Time Series Analysis of Major League Baseball Organizations’ Fan Attendance

Joseph Molis

Follow this and additional works at: https://vc.bridgew.edu/honors_proj

Part of the Physical Sciences and Mathematics Commons

Recommended Citation
Copyright © 2023 Joseph Molis

This item is available as part of Virtual Commons, the open-access institutional repository of Bridgewater State University, Bridgewater, Massachusetts.
Time Series Analysis of Major League Baseball Organizations’ Fan Attendance

Joseph Molis

Submitted in Partial Completion of the
Requirements for Commonwealth Honors in Mathematics

Bridgewater State University

May 7, 2023

Dr. Wanchunzi Yu, Thesis Advisor Date: 5/8/2023
Dr. Kevin Rion, Committee Member Date: 5/8/2023
Dr. Irina Seceleanu, Committee Member Date: 5/9/2023
Time Series Analysis of Major League Baseball Organizations’ Fan Attendance

Abstract
Throughout baseball’s rich and long history, fans have been one of the most integral parts of the game. However, in recent years, baseball has seen a decrease in fans, allegedly due to the pace of play, or the length of games. Baseball games can take up to four hours to complete, and in today’s fast-moving society where all information is at one’s fingertips, it is believed that baseball’s slower pace turns people away from the game. However, how true is that? The primary goal of this project is to build models to accurately forecast fan attendance for every Major League Baseball (MLB) team. With teams across the country with varying levels of success, can fan attendance be accurately predicted using time series analysis using fan attendance data from the 1988 to 2019 season. This range allows us to analyze 32 years of data for total fan attendance per season for all thirty teams. The 2020 and 2021 seasons were not used as fans were not allowed in stadiums during these years due to COVID-19. After creating time series models using the data, forecasts were made for the 2022 and 2023 seasons to determine the accuracy of the models.

Introduction
Major League Baseball organizations are always trying to put a good product on the field for their fans. Whether the team is in last in their division, or a World Series contender, owners are trying to build a team that can compete. Or so people think. A lot of teams that struggle begin to trade away top players or lose them to other teams in free agency in an attempt to ‘rebuild’ their roster. There are many reasons to do this, such as large contracts underperforming, or just a long stretch of mediocrity. When teams lose fan favorite players, fans will no longer want to watch that team play. Studying fan attendance trends may lead to some answers as to why fans may not be going to stadiums to watch baseball. Through the creation of time series analysis, this research attempts to use fan attendance data dating from 1988 to 2019 to build models that can forecast fan attendance for the 2022 season and beyond. As the 2022 season has concluded, there is a full dataset that can be used to compare the accuracy of the forecasts to the real results.

Methodology
The purpose of this research is to create an accurate model that can predict fan attendance at Major League Baseball games. The process of conducting this research consisted of collecting and organizing data from multiple sources into files for all thirty MLB teams. Originally, the data collected included more than fan attendance, but during the process of analysis, fan attendance is what was focused on. The research and data collection began during summer 2022 as part of the Adrian Tinsley Program. Here data was collected on fan attendance, payroll, wins per season, and a stat called weighted runs created plus (WRC+). It was all organized into an interactive dashboard displaying all data with the ability to select a range of years to view created in Tableau. The data for the fan attendance was found on the Ballparks Of Baseball Website (Ballparks Of Baseball). This site keeps records for attendance at all thirty major league ballparks by decade, going back to 1890. The reason this database was chosen over more popular databases such as Baseball Reference, Fan Graphs, and ESPN, is because all the data was exportable to an excel file. To check
the credibility of the site, a cross reference with Baseball Reference was performed. All the data matched with Ballparks Of Baseball’s data, so the research was continued. Attendance data dating back to the nineteenth century was too much, and unneeded, so the target years of study were 1988 to 2019. This avoided the labor strike shortened season in 1981, as well the COVID-19 shortened seasons in 2020 and 2021. There were other labor strikes in the early 90s, but none cancelled many games. During the 2021 season, fans were allowed after about two months of play. It was decided that it was still too long of an absence to be a full set of data.

The next step in data collection involved organizing the data into an excel file for each team. This was a simple, yet tedious, process that was necessary for the analysis portion of the research. Doing this allowed for easy fixes while coding and cleaner code as each file was only accessed in one section of the code. The language I used to write the code was R. R is a language whose functionality caters towards statistical analysis. Another option was SAS, but R is simpler and easier to read. In R, time series graphs were created and a full analysis of each one was conducted to find the best fitting model. Each team has their own time series graph, ACF, and PACF plots. When the time series was nonstationary differencing was required. Residual tests and tests for normality were also performed. Residual ACF and PACF plots were made first. If these were unsatisfactory, then more transformations needed to be made on the data. Next, the AIC and BIC of each model were calculated, then normal QQ plots and lines were drawn, as well as a p-value was calculated. These last tests were the basis of the chosen models. Another aspect of the models being weighed was the simplicity of the model. For example, if an AR(2) model was created, but the ARIMA(3,1,1) model had a better AIC, BIC, and p-value, as long as the difference between the values was not too large, the AR(2) may have been chosen. This is due to it being a simpler model that is easier to interpret. R was able to do all of this, so no math and analyses needed to be done by hand. As the models were created, it was noticed that most of the models were very similar. No team’s data needed to be differenced more than once, and many were simple AR(1) or AR(2) models. The simplicity of the models was shocking as some teams had abnormal looking data, such as the Cleveland Guardians, whose graph is pictured below.

![Cleveland Guardians Dataset](image.jpg)

Although all models were slightly different, they were all autoregressive integrated moving average (ARIMA) models. Benefits of using the ARIMA model are that it handles varying patterns in the data. This includes linearity, volatility, and seasonal changes in the data. With thirty different
data sets being analyzed, a model that can perform the same functions on each set no matter the type of trend is extremely beneficial. ANOVA groups differences by variances, which is ideal for the data being examined. With differencing being performed, a model with balanced differencing is ideal. The data was not uniform across all thirty teams, so differencing using the variance for each individual team is ideal. Other teams had polarizing data, such as the Washington Nationals, who had a very apparent split in their data and graph. The Nationals are the most recent team to join Major League Baseball, as they migrated south from Montreal. The organization changed its name from the Montreal Expos to the Washington Nationals in 2005. Due to this split, the time series graph and model were difficult to create. It was decided to split the data into the Expos and Nationals to create the models more easily. However, since the Expos are no longer an organization, when the forecasting was done, their model was excluded and only the Nationals model (from 2005 to 2019) was used. This model was the only model that needed to be split up into different eras due to such a large disparity in the data.

The final portion of the model creation process was the forecasting of fan attendance in 2022 and 2023. The 2022 season has concluded but forecasting the 2022 season and comparing the results is a necessary part of determining if the models are appropriate. In R, the output provides an exact value for the prediction, as well as an 80% confidence interval and a 95% confidence interval. The forecast also provides data for four years in the future, however, since the data ends in 2019, the first two rows are for 2020 and 2021. There is no interest in those years, so the data being analyzed are the last two rows of this output. An example of the output from the Boston Red Sox is shown below.
Organization of the forecast data was similar to the fan attendance data. Each team had a row in an excel sheet that included the lower and upper bounds for the 80% confidence interval, the 2022 total attendance taken from Baseball Reference (Baseball Reference) and the forecast for the 2023 season. The 80% confidence interval was chosen over the 95% because it is a narrower range. A narrower range better shows the accuracy of the models. The results of the forecasts will be discussed in greater detail later. Some teams had negative bounds, which proved to be an issue. After examining the code and data file, it was an error with the inputs for the model, and not bad data causing the issues. There were five teams that did not make the intervals predicted for them. Further analysis was conducted for those teams by looking at the team’s WRC+ and win totals for the 2020 and 2021 seasons that were not included in the model. This was to find a possible explanation as to why the fan attendance fell out of the predicted bounds.

**Results**

The results of the study were surprising as the vast majority of teams’ fan attendance for the 2022 season fell in 80% confidence interval bounds. To be exact, 24 teams were in the predicted range. This shows that despite having no fans the previous two seasons, most teams were able to retain fans over the pandemic shortened seasons. Rather than focusing on all the teams that fell in the bounds, focusing on the teams that did not fall into their respective ranges is a better indicator of what exactly brings fans to the ballpark.

The five teams that were not as successful in bringing in fans were the Arizona Diamondbacks, Chicago Cubs, Cincinnati Reds, Oakland Athletics (A’s), and the Pittsburgh Pirates. The fifth team was the San Diego Padres, but they are different from these other four and will be discussed later. Out of these four teams, the A’s and Pirates were the two lowest team payrolls in MLB, while the Diamondbacks were ranked twenty fifth, and the Cubs were ranked fourteenth. All four teams had winning percentages under 50%, with the A’s and Pirates also posting the two lowest winning percentages (Baseball Cube) The time series graphs for attendance and payroll for Oakland is below, as well as the bar graph of the number of wins per season.
Oakland’s payroll and fan attendance were actually slowly increasing throughout the 2019 season, as well as they were well above the 50% win rate mark. So the question is what happened to them? During the COIVD-19 seasons, they were still above a 50% winning percentage, but following that season, they decided to trade their two star players, Matt Olson to the Atlanta Braves, Matt Chapman to the Toronto Blue Jays, and their top pitcher Chris Bassitt signed as a free agent with the New York Mets. Along with these roster moves, they traded their next best pitcher Frankie Montas to the New York Yankees halfway through the 2022 season. All of this is in an attempt to rebuild their team as, although they were successful, they never made it deep into the playoffs, or even missed the playoffs in those years. As a result, their payroll plummeted, and their number of wins also plummeted. In Sports Illustrated’s 2023 ballpark rankings, they ranked Oakland’s Oakland Coliseum the second worst ballpark in the MLB (Sports Illustrated). Combining their lack of star power, payroll, and sufficient facility for fans, it makes sense as to why the Oakland A’s are struggling to bring in fans. Their current record in 2023 is 3-15, last in the American League West.

The model created for Oakland was an AR(2) model. No differencing was required to make this model, as the graph did not have any sudden increases or decreases. The AIC for the model was 870.96 and the BIC was 876.82. Both of these values seemed high, but are in line with the AIC and BIC of the other models. With a p-value at 0.825, the clear model choice for Oakland was the AR(2). Although other models had better AIC and BIC’s, the simplicity of the AR(2) model, along with the higher p-value made the choice easy. Just to solidify the decision, the normal qq plot was fantastic with all points sitting on or next to the line. The 85% confidence interval for their fan attendance was 1,414,220 to 2,375,960. The final number at the end of the season was a mere 787,902. This is just over half of the lower bound the model provided. In the 2021 season, fans were not allowed at ballparks until the end of May and beginning of June. This is 2 months into the season. They had 701,430 fans for that season, almost meeting for the full 2022 season. They are predicted to have a total of 1,906,736 fans in attendance for the 2023 season, a difficult number to achieve with their recent performance, as that is more than double their 2022 total. They are on pace to surpass the 2022 total, already having 132,308 fans come to games, despite their slow start to the season.

In contrast to the A’s, the Pirates were on a slight downward trend in attendance and payroll for a few years going into the 2020 season. They are attempting a rebuild process similar to the A’s, but started it a few years sooner, finishing last or second to last in the National League (NL) Central division every year since 2017 (Pittsburgh Baseball Reference). During those years, they
traded away stars like pitcher Gerrit Cole and outfielders Andrew McCutchen and Starling Marte. This lack of success combined with a payroll that is extremely low, there is an explanation as to why fans may not want to come to the ballpark to watch their team play, despite having the best ballpark in the MLB according to Sports Illustrated (Sports Illustrated). The Pirates’ current record in 2023 is 11-7, second in the NL Central. It is a very small sample size, but if this success can be maintained, the Pirates may have fans returning to PNC Park soon.

Pittsburgh’s model was also an AR(2) model, although not nearly as strong as Oakland’s. The AIC and BIC were higher at 900 and 905.86, respectively. Also, the p-value was lower at 0.723. The normal qq plot was also not nearly as strong, with a lot of the points at the beginning of the graph and end of the graph tailing away from the line. The Pirates did not miss the mark of their interval as bad as Oakland did. Their range is 1,324,749 to 2,221,894. The finished the season with a total of 1,257,458 fans, approximately 65,000 fans short of the lower bound the model provided. Unlike Oakland, the Pirates will not have to double their fans to reach their 2023 projection of 1,785,760 fans. They currently have a total of 101,182 fans at games this season.
The Arizona Diamondbacks fan attendance dropped a noticeable amount from 2008 to 2009, but since then, their rate is steady for the next decade. There are no big drops or spikes in the data. However, their payroll fluctuated greatly during those years. There was a big drop from 2014 to 2015 but has been increasing since. During those years, their winning percentage was well below 50%, but progressively got better, peaking in 2017 with 93 wins. Over the pandemic shortened seasons, the Diamondbacks really struggled, posting 25 wins in a 60-game season in 2020 and 52 wins, tied for the worst record in MLB that season, in 2021. The model created for the Diamondbacks was a ARIMA(2,1,1) model. They are the only team with a moving average parameter in the model. This is one of the more complicated models out of the thirty. It required two autoregressive parameters and 1 MA, as stated previously, along with one difference. The AIC and BIC for the model were 530.83 and 534.61, respectively. These are a lot lower than most teams because the Diamondbacks were not founded until 1998, 10 years after the start of the data. The normal qq plot was decent, with a lot of points on the line in the middle of the data, but a lot at the beginning not near it. There was a strong p-value of 0.712, as well. The AIC and BIC were the lowest that were achieved through the multiple models that were tested, and the p-value was strong, so this model was chosen as the best. The interval provided by the Diamondbacks’ model is 1,705,691 to 2,567,382. They missed the lower bound by about one hundred thousand fans with an end of season total of 1,605,199 fans. The 2023 projection from the model have the Diamondbacks bringing in 2,136,469 fans, a huge leap from the 2022 total.
The Diamondbacks never lost any key players on their roster. Starling Marte left in free agency, and they traded away pitcher Zack Greinke, but the core of the team remains intact. The case for the Diamondbacks decline is simply a lack of performance on the field. Currently, the Diamondbacks are 11-7 and first in the NL west division, bringing in 166,593 fans.

The Cincinnati Reds’ story is much like the previous three teams. They had a successful last decade, but towards the end started to fall behind other organizations and decided to rebuild. From their peak fan attendance in 2015 to 2019, fan attendance has dropped. Payroll dropped from 2015 to 2017 but has only increased since then. Surprisingly, despite being their best year for fan attendance, the Reds’ had their worst year on the field, collecting only 64 wins.

The model for the Reds is the simplest possible, an AR(1) model. The AIC and BIC are similar to the other models at 886.6 and 891. The normal qq plot is strong, as well, with every point being on the line. The measure that weakens the model is a small p-value at 0.465. The ideal p-value is 0.5 or higher. Although the p-value is not ideal, with the AIC and BIC appropriate in relation to the other models, the simplicity of the model, and the very strong norm qq plot, AR(1) seemed the most appropriate. Other models were tested, but either the AIC and BIC would increase, or the norm qq plot would become worse, and the p-value would not increase.
significantly. The model predicted that the Reds would have between 1,742,378 and 2,393,433 fans in 2022. They missed the lower bound by approximately 380,000 fans. The model forecast predicts that they will have 2,082,077 fans in 2023. They are currently 7-11 on the season and last in the NL central division. Although that is not a great start, they have a total fan attendance of 203,687 so far, the highest of the teams discussed so far.

The final team that did not meet their projected attendance is the Chicago Cubs. They differ from the past three teams, as they ranked fourteenth in payroll in 2022. They also differ as they have made the playoffs multiple times from 2015 to 2019, including winning the World Series in 2016. However, in the three seasons after that, the fan attendance has decreased by about 35,000, 17,000, and 90,000 in 2017, 2018, and 2019 respectively. Similarly, the number of games the Cubs have won has also decreased. Meanwhile, the payroll has only increased since 2016, although they are still well above a 50% winning percentage.

The analysis for the Cubs resulted in an ARIMA(1,1,0) model. The data needed to be differenced once, as there were big spikes and dips in the time series plot. The AIC was 860.99 and the BIC was 863.86. These values are not alarming as they are close to the AIC and BIC of the A’s and Pirates. The normal qq plot was strong, but not as convincing as Oakland’s. The first
and last points stray away from the line, as well as some points in the middle, but it curves right back to the line. The p-value is the measure that was the most convincing to choose this model. The p-value was a very strong 0.953. This measure, along with the model being relatively simple, made the model selection easy. Forecasting for the cubs was a lower bound of 2,653,239 and an upper bound of 3,573,823. The final 2022 total was 2,616,780 fans, just a mere 49,000 short. Of the five teams that did not meet the prediction, the Cubs were the closest. The 2023 forecast for the Cubs is also the largest of the group at over three million fans and is the fourth highest forecast of the thirty models.

In 2021, they only won 71 games, sending the organization who had too much money invested in players that were not performing, into a rebuild process similar to the Pirates and A’s. They lost fan favorite players such as Kris Bryant, Anthony Rizzo, Javier Baez, and Kyle Schwarber. The core of the 2016 championship team had been disassembled through free agency, trades, and retirement. The team that brought excitement to Cubs fans was gone and is a possible explanation as to why less fans are coming to Wrigley Field. In the short sample size of the 2023 season, the Cubs are currently 11-6, second in the NL central, and have reached 275,953 fans (Cubs Baseball Reference).

The final team whose fan attendance did not fall between the forecasted bounds is the San Diego Padres. They are an interesting case as their fan attendance was greater than predicted. The previous five teams all fell below, but the Padres are different. As previously mentioned, the A’s, Reds, Diamondbacks, Pirates, and Cubs were all on the decline in winning percentage leading into the 2022 season. They all entered a rebuild phase due to a lack of performance, high payrolls, or lack of exciting star players. The Padres are the opposite. They had the fifth highest payroll in 2022 at over 175 million dollars (Baseball Cube). They have only added to that this past offseason, ranking third for 2023 at 249 million (FanGraphs Roster Resource). They brought in all star players like Manny Machado, Juan Soto, Josh Hader, and Xander Bogaerts to play alongside their homegrown talent in Fernando Tatis Jr.

Looking at the time series graphs for fan attendance and payroll, both had peaks in 2015. From 2019 to 2015, fan attendance had been steady for the Padres and payroll had been increasing. After 2015, however, both payroll and fan attendance dropped back down to normal levels. Fan attendance remained steady, but the payroll quickly increased to near 2015 levels in 2017. During those years, they never finished above third place in the NL west division. So far, the story is similar, but the key difference is that the Padres did not rebuild their team. Ownership decided to
invest in good players and lock up talent with large and long contracts. This investment into success has led the Padres to draw in 2,987,470 fans in 2022 and an 89-win season with an appearance in the National League Championship Series (NLCS). This is about 600,000 more fans than the last complete season in 2019.

The model for the Padres is a simple AR(2) model, although it is on the weaker side. The AIC and BIC are higher than the typical AIC and BIC from other models. The values are 911.01 and 916.87, respectively. The p-value is strong at 0.872, which shows there is strength in the model. The norm qq plot is different than most others, as the beginning of the graph is centralized around the line, but the tail end of the graph really departs from it. An AR(1) was tested but the norm qq plot did not change much, and the AIC and BIC got larger. Due to this, the AR(2) was chosen. The forecast for the Padres had a lower bound of 1,662,514 fans, and an upper bound of 2,778,112 fans. As previously mentioned, the Padres beat that with a total of 2,987,470 fans. They are projected to bring in 2,189,841 fans in 2023. They currently have a
record of 9-11 and are in third place in the NL west. They have had an astounding 524,456 fans come to the ballpark in just twenty games to start the year. The Padres are the best example of how to attract fans to their team through star power and success. Their ownership and management want to build a winning team that can maintain a high-level of play for years to come while keeping the fans engaged.

**Conclusion**

Although no proper analysis was performed on the relationship between payroll, success, and fan attendance, through the six teams discussed, a pattern has seemed to emerge. Teams that invest in talent and perform on the field will draw more fans than teams that decide to tear down the team in an effort to restructure the roster to be more cost efficient. The models created in this research were successful in forecasting fan attendance data. Out of the thirty models, only six failed to fall within the predicted interval for the 2022 season. Of those six, five fell below the interval and one above. Four of the teams that fell below were bottom ten in payroll for 2022, and the other has just started to decline after multiple years of high-level success. As for the remaining 24 teams, the common denominator between all of them is that they either are successful enough to bring in fans or have an exciting roster with star players that people want to watch play. The teams that have the most fans attending games have a combination of both, which is what every team is striving for. At its core, Major League Baseball teams, and any professional sports teams, are in the entertainment business and what will draw fans to their product is the success of the team and excitement to see star players on the field.
Works Cited


Keywords

Baseball
Time Series
Fan Attendance
Models
Forecasting