Using Regression Modeling to Predict Single Season Total Winning Percentage for Baseball Teams

Shannon Leiss

March 15, 2022

Introduction

Motivation/Background

Baseball has many unique elements compared to other sports that provides ample data for analyzing and predicting the game using offensive and defensive elements.

Since the introduction of Sabermetrics and famous prediction formulas - introduced by Bill James in 1980 - many different attempts have been made to optimize a team's performance in these different areas using statistical modeling

Goals of Project

- Using `RetroSheet` and `Teams` Seasonal Baseball data, predicting a team's total single season winning percentage based on offensive and defensive data per season
- Currently, the industry standard for predicting total winning percentage is the Pythagorean Expectation

$$E(Win\%) = \frac{RS^2}{RS^2 + RA^2}$$

- Where RS stands for a team's overall runs scored and RA stands for a team's overall runs allowed
- Overall goal is to create a complex model that predicts better than the Pythagorean Expectation and a "simple" model that predicts similar to the Pythagorean Expectation

Improving on Pythagorean

- Pythagorean predicts within a ± 3 game range using only total runs scored and total runs allowed in a season
- Ignores the underlying offensive and defensive performance of a team that leads to their total runs scored and allowed
- Performs poorly in different situations due to few terms used

Data Cleaning

- `RetroSheet` data set has each individual game statistics for all seasons and all teams from 1871 - 2016
- Lahman's `Teams` data set provides each team's overall performance in a variety of variables per season
- Data sets were combined to create per game, split by home and away games, statistics for each team, each season

Definition of Baseball Terms

- Terms used in the analysis:
- **Total Win Percentage**, **Home Win Percentage**, **Previous Season Win Percentage**, **Previous Home Win Percentage**

Offensive Abbreviation	Description	Defensive Abbreviation	Description
H	Hits Batted in	**HA**	Hits Allowed
HR	Homeruns Batted in	**HRA**	Homeruns Allowed
RS	Runs Scored	**RA**	Runs Allowed
SO	Strikeouts By Batters	**SOP**	Strikeouts by Pitchers
BB	Walks By Batters	**BBP**	Walks allowed by Pitchers
RNS	Runners Stranded	**RNSP**	Runners Stranded by Pitchers
X3B	Triples Batted in	**X3P**	Triples Allowed by Pitchers
		FP	Fielding Percentage

- All variables will be split into per home game and per away game - besides FP - to make seasons with different games played comparable
- All of the above variables were also used to create differential statistics - besides FP. These are the difference between the corresponding offensive and defensive statistics that match for a team.

Methods

Linear Regression

- Different Linear Regression models will be considered to predict total winning percentage
- Models will be compared using AIC, BIC, and Adjusted R^2

Bootstrapping

- Will be used to find the optimal values of $\hat{\beta}_i$ terms in the selected models
- Randomly splitting the data into training sets of 80% of the seasons will be used to fit the different models with the predictions for these models being fit on the test sets of seasons
- Prediction performance of the models will be compared using the AME of the test seasons, where:

$$AME = \frac{1}{n} \sum_{i=1}^{n=20} |Y_i - \hat{Y}_i|$$

Penalized Regression

- Since all offensive statistics are highly correlated, the data suffers from multicollinearity
  - Multicollinearity does not violate model assumptions and can be ignored when predicting values
- Ridge and Lasso regression will be used to combat multicollinearity issues

Bootstrapping

- Many different simulations were run with slightly different constraints/tasks
  - For linear model comparison, 500 simulations using Cubs data to fit models on Cubs training seasons and predicting on the Cubs test seasons
  - For penalized regression comparisons, 100 simulations using predictors decided in linear comparison fit using Cubs training seasons and predictions on Cubs test seasons
  - For optimal coefficients for final models, 100 simulations fit on Cubs training seasons, predicting on all Teams all Seasons data to find coefficient values that minimize the AME for all teams all seasons

Procedure of Selecting/Fitting Models

- Due to size of data, models were fit using only the Chicago Cubs for the 1920-2016 seasons
- Potential complex models were created based on all available per game splits, offensive focused, defensive focused, and more
- Simple models were created based on variables with high correlation with total winning percentage
- 15 models were compared through simulation, 5 models - 2 complex and 3 simple - were chosen for final comparison

Terms of 5 Comparison Models - Complex Models

- **Model 1:** *Predicting Total Win Percentage* using the team's *previous season total win percentage, the team's home win percentage, the team's previous season home win percentage,fielding percentage*, along with the team's run, hit, homerun, strikeout, walk, triple, and runners stranded differentials per game - split into home game differentials and away game differentials. This model contained 18 explanatory variables.
- **Model 2:** *Predicting Total Win Percentage* using the team's *previous season total win percentage, the team's home win percentage*, *the team's previous season home win percentage*, *fielding percentage*, along with the teams run, hit, homerun, strikeout, walk, triple, and runners stranded differentials difference between home and away games. This model contained 11 explanatory variables.

Terms of 5 Comparison Models - Simple Models

- **Model 3:** *Predicting Total Win Percentage* using a team's overall *fielding percentage* and the team's run differential per game. This model contained 2 explanatory variables.
- **Model 4:** *Predicting Total Win Percentage* using a team's previous season total win percentage, *fielding percentage*, and the team's run differential during away games. This model contained 3 explanatory variables.
- **Model 5:** *Predicting Total Win Percentage* using a team's previous season total win percentage, *fielding percentage*, and the team's run differential during home games. This model contained 3 explanatory variables.

AIC Comparison Table

Table 1: AIC values for all 5 models fitted on the Cubs training seasons

Simulation	Model 1	Model 2	Model 3	Model 4	Model 5
1	**-396.078**	-376.184	-345.567	-282.172	-276.980
2	**-391.812**	-376.543	-346.906	-292.621	-286.723
3	**-388.984**	-361.243	-353.764	-288.818	-267.238
4	**-385.018**	-358.397	-346.548	-285.928	-279.726
5	**-386.840**	-365.945	-344.097	-292.095	-276.385

BIC Comparison Table

Table 2: BIC values for all 5 models fitted on the Cubs training seasons

Simulation	Model 1	Model 2	Model 3	Model 4	Model 5
1	**-351.546**	-348.058	-338.536	-272.797	-267.605
2	-347.280	**-348.417**	-339.875	-283.246	-277.348
3	-344.452	-333.117	**-346.733**	-279.442	-257.863
4	**-340.486**	-330.272	-339.517	-276.553	-270.350
5	**-342.308**	-337.819	-337.066	-282.720	-267.010

Model Summary Table

Table 3: Summary Table comparing all 5 models and the Pythagorean Expectation

	AME	Lowest AME	Lowest AIC	Lowest BIC
Model 1	0.0156513	79.2%	100%	93%
Model 2	0.0180447	16.0%	0%	5.2%
Model 3	0.0213956	3.4%	0%	1.8%
Model 4	0.0302909	0%	0%	0%
Model 5	0.0308722	0%	0%	0%
Pythagorean	0.0220278	1.4%		

Model Comparison Plot

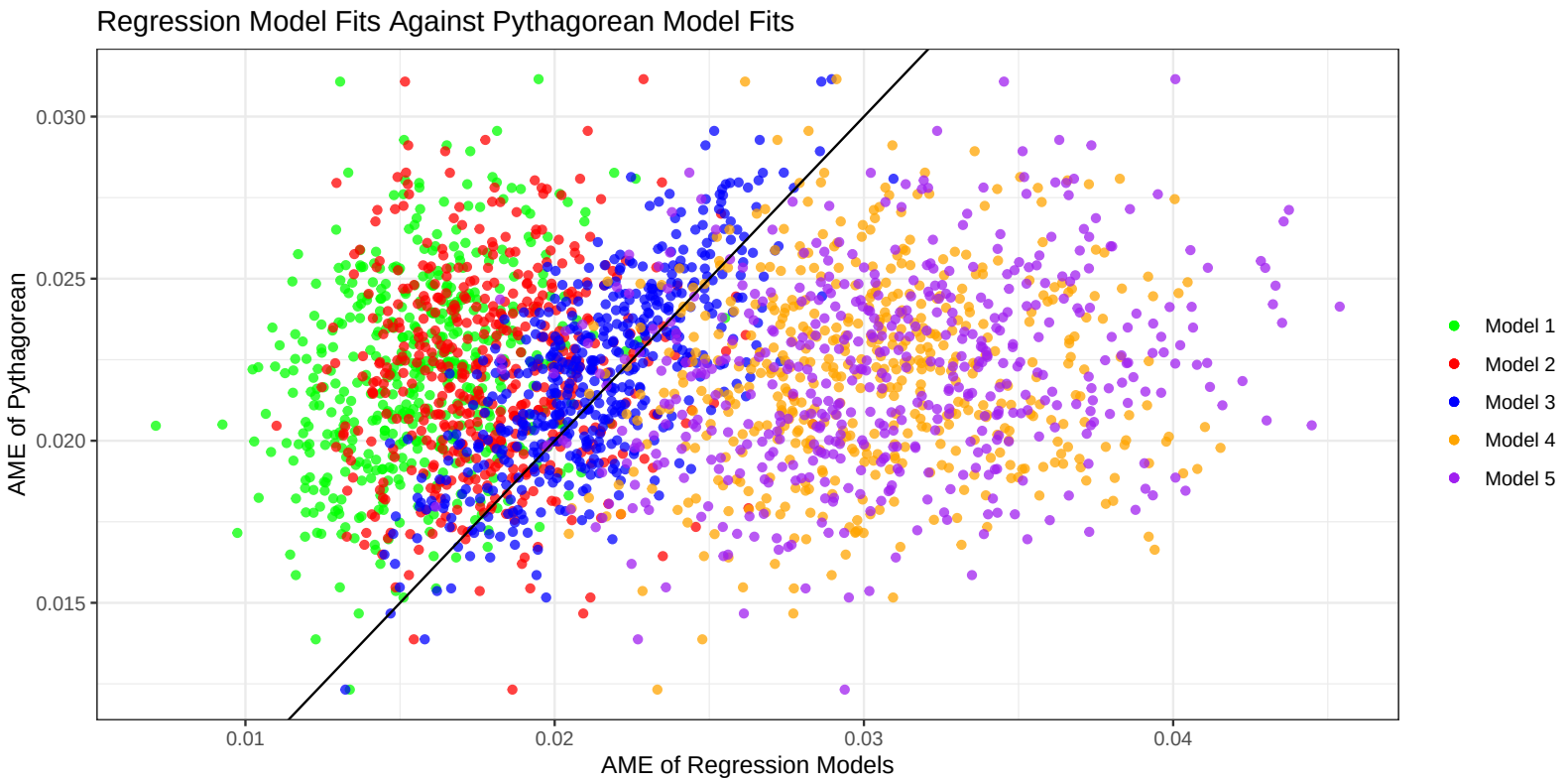


Figure 1: Linear Model AME values on Cubs test seasons compared to the Pythagorean Expectation AME for the test seasons

Final Models

- From comparing models, Model 1 was found to be the best complicated model while Model 3 was found to be the best simple model at predicting total winning percentage
- These models were then simulated again using Cubs data to find the optimal coefficient values
- Simulated models were used to predict all teams winning percentages from the 1920-2016 seasons

Table 4: Example total winning percentage predictions from simulation, predicted using Model 1, Model 3, and the Pythagorean Expectation

Team	Season	Total Win %	Model 1 Fit	Model 3 Fit	Pythagorean Fit	Model 1 Residual	Model 3 Residual	Pythagorean Residual
SFG	2016	0.5370	0.5286	0.5492	0.5622	0.0084	-0.0122	-0.0251
STL	2016	0.5309	0.4977	0.5365	0.5448	0.0331	-0.0057	-0.0140
TBD	2016	0.4198	0.4500	0.4723	0.4704	-0.0303	-0.0526	-0.0507
TEX	2016	0.5864	0.5443	0.5017	0.5053	0.0421	0.0847	0.0812
TOR	2016	0.5494	0.5414	0.5536	0.5650	0.0080	-0.0042	-0.0156
WSN	2016	0.5864	0.5974	0.5894	0.6085	-0.0110	-0.0030	-0.0221

Optimal Coefficient Value Analysis

- Looking at the AME for each simulation run, Simulation run 11 had the lowest AME for both models. With both Model 1 AME and Model 3 AME being lower than the Pythagorean Expectation AME

Table 5: Comparing Simulation Runs ordered by Model 1 AME

Simulation Run	Model 1 AME	Model 3 AME
11	0.01477134	0.02064096
75	0.01493020	0.02090482
93	0.01497031	0.02067588
58	0.01498123	0.02104307
67	0.01498159	0.02075509

Model Assumptions

- Linear Regression model assumptions of linearity, independence of observed values, normally distributed errors, and constant variance of the errors are met for both models

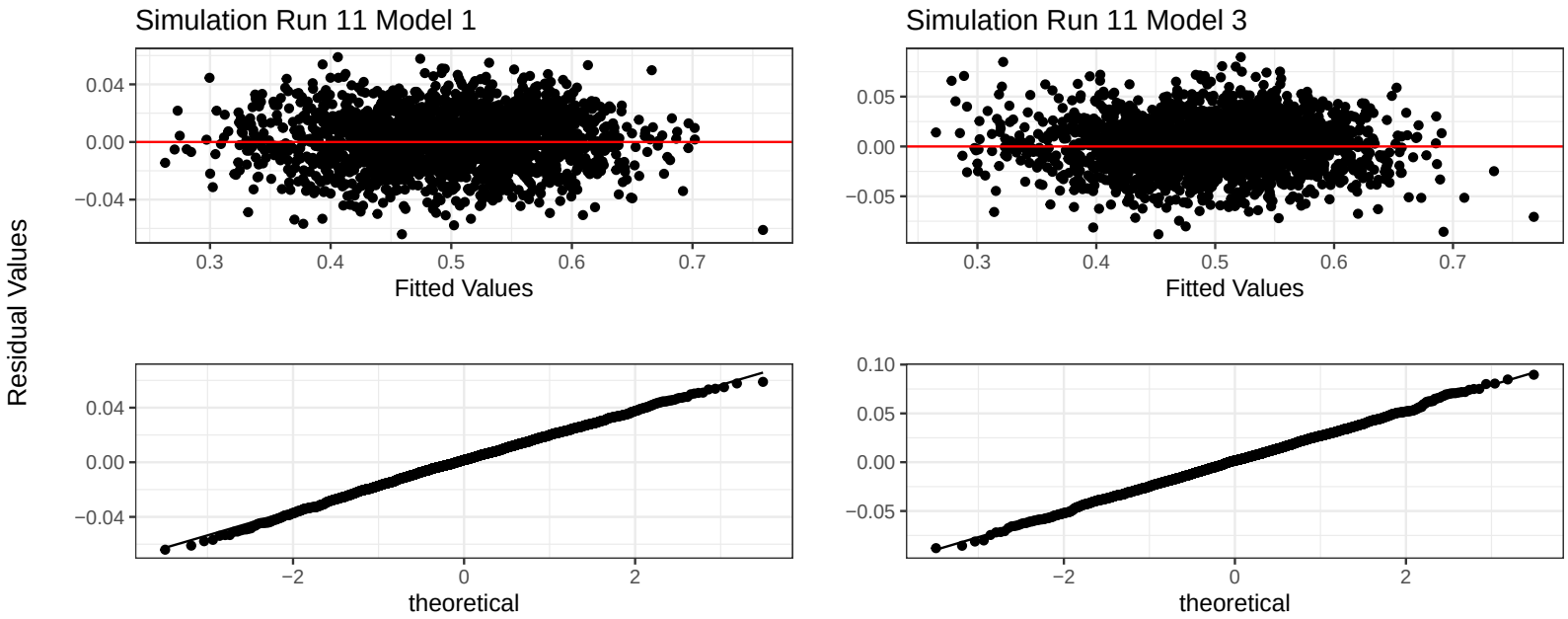


Figure 2: Residuals vs Fitted values plots and Normal QQ-plots for assessing model assumptions for Model 1 and Model 3

Results

Ridge and Lasso Regression

Table 6: Lambda values for Ridge and Lasso regression

Model	Ridge Lambda	Lasso Lambda	Linear AME	Ridge AME	Lasso AME	Pythagorean AME
1	0.00614	0.00216	0.01544	0.01647	0.01441	0.02161
1	0.00584	0.00393	0.01636	0.01776	0.01347	0.02271
1	0.00571	0.00265	0.01333	0.01693	0.01293	0.02827
3	0.00685	0.00349	0.02162	0.02470	0.02405	0.02542
3	0.00645	0.00188	0.01956	0.01621	0.01465	0.01779
3	0.00654	0.00030	0.02045	0.02003	0.01994	0.02160

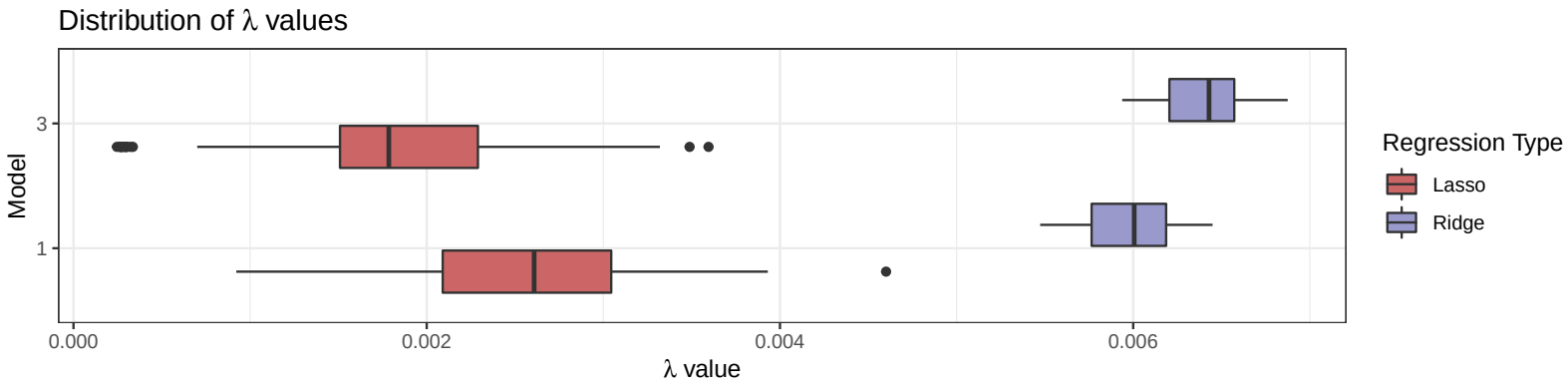


Figure 3: Distribution of lamba values for Ridge and Lasso regression

Ridge and Lasso Regression

- All λ values are extremely small, meaning that these models will not differ substantially from the linear models
- These small λ values lead to the predictions being close to the linear model predictions

Table 7: Overall AME for Cubs test seasons using Ridge, Lasso, and Linear regression compared to the overall AME for the Pythagorean Expectation

Model	Ridge	Lasso	Linear	Pythagorean
1	0.01643	0.01423	0.01569	0.02186
3	0.02189	0.02120	0.02135	0.02186

- Since the $\hat{\beta}$ produced using Ridge and Lasso regression are bias, the Linear regression model will be used

Model Predictions and Pythagorean Comparison

- Using the predictions for all teams between the 1920-2016 seasons, Model 1 had an overall AME of 0.01575, Model 3 had an overall AME of 0.02099, and the Pythagorean Expectation has an overall AME of 0.02067
- Model 1 predicted better than the Pythagorean Expectation 59.216% of the time
- Model 3 predicted better than the Pythagorean Expectation 48.244% of the time

Model Performance - Franchise Level

Table 8: Teams with the lowest AME for Model 1

Team	Model 1 AME	Model 3 AME	Pythagorean AME
ARI	0.01232	0.02274	0.02201
CHC	0.01257	0.02085	0.02199
COL	0.01257	0.02151	0.02062
TOR	0.01294	0.01831	0.01862
WSN	0.01360	0.01850	0.01823

Out of 30 Franchises, 11 of them had AME values for both models lower than the Pythagorean AME, with 6 of these teams also having both models predicting better than the Pythagorean more than 50% of the time

Model Perfomance - League and Decade Splits

- When split by decades, both Model 1 and Model 3 predict modern baseball the best - 2010's, 2000's, and 1990's

Table 9: Decade/League with both models predicting better than the Pythagorean Expectation

Decade	League	Model 1 AME	Model 3 AME	Pythagorean AME
1920	NL	0.01458	0.02040	0.02082
1940	AL	0.01600	0.02122	0.02247
1950	AL	0.01651	0.02288	0.02440
1960	AL	0.01585	0.01843	0.01852
1970	AL	0.01464	0.02067	0.02092
1970	NL	0.01587	0.02191	0.02341
2010	AL	0.01524	0.02078	0.02176

AME Distributions

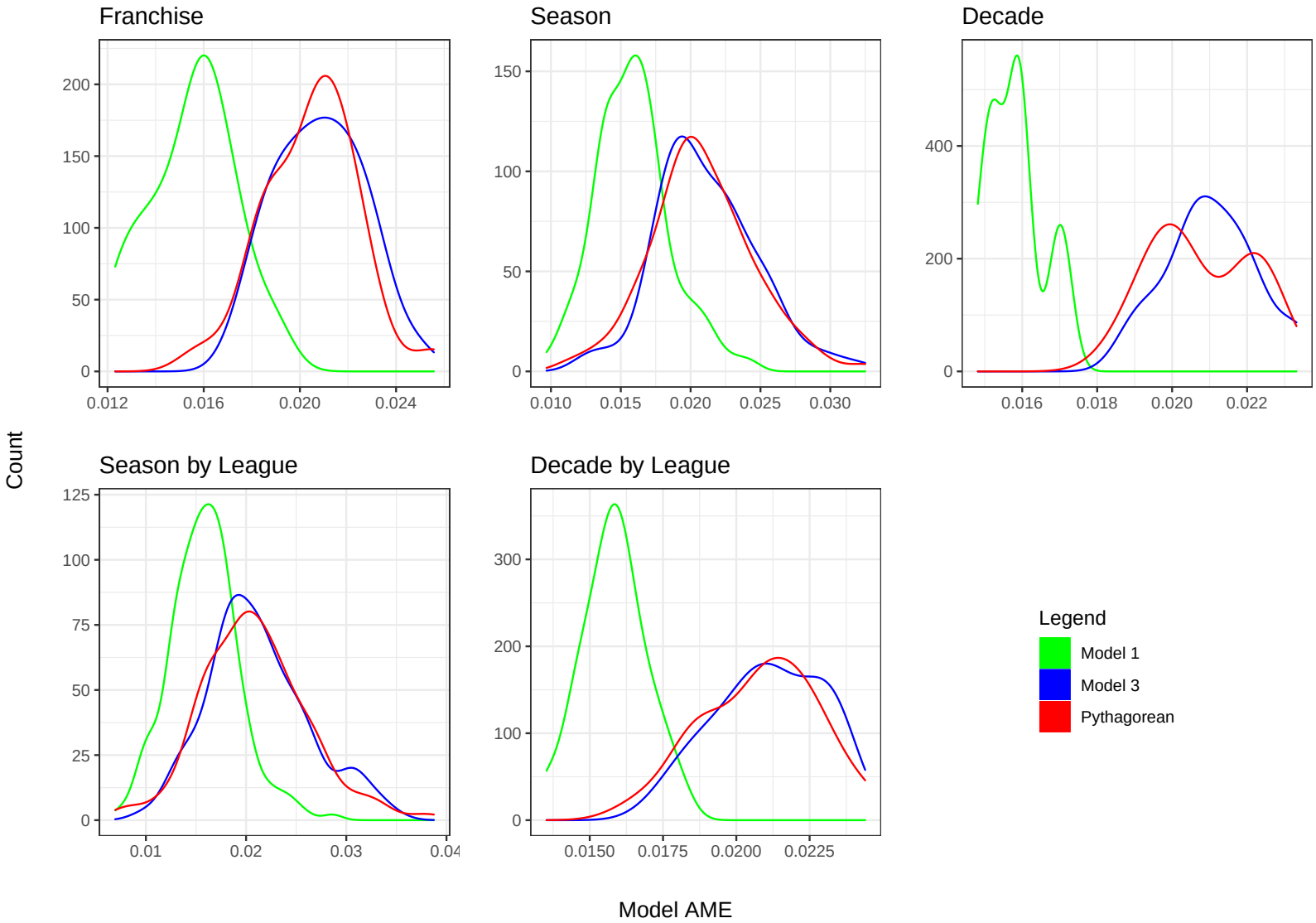


Figure 4: Distribution of AME values for predictions from 1920-2016

Applications of Model

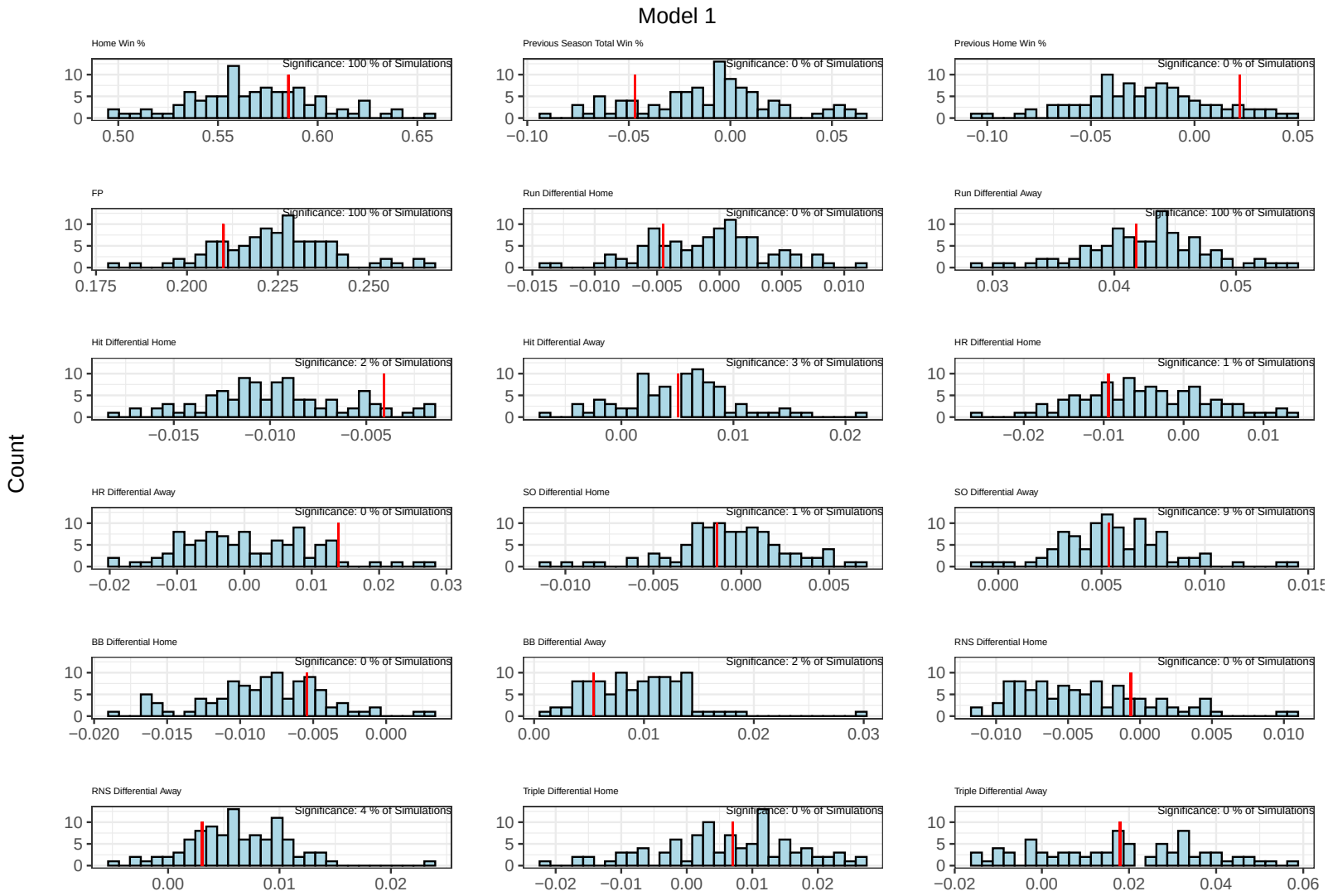
- Model 1 predicts consistently better than the Pythagorean Expectation, with Model 3 predicting similar to the Pythagorean Expectation
- Using terms other than runs scored and runs against allows teams to find specific areas of their team's that need to be improved to achieve a specific predicted winning percentage
- Of all the possible predictors, a team's fielding percentage had a surprisingly high impact on their overall winning percentage, as well a team's run differential for away games
- Model 1 also predicted a team's finish within their division better than the Pythagorean Expectation

Further Work

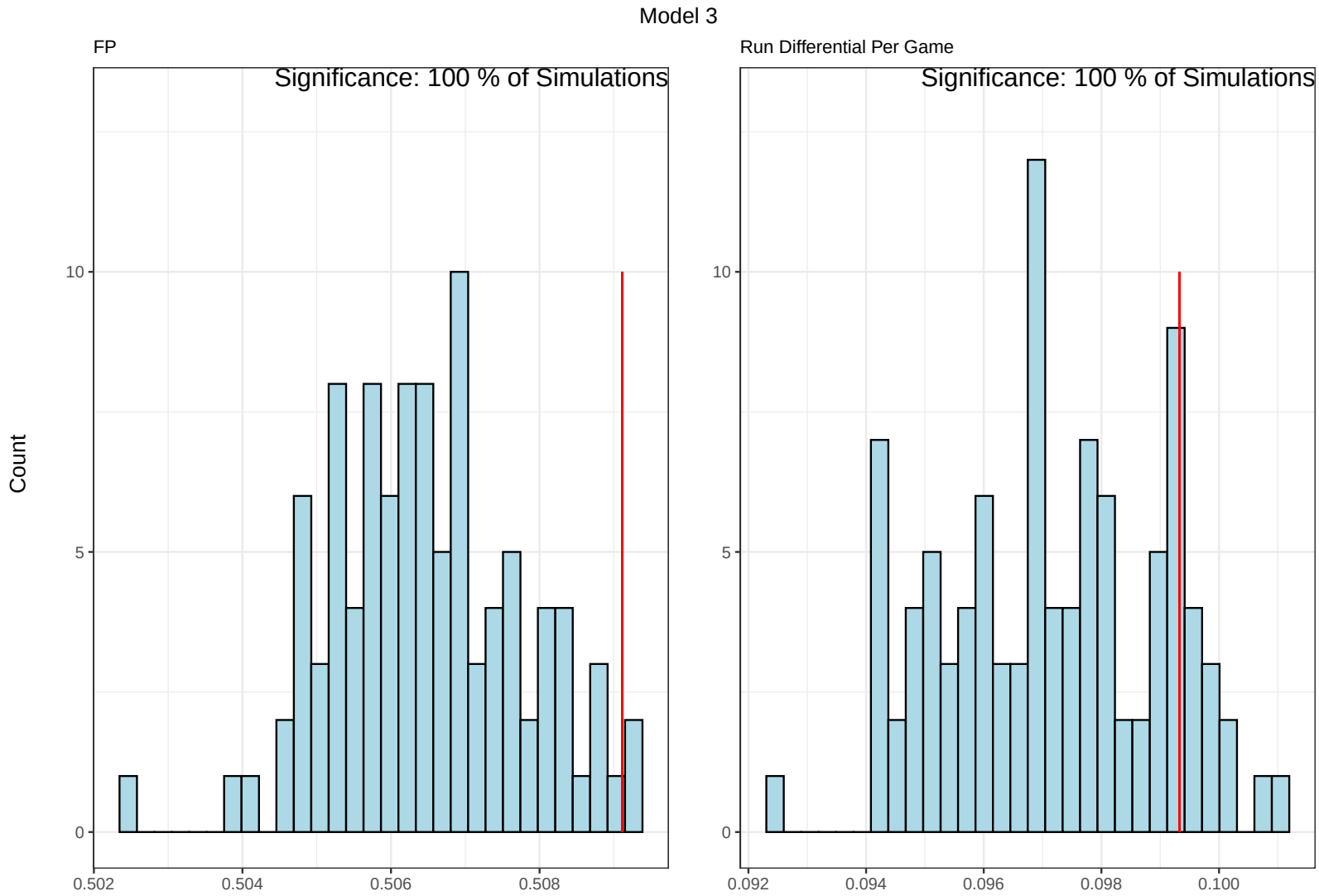
- Further testing of the model on mid-season data could be done to test the mid-season prediction against the end of season total winning percentage
- It could also be examined how many games a team must play in a season before the models have an AME around or below 0.02, as the earlier into the season an accurate prediction can be made the better
- Additional terms, such as difficulty of schedule and ballpark factors, could be investigated to be built into the models

Appendix Plots & Tables

Simulation 11 Against Other Coefficient Values - Model 1



Simulation 11 Against Other Coefficients - Model 3



Simulation 11 Coefficient Values Tables - Model 1

Table 10: Coefficient Values for Model 1

	Home_Win_Per	Previous_Season_Total_Win_Per	Previous_Home_Win_Per	FP	Run_Dif_HG	Run_Dif_AG
Simulation.11	0.5853348	-0.0469708	0.0218164	0.2100444	-0.0045270	0.0417995
Average	0.5693940	-0.0111603	-0.0235038	0.2244103	-0.0009043	0.0426175

	Hit_Dif_HG	Hit_Dif_AG	HR_Dif_HG	HR_dif_AG	SO_Dif_HG	SO_Dif_AG
Simulation.11	-0.0040903	0.0050781	-0.0094223	0.0139833	-0.0013648	0.0053648
Average	-0.0096557	0.0051570	-0.0047992	0.0005739	-0.0006252	0.0056080

	BB_Dif_HG	BB_Dif_AG	RNS_Dif_HG	RNS_Dif_AG	Trip_Dif_HG	Trip_Dif_AG
Simulation.11	-0.0054235	0.0054268	-0.0006435	0.0030601	0.0069812	0.0179530
Average	-0.0082724	0.0094120	-0.0037466	0.0063119	0.0061852	0.0186004

Simulation 11 Coefficient Values Tables - Model 3

Table 11: Coefficient Values for Model 3

	Simulation 11	Average
FP	0.5091103	0.5064585
Run Dif Per Game	0.0993291	0.0971794

Analysis: Simulation 11 against Averaged Coefficients

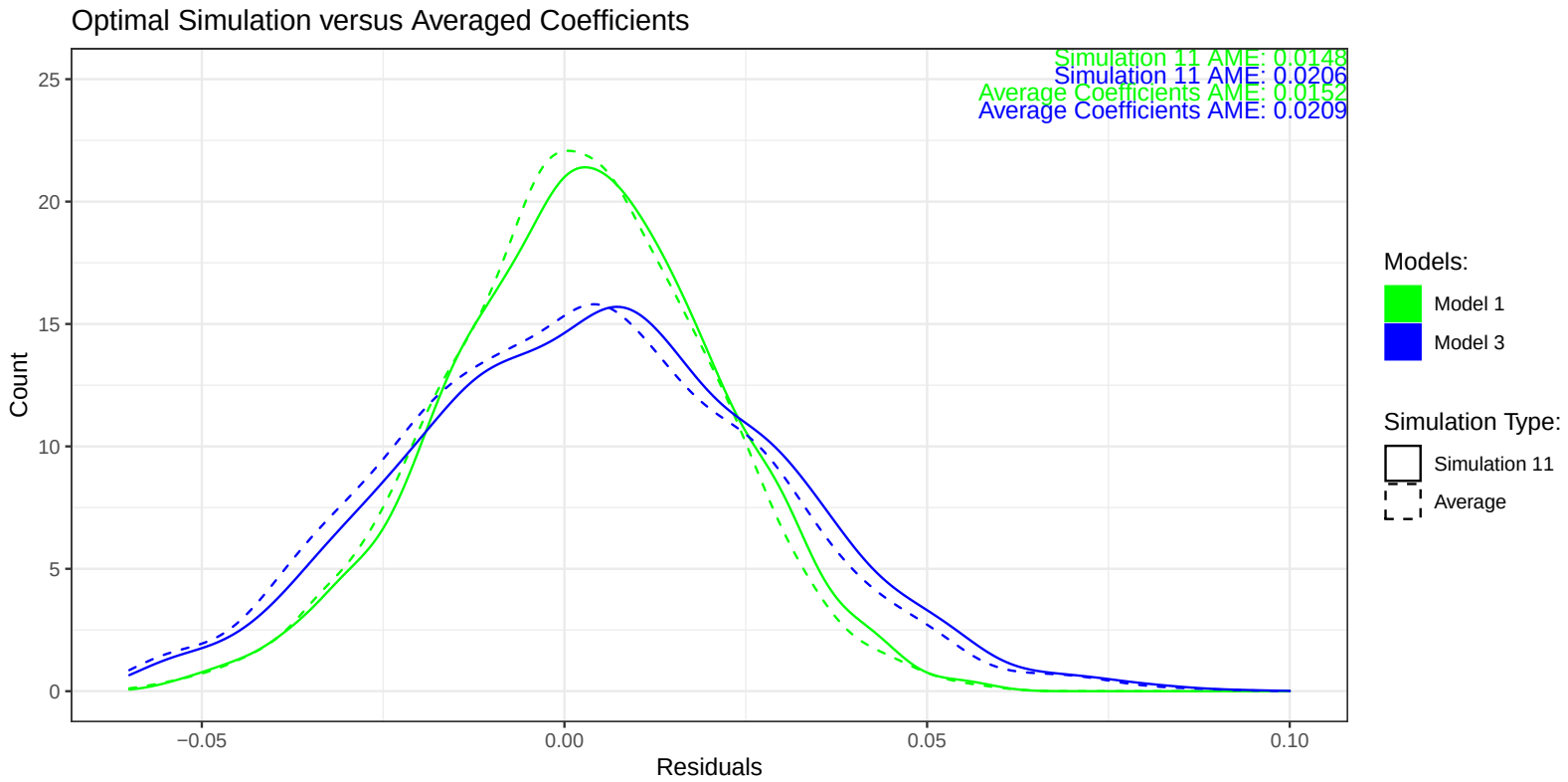


Figure 5: Distribution of residual values for the predicted total winning percentage of all teams from the 1920-2016 seasons for Optimal Simulation run and the Averaged Coefficient values

Team Rank Table for Lowest AME

	ANA	ARI	ATL	BAL	BOS	CHC	CHW	CIN	CLE	COL	DET	FLA	HOU	KCR	LAD
Model.1.AME	26	1	14	23	25	2	22	27	11	3	10	19	16	30	18
Model.3.AME	12	26	17	23	5	15	9	29	18	19	10	2	14	18	14
Pythagorean.AME	17	26	21	19	4	24	18	22	12	15	13	1	30	25	5

	MIL	MIN	NYM	NYY	OAK	PHI	PIT	SDP	SEA	SFG	STL	TBD	TEX	TOR	WSN
Model.1.AME	8	9	29	20	24	17	12	7	28	21	13	6	15	4	5
Model.3.AME	7	8	30	16	27	13	20	22	6	25	21	1	11	3	4
Pythagorean.AME	8	7	29	16	20	23	14	27	9	28	11	2	10	6	3

Division Finish Table

Table 12: Predicted Divison Finish for 2001 NL West

Team	Finish	Total Win Percentage	Model 1	Model 3	Pythagorean
ARI	1	56.79012	59.81951	58.84359	59.34828
SFG	2	55.55556	56.55137	53.07075	53.29312
LAD	3	53.08642	53.20508	50.80212	50.93201
SDP	4	48.76543	47.06960	48.27894	48.56369
COL	5	45.06173	48.91124	51.13880	50.92939

Table 13: Predicted Divison Finish for 2007 AL East

Team	Finish	Total Win Percentage	Model 1	Model 3	Pythagorean
CLE	1	59.25926	57.63253	56.71732	57.02765
DET	2	54.32099	54.63080	55.61474	55.32920
MIN	3	48.76543	49.51874	49.66726	49.51491
CHW	4	44.44444	43.53772	41.04275	40.55575
KCR	5	42.59259	43.50946	45.58001	45.15964