Eras in baseball: Change-Ups and Change Points: An Exploration of Baseball's Historic Eras

Mena Whalen
Department of Mathematics and Statistics
Center for Data Science and Consulting
Loyola University Chicago
Chicago, IL 60660
mwhalen3@luc.edu

Brian M. Mills
College of Education
University of Texas at Austin
Austin, TX 60660
brian.mills@austin.utexas.edu

Gregory J. Matthews
Department of Mathematics and Statistics
Center for Data Science and Consulting
Loyola University Chicago
Chicago, IL 60660
gmatthews1@luc.edu

Abstract

Baseball is some weird and wild shit.

Keywords: change point analysis, baseball,

1 Introduction

The first professional baseball team in the United States, the Cinicnnati Red Stockings, was formed in 1869 (Rothenberg (n.d.)). Many leagues came and went in the late 1800s, but National League (NL), formed in 1876, emerged as the predominant league of the time. A few decades later, the American League (AL) began growing in popularity and eventually reached an agreement with the NL to be the two major leagues of baseball with the winner of each league playing in the World Series starting in 1903.

Throughout the history of baseball in the United States, the game has gone through many changes and distinct eras. For example, the time period between approximately 1900-1919 is often referred to as the "Dead Ball Era" and was marked by low scoring games and dominant pitching. Another more recent example would be the "Steroid Era" which lasted from approximately 1994 through 2005 and was characterized by a rapid increase in power hitting largely attributed to players using performance enhancing drugs.

More specifically, Woltring and Jubenville (2018) mentions six eras of modern baseball": "Baseball has endured much change over the course of its history, and because of constant change, the modern era of baseball has been segmented into six distinct sub-eras. A common list presented at Baseball-Reference described the eras as the Dead Ball Era (1901-1919), the Live Ball Era (1920-1941), the Integration Era (1942-1960), the Expansion Era (1961-1976), the Free Agency Era (1977-1993) and the Long Ball/Steroid Era (1994-2005)." Woltring and Jubenville (2018) notes that they name a seventh era after 2006, which they term the Post Steroid Era.

While many baseball writers have attempted to define the different eras of baseball, there has also been some academic work that has sought to empirically define eras in baseball. Groothius, Rotthoff, and Strazicich (2017) looked for structural breaks in univariate times series of performance measures over the period from 1871-2020. They analyzed four statistics: slugging percentage (SLG, home run (HR) rate, batting average (BA), and runs batted in (RBI) rate. For each of these statistics, the computed the mean and standard deviation (SD) across all players who had at least 100 at bats in a given season to yield a univariate time series for each of these statistics. They then used the Lagrange Multiplier (LM) unit root test proposed in J. Lee and Strazicich (n.d.) to find structural breaks. They identified structural breaks in slugging percentage in 1921 and 1992, the first of which marks the end of the Dead Ball Era and the latter corresponding to the start of the steroid era.

Lee and Fort (2005) looks for structural changes in competitive balance of the two league American and National. Use methods from Andrews1993, Bai1997, 1999 and Bai and Perron 1998 and 2003 to look for break points between 1901 -1999. Theuy measure competitive balance in two ways: 1) Noll (1988) and Scully (1989) and 2) Lee 2004. They find break

point sin competitive balance in 1912, 1926, and 1933 for the NL and in 1926 and 1957 in the AL.

Baseball is not the only sport where this type of analysis has been applied. I. (2004) looked for structural changes in soccer using data from British soccer leagues through 1996. They had data from division I (Premier League) and II starting in the late 1800s, and data from lower divisions III and IV from just after WWII starting in 1947. They identify a number of change points. Notably they identify a change point in the mean of margin of victory in 1925 related to the change in the definition of offsides (changed from 3 players to 2 players), an change points in the variability of number of goals in the early 1980s and 1992 corresponding to the change in number of points for a win (i.e. went from 2 points to 3 points) and a change in the backpass rule, respectively. Fort and Lee (2007) looked for structural breaks in major North American sports other than baseball (i.e. basketball (NBA), hockey (NHL), and American football (NFL)). They identified a number of change points related to competitive balance in each sport that often, though not always, correspond to league expansion, league mergers, or other major events in a sport (e.g. increased number of foreign players in the NBA in the late 1990s/early 2000s).

All of the previous work mentioned here focuses on change point analysis in ***univariate*** time series. However, recent methodolgical developments in change point analysis allow for change point analysis in ***multivariate*** time series, which is the focus of this current manuscript. This work leverages techniques such as the Double CUSUM Binary Segmentation algorithm (H. Cho (2016)) and the Sparsified Binary Segmentation algorithm (H. Cho and Fryzlewicz (2014)) to look for change points in Major League Baseball at two main levels. First, we look for structural changes in the league as a whole across teams to empirically define different eras in baseball. Second, we look for change points within a team to determine eras of a team. This second analysis can be used to identify the beginning and end of so called "dynasties", periods of sustained excellent performance by a team. In addition, we can also identify the opposite, sustained periods of poor performance.

Y. H. Lee and Fort (2008): Attendance and the Uncertainty-of-Outcome Hypothesis in Baseball. Identify Break points in attendance in 1918 and 1945 for both leagues. For AL only: 1963, 1978, 1994. For NL only: 1976. Using Bai and Perron From table 3 in Mills and Fort 2014 MILLS and FORT (2014): LEAGUE-LEVEL ATTENDANCE AND OUTCOME UNCERTAINTY IN U.S. PRO SPORTS LEAGUES. Looks at NHL, NFL, NBA. Rottenburg 1956 looked at baseball. Identify break points in NBA, NFL, and NHL. Table 3 Mills and Salaga (2015): NCAA Basketball: League balance using stats.
Mills and Fort (2018): team-level: Attendence.
Salaga and Fort (2017): College football

SOME MORE STUFF

2 Methods

H. Cho and Fryzlewicz (2014) and H. Cho (2016)

R Pacakge: Haeran Cho and Fryzlewicz (2018)

A commonly employed method for detecting changes over time involves the CUSUM statistic in binary segmentation, which creates a tree diagram, dividing the time series after successfully identifying a change point until no further change points are detected. The magnitude of the CUSUM difference between two segments generally indicates a potential change point, depending on the assumptions and statistical tests applied to the time series. While this process works well for a single time series, detecting change points across a panel of multiple time series requires a more comprehensive approach. H. Cho (2016) introduced the double CUSUM (DC) algorithm, assuming there are n time series, all of the same length T, with unknown common change point locations. This algorithm utilizes the CUSUM values of multiple time series ($j = 1, \ldots, n$) to cross-sectionally compare time segments for potential change points.

To achieve this, Cho's DC statistic involves ordered CUSUM values for a given time location, favoring larger values as potential change point candidates. It incorporates a partial thresholding of the smaller ordered CUSUM values, effectively representing a scaled average of the most prominent values at a given location across all "n" time series. Maximizing the DC statistic across all time series and potential time locations results in a test statistic for detecting a change point. This test statistic is then compared against a specific testing criterion, denoted as $\pi^{\phi}_{n,T}$, leading to the partitioning of time series and the formation of a tree structure based on the detected change points.

To determine the testing criterion $\pi^{\phi}_{n,T}$, a Generalized Dynamic Factor Model (GDFM) bootstrapping algorithm is employed, accounting for potential correlations within and between high-dimensional time series. This methodology enables the detection of second-order change points, utilizing methodology from H. Cho and Fryzlewicz (2014) Haar wavelet periodograms and cross-periodograms, as opposed to the traditional CUSUM statistic. For further details, refer to H. Cho (2016) and H. Cho and Fryzlewicz (2014).

2.1 Data

We obtained multi-year baseball statistics from GREG??. Over the course of the sport's history, numerous statistics have been collected, and for our analysis, we focused on year-end statistics pertaining to the teams. To ensure consistency in our methodology, it was necessary for each team to have existed from 1900 until the end of the dataset in 2020. This timeframe of 120 years, denoted as T, encompassed a total of 16 franchise teams, namely

NYY, BOS, LAD, ATL, CHW, CHC, CIN, CLE, DET, BAL, SFG, OAK, PHI, PIT, STL, and MIN. The statistics of interest for these teams include runs, hits, home runs, balls, strikeouts, at-bats, stolen bases, number of games played in a season, attendance, runs against, hits against, home runs against, balls against, and strikeouts against. However, a few statistics, such as balls, stolen bases, strikeouts, and attendance, had missing data for certain years. To address this, we employed multiple imputations using classification and regression trees (cite mice/cart) to fill in the missing values and ensure consistent data across all analyses.

2.2 Eras and Dynasties

2.2.1 Eras

Since we are interested in finding where in time these different eras of baseball have occurred, we first examine each baseball statistic of interest as their own analysis for all 16 teams. Those baseball statistics are home runs, strikeouts, balls, stolen bases, runs, and attendance which were all standardized to the number of games played in a season since we are interested in seeing how the different statistics change from season to season. Using the methodology from Cho and Fry (CITE) a thresholding parameter was determined from bootstrapping and then used to find where change points in mean and variance existed in the panel of teams data of a given baseball statistic. To examine the idea of changes throughout teams and types of statistics (initially), for each season the average baseball statistic was found from all the teams that year and then an average time series was created of each of the types of statistics of interest. Then the change point methodology was performed on those 6 time series of average statistic types.

###Dynasties

Continuing our investigation, we explored the concept of "dynasties" within modern baseball teams, irrespective of their length of a teams existence. A dynasty, in this context, refers to a period characterized by significant success or potentially hardships. Such a dynasty is identified when a team experiences a noticeable change across all their statistics in a given season, whether it be in a positive or negative direction. This approach aligns well with our methodology, as it involves examining panels of time series data encompassing all statistics for a singular team. The statistics of interest for this analysis include runs, hits, home runs, balls, strikeouts, runs against, hits against, home runs against, balls against, and strikeouts against. To ensure comparability, these statistics were standardized using the season average and standard deviation derived from all teams present in each respective season. By standardizing the statistics, we capture a team's offensive and defensive performance. Notably, if a team exhibits a high and positive offensive performance, it can be balanced

to zero by large and negative defensive statistics when standardized. Therefore, if a team excels in both aspects of the game, it would have numerous time points above zero, whereas underperforming teams would display many below-zero values.For each team, we examined their ten statistics, analyzing both mean and variance changes to identify potential change points. By scrutinizing these indicators, we aimed to uncover shifts in a team's performance and assess their significance within the context of dynasties.

3 Results

3.1 Tables

Mean/Variance CPT	Franchise	Change Point Year
mean	ANA	1977
mean	ATL	1900, 1990
mean	BAL	1959, 1999
mean	BOS	1918, 1937
mean	CHC	1891
mean	CHW	1970
mean	CIN	1918, 1952
mean	CLE	1944, 1959, 1993
mean	DET	1958, 1988
mean	FLA	2011
mean	HOU	1999
mean	KCR	1997
mean	LAD	1940, 1961
mean	MIL	1977, 1996
mean	MIN	1959
mean	PHI	1917, 1949
mean	PIT	1939
mean	SDP	1977
mean	SEA	1986, 1999
mean	SFG	1903, 1973
mean	STL	1937
mean	TBD	2007
mean	TEX	1973, 1990
mean	WSN	2011
variance	ATL	1887

Mean/Variance CPT	Franchise	Change Point Year
variance	BOS	2008
variance	CHW	1969
variance	CLE	1966, 1981, 1993
variance	DET	1942
variance	HOU	2010
variance	NYM	1974
variance	PIT	1902, 2001
variance	SDP	1977
variance	SFG	1918

Mean/Variance CPT	Baseball Statistic	Change Point Year
mean	att_per_game	1945, 1975, 1992
mean	bb_per_game	1927
mean	hr_per_game	1920, 1946, 1967, 1993
mean	r_per_game	1919, 1940
mean	sb_per_game	1919
mean	so_per_game	1956, 1993, 2009
variance	att_per_game	1944, 2008
variance	hr_per_game	1927, 1944, 1995, 2007
variance	r_per_game	1935
variance	sb_per_game	1919, 1966
variance	so_per_game	1990

NYY 1918: Babe Ruth got to the Yankees in 1920. And starting in 1919 they had exactly one losing seasons between 1919 and 1965.
BOS 1917, 1929, 1948, 1994: Their last world series win in the 1900s was in 1918.
I don't know 1929. 1948 there was a big jump in runs?
1994: Strike year.

LAD: 1919, 1955

ATL 1918 1949

CHW: 1919: Blaack Sox Scandal

CHC: 1919 1940 1940: War.

Cin: 1929 1952

CLE 1920 1939 Cleveland won the world series in 1920. Major change in offensive output.

DET 1927 1940

BAL 1919 1942

SFG 1919 1937

OAK 1920 1936 1951

PHI 1918 1937

PIT 1920 1946

STL 1919 1941

MIN 1919 1942

Pearl Harbor was 1941. So US was in war in 1942.

Acknowledgements

We thank Michael Lopez for suggesting we do "something with change point analysis".

Supplementary Material

All code for reproducing the analyses in this paper is publicly available at https://github.com/menawhalen/baseball_cpt

References

Cho, H. 2016. "Change-Point Detection in Panel Data via Double CUSUM Statistic." *Electronic Journal of Statistics* 10: 2000–2038.

Cho, Haeran, and Piotr Fryzlewicz. 2018. *Hdbinseg: Change-Point Analysis of High-Dimensional Time Series via Binary Segmentation.* https://CRAN.R-project.org/package=hdbinseg.

Cho, H., and P. Fryzlewicz. 2014. "Multiple-Change-Point Detection for High Dimensional Time Series via Sparsified Binary Segmentation." *JRSSB* 77: 475–507.

Fort, and Lee. 2007. "Structural Change, Competitive Balance, and the Rest of the Major Leagues." *Economic Inquiry* 45 (3): 519–32.

Groothius, Peter A., Kurt W. Rotthoff, and Mark C. Strazicich. 2017. "Structural Breaks in the Game: The Case of Major League Baseball." *Journal of Sports Economics* 18 (6): 622–37.

I., Palacios-Huerta. 2004. "Structural Changes During a Century of the World's Most Popular Sport." *Statistical Methods and Applications* 12: 241–58.

Lee, J., and M. Strazicich. n.d. "Minimum Lagrange Multiplier Unit Root Test with Two Structural Breaks." *The Review of Economics and Statistics*, no. 4: 1082–89.

Lee, Y. H., and R. Fort. 2008. "Attendance and the Uncertainty-of-Outcome Hypothesis in Baseball." *Review of Industrial Organization* 33 (4): 281–95.

Lee, and Fort. 2005. "Structural Change in MLB Competitive Balance: The Depression, Team Location, and Integration." *Economic Inquiry* 43 (1): 158–69.

Mills, Brian M., and Rodney Fort. 2018. "Team-Level Time Series Analysis in MLB, the NBA, and the NHL: Attendance and Outcome Uncertainty." *Journal of Sports Economics* 19 (7): 911–33. https://doi.org/10.1177/1527002517690787.

Mills, Brian M., and Steven Salaga. 2015. "Historical Time Series Perspectives on Competitive Balance in NCAA Division i Basketball." *Journal of Sports Economics* 16 (6): 614–46. https://doi.org/10.1177/1527002515580925.

MILLS, BRIAN, and RODNEY FORT. 2014. "LEAGUE-LEVEL ATTENDANCE AND OUTCOME UNCERTAINTY IN u.s. PRO SPORTS LEAGUES." *Economic Inquiry* 52 (1): 205–18. https://doi.org/https://doi.org/10.1111/ecin.12037.

Noll, R. G. 1988. "Professional Basketball." *Stanford University Studies in Industrial Economics*, no. 144.

Rothenberg, Matt. n.d. *PRO BASEBALL BEGAN IN CINCINNATI IN 1869.* https://baseballhall.org/discover/pro-baseball-began-in-cincinnati-in-1869#:~:text=The%20Cincinnati%20Red%20Stockings%20made,professional%20baseball%20club%20in%201869.

Salaga, S., and R. Fort. 2017. "Structural Change in Competitive Balance in Big-Time College Football." *Review of Industrial Organization* 50: 27–41.

Scully, G. W. 1989. *The Business of Major League Baseball.* University of Chicago Press.

Woltring, Rost, M., and C. Jubenville. 2018. "Examining Perceptions of Baseball's Eras: A Statistical Comparison." *The Sport Journal.* https://thesportjournal.org/article/examining-perceptions-of-baseballs-eras/.