

# Modeling Election to the Major League Baseball Hall of Fame through the use of Genetic Algorithms

Abstract: This paper will use an alternate methodology for modeling called Genetic Algorithms. Using that method, several logical, rather than mathematical, rules for election to the baseball Hall of Fame will be found and examined. Predictions about future election, as well as past elections will be made. Ultimately, one rule will be picked as best, and examined in more depth than the others.

By David Cohen  
Friday, December 27, 2002

# TABLE OF CONTENTS

<b>INTRODUCTION</b> .....	<b>3</b>
<b>DATA ANALYSIS</b> .....	<b>4</b>
ELIGIBILITY .....	4
THE MEMBERS OF THE SAMPLE.....	5
<b>WHAT ARE GENETIC ALGORITHMS?</b> .....	<b>6</b>
ESTIMATOR BIAS VS. SEARCH BIAS: AN EXAMPLE WITH ORDINARY LEAST SQUARES REGRESSIONS.....	8
<b>HOW ARE GENETIC ALGORITHMS IMPLEMENTED?</b> .....	<b>9</b>
REPRESENTATION .....	9
THE SELECTION PROCESS: FITNESS FUNCTIONS AND SELECTION.....	11
<i>Examples of Binary Fitness Functions</i> .....	11
<i>The Parent Selection Process</i> .....	13
CROSSOVER AND MUTATION.....	14
<b>LITERATURE REVIEW</b> .....	<b>16</b>
<b>SPECIFIC'S OF HALL OF FAME IMPLEMENTATION</b> .....	<b>18</b>
BASE 2 REPRESENTATION OF BASE 10 INTEGERS .....	18
BASIC MATCHING.....	19
M OF N MATCHING .....	21
<b>SPECIFICATION OF THE MODEL</b> .....	<b>21</b>
<b>RESULTS</b> .....	<b>27</b>
ACCURACY.....	27
ODDS RATIO.....	33
<b>PREDICTIONS</b> .....	<b>36</b>
<b>INTERPRETATION</b> .....	<b>38</b>
ODDS RATIO VS ACCURACY .....	38
TREAT GAMES/AB AS GOOD OR BAD.....	40
<b>PROBLEMS AND ALTERNATIVES</b> .....	<b>42</b>
<b>CONCLUSION</b> .....	<b>43</b>
<b>APPENDIX A</b> .....	<b>45</b>
<b>BIBLIOGRAPHY</b> .....	<b>45</b>

## Introduction

Most baseball writers, at one point or another, have undoubtedly written an article describing whether or not a given player deserves to be in the Hall of Fame. It is a common question, among writers, fans, and sabermetricians<sup>1</sup>. A not too far off corollary of this is comparing players from different time periods. These comparisons occur any time a record, which has been held for a long time, falls, such as Barry Bond's recent home run record.

In the instructions for Hall of Fame voting, it says that, "Voting shall be based upon the player's record, playing ability, integrity, sportsmanship, character, and contributions to the team(s) on which the player played." This being the case it would seem entirely reasonable to look at the player's record and playing ability, via his career statistics. Using those statistics, it should be possible to build a model of election to the Hall of Fame. This has been done several times using standard regression analysis. The results from these regressions are generally probabilities that a player will be elected. A problem is that these results are not easily understood, and cutoff for election is not an inherent thing. The question present is, "At what percent do we become sure that a player is elected?" Is it 50%, 80% or higher? Part of this problem comes from the fact that regressions fit data to continuous functions. As such, the results are continuous as well. Even Logit and Probit models, specifically tailored to binary problems, produce continuous values.

By using an alternative methodology known as the Genetic Algorithm, a logical rule can be created. Genetic Algorithms excel at jumping around a large search space, to find near optimum solutions. The rule that will be searched for will be in the form of "if

(a & b & c), then the player will be elected to the Hall of Fame.” This eliminates the problem of a continuous dependant variable for binary questions.

This paper will present a short look at eligibility for the Hall of Fame, and what the Hall of Fame is. Next, an in depth description of the basics of GAs will be examined. Following that, previous work in the fields of predicting election to the Hall of Fame and Genetic Algorithms will be examined. The next section will look at the specific implementation used for the models in this paper. The results will then be presented followed by a series of predictions and some interpretations of the results. Finally, a discussion of some of possible pitfalls of using this approach are addressed.

## **Data Analysis**

The dataset used in this paper is a collection of career statistics of Major League Baseball players. In order to for a player to be included in the set, he had to fulfill several requirements. Firstly, the player had to be a position player, as opposed to a pitcher, for the majority of his career. Secondly, the player had to be eligible for the Hall of Fame.

### ***Eligibility***

In order to be eligible for election into the Hall of Fame in baseball there are several requirements. First, the player had to play for 10 years in the major leagues. Secondly, the player has to have been retired for five years. In addition to these requirements, the player cannot be “banned for life” from baseball, as Pete Rose is.

Not every eligible player appears on a ballot for the Hall of Fame, however. Before voting occurs there is a screening committee that decides which players should be placed on the ballot. The screening committees is made up of six members of the

---

<sup>1</sup> Sabermetrics was originally developed by Bill James, and is the study of baseball statistics.

Baseball Writers Association of America (BBWAA), and are elected to two year terms to serve on the committee, on a rotating basis. In order to appear on the ballot, a player must have been on the ballot in the previous year, or be in his first year of eligibility and have all six members of the screening committee approve his placement on the ballot. In addition, a player may only appear on 15 ballots, after which, he cannot be elected in the normal manner. In 1981, a change was made so that in order for a player to remain on the ballot he needed to receive at least five percent of the vote in the previous year.

Elections occurred every year until 1956, at which point elections were held every other year. Beginning in 1966, elections were returned to an annual schedule. Members of the BBWAA who have been members for at least 10 years are eligible to cast votes. Each elector may vote for no more than 10 players on a given ballot. Write in candidates are not allowed. In order to be elected, a player must receive a vote on at least 75 percent of the ballots returned<sup>2</sup>.

### ***The Members of the Sample***

For this specific sample, the player had to have retired in between 1957 and 1996, and have been eligible to appear on at least one ballot<sup>3</sup>. For each player, several statistics describing the player's performance in both the regular season and post-season are included. These statistics were compiled from Sean Lahman's database, which is used widely on the internet.<sup>4</sup> The compilation of this data was done exclusively inside of several computer programs written especially for this project to eliminate errors from the transfer.

---

<sup>2</sup> From the Baseball Hall of Fame website: <http://www.baseballhalloffame.org>

<sup>3</sup> This includes 1957 and 1996

<sup>4</sup> Sean Lahman's Database can be retrieved from the website <http://www.baseball1.com>

The only part of the compilation that was done outside of the computer was putting together the data on the Hall of Fame voting data. The only place to get the information of who was on each ballot was the actual ballots themselves. For this reason the names on each ballot had to be typed in, and matched with the number of votes each player received, in each year. In order to minimize errors, this data was typed in twice, and the two copies were compared with one another, to search for errors or omissions. Any mistakes found were then corrected, and incorporated into the data set.

## **What are Genetic Algorithms?**

Genetic algorithms (GA) are a type of machine learning loosely based on Darwin's theories of evolution. In this form of machine learning, there exists a group of hypotheses, or possible solutions to the problem, called the population. At each step of the learning procedure, often called a generation (to continue the biological analogy), the population is replaced by offspring from that generation. Deciding which members of the population are used as "parents" for the next generation is based upon the fitness of possible solutions in the parent generation. A relatively normal approach to the problem would be as follows:

- 1) Start by creating a population of random solutions to the problem. More than likely none of these original hypotheses will be great solutions, individually. However, it will serve as a good starting point for the search, providing a diverse base of possible solutions to start from.

- 2) The GA then proceeds through generations. In each generation:
- a. Assign a measure of how good each possible solution is by using a “fitness function.” This allows the GA to be able to judge which solutions are better than others.
  - b. Select parents by using the fitness associated with each of the possible solutions. Similar to the Darwin analogy, solutions with a higher fitness have a higher probability of “mating” to form offspring. In this manner a solution with a low fitness still has a chance of becoming a parent, just a smaller one than a solution with a high fitness.
  - c. Use operators such as crossover, which combines parts from the two parent solutions to form the child solutions, and mutation, which randomly changes parts of a given a solution, in order to form the new population.

This process will continue until a set number of generations has passed<sup>5</sup>.

The appeal for using GAs comes from many places, including the ability to model complex relationships between different attributes, not necessarily practical using standard econometric techniques such as regression analysis. In addition, the ever-reducing cost of computing power, makes the use of this computationally intense method more reasonable to use.

---

<sup>5</sup> Congdon et al. Genetic Algorithms in Cladistics.

## ***Estimator Bias vs. Search Bias: An Example with Ordinary Least Squares Regressions***

Different algorithms search for answers to different questions. In an economic student's first modeling class, he learns that the Ordinary Least Squares (OLS) method of estimation yields an unbiased estimator. While the estimator is unbiased, the search for the estimator is not. In the OLS problem, we bias the search for an equation that models our data well, while minimizing the sum of the squared residuals. In an unbiased search, one has a truly random search, where it would be just as likely to come up with an equation which models the data poorly, and has an exceptionally high sum of the squared residuals. This would obviously be counterproductive, and as such, we steer our search into the section of search space that we are interested in: mainly the section of the search space where the sum of the squared residuals is minimized. We will call this kind of bias a search bias. It is important to realize that is very possible, and most often necessary, to have a biased search in order to find an unbiased estimator.

In the OLS method, the eventual solution of parameters could be thought of as the best solution, or rule for fitting the data. In a GA, the task addressed is similarly to find the best rule. In a GA, the best rule can be thought of as the specific rule which optimizes a specified numerical measure for the specific problem. This measure is called the hypothesis fitness. For our purposes, the problem is that of estimating an unknown, or dependant value, given a set of known, independent variables. As such, it would make sense for the fitness to be a measure of how accurately the hypothesis models the data. Part of the appeal of GAs is the ability for the user to handcraft the function that calculates the fitness to his or her specific needs. In some fields, such as medicine, false

negatives (a test result of negative when in fact it should have been positive) can be very dangerous, and so someone using a GA for that purpose might use a metric in which hypothesis which creates large numbers of false negatives would have a relatively lower fitness than one that creates fewer false negatives. More about this will be discussed in a later section.

## **How are Genetic Algorithms Implemented?**

While certain things vary in different implementations, there are several factors, which are generally found in a GA. Genetic algorithms start with a pool of potential rules called the population, which is updated repeatedly. Each update is called a generation. In each generation, each of the member hypotheses is evaluated by a fitness function. From this initial population a new population is formed by selecting some of the best rules, and allowing them to continue on to the next generation. Those solutions, which are allowed to continue onto the next generation, unchanged, are known as the “elite” of the generation. The number of elite is a parameter that is set before the experiment begins. Applying common genetic operations called crossover and mutation, on the parent rules, forms the remainder of the child generation.

### ***Representation***

The hypotheses in the population are very simple at first glance. Every hypothesis is no more than a series of ones and zeros, put together into a bit string. For instance, a sample string might be 11001. It is up to the programmer of the fitness function to decide how to interpret the bit string. One possible example might be a program whose function is to determine whether it is a good day to play golf.

<b>TABLE 1</b>	<b>A SIMPLE HYPOTHESIS</b>	
Attribute Number	Value	Meaning
1. Weather	1	Cloudy =0, Sunny =1
2. Wind	1	Windy=0, No wind=1
3. Humidity	0	Humid = 0, not Humid=1
4. Temperature	0	Over 90 or below 68 =0, other wise =1
5. Play Golf?	1	No = 0, Yes = 1

The above table shows a simple hypothesis that can be interpreted as several if statements combined with and operators. In English this hypothesis would be, “If it is sunny, and there is no wind, and it is not humid, and the temperature is comfortable, then I should go play golf.” In the above example, each attribute can have only two possible values (0,1). In the real world, it can be useful to be allowed more than one bit per attribute. Below in table two a more complicated example is shown.

<b>TABLE 2</b>	<b>A MORE COMPLEX HYPOTHESIS</b>	
Attribute Number	Value	Meaning
1. Weather	10	Cloudy=00, Sunny=01, Don't Care =10 , 11
2. Wind	11	Windy=00, No wind=01, Don't Care =10 , 11
3. Humidity	10	Humid = 00, not Humid=01, Don't Care =10 , 11
4. Temperature	01	Over 85 =00, 75-84 =01, below 74 =10, don't care = 11
5. Play Golf?	1	No = 0, Yes = 1

Another hypothesis suggests that a golfer's desire to play can be shown by the hypothesis 101110011. According to our new definitions for the different values of the bit string, this would suggest that the golfer does not care about the weather, wind, or humidity, but the temperature has to be between 75 and 84 degrees for the golfer to want to play golf. A “Don't Care” value is a very useful tool in GAs, and can be used so that not every attribute made available to program is used in the rule for prediction.

## ***The Selection Process: Fitness functions and selection***

The fitness function is used as a way to rate the different hypotheses of the population. This function is custom tailored to the individual problem at hand. The interpretation of the individual bit strings takes place in the fitness function. As discussed earlier, most of the time, the string is interpreted as a complex “if... then” statement. When the problem being solved is an attempt to model data, then the fitness function generally includes some measure of how accurately the hypothesis describes the data.

## **Examples of Binary Fitness Functions**

As mentioned before, there are countless measures of how well a hypothesis describes the data. This section will go over four metrics for binary hypotheses, or hypotheses where one is trying to predict a yes or no answer. The names are sensitivity, specificity, Accuracy, and Odds Ratio.

In all of these examples the data set can be split up into positive examples and negative examples, for instance, a player was elected to the Hall of Fame, or not elected. In addition, any hypothesis can classify a positive as a positive, or a positive as a negative. It is also possible for a hypothesis, or rule, to classify as negative example as a negative, or a negative as a positive. In order to test each hypothesis, the data set is divided into 4 groups, true positives(a), false positives(b), true negatives(c), and false negatives(d)<sup>6</sup>.

---

<sup>6</sup> The table below is taken from Congdon PhD Thesis. Definitions of the metrics are from same source

		Data Set	
		In	not in
Rule	classified in	a	b
	classified out	<u>c</u>	<u>d</u>
		a+c	b+d

In order to better understand, consider an example where the data set contains several players including, Hank Aaron, who was elected to the Hall of Fame, and Joe Torre, who was not. Because of the fact that Hank Aaron was elected to the Hall of Fame, he is considered to be a positive example in the data and would be placed into one of the In categories: either a or c. In the same way, Joe Torre, a negative example (someone who did not get elected), would be a member of the “not in” category: either b or d. Now if the rule describing who gets into the Hall of Fame states that if a player hit more than 20 career home runs, he should be elected to the Hall of Fame, then both Joe Torre and Hank Aaron would be described as positive examples by the rule. This would make Hank Aaron a member of category “a”, or a true positive, as he is *correctly* identified as a positive example, while Joe Torre would be a member of the category “b” as he is falsely identified as being positive example.

Sensitivity measures the number of positive examples that are classified as positive, regardless of the number of false positives or true negatives. This formula for this metric is  $a/(a+c)$ .

Specificity is the opposite of the sensitivity, measuring the number of negative examples correctly identified as a negative example by the rule. The formula for this is  $b/(b+d)$ .

Accuracy is a metric that tries to address some of the possible shortcomings of sensitivity and specificity by combining the two using a weighted average. The sensitivity is weighted with the percent of positive examples in dataset, and the specificity is weighted with the percent of negative examples. The resulting metrics is  $(a+d)/(a+b+c+d)$ , where  $a+b+c+d$  is the number of observations in the data set.

One last metric is something called Odds Ratio. This particular metric has its roots in the field of epidemiology. This is the part of medicine that studies the causes and distribution of diseases. In this field, a false positive, is a very dangerous thing. If you are studying the causes of cancer, you don't want a false positive, saying that vitamin-C is a cause of cancer. As such, Odds Ratio is designed to minimize the number of false positives. The odds of an observation being positive is the probability that it is positive to the probability it is negative, or  $a/c$ . So too, the odds that a data point which is actually negative, is described by the rule as being positive can be shown as  $b/d$ . The ratio of the two of these  $(a/c)/(b/d)$  yields  $ad/bc$ , and this is called the Odds Ratio. This Odds Ratio tells us the odds that a rule classify positive as a positive against the odds the rule classifies a negative as a positive. Part of the reason why one would possibly prefer Odds Ratio is that it is not linear, unlike the other metrics. This non-linearity, creates a very high reward for describing the positive examples well, rather than describing the entirety of the positive examples.

## **The Parent Selection Process**

Once all of the hypotheses in a given generation are evaluated, it is time to choose which members go on to the next generation. This is accomplished though a method that resembles a giant roulette wheel, where the higher a hypothesis's fitness, the larger the

slice on the wheel. The key idea in this method is that no matter how low a specific example's fitness might be, there is still a chance that it can be chosen to move onto the next generation. This arrangement allows there to be more diversity kept in the population, and thus the resulting hypotheses will be more able to explore different sections of the search space. If the search space is narrowed to quickly, then it becomes difficult, and sometimes impossible to find the best results.

### **Crossover and Mutation**

In order to create a new generation of hypotheses there are two types of operations generally used. The first is called Crossover. This operation allows the algorithm to jump around the search space, rather than behave as a simple hill-climbing algorithm. There are three basic types of crossover, single-point, two-point, and uniform.

<b>TABLE 3</b>	Initial String	Crossover Mask	Offspring
Single Point Crossover	<b>11101001000</b>	11111000000	11101010101
	<b>00001010101</b>		00001001000
Two-Point Crossover	<b>11101001000</b>	00111110000	11001011000
	<b>00001010101</b>		00101000101
Uniform Crossover	<b>11101001000</b>	10011010011	10001000100
	<b>00001010101</b>		01101011001

Table three shows how each of the three types of crossover works<sup>7</sup>. An offspring's bits are gathered from the two parent strings. The parent strings are the

<sup>7</sup> Table 3 is an adaptation from Table 9.2 in Langley

parents which are chosen through the selection process discussed in the previous section. Crossover masks are generally randomly generated each time a crossover is needed. In a single point crossover, the mask is created such that it contains k continuous ones at the beginning of the mask and finishes with zeros. The first offspring is constructed by taking the bits of the first parent where there are ones in the crossover mask, and by taking the bits of the second parent where there are zeros. In order to form the second offspring the same pattern is reversed. Two-point crossover substitutes the middle of the strings into each other. In a two-point crossover, the mask begins with k continuous zeros, followed by j continuous ones, and is finished off with zeros. The construction of the uniform crossover mask uses a random string of ones and zeros, which follow no pattern. On average, in large samples, the uniform crossover will uniformly take an equal number of bits from each parent.

To better illustrate the crossover operator, consider a simple model in which home runs, batting average, and hits describe whether a player will be elected to the Hall of Fame:

	HR	AVE.	HITS
Solution #1	10	122	3000
Solution #2	300	310	31

Suppose that these two solutions are present in the population, and through the methods discussed earlier they are selected to be parent strings. Using a single point crossover, the following would occur:

		crossover point ↓							
		HR	AVE.	HITS	→	child #1	HR	AVE.	HITS
Parent	#1	10	122	3000	→	child #1	10	122	31
Parent	#2	300	310	31	→	child #2	300	310	3000

As one can see everything before the crossover point stays the same and the variable after the crossover point, HITS, has its values swapped. This operator is much different than the standard hill-climbing method of moving through a search space. This allows us to make a rather significant jump, as most likely, child #2 is better than either of its parents.

Mutation is a much simpler operator. In a mutation one parent string yields one child. A mutation is simply flipping a single bit in the string. One thing that is important to note is that this does not necessarily mean an increase or decrease of 1 to a value, as the bit is not necessarily the  $2^0$  bit<sup>8</sup>. Using our example above, let us say that parent number 1 is selected for a mutation. Let us further assume that the bit that is selected happens to be in the part of the string that describes the number of home runs hit:

$$\text{HR} = 10_{\text{base } 10} = 001010_{\text{base } 2}$$

If we were to change this string so it read  $101010_{\text{base } 2}$ , the base 10 value of the string would now be 42. Depending on the implementation, more than one mutation can happen in a given string in a given generation. This means that a similar change to the critical value for hits, could be changed in the same generation as the value for HR, by flipping a second bit.

## Literature Review

The following models were used in two separate articles by David Findlay and Clifford Reid. Findlay and Reid had two different dependant variables, with a model for each of them. The first dependant variable was the highest percentage of votes the player

---

<sup>8</sup> Base 2 representation is described in a later section

received. The second, dependant variable, which is more relevant to this paper, is a binary variable that is equal to one if the player was elected to the Hall of Fame and zero if the player wasn't.

The authors of this model claim that election is based on four different groups of variables. The model states that election is based upon a player's offensive capabilities, his defensive capabilities, the awards that he has won, as well as two race variables and a dummy variable to capture which league they played in.

Offensive capabilities can be measured in two different ways. The first way to measure this is through the use of a single performance index. This index is defined as  $(\text{Total bases from hits} + \text{walks} + \text{stolen bases}) / (\text{career at-bats} + \text{walks})$ . The second way uses individual career performance variables to model offensive productivity. In this method, offensive performance is measured with hits, doubles, triples, home runs, walks, stolen bases, and career at bats.

Defensive capabilities are measured using two variables. The first is the number of Gold Gloves received during the player's career. The second is a dummy variable that is equal to one if the player played the majority of his games at shortstop, second base, or catcher. The reason for this variable is that these positions are thought to be more skilled positions than the other positions.

The third category was the awards and post season play. The first variable captured the Rookie of the Year award. This was captured in a dummy variable that captured if the player won that award. The second award variable is a dummy variable that captures whether the player won an MVP award. It is a dummy variable, instead of the number of MVPs won, because of the fact that MVP awards are voted on by the

BBWAA, the same institution that votes on the Hall of Fame. The thinking was that any racial bias that showed up in the Hall of Fame voting would also show up in MVP voting<sup>9</sup>. A second reason was that when number of MVPs was included, some of the offensive capability statistics became insignificant. The last thing in this category is a variable that captures the number of post-season games played.

The last category was the racial category. There was one dummy denoting whether the player was black, and second denoting whether the player was born in a Latin American country. These were included because the authors were looking for significance in these variables. The thought was that these would be significant if there was racial bias in the voting. The last variable was a variable that denoted whether or not a player played most of his career in the American league or his time was split between the two leagues. This was included to capture the effect of the different styles of play, the designated hitter, and other differences between the leagues.

The results from the individual performance regression, predicting the binary election variable, found that all variables except for hits, Black and Latin were significant at at least the 10 percent level of significance. Eventually, the authors show the race variable to be significant by using several different time periods in the regression, but over the entire sample, these variables were not significant.

## **Specific's Of Hall Of Fame Implementation**

### ***Base 2 representation of Base 10 integers***

---

<sup>9</sup> The original purpose of this model was to determine if there was racial inequality in the Hall of Fame voting.

The first thing that has to be done in any GA is to determine a way to represent the question, or more appropriately, the answer to the question as a series of zeros and ones. While this might seem like a daunting task, we have one tool that makes this much easier. In our number system, we represent numbers in base 10. This means that the number 59 is  $5 \cdot 10^1 + 9 \cdot 10^0$ . As you can see in a base 10 number system, all numbers are expressed as the sum of powers of 10. It can be shown, mathematically, that any integer that can be expressed in base 10 can be expressed in base X, where X is any other integer. In addition, in base X, there are always X digits. In base two, there are only two digits, zero and one. In order to count to two then, we could count zero then one, but then would be out of digits. A similar thing happens in base 10 when counting to 10, and nine is reached. In order to move from X-1 to X in base X, X-1 becomes a zero and a one is placed in front of it. In the base 10 example, nine becomes zero, and a one is placed in front, for a final number of 10. In base two, in order to move from one to two, we make the one into a zero and place a one in front. Thus,  $10_{\text{base } 2}$  represents the number two. It can further be shown that the general format for a number in base two is the sum of powers of two. Therefore, in order to represent the number 59 in base two the number would be  $1 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0$ , or 111011. By converting base 10 integers to base two integers, all integers can be expressed as only ones and zeros<sup>10</sup>.

### ***Basic Matching***

Now that we have a method for representing every integer as a string of ones and zeros, we can reduce the question to, “how can one represent our answer as a series of numbers?” For our question, “what does a player need to accomplish during his career to

---

<sup>10</sup> It is much more difficult to express decimals as ones and zeros, as such, all decimals were pre-multiplied by 1000, and expressed as a 3 digit integer.

be voted into the Hall of Fame?” the answer can easily be expressed in integers. In this implementation, for every included variable, there is one integer in the answer that describes that value. Since in baseball, it is rare that more of a given stat, such as home runs or hits, is a bad thing, the integers in the answers will be a lower bound for the value. The few exceptions to this are dealt with explicitly in the fitness function<sup>11</sup>. Consider the simple model from the previous section, in which home runs, batting average, and hits describe whether a player will be elected to the Hall of Fame:

HR	AVE.	HITS
300	382	1500

In this example, the solution would be interpreted as “if the player has over 300 home runs, AND the player had a batting average over 382, AND had more than 1500 hits, then he should be elected to the Hall of Fame.” If any of these conditions are false then the entire statement is false. Consider the same example with different numbers:

HR	AVE.	HITS
800	382	1500

In such a case, nobody would satisfy the condition that they hit more than 800 home runs in a career, as Hank Aaron has the record with 755 career home runs. In this implementation the “don’t care”s that were talked about in the representation section, are any value where the lower boundary falls above the maximum occurring value. Given this, the above solution would be interpreted as, “it doesn’t matter how many home runs

---

<sup>11</sup> It is the case that an increase in errors and strikeouts is not viewed as a good thing. As such, there is an exception which treats the values for strikeouts or errors as upper boundaries instead of lower boundaries. There also is a discussion later about whether or not At Bats and Games are good or bad. This is accomplished during the matching procedure: rather than testing the expression  $(\text{Strikeouts}_{\text{player}} > \text{Strikeouts}_{\text{Rule}})$  we test  $(\text{Strikeouts}_{\text{player}} < \text{Strikeouts}_{\text{Rule}})$ . This is explicitly written into the fitness function as an exception.

the player hit, AND if the player's batting average is over 382, AND he had more than 1500 hits, then he should be elected to the Hall of Fame.”

### ***M of N matching***

Given that there is no actual rule that Hall of Fame voters are required to follow, it does not make sense to require that every single variable match the value in the rule, as described above. A modification, called “m of n matching,” can be made to improve this. A standard matching scheme is one where  $m=n$ , thus every attribute must match. However, by adjusting  $m$  downwards, in order for an example to be classified as positive not all  $n$  attributes have to match the rule. Consider our example from before:

HR	AVE.	HITS
300	382	1500

In this case,  $n$  would equal three, the total number of variables. In an  $m$  of  $n$  matching scheme, where  $m$  is equal to two, any example would be only have to match two of the three prescribed values to be considered a match. For instance,  $HR= 280$ ,  $AVE = 385$  and  $HITS= 1600$  would be a match.

### **Specification of the Model**

The model that I used to predict election into the Hall of Fame is based on a variation of the individual performance model, which essentially says that a player's election potential is based upon several variables representing his performance during his career<sup>12</sup>. Findlay and Reid (1999) attempt to model a players ability to become elected using hits, doubles, triples, home runs, walks, stolen bases, and career at bats, as the variables for offensive productivity. A dummy variable for middle infielders is included,

as well as a variable for the number of Gold Gloves the player won as well as whether or not the player won the rookie of the year award. There is also a variable to capture the effect of the number of postseason games played by the candidate.

While the basic idea that election is based on performance is not changed, several aspects of the model have. First of all, since the earlier papers were written, there has been a vast increase in the availability of data. As such, I was able to include several more variables to measure offensive performance. Additions such as batting average, slugging percentage, on base percentage, runs, and RBIs are just a few examples. In terms of awards, all the various MVP awards(world series, playoff, all star, regular season), as well as manager of the year were added to the previously included gold glove award winners and whether or not the player was the rookie of the year his first year. Whether or not the player won a triple crown was also included<sup>13</sup>. The number of All Star game elections was also introduced as a variable.

I could also improve upon the single post season variable by including post season performance variables, measuring the number of hits, doubles, triples, home runs, total bases, batting average, on base percentage, slugging percentage, as well as games and at bats. All of the post season variables are the totals from all rounds of playoffs. There are also variables specifying the number of games played in each type of playoff series (world series, divisional playoff, league championship).

In order to replace the middle infield dummy variable, 5 dummy variables referring to each position (1B,2B,3B,SS,C) are included. The way these dummy

---

<sup>12</sup> In the same paper, a second method, where total offensive performance is compiled into a 1 value index is discussed. It doesn't make sense to use this representation, as it would take away from the Genetic Algorithm's strengths.

variables should be interpreted as far the rule goes, is if the rule states that one of these is equal to one they are relevant, if they are equal to zero or greater than one then they are not relevant, thus their contributions are not positive or negative.

A list of all of the included variables, along with their contribution to a player's election follows. It is incorrect to call this a sign as one would in the OLS paradigm, because technically the rule generated is not a mathematical equation, rather it is a logical expression. Therefore, to clarify, if more of a variable is a good thing it will be said to be positive, and if more of a thing is bad, it will be negative.

---

<sup>13</sup> The triple crown is an award that is given to a player who leads his league in Batting Average, Home Runs, and RBIs in the same year. It is not an annual award.

**TABLE 4**

YearsPlayed	Games	AtBats	Runs	Hits	totalBases	doubles
+	+/-	+/-	+	+	+	+
Triples	HR	RBI	Sacs	StolenBases	Walks	Hit By Pitch
+	+	+	+	+	+	+
Strikeouts	Errors	FieldingPct	OnBase%	BattingAve	Slugging	NumAllStars
-	-	+	+	+	+	+
ManagerYear	Mvp	NlcsMVP	Rookie	Triple crown	wsMVP	AlcsMVP
+	+	+	+	+	+	+
AsMVP	NumGoldGlove	1Bdummy	2Bdummy	3Bdummy	Cdummy	Ssdummy
+	+	=	=	=	=	=
Psgames	Psab	Psrns	PS-Hits	Psdoubles	PStriples	PShr
+	+	+	+	+	+	+
Psrbi	PsstolenBases	PSbb	Psstrikeouts	PstotalBases	numWS	NumALCS
+	+	+	-	+	+	+
NLCS	NumAEDIV	numAWDIV	NumNEDIV	NumNWDIV		
+	+	+	+	+		
PsbatingAve	Psslugging%	PSOnBase%				
+	+	+	A given variables contribution is listed below the variable name			

In total there are 58 variables that were included in the model. One important thing to note is that in the final rule it is unlikely that all 58 variables will play a part, as the GA will most likely find that some of the variables do not do a good job describing the dependent variable. However, just because several might prove to not be significant in the final rule, does not mean they should not be included in the model. In fact, a GA tends to perform better the more variables it has to work with.

There are several variations of each of the rules to be looked at. First of all a decision has to be made on which fitness function should be used to evaluate the rules. The best way to do this, is obviously with pure theory, so we will start there. It clearly doesn't make sense to use sensitivity or specificity, as both types of evaluation metric

ignore the consequences of either false positives or false negatives. This leaves Odds Ratio and Accuracy as possible choices. By selecting Odds Ratio we are saying that a false negative is more acceptable than a false positive. In terms of Hall of Fame voting, this means we prefer rules that say some people actually in the Hall of Fame don't belong, then rules which say people who aren't in the Hall of Fame actually do belong in. The other option is Accuracy, which treats the two situations mentioned as equal. Rules created using both metrics will be examined later.

Two variables that do need more explanation are the at bats variable, and the games variable. As discussed in their paper on predicting election to the Hall of Fame, Reid and Findlay (1999) point out that career at bats could be a thought of as either a good thing or a bad thing. The same could be said for games. The logic behind this is that some people might see it as a positive that players could compete for a long time. The opposite side would say that given two players with identical statistics, the one who played fewer games was the more dominant player. Because the contribution of the variables are built into the fitness function(whether they are + or -), they have to be specified before the model is built. In an OLS model the hypothesis can be that the variable is not equal to 0, however that is not the case in this model, as such, two separate rules had to be generated, one saying that games and at bats are a good thing, the other saying the opposite<sup>14</sup>. The rules can then be compared afterwards.

The last thing that is important to address is the fact that *a priori* I have no reason to assume a particular value of M in the M of N matching scheme is correct. I believe that M should be less than N, signifying that a voter does not vote exactly according to

the generated rule, but rather votes in accordance with a general opinion of a player which is approximated by the generated rule. This allows the rule to possess some variance, which will naturally occur in the voting for two main reasons. Firstly, humans who have different preferences and judgments from one another, cast the votes. Secondly, election is not based on a unanimous vote, but rather on a 75% requirement, as discussed earlier. These two things would suggest that few players would satisfy the original rule, while several might satisfy a close variant of the rule. How close the variants must be to the original rule is determined by the value of  $M$ . Because there is no predetermined value of  $M$  that is best to use, in order to determine this, several models, each with a different value of  $M$  were generated and compared. More details of this will be given in the Results section.

---

<sup>14</sup> A third option, not examined in this project is the possibility that games/AB have a positive contribution until some critical level after which point the player begins to fall towards the statistics compiler category. This combination approach could be accommodated with a slightly more complex model.

## Results

Earlier it was discussed that there were two possible types of fitness functions that we could use: Accuracy and Odds Ratio. In the following sections, we will discuss each of the two types of fitness functions separately. Within each type of fitness function, there are two cases which we need to look at: one case where games/AB are a good thing, and one case where they are bad things.

Also, for each case, there were experiments run with 10 different values for M(the number of matches required). We will use the first example in the Accuracy section to define a method of picking which rule is best.

### **Accuracy**

In constructing the following rules, the assumption was made that games and AB both positively contribute to a player's election. At the end of an experiment one collects the fitnesses from the best rules found for each value of M, as shown in the table below:

**TABLE 5**

M	Fitness	False Classifications	M	Fitness	False Classifications
<b>57</b>	0.992	6	<b>52</b>	0.996	3
<b>56</b>	0.994	4	<b>51</b>	0.996	3
<b>55</b>	0.996	3	<b>50</b>	0.997	2
<b>54</b>	0.996	3	<b>49</b>	0.997	2
<b>53</b>	0.996	3	<b>48</b>	0.996	3

The value of  $N-M$  is the number of attributes that can not match before an example is counted as not matching the rule. In our data,  $N$ , the number of attributes, is 58, therefore if  $N-M$  is equal two then if a player matches a rule in all but two attributes it is a match, and the player would be predicted to be elected. If however, a player matches the same rule in all but three attributes, it would be considered a miss, and the player would be predicted to not be elected. In Accuracy, the fitness is the percent of

observations correctly explained. And lastly, the False Classifications column describes the number of players who were actually elected to the Hall of Fame, but predicted not to be elected, or vice versa. There is no need to break this column down into two columns of false positives and false negatives, as they are treated equally in this fitness function.

While we do want the rule to be able to describe as many examples as possible, in this case, better explanations come with a price, less understandability. It is very easy to understand the mechanics of a rule when every single attribute of a given observation must match a specified value in order for the observation to be classified as a positive. The understanding becomes much more difficult when rather than having to match every single attribute, any given observation can not match between zero and  $M$  attributes, and yet still the observation would be considered a positive example. Also, at the extreme, if  $M$  equals zero, then  $N-M$  would equal  $N$ , which would mean that regardless of every attribute, the example would be classified as a positive. From this we can say that at some point, increasing  $M$  will hurt the fitness.

It makes sense, that the best rule should compromise both fitness and understandability. As such, the way that I will choose which level of  $M$  to use is to examine the results, and keep allowing  $M$  to increase until an increase of one in  $M$  leads to less than one more example classified. In Table five, we can see that the first decrease of  $M$  leads to an increase in two, the second increment leads an increase of one. The next several increments leads to an increase of zero in the number of examples correctly classified, as such we will use a value of  $M=55$  for the rule.

Using an  $M$  value of 55 a rule is generated. Keep in mind that for an  $M$  of 55,  $M-N$ , or the number of missed attributes allowed, is three. Below is the rule that is

generated, any value that was a “don’t care” was left out of the following table, so only attributes which contribute are included. All others were left out for ease of understanding.

**RULE 1**

YearsPlayed >11	Games >2093	AB >9491	Runs >1199	Hits >1325	Triples >15	HR >18
RBI >1328	StolenBases >6	BB >700	Strikeouts <2199	Errors <374	NumAllStars >10	Mvp >1
BattingAverage >.252	Pshits >10	PSstrikeout >48	PstotalBases >2	Psbating >.187	PSslugging >.081	PSOBP >.093

The surprising thing in this rule is the fact that home runs, triples, and stolen bases are essentially don’t cares. Its is most likely that each of those particular conditions exclude one, or maybe two, players from the Hall of Fame.

Aside from seeing the actual form of the rule, we can use the rule to classify players and see whom it predicts to be elected into the Hall Of Fame. Given that we have an M of 55, it will be interesting to see who is elected at each level of M between 58, and 55. It would seem to make sense that players who match the rule with fewer misses could be considered to be “better” Hall of Fame members than someone who matches with more attributes that don’t match. By using this method, we can see which players are allowed in as constraints on the rule are relaxed<sup>15</sup>.

---

<sup>15</sup> Note that the when we relax and tighten constraints on a rule we keep the same basic rule. This is not the case when we were picking from the rules, earlier in the section. When a rule is built with a different value of M, each attribute can be different, not just the number of misses allowed.

**TABLE 6**

<b>M=58</b>	<b>M=55</b>	<b>M=54</b>
CARL YASTRZEMSKI	TED WILLIAMS	RUSTY STAUB *
FRANK ROBINSON	WILLIE STARGELL	ENOS SLAUGHTER ^
BROOKS ROBINSON	DUKE SNIDER	PEE WEE REECE^
STAN MUSIAL	OZZIE SMITH	DAVE PARKER *
WILLIE MAYS	TONY PEREZ	STEVE GARVEY *
GEORGE BRETT	WILLIE MCCOVEY	NELLIE FOX ^
HANK AARON	EDDIE MATHEWS	DWIGHT EVANS *
<b>M=57</b>	MICKEY MANTLE	ANDRE DAWSON *
NONE	HARMON KILLEBREW	GARY CARTER *
	REGGIE JACKSON	DON BAYLOR *
<b>M=56</b>	CARLTON FISK	ANDRE DAWSON *
ROBIN YOUNT	ROD CAREW	GARY CARTER *
DAVE WINFIELD	LOU BROCK	DON BAYLOR *
MIKE SCHMIDT	YOGI BERRA	<i>All players with ^ were elected by the veterans committee.</i>
AL KALINE	ERNIE BANKS	<i>All players with * following there names were never</i>
JOHNNY BENCH	LUIS APARICIO	<i>elected to the Hall of Fame.</i>

What we can see from Table six is the fact that, according to this rule, people who were elected by the Veteran's Committee don't deserve to be in the Hall of Fame, according to the general voting pattern of the writers. This makes some sense, as the only reason a player makes it to be in front of the Veteran's Committee is after he wasn't elected by BBWAA. A second observation is that The first group of players, the M=58 players, can be said to be a cut above the rest of the players as it takes two relaxed constraints before they are let in.

The next situation to look at is if games and AB are bad things. As before, the first thing that is necessary to do is to pick a level of M that is best for our fitness function. To do so we will look at a table similar to the one we examined before.

**TABLE 7**

M	Fitness	False Classifications	M	Fitness	False Classifications
<b>57</b>	0.996	3	<b>52</b>	0.996	3
<b>56</b>	0.994	4	<b>51</b>	0.996	3
<b>55</b>	0.996	3	<b>50</b>	0.996	3
<b>54</b>	0.996	3	<b>49</b>	0.996	3
<b>53</b>	0.996	3	<b>48</b>	0.997	2

Table seven shows one of the interesting things about using M of N matching. It is not necessarily the case that such as smooth improvement, as we saw in the Table five, will always happen. Due to the random nature of the starting point of GAs, they don't always behave as we would expect. The question we are looking to answer is which value of M is best. From the table above, we see that an M of 57, does equally as well as an M of 55,54,53 and so on. According to the rule laid out in the last example, we should choose an M of 57 because its value of M is closest to the value of N. The other thing that would be lost is an ability to strictly compare the rule, and more importantly, which players would be elected at each level of M. This comparison will help us interpret the difference in the rules. Because we used an M of 55 in the other case, and the difference in understandability is minimal between an M of 55 and 57, a slight compromise with rule will be made, and the rule for M of 55 will be used<sup>16</sup>. As before, in the following rule, only the attributes found not to be “don't cares” were included.

**RULE 2**

Games <3558	AB <12174	Runs >1206	TotalBases >3224	Doubles >149	Triples >47	HR >26
RBI >176	StolenBases >13	BB >688	HBP >3	Strikeouts <2518	Errors <397	NumAllStar >9
Mvp >1	Psab >23	Psrns >1	PShr >14	Pstrikeouts <63	Psbating >0.185	Psslugging >0.171
PSOBP >0.254						

In the above rule, it can be noted, that despite the fact that regular season games and AB are not attributes that make positive contributions to a player's election, post season at bats are still thought of as a good thing. This is because of the fact that there is very little talk of talk of players sticking around to “compile” post season statistics. This

---

<sup>16</sup> As can be seen there is some subjectivity in choosing the rule. However, as long as justification for deviating from a prescribed method can shown, then it presents less of a problem. Furthermore, given that the fitnesses of the 2 rules are the same, I find this choice even less problematic.

is partly due to the fact that a high level of play is necessary to reach the playoffs, and get at bats.

The last thing to look at in the Accuracy paradigm is the players that model predicts should be elected, at various levels of strictness.

**TABLE 8**

<b>M=58</b>	<b>M=55</b>	<b>M=54</b>
NONE	ROBIN YOUNT	LOU WHITAKER *
	DAVE WINFIELD	ALAN TRAMMELL *
<b>M=57</b>	WILLIE STARGELL	JOE TORRE *
MIKE SCHMIDT	DUKE SNIDER	JIM RICE *
STAN MUSIAL	OZZIE SMITH	WEE PEE ^
WILLIE MAYS	ENOS SLAUGHTER ^	DAVE PARKER *
MICKEY MANTLE	RED SCHOENDIENST ^	FRED LYNN *
HANK AARON	TONY PEREZ	GIL HODGES *
	WILLIE MCCOVEY	KEITH HERNANDEZ *
<b>M=56</b>	HARMON KILLEBREW	STEVE GARVEY *
CARL YASTRZEMSKI	AL KALINE	GEORGE FOSTER *
TED WILLIAMS	REGGIE JACKSON	DWIGHT EVANS *
FRANK ROBINSON	CARLTON FISK	ANDRE DAWSON *
BROOKS ROBINSON	LOU BROCK	JOSE CRUZ *
JOE MORGAN	YOGI BERRA	DAVE CONCEPCION *
EDDIE MATHEWS	JOHNNY BENCH	GARY CARTER *
NELLIE FOX	ERNIE BANKS	KEN BOYER *
ROD CAREW	LUIS APARICIO	DON BAYLOR *
GEORGE BRETT		

\*s represent players not elected by the BBWAA while ^s mean the player was elected by the veterans committee.

We can see from table eight, above, that this rule, where games are a bad thing, behaves differently then the one where games are a good thing. First of all, there is no player that matches the rule exactly. There also isn't the jump from one level of players to another level of players as could be seen in the first example. This could be interpreted as meaning that there was more of a continuous talent gradient. On the surface, it can be seen that one major difference between the rules is the fact that in this rule, two players not elected by BBWAA are classified as players that should have been elected.

### **Odds Ratio**

Earlier in the paper, it was discussed that there were two different paradigms that would be looked at as tools for evaluating the fitness of each member of the population. This method for evaluating fitness minimizes the number of false positives. As before, there are two possible classifications to look at, the case where games are good, and the case where they are bad. The first case that will be examined is the one where more games have a positive effect on a player’s election to the Hall of Fame. When examining this situation, the first thing that we do is collect the fitness data from 10 experiments, varying the levels of M, in order to pick which level to use for the rule.

**TABLE 9**

M	Fitness	False Classifications	M	Fitness	False Classifications
<b>57</b>	4193	7	<b>52</b>	17629	2
<b>56</b>	9151	4	<b>51</b>	17629	2
<b>55</b>	9151	4	<b>50</b>	30345	1
<b>54</b>	12179	3	<b>49</b>	30345	1
<b>53</b>	17629	2	<b>48</b>	30345	1

In the above chart the fitness is the odds of a positively classified example being positive, against the odds that the example is truly a negative. The same tradeoff present in the Accuracy fitness function, trading classification for ease of understanding. As we relax the number of variable matches required, we continue to have an average gain of at least one gained example classified, for each additional allowed miss, until we hit a value of 53. Therefore, we will use the value of 53 for M, meaning that a given rule can miss five attributes before it is considered to have not matched the rule.

### **RULE 3**

Games >2794	AB >3967	Runs >1145	Hits >1297	TotalBases >4432	Doubles >252	Triples >14
HR >355	RBI >325	StolenBases >341	BB >658	HBP >9	Strikeouts <2087	Errors <399
NumAllStars >9	Psbb >1	Psstrikeouts <42	Psslugging >0.269	PSOBP >0.010		

As before, only the variables that were not determined to be “don’t care” are included in the preceding rule. This rule seems to agree with the thinking of most fans. Offensive performance is important, as is participation in the post season. The rule is more concise than either of the Accuracy rules, and variables that the typical fan argues over, seem to drive the model. Home runs, number of all star games elected to, and total bases all seem to be rather large and difficult to obtain for the average player. The player needs to have a good knowledge of the strike zone, as can be seen from the number of walks required.

As before, we can apply this rule, with different values of M to gain insight into how this rule ranks the Hall of Fame eligible players.

**TABLE 10**

<b>M=58,57</b>	<b>M=53</b>	<b>M=52</b>
NONE	ROBIN YOUNT	RUSTY STAUB *
	TED WILLIAMS	ENOS SLAUGHTER ^
<b>M=56</b>	BILLY WILLIAMS	RED SCHOENDIENST ^
STAN MUSIAL	WILLIE STARGELL	JIM RICE *
WILLIE MAYS	DUKE SNIDER	PEE WEE REECE ^
HANK AARON	BROOKS ROBINSON	GRAIG NETTLES *
	WILLIE MCCOVEY	DALE MURPHY *
<b>M=55</b>	MICKEY MANTLE	GIL HODGES *
CARL YASTRZEMSKI	HARMON KILLEBREW	JIM GILLIAM *
DAVE WINFIELD	REGGIE JACKSON	NELLIE FOX
FRANK ROBINSON	CARLTON FISK	DWIGHT EVANS *
AL KALINE	ROD CAREW	ANDRE DAWSON *
	LOU BROCK	DAVE CONCEPCION *
<b>M=54</b>	GEORGE BRETT	GARY CARTER *
OZZIE SMITH	YOGI BERRA	
MIKE SCHMIDT	JOHNNY BENCH	
TONY PEREZ	LUIS APARICIO	
JOE MORGAN		
EDDIE MATHEWS		
ERNIE BANKS		

All players marked with a \* were not elected to the Hall of Fame. Players marked with a ^ were elected by the Veteran’s Committee. All unmarked players were elected by the Hall of Fame.

This rule is stricter than any of the rules created by the Accuracy metric. This can be seen through the fact that no players match the rule exactly or with one miss. It can

also be seen that following that, there is a continuous performance gradient, and no jumps as seen in Rule two. In the end, every player with an M of 53 or higher, is a player who the rule predicts should be elected to the Hall of Fame.

The next thing in the Odds Ratio paradigm is to look at the case where games and at bats are bad things. Here too, 10 experiments with varying levels of M are run and fitnesses are collected.

**TABLE 11**

M	Fitness	False Classifications	M	Fitness	False Classifications
<b>57</b>	5189	6	<b>52</b>	17629	2
<b>56</b>	7225	5	<b>51</b>	17629	2
<b>55</b>	9151	4	<b>50</b>	30345	1
<b>54</b>	12179	3	<b>49</b>	30345	1
<b>53</b>	17629	2	<b>48</b>	30345	1

It can be seen in the above table that there is a consistent gain of more classified example for each additional allowed miss until M=53. This is the rule that will be used. It is also worthwhile to note that this rule finds the same level of M as the first rule that used Odds Ratio, and the assumption that games are good.

**RULE 4**

Games <3343	AB <12642	Runs >1219	Hits >982	TotalBases >145	Doubles >391	Triples >17
HR >67	RBI >1505	Sacs >5	StolenBases >333	BB >418	Strikeouts <2258	Errors >372
NumAllStars >11	BattingAverage >.179	Slugging >.475	OnBase% >.211	Psab >14	Psstrikeouts <18	NumWS >2
Psbating >.188	Psslugging >.320	PSOBP >.228				

This rule, where games and at bats are bad things, has considerably more components that are not don't cares than Rule 3, the Odds Ratio rule, where games are good. The components that the typical fan would expect to be important, like HR and total bases are miniscule numbers. Both cutoffs have been surpassed by individual

players in a single season. This model seems to be driven by run, doubles, all stars, as well as the hitting performance variables<sup>17</sup>.

The next thing to look at is what happens when this rule is applied to the players in the sample.

**TABLE 12**

M=58	M=54	M=53
WILLIE MAYS	CARL YASTRZEMSKI	ROBIN YOUNT
	TED WILLIAMS	DUKE SNIDER
M=57	WILLIE STARGELL	KIRBY PUCKETT
STAN MUSIAL	OZZIE SMITH	JOE MORGAN
	ENOS SLAUGHTER <sup>^</sup>	HARMON KILLEBREW
M=56	RED SCHOENDIENST <sup>^</sup>	ORLANDO CEPEDA
DAVE WINFIELD	BROOKS ROBINSON	ROD CAREW
MIKE SCHMIDT	TONY PEREZ	LOU BROCK
FRANK ROBINSON	EDDIE MATHEWS	YOGI BERRA
	MICKEY MANTLE	JOHNNY BENCH
M=55	REGGIE JACKSON	LUIS APARICIO
AL KALINE	CARLTON FISK	
LOU BROCK		M=52
GEORGE BRETT		<i>Because of the size of this group</i>
ERNIE BANKS		<i>It is shown in appendix A instead of In this chart</i>

*Players marked with a ^ were elected by the Veterans Committee. No marking means they were elected.*

Once again, we can see a smooth performance gradient throughout the entire range of the rule. As before, all players in the M=53 group, or higher are the people that deserved to be elected to the Hall of Fame.

## Predictions

One last thing that we would hope that all of our models would be able to do is to predict which players would be elected to the Hall of Fame, that haven't been voted on. This is a question that often comes up in newspaper columns, and sports television. Sports fans are always talking about which players are going to get elected to the Hall of Fame when they retire, and which players won't. Below is a table of several players

<sup>17</sup> Hitting performance variables are being defined as batting average, on base percentage, slugging percentage, and their post season counterparts.

who have either, retired since the sample in the paper ended, or are still actively playing the major leagues. For those players still in the major leagues, there is no projecting done. The forecasts are based on what would happen if the player retired with only the statistics compiled up to the end of the 2001 season.

**TABLE 13**

	Accuracy Games	Accuracy	Odds Ratio Games	Odds Ratio
ROBERTO ALOMAR	YES	NO	YES	YES
HAROLD BAINES	NO	YES	YES	YES
WADE BOGGS	YES	NO	YES	NO
BARRY BONDS	YES	YES	YES	YES
JOE CARTER	NO	NO	NO	NO
WILL CLARK	NO	NO	YES	NO
JOSE CANSECO	NO	NO	NO	NO
MARK GRACE	NO	NO	NO	NO
KEN GRIFFEY	YES	YES	YES	YES
TONY GWYNN	YES	YES	YES	YES
RICKEY HENDERSON	YES	YES	YES	YES
BARRY LARKIN	YES	YES	YES	YES
TINO MARTINEZ	NO	NO	NO	NO
FRED MCGRIFF	NO	NO	NO	NO
MARK MCGWIRE	NO	NO	NO	NO
PAUL MOLITOR	NO	NO	YES	YES
EDDIE MURRAY	NO	YES	YES	YES
RAFAEL PALMIERO	NO	NO	NO	NO
MIKE PIAZZA	NO	NO	NO	NO
TIM RAINES	NO	NO	YES	NO
CAL RIPKEN	YES	YES	YES	YES
IVAN RODRIGUEZ	NO	NO	NO	NO
RYNE SANDBERG	YES	YES	YES	YES
SAMMY SOSA	NO	NO	NO	NO

*In the above table, Accuracy Games is the Accuracy rule where games are bad. Odds Ratio Games is the Odds Ratio rule where games are bad.*

In the above table, one should notice the fact that in most all cases, the all of the different rules do agree on a player's potential of getting into the Hall of Fame. One thing that is also apparent is that arguably the two greatest catchers of the last decade, Mike Piazza and Ivan Rodriguez, are not listed as being elected into the Hall of Fame. Sportswriters would claim that both will eventually be elected. What could cause this disparity? One possible explanation is the fact that these two are catchers, and as such, are generally elected with fewer pure statistics than other positions. It is striking that

even shortstops, who generally have worse numbers than other positions players are classified as being elected, but those two catchers are not. My guess is that in a few more years, they both might be better qualified for the Hall of Fame, and that as of now, they are not. This having been said, I do not believe that this can be linked to the fact that positions are not differentiated in the rules, since the “weaker” producing shortstops are classified as players that should be elected<sup>18</sup>.

## Interpretation

### *Odds Ratio vs Accuracy*

The reason that both Odds Ratio and Accuracy rules were looked at was because there was no way to know which would produce a better rule. After looking at the result in the previous section, we are now in a better position to make a judgement about which fitness function appears to work best for our purposes. There are several ways we can judge between the two sets of rules.

When trying to decide which value of M to use, to create the rule, tables were used to determine which value had the fewest misclassifications. A similar approach seems to make sense here.

**TABLE 14**

	Games Good	Games Bad
Accuracy	3	3
Odds Ratio	2	2

Table 14 shows the number of misclassifications in each of the four rules at the optimum value of M. The number of misclassifications in each rule is a way to measure

---

<sup>18</sup> Findlay and Reid(1999) use a middle infield variable to capture this effect and find that it is significant at the 1 percent level in predicting election to the Hall of Fame using a Logit and Probit model.

the in-sample accuracy of the rules<sup>19</sup>. In standard econometrics, there are several methods for measuring the in-sample accuracy, the most commonly used being the  $R^2$  statistic. Generally, it is the case in econometrics, that the  $R^2$  statistic should not be compared between different equations. This is because it is not a standardized measurement, and thus will change, as the dependent variable changes. While this is a problem in standard econometrics, the idea of comparing a statistic like the  $R^2$  doesn't present as much of a problem for us. The reason for this is the fact that in all the models we are using have the same sample, with the same dependant variable. By comparing the number of misclassifications, it can be seen that in both cases, the number of misclassifications is fewer in the Odds Ratio paradigm.

This realization might make it seem that odds ratio is the clear choice for which paradigm is the better of the two. However, before making this choice there is something else to consider. When choosing the rules, it was determined that while the accuracy was important, the understandability was important as well. This means before it is decided that odds ratio, rather than accuracy, is the fitness of choice, we need to make that the resulting rule from odds ratio is at least as understandable as the one from Accuracy. The simplest measure for understandability, as used earlier, was to use the value of  $M$ . As described in the results section, the closer  $M$ , the number of required matches, is to  $N$ , the number of total variables, the easier it is for the rule to be understood. At this simple level, the Accuracy fitness function, with an  $M$  of 55, is better than the Odds Ratio fitness function with an  $M$  of 53. However, the difference is so slight that there is little problem with saying that Odds Ratio is superior.

---

<sup>19</sup> Accuracy, in this case, refers to the correctness of the rule, not the metric for determining fitness.

There is one other thing about Odds Ratio that makes sense. Consider exactly what minimizes false positives mean. If we have a borderline case, Odds Ratio will classify the example as false, while, others might classify it as true. This makes sense in our situation since a player has to prove, through his statistics and career that he deserves to be. Minimizing false positives provides this type of decision preference.

### ***Treat Games/AB as Good or Bad***

The other decision to make is whether games should be treated as a good thing or a bad thing. In order to do this we will again look at how understandable the rules are. The typical fan looks at very few statistics when they make their arguments about who should make the Hall of Fame. The most recognizable records in baseball are probably the homerun records. The Odds Ratio rule, where games are good, rule number three, has a high requirement for homeruns. Fans are enamored by offensive statistics, and rule number three shows this, as both total bases, and runs are difficult plateaus to reach.

Much of fan preferences and their opinions is also shown in this rule. The requirement that the player be elected nine times to the All Star game captures two separate things. First of all, fans have always elected the all stars. Therefore, the number of All Star games played in captures a fan sentiment towards the player in question. A second thing that All Star games captures is a relative ranking amongst players of the time. There are not a limited number of home runs in a season, nor are there a limited number of hits, runs, or any other pure statistic. There are a limited number of spots on an All Star team roster, however. As such, election means that a player is one of the top players in the league, at the time.

From Table 10 in the Results section, we can see that no players match the rule number three exactly, or even with one allowed missed match. Does this make the rule any less efficient? No, because it makes sense that there are several variations of a good player. No two players are known for being excellent at exactly the same things. This rule shows this, and as such, even the best players ever, have some places where they don't satisfy the requirements set forth by the rule.

The question still remains, why should games and at-bats be positive attributes? One answer is that longevity comes from being able to perform at a high level for a long time. Secondly, we know that people didn't that Lou Gerig hung around too long and should of retired sooner. In fact, what most people know him for was his longevity.

A second answer is that the correct representation wasn't looked at. Most likely, there is some level, after which a player is seen as being a compiler. However, until that point is reached, more games are good things. This does not mean that rule number 3 is not the best option of the two possibilities to pick from, as it is most likely that the level of games and at-bats where players become compilers is rather high. Consider rule number four where the number of games that players must be under is so high that only 2 players in the sample are above that level<sup>20</sup>.

Another very important question that needs to be addressed is how the rule takes a player's position into account. As mentioned earlier, the middle infielders often have much worse power numbers, and so one would expect that the middle infielders would be misclassified, and predicted to not get elected. The number of All Star games, however, shows a player's relative ability to his peers. Since the number of All Star games

---

<sup>20</sup> Only Hank Aaron and Carl Yastrzemski surpassed the boundaries of 12,642 at-bats, and 3,343 games. Pete Rose also surpasses these marks but was not included in the sample for reasons discussed earlier.

required is high, this says that a player has to be the best in his position for a long period of time. This is how position is taken into account.

## **Problems and Alternatives**

The most noticeable attribute of the model is the way that it matches a player to the given rule. For instance, in rule three the required number of home runs is 355. The way this rule is interpreted means that 354 HR is the same as 254 HR is the same as zero HR. On the other side, 355 HR is the same as 500 HR and so on. The question present, is this a problem? To answer this, we can look at a possible solution along with its implications.

The way to modify the matching algorithm lies in the way that the M of N matching works. As it stands, during matching, each player is allowed a fixed number of missed attributes. At the same time, each variable is turned into a binary variable of sorts. For home runs, this binary variable would be a one if the player hit more than 355 home runs, and zero if the player hit fewer. This throws away a lot of information that could possibly be used for matching.

One possible way to modify this matching algorithm is to change the rule so that if a player surpasses a given rule, for instance, home runs, by some scalar multiple,  $k$ , of the standard deviation, then the player would be allowed additional missed matches<sup>21</sup>. In the same sense, when a player misses a required rule, like the 355 home runs, then if that miss was by more than  $k$  standard deviations of HR then the player would lose multiple misses.

---

<sup>21</sup> The actual number of additional misses that would be allowed would depend on  $k$ , the scalar multiple for the standard deviation. If the difference between the rule and the players attributes, is equal to  $2 * (k * \text{stdDev})$ , then the player would be allowed 2 additional misses. The way to determine the number of

There is a problem with this approach. One of the biggest gains from the GA rules is the understandability. Consider the fact that in order to interpret a rule, it would be necessary to first determine which players, all 755 of them, were eligible for the Hall of Fame. Next it would be necessary to calculate 58 different standard deviations. Lastly, when matching a given player several differences would have to be calculated, and divisions would have to be done. Clearly, this is much more complex. And while it might be more accurate, the rule is nearly totally void of intuitive understanding. This is a huge shortfall.

A second thing to be very aware of is the time constraint on running GA experiments. A typical experiment in this paper required over 580 million comparisons, and possibly another several hundred million operations are required. These numbers are only in the fitness function. There is also a huge number of operations used to keep track of all the members of the population. All of this translates to a process that takes several hours to run. In this paper there were 10 values of M that were run for each possible rule. For each value of M, it took just over five hours to find its respective rule.<sup>22</sup>. This means that significant planning should go into setting up and researching a GA before one starts to run the experiments.

## Conclusion

The goal of this paper was to develop a model for election to the Baseball Hall of Fame, that is easier to understand than simple regression analysis. After deciding that the best way to proceed was through the usage of Genetic Algorithms, the next thing to

---

additional misses is to divide the difference by  $(k * \text{stdDev})$ . If k is 1 then this would just be the difference divided by the stdDev.

<sup>22</sup>A single set of experiments took about 5 hours to run on a 2 GHz Intel Pentium 4 computer, running the Red Hat Linux Operating System, with 1 GB of memory.

accomplish was to represent our solution in only 1s and 0s. This was accomplished by converting integers to base 2. The representation of our model became a collection of required statistics.

Because of the fact that there were two possible fitness functions and 2 possible interpretations for how at-bats and games affected election it was necessary to create 4 different rules covering all the possible situations, and compare them to determine which one was best.

After looking at various rules, and weighing the explanatory value, the ease of understanding, and the how much the rule agrees with typical baseball logic, it was decided that the odds ratio rule where games are a good thing is the best rule.

The rule does a good job at describing the data set, correctly identifying the status of all but two players in the data set. The rule even captures middle infielders, players that require special consideration in other models. The results reported in this paper seem to accomplish the job of explaining Hall of Fame election in a more easily understandable method than standard regression analysis.

## Appendix A

	M= 52	
JIM WYNN	FRED LYNN	LARRY DOBY ^
LOU WHITAKER	DAVEY LOPES	ANDRE DAWSON
ALAN TRAMMELL	ED KRANEPOOL	ALVIN DARK
EARL TORGESON	JAY JOHNSTONE	JOSE CRUZ
RUSTY STAUB	GIL HODGES	DEL CRANDALL
LONNIE SMITH	KEITH HERNANDEZ	DAVE CONCEPCION
JIM RICE	KEN GRIFFEY SR.	JACK CLARK
PEE WEE REECE	STEVE GARVEY	CHRIS CHAMBLISS
DAVE PARKER	CARL FURILLO	ORLANDO CEPEDA ^
AMOS OTIS	GEORGE FOSTER	NORM CASH
AL OLIVER	RON FAIRLY	GARY CARTER
TIM MCCARVER	DWIGHT EVANS	ROY CAMPANELLA *
LEE MAY	DAN DRIESSEN	DUSTY BAKER

All Players with out marks were not elected. ^ indicates elections by the Veteran's Committee. \* indicates election by the BBWAA.

## Bibliography

1. Langley, Pat. Elements of Machine Learning. San Francisco, CA. Morgan Kaufmann, 1996.
  2. Findlay, David W; Reid, Clifford E. A Comparison of Two Voting Models to Forecast Election into the National Baseball Hall of Fame.. [Journal Article] Managerial & Decision Economics. Vol. 23 (3). p 99-113. April-May 2002.
  3. Findlay, David W; Reid, Clifford E. Voting Behavior, Discrimination and the National Baseball Hall of Fame.. [Journal Article] Economic Inquiry. Vol. 35 (3). p 562-78. July 1997.
- Congdon – PhD.
5. Congdon, Clare Bates, Greenfest, Emily F. Genetic Algorithms in Cladistics, Proceedings: Grace Hopper Celebration of Women in Computing. Cape Cod, MA. September 2000