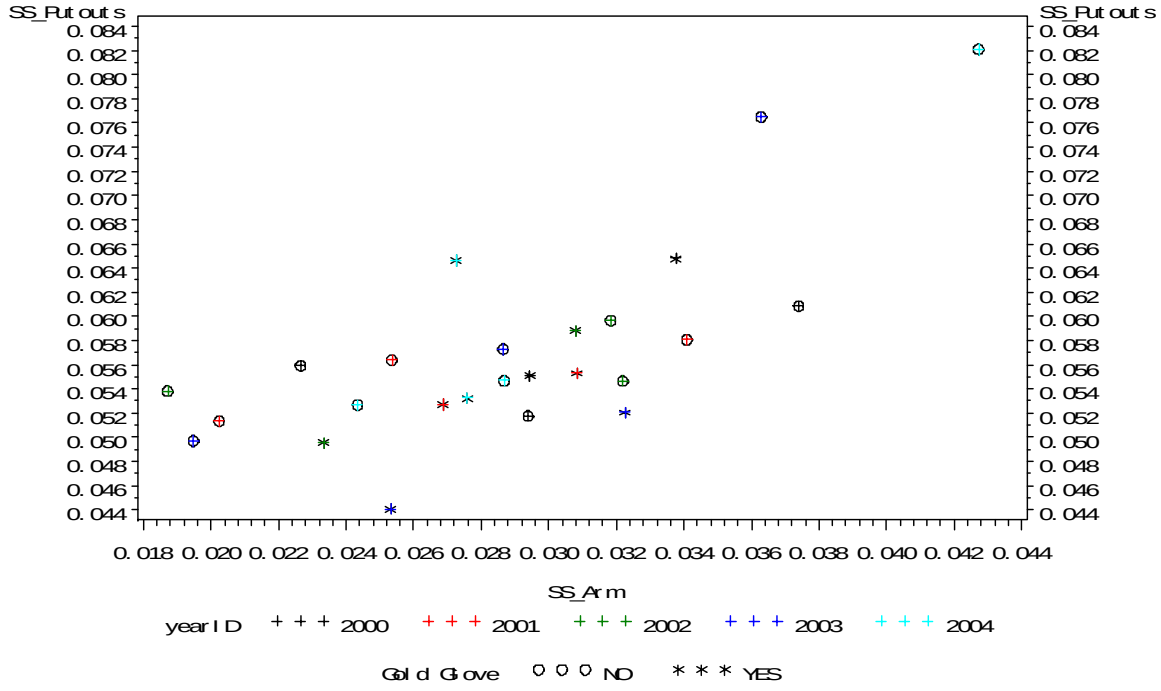


Figure 2: Cluster Analysis - Shortstop Putouts vs. Arm

This relation was not always apparent, however, as some plots displayed no association. Since the factors within each position are uncorrelated with each other, this is only to be expected in at least some instances. As shown in Figure 3, a shortstop's arm, or the player's ability to make relay throws, seems to have no effect on the number of errors he makes.

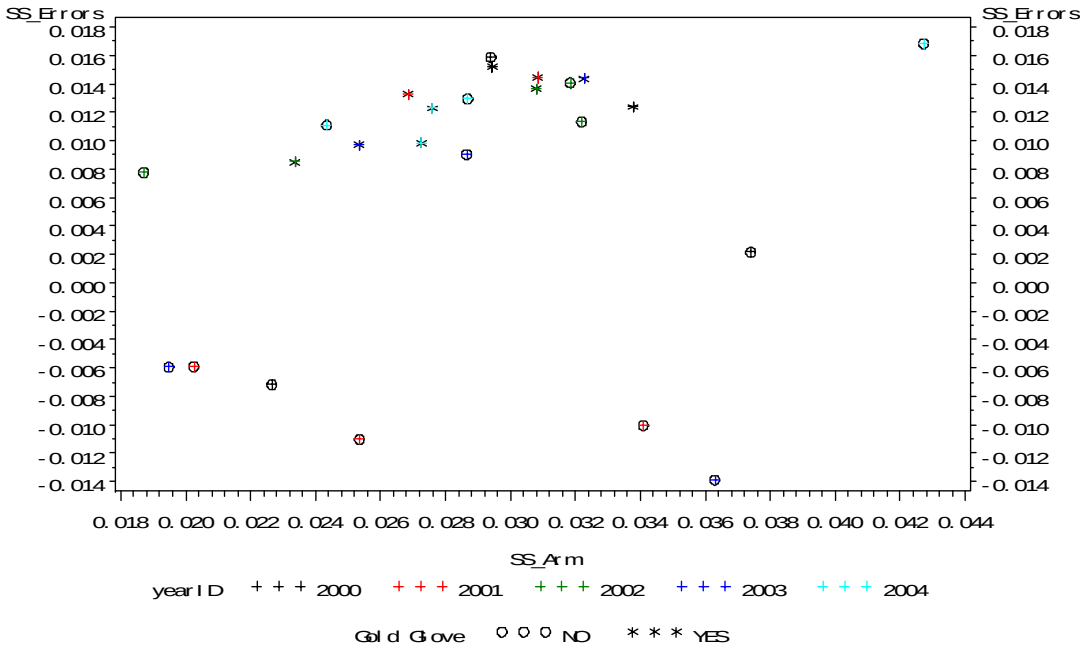


Figure 3: Cluster Analysis - Shortstop Errors vs. Arm

Additionally, some plots were not consistent with the initial expectations, where in some instances, a positive relation existed between errors and a positive characteristic, such as teamwork. Under further assessment, however, this may be explained by the relation to a player's number of opportunities on the field. For example, one may think an outfielder with a large number of successful assists may have a small number of errors, but the converse may be true due to the two quantities being proportional to the amount of opportunities that player has. In other words, an outfielder's number of assists and errors are both likely to increase with increased opportunities to field the ball, which is illustrated in Figure 4.

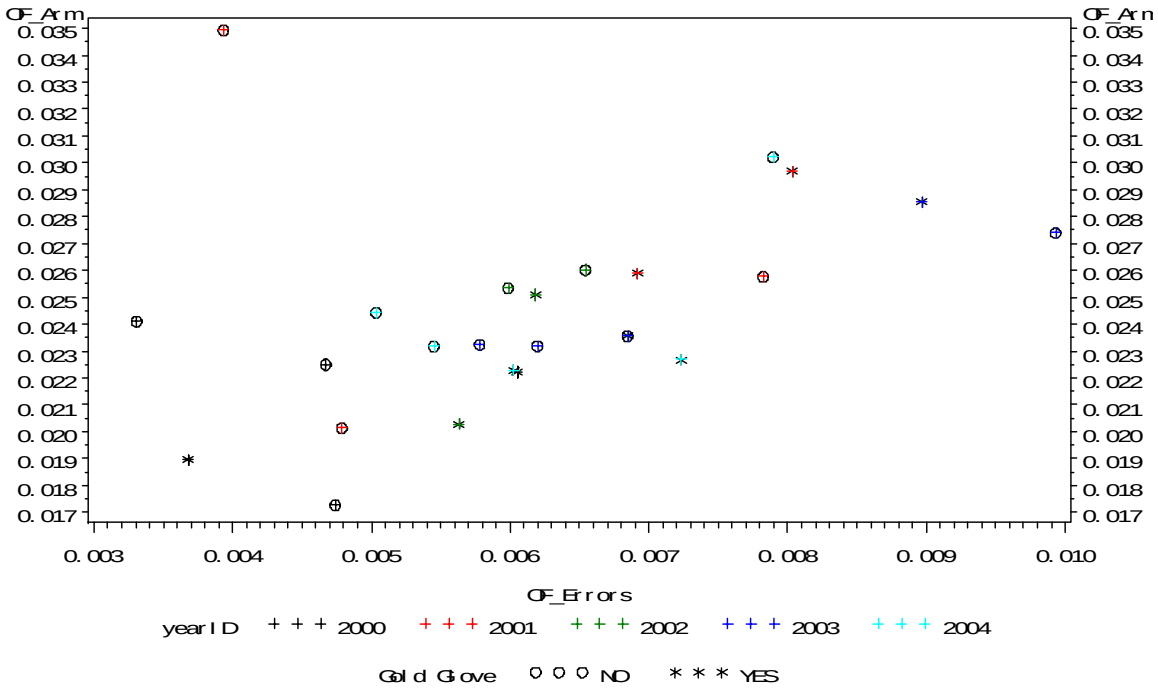


Figure 4: Cluster Analysis - Outfield Errors vs. Arm

The Gold Glove recipients were mixed with the other players in many of the plots. This may have been caused by the manner in which the other players were selected. Since the project group was unable to define an objective way to select a sample of "bad" defensive players, a random selection of players who had appeared in the most games was used; for each team, the player with the greatest number of innings played in the field for each of the 9 positions was included in the group from which the sampling was made. Even though these players were not awarded the Gold Glove for the season, undoubtedly they did not see the most time in the field without just cause. The majority of these players would be the best defensive players available for their position, and so their numbers would not be too different from Gold Glove winners. In addition, Gold Glove recipients are selected through a vote, where popularity and other subjective factors and biases can influence the outcome. Therefore, since Gold Glove players may

not be significantly better than the other players included in the validation procedure, the mixed results of the cluster analysis are not of great concern.

7. Conclusion

At the outset of this project, the project group set several goals to which it focused its efforts. First, the group wanted to identify the variables that are most important to a baseball team's defensive production. A second goal was to explore defensive characteristics in terms of the individual positions. Additionally, the project group sought to develop a way to model defensive contribution to a team's overall success. And finally, the resulting model was to be validated through cluster analysis and other appropriate statistical procedures. Of these four generalized goals, the majority were met, namely, the identification of variables related to team defensive production, the exploration of positional defensive characteristics, and the execution of validation procedures.

7.1 Discussion of Results

While the overall model did not completely satisfy all of the goals put forth at the start of the project, there were several significant findings that are interesting to note.

7.1.1 Positional Defensive Characteristics

Before compiling all of the positional factors with the pitching and hitting components, the project group looked at the defensive factors for each position separately. In general, the number and composition of factors determined for each position appeared to be comprehensive enough to fully describe the responsibilities of that position. Intuitively, the factors for each position made sense, and with support from the cluster analysis performed, it appears that the project group successfully identified the most relevant factors for each defensive position. Even though their contribution to the

overall success of a team may not be significant, the factors do describe the defensive roles of individual positions well.

7.1.2 Importance of Defensive Contribution

One important observation that the project group was able to make during the course of the project was the importance of defense relative to the importance of offense and pitching. That is to say, that the results of this analysis suggest that the defense is far less significant than either offense or hitting. In the final model, none of the factors related to defense were significant, however both pitching (earned runs) and hitting (runs scored) factors were. In recent years, teams have weighed the benefits of signing players known more for their defensive abilities or players that are known to generate a lot of offense. The findings suggest that teams ought to focus on acquiring players that will be able to generate more runs, versus a player who excels defensively.

7.1.3 Model Weaknesses

There are some weaknesses with the model that can and should be addressed. As discussed in the previous chapter, so much of the variation in the data is explained by offensive and pitching contributions. With little left to explain, the defensive portion of the model is not as significant as had been anticipated at the start of this project. Also, concerns with the model's applicability to future seasons is questionable; with so many variables involved in the model, this measurement would probably never be used by the everyday fan. That does not mean, however, that baseball statisticians and professionals associated with baseball and the major leagues would not use such a model, as their resources are greater.

7.2 Recommendations for Further Study

Unfortunately, certain constraints existed that prevented the project group from addressing many of the shortcomings with the model. The project group had a limited amount of time in which to complete the model, and so not every idea or refinement on the model could be executed. Therefore, there are several suggested recommendations for future work on or possible continuation of the project.

7.2.1 Exploration of the Ballpark Factor

Early in the process, the project group considered investigating the effect that a team's ballpark has on the defensive production of a team and its opponents. Each ballpark is unique; each has different dimensions, distinct obstacles and features, varying ground conditions, and a limitless number of characteristics that make it one of a kind. Accordingly, players cannot be expected to perform the same defensively from park to park. In addition, hitters often adjust their hitting style in an attempt to put the ball in play in the most advantageous location on the field, and hence different defensive positions may see a greater number of plays than others in a particular ballpark. Because of these considerations, the project group recognizes this weakness in its model, and recommends that future studies be conducted on the effect that ballparks have on defensive performance.

7.2.2 Exploration of the Relation between Defense and Pitching

The type of play that results from each at bat is dependent not just on the hitter, but on the pitcher as well. Pitchers are often categorized into various groups according to their repertoire of pitches that they typically use (curve ball, slider, knuckleball, etc.), the

type of out they most commonly produce (groundball pitcher, strikeout pitcher, etc.), the average length of appearance (starters, relievers, specialists, etc.) and whether they are left- or right-handed. All of these factors have an effect on the number and type of balls each fielder sees. Therefore, the relation between the defensive performance of a particular player or team and the pitchers that they play behind would be an interesting element to add to the model produced by this project.

7.2.3 Consideration of Distribution of Plays

The project group also considered an alternate method of weighting the final model. Concerns were voiced regarding the uneven distribution of plays involving different positions, which could potentially make the factors relating to one position more important than another's factors. In order to investigate whether this concern was justified, members of the project group determined the total number of plays that each position took part in during an entire season; these totals included every play made by any player for any team for a given position (Appendix J). Although this information was not utilized further during the course of the project, the project group acknowledges that enough disparity exists between the positions in the amount of plays in which each is involved in that further exploration of this topic could be worthwhile.

7.3 Concluding Remarks

Inclusion of any of the abovementioned recommendations has the potential to improve both the overall and position-specific models proposed by the project group. At the same time, the nature of the data itself may not lend itself to many improvements. As discussed in the previous chapter, so much of the variation in winning percentage can be

described by pitching and offensive production that it is possible that no measure of this type can accurately model the relatively small contributions made by defensive production. Nonetheless, the project group is confident that if an ideal defensive model does exist, that the results of this MQP are at the very least, a strong foundation on which to find such a model.

Glossary

At bat (AB) – Every time a player bats except when the following takes place: a walk, hit by pitch, sacrifice bunt, sacrifice fly ball, advancement to first base due to interference, the inning ends while still at bat, replacement by another player while at bat

Balk (BK) – A rule which penalizes the pitcher if he deceives the runner and intends to catch him off balance. If called, each base runner advances one base.

Bunt – A soft hit of the ball into a few feet of the infield which causes the catcher, pitcher, or infielder to field it.

Earned Runs (ER) - A run that is scored without an error, passed ball, obstruction, or catcher's interference occurring.

Error (E) – A mistake made in defense which aids the opposing team

Gold Glove – An annual award given to the best defensive player in each league for each position

Ground Ball – A ball hit which rolls on the ground

Hit by Pitch (HBP) – When a batter is struck by a pitch while within the batter’s box and is advanced to first base.

Hit (H) – A ball which is batted into fair territory and lets the batter reach a base without an error occurring.

Home Run (HR) - A base hit where the batter is able go around all bases before being called out.

Opponent Batting Average (OBA) - Combined batting average of all batters against a particular pitcher

Pop Fly – A hit into the air that often results in an out but can cause base runners to either advance or score runs

Relief Pitcher – A pitcher who comes into the game after the starting pitcher is taken out. He is expected to pitch for shorter periods of time, such as a couple of innings or even just to one batter.

Sacrificing – A hit which results in an out but advances a runner already on base

Save (S) – Earned by a relief pitcher when he enters the game with his team leading by less than 3 runs and is still pitching at game's end with his team victorious

Starting Pitcher – A pitcher who begins playing in the first inning and is expected to pitch for most of the game.

Strikeout (SO) – An out created when three strikes are called when a player is at bat.

Strike zone – The area where a pitch is counted as a strike to a batter, defined as the area directly over the homeplate which ranges from the batter's knees to his mid-chest.

Total Chances – The total number of putouts, assists, and errors made by a fielder or, the total play opportunities

Walk (W) – An advancement of a batter to first base when the pitcher throws four pitches outside of the strike zone and the batter does not swing at them.

Walks and Hits Per Innings Pitched (WHIP) – The number of hits and walks (totaled) per inning

Wild Pitch (WP) – The number of pitches determined to be too erratic for the catcher to handle and result in the advancement of a base runner

Bibliography

Albert, J. (2001a). Hitting with Runners in Scoring Position. Technical report, Department of Mathematics and Statistics, Bowling Green State University.

Albert, J. (2001b). Using Play-by-Play Baseball Data to Develop a Better Measure of Batting Performance. Technical report, Department of Mathematics and Statistics, Bowling Green State University.

Albert, J. (2002). Smoothing Career Trajectories of Baseball Hitters. Technical report, Department of Mathematics and Statistics, Bowling Green State University.

Albert, J., & Bennett, J. (2001). *Curveball*. New York: Springer-Veriag New York, Inc.

Albert, J. , and Cochran, J. J. (2005). Introduction to the Baseball Chapters. In J. Albert, and Cochran, J. J. (Eds.), *Anthology of Statistics in Sports* (pp. 61-66) Philadelphia, PA: SIAM.

Allison, P. D. (1999). Logistic Regression Using the SAS System. Cary, NC: SAS Institute Inc.

Baseball statistics. (2005). *Wikipedia*. Retrieved September 29, 2005, from http://en.wikipedia.org/wiki/Baseball_statistics

Bennett, J. M., & Flueck, J. A. (1983). An Evaluation of Major League Baseball Offensive Performance Models. *American Statistician.* **37**(1): 76-82.

Grabiner, D. The Sabermetric Manifesto. http://baseball1.com/bb-data/grabiner/manifesto.html

Hakes, J. K., & Sauer, R. D. (2004). A Probability Based Measure of Productivity in Major League Baseball With Application to the Questions of Clutch Performance & the Value of Pitching. Technical report, Department of Economics, Clemson University.

Johnson, R.A. & Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis*, Upper Saddle River, NJ: Prentice-Hall Inc.

Khattree, R.& Naik. D. N. (2000). *Multivariate Data Reduction and Discrimination with SAS Software*, Cary, NC: SAS Institute Inc.

MLB.com. (2005). Retrieved September 25, 2005, from http://mlb.mlb.com/NASApp/mlb/index.jsp

Montgomery, D.C. & Myers, R. H. (1997). A Tutorial on Generalized Linear Models. *Journal of Quality Technology*. **29**(3): 274-291.

Neyer, R. (1999). Clutch Hitting. http://www.diamond-mind.com/articles/neyerclutch.htm

Nutting, A. W. (2004). The Determinants of Winning a Gold Glove. *Cornell University: Department of Labor Economics.*

Rechner, A. C. (2002). *Methods of Multivariate Analysis*, New York: John Wiley & Sons, Inc.

Roney, S. (2003). Bill James' Abstracts. http://www.sabr.org/sabr.cfm?a=cms,c,191,3,0

Rosciam, C. (2005). A Look At Retrosheet's Triple Plays. http://www.retrosheet.org/Research/RosciamC/Triple%20Plays%20Analysis.pdf

Sandalow, M., & Sutton, J. (2004). *Ballparks: a Panoramic History*. Edison, NJ: Chartwell Books, Inc.

Spalding, A. G. (1992). *America's National Game*. New York, NY: University of Nebraska Press.

Tangotiger. (2004). Evaluating Catchers. http://www.retrosheet.org/Research/Tangotiger/Catchers.pdf

Weigand, J. (2004). Rating the Catchers. http://www.retrosheet.org/Research/WeigandJ/Rating%20The%20Catchers.pdf

Appendices

Appendix A: Retrosheet Event Files

The following is a list of the available descriptors found in the event codes from project Retrosheet. Each listing corresponds to a column heading in the database that describes a particular aspect of every play.

(0*) - Game id: Game ID following the format described in the "data.doc" file.

(1*) - Visiting team: Visiting team name

(2*) - Inning: Inning in which this play took place.

(3*) - Batting team: A one-character identification of the team at bat ("0" for the visiting team and "1" for the home team).

(4*) - Outs: Number of outs before this play.

(5*, 6*, 7) -Balls, Strikes, Pitch sequence: These three consecutive fields present the pitch information for this play.

(8*) - Vis score: Number of runs for the visiting team before this play.

(9*) - Home score. Number of runs for the home team before this play.

(10) - Batter: Player ID codes for the batter.

(11) - Batter hand: One character which describes how the batter batted for this event (L or R).

(12*, 13*) - Res batter and res batter hand: These fields are almost always the same as batter and batter hand. They only differ if the batter is replaced during the time at bat and the final event is charged to the previous batter. For example, if a pinch-hitter is inserted with two strikes and then takes strike three, the strikeout is charged to the first batter (the responsible batter).

(14) - Pitcher: Player ID code for the pitcher.

(15) - Pitcher hand: The hand with which the pitcher throws (L or R).

(16*, 17*) - Res pitcher and res pitcher hand: Counterparts to res batter and res batter hand for those occasions when a pitcher is changed during an at-bat and the first pitcher is charged with the result. For example, if a relief pitcher enters with a three-ball, no-strike count and throws ball four, then the walk is charged to the first pitcher.

(18, 19, 20, 21, 22, 23, 24, 25) - Positions: The next eight fields contain the Player ID codes for the players at each of the eight fielding positions, in numerical sequence by position number.

(26*, 27*, 28) - First runner, second runner, third runner. These three consecutive fields contain the Player ID codes for the runner at each base. If a base is not occupied, then the field has no width and there will be a pair of double quotes with no space between them. For example, Bill Ripken on first as the only runner would look like this:

"ripkb001","","",

With Joe Orsulak on first and Cal Ripken on third, these fields would look like::

"orsuj001","","ripkc001"

(29*) - Event text: The complete description of the play using the format described for the event files.

(30*) - Leadoff flag: A one character descriptor which is T for the first batter of each inning and F for all others.

(31*) - Pinch-hit flag: Another one character flag which is T for pinch-hitters and F for all others.

(32*) - Defensive position: The defensive position currently being played by this batter. It is pinch-hitter (position 11) for pinch-hitters.

(33*) - Lineup position: Position in the batting order for this batter.

(34*) - Event type: There are 25 different numeric codes to describe the type of event. They are:

Code	Meaning	Code	Meaning
0	Unknown event	13	Foul error
1	No event	14	Walk
2	Generic out	15	Intentional walk
3	Strikeout	16	Hit by pitch
4	Stolen base	17	Interference
5	Defensive indifference	18	Error
6	Caught stealing	19	Fielders choice
7	Pick off error	20	Single
8	Pick off	21	Double
9	Wild pitch	22	Triple
10	Passed ball	23	Home run
11	Balk	24	Missing Play
12	Other advance		

(35*) - Batter event flag: A one character indication of whether or not the event terminated the batter's appearance. T = yes, which is most common; F = no, meaning the same batter stayed at the plate, such as after a stolen base.

(36*) - Ab flag: A one character indication of whether batter was charged with at-bat (T = yes, F = no).

(37*) - Hit value: One number indicating value of hit (0 = no hit; 1 = single; 2 = double; 3 = triple; 4 = home run).

(38*) - SH flag: One character indicating sacrifice hit (T = yes; F = no).

(39*) - SF flag: One character indicating sacrifice fly (T = yes; F = no).

(40*) - Outs on play: Number of outs recorded on this play.

(41) - Double play flag: One character field of DP or not.

(42) - Triple play flag: nce character field of TP or not.

(43*) - RBI on play: Number of RBI credited to batter on this play.

(44*, 45*) - Wild pitch flag, Passed ball flag: Two records with indication of whether there was a WP or PB on this play.

(46) - Fielded by: Identity of the fielder who played the ball. This is especially important for base hits when no formal fielding credit is given.

(47) - Batted ball type: Descriptor, which is F (fly ball), L (line drive), P (pop-up), or G (ground ball).

(48) - Bunt flag: Descriptor for whether or not play was a bunt.

(49) - Foul flag: Descriptor for whether or not ball was played in foul ground.

(50) - Hit location: The zone on the field where the ball was hit. Refer to the Scoring System attachments for a diagram of all locations.

(51*) - Num errors: Number of errors on this play (a maximum of three is allowed).

(52, 53, 54, 55, 56, 57) - Error players and types: These are 6 consecutive fields which identify the player committing the 1st, 2nd or 3rd errors on the play and the type of error each was (throw or drop).

(58*) - Batter dest: The base that the batter reached at the conclusion of the play. If he was out, the base is 0.

(59*, 60*, 61*) - Runner dest: The next three fields contain the base reached by each of the three runners at the conclusion of the play. If there was no advance, then the base shown will be the one where the runner started. Note that these runner fields are not updated on plays which end an inning, even if the inning-ending play would have resulted in an advance of one or more runners had it occurred earlier in the inning (5 if scores and unearned, 6 if team unearned).

(62, 63, 64, 65) - Plays: The next four fields indicate the play (if any) made on the batter and each of the runners (if any).

(66, 67, 68) - Stolen base flags: The next three fields contain single character descriptors for each of the runners indicating whether he had a stolen base.

(69, 70, 71) – Caught stealing flags: The next three fields contain single character descriptors for each of the runners indicating whether he was caught stealing.

(72, 73, 74) – Picked off flags: The next three fields contain single character descriptors for each of the runners indicating whether he was picked off.

(75, 76, 77) - Responsible pitcher for runner: The next three fields indicate which pitcher was responsible for the runners on each base, if any. This assignment reflects responsibility should the runner score.

(78, 79) - New game and end game flags: The next two fields set a flag if this is the first record of a new game or the last record of the game.

(80, 81, 82) - Pinch-runners: The next three fields indicate if a pinch-runner has entered the game and at which base.

(83, 84, 85) - Removed runners: The next three fields contain the player ID of the runner who was just run for, one field for each base. If there is no pinch-runner at that base, the field contains the NULL string "".

(86) - Removed batter: If there is a pinch-hitter, this field contains the player ID of the batter removed. If there is no pinch-hitter, this field contains the NULL string "".

(87) - Removed batter position: If there is a pinch-hitter, this field contains the fielding position of the removed batter. If there is no pinch-hitter, this value is 0.

(88, 89, 90) - Fielder putouts: The next three fields indicate the first, second, and third fielders credited with putouts on the play (0 if none).

(91, 92, 93, 94, 95) - Fielder assists: The next five fields indicate which fielders got credited with assists on the play (maximum of five fielders) (0 if none).

(96) - Event num: All events are numbered consecutively throughout each game for easy reference.

Source: Retrosheet. <http://www.retrosheet.org/datause.doc>

Appendix B: Player Qualities and Responsibilities

The project group identified various qualities and responsibilities specific to each position that measure a player's defensive abilities. The following list names these qualities, and where appropriate, identifies the associated statistical measures for each.

Pitcher

Responsibilities

- Keeping runners on their base - # stolen base attempts
- Throwing out runners that try to steal – # of stolen base attempts, # of pickoffs
- Covering 1st base - # of plays made by pitcher at first base*

Qualities

- Few wild pitches while men are on base
- Few balks
- Few errors
- Assists
- Putouts
- Quick reactions - # of ground balls fielded, # of line drives caught

Catcher

Responsibilities

- Ability to block the plate - # of players tagged out at home
- Preventing stolen bases - # of players thrown out by a catcher, # of attempted stolen bases
- Get one of more outs on a bunt attempt
- Game calling*

Qualities

- Few errors
- Few passed balls
- Assists
- Putouts

1st Base

Responsibilities

- Get one of more outs on a bunt attempt
- Ability to field a ground ball - # of errors related to ground balls
- Outs in foul ground

Qualities

- Few errors
- Assists
- Putouts
- Range - # of ground balls to the right fielder*

2nd Base

Responsibilities

- Turning double plays - # of double plays
- Relay throws - # of putouts originating with a 7,8,9 player going to 2nd baseman and then resulting in a putout by a third player
- Ability to field a ground ball - # of errors related to ground balls

Qualities

- Few errors
- Assists
- Putouts
- Range - # of ground balls to the center and right fielders*

3rd Base

Responsibilities

- Get one of more outs on a bunt attempt
- Ability to field a ground ball - # of errors related to ground balls
- Outs in foul ground

Qualities

- Few errors
- Assists
- Putouts
- Range - # of ground balls to the left fielder*

Shortstop

Responsibilities

- Turning double plays - # of double plays
- Relay throws - # of putouts originating with a 7,8,9 player going to shortstop and then resulting in a putout by a third player
- Ability to field a ground ball - # of errors related to ground balls

Qualities

- Few errors
- Assists
- Putouts
- Range - # of ground balls to the center and left fielders*

Outfielders

Qualities

- Arm strength - # of assists
- Speed - # of balls reached*
- Few errors
- Putouts

* Indicates that the quality or measure was not included in calculations due to limitations in availability of data

Appendix C: Player Data for Validation

For each position, 25 players, the 10 Gold Glove recipients from 2000-2004 and 15 non-Gold Glove recipients from the same seasons, were selected for inclusion in the player validation aspect of the project. The following tables contain the relevant defensive statistics for these players.

C.1 Pitchers

nameLast	nameFirst	teamID	yearID	InnOuts	P_Balks	P_Pickoffs	P_Putouts	P_Errors	P_Assists	P_WPMenON	P_GBLD
Batista	Miguel	TOR	2004	596	0	1	18	1	30	12	32
Colon	Bartolo	CLE	2001	667	1	3	12	1	27	4	28
Estes	Shawn	COL	2004	606	2	5	14	1	27	4	26
Lieber	Jon	CHN	2001	697	2	0	16	0	35	4	36
Morris	Matt	SLN	2002	631	0	0	11	0	22	3	25
Moyer	Jamie	SEA	2003	645	0	5	19	1	31	0	30
Ponson	Sidney	BAL	2004	647	2	0	21	1	28	8	31
Sabathia	C.C.	CLE	2002	630	3	1	2	1	19	6	19
Suppan	Jeff	KCA	2001	655	0	3	18	1	33	6	32
Maddux	Greg	CHN	2004	638	0	0	21	1	55	2	60
Maddux	Greg	ATL	2002	598	0	2	21	1	48	1	50
Maddux	Greg	ATL	2001	699	0	4	19	1	54	2	58
Maddux	Greg	ATL	2000	748	2	0	25	2	68	1	73
Rogers	Kenny	TEX	2002	632	1	2	22	3	40	5	45
Rogers	Kenny	TEX	2004	635	1	6	16	1	49	1	45
Rogers	Kenny	TEX	2000	682	1	9	18	2	46	1	42
Mussina	Mike	NYA	2003	644	0	2	14	0	35	4	39
Mussina	Mike	NYA	2001	686	0	2	18	1	25	6	27
Hampton	Mike	ATL	2003	570	1	1	15	1	52	10	55
Lawrence	Brian	SDN	2002	630	1	0	22	0	31	2	37
Thomson	John	TEX	2003		0	0	17	2	36	5	40
Zambrano	Victor	TBA	2003	1	0	10	2	16	565	15	75
Hudson	Tim	OAK	2000	0	3	15	4	20	607	6	22
Kile	Darryl	SLN	2000	1	1	12	1	30	697	8	36
Hampton	Mike	NYN	2000	0	7	9	2	45	653	10	40

nameLast	nameFirst	teamID	yearID	InnOuts	C_PassedBalls	C_Putouts	C_Errors	C_Assists	C_OutsOnBunts	C_TagOutAtHome	C_CaughStealing
Hundley	Todd	LAN	2000	2102	8	554	13	38	8	6	19
Mayne	Brent	COL	2000	2429	5	582	6	35	3	9	17
Alomar Jr.	Sandy	CHA	2001	1635	5	367	4	19	2	8	9
Girardi	Joe	CHN	2002	1847	6	554	6	43	6	5	20
Inge	Brandon	DET	2002	2461	10	484	1	47	2	16	13
Santiago	Benito	SFN	2002	3200	7	739	4	53	9	9	22
Bard	Josh	CLE	2003	2147	4	486	5	55	5	11	19
Kendall	Jason	PIT	2003	3835	9	841	10	48	16	12	16
Posada	Jorge	NYA	2003	3495	13	933	6	75	6	11	26
Matheny	Mike	SLN	2004	2933	2	742	1	58	10	15	10
Matheny	Mike	SLN	2003	3290	5	774	0	49	6	15	10
Matheny	Mike	SLN	2000	3095	4	803	5	75	6	9	42
Rodriguez	Ivan	TEX	2000	2209	2	507	2	34	3	9	9
Rodriguez	Ivan	TEX	2001	2566	2	631	7	52	5	9	18
Rodriguez	Ivan	DET	2004	3153	3	770	11	52	9	19	14
Ausmus	Brad	HOU	2001	3170	1	949	3	62	7	14	25
Ausmus	Brad	HOU	2002	3237	2	942	3	35	9	11	25
Molina	Ben	ANA	2003	2849	4	672	5	62	2	7	22
Molina	Ben	ANA	2002	3043	5	707	1	60	1	10	27
Molina	Ben	ANA	2000	3276	6	685	7	60	3	16	33
Lo Duca	Paul	LAN	2001	2404	4	643	6	53	4	9	28
Hatteberg	Scott	BOS	2001	1745	13	491	4	29	2	7	8
Hernandez	Ramon	SDN	2004	2776	7	753	6	35	8	14	18
Molina	Ben	ANA	2004	2286	6	597	3	56	0	13	17
Johnson	Charles	COL	2004	2239	7	523	7	44	7	10	11

C.3 First Basemen

nameLast	nameFirst	teamID	yearID	InnOuts	FB_OutsInFoul	FB_ErrorsOnGroundBalls	FB_Putouts	FB_Errors	FB_Assists	FB_OutsOnBunts
Colbrunn	Greg	ARI	2000	2197	25	4	647	8	54	8
Galarraga	Andres	ATL	2000	3234	32	9	1104	14	61	5
Giambi	Jason	OAK	2000	3193	36	5	1161	6	59	5
Delgado	Carlos	TOR	2002	3687	39	9	1231	12	95	1
Lee	Derrek	FLO	2002	4324	52	9	1312	12	121	3
Giambi	Jason	NYA	2003	2228	27	4	748	4	19	1
Helton	Todd	COL	2003	4107	41	9	1418	11	156	26
Delgado	Carlos	TOR	2004	3116	33	3	1041	5	88	2
Wilkerson	Brad	MON	2004	2105	24	2	694	4	67	6
McGriff	Fred	CHN	2002	3448	28	6	1004	7	60	0
Snow	T.J.	SFN	2000	3820	41	4	1198	6	92	12
Olerud	John	SEA	2000		42	5	1271	5	133	7
Mientkiewicz	Doug	MIN	2001	3808	48	3	1263	4	69	1
Helton	Todd	COL	2001	4110	37	1	1306	2	119	16
Helton	Todd	COL	2002	4026	46	5	1358	7	112	0
Olerud	John	SEA	2002	3953	46	4	1169	5	101	0
Lee	Derrek	FLO	2003	4061	55	4	1279	5	97	14
Olerud	John	SEA	2003	3861	52	3	1096	3	125	8
Erstad	Darin	ANA	2004	3196	39	2	986	4	66	2
Helton	Todd	COL	2004	3962	37	3	1356	4	114	27
Phillips	Jason	NYN	2003	2047	27	3	666	7	44	2
Konerko	Paul	CHA	2001	3781	36	9	1277	8	90	3
Hatteberg	Scott	OAK	2004	3840	52	6	1281	10	86	1
Cox	Steve	TBA	2001	1923	30	4	569	4	48	0
Brogna	Rico	ATL	2001	1468	15	0	431	3	42	3

nameLast	nameFirst	teamID	yearID	InnOuts	SB_RelayThrows	SB_ErrorsOnGroundBalls	SB_Putouts	SB_Errors	SB_Assists	SB_DoublePlays
Biggio	Craig	HOU	2000	2558	1	6	183	6	282	57
Grudzielanek	Mark	LAN	2000	3854	1	16	288	17	414	97
Lansing	Mike	COL/BOS	2000	2241	1	5	249	7	326	70
Easley	Damien	DET	2001	3988	5	13	278	14	496	113
Walker	Todd	CIN/COL	2001	1629	0	9	293	11	366	82
Butler	Brent	COL	2002	1711	0	7	121	8	178	46
Hart	Bo	SLN	2003	1758	0	4	167	4	180	35
Jimenez	D'Angelo	CHA	2003	1760	5	8	119	11	175	56
Hudson	Orlando	TOR	2004	3374	5	11	276	12	449	90
Sanchez	Rey	TBA	2004	2088	2	4	157	5	234	55
Boone	Bret	SEA	2004	3926	3	13	280	14	350	90
Boone	Bret	SEA	2002	3952	1	4	251	7	387	84
Boone	Bret	SEA	2003	4125	2	6	268	7	426	107
Castillo	Luis	FLO	2004	2853	1	5	275	6	406	97
Castillo	Luis	FLO	2003	3937	1	9	286	10	433	99
Vina	Fernando	SLN	2001	3898	2	7	315	9	383	100
Vina	Fernando	SLN	2002	3937	1	11	287	13	401	104
Alomar	Roberto	CLE	2000	3928	2	8	293	15	436	109
Alomar	Roberto	CLE	2001	3972	4	3	269	5	424	89
Reese	Pokey	CIN	2000	3387	5	8	289	14	393	88
Mike	Young	TEX	2002	3776	0	7	298	9	419	97
Alfonso	Soriano	NYA	2001	4153	2	15	319	19	366	93
Tony	Graffanino	KCA	2004	1891	1	3	185	5	219	67
Luis	Castillo	FLO	2002	3773	1	11	270	13	390	93
Walker	Todd	BOS	2003	3562	0	12	235	16	391	78

C.5 Third Basemen

nameLast	nameFirst	teamID	yearID	InnOuts	TB_OutsInFoul	TB_ErrorsOnGroundBalls	TB_Putouts	TB_Errors	TB_Assists	TB_OutsOnBunts
Batista	Tony	TOR	2000	4101	43	13	120	17	317	11
Chavez	Eric	OAK	2000	3619	33	17	91	18	256	11
Hernandez	Jose	MIL	2000	2259	30	13	81	13	163	7
Fryman	Travis	CLE	2002	2785	17	10	53	10	185	7
Boone	Aaron	CIN/NYA	2003	2147	28	19	98	20	290	14
Burroughs	Sean	SDN	2003	3434	27	12	105	12	238	13
Cirillo	Jeff	SEA	2003	2018	27	4	65	4	108	5
Burroughs	Sean	SDN	2004	3180	27	13	100	14	209	6
Munson	Eric	DET	2004	2222	11	16	51	16	177	9
Rolen	Scott	PHI	2000	3240	29	7	89	10	245	12
Fryman	Travis	CLE	2000	4048	24	5	79	8	276	8
Rolen	Scott	PHI	2001	3988	35	12	104	12	325	20
Chavez	Eric	OAK	2001	3908	29	10	100	12	321	7
Rolen	Scott	PHI/SLN	2002	4080	46	15	133	16	335	9
Chavez	Eric	OAK	2002	3798	33	15	120	17	301	9
Rolen	Scott	SLN	2003	4017	36	9	109	13	298	11
Chavez	Eric	OAK	2003	4000	42	12	125	14	343	12
Rolen	Scott	SLN	2004	3684	32	9	93	10	325	9
Chavez	Eric	OAK	2004	3400	33	11	113	13	276	12
Spiezio	Scott	SEA	2004	1763	22	6	56	7	131	7
Polanco	Placido	SLN	2002	1761	25	6	52	6	139	12
Hillenbrand	Shea	BOS	2002	4098	37	20	101	23	283	3
Boone	Aaron	CIN	2001	2654	15	13	73	19	207	12
Ripken Jr.	Cal	BAL	2001	2945	36	14	97	14	209	4
Koskie	Corey	MIN	2001	3924	32	10	95	15	306	15

nameLast	nameFirst	teamID	yearID	InnOuts	SS_DoublePlays	SS_RelayThrows	SS_ErrorsOnGroundBalls	SS_Putouts	SS_Errors	SS_Assists
Loretta	Mark	MIL	2000	2251	54	3	2	122	2	255
Valentin	Jose	CHA	2000	3637	118	4	33	234	36	456
Clayton	Royce	CHA	2002	2650	72	3	5	166	5	292
Larkin	Barry	CIN	2002	3271	89	0	11	190	12	370
Lugo	Julio	HOU	2002	2148	32	1	8	121	8	205
Reyes	Jose	NYN	2003	1789	42	1	9	108	9	213
Cabrera	Orlando	MON	2004	2603	92	3	11	225	15	437
Clayton	Royce	COL	2004	3723	88	5	9	213	9	418
Eckstein	David	ANA	2004	3575	75	2	5	198	6	309
Vizquel	Omar	CLE	2001	3962	88	4	6	219	7	414
Vizquel	Omar	CLE	2000	3986	99	3	3	231	3	416
Perez	Neifi	COL	2000	4208	120	2	15	288	18	524
Cabrera	Orlando	MON	2001	4220	106	3	9	246	11	515
Rodriguez	Alex	TEX	2003	4109	111	1	7	227	8	464
Rodriguez	Alex	TEX	2002	4172	108	2	9	259	10	472
Renteria	Edgar	SLN	2002	3862	72	1	13	202	19	410
Renteria	Edgar	SLN	2003	4102	83	2	13	191	16	439
Izturis	Cesar	LAN	2004	4158	96	1	8	234	10	430
Jeter	Derek	NYA	2004	4025	96	3	13	273	13	392
Jeter	Derek	NYA	2003	3101	51	2	11	160	14	271
Womack	Tony	ARI	2000	3732	72	2	13	217	18	365
Uribe	Juan	COL	2003	1795	57	1	6	143	11	242
Jeter	Derek	NYA	2001	3937	68	0	12	211	15	344
Gonzalez	Alex	TOR	2001	4123	120	4	7	248	10	511
Jimenez	D'Angelo	SDN	2001	2222	47	3	14	130	21	255

C.7 Outfielders

POS	nameLast	nameFirst	teamID	yearID	InnOuts	OF_Assists	OF_Errors	OF_Putouts
RF	Bell	Derek	NYN	2000	3504	5	3	252
CF	Everett	Carl	BOS	2000	3191	11	6	277
LF	Sanders	Reggie	ATL	2000	1694	7	6	155
CF	Erstad	Darin	ANA	2001	3808	10	1	399
CF	Hunter	Torii	MIN	2002	3704	7	3	364
RF	Cedeno	Roger	NYN	2003	2409	5	3	231
CF	Duncan	Jeff	NYN	2003	1098	0	0	136
RF	Kapler	Gabe	BOS	2004	1772	6	4	170
CF	Kotsay	Mark	OAK	2004	3765	11	6	347
LF	Sledge	Terrmel	MON	2004	1739	5	3	216
CF	Cameron	Mike	SEA	2003	3852	3	4	485
RF	Dye	Jermaine	KCA	2000	3781	11	7	277
CF	Edmonds	Jim	SLN	2002	3480	11	5	347
CF	Hunter	Torii	MIN	2001	3886	14	4	460
CF	Jones	Andruw	ATL	2003	3987	8	3	390
CF	Jones	Andruw	ATL	2001	4306	10	6	461
RF	Suzuki	Ichiro	SEA	2002	3925	8	3	325
CF	Finley	Steve	ARI	2000	3851	10	3	343
RF	Suzuki	Ichiro	SEA	2004	4216	12	3	372
CF	Wells	Vernon	TOR	2004	3405	5	1	327
CF	Beltran	Carlos	KCA	2002	3925	12	7	398
CF	Cedeno	Roger	DET	2001	1614	5	12	236
RF	Cruz Jr.	Jose	SFN	2003	3998	18	2	340
LF	Jenkins	Geoff	MIL	2001	2692	8	3	210
LF	Justice	David	OAK	2002	1172	3	2	125

Appendix D: Ranking of Factor Analysis Results for Each Position

The following tables describe the different factor analysis methods that were considered for each position. Rankings for each, including consistency rankings, interpretability rankings, AIC rankings, and overall rankings are included. Note that the selected method is bolded for each position.

D.1 Pitcher Rankings

Method	# of Factors	Reason for # of Factors	Consistency Description	Interpretability Description	AIC	Consistency Rank	Interpretability Rank	Log Likelihood Rank	Final Ranking
cov	3	mineigen	5/5	putouts; gbld; assists	0.673	1	4	8	13
cov	4	95%	4/5	putouts; assists/gbld; wp	0.774	2	1	9	12
Corr	**3**	**mineigen**	**3/5**	**errors(.72); assists/gbld; putouts(.72)**	**0.477**	**3**	**2**	**3**	8
log	3	mineigen	3/5	errors; pickoffs; balks	0.513	3	5	4	12
log	5	90%	3/5	putouts; wp; balks; errors; pickoffs	0.409	3	6	1	10
rank	3	mineigen	3/5	assists; putouts; wp	0.938	3	3	10	16
rank	5	90%	2/5	assists; errors; wp; putouts; pickoffs	0.449	4	7	2	13
norm_v	3	mineigen	2/5	errors; pickoffs; balks	0.534	4	5	5	14
norm_v	5	90%	4/5	gbld; wp; errors; balks; pickoffs	0.549	2	7	6	15
norm_cv	3	mineigen	2/5	errors; pickoffs; balks	0.612	4	5	7	16

D.2 Catcher Rankings

Method	# of Factors	Reason for # of Factors	Consistency Description	Interpretability Description	AIC	Consistency Rank	Interpretability Rank	Log Likelihood Rank	Final Ranking
Cov	1	mineigen/95%	5/5	Putouts	0.893	1	7	9	17
corr	3	mineigen	2/5	assists; errors(.73); outsonbunts	0.346	4	2	3	9
log	3	mineigen	2/5	errors,; outsonbunts; passedballs	0.314	4	3	2	9
log	4	90%	0/5	Errors; outsonbunts; tagoutathome; passedballs	0.466	5	4	6	15
rank	**3**	**mineigen**	**0/5**	**assists/caughtstealing; errors; outsonbunts**	**0.246**	**5**	**1**	1	**7**
norm_v	4	mineigen/90%	3/5	errors; outsonbunts; tagoutathome; passedballs	0.360	3	4	4	11
norm_v	5	95%	3/5	caughtstealing/assists; outsonbunts; tagoutathome; errors; passedballs	0.496	3	6	7	16
norm_cv	4	mineigen	2/5	caughtstealing/errors; outsonbunts; tagoutathome; passedballs	0.398	4	5	5	14
norm_cv	5	90%	3/5	caughtstealing/assists; outsonbunts; tagoutathome; errors; passedballs	0.521	3	6	8	17

D.3 First Base Rankings

Method	# of Factors	Reason for # of Factors	Consistency Description	Interpretability Description	AIC	Consistency Rank	Interpretability Rank	Log Likelihood Rank	Final Ranking
cov	1	mineigen	5/5	putouts	0.395	1	8	13	21
cov	2	95%	5/5	putouts; assists	0.697	1	5	14	19
corr	2	mineigen	4/5	errors/err on gb; assists	0.274	2	1	4	5
log	2	mineigen	5/5	errors/err on gb; outsonbunts	0.359	1	3	10	13
log	3	90%	3/5	errors/err on gb; outsinfoul; outsonbunts	0.285	3	2	6	8
log	4	95%	4/5	errors/err on gb; outsonbunts; outsinfoul; assists	0.342	2	6	9	15
rank	2	mineigen	2/5	errors/err on gb; assists	0.082	4	1	1	2
rank	4	90%	2/5	errors/err on gb; outsonbunts; outsinfoul; putouts	0.277	4	6	5	11
rank	5	95%	5/5	errors/err on gb; outsonbunts; assists; outsinfoul; putouts	0.370	1	7	11	18
norm_v	2	mineigen	5/5	errors/err on gb; outsonbunts	0.302	1	3	7	11
norm_v	3	90%	4/5	errors/err on gb; gbld; outsonbunts	0.256	2	4	3	9
norm_v	4	95%	4/5	errors/err on gb; outsonbunts; gbld; assists	0.316	2	6	8	16
norm_cv	**2**	**mineigen**	**5/5**	**errors/err on gb; outsonbunts**	**0.243**	**1**	**3**	**2**	**6**
norm_cv	4	90%	5/5	errors/err on gb; outsonbunts; assists; gbld	0.374	1	6	12	19

D.4 Second Base Rankings

Method	# of Factors	Reason for # of Factors	Consistency Description	Interpretability Description	AIC	Consistency Rank	Interpretability Rank	Log Likelihood Rank	Final Ranking
cov	2	mineigen	5/5	assists; putouts	0.199	1	3	4	8
cov	3	0.95	4/5	doubleplays; putouts; assists	0.193	2	2	3	7
corr	3	mineigen	2/5	errors/err on gb; doubleplays; relaythrows	0.133	4	5	1	10
log	2	mineigen/90%	5/5	errors/err on gb; relaythrows	0.370	1	4	10	15
log	3	0.95	3/5	errors/err on gb; relaythrows; --	0.339	3	9	8	20
rank	**2**	**mineigen**	**3/5**	**errors/err on gb; assists/doubleplays**	**0.174**	**3**	**1**	**2**	**6**
rank	4	0.9	2/5	errors/err on gb; assists/doubleplays; putouts; relaythrows	0.412	4	7	11	22
rank	5	0.95	5/5	errors/err on gb; assists; putouts; relaythrows; doubleplays	0.483	1	8	12	21
norm_v	2	mineigen/90%	5/5	errors/err on gb; relaythrows	0.292	1	4	6	11
norm_v	3	0.95	2/5	errors/err on gb; relaythrows; --	0.277	4	9	5	18
norm_cv	2	mineigen	5/5	errors/err on gb; relaythrows	0.302	1	4	7	12
norm_cv	4	0.9	2/5	errors/err on gb; assists/doubleplays; relaythrows; putouts	0.347	4	6	9	19

D.5 Third Base Rankings

Method	# of Factors	Reason for # of Factors	Consistency Description	Interpretability Description	AIC	Consistency Rank	Interpretability Rank	Log Likelihood Rank	Final Ranking
cov	2	mineigen	5/5	putouts; assists	0.285	1	4	9	14
cov	3	0.95	5/5	errors/err on gb; putouts; assists	0.299	1	1	12	14
corr	**3**	**mineigen**	**3/5**	**errors/err on gb; utouts/outsinfoul; outsonbunts/assists(.74)**	**0.161**	**3**	**3**	**2**	**8**
log	2	mineigen	5/5	errors/err on gb; outsonbunts	0.165	1	5	3	9
log	3	0.9	5/5	errors/err on gb; outsinfoul/putouts; outsonbunts	0.295	1	7	10	18
log	4	0.95	2/5	errors/err on gb; outsinfoul/putouts; outsonbunts; --	0.388	4	10	13	27
rank	3	mineigen	2/5	errors/err on gb; putouts/outsinfoul; assists/outsonbunts	0.181	4	2	4	10
rank	4	0.9	5/5	errors/err on gb; putouts/outsinfoul; outsonbunts; assists	0.297	1	6	11	18
rank	5	0.95	4/5	errors/err on gb; assists; outsinfoul; outsonbunts; putouts	0.428	2	8	14	24
norm_v	2	mineigen	5/5	errors/err on gb; outsonbunts	0.123	1	9	1	11
norm_v	3	0.9	4/5	errors/err on gb; gbld/putouts; outsonbunts	0.232	2	7	7	16
norm_v	4	0.95	2/5	errors/err on gb; gbld/putouts; outsonbunts; --	0.247	4	10	8	22
norm_cv	3	mineigen	5/5	errors/err on gb; gbld/putouts; outsonbunts	0.209	1	7	6	14
norm_cv	4	0.9	5/5	errors/err on gb; gbld/putouts; outsonbunts; assists	0.189	1	6	5	12

D.6 Shortstop Rankings

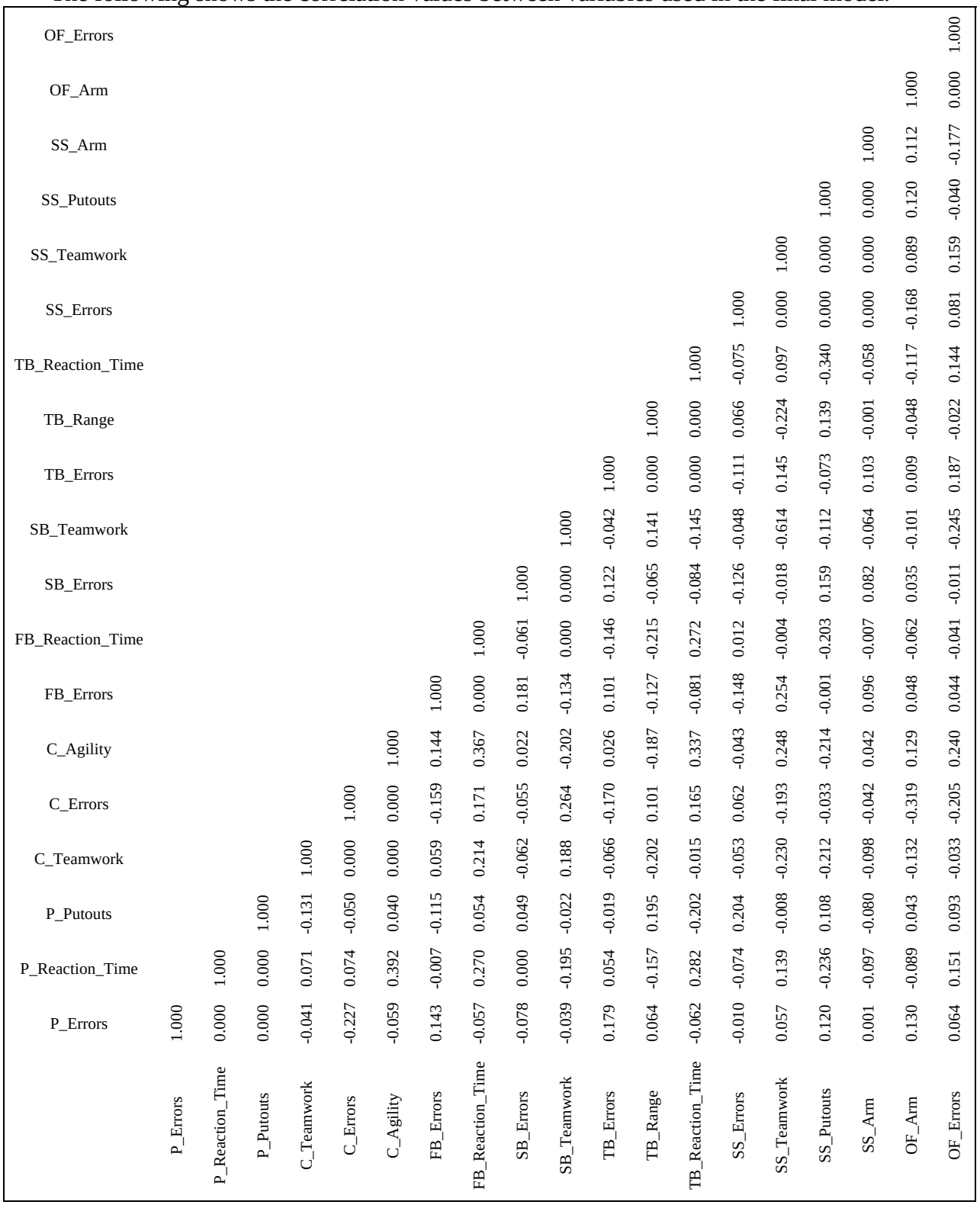
Method	# of Factors	Reason for # of Factors	Consistency Description	Interpretability Description	AIC	Consistency Rank	Interpretability Rank	Log Likelihood Rank	Final Ranking
cov	2	mineigen	3/5	assists; putouts	0.302	3	4	12	19
cov	3	0.95	2/5	assists/double plays; putouts; --	0.479	4	8	13	25
corr	3	mineigen	0/5	errors/err on gb; doubleplays/assists; relaythrows	0.269	5	1	10	16
log	2	Mineigen 90%	5/5	errors/err on gb; relaythrows	0.186	1	6	2	9
log	3	0.95	3/5	errors/err on gb; doubleplays; relaythrows	0.187	3	5	3	11
rank	3	mineigen	0/5	doubleplays/assists; errors/err on gb; relaythrows	0.189	5	1	4	10
rank	**4**	**0.9**	**3/5**	**errors/err on gb; assists/doubleplays; putouts; relaythrows**	**0.141**	**3**	**3**	**1**	**7**
rank	5	0.95	4/5	errors/err on gb; assists; putouts; relaythrows; doubleplays	0.227	2	7	6	15
norm_v	2	mineigen/90%	5/5	errors/err on gb; relaythrows	0.246	1	6	7	14
norm_v	3	0.95	2/5	errors/err on gb; doubleplays; relaythrows	0.249	4	5	9	18
norm_cv	2	mineigen	5/5	errors/err on gb; relaythrows	0.249	1	6	8	15
norm_cv	3	0.9	2/5	errors/err on gb; doubleplays; relaythrows	0.195	4	5	5	14
norm_cv	4	0.95	3/5	errors/err on gb; doubleplays/assists; putouts; relaythrows	0.281	3	3	11	17

D.7 Outfield Rankings

Method	# of Factors	Reason for # of Factors	Consistency Description	Interpretability Description	AIC	Consistency Rank	Interpretability Rank	Log Likelihood Rank	Final Ranking
cov	1	mineigen/95%	5/5	putouts	0.896	1	4	7	12
corr	2	mineigen	2/5	putouts; errors	0.743	4	1	5	10
log	**2**	**mineigen/95%**	**5/5**	**assists; errors**	**0.429**	**1**	**1**	**1**	**3**
rank	2	mineigen	2/5	putouts/assists; errors	0.860	4	3	6	13
norm_v	1	mineigen	4/5	errors	0.494	2	4	2	8
norm_v	2	0.95	4/5	assists; errors	0.655	2	1	3	6
norm_cv	2	mineigen	1/5	assists; errors	0.668	5	1	4	10

Appendix E: Pearson Correlation Chart of Factors Used in Final Model

The following shows the correlation values between variables used in the final model.



	P_Errors	P_Reaction_Time	P_Putouts	C_Teamwork	C_Errors	C_Agility	FB_Errors	FB_Reaction_Time	SB_Errors	SB_Teamwork	TB_Errors	TB_Range	TB_Reaction_Time	SS_Errors	SS_Teamwork	SS_Putouts	SS_Arm	OF_Arm	OF_Errors
P_Errors	1.000																		
P_Reaction_Time	0.000	1.000																	
P_Putouts	0.000	0.000	1.000																
C_Teamwork	-0.041	0.071	-0.131	1.000															
C_Errors	-0.227	0.074	-0.050	0.000	1.000														
C_Agility	-0.059	0.392	0.040	0.000	0.000	1.000													
FB_Errors	0.143	-0.007	-0.115	0.059	-0.159	0.144	1.000												
FB_Reaction_Time	-0.057	0.270	0.054	0.214	0.171	0.367	0.000	1.000											
SB_Errors	-0.078	0.000	0.049	-0.062	-0.055	0.022	0.181	-0.061	1.000										
SB_Teamwork	-0.039	-0.195	-0.022	0.188	0.264	-0.202	-0.134	0.000	0.000	1.000									
TB_Errors	0.179	0.054	-0.019	-0.066	-0.170	0.026	0.101	-0.146	0.122	-0.042	1.000								
TB_Range	0.064	-0.157	0.195	-0.202	0.101	-0.187	-0.127	-0.215	-0.065	0.141	0.000	1.000							
TB_Reaction_Time	-0.062	0.282	-0.202	-0.015	0.165	0.337	-0.081	0.272	-0.084	-0.145	0.000	0.000	1.000						
SS_Errors	-0.010	-0.074	0.204	-0.053	0.062	-0.043	-0.148	0.012	-0.126	-0.048	-0.111	0.066	-0.075	1.000					
SS_Teamwork	0.057	0.139	-0.008	-0.230	-0.193	0.248	0.254	-0.004	-0.018	-0.614	0.145	-0.224	0.097	0.000	1.000				
SS_Putouts	0.120	-0.236	0.108	-0.212	-0.033	-0.214	-0.001	-0.203	0.159	-0.112	-0.073	0.139	-0.340	0.000	0.000	1.000			
SS_Arm	0.001	-0.097	-0.080	-0.098	-0.042	0.042	0.096	-0.007	0.082	-0.064	0.103	-0.001	-0.058	0.000	0.000	0.000	1.000		
OF_Arm	0.130	-0.089	0.043	-0.132	-0.319	0.129	0.048	-0.062	0.035	-0.101	0.009	-0.048	-0.117	-0.168	0.089	0.120	0.112	1.000	
OF_Errors	0.064	0.151	0.093	-0.033	-0.205	0.240	0.044	-0.041	-0.011	-0.245	0.187	-0.022	0.144	0.081	0.159	-0.040	-0.177	0.000	1.000

Appendix F: SAS Program Files

Found in this appendix is a listing of all SAS programs used throughout the duration of the project. A description of each program and sample code are included whenever appropriate.

F.1 Factor Analysis using Covariance

F.1.1 factor_cov

This code took all the statistics per position and created a specific number of factors for each year using the covariance matrix.

F.1.2 Sample Code

```
proc factor data=sasuser.byteambypos_2004 method=principal scree cov
                rotate=varimax nfact=3 out=outp32004;
        var P_Assists  P_Errors  P_Putouts  P_Pickoffs  P_Balks  P_WPMenON
                P_GBLD;
run;
```

F.2 Factor Analysis using Correlation

F.2.1 factor_corr

This code took all the statistics per position and created a specific number of factors for each year using the correlation matrix.

F.2.2 Sample Code

```
proc factor data=sasuser.byteambypos_2000_2004 method=principal scree
                rotate=varimax nfact=3 out=outp32004;
        var P_Assists  P_Errors  P_Putouts  P_Pickoffs  P_Balks
                P_WPMenON  P_GBLD;
run;
```

F.3 Logarithmic Transformations

F.3.1 logtranformP

This code transformed the raw data for pitchers to logs of the original numbers and then ran a factor analysis on the transformed data.

F.3.2 logtranformC

This code transformed the raw data for catchers to logs of the original numbers and then ran a factor analysis on the transformed data.

F.3.3 logtranformFB

This code transformed the raw data for first basemen to logs of the original numbers and then ran a factor analysis on the transformed data.

F.3.4 logtranformSB

This code transformed the raw data for second basemen to logs of the original numbers and then ran a factor analysis on the transformed data.

F.3.5 logtranformTB

This code transformed the raw data for third basemen to logs of the original numbers and then ran a factor analysis on the transformed data.

F.3.6 logtranformSS

This code will transformed the raw data for shortstops to logs of the original numbers and then ran a factor analysis on the transformed data.

F.3.7 logtranformOF

This code will transformed the raw data for outfielders to logs of the original numbers and then ran a factor analysis on the transformed data.

F.3.8 Sample Code

```
data log2000_2004OF;
        set sasuser.byteambypos_2000_2004;
        logOF_Assists=log(OF_Assists + 1);
        logOF_Errors=log(OF_Errors + 1);
        logOF_Putouts=log(OF_Putouts + 1);
run;

proc factor data=log2000_2004OF method=principal rotate=varimax cov scree
                nfact=2 out=outOF2;
        var logOF_Assists logOF_Errors logOF_Putouts;
run;
```

F.4 Rank Transformation

F.4.1 ranktransformP

This program transformed raw data for pitchers to ranks using the rank procedure. Then a factor analysis was run on the transformed data.

F.4.2 ranktransformC

This program transformed raw data for catchers to ranks using the rank procedure. Then a factor analysis was run on the transformed data.

F.4.3 ranktransformFB

This program transformed raw data for first basemen to ranks using the rank procedure. Then a factor analysis was run on the transformed data.

F.4.4 ranktransformSB

This program transformed raw data for second basemen to ranks using the rank procedure. Then a factor analysis was run on the transformed data.

F.4.5 ranktransformTB

This program transformed raw data for third basemen to ranks using the rank procedure. Then a factor analysis was run on the transformed data.

F.4.6 ranktransformSS

This program transformed raw data for shortstops to ranks using the rank procedure. Then a factor analysis was run on the transformed data.

F.4.7 ranktransformOF

This program transformed raw data for outfielders to ranks using the rank procedure. Then a factor analysis was run on the transformed data.

F.4.8 Sample Code

```
data rank2000_2004SS;
        set sasuser.byteambypos_2000_2004;
run;
proc rank data=rank2000_2004SS out=temp;
        var SS_AssistsSS_Errors  SS_Putouts  SS_ErrorsOnGroundBalls
                SS_DoublePlays  SS_RelayThrows;
run;
proc factor data=temp method=principal rotate=varimax   cov scree nfact=5
                out=outss5;
        var SS_AssistsSS_Errors  SS_Putouts  SS_ErrorsOnGroundBalls
                SS_DoublePlays  SS_RelayThrows;
run;
```

F.5 Coeffiecient of Variation Transformations

F.5.1 normtransformP

This code took the raw data for pitchers and applied two transformations on it: first by adjusting for the coefficient of variation and second by the coefficient of variation squared. Factor analysis was then used on the transformed data.

F.5.2 normtransformC

This code took the raw data for catchers and applied two transformations on it: first by adjusting for the coefficient of variation and second by the coefficient of variation squared. Factor analysis was then used on the transformed data.

F.5.3 normtransformFB

This code took the raw data for first basemen and applied two transformations on it: first by adjusting for the coefficient of variation and second by the coefficient of variation squared. Factor analysis was then used on the transformed data.

F.5.4 normtransformSB

This code took the raw data for second basemen and applied two transformations on it: first by adjusting for the coefficient of variation and second by the coefficient of variation squared. Factor analysis was then used on the transformed data.

F.5.5 normtransformTB

This code took the raw data for third basemen and applied two transformations on it: first by adjusting for the coefficient of variation and second by the coefficient of variation squared. Factor analysis was then used on the transformed data.

F.5.6 normtransformSS

This code took the raw data for shortstops and applied two transformations on it: first by adjusting for the coefficient of variation and second by the coefficient of variation squared. Factor analysis was then used on the transformed data.

F.5.7 normtransformOF

This code took the raw data for outfielders and applied two transformations on it: first by adjusting for the coefficient of variation and second by the coefficient of variation squared. Factor analysis was then used on the transformed data.

F.5.8 Sample Code

```
data allyears;
        set sasuser.byteambypos_2000_2004;
run;
proc means data=allyears noprint;
        var FB_Assists  FB_Errors  FB_Putouts  FB_ErrorsOnGroundBalls
                FB_OutsonBunts FB_OutsInFoul;
        output out=temp mean=m1-m6 stddev=s1-s6;
run;
proc iml;
        use temp;
        read all var {m1 m2 m3 m4 m5 m6} into means;
        use allyears;
        read all var {FB_Assists  FB_Errors  FB_Putouts
                FB_ErrorsOnGroundBalls  FB_OutsonBunts  FB_OutsInFoul}
                into datamatrix;
        newdata=datamatrix*diag(100/means);
        varnames={'N2FB_Assists' 'N2FB_Errors' 'N2FB_PO' 'N2FB_EGB'
                'N2FB_OutsonBunts' 'N2FB_OIF'};
        create new from newdata [Colname = varnames];
        append from newdata;
run;
proc factor cov data=new method=principal rotate=varimax scree nfact=2
                out=outN2_fb2;
        var N2FB_Assists  N2FB_Errors  N2FB_PO  N2FB_EGB
                N2FB_OutsonBunts  N2FB_OIF;
run;
proc iml;
        use temp;
        read all var {m1 m2 m3 m4 m5 m6} into means;
        read all var {s1 s2 s3 s4 s5 s6} into sds;
        use allyears;
        read all var {FB_Assists  FB_Errors  FB_Putouts
                FB_ErrorsOnGroundBalls  FB_OutsonBunts  FB_OutsInFoul}
                into datamatrix;
        newdata=datamatrix*diag(10/sqrt(means))*diag(1/sqrt(sds));
        varnames={'NFB_Assists' 'NFB_Errors' 'NFB_PO' 'NFB_EGB'
                'NFB_OutsonBunts' 'NFB_OIF'};
        create new2 from newdata [Colname = varnames];
        append from newdata;
run;
proc factor cov data=new2 method=principal rotate=varimax scree nfact=2
        out=outN_fb2;
        var NFB_Assists  NFB_Errors  NFB_PO  NFB_EGB  NFB_OutsonBunts
                NFB_OIF;
run;
```

F.6 Cluster Analysis

F.6.1 clusterP

This code produced 3 factors using the correlation matrix for pitchers. Then a cluster analysis was run in addition to creating several plots (both 2-D and 3-D) including gold glove, year, and player indicators.

F.6.2 clusterC

This code produced 3 factors for catchers utilizing the rank transformation. Then a cluster analysis was run in addition to creating several plots (both 2-D and 3-D) including gold glove, year, and player indicators.

F.6.3 clusterFB

This code produced 2 factors for first basemen utilizing normalization by the coefficient of variation. Then a cluster analysis was run in addition to creating several 2-D plots including gold glove, year, and player indicators.

F.6.4 clusterSB

This code produced 2 factors for second basemen utilizing the rank transformation. Then a cluster analysis was run in addition to creating several 2-D plots including gold glove, year, and player indicators.

F.6.5 clusterTB

This code produced 3 factors using for second basemen the correlation matrix. Then a cluster analysis was run in addition to creating several plots (both 2-D and 3-D) including gold glove, year, and player indicators.

F.6.6 clusterSS

This code produced 4 factors for shortstops utilizing the rank transformation. Then a cluster analysis was run in addition to creating several plots (both 2-D and 3-D) including gold glove, year, and player indicators.

F.6.7 clusterOF

This code produced 2 factors for outfielders utilizing the log transformation. Then a cluster analysis was run in addition to creating several 2-D plots including gold glove, year, and player indicators.

F.6.8 Sample Code

```
data outfield;
        set outfield;
        length colorval $40.;
        if gold_glove='no' then
                do;
                colorval='red';
        end;
        if gold_glove='yes' then
                do;
                colorval='blue';
        end;
        label col1='OF_Arm';
        label col2='OF_Errors';
run;
proc gplot data=outfield;
    symbol1
        value='circle'
        cv='black';
    symbol2
        value='star'
        cv='black';
    plot col1*col2=gold_glove;
    plot2 col1*col2=yearid;
run;
proc gplot data=outfield;
        plot col1*col2=id;
run;
proc gplot data=outfield;
        plot col1*col2=gold_glove;
run;
```

F.7 Other Program Files

F.7.1 corr_plots

This program will create the factors produced for position. Then it runs a logistic model on all the factors produced. A chart is then produced with all the Pearson Correlations along with the p-values for each correlation. Then a box plot of the residuals is created. Finally a three dimensional plot of runs vs. earned runs vs. winning percentage is created for each year comparing all the teams for that season.

F.7.2 final_model

This code will create the factors for each position using varied methods. Then a logistic model will fit the factors all together, then by position, then individually.

Appendix G: SAS Data Files

Listed below are the names and descriptions of all SAS data files used throughout the duration of the project.

G.1 Team Data

G.1.1 byteambypos_2000
: This data set contains the sum of all defensive statistics for each team, grouped by position for the year 2000.

G.1.2 byteambypos_2001
: This data set contains the sum of all defensive statistics for each team, grouped by position for the year 2001.

G.1.3 byteambypos_2002
: This data set contains the sum of all defensive statistics for each team, grouped by position for the year 2002.

G.1.4 byteambypos_2003
: This data set contains the sum of all defensive statistics for each team, grouped by position for the year 2003.

G.1.5 byteambypos_2004
: This data set contains the sum of all defensive statistics for each team, grouped by position for the year 2004.

G.1.6 byteambypos_2000_2004
: This data set contains the sum of all defensive statistics for each team, grouped by position for the years 2000 - 2004.

G.1.7 Team_data
: This file contains the offensive and defensive statistics as well as the response variable used in the final model for each team from 2000 to 2004.

G.2 Player Data

G.2.1 pitchers
: This data set holds all statistics of gold glove pitchers as well as other pitchers randomly selected for the validation process.

G.2.2 catchers
: This data set holds all statistics of gold glove catchers as well as other catchers randomly selected for the validation process.

G.2.3 firstbase
: This data set holds all statistics of gold glove first basemen as well as other first basemen randomly selected for the validation process.

G.2.4 secondbase
: This data set holds all statistics of gold glove second basemen as well as other second basemen randomly selected for the validation process.

G.2.5 thirdbase

This data set holds all statistics of gold glove third basemen as well as other third basemen randomly selected for the validation process.

G.2.6 shortstop

This data set holds all statistics of gold glove shortstops as well as other shortstops randomly selected for the validation process.

G.2.7 outfielders

This data set holds all statistics of gold glove outfielders as well as other outfielders randomly selected for the validation process.

Appendix H: SAS Variables

Listed below are the names and descriptions of all variables used in the SAS data files.

H.1 Variables

P_Assists
: Number of pitcher assists

P_Errors
: Number of pitcher errors

P_Putouts
: Number of pitcher putouts

P_Pickoffs
: Number of pitcher pickoffs

P_Balks
: Number of pitcher balks

P_WPMenON
: Number of pitcher wild pitches made with men on base

P_GBLD
: Number of ground balls and line drives fielded by pitcher

C_Assists
: Number of catcher assists

C_CaughtStealing
: Number of runners caught stealing by catcher

C_Errors
: Number of catcher errors

C_PassedBalls
: Number of catcher passed balls

C_Putouts
: Number of catcher putouts

C_OutsonBunts
: Number of bunt attempts in which an out was made by catcher

C_TagOutAtHome
: Number of catcher tag outs at home plate

FB_Assists
: Number of first base assists

FB_Errors
: Number of first base errors

FB_Putouts
: Number of first base putouts

FB_ErrorsOnGroundBalls
: Number of first base errors made on ground balls

FB_OutsonBunts
: Number of bunt attempts in which an out was made by first base

FB_OutsInFoul
: Number of outs made in foul ground by first base

SB_Assists
: Number of second base assists

SB_Errors
: Number of second base errors

SB_Putouts
: Number of second base putouts

SB_ErrorsOnGroundBalls
: Number of second base errors made on ground balls

SB_DoublePlays
: Number of second base double plays

SB_RelayThrows
: Number of second base relay throws

TB_Assists
: Number of third base assists

TB_Errors
: Number of third base errors

TB_Putouts
: Number of third base putouts

TB_ErrorsOnGroundBalls
: Number of third base errors made on ground balls

TB_OutsInFoul
: Number of outs made in foul ground by third base

TB_OutsOnBunts
: Number of bunt attempts in which an out was made by third base

SS_Assists
: Number of shortstop assists

SS_Errors
: Number of shortstop errors

SS_Putouts
: Number of shortstop putouts

SS_ErrorsOnGroundBalls
: Number of shortstop errors made on ground balls

SS_DoublePlays
: Number of shortstop double plays

SS_RelayThrows
: Number of shortstop relay throws

OF_Assists
: Number of outfield assists

OF_Errors
: Number of outfield errors

OF_Putouts
: Number of outfield putouts

logP_Assists
: Logarithm of pitcher assists

logP_Errors
: Logarithm of pitcher errors

logP_Putouts
: Logarithm of pitcher putouts

logP_Pickoffs
: Logarithm of pitcher pickoffs

logP_Balks
: Logarithm of pitcher balks

logP_WPMenOn
: Logarithm of pitcher wild pitches made with men on base

logP_GBLD
: Logarithm of ground balls and line drives fielded by pitcher

logC_Assists
: Logarithm of catcher assists

logC_Errors
: Logarithm of catcher errors

logC_Putouts
: Logarithm of catcher putouts

logC_PassedBalls
: Logarithm of catcher passed balls

logC_TagOutAtHome
: Logarithm of catcher tag outs at home plate

logC_OutsOnBunts
: Logarithm of bunt attempts in which an out was made by catcher

logC_CaughtStealing
: Logarithm of runners caught stealing by catcher

logFB_Assists
: Logarithm of first base assists

logFB_Errors
: Logarithm of first base errors

logFB_Putouts
: Logarithm of first base putouts

logFB_ErrorsOnGroundBalls
: Logarithm of first base errors made on ground balls

logFB_OutsInFoul
: Logarithm of outs made in foul ground by first base

logFB_OutsOnBunts
: Logarithm of bunt attempts in which an out was made by first base

logSB_Assists
: Logarithm of second base assists

logSB_Errors
: Logarithm of second base errors

logSB_Putouts
: Logarithm of second base putouts

logSB_ErrorsOnGroundBalls
: Logarithm of second base errors made on ground balls

logSB_DoublePlays
: Logarithm of second base double plays

logSB_RelayThrows
: Logarithm of second base relay throws

logTB_Assists
: Logarithm of third base assists

logTB_Errors
: Logarithm of third base errors

logTB_Putouts
: Logarithm of third base putouts

logTB_ErrorsOnGroundBalls
: Logarithm of third base errors made on ground balls

logTB_OutsInFoul
: Logarithm of outs made in foul ground by third base

logTB_OutsOnBunts
: Logarithm of bunt attempts in which an out was made by third base

logSS_Assists
: Logarithm of shortstop assists

logSS_Errors
: Logarithm of shortstop errors

logSS_Putouts
: Logarithm of shortstop putouts

logSS_ErrorsOnGroundBalls
: Logarithm of shortstop errors made on ground balls

logSS_DoublePlays
: Logarithm of shortstop double plays

logSS_RelayThrows
: Logarithm of shortstop relay throws

logOF_Assists
: Logarithm of outfield assists

logOF_Errors
: Logarithm of outfield errors

logOF_Putouts
: Logarithm of outfield putouts

NP_Assists
: Coefficient of variation for pitcher assists

NP_Errors
: Coefficient of variation for pitcher errors

NP_PO
: Coefficient of variation for pitcher putouts

NP_Pickoffs
: Coefficient of variation for pitcher pickoffs

NP_Balks
: Coefficient of variation for pitcher balks

NP_WPMO
: Coefficient of variation for pitcher wild pitches made with men on base

NP_GBLD
: Coefficient of variation for ground balls and line drives fielded by pitcher

NC_Assists
: Coefficient of variation for catcher assists

NC_Errors
: Coefficient of variation for catcher errors

NC_PO
: Coefficient of variation for catcher putouts

NC_PassedBalls
: Coefficient of variation for catcher passed balls

NC_OutsonBunts
: Coefficient of variation for bunt attempts in which an out was made by catcher

NC_TagOutAtHome
: Coefficient of variation for catcher tag outs at home plate

NC_CaughtStealing
: Coefficient of variation for runners caught stealing by catcher

NFB_Assists
: Coefficient of variation for first base assists

NFB_Errors
: Coefficient of variation for first base errors

NFB_PO
: Coefficient of variation for first base putouts

NFB_EGB
: Coefficient of variation for first base errors made on ground balls

NFB_OutsonBunts
: Coefficient of variation for bunt attempts in which an out was made by first base

NFB_OIF
: Coefficient of variation for outs made in foul ground by first base

NSB_Assists
: Coefficient of variation for second base assists

NSB_Errors
: Coefficient of variation for second base errors

NSB_PO
: Coefficient of variation for second base putouts

NSB_EGB
: Coefficient of variation for second base errors made on ground balls

NSB_DB
: Coefficient of variation for second base double plays

NSB_RelayThrows
: Coefficient of variation for second base relay throws

NTB_Assists
: Coefficient of variation for third base assists

NTB_Errors
: Coefficient of variation for third base errors

NTB_PO
: Coefficient of variation for third base putouts

NTB_EGB
: Coefficient of variation for third base errors made on ground balls

NTB_OutsonBunts
: Coefficient of variation for bunt attempts in which an out was made by third base

NTB_OIF
: Coefficient of variation for outs made in foul ground by third base

NSS_Assists
: Coefficient of variation for shortstops assists

NSS_Errors
: Coefficient of variation for shortstops errors

NSS_PO
: Coefficient of variation for shortstops putouts

NSS_EGB
: Coefficient of variation for shortstops errors made on ground balls

NSS_DB
: Coefficient of variation for shortstops double plays

NSS_RelayThrows
: Coefficient of variation for shortstops relay throws

NOF_Assists
: Coefficient of variation for outfield assists

NOF_Errors
: Coefficient of variation for outfield errors

NOF_PO
: Coefficient of variation for outfield putouts

N2P_Assists
: Coefficient of variation squared for pitchers assists

N2P_Errors
: Coefficient of variation squared for pitchers errors

N2P_PO
: Coefficient of variation squared for pitchers putouts

N2P_Pickoffs
: Coefficient of variation squared for pitchers pickoffs

N2P_Balks
: Coefficient of variation squared for balks

N2P_WPMO
: Coefficient of variation squared for pitcher wild pitches made with men on base

N2P_GBLD
: Coefficient of variation squared for ground balls and line drives fielded by pitcher

N2C_Assists
: Coefficient of variation squared for catcher assists

N2C_Errors
: Coefficient of variation squared for catcher errors

N2C_PO
: Coefficient of variation squared for catcher putouts

N2C_PassedBalls
: Coefficient of variation squared for catcher passed balls

N2C_OutsonBunts
: Coefficient of variation squared for bunt attempts in which an out was made by catcher

N2C_TagOutAtHome
: Coefficient of variation squared for catcher tag outs at home plate

N2C_CaughtStealing
: Coefficient of variation squared for runners caught stealing by catcher

N2FB_Assists
: Coefficient of variation squared for first base assists

N2FB_Errors
: Coefficient of variation squared for first base errors

N2FB_PO
: Coefficient of variation squared for first base putouts

N2FB_EGB
: Coefficient of variation squared for first base errors made on ground balls

N2FB_OutsonBunts
: Coefficient of variation squared for bunt attempts in which an out was made by first base

N2FB_OIF
: Coefficient of variation squared for outs made in foul ground by first base

N2SB_Assists
: Coefficient of variation squared for second base assists

N2SB_Errors
: Coefficient of variation squared for second base errors

N2SB_PO
: Coefficient of variation squared for second base putouts

N2SB_EGB
: Coefficient of variation squared for second base errors made on ground balls

N2SB_DB
: Coefficient of variation squared for second base double plays

N2SB_RelayThrows
: Coefficient of variation squared for second base relay throws

N2TB_Assists
: Coefficient of variation squared for third base assists

N2TB_Errors
: Coefficient of variation squared for third base errors

N2TB_PO
: Coefficient of variation squared for third base putouts

N2TB_EGB
: Coefficient of variation squared for third base errors made on ground balls

N2TB_OutsonBunts
: Coefficient of variation squared for bunt attempts in which an out was made by third base

N2TB_OIF
: Coefficient of variation squared for outs made in foul ground by third base

N2SS_Assists
: Coefficient of variation squared for shortstops assists

N2SS_Errors
: Coefficient of variation squared for shortstops errors

N2SS_PO
: Coefficient of variation squared for shortstops putouts

N2SS_EGB
: Coefficient of variation squared for shortstops errors made on ground balls

N2SS_DB
: Coefficient of variation squared for shortstops double plays

N2SS_RelayThrows
: Coefficient of variation squared for shortstops relay throws

N2OF_Assists
: Coefficient of variation squared for outfield assists

N2OF_Errors
: Coefficient of variation squared for outfield errors

N2OF_PO
: Coefficient of variation squared for first outfield putouts

norm_P_Assists
: Normalized number of pitchers assists

norm_P_Errors
: Normalized number of pitchers errors

norm_P_Putouts
: Normalized number of pitchers putouts

norm_P_Pickoffs
: Normalized number of pitchers pickoffs

norm_P_Balks
: Normalized number of pitchers balks

norm_P_WPMenON
: Normalized number of pitcher wild pitches made with men on base

norm_P_GBLD
: Normalized number of ground balls and line drives fielded by pitcher

norm_C_Assists
: Normalized number of catching assists

norm_C_Errors
: Normalized number of catching errors

norm_C_Putouts
: Normalized number of catching putouts

norm_C_PassedBalls
: Normalized number of catching passed balls

norm_C_OutsOnBunts
: Normalized number of bunt attempts in which an out was made by catcher

norm_C_TagOutAtHome
: Normalized number of catching tag outs at home plate

norm_C_CaughStealing
: Normalized number of runners caught stealing by catcher

norm_FB_Assists
: Normalized number of first base assists

norm_FB_Errors
: Normalized number of first base errors

norm_FB_Putouts
: Normalized number of first base putouts

norm_FB_ErrorsOnGroundBalls
: Normalized number of first base errors made on ground balls

norm_FB_OutsInFoul
: Normalized number of outs made in foul ground by first base

norm_FB_OutsOnBunts
: Normalized number of bunt attempts in which an out was made by first base

norm_SB_Assists
: Normalized number of second base assists

norm_SB_Errors
: Normalized number of second base errors

norm_SB_Putouts
: Normalized number of second base putouts

norm_SB_ErrorsOnGroundBalls
: Normalized number of second base errors made on ground balls

norm_SB_RelayThrows
: Normalized number of second base relay throws

norm_SB_DoublePlays
: Normalized number of second base double plays

norm_TB_Assists
: Normalized number of third base assists

norm_TB_Errors
: Normalized number of third base errors

norm_TB_Putouts
: Normalized number of third base putouts

norm_TB_ErrorsOnGroundBalls
: Normalized number of third base errors made on ground balls

norm_TB_OutsInFoul
: Normalized number of outs made in foul ground by third base

norm_TB_OutsOnBunts
: Normalized number of bunt attempts in which an out was made by third base

norm_SS_Assists
: Normalized number of shortstops assists

norm_SS_Errors
: Normalized number of shortstops errors

norm_SS_Putouts
: Normalized number of shortstops putouts

norm_SS_ErrorsOnGroundBalls
: Normalized number of shortstops errors made on ground balls

norm_SS_RelayThrows
: Normalized number of shortstops relay throws

norm_SS_DoublePlays
: Normalized number of shortstops double plays

norm_OF_Assists
: Normalized number of outfield assists

norm_OF_Errors
: Normalized number of outfield errors

norm_OF_Putouts
: Normalized number of outfield putouts

H.2 Factor Names

P_Errors
: Final model factor describing pitcher errors

P_Reaction_Time
: Final model factor describing pitcher reaction time

P_Putouts
: Final model factor describing pitcher putouts

C_Agility
: Final model factor describing catcher agility

C_Errors
: Final model factor describing catcher errors

C_Teamwork
: Final model factor describing catcher teamwork

FB_Errors
: Final model factor describing first base errors

FB_Reaction_Time
: Final model factor describing first base reaction time

SB_Errors
: Final model factor describing second base errors

SB_Teamwork
: Final model factor describing second base teamwork

TB_Errors
: Final model factor describing third base errors

TB_Range
: Final model factor describing third base range

TB_Reaction_Time
: Final model factor describing third base reaction time

SS_Errors
: Final model factor describing shortstop errors

SS_Teamwork
: Final model factor describing shortstop teamwork

SS_Putouts
: Final model factor describing shortstop putouts

SS_Arm
: Final model factor describing shortstop arm

OF_Arm
: Final model factor describing outfield arm

OF_Errors
: Final model factor describing outfield errors

H.3 Other Variables

Color
: Color used to distinguish Gold Glovers and non-Gold Glovers

divisionid
: Division that a team belongs to

games
: Number of games played by each team in a year

gold_glove
: Gold Glove tag

id
: Player identification

Shape
: Shape used to represent variables in plots

Sumofw
: Sum of wins for a team

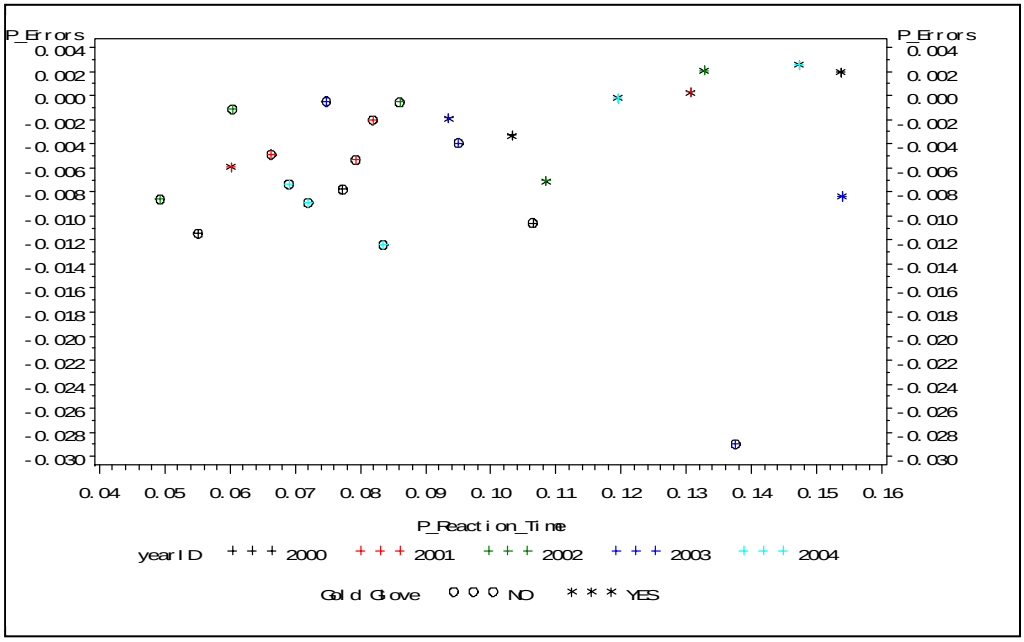
sumofer
: Sum of earned runs for a team

sumofr
: Sum of runs for a team

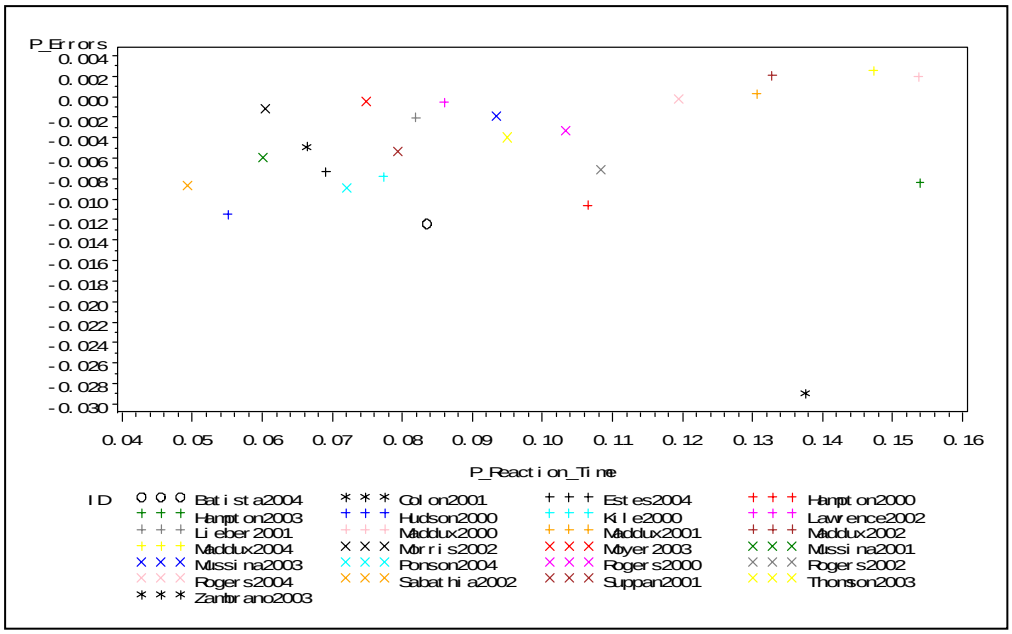
Appendix I: SAS Output

This appendix contains all graphical results of the model validation analysis conducted by the project team using each position's optimal factor method.

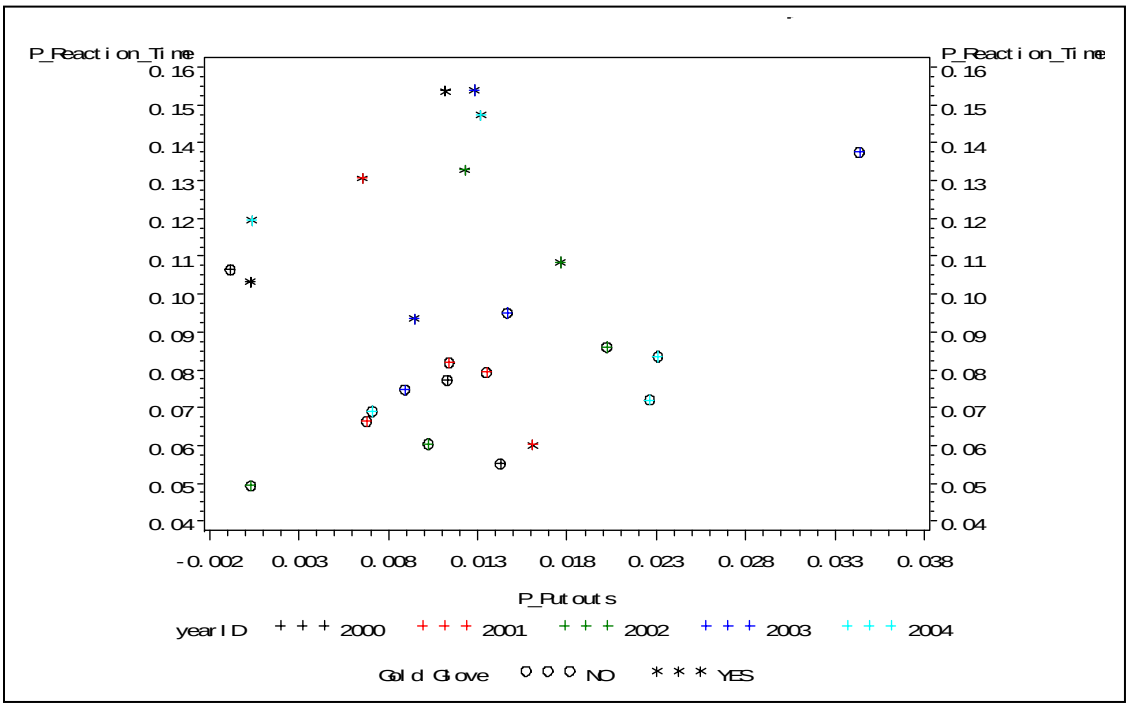
I.1 Pitcher: Errors vs. Reaction Time by Year and Gold Glove



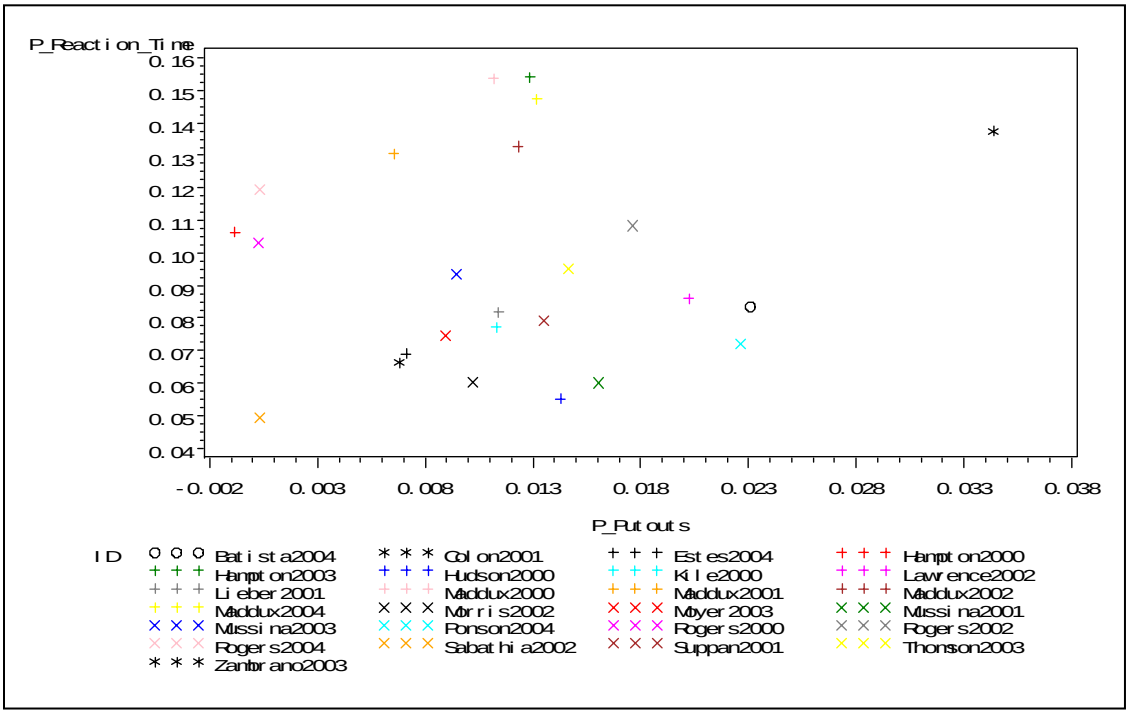
I.2 Pitcher: Errors vs. Reaction Time by Player



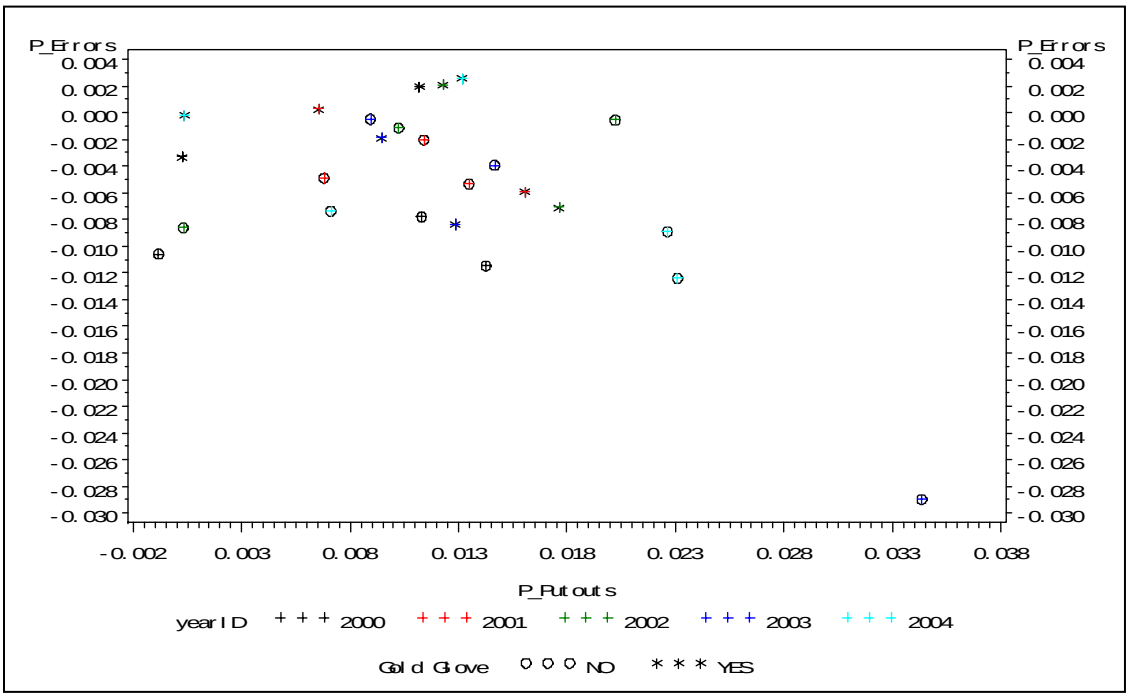
I.3 Pitcher: Reaction Time vs. Putouts by Year and Gold Glove



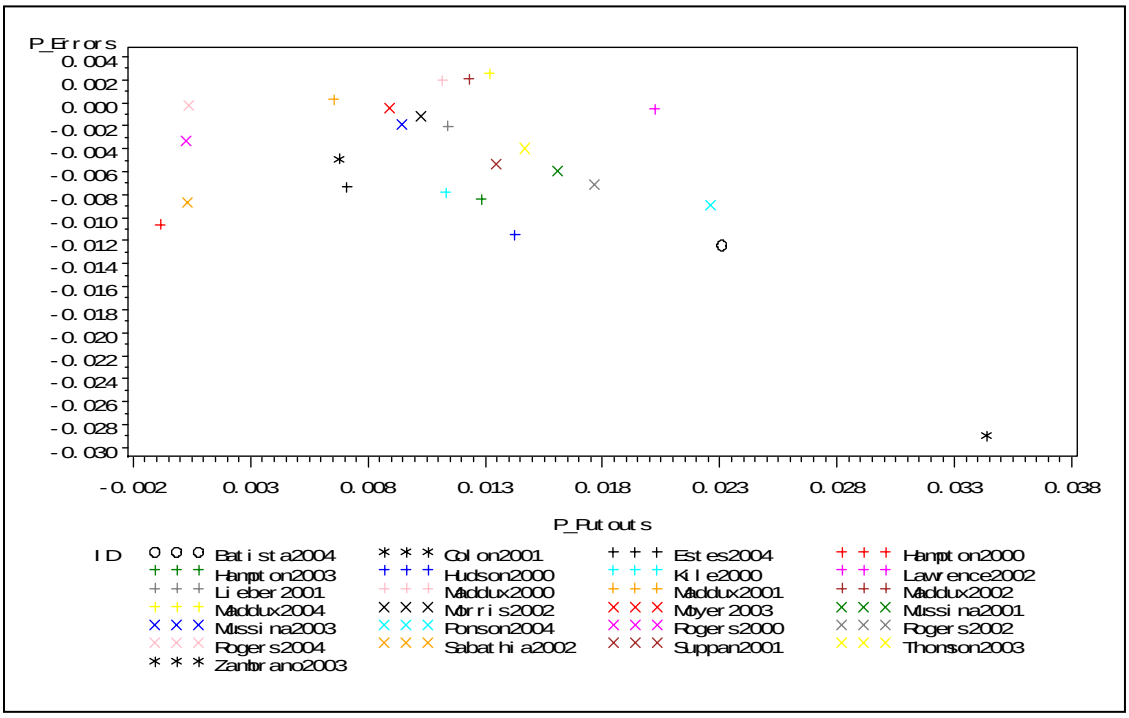
I.4 Pitcher: Reaction Time vs. Putouts by Player



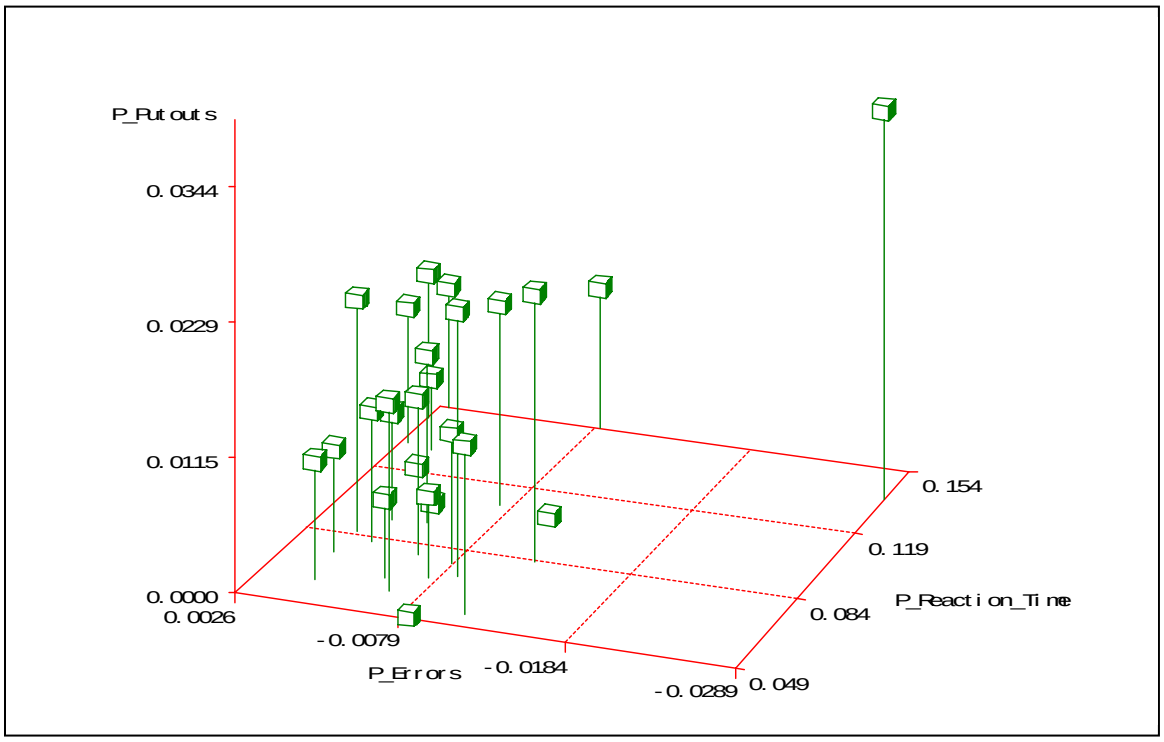
I.5 Pitcher: Errors vs. Putouts by Year and Gold Glove



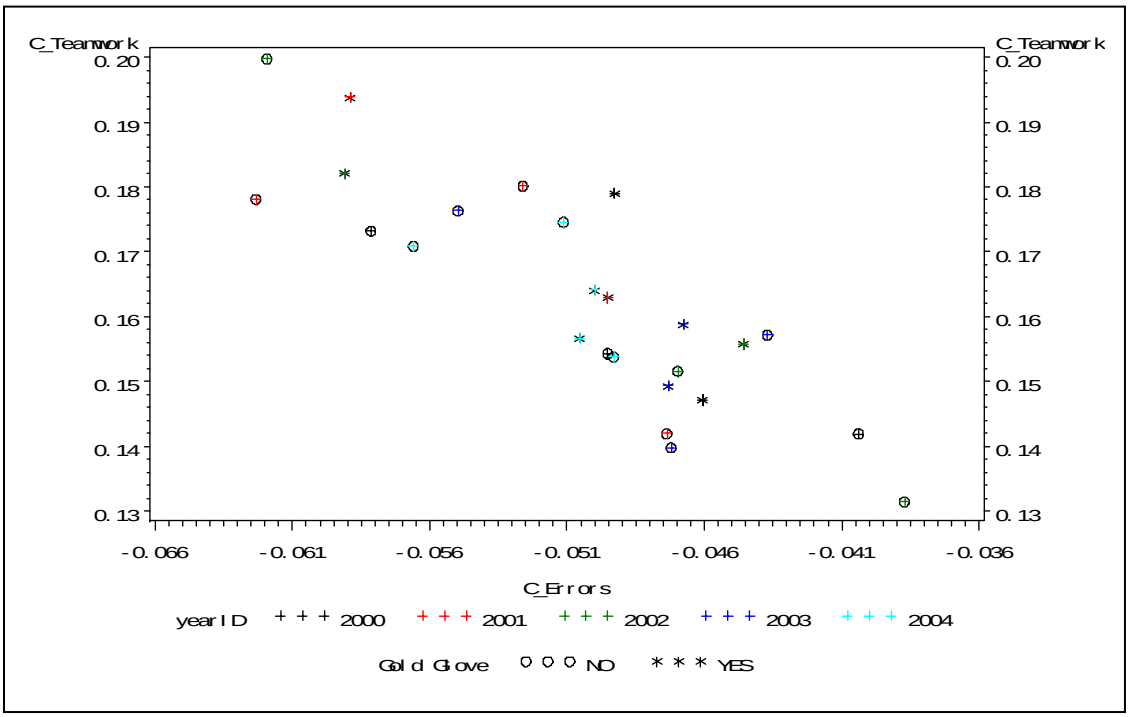
I.6 Pitcher: Errors vs. Putouts by Player



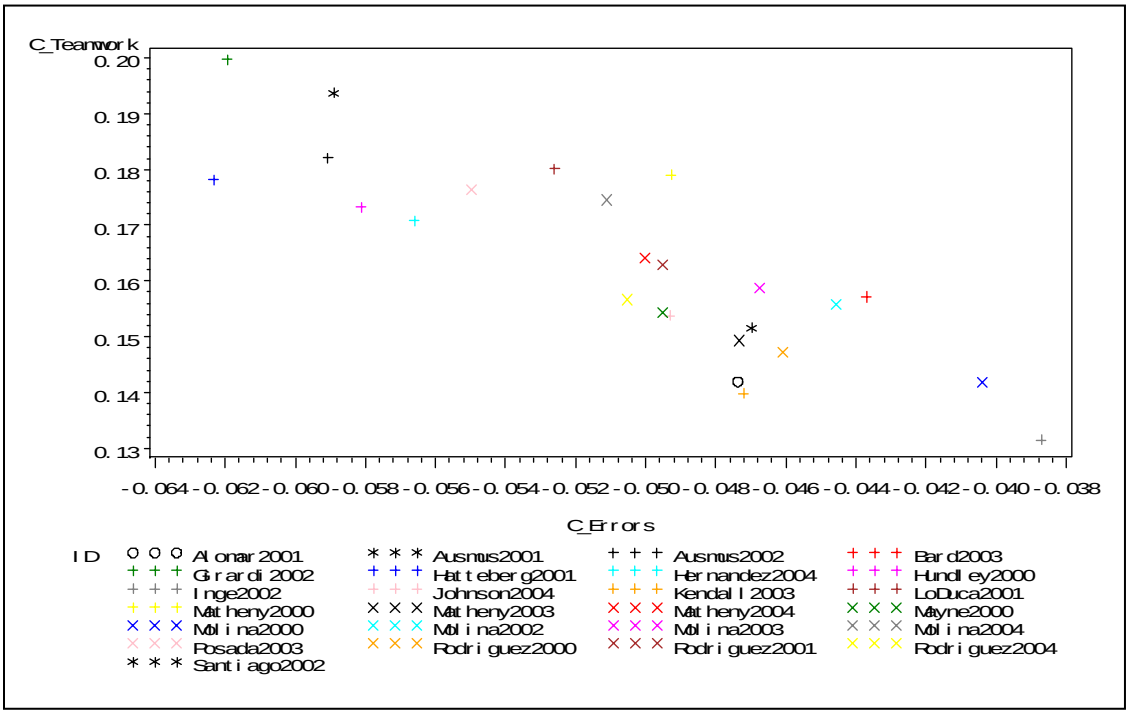
I.7 Pitcher: Putouts vs. Errors vs. Reaction Time



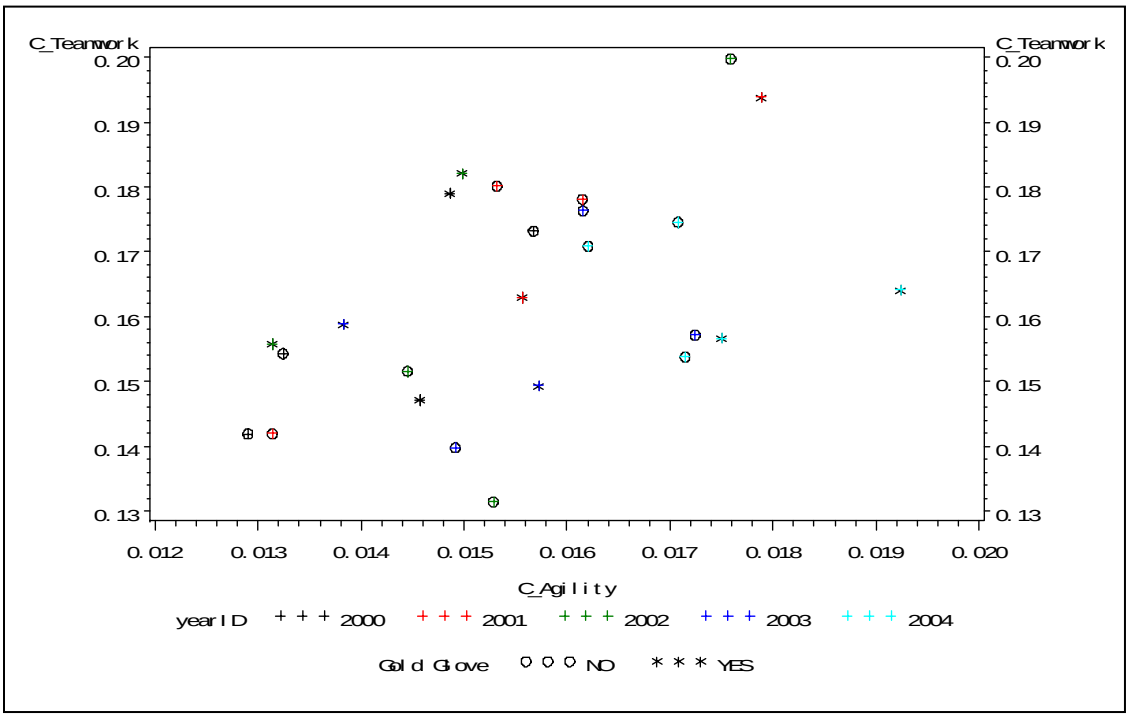
I.8 Catcher: Teamwork vs. Errors by Year and Gold Glove



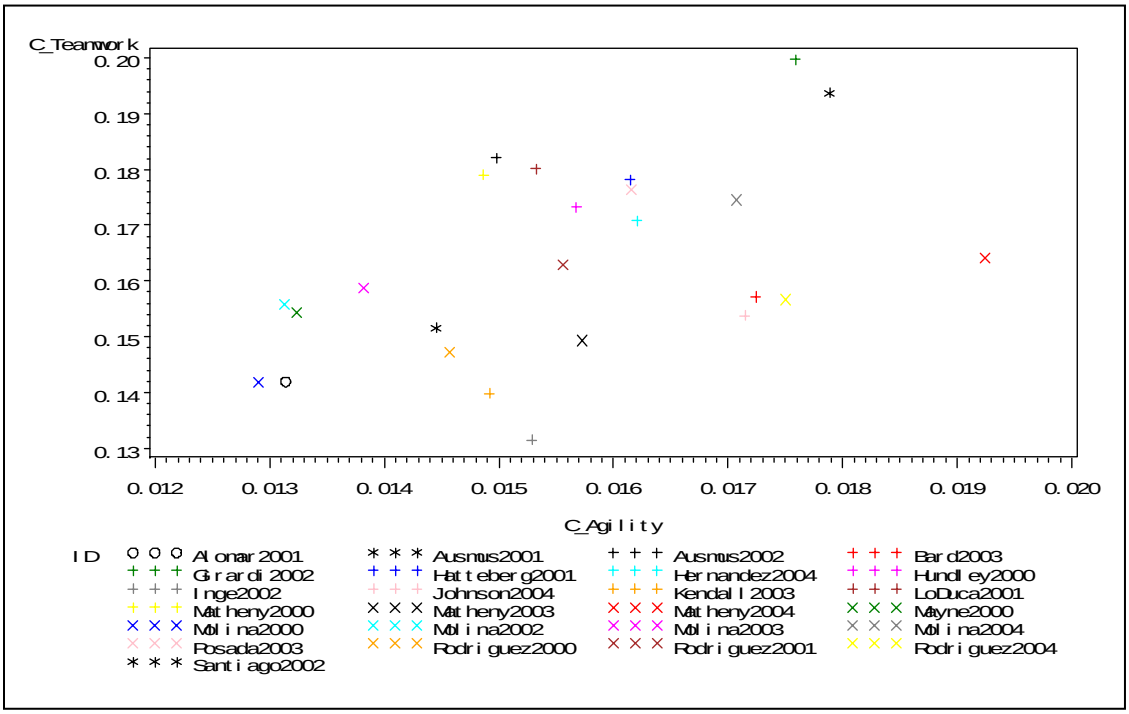
I.9 Catcher: Teamwork vs. Errors by Player



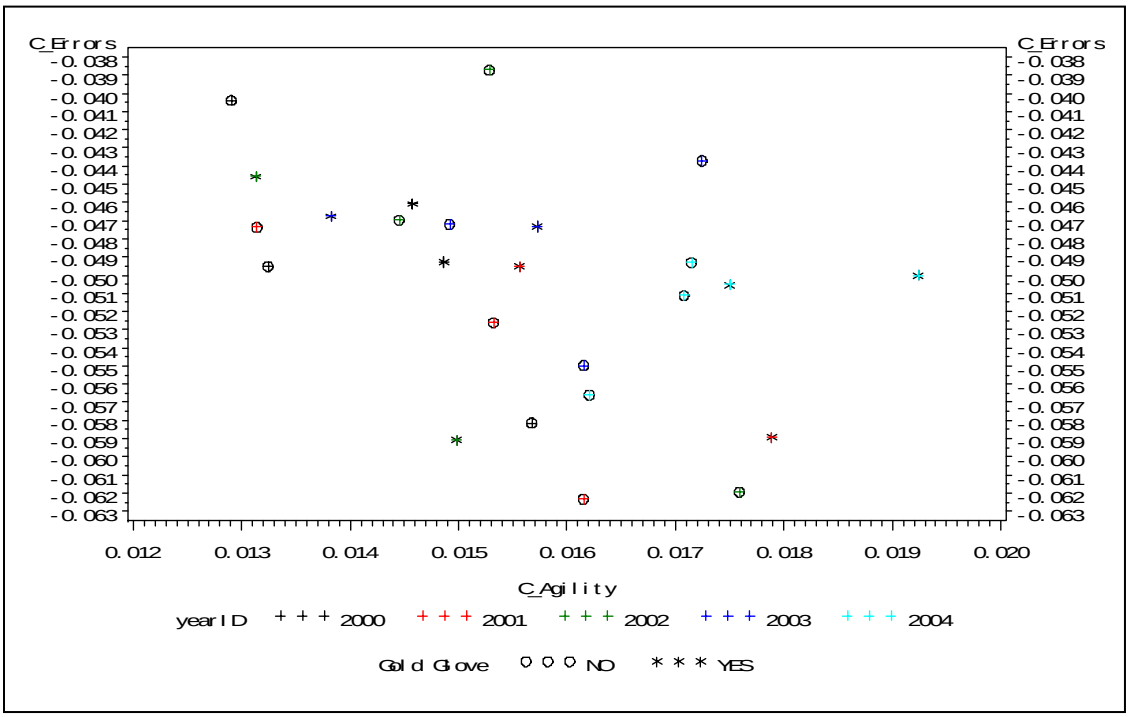
I.10 Catcher: Teamwork vs. Agility by Year and Gold Glove



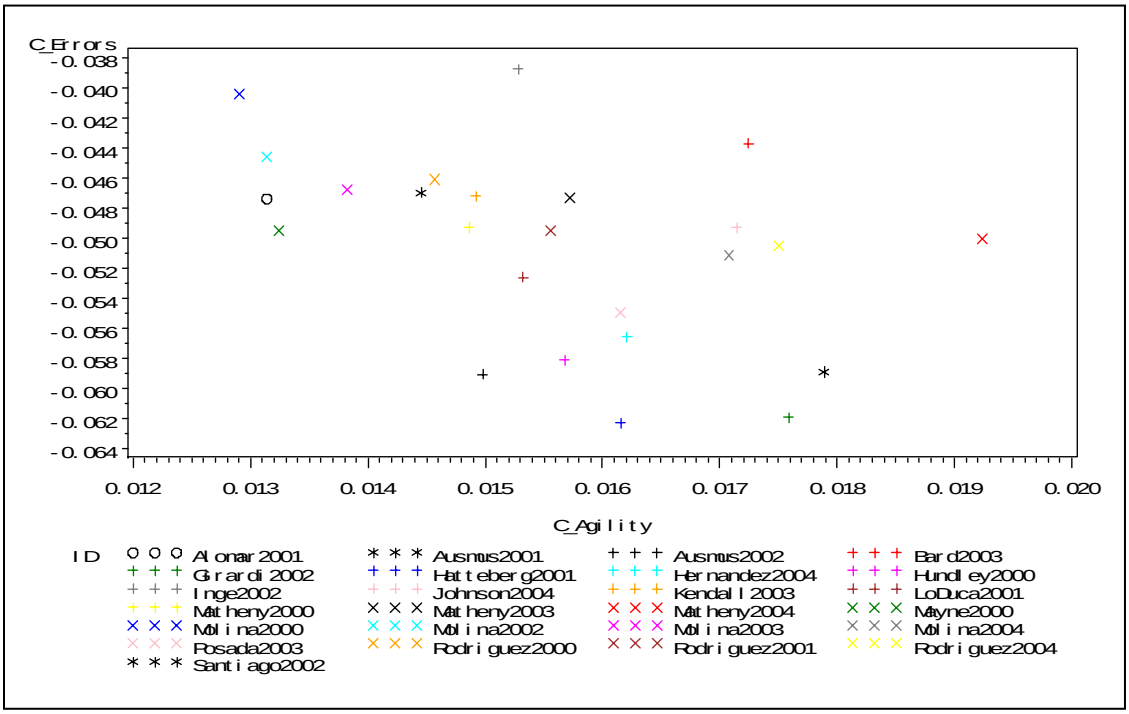
I.11 Catcher: Teamwork vs. Agility by Player



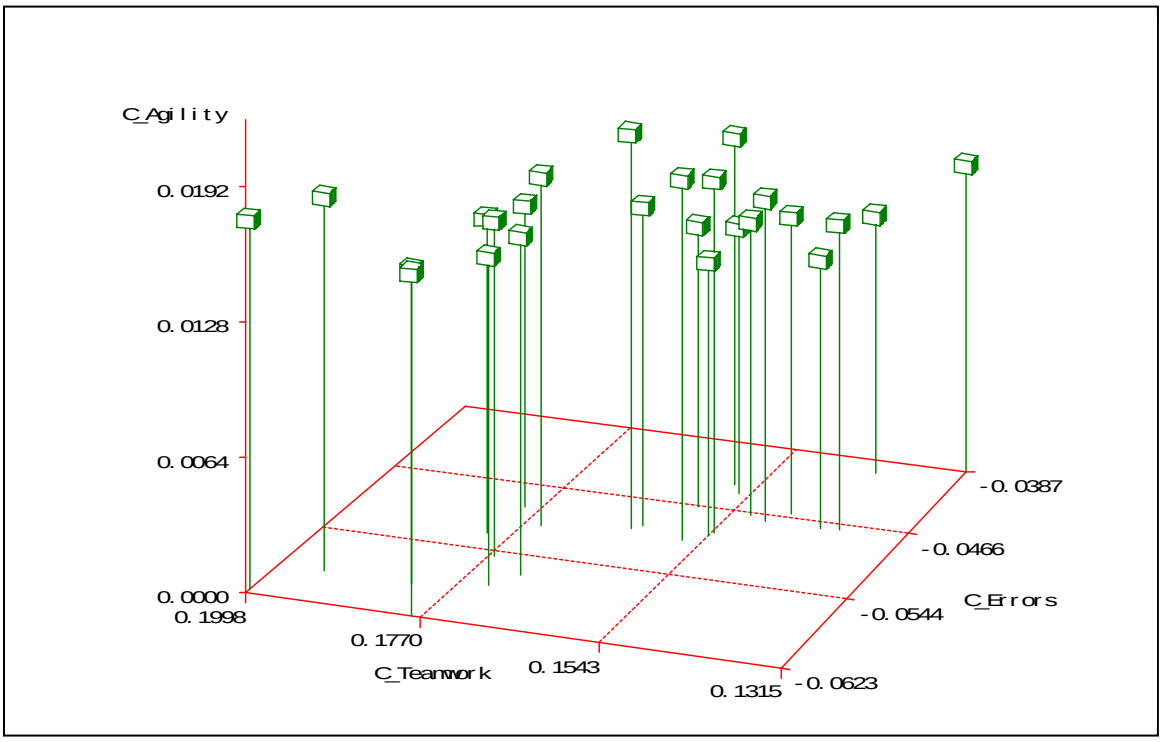
I.12 Catcher: Errors vs. Agility by Year and Gold Glove



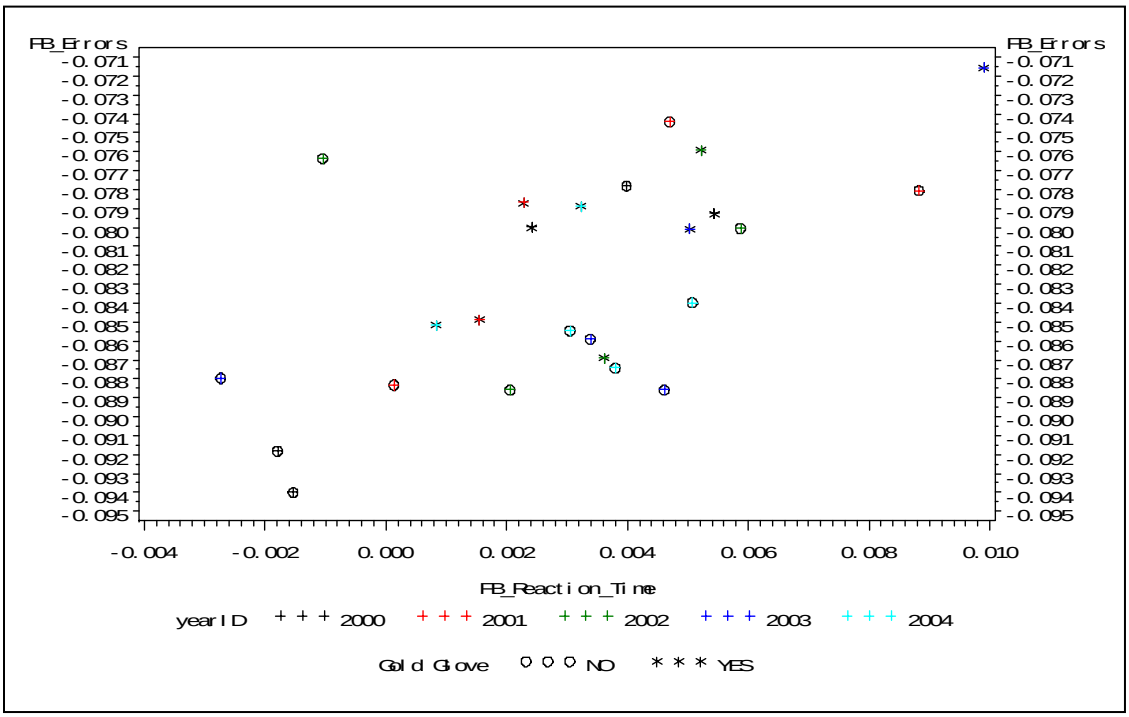
I.13 Catcher: Errors vs. Agility by Player



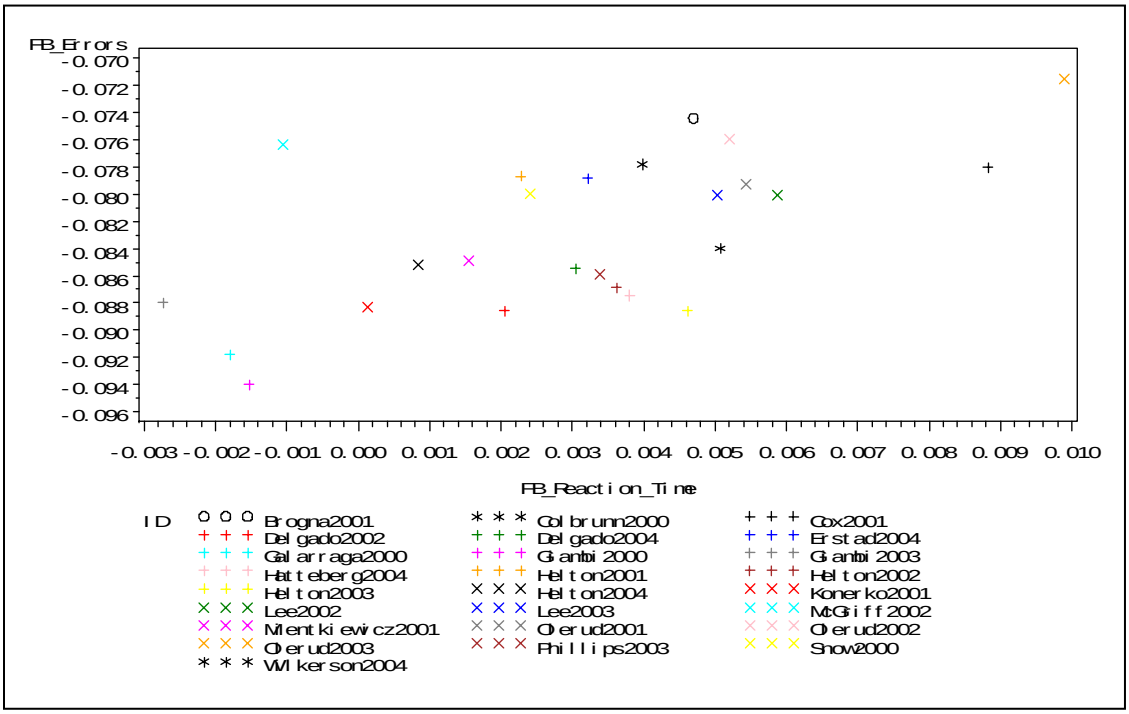
I.14 Catcher: Agility vs. Teamwork vs. Errors



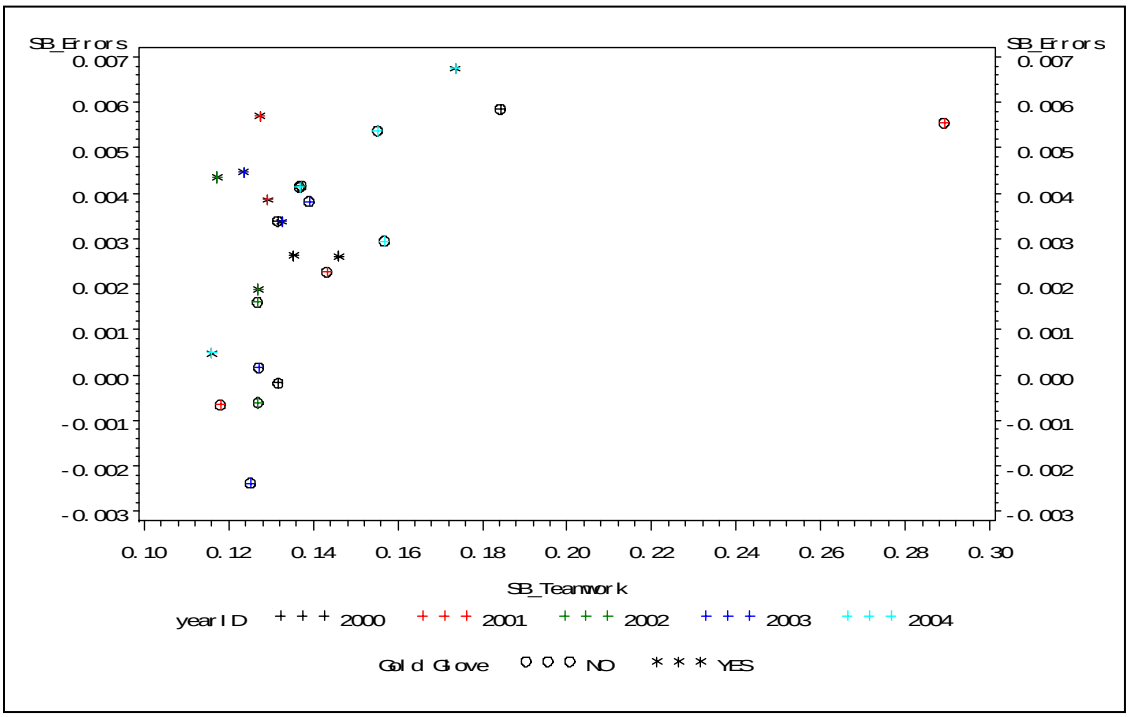
I.15 First Base: Errors vs. Reaction Time by Year and Gold Glove



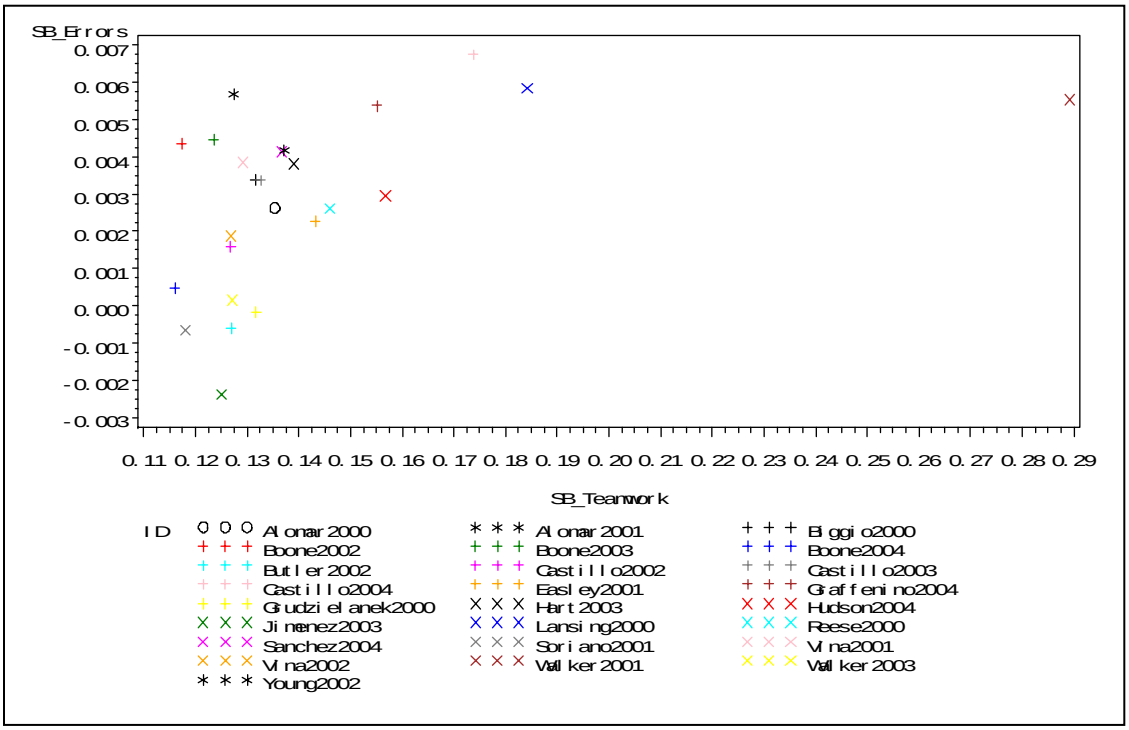
I.16 First Base: Errors vs. Reaction Time by Players



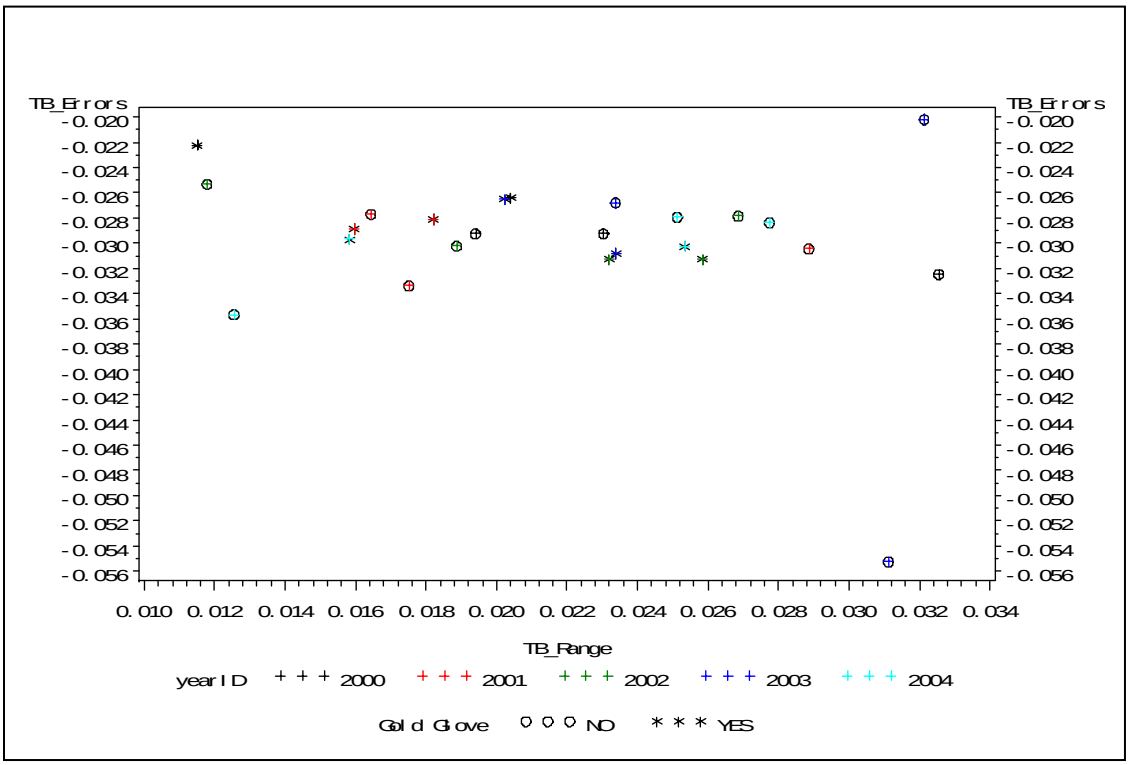
I.17 Second Base: Errors vs. Teamwork by Year and Gold Glove



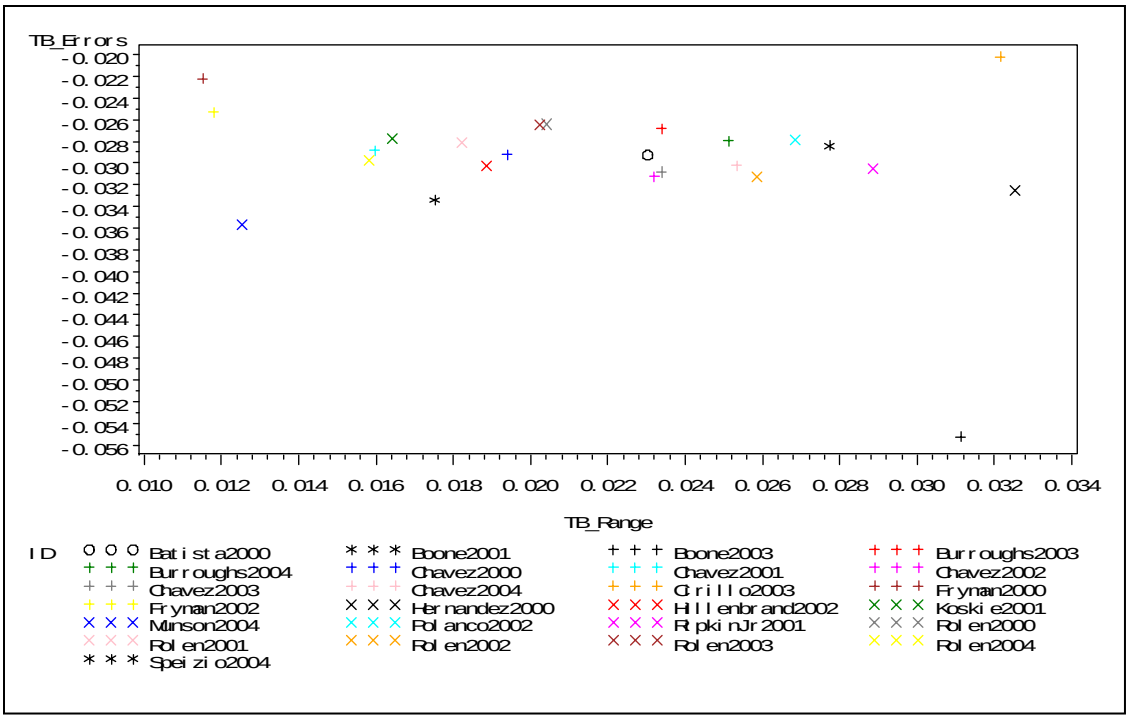
I.18 Second Base: Errors vs. Teamwork by Player



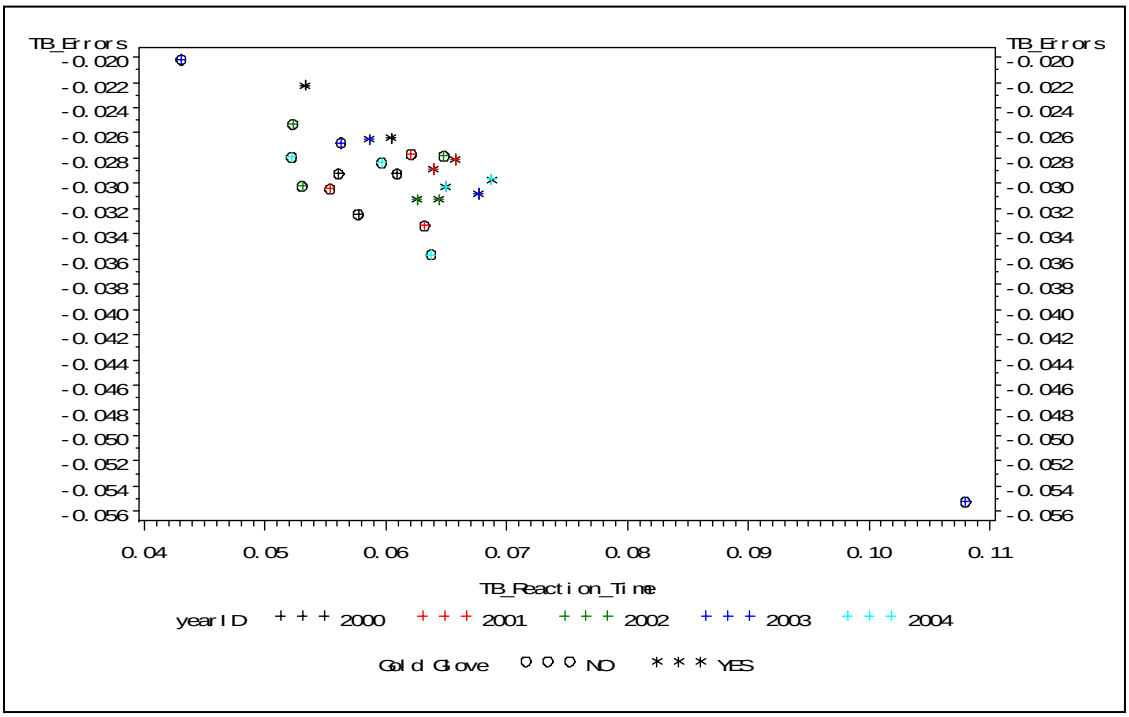
I.19 Third Base: Errors vs. Range by Year and Gold Glove



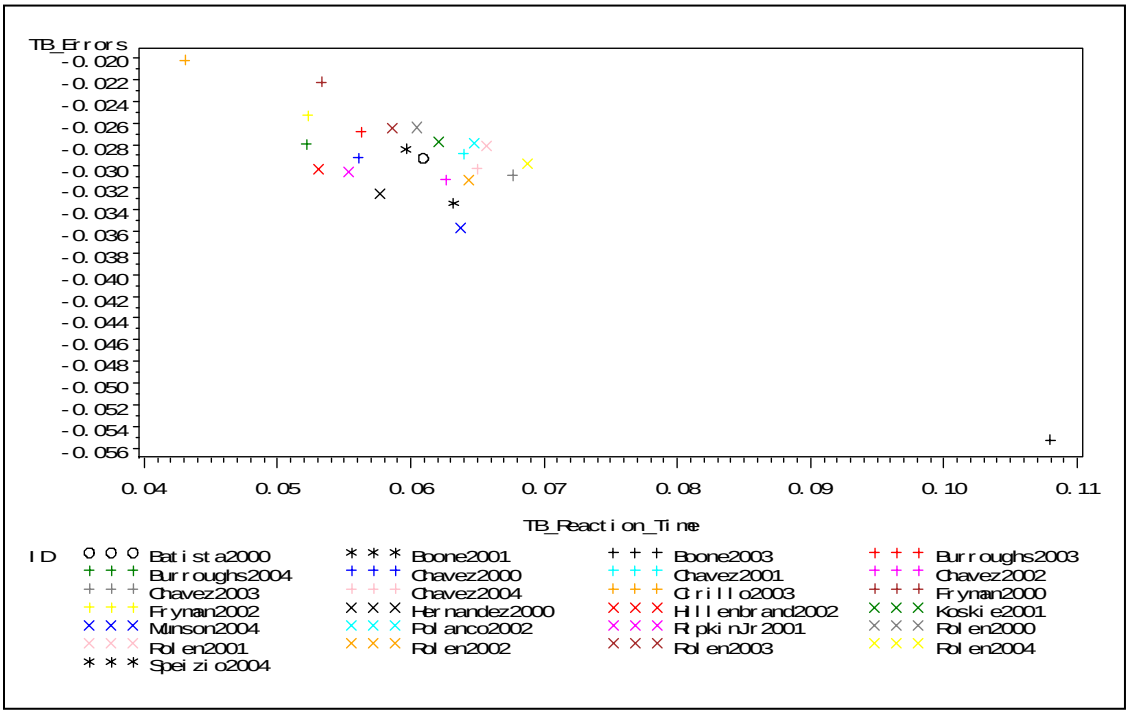
I.20 Third Base: Errors vs. Range by Player



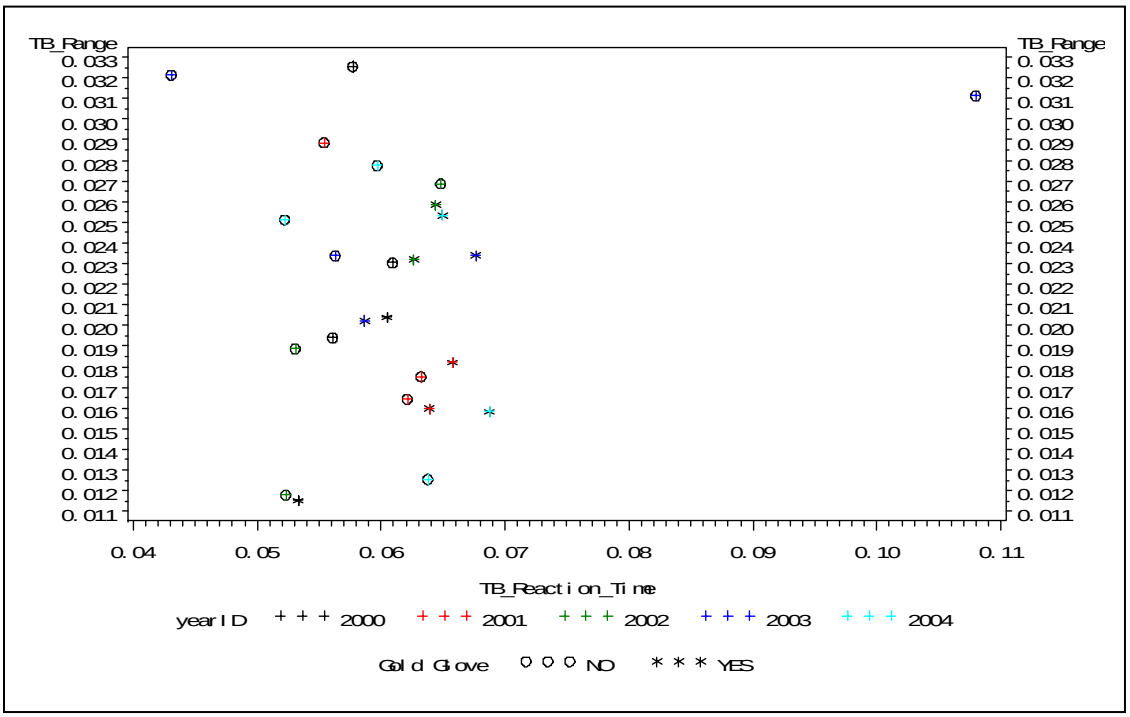
I.21 Third Base: Errors vs. Reaction Time by Year and Gold Glove



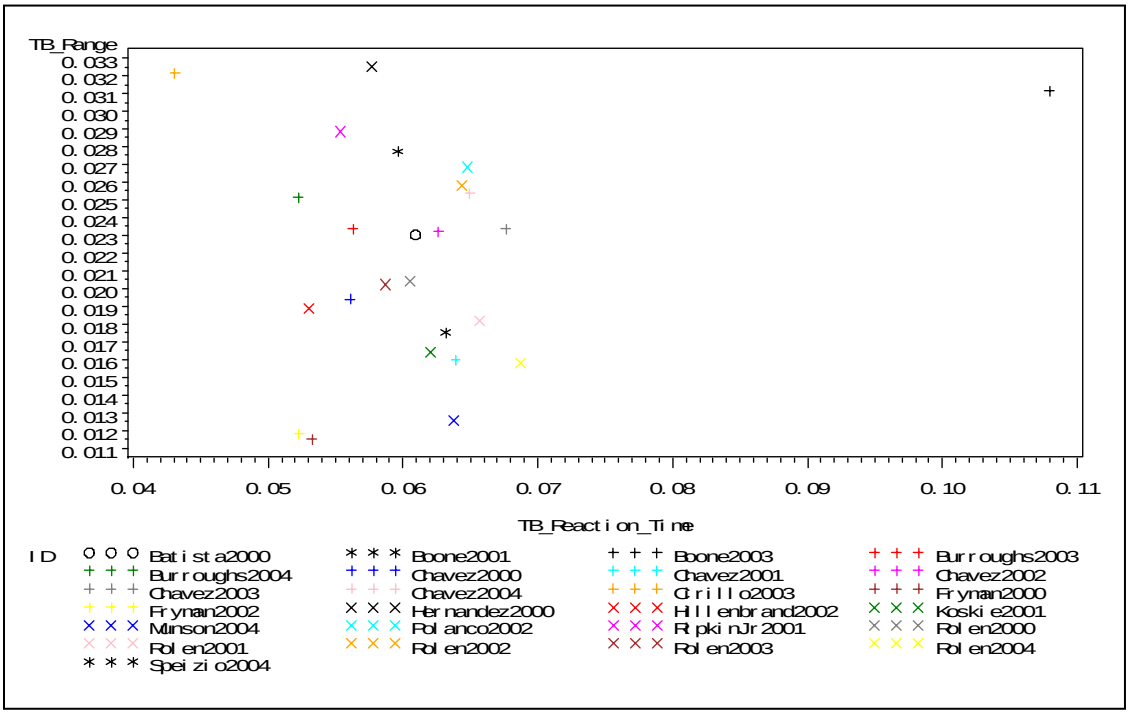
I.22 Third Base: Errors vs. Reaction Time by Player



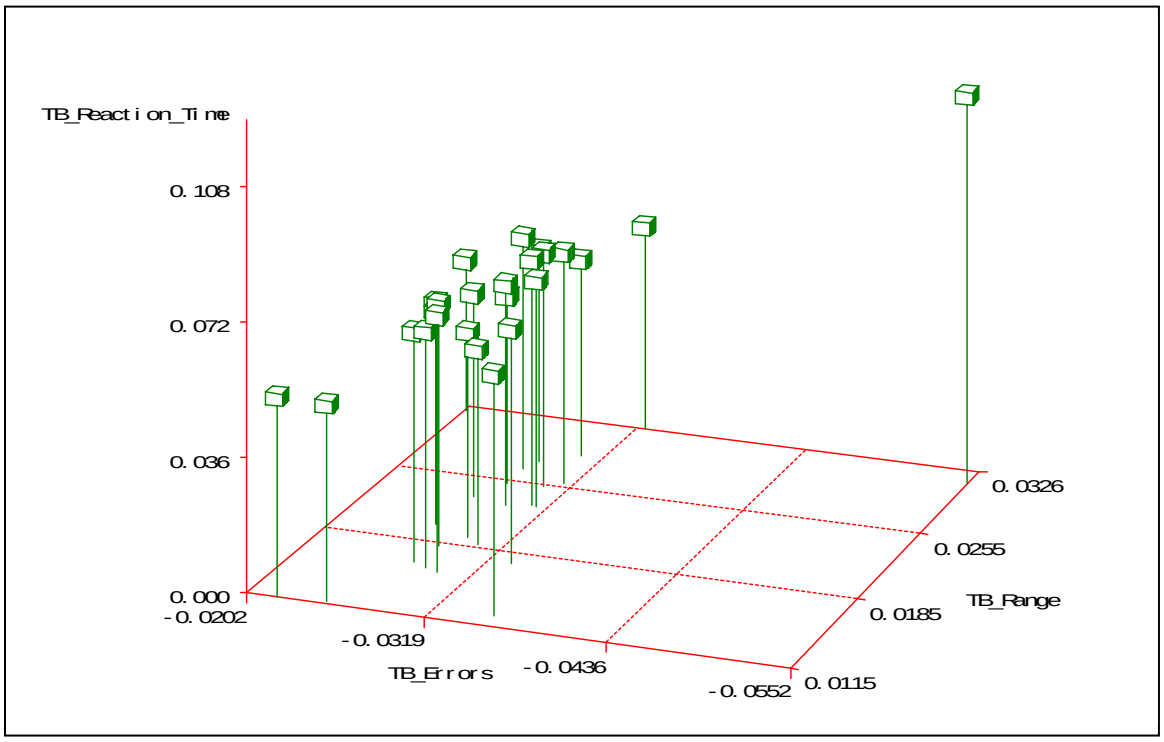
I.23 Third Base: Range vs. Reaction Time by Year and Gold Glove



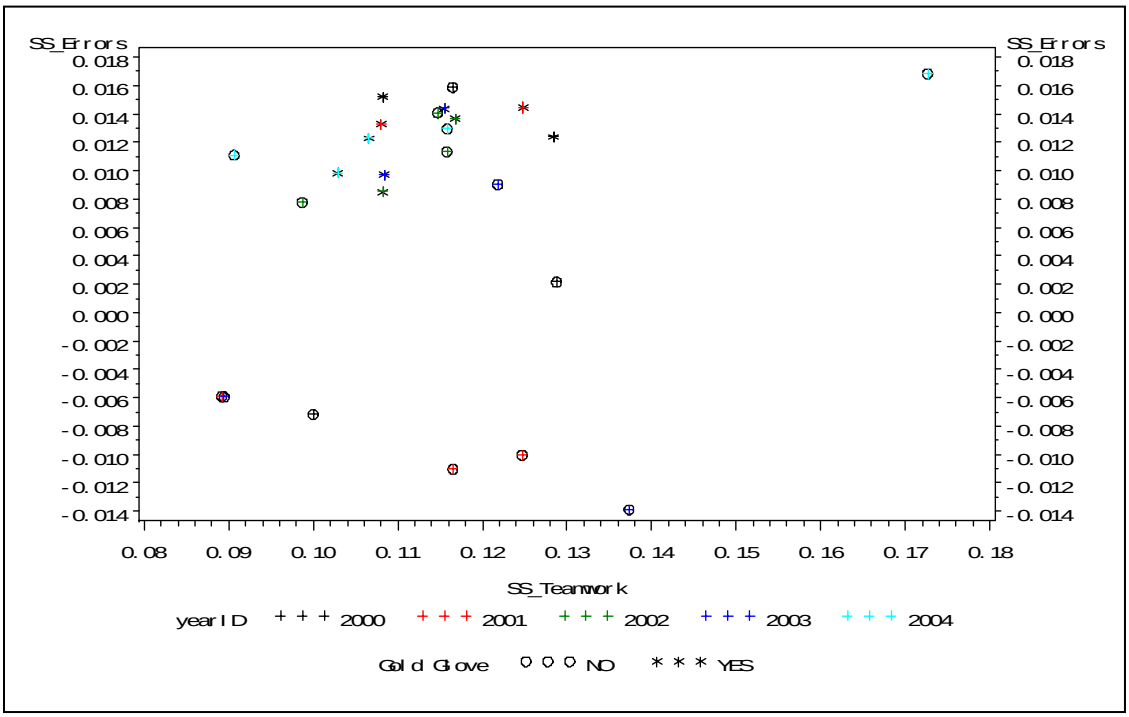
I.24 Third Base: Range vs. Reaction Time by Player



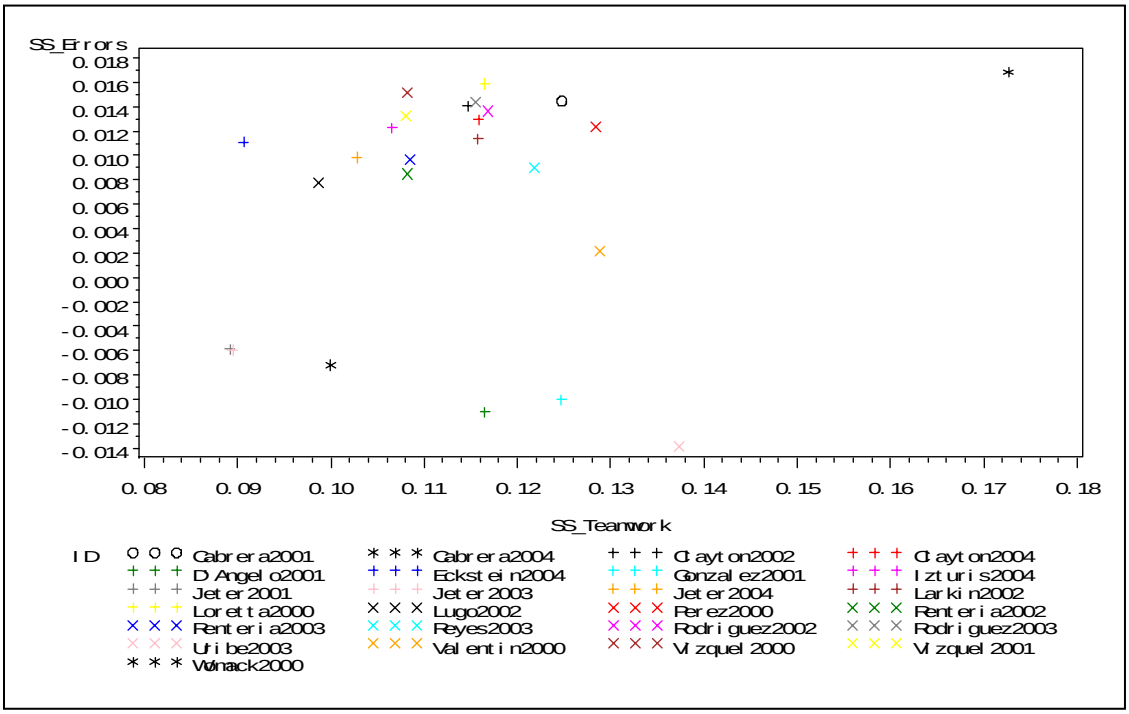
I.25 Third Base: Reaction Time vs. Errors vs. Range



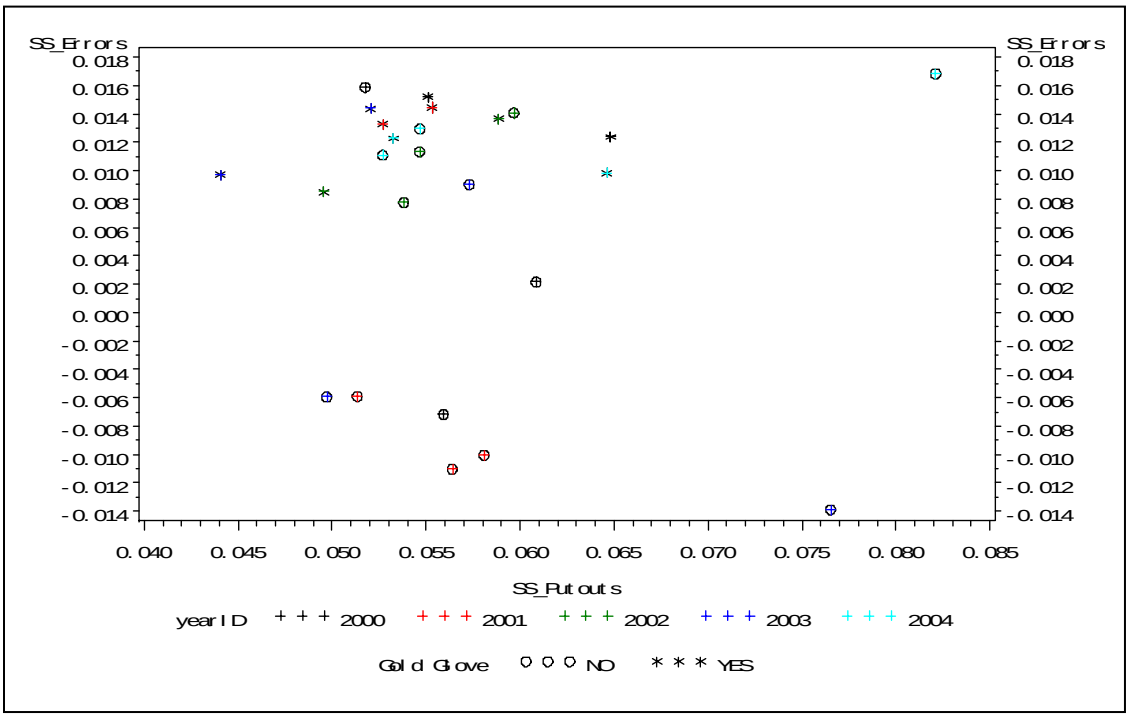
I.26 Shortstop: Errors vs. Teamwork by Year and Gold Glove



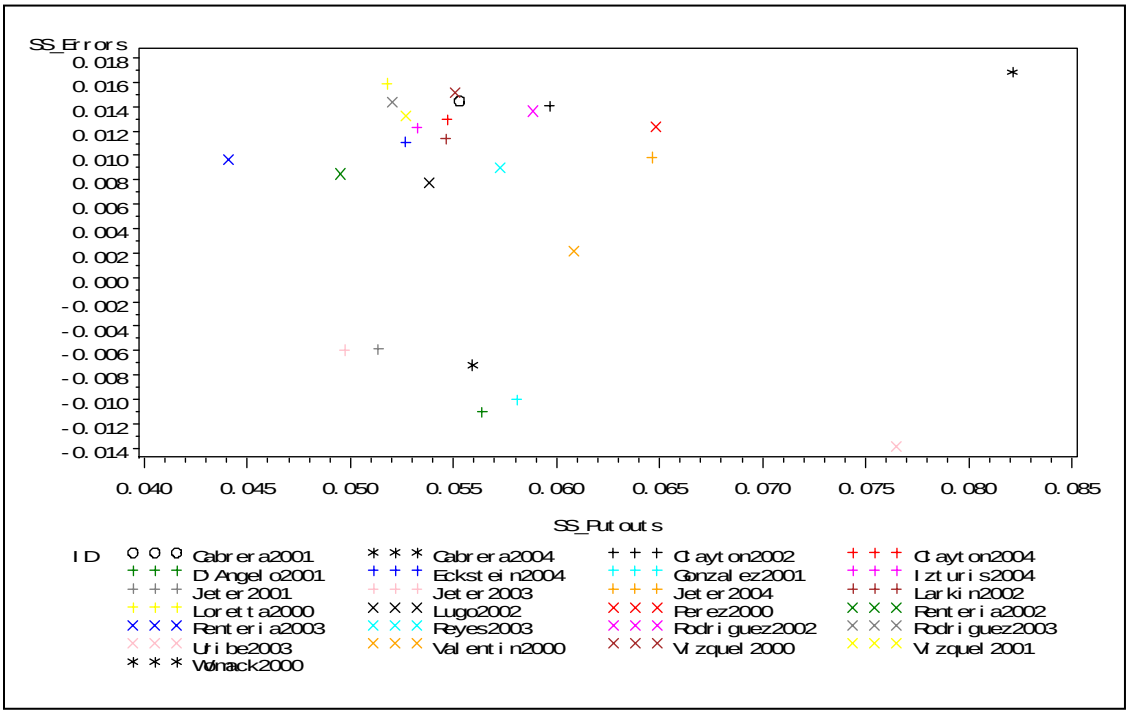
I.27 Shortstop: Errors vs. Teamwork by Player



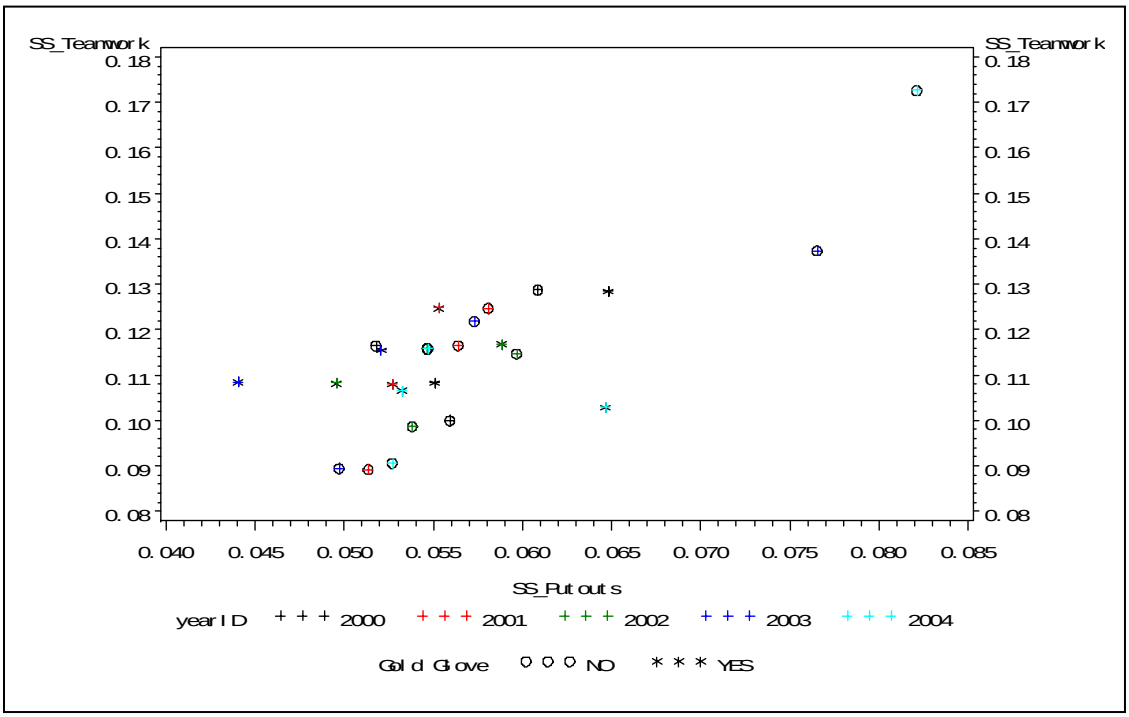
I.28 Shortstop: Errors vs. Putouts by Year and Gold Glove



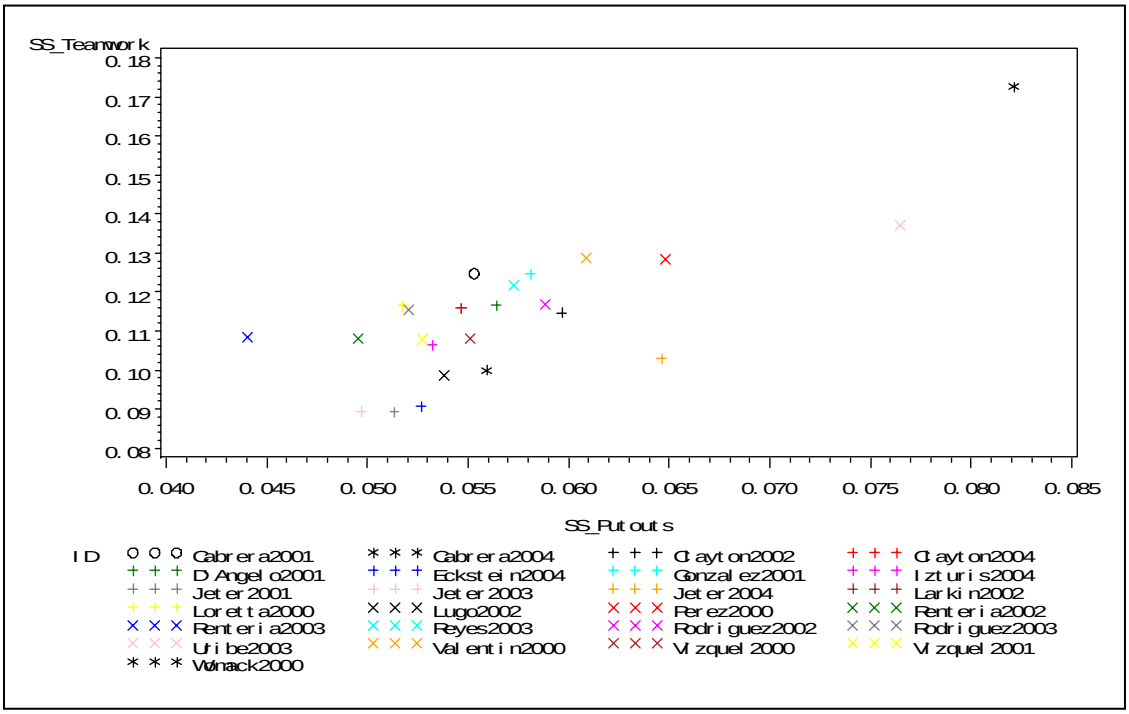
I.29 Shortstop: Errors vs. Putouts by Player



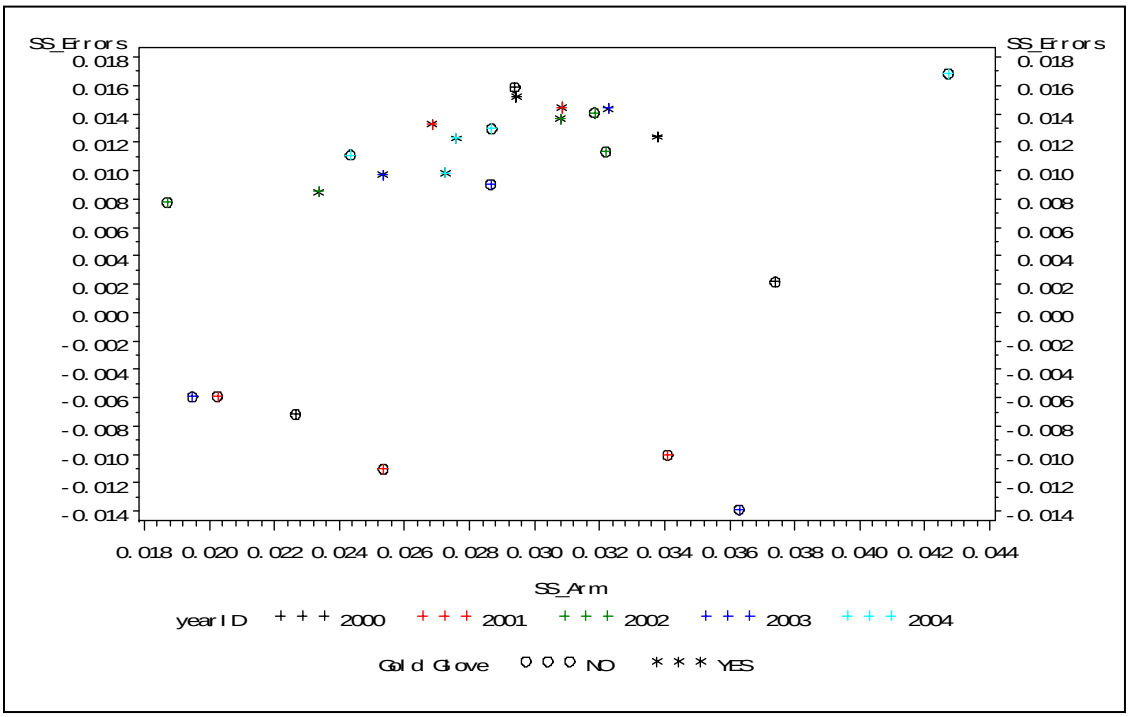
I.30 Shortstop: Teamwork vs. Putouts by Year and Gold Glove



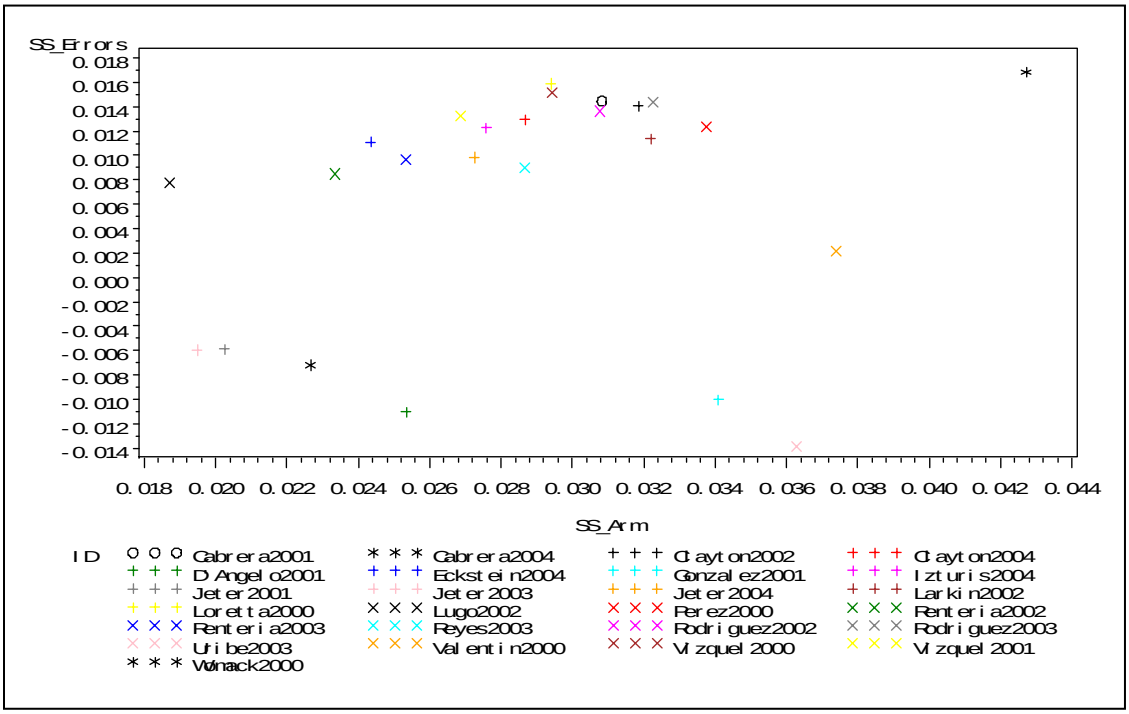
I.31 Shortstop: Teamwork vs. Putouts by Player



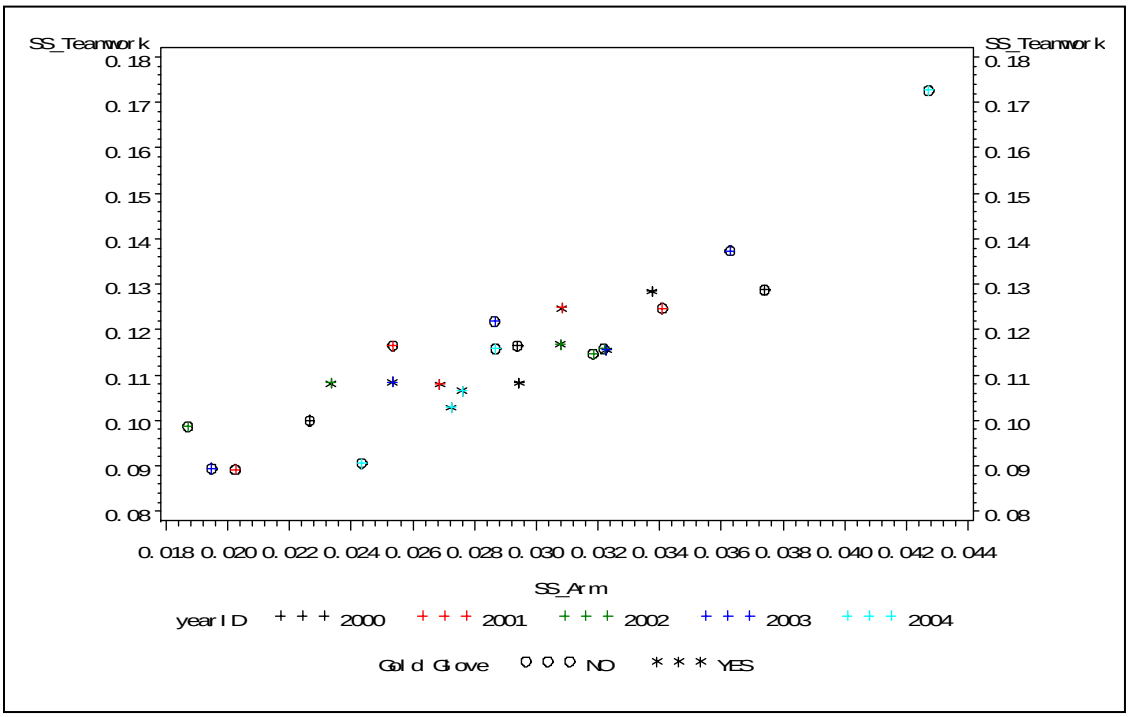
I.32 Shortstop: Errors vs. Arm by Year and Gold Glove



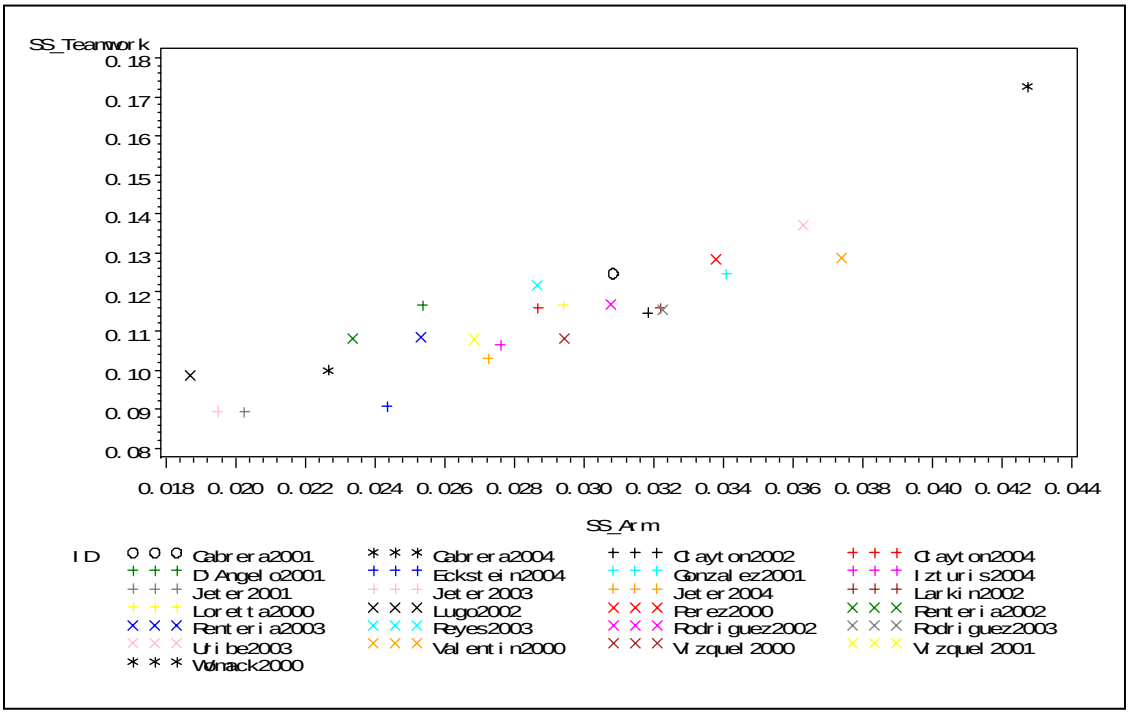
I.33 Shortstops: Errors vs. Arm by Player



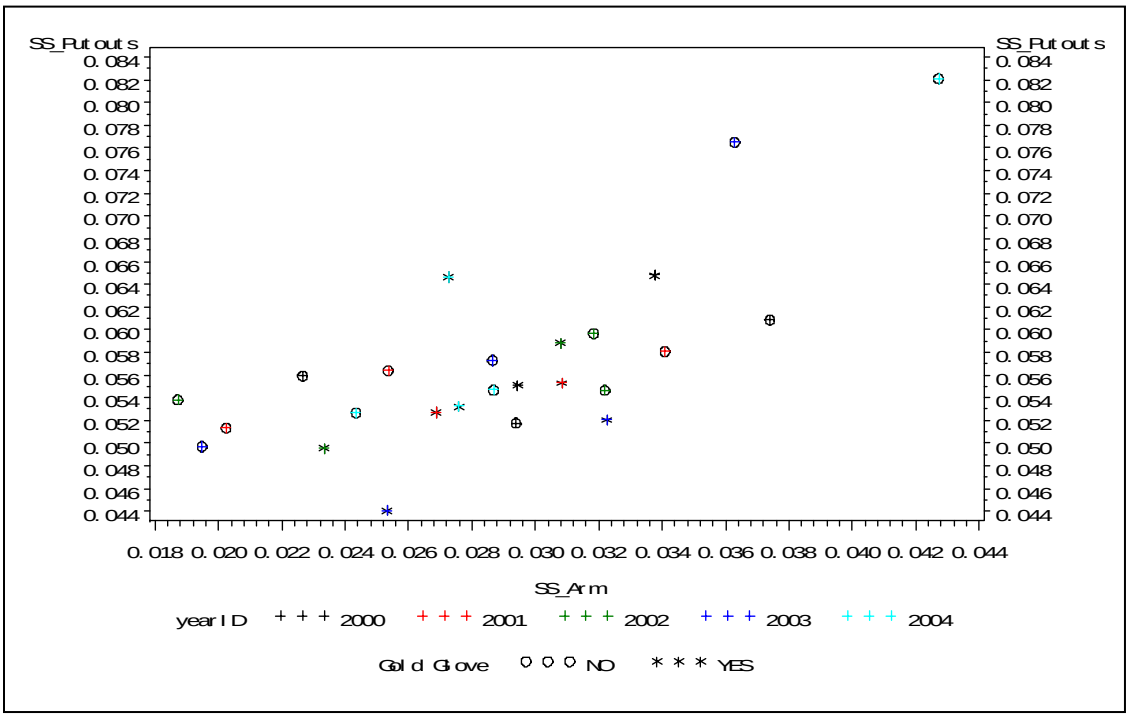
I.34 Shortstop: Teamwork vs. Arm by Year and Gold Glove



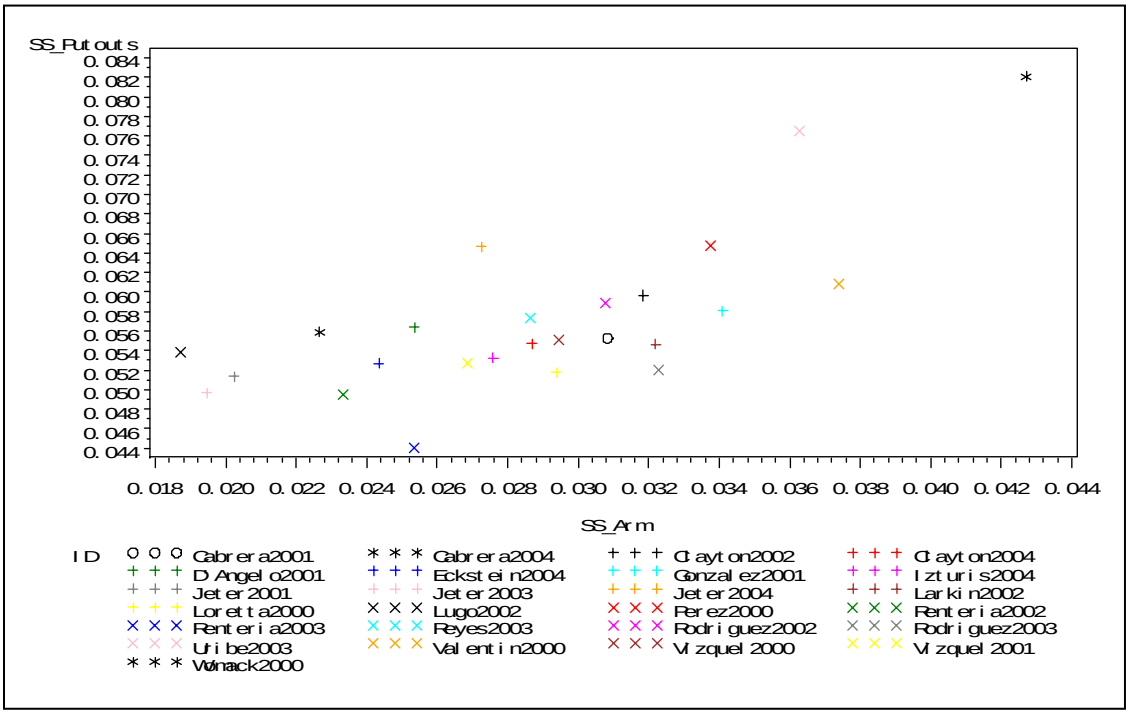
I.35 Shortstop: Teamwork vs. Arm by Player



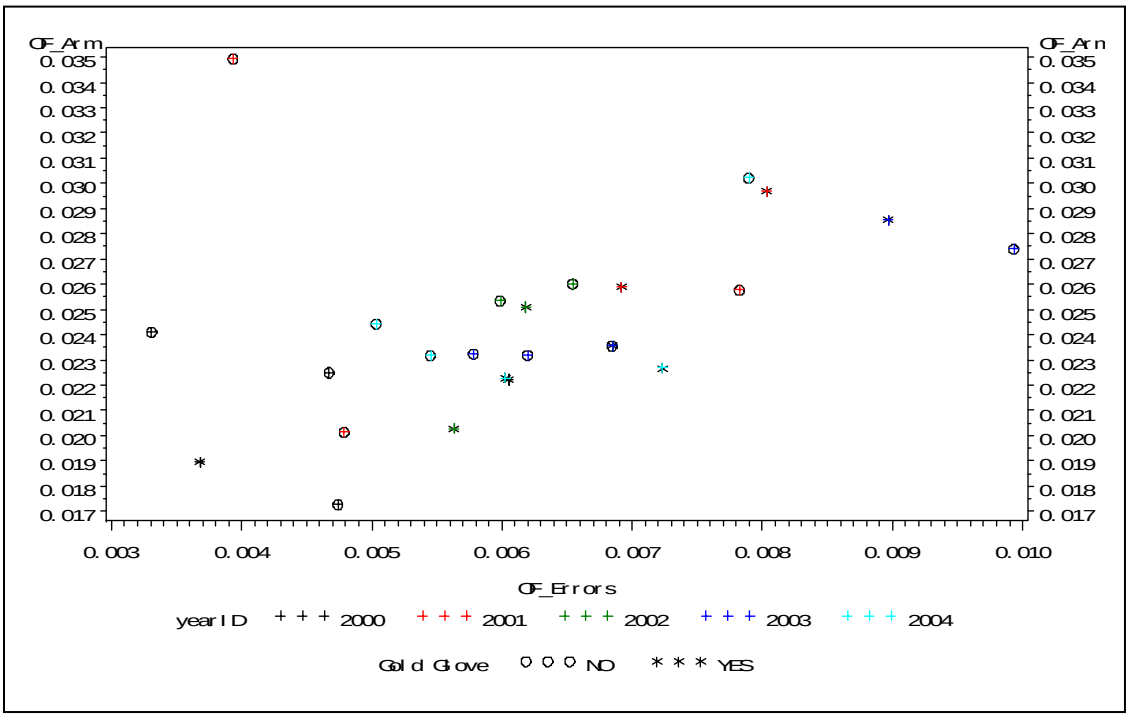
I.36 Shortstop: Putouts vs. Arm by Year and Gold Glove



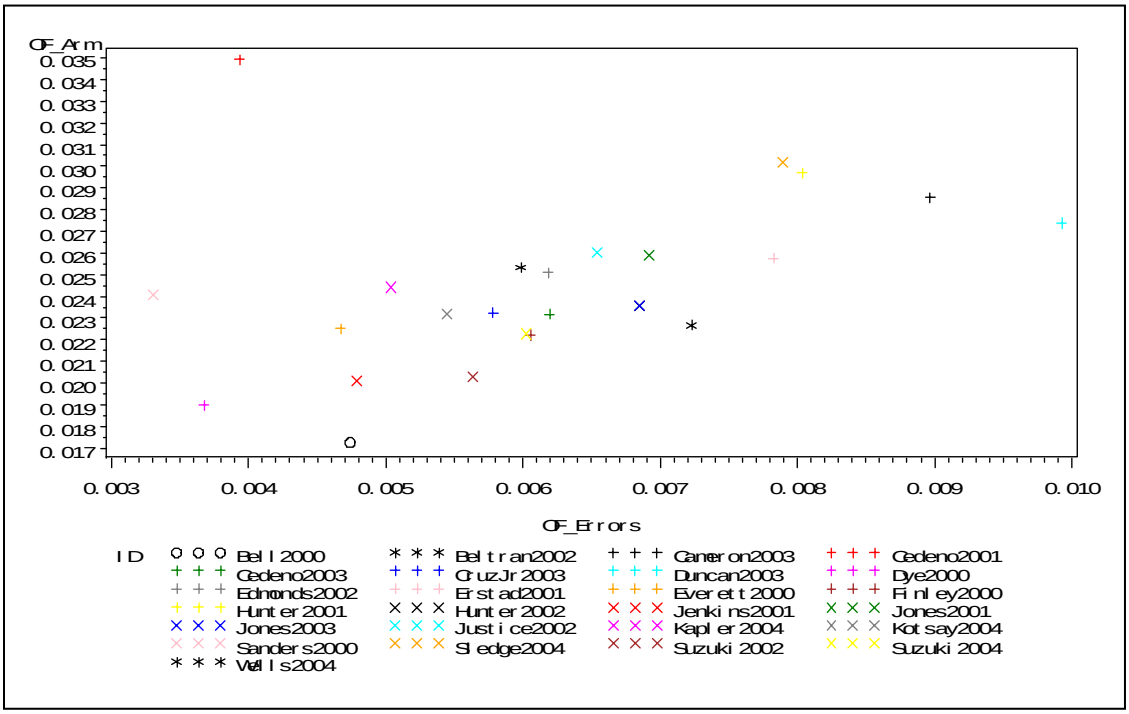
I.37 Shortstop: Putouts vs. Arm by Player



I.38 Outfield: Arm vs. Errors by Year and Gold Glove



I.39 Outfield: Arm vs. Errors by Player



Appendix J: Plays per Position Totals by Year

The following chart displays the total number of plays that each position was involved in for the 2000-2004 seasons.

	Pitcher	Catcher	1st Base	2nd Base	3rd Base	Shortstop	Left Field	Center Field	Right Field
2000	8351	4387	44801	22258	13513	20824	20177	21693	18838
2001	8866	4687	46029	21900	13568	20470	20968	21257	19429
2002	6979	4202	45528	21231	13647	20501	20353	21226	19331
2003	9457	4501	46474	23509	14548	21842	20749	24360	20321
2004	9249	4511	45133	23346	14659	21930	20173	24247	20838

Appendix K: Factor Names and Description

Below is a chart that contains the name and variable components of every factor found in the final model.

Position	Factor Name	Components
P	P_Errors	Errors
P	P_Reaction_Time	Assists; GBLD
P	P_Putouts	Putouts
C	C_Teamwork	Assists; Caught Stealing
C	C_Errors	Errors
C	C_Agility	Outs On Bunts
1B	FB_Errors	Errors; Errors on Ground Balls
1B	FB_Reaction_Time	Outs On Bunts
2B	SB_Errors	Errors; Errors on Ground Balls
2B	SB_Teamwork	Assist; Double Plays
3B	TB_Errors	Errors; Errors on Ground Balls
3B	TB_Range	Putouts; Outs in Foul Ground
3B	TB_Reaction_Time	Outs on Bunt; Assists
SS	SS_Errors	Errors; Errors on Ground Balls
SS	SS_Teamwork	Assists; Double Plays
SS	SS_Putouts	Putouts
SS	SS_Arm	Relay Throws
OF	OF_Arm	Assists
OF	OF_Errors	Errors