Time Series Analysis of Major League Baseball

Danielle Sebring & Steven Barnett

5/8/2022

Problem Statement

Major League Baseball (MLB) has collected data on its players and their performances since 1871. This rich supply of data allows one to investigate and learn about trends in performance over time. Additionally, the game of baseball itself has evolved over the years. This is due to a variety of factors, ranging from the introduction of new technology to changes in the rules established by MLB. For example, the advent of sports analytics in the early 2000s has had an enormous impact on the game. Recently there has been increased concern that with player improvements, particularly in the area of pitcher arm strength and pitch speed. This has the potential to make the game of baseball less entertaining and therefore cause fans to lose interest. As a result, MLB is concerned about major revenue losses from a decrease in in-person game attendance as well as TV viewership. Therefore, we are going to investigate some of these trends that lead to a less strategic, nuanced, and entertaining game of baseball. We will look at several statistics over the last century that are indicative of these trends, find the best models that fit these time series, and perform forecasting to reveal what the future of MLB holds considering the current direction.

Data Collection

Our data source is Lahman's Baseball Database. This database has batting, pitching, and fielding statistics dating back to 1871. Lahman's data allows for a high-level view of trends across the league, as well as potential for in-depth analysis down to an individual team-by-team or player-by-player level. We also referred to the Baseball Reference via Bill Petti's baseballr package, although there is restricted access to the database. Baseball Reference provided some insight on statistics of interest that were not readily available in Lahman's Baseball Database, but that could be calculated through simple data wrangling procedures.

Our data has over 100 different statistics measured over the last century. An analysis of all of these would go beyond the scope of this project. Therefore, we identified six statistics that are indicative of the changes in baseball detrimental to the game's entertainment value: strikeouts, home runs, hits, wild pitches, batters hit by pitch, and sacrifices. Each of these has either increased or decreased over time due to the greater arm strength of pitchers and the resulting higher velocity of pitches. For example, as velocity of pitch has increased, the number of strikeouts naturally increases because batters have a harder time connecting with the ball. Additionally, home runs increase when pitch speed increases because the exit velocity of the ball from the bat is higher, making a home run more likely.

Theses six statistics are all tracked by count (e.g. number of hits in the league per season). However, it is common in sports analytics to look at count variables as averages or rates. This practice reduces extreme values (e.g. the 2020 MLB season was shortened due to COVID-19, resulting in a drastic drop in all statistics) and produces more well-behaved residuals in model fitting. Therefore, for each of these six statistics, we calculate the "per nine innings" rate.

We also found that data from the early years is highly variable. We attribute this to the rudimentary methods of data collection available in the late 1800s as well as the changes that often come as a sports organization forms and manages the development of its rules and procedures. The National League of MLB officially began in 1876. Due to this early variability, we only consider the last 100 years (1922-2021). Therefore, after all data manipulation and cleanup, we begin our analysis with statistics for the years 1922-2021 listed in Table 1.

Table 1: Generated MLB statistics for Time Series Analysis

Code	Definition
SO9	Strikeouts per Nine Innings
HR9	Home Runs per Nine Innings
H9	Hits per Nine Innings
WP9	Wild Pitches per Nine Innings
HBP9	Hit by Pitch per Nine Innings
SAC9	Sacrifices (bunts and fly outs) per Nine Innings

Model Selection

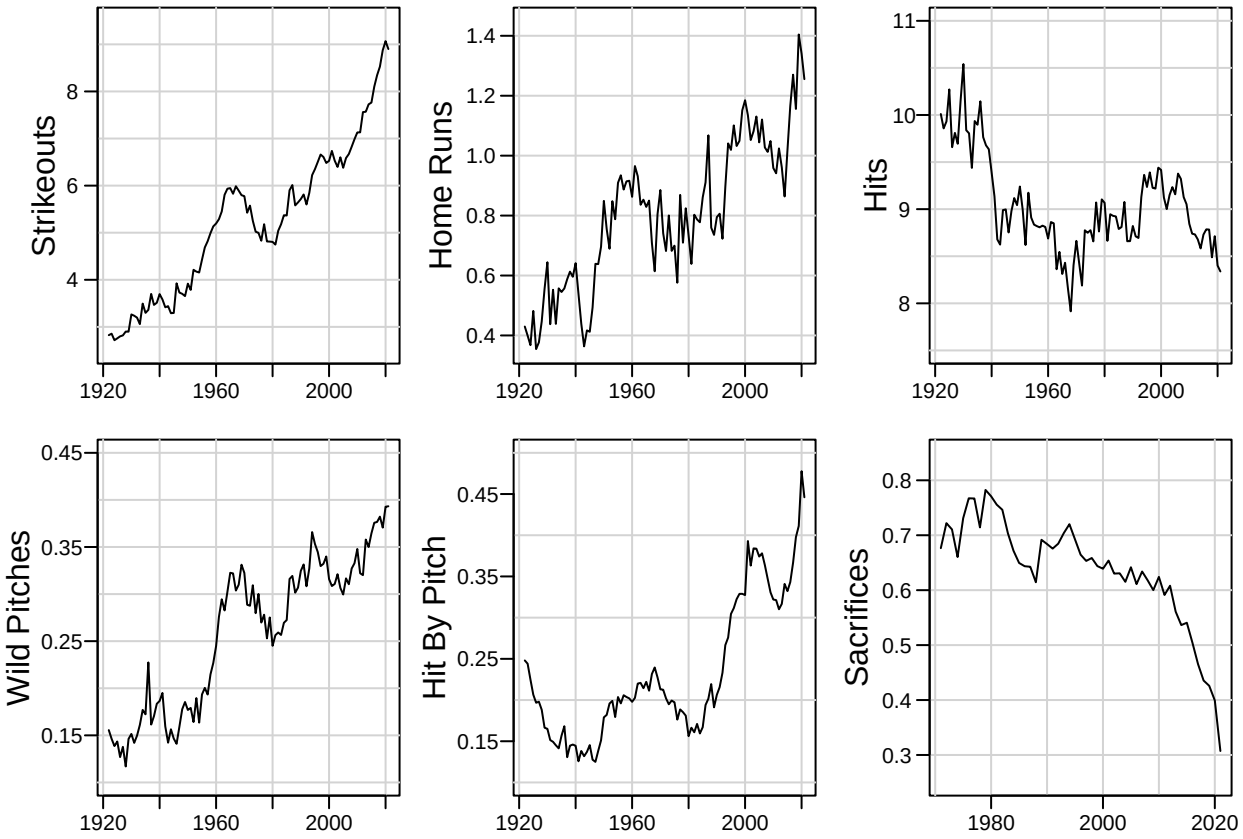

Figure 1: MLB league-wide statistics collected from 1922-2021

Figure 1 displays the six time series of interest. It is clear at first glance that none of these are stationary time series. Therefore, in order to fit a model to these we will either have to consider a Dynamic Linear Model (DLM) or perform some sort of differencing in order to utilize an Auto Regressive or Moving Average model.

Differencing

First, we will calculate a first-order difference on each time series in order to consider fitting Auto Regressive Integrated Moving Average (ARIMA) models. The first-order differencing is displayed in Figure 2.

For the most part, it appears that the first-order differencing took out all the non-stationarity from these time series. There exists a hint of a positive trend or a prominent increase in variability as time goes on in

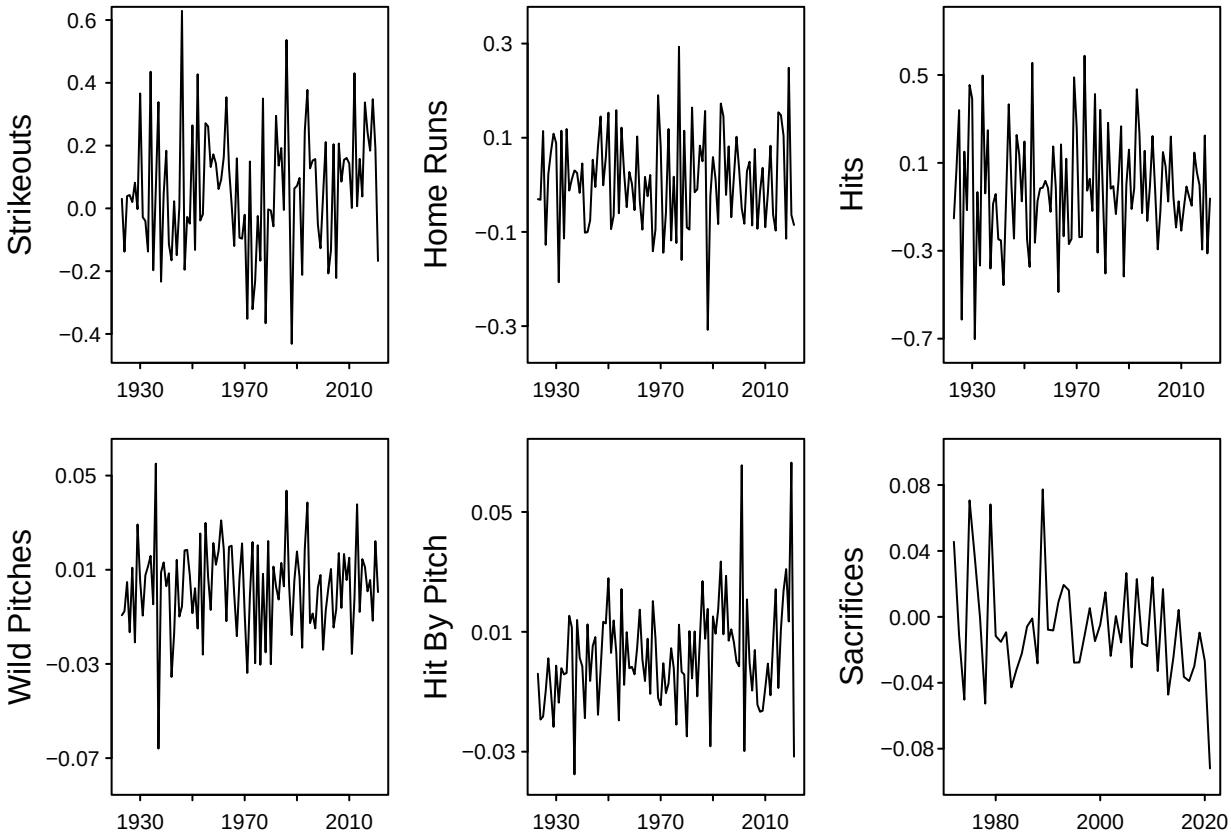

Figure 2: First-order differencing of MLB time series data

the Hit By Pitch time series. Additionally, the Sacrifices time series appears to be trending downwards, but that could just be the final time point, which appears to be a negative outlier. However, we will proceed as if the first-order differencing is sufficient.

Auto Regressive Integrated Moving Average Models

We will next attempt to fit different ARIMA models for these time series. First, we plot the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for each time series to get a sense of any significant lags that would translate into auto regression or moving average coefficients.

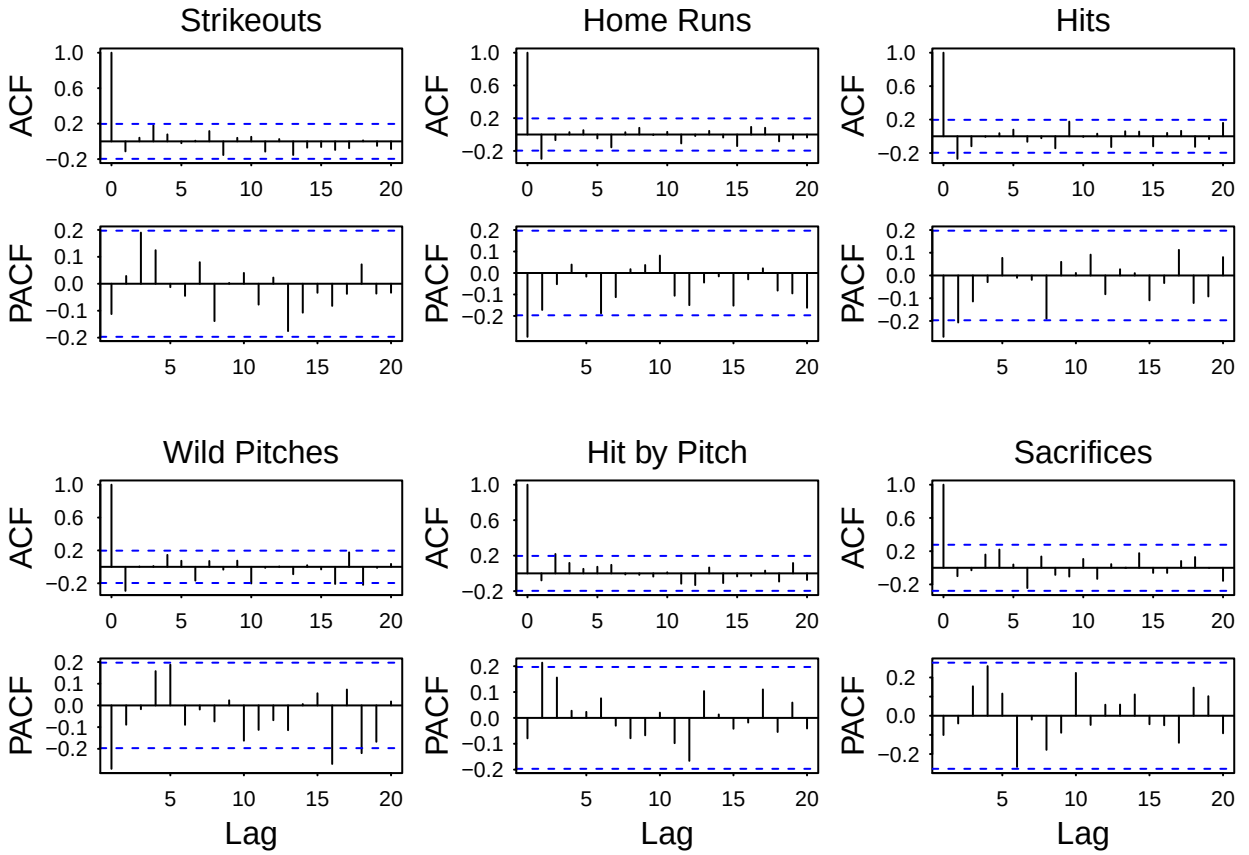

Figure 3: Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for differenced MLB time series

As shown in Figure 3, for the most part we do not see any significant lags in the ACF or PACF for any of the differenced time series. In the Hits time series, it appears that the first lag is barely significant in the ACF and the PACF. As such, we will consider an ARIMA model with an auto regressive order of 0 and 1 and a moving average order of 0 and 1. The Wild Pitches times series also has some interesting behavior. The first lag in the ACF is slightly significant, indicating that we should consider an auto regression coefficient of order 1. However, the most intriguing part is the significant lags of 16 and 18. This indicates that there may be some seasonality in this time series. Looking back at the original time series, we do see some cyclical pattern. It remains to be seen if that is of a regular nature or happens by random chance.

Now that we have some indication of what lags are significant in the models, we fit all combinations of possible order auto regressive, moving average, and differencing components. We calculate the Bayesian Information Criterion (BIC) for each as well as forecasting error for one, two and three step ahead forecasts. We use mean absolute error (MAE), mean squared error (MSE), and mean absolute percentage error (MAPE). Based on these different metrics, we select what we believe to be the best model between performance and simplicity. For brevity's sake, we will only display the output for the Strikeout time series.

Table 2: ARIMA models and their performance metrics for Strikeout time series

	BIC	MAE1	MSE1	MAPE1	MAE2	MSE2	MAPE2	MAE3	MSE3	MAPE3
(0, 0, 0)	386.688	2.579	7.170	0.338	2.607	7.321	0.342	2.633	7.468	0.345
(0, 0, 1)	273.199	1.363	2.067	0.179	2.603	7.299	0.341	2.630	7.448	0.345
(0, 1, 0)	-24.152	0.181	0.044	0.024	0.273	0.108	0.035	0.397	0.210	0.051
(0, 1, 1)	-19.562	0.187	0.045	0.025	0.278	0.111	0.036	0.403	0.216	0.052
(1, 0, 0)	-19.562	0.187	0.045	0.025	0.278	0.111	0.036	0.403	0.216	0.052
(1, 0, 1)	-19.562	0.187	0.045	0.025	0.278	0.111	0.036	0.403	0.216	0.052
(1, 1, 0)	-19.566	0.187	0.045	0.025	0.278	0.112	0.036	0.403	0.216	0.052
(1, 1, 1)	-15.039	0.187	0.045	0.025	0.277	0.111	0.036	0.402	0.215	0.052

Table 3: ARIMA models selected for each MLB time series

	SO9	HR9	H9	WP9	HBP9	SAC9
Best ARIMA Model	(1, 0, 1)	(0, 1, 1)	(0, 1, 1)	(1, 1, 0)	(0, 1, 0)	(0, 1, 0)
BIC	-24.152	-171.22	8.199	-500.661	-517.684	-196.273
MAE1	0.181	0.08	0.142	0.012	0.017	0.027
MSE1	0.044	0.01	0.027	<0.001	0.001	0.001
MAPE1	0.024	0.073	0.016	0.034	0.046	0.056
MAE2	0.273	0.091	0.19	0.016	0.025	0.032
MSE2	0.108	0.015	0.048	<0.001	0.001	0.002
MAPE2	0.035	0.081	0.022	0.046	0.066	0.07
MAE3	0.397	0.101	0.214	0.016	0.035	0.044
MSE3	0.21	0.019	0.066	<0.001	0.002	0.003
MAPE3	0.051	0.088	0.024	0.046	0.091	0.096

It's clear from Table 2 that the best model for the Strikeout time series is an ARIMA(0, 1, 0). This is interesting, as it doesn't rely on any previous observations or averages once the first-order difference is calculated. Essentially, the differenced time series is a random walk based only on the error term in the model.

Although the best model for the strikeout time series was very clear, this is not the case with the other five time series. Decisions need to be made based on forecast performance and the likelihood of the model. We do not display all the metrics for all possible models to reduce output. Instead, we display the metrics for our selected ARIMA models for each time series in Table 3.

Dynamic Linear Models

We next consider Dynamic Leader Models (DLMs). DLMs allow for better interpretation of parameters in the model than ARIMA models. We are particularly interested in this because our goal is to understand how the time series' we are investigating change over time. As we can see from the original plots of each time series, a clear trend exists in all of them. The trend seems fairly constant, so we believe a second-order polynomial DLM will be sufficient for most, if not all, of the six time series of interest. However, there could be some exceptions to this constant trend. For example, after the year 2010, the trend of the Strikeouts time series potentially increases up until 2021. Therefore, for a thorough analysis, we will consider first-order and third-order polynomial DLMs in addition to second-order polynomials. Once these DLMs are fit, we will select the best model for each time series based on BIC and one, two, and three step ahead forecast errors. We will also compare these metrics to the performance of the ARIMA models fit previously.

The results of three different polynomial order DLMs for each time series across ten metrics are listed below:

As shown in Tables 4-9, there is no consistent pattern in which polynomial order model performs the best. It is consistent that the first-order polynomial DLM always has the lowest BIC value for each time series.

Table 4: Strikeouts

	BIC	MAE1	MSE1	MAPE1	MAE2	MSE2	MAPE2	MAE3	MSE3	MAPE3
DLM(1)	-185.393	0.219	0.063	0.027	0.381	0.175	0.046	0.568	0.355	0.068
DLM(2)	-167.215	0.162	0.046	0.020	0.277	0.090	0.034	0.371	0.170	0.045
DLM(3)	-135.298	0.148	0.042	0.018	0.234	0.071	0.028	0.306	0.114	0.037

Table 5: Home Runs

	BIC	MAE1	MSE1	MAPE1	MAE2	MSE2	MAPE2	MAE3	MSE3	MAPE3
DLM(1)	-337.051	0.108	0.016	0.094	0.126	0.025	0.106	0.141	0.033	0.116
DLM(2)	-308.156	0.106	0.015	0.093	0.119	0.022	0.101	0.133	0.028	0.111
DLM(3)	-260.640	0.120	0.017	0.104	0.135	0.027	0.116	0.120	0.029	0.100

Table 6: Hits

	BIC	MAE1	MSE1	MAPE1	MAE2	MSE2	MAPE2	MAE3	MSE3	MAPE3
DLM(1)	-157.632	0.136	0.024	0.016	0.165	0.035	0.019	0.202	0.055	0.024
DLM(2)	-129.653	0.131	0.022	0.015	0.151	0.029	0.018	0.182	0.045	0.021
DLM(3)	-94.006	0.137	0.025	0.016	0.175	0.040	0.020	0.227	0.070	0.026

Table 7: Wild Pitches

	BIC	MAE1	MSE1	MAPE1	MAE2	MSE2	MAPE2	MAE3	MSE3	MAPE3
DLM(1)	-666.480	0.014	0	0.038	0.018	0	0.048	0.018	0	0.050
DLM(2)	-627.316	0.013	0	0.034	0.016	0	0.044	0.015	0	0.042
DLM(3)	-587.418	0.012	0	0.033	0.015	0	0.043	0.014	0	0.039

Table 8: Hit by Pitch

	BIC	MAE1	MSE1	MAPE1	MAE2	MSE2	MAPE2	MAE3	MSE3	MAPE3
DLM(1)	-679.175	0.023	0.001	0.057	0.032	0.002	0.079	0.042	0.003	0.104
DLM(2)	-655.137	0.022	0.001	0.055	0.028	0.001	0.074	0.042	0.002	0.110
DLM(3)	-612.009	0.023	0.001	0.059	0.027	0.001	0.068	0.037	0.002	0.092

Table 9: Sacrifices

	BIC	MAE1	MSE1	MAPE1	MAE2	MSE2	MAPE2	MAE3	MSE3	MAPE3
DLM(1)	-268.260	0.036	0.002	0.085	0.055	0.004	0.130	0.074	0.006	0.172
DLM(2)	-244.153	0.028	0.001	0.066	0.043	0.003	0.099	0.060	0.004	0.137
DLM(3)	-209.001	0.025	0.001	0.059	0.027	0.001	0.065	0.041	0.002	0.092

Table 10: Results of diagnostics test for normality and independence of residuals

	W	p-value	Chi-square	p-value
SO9	0.987	0.417	14.803	0.788
HR9	0.991	0.711	14.949	0.779
H9	0.989	0.620	16.551	0.682
WP9	0.979	0.120	37.254	0.011
HBP9	0.954	0.001	18.565	0.550
SAC9	0.920	0.002	18.253	0.571

This is due to the differences in likelihoods and is exacerbated by the increased complexity from higher order polynomial models and the penalty BIC puts on more parameters. However, the model with the lowest BIC does not always perform the best in terms of forecast. In our selection of the best model, we put a premium on forecasting error. If there existed little to no difference in performance in forecasting, we fell back on BIC as a decision maker. With that said, we selected a second-order polynomial model for each time series other than Strikeouts and Sacrifices. For Strikeouts and Sacrifices, we we picked a third-order polynomial model.

Once these were chosen, we compared the results to the ARIMA models chosen early to pick the best performing model for each time series. We originally thought that the DLM model would outperform the ARIMA model in all cases. When using BIC as a criterion, that is certainly true. But as stated previously, we put more emphasis on accuracy in forecasting. Therefore, for two of our time series we selected simple ARIMA models. Our selected models are listed below:

- Strikeouts: DLM(3)
- Home Runs: ARIMA(0, 1, 1)
- Hits: DLM(2)
- Wild Pitches: DLM(2)
- Hit by Pitch: ARIMA(0, 1, 0)
- Sacrifices: DLM(3)

Diagnostics

Now that we have our selected models, we need to perform some diagnostics to ensure that the assumptions these models need are satisfied. First we plot the ACF and PACF of the residuals from each model. This gives us a sense if there exists any correlation between the observations. As seen in Figure 4, for the most part there are no significant lags in both the ACF and PACF of each time series' residuals. The only exception to this is the Wild Pitches time series. There appears to be significant lags of 15 and perhaps 17. This actually lines up with what we found earlier in our analysis, that there may be some cyclical or seasonal behavior. At this point we are satisfied with these results, but we'd like to investigate further in the future.

In Table 10 we list the results of formal tests for normality and independence of error. For the normality assumption, we employ the Shapiro-Wilk test. For the testing of correlation among the residuals, we use the Box-Ljung test. For the most part, we are unable to reject the null hypotheses of normality and independence. However, Wild Pitches rejects the null hypothesis, indicating that there exists some correlation among the residuals. This is expected as we saw this in the ACF and PACF. Surprisingly, we also see rejections of normality for the fitted model for Sacrifices and Hit by Pitch. For the Hit by Pitch time series, we employed an ARIMA model. In the future we may go back and use a DLM or other ARIMA model if it satisfies the assumptions better. For the Sacrifices time series, if we instead use a second-order polynomial DLM, the assumptions are satisfied. This will result in a drop in forecasting performance, but the assumptions will be satisfied.

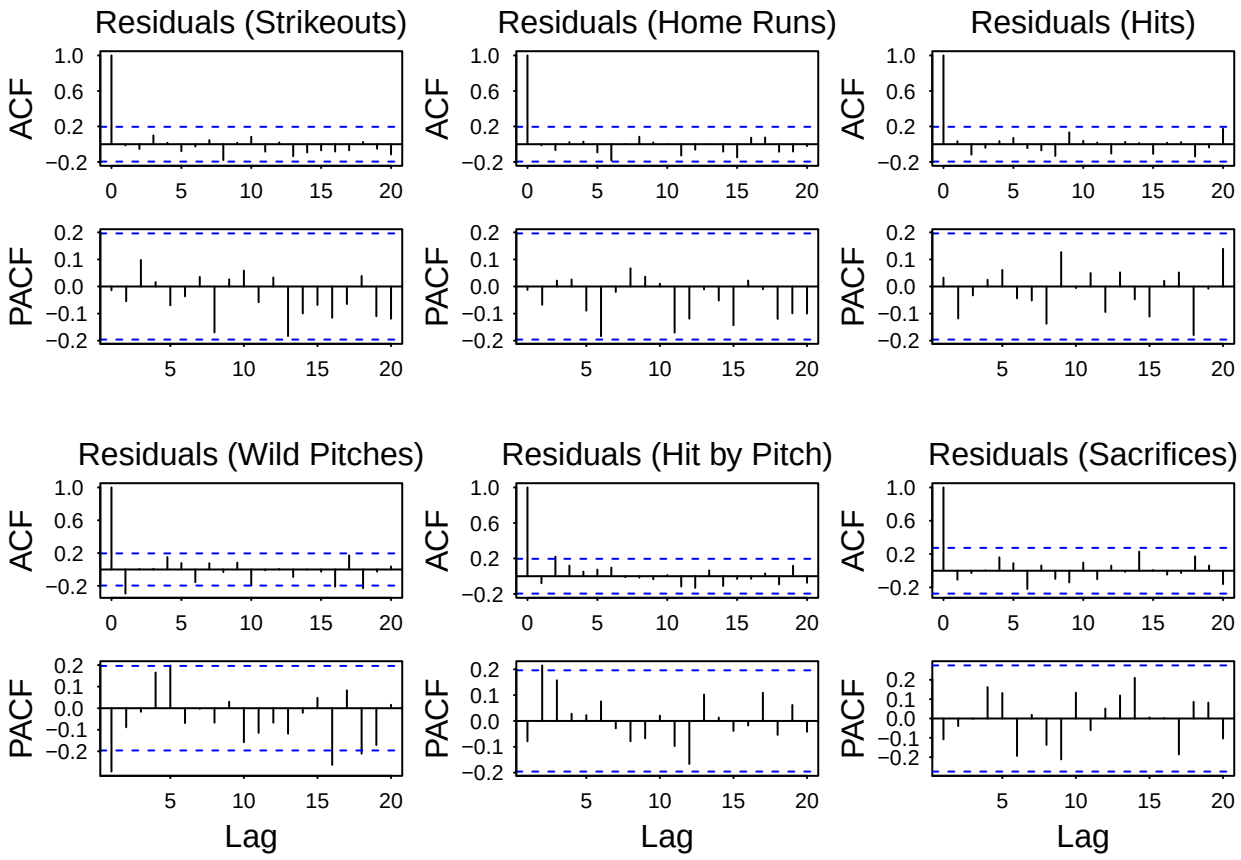

Figure 4: Diagnostic plots for correlation of residuals from fitted models

Forecasting

Next we calculate the 20-year forecast for each of our time series. These forecasts are displayed alongside the time series in Figure 5. As we can see, the difference in the forecasts for the Dynamic Linear and ARIMA models is quite stark. Because the DLMs account for the trend in the time series, that trend is present in the forecast. However, the ARIMA models only project the most recent value. This is likely due to the fact that the models for Home Runs and Hit by Pitch are very simple. The Home Runs time series is modeled using an ARIMA(0, 1, 1). There is no autoregressive component. The Hit by Pitch time series is modeled using an ARIMA(0, 1, 0). There is only a first-order difference taken, and so the resulting time series is a random walk.

We can see that the behavior of the confidence intervals of forecasts varies by time series. Some increase exponentially as time increases, such as Strikeouts and Sacrifices. However, the rest seem to be headed toward a constant interval after some time. We believe this is due to the magnitude of the trend in each time series. Strikouts are increasing at an increased rate, so the uncertainty is higher, whereas Wild Pitches has a much milder increase. Lastly, we see that Sacrifices are projected to go below zero. Obviously this does not make sense, so we should adjust our model to restrict the range of our response.

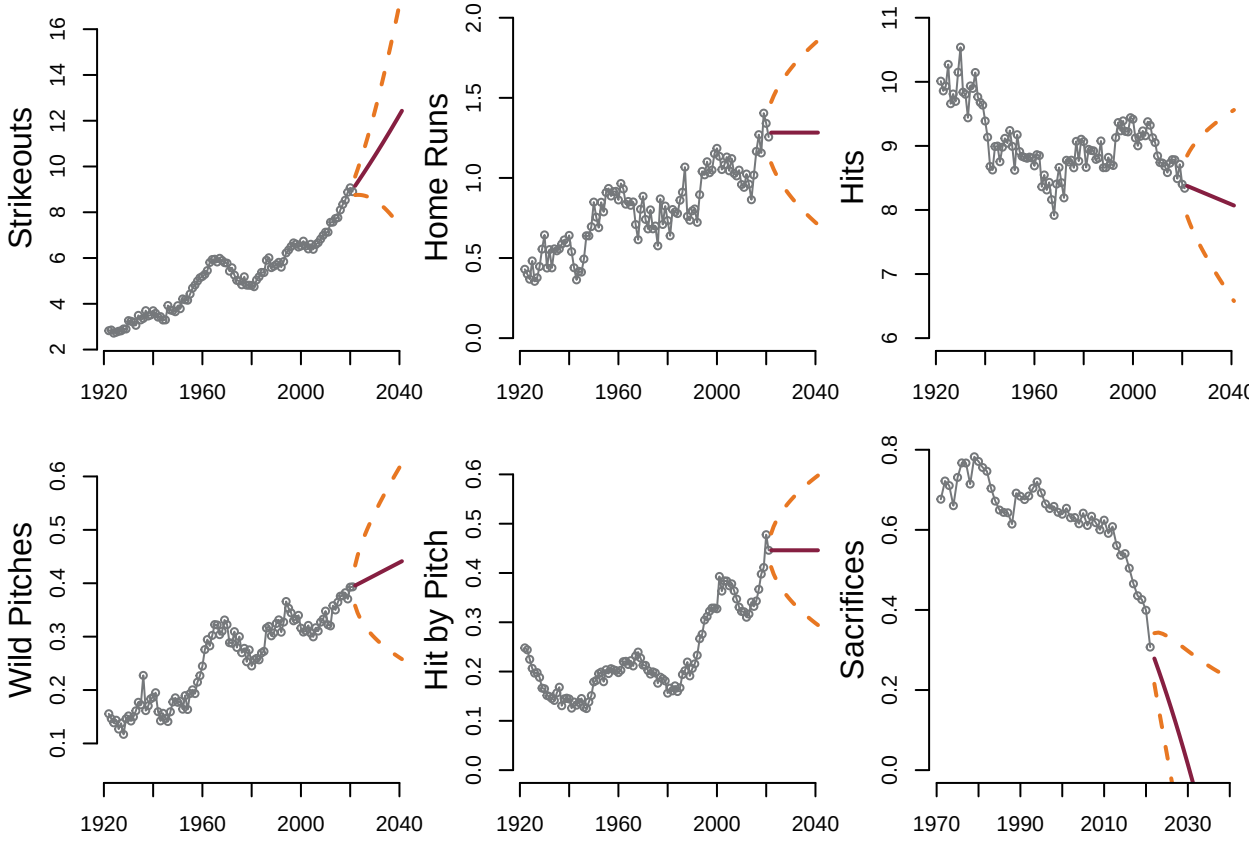

Figure 5: 20-year forecasts for each MLB time series

Discussion

It appears from our analysis that Major League Baseball has become less strategic over the years and more brute strength based. This is easy to see from the basic plot of each of these time series. Strikeouts, Home Runs, Wild Pitches, and Hit by Pitches (all per 9 innings) have increased over time. At the same time, Hits and Sacrifices (sacrifice hits & fly outs) (both per 9 innings) have decreased over the years.

Forecasting using the models we selected confirms the trend seen in the data. Forecasting indicates that Major League Baseball will continue to become less strategic over time and rely more and more on brute strength. In some instances, the trends are forecast to continue at an alarming rate. For instance, about 8-9 strikeouts currently occur over a nine inning game. However, if the trend continues as it has, within 15-20 years, over half of the outs in an average baseball game will come from strikeouts.

We strongly recommend that Major League Baseball investigate these changes and determine how to fix it. In fact, we know that Major League Baseball is already aware of this problem and are testing rule changes or game play procedures in order to address it. For instance, many rule changes are being introduced to Minor League Baseball as a testing ground to measure their effectiveness at resolving the issues. One rule change that has made its way to MLB is that every team begins with a runner on second base in extra innings. This prevents games from stretching to abnormal lengths and boring fans. However, we don't believe this addresses the real problem. Many of the issues we see likely stem from increased arm strength and pitch velocity. As such, we recommend Major League baseball attempt to move the pitching mound further from home plate in order to give hitters more control and ability to bat strategically.

Lastly, we believe that this analysis shows the usefulness of Dynamic Linear Models. The Hit by Pitch time series illustrates this. The best performing Auto Regressive Integrated Moving Average model was an ARIMA(0, 1, 0). All that was done was to compute a first-order difference that removed the non-stationarity from the time series. But no autoregressive or moving average coefficients were included. When we fit a DLM to the Hit by Pitch time series, the best performing model was a second-order polynomial model. A second-order polynomial DLM models the trend, and the level of the time series acts as a random walk. In fact, that is exactly what occurs in an ARIMA(0, 1, 0) model. Therefore, using a DLM in this case is incredibly helpful to understand the trend and how it varies, something that may be lost when restricted to an ARIMA model.

Future Work

This analysis is far from comprehensive and has only shown us how much more that can be done. As stated at the beginning of the report, this dataset contains over 100 statistics tracked over the last 150 years. This analysis only covers six of those. In the future we would like to investigate the following:

- Slugging percentage and On-base percentage: these metrics track how often a player gets on base and puts weight on what type of hit it is (e.g. double, triple). The increase in popularity of data analytics has had a major effect on these statistics. Therefore, we'd like to see what change occurred and what the forecast for these statistics holds
- Time-varying models: as we've seen, the trends of some of our statistics change throughout time. We've only scrateched the service of exploring time-varying components. We'd like to explore more sophisticated time-varying models that may fit our data better.
- Spatiotemporal models: all our analysis was of a temporal nature. We'd like to incorporate any spatial components into our models and projects. For instance, where the pitches predominantly fall in the strike zone most likely has evolved over time. Additionally, where balls land after hits on the field has likely also changed. There is great potential for insight from this data on how the game has changed and how it will likely continue to change.
- Wild Pitches: our analysis did not account for any cyclical or seasonal trend in any of our time series. We felt justified in this because of the lack of evidence and knowledge of the game. However, in the Wild Pitches time series, there appears to be some cyclical behavior. We'd like to explore that further.
- Other variables: there are potential factors that influence a time series outside of the time series itself. We would like to consider other variables that may be helpful in model fit and lead to improved forecasting.

References

- Lahman, Sean. "Lahman's Baseball Database." SeanLahman.com, 9 Mar. 2022, https://www.seanlahman.com/baseball-archive/statistics/.
- "MLB Stats, Scores, History, & Records." Baseball, https://www.baseball-reference.com/.