

# AggPro: The Aggregate Projection System

Ross J. Gore, Member, *The Society For American Baseball Research (SABR)* and Cameron T. Snapp

**Abstract**— Currently there exist many different systems to predict the performance of Major League Baseball (MLB) players in a variety of statistical categories. We propose, AggPro, an aggregate projection system that forms a projection for a MLB player’s performance by weighting the player’s projections from these projection systems. Using automated search methods each projection system is assigned a weight. The determined weight for a system is then applied to all the projections from that system. Then, an AggPro projection is formed by summing the different weighted projections for a player across all the projection systems. The AggPro projections are more accurate than the constituent projection systems when evaluated by average error, root mean square error (RMSE) and Pearson correlation coefficient from actual player performance for the 2006, 2007, 2008 and 2009 MLB seasons.

## I. INTRODUCTION

Many different methods for projecting the performance of Major League Baseball (MLB) players in a variety of statistical categories for an upcoming MLB season exist. These projection systems include: Brad Null [1], Bill James Handbook [2], CAIRO [3], CBS [4], CHONE [5], ESPN [6], Hardball Times [7], Hit Tracker [8], KFFL [9], Marcel [10], Oliver [11], PECOTA [12], RotoWorld [13], and ZiPS [14]. Despite the availability and prevalence of these systems there has been relatively little research on the evaluation of these systems’ accuracy. Furthermore, there has been no research that attempts to compose these projection systems to create a single more accurate projection.

We propose, AggPro, an aggregate projection system that forms a projection for a MLB player’s performance by weighting the player’s projections from these other projection systems. Using automated search methods each projection system is assigned a weight. The determined weight for a system is then applied to all the projections from that system. Then, an AggPro projection is formed by summing the different weighted projections for a player across all the projection systems.

We believe the aggregate projections contain the best parts of each projection system resulting in a system that is more accurate than any of the constituent systems in the AggPro projection. The AggPro projections are evaluated against all

the constituent system by measuring the average error, root mean square error (RMSE) and Pearson correlation coefficient of the projections from actual player performance for the 2006, 2007, 2008 and 2009 MLB seasons.

It is important to note that AggPro is not just another projection system. Instead it is a methodology for aggregating effective projections from different systems into a single more accurate projection. Furthermore, Greg Rybarczyk [8] believes paradigm shifts that will improve the accuracy of projection systems are on the horizon. If paradigm shifting projection systems are developed, the AggPro methodology will be applicable and improve the projections from these systems as well.

In the next section we describe work related to AggPro. Then, AggPro is presented and evaluated. Finally we conclude the paper and present directions for future work with AggPro.

## II. RELATED WORK

Research efforts in the areas of baseball, computer science, and artificial intelligence have all contributed to AggPro. We review these related works here.

### A. BellKor and The Netflix Prize

The strategy of applying different weights to different predictions from effective projection systems has been used successfully by the winning solution for the Netflix prize [15], BellKor by AT&T labs [16]. In October, 2006 Netflix released a dataset of anonymous movie ratings and challenged researchers to develop systems that could beat the accuracy of its recommendation system, Cinematch. A grand prize, known as the Netflix Prize, of \$1,000,000 was awarded to the first system to beat Cinematch by 10%. The BellKor prediction system was part of the winning solution, with 10.05% improvement over Cinematch.

BellKor employs 107 different models of varying approaches to generate user ratings for a particular movie. Then BellKor weights each model’s prediction to create an aggregate prediction for the movie [16]. AggPro applies this prediction strategy to projecting the performance for MLB players by employing the different existing MLB projection systems.

### B. Nate Silver’s 2007 Evaluation of Projection Systems

In 2007 Nate Silver performed a quick and dirty evaluation of the on-base percentage (OPS) statistic projection from eight 2007 MLB projection systems [17]. Silver’s work offers several evaluation metrics including average error, RMSE and Pearson’s correlation coefficient, which we employ to evaluate

Ross. J. Gore is with the Department of Computer Science at The University of Virginia, Charlottesville, VA 22901 USA (phone: 703-887-8060; e-mail: rjg7v@virginia.edu).

Cameron T. Snapp is with CapTech Ventures, Inc. Richmond, VA 23220 USA. (phone: 804—355-0511; e-mail: cameron.snapp@gmail.com).

AggPro. However, Silver also offers a metric to determine which system provides the best information. The metric is based on performing a regression analysis on all the systems for the past year and identifying "which systems contribute the most to the projection bundle [17]." AggPro performs this same regression analysis using the projections of systems for the past several years. Then AggPro applies each metric identified by the analysis as a weight to system's upcoming projections for the year. This methodology identifies most accurate parts of each projection system and combines these parts in one aggregate projection produced by AggPro.

### III. AGGPRO

The AggPro projections were generated through a three part process. First, we collected projections from five different systems for the years 2006-2009. Next, for each year we identified the players that were common among all five systems and the MLB actual data. We also identified the statistical categories that were common among all five projection systems and MLB actual data. Finally, we performed an automated search over all the combinations of possible weights to apply to the five systems. The automated search identified the weight combination that minimized the root mean squared error (RMSE) of the aggregate projection from the actual player performance for 2007 and 2008. We review each part of this process in this section.

#### A. Projection and MLB Actual Data Collection

We collected projections from Bill James Handbook [2], CHONE [5], Marcel [10], PECOTA [12] and ZiPS [14] for the years 2006-2009. We collected the actual MLB performance data for 2006-2009 from Baseball Prospectus. The 2009 data is dynamic due to the season being in progress at the time our work. Also, the CHONE projection system did not have formalized projections for 2006. These projection systems are a representative sample of the many different systems that exist. If AggPro can successfully create an aggregate projection from these systems that is more accurate than any of the individual constituent projection systems then the AggPro methodology will have been shown to be successful. Given a successful methodology the reader can apply AggPro to any projection systems s/he chooses.

#### B. Identification of Players and Statistics to Project

Recall that each year AggPro only projects the performance of those players common to all five systems and the MLB actual player performance data. The player list for each year is available at [18]. Also recall that AggPro can only project those statistical categories that are common to all five systems and the MLB actual player performance data. The hitter categories are: At Bats, Hits, Runs, Doubles, Triples, Home Runs, RBIs, Stolen Bases, Walks, and Strikeouts. The pitcher categories are: Innings Pitched, Earned Runs, Strikeouts, Walks, and Hits. These sets of players and statistics represent the largest possible set that was common to all the systems and the MLB actual performance data.

#### C. Automated Search To Identify AggPro Weights

Given the five projection systems, identified common players and statistical categories we performed a brute force automated search over all the possible combinations of weights for the five projection systems. The search identified the weights that minimized the RMSE of the projections for 2007 and 2008 from the 2007 and 2008 MLB actual performance data. We could not program the search to identify the weights that minimized the RMSE of the projections for 2006 due to the lack of CHONE projections for the year.

The set of weights which minimized the RMSE from the 2007 and 2008 MLB actual performance was: Bill James Handbook = 0.56, CHONE = 0.00, Marcel = 0.21, PECOTA = 0.23, and ZiPS = 0.00. This is approximately equivalent to a projection cocktail consisting of 2 parts Bill James, 1 part Marcel and 1 part PECOTA. Applying these weights to the projection systems for 2006, 2007, 2008 and 2009 generates AggPro projections for each year. In the next section we evaluate the accuracy of the AggPro projections for each year using average error, RMSE and Pearson's correlation coefficient as evaluation criteria.

### IV. EVALUATION

AggPro and the five constituent projection systems were evaluated by computing the average error, RMSE, and Pearson correlation coefficient for each year for each statistical category from the MLB actual data. All of this evaluation data is shown and discussed in the Appendix.

For each system, for each year we also computed the average of each evaluation criterion over all the statistical categories. Each year we identified the best constituent projection system (BCPS). The BCPS is the constituent system for a given year which had the best average evaluation criterion over all the statistical categories. Furthermore, we identified the best constituent projection in each statistical category for each evaluation criterion. Combining these best constituent projections of each category forms the theoretical projection system (TPS). Due to how it is constructed the TPS is guaranteed to be at least as accurate as the BCPS. We also computed the average of each evaluation criterion over all the statistical categories in the TPS. AggPro's percent improvement over the BCPS and the TPS for the average of each evaluation criterion for each year is shown in Table 1-3. The 2009 projections are evaluated through MLB games completed on July 27<sup>th</sup>, 2009.

**Average Error**

Year	Percent Improvement over BCPS (Bill James)	Percent Improvement over TPS
2006	3.3	2.5
2007	4.3	3.5
2008	6.0	3.0
2009	0.8	0.7

**Table 1: The average error evaluation of AggPro.**

## RMSE

Year	Percent Improvement over BCPS (Bill James)	Percent Improvement over TPS
2006	5.2	5.0
2007	4.3	2.7
2008	5.9	4.6
2009	2.2	1.9

Table 2: The RMSE evaluation of AggPro.

## Pearson Correlation Coefficient

Year	Percent Improvement over BCPS (Bill James)	Percent Improvement over TPS
2006	1.8	1.6
2007	1.8	1.6
2008	2.2	2.2
2009	1.0	0.9

Table 3: The Pearson correlation coefficient evaluation of AggPro.

AggPro is an improvement over both the BCPS and TPS for each evaluation criterion. This result is surprising. Since the TPS is constructed to contain the best constituent projection for each statistical category we did not anticipate that AggPro would outperform it. Instead, we had anticipated that the TPS would be a baseline for the best theoretical improvement AggPro could achieve. However, it appears the weighting of the different projections creates an aggregate projection that is more than the sum of the best parts of the constituent projection systems. This bodes well for future work with the AggPro methodology.

## V. CONCLUSION

There exist many different systems to predict the performance of Major League Baseball (MLB) players in a variety of statistical categories. We have shown that our methodology, AggPro, can aggregate these existing projection systems into a single aggregate projection that is more accurate than any of AggPro's constituent project systems. Furthermore, AggPro is more accurate than the TPS when measured by any of the three evaluation criteria for the years 2006-2009. In other words, even if a reader was given a fictional oracle function at the beginning of the season which could pick the most accurate projection from the five systems for each statistical category, AggPro's predictions would still be more accurate for the upcoming season.

In future work with AggPro we will explore using distinct weight sets for the constituent projection systems for hitting statistic categories and pitching statistic categories. We will also explore how often the AggPro weight sets need to be recalibrated to maintain the level of accuracy we have achieved in this work.

## APPENDIX

The evaluation of each system for each statistical category, for each evaluation criterion is listed in the following tables.

AggPro is abbreviated with AP, Bill James Handbook is abbreviated with BJ, CHONE is abbreviated with CH, Marcel is abbreviated with M, PECOTA is abbreviated with P and ZiPS is abbreviated with Z. The system that performs the best for the given evaluation criteria for the given year is bolded.

Average error is the measure of the average absolute (without regard to sign) error of the player projections. Average error is measured for each statistical category. The system with the smallest average error in each category in each year is bolded to indicate that it is the most accurate. The 2009 projections are evaluated through MLB games completed on July 27<sup>th</sup>, 2009.

## Average Error: Hitter At Bats

Year	AP	BJ	CH	M	P	Z
2006	<b>102.0</b>	102.7	NA	122.9	114.4	133.7
2007	<b>97.7</b>	98.6	128.8	122.8	116.1	126.8
2008	<b>109.9</b>	112.8	151.4	128.4	129.3	139.6
2009	65.4	<b>63.7</b>	84.8	78.7	75.2	80.2

## Average Error: Hitter Hits

Year	AP	BJ	CH	M	P	Z
2006	<b>31.1</b>	31.8	NA	36.4	34.3	39.1
2007	<b>29.9</b>	30.9	37.3	36.0	34.9	37.0
2008	<b>32.9</b>	34.6	42.6	37.2	37.4	40.6
2009	<b>20.1</b>	<b>20.1</b>	24.7	23.1	22.5	31.2

## Average Error: Hitter Runs

Year	AP	BJ	CH	M	P	Z
2006	<b>17.8</b>	18.2	NA	20.6	19.1	22.3
2007	<b>16.8</b>	17.5	19.4	19.8	19.2	20.2
2008	<b>18.8</b>	19.9	23.4	20.6	20.6	21.8
2009	<b>11.3</b>	11.6	13.9	12.8	12.2	38.6

## Average Error: Hitter Doubles

Year	AP	BJ	CH	M	P	Z
2006	<b>7.7</b>	8.1	NA	8.6	8.1	9.3
2007	<b>7.3</b>	7.5	8.7	8.6	8.2	9.0
2008	<b>8.0</b>	8.5	9.6	8.6	8.7	9.3
2009	<b>5.1</b>	5.3	6.0	5.5	5.6	5.7

## Average Error: Hitter Triples

Year	AP	BJ	CH	M	P	Z
2006	<b>1.4</b>	<b>1.4</b>	NA	1.5	1.5	1.5
2007	<b>1.4</b>	<b>1.4</b>	<b>1.4</b>	1.5	1.5	1.5
2008	<b>1.3</b>	<b>1.3</b>	1.5	1.5	<b>1.3</b>	1.5
2009	<b>1.0</b>	<b>1.0</b>	<b>1.0</b>	<b>1.0</b>	<b>1.0</b>	<b>1.0</b>

## Average Error: Hitter Home Runs

Year	AP	BJ	CH	M	P	Z
2006	<b>5.2</b>	5.3	NA	5.9	5.6	5.8
2007	<b>5.2</b>	5.5	5.6	5.8	5.3	5.6
2008	<b>4.9</b>	5.3	5.9	5.4	5.3	5.5
2009	<b>3.3</b>	3.5	3.8	3.7	3.6	3.7

Average Error: Hitter RBIs

Year	AP	BJ	CH	M	P	Z
2006	<b>17.2</b>	17.4	NA	20.1	19.1	21.6
2007	<b>16.7</b>	17.6	19.3	19.7	18.7	21.0
2008	<b>18.0</b>	19.2	23.0	19.1	19.9	21.4
2009	<b>11.1</b>	11.2	13.6	12.6	12.2	13.0

Average Error: Pitcher Hits (Given up)

Year	AP	BJ	CH	M	P	Z
2006	<b>32.9</b>	33.8	NA	36.0	37.1	43.1
2007	<b>33.9</b>	36.0	42.6	37.2	36.1	43.4
2008	<b>34.4</b>	37.1	41.0	36.3	34.6	42.7
2009	<b>17.7</b>	18.1	21.4	20.0	18.9	21.6

Average Error: Hitter Stolen Bases

Year	AP	BJ	CH	M	P	Z
2006	<b>3.6</b>	3.7	NA	3.9	3.7	4.2
2007	3.5	<b>3.4</b>	3.9	3.9	3.9	3.8
2008	<b>3.7</b>	<b>3.7</b>	4.4	4.1	4.3	4.1
2009	<b>2.7</b>	<b>2.7</b>	3.0	2.9	2.8	3.1

Average Error: Hitter Walks

Year	AP	BJ	CH	M	P	Z
2006	<b>12.4</b>	12.7	NA	14.8	13.3	15.8
2007	<b>12.0</b>	12.5	14.3	14.2	13.3	14.1
2008	<b>13.1</b>	13.7	15.8	14.5	14.5	14.9
2009	<b>9.0</b>	9.2	10.5	9.7	9.6	9.7

Average Error: Hitter Strikeouts

Year	AP	BJ	CH	M	P	Z
2006	<b>19.2</b>	20.0	NA	23.1	21.4	25.3
2007	<b>19.0</b>	19.5	24.6	24.3	21.5	25.4
2008	<b>21.5</b>	22.6	29.6	25.1	25.1	27.5
2009	<b>16.7</b>	17.6	19.2	17.9	17.6	19.4

Average Error: Innings Pitched

Year	AP	BJ	CH	M	P	Z
2006	<b>32.5</b>	34.7	NA	34.8	36.4	43.6
2007	<b>32.9</b>	35.5	40.5	35.8	35.6	42.6
2008	<b>34.4</b>	37.3	40.6	35.8	34.6	41.5
2009	<b>19.2</b>	<b>19.2</b>	23.0	21.1	21.4	22.7

Average Error: Pitcher Earned Runs

Year	AP	BJ	CH	M	P	Z
2006	<b>15.8</b>	16.2	NA	17.1	36.4	43.6
2007	<b>16.1</b>	17.0	19.2	17.7	35.6	42.6
2008	<b>15.9</b>	17.2	19.6	16.9	34.6	41.6
2009	<b>8.7</b>	9.0	10.0	9.7	9.0	10.5

Average Error: Pitcher Strikeouts

Year	AP	BJ	CH	M	P	Z
2006	<b>26.9</b>	29.6	NA	28.2	27.8	35.1
2007	<b>27.6</b>	30.4	30.6	28.4	28.2	32.7
2008	<b>28.9</b>	32.3	32.8	29.6	28.8	34.0
2009	<b>16.7</b>	17.6	19.2	17.9	17.6	19.4

Average Error: Pitcher Walks

Year	AP	BJ	CH	M	P	Z
2006	<b>12.8</b>	13.9	NA	12.9	13.7	16.7
2007	<b>12.7</b>	13.7	15.6	13.8	13.6	16.0
2008	<b>12.1</b>	13.6	15.3	12.8	12.5	15.4
2009	<b>7.7</b>	8.2	8.6	8.1	8.0	8.8

Root Mean Squared Error (RMSE) is the frequently-used measure of the differences between values predicted by a model or an estimator and the values actually observed from the phenomenon being modeled or estimated. RMSE is known as the best measure of accuracy for prediction models. RMSE is measured for each statistical category. The system with the smallest RMSE in each category in each year is bolded to indicate that it is the most accurate. The 2009 projections are evaluated through MLB games completed on July 27<sup>th</sup>, 2009.

RMSE: Hitter At Bats

Year	AP	BJ	CH	M	P	Z
2006	<b>129.0</b>	135.6	NA	151.2	147.3	155.7
2007	<b>120.0</b>	121.8	146.8	147.4	146.7	147.5
2008	<b>135.7</b>	142.3	165.8	151.0	159.3	159.8
2009	<b>83.6</b>	84.2	98.4	95.8	97.4	99.5

RMSE: Hitter Hits

Year	AP	BJ	CH	M	P	Z
2006	<b>39.3</b>	41.2	NA	44.6	43.6	45.3
2007	<b>36.3</b>	36.9	42.5	42.9	43.3	42.6
2008	<b>40.0</b>	41.9	46.5	43.4	45.3	45.6
2009	<b>25.4</b>	25.8	28.2	27.9	28.5	30.2

RMSE: Hitter Runs

Year	AP	BJ	CH	M	P	Z
2006	<b>22.3</b>	23.3	NA	25.2	24.5	26.7
2007	<b>19.7</b>	20.5	21.4	23.1	22.8	22.3
2008	<b>22.5</b>	24.1	24.8	23.9	24.7	24.7
2009	<b>14.1</b>	14.5	15.5	15.4	15.7	19.7

RMSE: Hitter Doubles

Year	AP	BJ	CH	M	P	Z
2006	<b>9.6</b>	10.2	NA	10.5	10.1	10.9
2007	<b>8.8</b>	8.9	10.1	10.2	10.2	10.4
2008	<b>9.4</b>	9.9	10.6	10.0	10.4	10.6
2009	<b>6.3</b>	6.5	6.9	6.7	7.0	6.9

RMSE: Hitter Triples

Year	AP	BJ	CH	M	P	Z
2006	<b>2.0</b>	2.1	NA	2.1	2.1	2.1
2007	<b>2.0</b>	<b>2.0</b>	6.5	2.1	2.1	2.2
2008	<b>1.7</b>	1.8	7.1	1.9	1.8	2.0
2009	<b>1.3</b>	1.4	<b>1.3</b>	1.4	<b>1.3</b>	1.4

RMSE: Hitter Home Runs

Year	AP	BJ	CH	M	P	Z
2006	<b>7.0</b>	7.3	NA	7.8	7.5	7.5
2007	<b>6.3</b>	6.8	6.5	6.9	6.6	6.6
2008	<b>6.6</b>	7.1	7.1	6.9	7.0	7.1
2009	<b>4.5</b>	4.6	4.8	4.9	4.9	4.9

RMSE: Pitcher Walks

Year	AP	BJ	CH	M	P	Z
2006	<b>16.1</b>	17.6	NA	16.8	17.5	19.2
2007	<b>16.9</b>	18.1	19.0	18.1	17.5	19.4
2008	<b>15.6</b>	17.0	17.6	16.3	16.2	17.9
2009	<b>10.2</b>	10.8	11.6	10.8	10.4	11.7

RMSE: Hitter RBIs

Year	AP	BJ	CH	M	P	Z
2006	<b>22.0</b>	22.8	NA	25.2	24.4	24.9
2007	<b>20.6</b>	21.7	23.5	23.6	23.2	22.7
2008	<b>22.1</b>	23.7	24.9	23.3	24.5	24.6
2009	<b>13.8</b>	14.2	15.5	15.1	15.3	15.4

RMSE: Pitcher Hits (Given up)

Year	AP	BJ	CH	M	P	Z
2006	<b>43.8</b>	45.8	NA	48.4	49.5	53.3
2007	<b>47.3</b>	50.2	53.0	50.8	48.8	52.7
2008	<b>46.2</b>	49.0	49.9	47.6	48.1	50.6
2009	<b>25.2</b>	25.8	20.5	27.7	26.7	30.3

RMSE: Hitter Stolen Bases

Year	AP	BJ	CH	M	P	Z
2006	<b>5.7</b>	6.1	NA	6.1	6.0	6.8
2007	<b>5.5</b>	5.6	6.2	6.0	6.0	6.0
2008	<b>5.8</b>	6.3	6.8	5.9	6.5	6.8
2009	<b>4.4</b>	4.5	4.7	4.7	4.6	4.9

The Pearson correlation coefficient is a measure of the correlation (linear dependence) between two variables. The Pearson correlation coefficient is measured for each statistical category. The system with the highest Pearson correlation coefficient in each category in each year is bolded to indicate that it is the most accurate. The 2009 projections are evaluated through MLB games completed on July 27<sup>th</sup>, 2009.

RMSE: Hitter Walks

Year	AP	BJ	CH	M	P	Z
2006	<b>16.2</b>	17.0	NA	18.8	17.5	18.7
2007	<b>15.1</b>	16.0	16.5	17.1	17.0	16.1
2008	<b>16.2</b>	17.2	17.7	17.5	17.8	18.0
2009	<b>11.4</b>	11.8	12.5	12.3	12.2	12.1

Pearson Correlation Coefficient: Hitter At Bats

Year	AP	BJ	CH	M	P	Z
2006	<b>.73</b>	.71	NA	.60	.63	.57
2007	<b>.78</b>	.77	.63	.63	.63	.62
2008	<b>.70</b>	.67	.47	.59	.55	.53
2009	<b>.71</b>	<b>.71</b>	.57	.59	.58	.54

RMSE: Hitter Strikeouts

Year	AP	BJ	CH	M	P	Z
2006	<b>25.1</b>	26.6	NA	29.6	28.6	31.8
2007	<b>24.3</b>	25.3	28.8	29.7	27.8	30.1
2008	<b>28.1</b>	30.2	34.7	32.9	32.9	33.8
2009	<b>18.1</b>	18.7	22.5	20.0	21.6	21.4

Pearson Correlation Coefficient: Hitter Hits

Year	AP	BJ	CH	M	P	Z
2006	<b>.72</b>	.71	NA	.63	.65	.61
2007	<b>.78</b>	.77	.68	.67	.66	.67
2008	<b>.69</b>	.68	.55	.63	.60	.58
2009	<b>.71</b>	.70	.63	.63	.61	.57

RMSE: Innings Pitched

Year	AP	BJ	CH	M	P	Z
2006	<b>42.3</b>	44.2	NA	46.0	47.8	49.8
2007	<b>46.1</b>	49.1	50.1	48.8	47.5	51.1
2008	<b>46.1</b>	48.8	49.0	47.6	47.3	49.3
2009	<b>26.8</b>	27.2	31.3	28.8	28.2	30.8

Pearson Correlation Coefficient: Hitter Runs

Year	AP	BJ	CH	M	P	Z
2006	<b>.74</b>	.73	NA	.66	.68	.60
2007	<b>.79</b>	.78	.75	.70	.71	.72
2008	<b>.70</b>	.68	.61	.64	.63	.63
2009	<b>.71</b>	.70	.63	.64	.63	.58

RMSE: Pitcher Earned Runs

Year	AP	BJ	CH	M	P	Z
2006	<b>20.9</b>	21.9	NA	23.2	23.8	26.2
2007	<b>22.4</b>	23.7	24.8	24.3	23.4	25.6
2008	<b>21.2</b>	22.4	23.4	22.1	22.6	24.4
2009	<b>12.0</b>	12.4	14.3	13.3	12.5	14.5

Pearson Correlation Coefficient: Hitter Doubles

Year	AP	BJ	CH	M	P	Z
2006	<b>.65</b>	.64	NA	.55	.60	.53
2007	<b>.73</b>	.73	.62	.61	.61	.59
2008	<b>.67</b>	.65	.54	.60	.57	.56
2009	<b>.64</b>	<b>.64</b>	.55	.58	.54	.55

RMSE: Pitcher Strikeouts

Year	AP	BJ	CH	M	P	Z
2006	<b>34.0</b>	36.3	NA	36.2	36.2	38.3
2007	<b>36.6</b>	39.4	36.9	37.8	37.7	38.6
2008	<b>38.0</b>	40.4	39.7	38.9	38.7	40.9
2009	<b>22.9</b>	23.7	25.7	24.4	23.3	25.8

Pearson Correlation Coefficient: Hitter Triples

Year	AP	BJ	CH	M	P	Z
2006	<b>.65</b>	.62	NA	.61	.60	.53
2007	<b>.63</b>	.62	.62	.57	.58	.59
2008	<b>.64</b>	.62	.54	.55	.58	.56
2009	<b>.50</b>	<b>.50</b>	.48	.43	.47	.46

Pearson Correlation Coefficient: Hitter Home Runs

Year	AP	BJ	CH	M	P	Z
2006	<b>.78</b>	.77	NA	.73	.74	.74
2007	<b>.78</b>	.77	.77	.73	.76	.76
2008	<b>.77</b>	.75	.72	.74	.73	.73
2009	.74	<b>.75</b>	.71	.70	.70	.70

Pearson Correlation Coefficient: Pitcher Walks

Year	AP	BJ	CH	M	P	Z
2006	<b>.70</b>	.69	NA	.72	.73	.69
2007	<b>.66</b>	.64	.69	.67	.69	.66
2008	<b>.71</b>	.69	.65	.66	.67	.63
2009	.71	.70	.61	.67	.70	.62

Pearson Correlation Coefficient: Hitter RBIs

Year	AP	BJ	CH	M	P	Z
2006	<b>.75</b>	<b>.75</b>	NA	.66	.69	.67
2007	<b>.78</b>	.77	.72	.70	.71	.73
2008	<b>.73</b>	.71	.63	.68	.66	.65
2009	<b>.74</b>	.73	.65	.68	.66	.66

Pearson Correlation Coefficient: Pitcher Hits (Given up)

Year	AP	BJ	CH	M	P	Z
2006	<b>.80</b>	.79	NA	.74	.74	.69
2007	<b>.73</b>	.71	.66	.68	.71	.67
2008	<b>.73</b>	.72	.69	.71	.70	.69
2009	<b>.80</b>	.79	.69	.75	.77	.70

Pearson Correlation Coefficient: Hitter Stolen Bases

Year	AP	BJ	CH	M	P	Z
2006	<b>.83</b>	.80	NA	.82	.69	.67
2007	<b>.84</b>	<b>.84</b>	.79	.82	.71	.73
2008	<b>.80</b>	.79	.74	.74	.66	.65
2009	<b>.75</b>	.74	.70	.71	.72	.70

Pearson Correlation Coefficient: Hitter Walks

Year	AP	BJ	CH	M	P	Z
2006	<b>.78</b>	.77	NA	.70	.74	.70
2007	<b>.81</b>	.80	.77	.75	.75	.78
2008	<b>.76</b>	.74	.70	.71	.70	.70
2009	<b>.73</b>	.71	.66	.69	.68	.69

Pearson Correlation Coefficient: Hitter Strikeouts

Year	AP	BJ	CH	M	P	Z
2006	<b>.74</b>	.73	NA	.61	.65	.57
2007	<b>.77</b>	.76	.66	.63	.69	.64
2008	<b>.71</b>	.68	.53	.64	.61	.58
2009	<b>.71</b>	<b>.71</b>	.52	.63	.59	.59

Pearson Correlation Coefficient: Innings Pitched

Year	AP	BJ	CH	M	P	Z
2006	<b>.79</b>	<b>.79</b>	NA	.74	.73	.70
2007	<b>.71</b>	.70	.65	.66	.69	.64
2008	<b>.72</b>	.70	.68	.68	.69	.67
2009	<b>.78</b>	.77	.67	.74	.75	.69

Pearson Correlation Coefficient: Pitcher Earned Runs

Year	AP	BJ	CH	M	P	Z
2006	<b>.78</b>	.77	NA	.72	.72	.70
2007	<b>.71</b>	.68	.64	.65	.69	.64
2008	<b>.73</b>	.72	.68	.70	.69	.67
2009	<b>.78</b>	.77	.67	.72	.76	.67

Pearson Correlation Coefficient: Pitcher Strikeouts

Year	AP	BJ	CH	M	P	Z
2006	<b>.77</b>	.76	NA	.72	.73	.69
2007	<b>.71</b>	.69	.65	.67	.69	.66
2008	<b>.69</b>	.68	.67	.66	.67	.63
2009	<b>.75</b>	.74	.68	.72	.74	.67

## ACKNOWLEDGMENT

Ross J. Gore would like to thank Michael Spiegel for helping to hone this idea and referring us to the BellKor literature despite his “healthy distaste” for sports. We would also like to thank Chone Smith for his prompt reply to our query about the availability of the 2006 CHONE projections.

## REFERENCES

- [1] <http://www.bradnull.blogspot.com/> accessed 12 March 2009.
- [2] <http://bis-store.stores.yahoo.net/bijahapr203.html> accessed 12 March 2009.
- [3] [http://www.replacementlevel.com/index.php/RLYW/comments/cairo\\_projections\\_v01](http://www.replacementlevel.com/index.php/RLYW/comments/cairo_projections_v01) accessed 12 March 2009.
- [4] <http://fantasynews.cbssports.com/fantasybaseball/stats/sortable/points/1B/standard/projections> accessed 12 March 2009.
- [5] <http://www.baseballprojection.com/> accessed 12 March 2009.
- [6] <http://games.espn.go.com/flb/tools/projections> accessed 12 March 2009.
- [7] <http://www.actasports.com/detail.html?id=019> accessed 12 March 2009.
- [8] [http://baseballanalysts.com/archives/2009/02/2009\\_projection.php](http://baseballanalysts.com/archives/2009/02/2009_projection.php) accessed 12 March 2009.
- [9] <http://www.kffl.com/fantasy-baseball/2009-baseball-draft-guide.php> accessed 12 March 2009.
- [10] <http://www.tangotiger.net/marcel/> accessed 12 March 2009.
- [11] <http://statspeak.net/2008/11/2009-batter-projections.html> accessed 12 March 2009.
- [12] <http://www.baseballprospectus.com/pecota/> accessed 12 March 2009.
- [13] [http://www.rotoworld.com/premium/draftguide/baseball/main\\_page.aspx](http://www.rotoworld.com/premium/draftguide/baseball/main_page.aspx) accessed 12 March 2009.
- [14] <http://www.baseballthinkfactory.org/> accessed 12 March 2009.
- [15] J. Bennet and S. Lanning, “The Netflix Prize”, KDD Cup and Workshop, 2007.
- [16] R. Bell, Y. Koren, and C. Volinsky, “Chasing \$1,000,000: How we won the netflix progress prize”, ASA Statistical and Computing Graphics Newsletter 18(2):4-12, 2007.
- [17] <http://www.baseballprospectus.com/unfiltered/?p=564> accessed 28 July 2009.
- [18] <http://www.cs.virginia.edu/~rjg7v/aggpro/> accessed 28 July 2009.

**Ross J. Gore** is a PhD candidate in computer science at the University of Virginia. He received a Bachelors of Science Degree from the University of Richmond in 2003 and a Masters of Computer Science Degree from the University of Virginia in 2007. He has been a member of Phi Beta Kappa since 2003 and SABR since 2009.

**Cameron T. Snapp** is a Senior IT Consultant from Richmond, VA. Currently, he works for CapTech Ventures, a Richmond based IT Consulting firm. Cameron received a Bachelors of Arts Degree in Computer Science from the University of Richmond and a Masters of Information Technology Management Degree from the University of Virginia.