Time Series Final Project

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 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.

Initial data manipulation required aggregating the player level statistics to the league level. Some statistics present in the Baseball Reference database (e.g. batting average) were not present in Lahman's, so we had to do some calculations in order to maintain the consistency.

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.

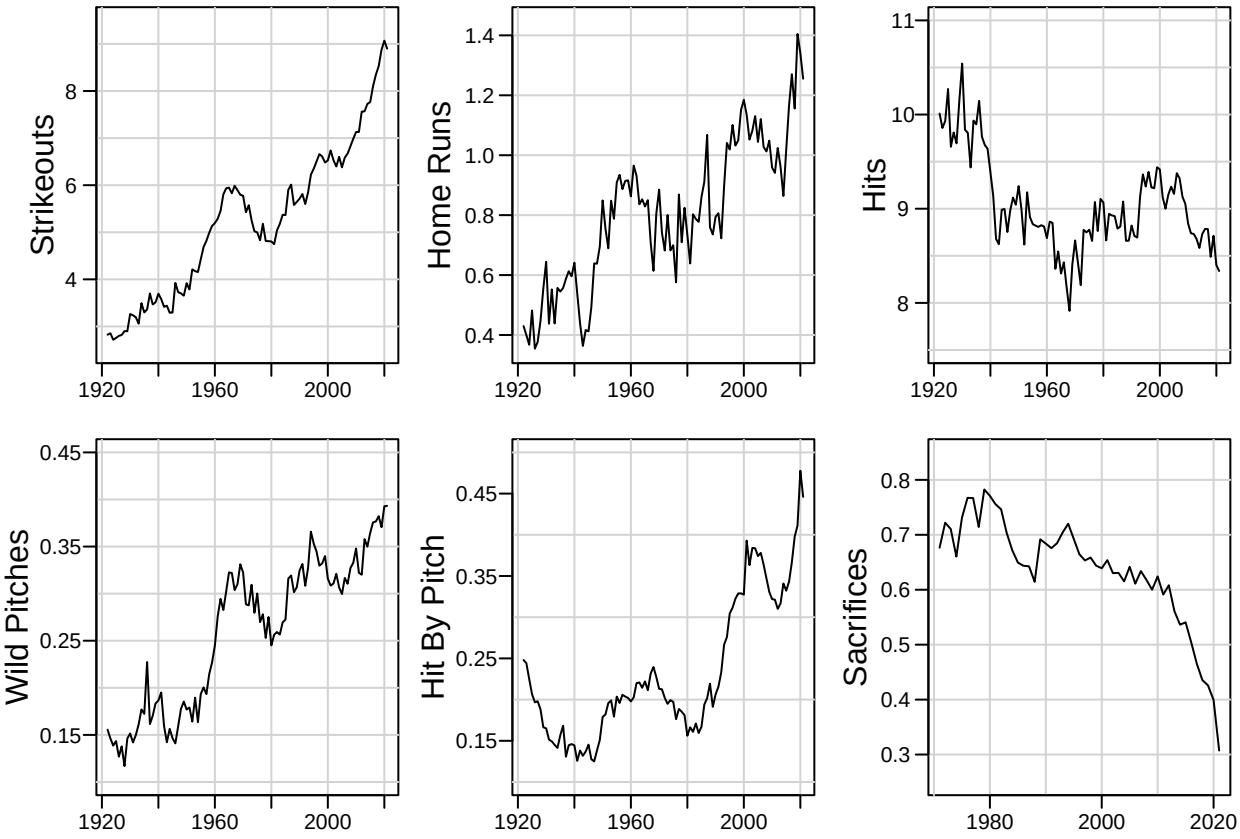
Theses six statistics are all tracked by count (e.g. number of hits in the league per season). However, it is common in sports analytics to look at count variables as averages or rates. This practice reduces extreme values (e.g. the 2020 MLB season was shortened due to COVID-19, resulting in a drastic 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 an organization begins.

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 have the following statistics for the years 1922-2021:

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

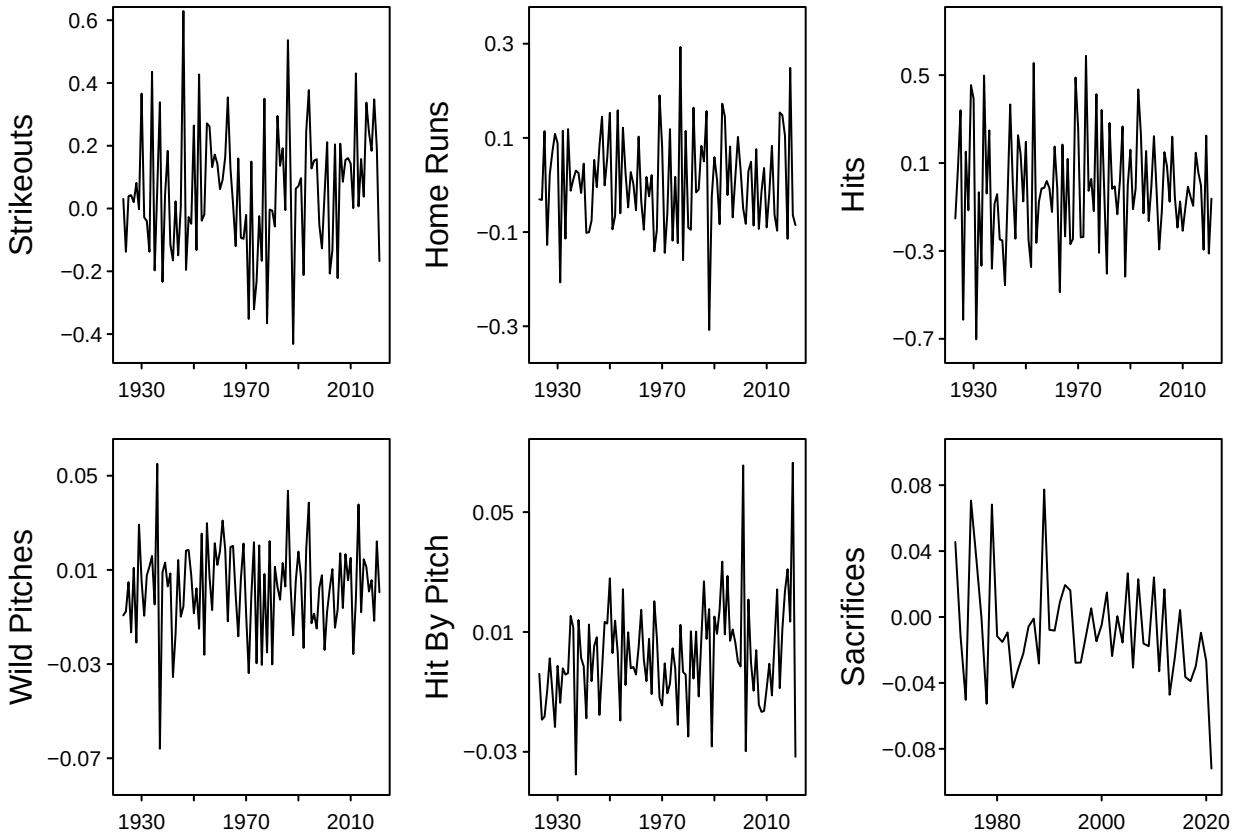
Model Selection



Above are displayed 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

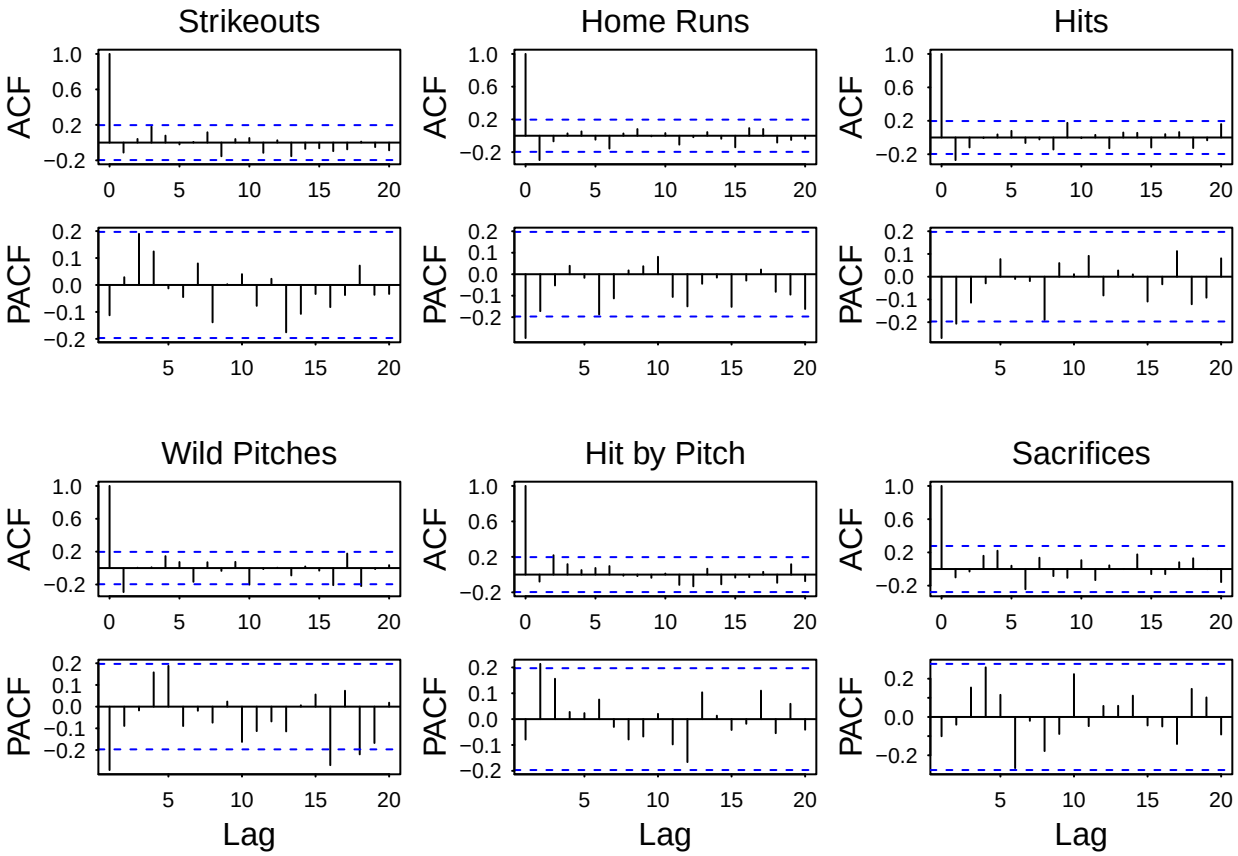
First, we will calculate a first-order difference on each time series in order to consider fitting Auto Regressive Integrated Moving Average (ARIMA) models. The first-order differencing is displayed below.



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 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.

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.



As shown above, 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 ACF and the PACF. As such, we will consider an ARIMA model with auto regressive order of 0 and 1 and 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 Strikeour 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

It's clear 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. Instead, we display the metrics for our selected ARIMA models for each time series below.

	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

Diagnostics

Forecasting

Discussion

Future Work

- Seasonal pattern in Wild Pitches