International Journal of Performance Analysis in Sport

ISSN: 2474-8668 (Print) 1474-8185 (Online) Journal homepage: http://www.tandfonline.com/loi/rpan20

A predictive model for analysing the starting pitchers' performance using time series classification methods

César Soto-Valero , Mabel González-Castellanos & Irvin Pérez-Morales

To cite this article: César Soto-Valero , Mabel González-Castellanos & Irvin Pérez-Morales (2017): A predictive model for analysing the starting pitchers' performance using time series classification methods, International Journal of Performance Analysis in Sport, DOI: 10.1080/24748668.2017.1354544

To link to this article: http://dx.doi.org/10.1080/24748668.2017.1354544

Published online: 01 Aug 2017.

INTERNATIONAL JOURNAL OF PERFORMANCE ANALYSIS IN SPORT, 2017
https://doi.org/10.1080/24748668.2017.1354544

A predictive model for analysing the starting pitchers' performance using time series classification methods

César Soto-Valero[a], Mabel González-Castellanos[a] and Irvin Pérez-Morales[b]

[a]Department of Computer Science, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba;
[b]CIMCNI, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba

ABSTRACT

Pitcher's performance is a key factor for winning or losing baseball games. Predicting when a starting pitcher will enter into an unfortunate pitching sequence is one of the most difficult decision-making problems for baseball managers. Since 2007, vast amounts of pitch-by-pitch records are available for free via the PITCHf/x system, but obtaining useful knowledge from this huge amount of data is a complex task. In this paper, we propose a novel model for analysing the performance of starting pitchers, determining when they should be removed from the game and replaced by a reliever. Our approach represents pitch-by-pitch sequences as time series data using baseball's linear runs and builds an instance-based model that learns from past experience using the k-Nearest Neighbours classification method. In order to compare time series of pitcher's performance, Dynamic Time Warping is used as the dissimilarity measure in conjunction with the Keogh's lower bound technique. We validate the proposed model using real data from 20 Major League Baseball starting pitchers during the 2009 regular season. The experimental results show a good performance of the predictive model for all pitchers; with values of Precision, Recall and F1 near to 0.9 when the outcomes of their last 10 throws are unknown.

ARTICLE HISTORY
Received 15 June 2017
Accepted 10 July 2017

KEYWORDS
Baseball; starting pitcher; performance analysis; time series classification; DTW; k-NN

1. Introduction

Baseball is a complex and unique sport. This is mostly because of its specific and discrete game-play structure, which allows recording and manipulating a huge amount of statistical data during each game (Wolf, 2015). Major League Baseball (MLB) is recognised as one of the most important professional sports organisations around the world. The popularity of baseball in the USA and Canada has transformed the MLB in a multimillionaire business (Peach, Fullerton, & Fullerton, 2016).

Predicting the future performance of baseball players based on their historical records and statistics is a very active field of research today because of its numerous advantages for managerial and decision-making. Since Lewis (2004) published his best-seller "Moneyball" in 2003, sabermetric (recognised as the science of learning about baseball through the use of the empirical evidence obtained) has been gaining more and more interest in the sports analytics community (Albert, 2010b).

CONTACT César Soto-Valero cesarsotovalero@gmail.com

In baseball, pitching is considered a very difficult skill to learn. Many experts believe that it is an essential component for obtaining victories (Pavitt, 2011). One of the most complex decision problems that baseball managers have to handle during games consists in deciding when a fatigued, or faltering, pitcher should be removed from the game and replaced by a reliever. This is a decision-making problem, which is made more difficult by the fact that the substitute pitcher needs approximately 10 min to "warm up" before he can enter the game. Furthermore, the reliever should not warm up for an excessive length of time because he could become too exhausted. There is not a definitive formula for making this decision correctly, and managers rely on various heuristics (e.g. pitch count, game score, own experience or intuition) to decide the exact moment when a starting pitcher should be relieved (Keeley, Oliver, Torry, & Wicke, 2014; Zimniuch, 2010).

In the last years, some models have been proposed for analysing and estimating pitcher's performance. For example, Piette, Braunstein, McShane, and Jensen (2010) study the reliability and consistency of various statistics to evaluate the effectiveness of pitchers in the MLB using a Bayesian Random Effect model, Sidhu and Caffo (2014) explore pitchers' decision-making by modelling the at-bat information (pitch selection and counts) as a Markov Decision Process, while Hoang, Hamilton, Murray, Stafford, and Tran (2015) introduce a novel adaptive strategy using Linear Discriminant Analysis to predict binary pitch types (Fastball vs. Non-fastball).

Sabermetric has proven that taking advantage from historical statistics and past data could reveal important patterns in many baseball scenarios. Accordingly, it is viable to use past pitches records and outcomes to learn and predict the future pitcher's performance (Chih-Cheng, Yung-Tan, & Chung-Ming, 2014). For this aim, we propose modelling and analysing pitch-by-pitch data as time series data.

Time series analysis is an active research area because it comprises a vast field of applications (Gooijer & Hyndman, 2006). In the last decade, time series data mining (Fu, 2011), and especially time series classification, have been gaining particular attention (Batista, Hao, Keogh, & Mafra-Neto, 2011; Flesca, Manco, Masciari, Pontieri, & Pugliese, 2007). However, to the best of our knowledge, no attempts of analysing and predicting pitcher's performance using time series classification methods have been made in the literature.

The main contribution of this paper consists of presenting a predictive model for determining when a starting pitcher is about to falter using time series classification methods. With this aim, we model pitch events as time series data by assigning different scores to each possible pitcher's thrown outcome. Once the pitch time series have been created, the k-Nearest Neighbours (k-NN) classification algorithm is then used for predicting the future performance (which is a binary result labelled as "High Performance" or "Low Performance"). In order to compare time series properly during the k-NN processing, Dynamic Time Warping (DTW) has been selected as the distance measure for k-NN. Furthermore, the lower bounding technique of Keogh and Ratanamahatana (2005) is also integrated in all DTW calculations to speed up the predictive algorithm.

With the purpose of evaluating the performance of our model, we conduct experiments involving all pitches throws for a total of 20 MLB starting pitchers during the 2009 regular season using the data provided by Albert (2010a). We reduce the length of the testing time series, from 5 to 50 throws, in order to evaluate the reliability of the prediction obtained.

The F1 measure and 10-fold cross-validation method are selected as the main criteria for evaluating the global performance of the model.

The rest of this paper is organised as follows. In Section 2, we describe our method for modelling pitch-by-pitch outcomes as time series data. Section 3 offers details about the predictive model proposed as well as some theoretical considerations related to this specific time series classification problem. In Section 4, we present a characterisation of the dataset and the experimental framework used to perform the model validation process. Sections 5 and 6 present the results obtained and offer a discussion about the model applicability and future issues, respectively. Finally, in Section 7, we give some conclusions about this work.

2. Data modelling

2.1. Run value of pitches based on linear runs

Weighted On-Base Average (wOBA) is a baseball statistic created by Tom Tango for measuring the hitters overall offensive contribution to their teams. It represents an empirical measure based on the relative values of each possible offensive event (Tango, Lichtman, & Dolphin, 2007). wOBA is considered as one of the most complete offensive statistics in baseball, combining all the different aspects of hitting into one metric and weighting each outcome in proportion to their actual run value. While, the traditional Batting Average (AVE), On-Base Percentage (OBP) and Slugging Percentage (SLG) fall short in accuracy and scope, wOBA measures and captures offensive value more accurately and comprehensively.

Equation (1) shows the general formula to calculate wOBA. First, it is necessary to find the specific weights of each offensive event in the season (weights are denoted as α, β, γ, δ, ϵ and θ) and then multiply these weights by the batter's unintentional bases on balls (UBB), singles (1B), doubles (2B), triples (3B) and homeruns (HR). These weights change annually and it is possible to find the specific wOBA weights for every year from 1871 to the present in the Fangraph website[1]. Next, dividing that number by the sum of his at bats (AB), walks (BB) excluding intentional walks (IBB), hit by pitches (HBP) and sacrifice flies (SF), that is the wOBA of the batter for the season.

$$\text{wOBA} = \frac{\alpha \cdot \text{UBB} + \beta \cdot \text{HBP} + \gamma \cdot \text{1B} + \delta \cdot \text{2B} + \epsilon \cdot \text{3B} + \theta \cdot \text{HR}}{\text{AB} + \text{BB} - \text{IBB} + \text{SF} + \text{HBP}} \tag{1}$$

From the pitcher's perspective, it is viable to tabulate the average of wOBA values at each step in the count and then convert them into a run value for a strike or a ball in any count. Assigning run values to a strike at each step in the count is not a novel concept. For example, using data from Tom Tippet's Diamond Mind Baseball, Burley (2004) calculated AVE, OBP and SLG values at each count after a ball and after a strike. Then, using linear weights he obtained the run value associated with a ball or strike in each count.

Tables 1 and 2 show the negative and positive run values (from the pitcher's perspective) of almost anything that could happen to a pitch after the ball leaves the pitcher's hand. Using the wOBA of every ball or strike count, we subtracted the league average wOBA (in 2009, that value was 0.329) from each count to determine how much above or below average the count affect wOBA. Then, using those wOBA values we determined how many runs were added or subtracted in every possible count according to the pitch outcome. For example, if a strike is thrown in a two-strike count, then the resulting wOBA for the batter

Table 1. Negative linear run value for each pitch outcome based on wOBA (regular season of 2009).

	New count							
Count	Ball	Single	Double	Triple	Homerun	Error	HBP	Walk
0–0	−0.038095	−0.508140	−0.888141	−1.058143	−1.438144	−0.508143	−0.368142	NA
0–1	−0.024620	−0.550985	−0.930985	−1.100985	−1.480985	−0.550985	−0.410985	NA
0–2	−0.031358	−0.602511	−0.982512	−1.152512	−1.532512	−0.602512	−0.462512	NA
1–0	−0.060137	−0.470044	−0.850045	−1.020045	−1.400045	−0.470045	−0.330045	NA
1–1	−0.058002	−0.526365	−0.906365	−1.076365	−1.456365	−0.526365	−0.386365	NA
1–2	−0.040629	−0.589903	−0.969903	−1.139903	−1.519903	−0.589903	−0.449903	NA
2–0	−0.106522	−0.409907	−0.789908	−0.959908	−1.339908	−0.409908	−0.401325	NA
2–1	−0.100481	−0.468363	−0.848364	−1.018364	−1.398364	−0.468364	−0.328364	NA
2–2	−0.096860	−0.549274	−0.929274	−1.099274	−1.479274	−0.549274	−0.409274	NA
3–0	NA	−0.303386	−0.683386	−1.069499	−1.233386	−0.303386	−0.368819	−0.163386
3–1	NA	−0.367882	−0.747883	−0.917883	−1.297883	−0.367883	−0.227883	−0.227883
3–2	NA	−0.452414	−0.832414	−1.002414	−1.382414	−0.452414	−0.312414	−0.312414

Table 2. Positive linear run value for each pitch outcome based on wOBA (regular season of 2009).

	New count						
Count	Strike	Strikeout	Bunt out	Fly out	Groundout	Line out	Pop out
0–0	0.04285	NA	0.26186	0.26186	0.261859	0.261859	0.261859
0–1	0.05153	NA	0.21901	0.21901	0.219014	0.219014	0.219014
0–2	0.01261	0.16749	0.16749	0.16749	0.167488	0.167488	0.167488
1–0	0.05632	NA	0.29996	0.29996	0.299955	0.299955	0.299955
1–1	0.06354	NA	0.24363	0.24363	0.243634	0.243634	0.243634
1–2	NA	0.1801	0.18010	0.18010	0.180096	0.180096	0.180096
2–0	0.05846	NA	0.22868	0.36009	0.360092	0.360092	0.360092
2–1	0.08091	NA	0.30164	0.30164	0.301636	0.301636	0.301636
2–2	NA	0.22073	0.22073	0.22073	0.220725	0.220725	0.220725
3–0	0.06450	NA	NA	0.46661	0.466613	0.466613	0.466613
3–1	0.08453	NA	0.40212	0.40212	0.402117	0.402117	0.402117
3–2	NA	0.31759	NA	0.31759	0.317585	0.317585	0.317585

is 0.000 (or a positive run value of 0.16749 for the pitcher). Thus, a strikeout in a two-strike count transitions the batter from his starting wOBA in the two-strike count to a wOBA of 0.000. Similarly, if a ball is thrown in a three-ball count then this transitions the batter to a walk (or a negative run value of 0.312414 for the pitcher). Notice that a strike thrown in a two-strike count decreases this run value in a different way than a strike thrown in a one-strike count. Thus, if the batter is up three-ball, but grounds out, then the pitch that created the groundout gets more credit than if it had grounded out in a two-strike count.

2.2. Time series of pitching performance

A time series T can be defined as a sequence $(t_1, \ldots, t_i, \ldots, t_n)$ of n data points measured typically at successive time intervals. The inherent temporal ordering of time series makes its study distinct from other common data analysis problems, in which there is no natural ordering of the observations. A time series model will generally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time series models will often make use of the natural one-way ordering of time. As a consequence, values for a given period will be expressed as deriving in some way from past values, rather than from future values.

We propose to model pitch-by-pitch data sequences as time series data using the run value of each thrown outcome. For this aim, we implemented a new metric, which is inspired by the previous work of Sidran (2005), to evaluate pitcher's performance during the baseball game. The metric, called Linear Run Pitcher's Performance (LRPP), extracts information from previous pitch outcomes. It consists of an accumulated pitcher's score, which is based on the linear run outcome of every pitch thrown (see Tables 1 and 2). The LRPP metric has a numeric output, in a format which allows us to create a graphical representation of the actual pitcher's performance in any moment of the game.

Let U be the set of defined scores corresponding to each possible pitcher's throw outcome p_i. Tables 1 and 2 show the U values used to calculate LRPP for the 2009 season, positive and negative outcomes are scored according to its impact on the result of the game. In other words, scores are defined in a way that allows balancing the output probabilities of all possible plate appearances.

LRPP scores can be intuitively modelled as a time series data, and the idea is as follows. At the beginning of the game (instant $i = 0$) the pitcher initiates with a score of performance $p_0 = 0$. Then, the associated score p_i in U of each pitch i is added to the accumulated LRPP(i) value. The Equation (2) resumes this procedure.

$$\mathrm{LRPP}(i) = \sum_{n=0}^{i} p_i \tag{2}$$

During each pitcher's throw, his total score is updated and saved in order to construct a time series of his performance during the game. For predictive purposes, we labelled the pitcher's performance in the moment i as a High Performance (HP) or Low Performance (LP), according to the following function:

$$\mathrm{Performance}(i) = \begin{cases} \mathrm{HP} \text{ if } \mathrm{LRPP}(i) \geqslant 0, \\ \mathrm{LP} \quad \text{otherwise} \end{cases}$$

As an example, Figure 1 represents two different time series constructing from the Justin Verlander performance, for the Detroit Tigers, in games played during the season of 2009. In the game labelled as HP, his LRPP score increased steadily until the pitch number 75 and then decreased just a few points, but finished with a very positive score of 2.45 points. In the game labelled as LP, the performance of Mr Verlander is considered positive up to his pitch number 76, but then he began to falter until he was removed in the sixth inning and replaced by a substitute pitcher. In that moment his LRPP score was of –2.34 points.

3. Methods

3.1. Classification of pitch time series data

Time series classification is a traditional data mining task which has attracted great interest in the data mining community, finding applications in several domains such as medicine, finance, entertainment and industry (Gooijer & Hyndman, 2006; Keogh & Kasetty, 2003; Nanopoulos, Alcock, & Manolopoulos, 2001). In the sport sciences context, time series classification is a field of research still under development. It has been applied mostly for

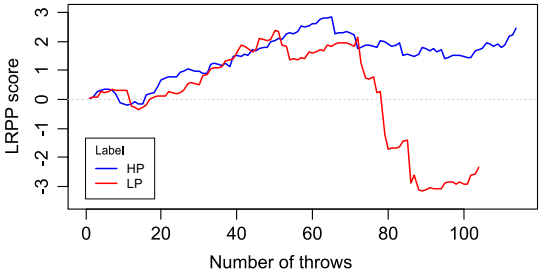

Figure 1. Graphical representation of pitcher's performance as time series data.
Note: The time series are labelled according to the value of its LRPP score when the starting pitcher leaves the game.

analysing video data in some specific sports such as table tennis (Maeda, Fujii, Hayashi, & Tasaka, 2014) and baseball (Fleischman, Roy, & Roy, 2007).

Time series classification is a supervised learning problem, where the objective is to predict the class membership of time series as accurately as possible (Xing, Pei, & Keogh, 2010). All the time series classification approaches first build a classification model based on labelled time series. In this case, "labelled time series" means that it uses a training dataset with correctly classified observations or time series sequences for some model building. Then, the built model is used to predict the label of a new unlabelled observation or time series.

Geurts (2001) illustrates one possible idea and the necessary steps to perform time series classification accurately. The first essential step is to find local properties and patterns from the series. In a second step, these patterns are combined to build classification rules using classification and machine learning methods.

Among many methods that can be used for this problem, the group of nearest neighbour classifiers has the simplest classification idea: to assign a new time series object or time series sequence to the most common class among its neighbourhoods. As indicated by the name, the k-NN classifier takes the k nearest neighbours into account.

Due to its effectiveness and simplicity, our model uses the 1-NN classification method as base learner for prediction. It is based on learning by analogy, that is, by comparing a given time series of pitcher's performance with others in order to find the most similar to it. For example, given an unlabelled pitch time series P and a training set of labelled time series $S = (s_1, \ldots, s_n)$, it searches for the series in S that is more similar to P using some comparison criterion. This time series is the "nearest neighbour" of P in S.

The combination of the 1-NN classification algorithm with DTW as the dissimilarity measure and comparison criterion has proven to be exceptionally accurate in practice and very difficult to beat in the time series domain (Wang et al., 2013). Furthermore, this method is parameter-free and does not require feature selection and discretisation. We provide more details about our predictive model in the next subsections.

3.2. Comparison of pitch time series data

As we comment in the previous section, for time series classification problems the 1-NN classifier with DTW as dissimilarity measure has shown to outperform most of the other distance measures (Wang et al., 2013). DTW overcome the weakness of Euclidean

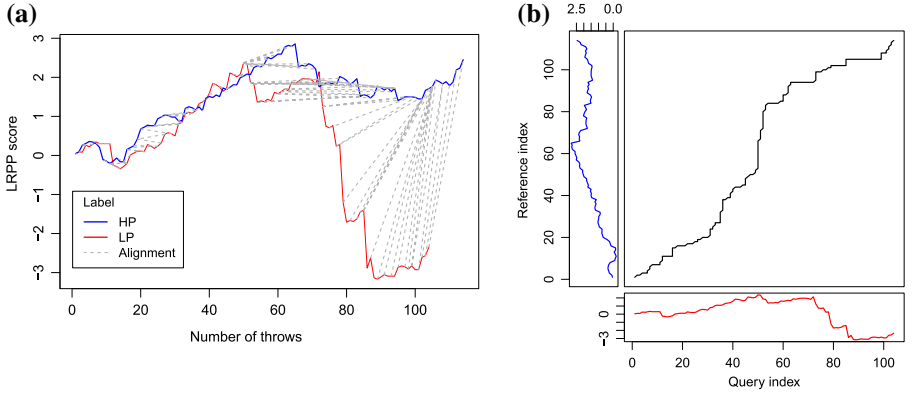

Figure 2. Aligning two pitch time series using the DTW similarity measure. (a) Aligning indexes between a query (red) and a reference (blue) time series. (b) The warping path constructed as a result of the alignment.

metric in measuring the similarity between time series, where time phases of different series are distinct. Our predictive model implements DTW as the comparison criterion for measuring the similarity among pitch time series.

DTW implements dynamic programming to find an optimal warping path between two time series sequences. To calculate the path value, it first creates a distance matrix, where each element in the matrix is a cumulative distance of a minimum of three surrounding data points. Let $P = (p_1, \ldots, p_i, \ldots, p_n)$ and $P' = (p'_1, \ldots, p'_j, \ldots, p'_m)$ be two pitch time series. First, we create an n x m matrix, where each (i, j) element $\delta_{i,j}$ of the matrix is defined as shown in Equation (3).

$$\delta_{i,j} = |p_i - p'_j|^d + \min\{\delta_{i-1,j-1}, \delta_{i-1,j}, \delta_{i,j-1}\} \tag{3}$$

Here, $\delta_{i,j}$ is the summation of $|p_i - p'_j|^d$ and a minimum cumulative distance of three elements surrounding the (i, j) element, and d is the dimension of L_p-norms (typically $p = 2$ for the time series domain). Then, when all elements in the matrix are filled, the DTW measure represents the total cost of the alignment and is determined from the last element $\delta_{n,m}$ of the matrix.

Figure 2 shows the alignment between the two pitch time series of Figure 1 using DTW. As this example illustrates, DTW compares effectively both time series even when they have different lengths and are out of phase. In this example, the global cost of the alignment is 205.04.

Although DTW outperforms many other distance measures, it is known to consume a huge computational cost with a time complexity of $O(n^2)$. Due to this situation, the lower bound of Keogh has been proposed to speed up the similarity search (Keogh & Ratanamahatana, 2005). The LB_Keogh(P, P') between the query time series P and a candidate time series $P' = (p'_1, \ldots, p'_i, \ldots, p'_n)$ can be computed by the following function:

$$\text{LB\_Keogh}(P, P') = \sum_{i=1}^{n} \begin{cases} |p'_i - u_i| & \text{if } p'_i > u_i, \\ |l_i - p'_i| & \text{if } p'_i < l_i, \\ 0 & \text{otherwise} \end{cases}$$

where $u_i = \max\{p_{i-r_i}, \ldots, p_{i+r_i}\}$ and $l_i = \min\{p_{i-r_i}, \ldots, p_{i+r_i}\}$ are envelope elements calculated from a global constraint $R = (r_1, \ldots, r_i, \ldots, r_n)$. Ratanamahatana and Keogh (2005) report that LB_Keogh allows them to prune out over 90% of all DTW computations on several datasets.

3.3. Predictive algorithm

According to our pitch-by-pitch data modelling, the length of pitch time series increases dynamically with each new throw. It is clear that as far as the length of the testing time series becomes larger then more information about the current performance of the pitcher could be used for testing and better the prediction will be. Thus, our model allows predicting the future performance of the pitcher, while the game is in progress.

Due to the characteristics of this prediction, it is reasonable to weight more recent information over older information in the time series. We accomplished this issue during the computation of DTW, multiplying the cost of the alignment between each pair of point by a weighed factor. Let $(a_1, \ldots, a_i, \ldots, a_n)$ and $(a'_1, \ldots, a'_i, \ldots, a'_n)$ be the values of the aligned points between the time series P and P' settled by DTW, then the Equation (4) defines a weighted DTW function.

$$\mathrm{wDTW}(P, P') = \sum_{i=1}^{n} \frac{i \cdot (|a_i - a'_i|)}{n} \tag{4}$$

The prediction is based on the pitcher's outcomes corresponding to similar situations from the past. Our model implements 1-NN and uses wDTW as the similarity measure in conjunction with the Keogh's lower bound to find a nearest neighbour. Detailed pseudo-code of the method is presented in Algorithm 1.

Algorithm 1 Predicts pitcher's performance using 1-NN and DTW as similarity measure. Implements Keogh's lower bound to speed up the computation process.

```
1: function PREDICTPERFORMANCE(P, S)
     P: A pitch time series for prediction.
     S: Training set of pitch time series.
2:     minDist ← ∞
3:     minLB ← ∞
4:     class ← ∅
5:     while (S ≠ ∅) do
6:         p ← any element in S
7:         distLB ← lbKeogh(P, p)
8:         if (distLB < minLB) then
9:             dist ← wDTW(P, p)
10:            if (dist < minDist) then
11:                class ← getClass(p)
12:                minDist ← dist
13:            end if
14:            bestLB ← minLB
15:        end if
16:        S ← S \ {p}
17:    end while
18:    return class
19: end function
```

The lower bound between the query pitch time series P and a candidate time series p is computed in line 7. If the lower bound is sufficiently large then the DTW is not computed (see line 8). The function wDTW in line 9 measures the distance of the optimal alignment between P and p. If this value is lower than minDist then P is labelled (HP or LP) accordingly with the most similar pitch time series p in S.

4. Procedures

4.1. Dataset

Since the season of 2007, Sportvision's PITCHf/x system has recorded (in real time) detailed data about each pitch that is thrown during each MLB game (Fast, 2010). This data are available for free from the MLB GameDay website[2] and includes details about every pitch thrown, as well as the outcome of the plate appearance associated with each pitch. The PITCHf/x system has resulted in a huge amount of fine-grained data, which has proven to be especially useful for pitching trainers, sports analysts and fans of baseball worldwide.

Data were directly obtained from Albert (2010a). He collected pitch-by-pitch data using the PITCHf/x system for 20 starting pitchers that played in the 2009 season. This represents the regular season games in which these pitchers participated as starters (649 in total). Table 3 shows information about the number of games played and the classification of their performance in the moment they left the game according to our pitch data modelling criterion (297 classified as HP and 352 classified as LP). The average number of throws per game was 101. Nine of those pitchers (marked with *e*) are considered among the elite pitchers since each received or was nominated for the prestigious Cy Young pitching award.

Table 3. Summary of the classification of performance for the 20 starting pitchers considered in this study, in the moment when they left the game.

Pitcher	Mean ± SD of throws per game	Classified as HP	Classified as LP	Total games
Brett Anderson	93.80 ± 14.60	13	17	30
Bronson Arroyo	103.24 ± 13.98	11	22	33
Scott Baker	98.73 ± 11.45	16	17	33
Joe Blanton	104.87 ± 8.29	10	21	31
Scott Feldman	91.15 ± 23.78	15	19	34
Gavin Floyd	99.37 ± 14.10	14	16	30
Zack Greinke[e]	106.48 ± 10.02	21	12	33
Roy Halladay[e]	106.00 ± 14.71	13	19	32
Cole Hamels	97.38 ± 19.99	11	21	32
Danny Haren[e]	97.38 ± 9.09	19	14	33
Felix Hernandez[e]	106.82 ± 7.63	19	15	34
Cliff Lee[e]	103.91 ± 15.44	17	17	34
Tim Lincecum[e]	107.47 ± 11.79	25	7	32
Derek Lowe	94.50 ± 15.89	9	25	34
Ricky Nolasco	97.90 ± 12.43	13	18	31
Roy Oswalt	92.70 ± 23.47	11	19	30
Andy Pettitte	102.63 ± 8.00	10	22	32
C C Sabathia[e]	106.53 ± 16.24	14	18	32
Justin Verlander[e]	112.49 ± 13.16	19	16	35
Adam Wainwright[e]	106.29 ± 9.65	17	17	34
Total		297	352	649

Note: [e]Nominated for the prestigious Cy Young pitching award.

4.2. Experimental framework

We follow the main steps of the CRISP-DM methodology (Shearer, 2000), which provides a structured way of conducting the data analysis, with the consequent improvement in the probability of obtaining accurate and reliable results. Figure 3 shows the methodology used to evaluate the proposed model. First, the pitch-by-pitch dataset is transformed into time series by means of our LRPP-based model. Then, the dataset is partitioned into two independent sets: training and testing sets. The training set is composed by entire time series of pitchers' games and it is used by the learning algorithm to derive the model, whose performance is estimated using the testing set of time series.

Once the training and testing partitions of the dataset have been properly defined, we reduced the length of the testing time series in order to evaluate the predictive performance of the model. That is, we sequentially decreased the number of pitches in the testing time series, from last to first, for validation purposes. According to this assessment method, the class value of the reduced testing series is maintained the same as the class membership of the full testing time series.

For a more general evaluation, and in order to avoid a possible over-fitting, we decided to use the stratified 10-fold cross-validation methodology (Han, Pei, & Kamber, 2011). This is a popular statistical technique widely used for comparing the predictive capabilities of data mining methods, which has become the standard in practical terms (Witten, Frank, & Hall, 2011). During the 10-fold cross-validation, the complete dataset is randomly split into 10 mutually exclusive partitions or "folds" of approximately equal size. The classification method is trained and tested 10 times, each time it is trained on all but one-fold and tested on the remaining single-fold. That is, in iteration i, partition T_i is reserved as the testing set, and the remaining partitions are collectively used to train the model. The cross-validation estimate of the overall performance is calculated as the average of the 10 individual results according to some statistical measure of accuracy. Since the cross-validation results depend on the random assignment of the individual samples to distinct folds, a common practice is to stratify the folds themselves. This stratification ensures that each fold has the same proportion of each class value (HP or LP). Empirical studies have shown that stratified 10-fold cross-validation is a recommended method for estimating model performance (even if computation power allows using more folds) due to its relatively low bias and variance (Zeng & Martinez, 2000).

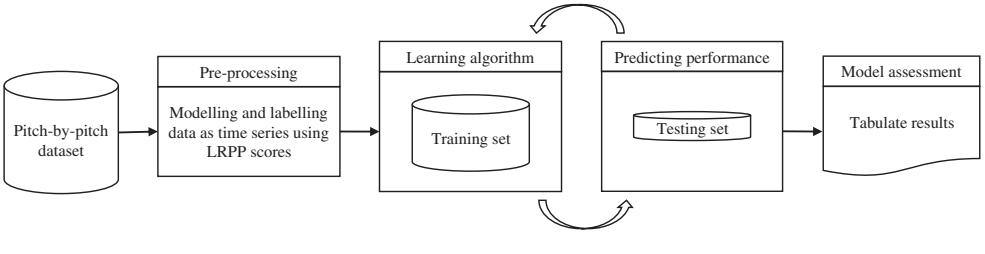

Figure 3. Graphical representation of the methodology used to validate the predictive model proposed.

4.3. Evaluation criteria

We are more interested in predicting when a starting pitcher is about to falter, which is directly reflected by the decreasing of his LPPR score. Accordingly, we selected Precision, Recall and F1 as the statistics criteria to test the performance of the model. The first measures the fraction of time series predicted as LP that are actually labelled as LP, the latter measures the fraction of time series predicted as LP from the total time series labelled as LP. A high value of Precision means that our algorithm returned substantially more relevant results than irrelevant ones, while high Recall means that our algorithm returned most of the relevant results. Equations (5) and (6) present this statistics. In the binary classification context, TP, TN, FP and FN denote true positive (accurate prediction of low pitch performance), true negative (accurate prediction of high pitch performance), false positive (inaccurate prediction of low pitch performance as high) and false negative (inaccurate prediction of high pitch performance as low), respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

Precision and Recall give information about the proportion of correctly and incorrectly classified pitch time series. However, we have decided to add a single measure for characterising the model performance in a more general way. For this aim, we selected the F1 statistic (Equation (7)) as our general evaluative criterion, because combines both Precision and Recall into a single measure of prediction performance.

$$\text{F1} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{7}$$

5. Results

In this section, we present the results obtained by our predictive model for the 20 starting pitchers selected as study case. We used the methodology presented in Section 4.2 during the experiments and the evaluation criteria of Section 4.3. The aim is comparing the predictions of performance when less information about the outcomes of future pitches is known.

First, we conducted experiments in order to assess the performance of the prediction for each pitcher individually. This validation procedure involves sequentially removing time series of one pitcher from the dataset, training with the rest of pitchers' time series and then evaluating the quality of the prediction using the time series of the removed pitcher. We selected data of games played by each particular pitcher for testing and then use the rest of pitchers' data for training the model. Figure 4 shows the results of Precision, Recall and F1 obtained for each pitcher in the dataset following this procedure.

Overall, the results show that the model's performance improves steadily with the increase of the testing time series length. This is an expected result because this increase gives more information to the learning algorithm about the future behaviour of the pitch sequence during the testing. The reduction of time series length by 10 pitches produced the better Precision, Recall and F1 values, with means of 0.90, 0.89 and 0.89, respectively. As it

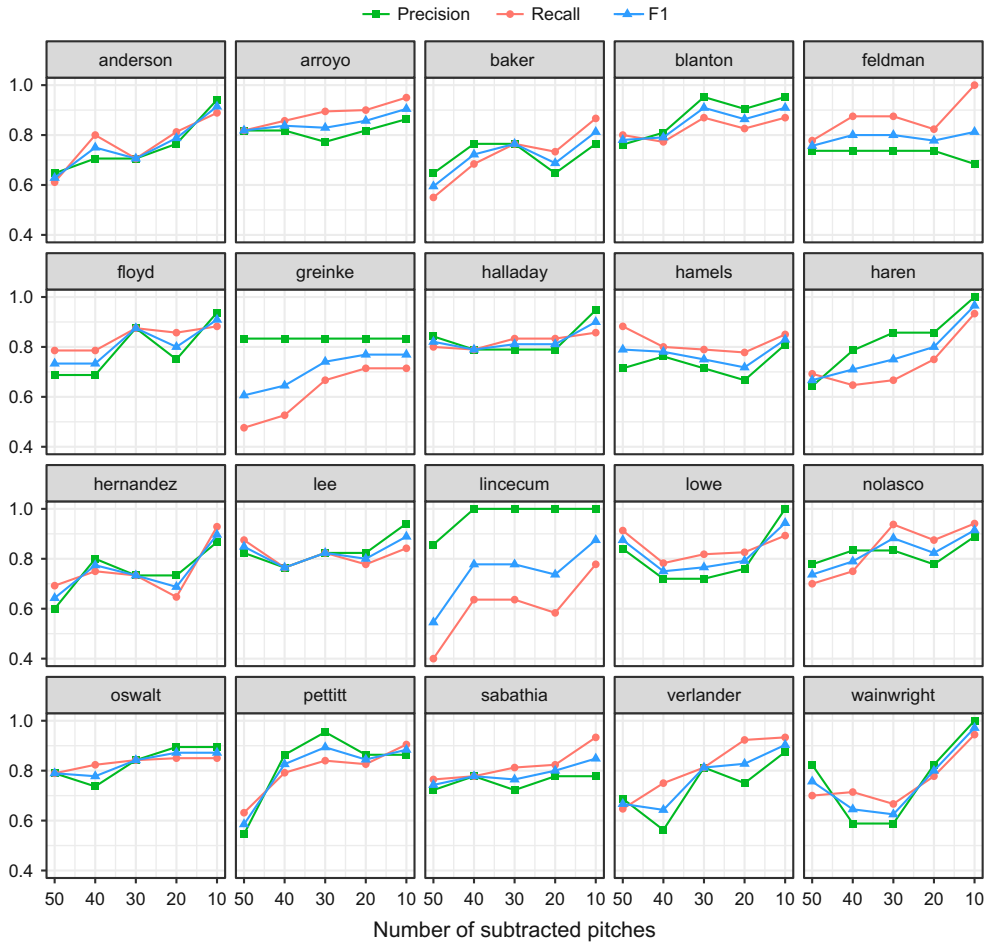

Figure 4. Results of Precision, Recall and F1 for the 20 starting pitchers included in this study using a reduced number of throws for testing.

shows, the model performs well even when 30 of the future pitcher's throws are unknown (F1 values nearly 0.80). Further, it is noticeable the perfect value of Recall obtained for the pitcher Tim Lincecum even 40 throws before he left the game.

As an additional comparison, we applied a two-sample Wilcoxon (or Mann–Whitney) test in order to compare the prediction between elite and non-elite pitchers. Table 4 shows the results of the test for the five time series lengths reduction used. Overall, the results show that p-values are similarly distributed, thus, we can accept the null hypothesis that predictions are the same for elite and non-elite pitchers (significance level of $\alpha = 0.05$). This result show that the performance of our predictive model does not differ significantly for both categories of studied pitchers.

In order to obtain a more general evaluation of the model, we use the stratified 10-fold cross-validation procedure for the complete dataset of studied pitchers. Table 5 presents the results obtained. As additional information, we include the classical Accuracy data mining measure and the confusion matrix of classified pitchers' games; both assess the model ability of correctly predicting the class label of a new or previously unseen pitch

Table 4. Two-sample Wilcoxon test of prediction results to the elite and non-elite pitchers.

Pitchers	Throws left	Precision *p*-value	Recall *p*-value	F1 *p*-value
	10	0.2853	0.6479	1
Elite	20	0.3614	0.0732	0.4669
vs.	30	1	0.0096*	0.0680
Non-elite	40	0.7037	0.0077*	0.0521
	50	0.3417	0.2236	0.5684

*p-value <0.05.

Table 5. General predictive results for the complete dataset using stratified 10-fold cross-validation.

Throws left	Confusion matrix*	LP	HP	Accuracy	Precision	Recall	F1
5	LP	317	35	0.906009	0.924198	0.900568	0.91223
	HP	26	271				
10	LP	305	47	0.869029	0.889213	0.866477	0.877698
	HP	38	259				
15	LP	285	67	0.821264	0.853293	0.809659	0.830904
	HP	49	248				
20	LP	270	82	0.791988	0.835913	0.767046	0.8
	HP	53	244				
25	LP	271	81	0.77812	0.811377	0.769886	0.790088
	HP	63	234				
30	LP	268	84	0.771957	0.807229	0.761364	0.783626
	HP	64	233				
35	LP	257	95	0.744222	0.783537	0.730114	0.755882
	HP	71	226				
40	LP	267	85	0.734977	0.754237	0.758523	0.756374
	HP	87	210				
45	LP	264	88	0.730354	0.752137	0.75	0.751067
	HP	87	210				
50	LP	261	91	0.701079	0.717033	0.741477	0.72905
	HP	103	194				

*Rows show actual class, columns show predictions.

time series. The results confirm that the model performs better as the length of the testing time series increases. It is noticeable that all the measures rise over the 0.9 value when only five of the pitcher's throws outcomes are unknown.

As an additional comparison and as a benchmark, we evaluate the results of our model against a naïve, but not trivial, baseline approach. Due to the fact that our scoring measure is constructed as an aggregate summation of pitching events, it is rational to classifying directly a testing time series as the final result of its LRPP score. Thus, if the LRPP value of the testing time series, in the moment of the cut-off of the time series, is greater than 0 then it is classified as HP, else the performance of the pitcher is classified as LP. Table 6 shows a comparison between the 10-fold cross-validation results of F1 obtained with both, this naïve approach and our model. The application of the Wilcoxon Signed Rank Test to these results shows that our model clearly outperforms the baseline approach (p-value = 0.001953).

Figure 5 shows an example of a testing time series composed by 100 pitches. According to our validation methodology, this time series has been cutting off by 10 throws from the last pitch. The first 90 throws were used for testing the prediction of both the proposed model

Table 6. Comparison of F1 results between the proposed model and a baseline approach.

Throws left	5	10	15	20	25	30	35	40	45	50
Baseline	0.8075	0.79	0.7842	0.7703	0.7645	0.747	0.7411	0.7106	0.7076	0.6819
Our model	0.9122	0.8776	0.8309	0.8	0.79	0.7836	0.7558	0.7563	0.751	0.729

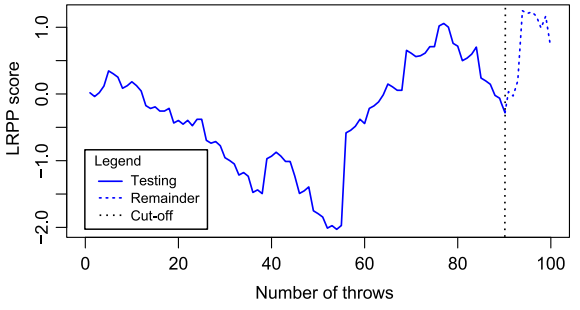

Figure 5. An example of a time series correctly classified as HP by our model but incorrectly classified as LP by the baseline approach.

and the baseline approach. The remainder 10 throws represents the pitcher's performance that we are trying to predict and were used for validating the quality of the prediction. The baseline approach classifies this time series as LP because, in the point of the cut-off, its LRPP value is negative. However, our 1-NN-based model learns from other time series and predicts a good performance for the remainder 10 pitches. Hence, our method correctly classifies the performance of the pitcher as HP.

6. Discussion

The analysis of the experimental results shows that the in-play performance of a starting pitcher is not a homogeneous process. We demonstrate that modelling and visualising these non-homogeneous parts of the game as time series data, using some reasonable criterion of performance, could significantly improve our comprehensibility about pitching. Furthermore, we show that the application of time series classification methods could be a suitable tool for the particular problem of predicting the future performance of a baseball pitcher.

Although our time series classification model proved to be a suitable predictor of performance for a margin of 10 throws of anticipation, it is evident that managers do not dispose of an a-priori knowledge of the total number of throws that a starting pitcher will make during the game. They do not need a model that tells that a pitcher is more likely to falter deeper in the game. Hence, it could be unclear for a manager when is the exact moment for applying this model. However, we believe that the model should perform well when starting pitchers exceed the throw 50, given that the average number of throws per game of the 20 studied pitchers is equal to 101.

It is important to note that predicting the outcome of a single pitch is of dubious value. Managers never pull pitchers after a particular pitch because they are always willing to allow the pitcher to finish the at-bat. Trying to predict a single pitch is irrelevant to their

decision-making process, which is otherwise focused on outcomes of plate appearances and its corresponding pitches sequences. The graphical representation of pitching performance as time series could tell us much about this decision point.

Determining thresholds at the individual pitch level could be more useful in identifying what criteria should be considered for taking the pitcher out. As an example, if they have performed extremely positive in all game, but have a run of 10 consecutive hits and are still with a positive LRPP, how would we decide to take them out or not based on this procedure? Or perhaps at other known points (innings changes, etc.). Once more, the graphical time series representation of performance could be a good starting point for tackling with this particular issue.

Another significant concern is that pitchers, especially when they are tiring late in games, will often "waste" pitches (i.e. throw an unhittable pitch with a remote possibility that the batter will swing). As a consequence of this behaviour, the majority of these "waste" pitches will show up as negative values for our model, when they are in fact part of an intentional strategy. Furthermore, it is not clear what a high performance would be here. That depends on the counter-factual of the outcome for the given pitcher removal decision, and the outcome absent in the decision. Of course, this also depends on bullpen availability and expectations for the given game, which complicates things considerably. The decision of managers always will be based on their knowledge about pitchers and the dynamic of the game. We believe that a threshold of the performance should be considered relative to some average expectation of bullpen performance when replaced.

On the other hand, the proposed predictive approach on this paper could be easily extended to other sports where play-by-play data of individual players are available, such as basketball (Vračar, Štrumbelj, & Kononenko, 2016) or cricket (Iyer & Sharda, 2009). The model makes feasible its inclusion into any expert system and decision-making process that requires the ranking and evaluation of players. In addition, the scoring system could be tested for predictive ability of run scoring and game outcomes.

In our opinion, further research on this predictive model should consider the following lines:

- Compare the predictive performance of the 1-NN algorithm with other state-of-the-art time series classification methods (e.g. support vector machines, artificial neural networks or decision trees).
- Estimate empirically the necessary number of throws that must be known for achieving a competitive prediction of performance.
- Assess the feasibility of using this model in other sports domains, especially in those producing large amount of play-by-play data.

7. Conclusion

In this paper, we presented a time series classification model for analysing the performance of a starting pitcher and predicting when he should be removed from the game and replaced by a reliever. We transformed and labelled pitch-by-pitch data, obtained from the PITCHf/x system, into time series using an accumulative metric of pitcher's performance, which is based on the well-known linear runs baseball metric. The k-NN algorithm, in conjunction with the dissimilarity measure DTW, was used for classifying pitch time series data and predicting the future performance during the game. In order to validate the

model, 20 MLB starting pitchers were selected for model analysis. The experimental results show that the model performs significantly accurate (with mean values of Precision, Recall and F1 near to 0.9) when the 10 last pitcher's throws outcomes are unknown. Furthermore, the results of the model show no predictive differences between elite and normal pitchers. The development and application of this model is of interest, not only because it attempts to answer one of the most difficult questions in baseball, but also, as a novel approach for modelling, analysis and predicting performance in baseball using time series classification and data mining methods.

Notes

1. http://www.fangraphs.com/guts.aspx?type=cn
2. http://gd2.mlb.com/components/game/mlb/

Acknowledgements

The authors would like to thank Mr Jim Albert for his encouragement and constructive suggestions, which considerably helped to improve the quality of this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

César Soto-Valero http://orcid.org/0000-0003-0541-6411
Mabel González-Castellanos http://orcid.org/0000-0003-0152-444X
Irvin Pérez-Morales http://orcid.org/0000-0001-6211-1468

References

Albert, J. (2010a). Baseball data at season, play-by-play, and pitch-by-pitch levels. *Journal of Statistics Education, 18*(3), 1–27.
Albert, J. (2010b). Sabermetrics: The past, the present, and the future. *Mathematics and Sports, Chapter 1,* 3–15.
Batista, G., Hao, Y., Keogh, E., & Mafra-Neto, A. (2011). Towards automatic classification on flying insects using inexpensive sensors. In *10th International Conference on Machine Learning and Applications (ICMLA)* (Vol. 1, pp. 364–369). Washington, DC: IEEE.
Burley, C. (2004). The importance of strike one. Retrieved from http://www.hardballtimes.com/the-importance-of-strike-one-part-two/
Chih-Cheng, C., Yung-Tan, L., & Chung-Ming, T. (2014). Professional baseball team starting pitcher selection using ahp and topsis methods. *International Journal of Performance Analysis in Sport, 14*, 545–563.
Fast, M. (2010). What the heck is pitchf/x. *The Hardball Times Annual, 2010*, 153–158.
Fleischman, M., Roy, B., & Roy, D. A. (2007). Temporal feature induction for baseball highlight classification. In *Proceedings of the 15th international conference on Multimedia* (pp. 333–336). Augsburg: ACM.
Flesca, S., Manco, G., Masciari, E., Pontieri, L., & Pugliese, A. (2007). Exploiting structural similarity for effective web information extraction. *Data & Knowledge Engineering, 60*, 222–234. Intelligent Data Mining.

Fu, T.-C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence, 24*, 164–181.
Geurts, P. (2001). Pattern extraction for time series classification. In *Principles of Data Mining and Knowledge Discovery: 5th European Conference (PKDD 2001)* (pp. 115–127). Berlin, Heidelberg: Springer.
Gooijer, J. G. D., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting, 22*, 443–473. Twenty five years of forecasting.
Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques* (2nd ed.). Waltham, MA: Elsevier.
Hoang, P., Hamilton, M., Murray, J., Stafford, C., & Tran, H. (2015). *A dynamic feature selection based LDA approach to baseball pitch prediction*. Cham: Springer International Publishing. (pp. 125–137).
Iyer, S. R., & Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection. *Expert Systems with Applications, 36*, 5510–5522.
Keeley, D., Oliver, D. G., Torry, R. M., & Wicke, J. (2014). Validity of pitch velocity and strike percentage to assess fatigue in young baseball pitchers. *International Journal of Performance Analysis in Sport, 14*, 55–366.
Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery, 7*, 349–371.
Keogh, E., & Ratanamahatana, C. (2005). Exact indexing of dynamic time warping. *Knowledge and Information Systems, 7*, 358–386.
Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. New York, NY: WW Norton & Company.
Maeda, T., Fujii, M., Hayashi, I., & Tasaka, T. (2014). Sport skill classification using time series motion picture data. In *Industrial Electronics Society, IECON 2014–40th Annual Conference of the IEEE* (pp. 5272–5277). Dallas, TX: IEEE.
Nanopoulos, A., Alcock, R., & Manolopoulos, Y. (2001). Feature-based classification of time-series data. *International Journal of Computer Research, 10*, 49–61.
Pavitt, C. (2011). An estimate of how hitting, pitching, fielding, and basestealing impact team winning percentages in baseball. *Journal of Quantitative Analysis in Sports, 7*(4), 1–20.
Peach, J. T., Fullerton, S. L., & Fullerton, T. M. (2016). An empirical analysis of the 2014 major league baseball season. *Applied Economics Letters, 23*, 138–141.
Piette, J., Braunstein, A., McShane, B. B., & Jensen, S. T. (2010). A point-mass mixture random effects model for pitching metrics. *Journal of Quantitative Analysis in Sports, 6*(3), 1–15.
Ratanamahatana, C., & Keogh, E. (2005). Three myths about dynamic time warping data mining. In *Proceedings of the Fifth SIAM International Data Mining Conference (SDM'05)*, Philadelphia, PA, USA. (pp. 506–510).
Shearer, C. (2000). The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing, 5*, 13–22.
Sidhu, G., & Caffo, B. (2014). MONEYBaRL: Exploiting pitcher decision-making using reinforcement learning. *The Annals of Applied Statistics, 8*, 926–955.
Sidran, D. E. (2005). A method of analyzing a baseball pitcher's performance based on statistical data mining. Unpublished. https://doi.org/10.13140/rg.2.1.1653.7041
Tango, T., Lichtman, M., & Dolphin, A. (2007). *The book: Playing the percentages in baseball*. Washington, DC: Potomac Books.
Vračar, P., Štrumbelj, E., & Kononenko, I. (2016). Modeling basketball play-by-play data. *Expert Systems with Applications, 44*, 58–66.
Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery, 26*, 275–309.
Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining practical machine learning tools and techniques*. Vol. 3, Burlington, ON: Morgan Kaufmann Publishers.
Wolf, G. H. (2015). The sabermetric revolution: Assessing the growth of analytics in baseball. *Journal of Sport History, 42*, 239–241.

Xing, Z., Pei, J., & Keogh, E. (2010). A brief survey on sequence classification. *SIGKDD Exploration Newsletter, 12*, 40–48.
Zeng, X., & Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence, 12*, 1–12.
Zimniuch, F. (2010). *Fireman: The evolution of the closer in baseball*. Chicago, IL: Triumph Books.