The Korean Journal of Applied Statistics (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.2.171
33(2), 171–184

Evaluating the quality of baseball pitch using PITCHf/x

Sungmin Park[a] · Woncheol Jang[a,1]

[a]Department of Statistics, Seoul National University

(Received February 26, 2020; Revised March 20, 2020; Accepted March 20, 2020)

Abstract

Major League Baseball (MLB) records and releases the trajectory data for every baseball pitch, called the PITCHf/x, using three high-speed cameras installed in every stadium. In a previous study, the quality of the pitch was assessed as the expected number of bases yielded using PITCHf/x data. However, the number of bases yielded does not always lead to baseball scores, or runs. In this paper, we assess the quality of a pitch by combining baseball analytics metric Run Expectancy and Run Value using a Random Forests model. We compare the quality of pitches evaluated with Run Value to the quality of pitches evaluated with the expected number of bases yielded.

Keywords: Major League Baseball, quality of pitch, expected number of bases yielded, run expectancy, run value, random forests model

1. 서론

PITCHf/x는 야구 경기의 투구를 세 개의 정밀한 카메라로 관측해 기록하는 시스템으로 2006년에 도입되어 현재 미국의 모든 메이저리그 야구(Major League Baseball; MLB) 경기장에 설치되어 있다. 각 경기장에 설치된 세 대의 카메라는 투수가 던지는 모든 공을 촬영해 투구의 구종, 속도, 가속도, 회전속도, 위치 등의 정보를 추적해 정보를 기록한다. 그리고 그 기록은 메이저리그에서 MLB Gameday Data의 항목으로 공개하기 때문에 누구나 접근하고 활용할 수 있다. 한국 프로 야구(Korea Baseball Organization; KBO)에서는 투구추적시스템(Pitch Tracking System; PTS)을 도입해 PITCHf/x와 유사한 투구 추적 데이터를 수집하고 있다. MLB Gameday를 통해 공개된 PITCHf/x 데이터는 야구 팬들과 전문가들에 의해 활발하게 분석 주제로 활용되고 있는데, 이 논문에서는 PITCHf/x 데이터를 이용해 투구의 질을 평가한 선행 연구 (Swartz 등, 2017)를 바탕으로, 그에 대한 개선 방향과 응용 가능성에 대해서 논의하는 내용을 담고 있다. 특히, 투수 자원은 야구 경기에서 핵심적인 포지션으로, 투구의 질에 대한 연구를 활용해 투수 운용에 대한 개선점을 발견할 수 있다면 팀의 성적에 커다란 기여를 할 것으로 예상할 수 있다.

This research was supported by the National Research Foundation of Korea (NRF) grant and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Korea government (MSIT) and the Ministry of Health & Welfare, Republic of Korea (No. 2017R1A2B2012816, HI19C1234).

[1]Corresponding author: Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-Gu, Seoul 08826, Republic of Korea. E-mail: wcjang@snu.ac.kr

Table 1.1. Expected number of Bases yielded for balls and strikes at each ball count

Count	Ball	Strike
0-0	0.52	0.38
0-1	0.42	0.28
0-2	0.32	0.00
1-0	0.63	0.43
1-1	0.52	0.32
1-2	0.39	0.00
2-0	0.82	0.52
2-1	0.69	0.39
2-2	0.55	0.00
3-0	1.00	0.69
3-1	1.00	0.55
3-2	1.00	0.00

투구의 질은 주관적인 개념이기 때문에 연구를 진행하기에 앞서 반응 변수에 대해 측정 가능하고 구체적인 조작적 정의를 필요로 한다. Swartz 등 (2017)에서는 투구의 질을 주자가 진루한 베이스의 수, 즉 피루타수를 바탕으로 결정하였다. 그리고 피루타수 정의에 안타뿐만 아니라 볼넷과, 몸에 맞는 공 등을 모두 포함시켜 모든 타석에서 피루타수가 산출되도록 정의했다. 따라서, 투구의 질은 아웃의 경우 0, 1루타와 볼넷, 몸에 맞는 공은 1, 그리고 2루타, 3루타, 홈런은 각각 2, 3, 4의 가치를 지닌다고 정의했다.

스트라이크 아웃과 볼넷이 아닐 경우 직접적으로 진루를 야기하지 않기 때문에 실제 피루타수를 반응 변수로 사용할 수 없어 Swartz 등 (2017)에서는 기대 피루타수로 반응 변수를 정의했다. 2013년부터 2015년까지 스트라이크와 볼이 발생한 타석을 아웃카운트별로 종합한 뒤 평균을 이용하여 아웃카운트별 스트라이크와 볼의 피루타수 기댓값을 계산해 볼과 스트라이크의 반응 변수로 표현하였다. 즉, 볼카운트 0-0 상황에서 볼로 인해 볼카운트가 0-1이 된 총 N개의 타석 중 최종적으로 허용한 루타수를 B라고 하면, 볼카운트 0-0의 기대피루타수는 B를 N으로 나눈 값으로 Table 1.1에서 0.52의 값이다. 2 스트라이크 상황에서 스트라이크는 스트라이크아웃으로 피루타수 0, 3 볼 상황에서 볼은 볼넷으로 피루타수 1이기 때문에, 피루타수 기댓값을 계산하지 않았다. 위의 방법에 의한 모든 볼카운트에 대한 피루타수 기댓값은 Table 1.1에 나타나 있다.

그 다음 단계는 피루타수 기댓값과 변수들 사이의 관계를 표현하는 모형 선택이다. 선행 연구에서는 투구의 질을 예측하는 모형을 구축하기에 앞서, 투구의 위치나, 투구수 등의 변수가 피루타수와 비선형적인 관계를 취할 것이며, 그리고 변수들 사이의 교호작용이 작용할 것이라는 가정을 했다. 때문에 선형모형으로는 변수들과 피루타수 사이의 비선형 관계를 표현할 수 없을 뿐만 아니라, 교호작용을 선형 모형에 포함할 경우 변수가 크게 늘어나 발생하게 될 다중공선성 우려가 있다. 때문에, 그 대안으로 선행연구에서는 Random Forests 방법을 이용한 회귀나무 모형을 적합해 피루타수를 추정했다.

선행 연구에서는 투구의 질을 투구의 결과로 진루하게 된 베이스의 수로 정의했는데, 진루가 득점으로 연결되지 않으면 승리에는 도움이 되지 않는다. 따라서 본 연구에서는 새로운 지표의 필요성을 설명하고 이에 대한 Random Forests 회귀나무 모형을 적합하고자 한다. 실제 투수의 능력은 위기관리 능력도 포함해서 평가해야 한다는 생각에 본 연구에서는 피루타수 대신 득점을 기준으로 투구의 질을 평가하고자 한다. 투구의 질이 승리에 기여하는 영향을 예측하기 위해서는 투구의 질을 득점수로 정의해 모형을 학습시키는 것이 적합할 수 있다. 이어지는 논문의 본문 구성은 다음과 같다. 2.1절에서는 분석에 사용된 데이터에 대해 소개하고 설명한다. 2.2절에서는 선행연구의 내용과 차별화할 수 있는 새로운 지표를 확인한 뒤 그것을 이용한 새로운 모형을 구현, 선행연구의 모형과 비교해본다. 2.3절에서는 사용된 방법

Table 2.1. Data used in the analysis

Period	2013.03.31∼2015.10.31
Total number of games	7,465
Total number of pitches	2,152,747
Total number of at bats	566,465
Average number of pitches per game	144
Average number of at bats per game	38

의 한계 및 개선방향에 대해 살펴본다. 그리고 결론을 통해 연구의 의의와 응용 방향에 대해 제안해보며 논문을 마무리한다.

2. 본론

2.1. 데이터 소개

데이터는 MLB 2013년부터 2015년까지 시즌의 MLB Gameday Data를 이용했다. 데이터는 MLB 공식 홈페이지에 XML 형식으로 제공된다. MLB Gameday Data의 URL을 이용해 접속하면 누구나 쉽게 과거 MLB 야구 경기 기록을 열람할 수 있으며, XML 정보를 손쉽게 web scraping하여 저장할 수 있다. R의 경우에는 Carson Sievert가 2015년 12월에 공개한 pitchRx 패키지를 이용해 MLB Gameday Data를 저장할 수 있으며, 대용량 데이터를 수월하게 저장하기 위해, 일반적인 CSV나 TXT 확장자 형태의 파일이 아닌 SQLite 등의 데이터베이스에도 바로 연결해 저장할 수 있다. 분석에 사용한 데이터의 양이 방대해, SQLite 데이터베이스에 자료를 저장한 뒤 본 연구에서는 활용하였다.

간혹 MLB Gameday Data를 살펴보면 특정 변수들의 데이터가 없는 경우가 존재한다. 일부 자료 결측의 내용을 살펴보기 위해 URL로 직접 접속해 MLB Gameday Data를 확인한 결과, 결측치의 많은 경우 전산 시스템 상의 오류로 해당 페이지에 자료가 누락된 경우가 대부분이었다. 한 편 자료의 값이 명백하게 잘못 기입된 경우, 가령 스트라이크의 값이 0, 1, 2를 벗어나거나, 볼의 값이 0, 1, 2, 3을 벗어나는 경우 자료에 오류가 있음으로 판단하였다. 이러한 결측 및 오류를 제외한 2013–2015 시즌 유효한 데이터는 총 2,152,747개 중 2,134,803개로 약 0.8%의 기록이 누락되었다. 투구 정보는 총 26개의 변수를 활용했는데, 공을 던져지는 지점과 공이 홈플레이트를 통과하는 지점에서의 속도, 가속도, 회전속도, 그리고 위치와 더불어 구종과 구종에 대한 신뢰도 등이 포함된다. 그리고 그 투구에 대한 경기 결과 또한 MLB Gameday Data에 제공되어 있어, 이 정보를 투구에 대한 정보와 연결 지어 결과를 확인했다. 사용된 데이터에 대한 요약 통계량은 Table 2.1에 제공되어 있고, Table 2.2와 Figure 2.1에는 PITCHf/x 변수에 대한 설명을 수록하였다. PITCHf/x 변수에 대한 소개를 다룬 Table 2.2와 Figure 2.1은 Fastballs 웹사이트에 게재된 Fast (2007)의 게시물을 인용하였다.

2.2. 득점 수를 이용한 예측 모형 구현

피루타수로 투구의 질을 정의하는 경우, 투구의 결과는 타석별로 결정되지만, 득점 수로 투구의 질을 정의하는 경우에는 투구의 결과가 이닝별로 결정되며 투구나 타석의 결과로 즉각적으로 결정되지 않는다. 득점 수로 투구의 결과를 정의하게 되면 주자 상황과, 아웃카운트를 비롯해 추가적으로 고려해야 하는 조건들이 발생하는데 이를 효과적으로 해결할 수 있어야 한다.

한 이닝 내부에서 누적된 결과가 각각 득점에 기대하는 효과를 살펴보기 위해, Tango 등 (2007)는 run expectancy, run value라는 개념을 제공하고, 이를 바탕으로 득점 기댓값(run expectancy; RE)와 득점 가치(run value; RV)에 실제 득점을 추가한 run expectancy based on 24 base-out states (RE24)라는

Table 2.2. PITCHf/x variables

des	투구 결과에 대한 요약
id	투구 고유 번호 (PITCHf/x 도입 이후에는 sv_id로 대체)
type	투구 결과에 대한 요약 (B: 볼, S: 스트라이크, X: 인플레이)
x	공이 홈플레이트를 통과할 때 수평 축 좌표
y	공이 홈플레이트를 통과할 때 수직 축 좌표
sv_id	투구 고유 번호
start_speed	공을 던지는 시점에서 투구 속도
end_speed	공이 홈플레이트 통과하는 시점에서 속도
sz_top	스트라이크 존의 상한 높이
sz_bot	스트라이크 존의 하한 높이
pfx_x	공을 던지는 시점에서 홈플레이트까지 수평 축 움직임
pfx_z	공을 던지는 시점에서 홈플레이트까지 수직 축 움직임
px	공이 홈플레이트 통과하는 시점에서 수평 축 좌표
pz	공이 홈플레이트 통과하는 시점에서 수직 축 좌표
x0	공을 던지는 시점에서 수평 축 좌표
y0	홈플레이트부터 PITCHf/x 관측 장비의 거리
z0	공을 던지는 시점에서 수직 축 좌표
vx0	공을 던지는 시점에서 수평 축 속력
vy0	공을 던지는 시점에서 홈플레이트 방향 속력
vz0	공을 던지는 시점에서 수직 축 속력
ax	공을 던지는 시점에서 수평 축 가속력
ay	공을 던지는 시점에서 홈플레이트 방향 가속력
az	공을 던지는 시점에서 수직 축 가속력
break_y	투구의 곡선 궤적과 직선 궤적 사이의 거리가 최대가 되는 공의 위치에서의 수평 축 움직임, Figure 2.1 참조
break_angle	공이 홈 플레이트를 지나간 점을 이은 직선과, 수직으로 홈 플레이트를 지나갔을 경우의 궤적 이루는 각, Figure 2.1 참조
break_length	투구의 곡선 궤적과 직선 궤적 사이의 최대 거리, Figure 2.1 참조
pitch_type	투구 종류
type_confidence	투구 종류에 대한 정확도 추정치
zone	홈플레이트를 지날 때 공이 통과한 위치 구분
nasty	공이 통과한 zone에 기반한 점수
spin_dir	공의 회전 방향
spin_rate	공의 회전수

방법을 제안하였다. 각각의 용어의 정의와 계산 방법은 다음과 같다.

2.2.1. 득점 기댓값(run expectancy; RE) 기대 득점은 8가지 주자 상황과 3가지 아웃 카운트의 조합으로 발생하는 24가지 주자와 아웃 상황에서 발생한 득점 평균을 의미한다. 주자상황이 b, 아웃카운트가 o인 조합에 대해 득점 기댓값을 계산하는 공식은 다음과 같다.

$$\mathrm{RE} = \frac{\text{주자와 아웃 상황 (b, o)부터 이닝 종료까지 발생한 득점의 합}}{\text{주자와 아웃 상황 (b, o)가 발생한 이닝 수}}. \tag{2.1}$$

2013년부터 2015년의 MLB Gameday 자료에 위 공식을 적용해 득점 기댓값을 구한 결과는 Table 2.3과 같다. 결과표를 이용해 예시를 살펴보면, 주자가 없고 2 아웃인 상황부터 이닝이 종료되기 전까지

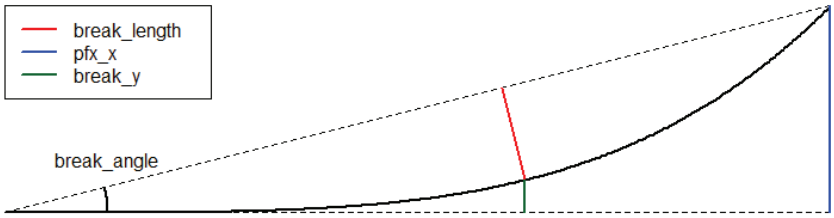

Figure 2.1. Graphic explanation of PITCHf/x variables when viewed from above. The left end point is the position of the pitcher and the right end point is the location of the ball as it crosses the home plate. The variable "break_length" refers to the maximum distance between the arc and the chord of ball trajectory. The variable "break_y" refers to the horizontal length at the location of the ball trajectory where maximum distance between the arc and the chord is achieved. The variable "pfx_x" refers to the length of horizontal movement of the ball as it crosses the home plate. Lastly, the variable "break_angle" refers to the angle between two lines, the first line connecting the initial point and the end point of the thrown ball trajectory and the second line connecting the initial point and the end point of an imaginary ball trajectory if the pitch did not have any curve.

Table 2.3. Run expectancy in each base-out state

Base			0 Out	1 Out	2 Out
-	-	-	0.461	0.242	0.093
1B	-	-	0.835	0.492	0.204
-	2B	-	1.084	0.633	0.298
1B	2B	-	1.418	0.849	0.392
-	-	3B	1.374	0.919	0.341
1B	-	3B	1.756	1.073	0.425
-	2B	3B	1.984	1.343	0.534
1B	2B	3B	2.251	1.527	0.715

평균 0.093 득점이 발생했고, 주자 만루에 0 아웃인 상황부터 이닝이 종료되기 전까지는 평균 2.251 득점이 발생했음을 의미한다. 이 때, 도루와 견제구로 발생하는 주자 아웃 타석 상황의 변화도 하나의 상황으로 자료에 기록되어 있으며, RE24 계산에도 반영되었다.

2.2.2. 득점 가치(run value; RV) 각 타석에서 마지막으로 던진 투구는 주자와 아웃 상황의 득점 기댓값을 변화시킨다. 득점 가치는 투구의 결과로 발생하는 득점 기댓값의 변화량을 의미한다. 예를 들어, 주자가 없고 0 아웃인 상황에서 아웃을 당하게 되면 득점 기댓값은 0.461에서 0.242로 감소한다. 이 때 RV는 -0.219 $(= 0.242 - 0.461)$가 된다. RV를 계산하는 공식은 다음과 같이 표현할 수 있다.

$$\text{RV} = \text{RE End State} - \text{RE Beginning State}. \tag{2.2}$$

그러나 RV를 계산하는 위 공식은 득점이 발생하지 않은 상황에서의 RV를 정확하게 반영하지만, 실제로 득점이 발생했을 때, 해당 득점을 반영하지 못할뿐더러 오히려 RV의 방향을 반대로 계산한다. 가령 주자 만루 0 아웃 상황에서 만루 홈런을 터뜨릴 경우, 위 공식에 의한 RV는 -1.790 $(= 0.461 - 2.251)$가 된다. RE24는 RV에 각 타석에서 발생하는 실제 득점을 더해서 계산되고 그 공식은 아래와 같다. 이로서 타석 내 마지막 투구에 대해서는 RV를 모두 부여할 수 있다. 본 논문에서 정의하고 사용하는 RV는 모두 RE24의 산출식 (2.3)을 참조해 해당 타석에서의 득점을 포함해 계산하였다.

$$\text{RE24} = \text{RE End State} - \text{RE Beginning State} + \text{Runs Scored}. \tag{2.3}$$

Table 2.4. Run expectancy in each ball count

B	S	RE
0	0	0.456
0	1	0.405
0	2	0.349
1	0	0.495
1	1	0.431
1	2	0.370
2	0	0.557
2	1	0.480
2	2	0.404
3	0	0.665
3	1	0.579
3	2	0.492

그러나 전체 투구 중 위의 방법에 의해 계산할 수 있는 비율은 25% 뿐이고, 스트라이크와 볼에 대해 RV를 계산할 수 있는 방법을 추가적으로 구현해야 한다. 스트라이크와 볼의 RV를 계산하기 위해서 선행 연구의 기대 피루타수와 득점 기댓값의 개념을 조합해 볼 b, 스트라이크 s에 대응하는 각 볼카운트 상황(b, s)에 대응하는 득점 기댓값을 2013년부터 2015년까지 3개 시즌에 해당하는 경기 기록을 바탕으로 계산한 뒤 Table 2.4에 수록하였다.

$$\text{RE} = \frac{\text{볼카운트(b, s)부터 이닝 종료까지 발생한 득점의 합}}{\text{볼카운트(b, s) 발생 횟수}}. \tag{2.4}$$

각 볼카운트 상황에 대해 스트라이크와 볼의 RV를 주자와 아웃 상황별 RV를 계산한 방법과 동일하게 계산했다. 가령 볼카운트 0-0 상황에서 스트라이크를 던져 볼카운트가 0-1로 바뀌는 경우 RV는 -0.051 $(= 0.405 - 0.456)$가 되고, 볼을 던져 볼카운트가 1-0으로 바뀌는 경우 득점 가지는 0.039 $(= 0.495 - 0.456)$가 된다. 각 볼카운트 상황별 RV를 정리한 결과는 다음의 Table 2.5에 나타나 있다. 스트라이크가 2개인 볼카운트에서 스트라이크의 RV와 볼이 3개인 경우에서의 RV는 주자와 아웃 상황이 바뀌는 경우에 해당하기 때문에 RE24의 산출식 (2.3)에 의해 계산되며 볼카운트 상황별 RV와는 무관한 것으로 판단하였다. 다만, 2 스트라이크 상황에서 파울이 선언될 경우는 주자와 아웃 상황 및 볼카운트 상황이 아무런 변화를 하지 않았기 때문에 RV를 0으로 간주했다. Table 2.5를 살펴보면 꽉 찬 볼카운트에서 스트라이크와 볼의 가치가 상대적으로 더 큰 것을 확인할 수 있다.

2013년부터 2015년까지의 MLB Gameday 자료에 RV 계산 방법을 적용해 각 투구의 RV를 부여했다. 한 타석에 볼이 다섯 개인 경우나, 스트라이크가 4개 이상 기록된 경우로, 자료 상에 오류가 있었던 경우는 분석에서 제외하였다. Table 2.6은 부여된 RV의 요약 통계량을 보여주고 있다.

2013년 시즌 8월까지의 투구 자료에 대해 RV를 계산한 뒤 9월 이후의 투구 자료에 대한 RV를 예측하는 Random Forests 모형을 적합하였다. 데이터의 양이 방대해, 전체 데이터를 모두 적합하지 않고 주어진 시즌의 8월 이전 투구 데이터 중 10,000개의 표본을 추출해 그에 대해 1,000개의 의사결정나무를 적합해 Random Forest 모형을 구현하였다.

Figure 2.2는 세 명의 투수 Adam Wainwright, 류현진, 그리고 Matt Harvey의 2013년 투구 수별 평균 RV를 나타낸 자료이다. Adam Wainwright는 앞서 언급했듯 2013년 최다 이닝(241.2 이닝)을 던지며 수비무관 평균자책점(fielding independent pitching; FIP) 4위를 기록한 훌륭한 선발 투수이고, Matt Harvey는 178.1 이닝으로 다소 적은 이닝을 던졌지만 FIP 2.00으로 1위를 기록하였다. 류현진은

Table 2.5. Run value in each ball count

Count	RV	
	Ball	Strike
0-0	0.038	−0.051
0-1	0.026	−0.056
0-2	0.021	-
1-0	0.063	−0.064
1-1	0.049	−0.061
1-2	0.033	-
2-0	0.108	−0.077
2-1	0.099	−0.076
2-2	0.088	-
3-0	-	−0.086
3-1	-	−0.087
3-2	-	-

Table 2.6. Summary statistics of run value

Min	-1.911
Q1	-0.061
Median	0.000
Q3	0.049
Max	3.752
Mean	0.024
SD	0.243

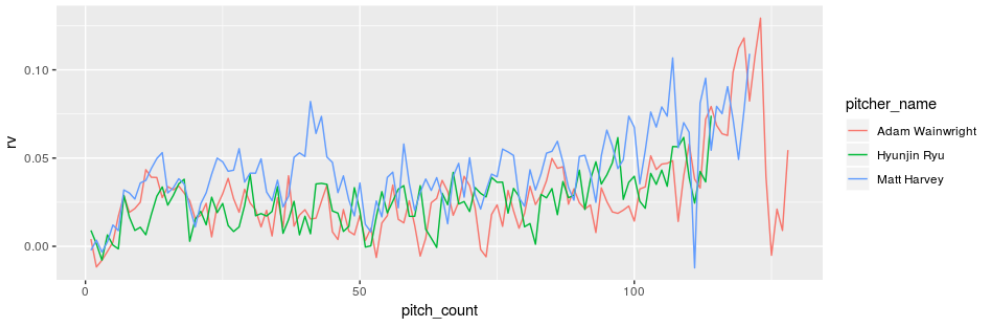

Figure 2.2. Average run value per pitch count for Adam Wainwright, Hyunjin Ryu, Matt Harvey in 2013.

2013년 메이저 리그에 데뷔해 192 이닝을 투구하고 메이저 리그 전체 선발 투수 79명 중 FIP 17위를 기록하는 준수한 성적으로 시즌을 마무리 하였다. 이들 세 명의 투구별 평균 RV를 살펴보면 경기 후반부로 접어들수록, 즉, 투구 수가 많아질수록 Adam Wainwright의 RV가 가장 낮게 유지되고, 류현진 선수의 RV가 중간을 유지하며, Matt Harvey의 RV 추정치가 가파르게 상승하는 것을 볼 수 있다.

Figure 2.3은 세 명의 투수가 경기별로 던진 투구 수를 상자 그림으로 표현한 자료다. 세 선수 모두 경기 당 100개 전후의 투구 수를 유지하였다. 그 중 Matt Harvey의 투구 수 분포의 1사분위가 다른 두 선수에 비해 다소 낮고, Adam Wainwright는 투구 수가 많았던 경기가 두 선수에 비해 많은 것으로 나타난다. Figure 2.2를 살펴보면 파란색 선, 즉, Matt Harvey 선수의 투구 수별 평균 RV가 가장 높게 나

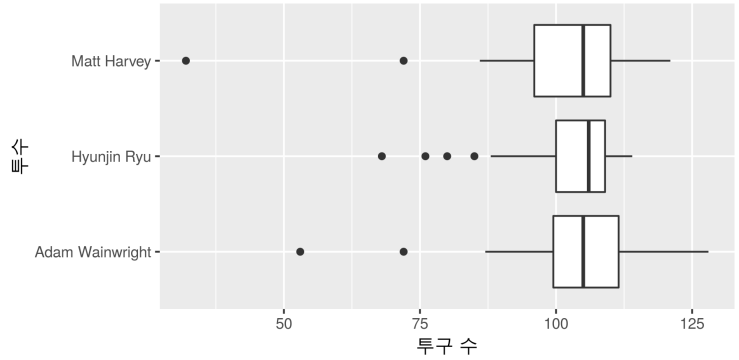

Figure 2.3. Pitch count per game for Adam Wainwright, Hyunjin Ryu, Matt Harvey in 2013.

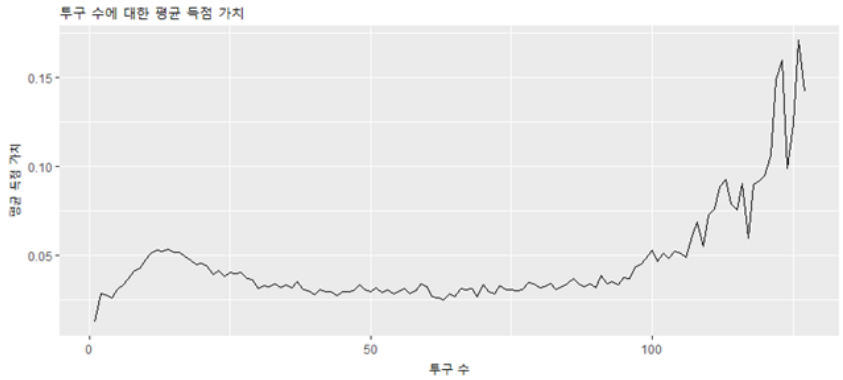

Figure 2.4. Average run value per pitch count thrown in 2013.

타나는데 세 명의 선수 중 2013년 전체 RV 평균이 가장 낮은 의아한 현상을 발견할 수 있다. 이는 상자 Figure 2.3을 통해 설명할 수 있다. Matt Harvey 선수는 RV가 다른 선수들보다 다소 높더라도 보통 다른 투수들에 비해 적은 이닝을 소화하고 마운드에서 내려왔기 때문에 낮은 RV 평균을 유지할 수 있었다고 해석할 수 있다.

Random Forest 모형으로 투구의 RV를 추정한 결과를 투구의 질로 평가할 수 있다. 2013년 시즌 시작부터 8월까지의 투구 결과를 이용해 모형을 적합하고 9–10월의 결과에 대해 투구의 RV를 예측했다. 투구 수에 대한 평균 예측 RV를 Figure 2.4에 요약했다. 결과를 살펴보면 경기 시작 후에 20–30 구까지는 몸이 덜 풀려 투구의 RV가 다소 상승했다가 30–90 구 사이에서는 투구의 RV가 낮게 유지됨이 확인된다. 그리고 100구 이후로 평균 RV가 급격히 치솟기 시작하는 것을 확인할 수 있다. 100구 이후의 평균 RV에 대한 자료는 많지 않기 때문에 조심스럽게 해석해야 하지만, 가파른 증가세를 보이는 현상이 뚜렷하게 관찰된다.

2013년 전체 투구 자료에 대한 자료의 숫자가 충분히 많기 때문에 Figure 2.4의 전체 투구 수에 대한 RV 평균에는 뚜렷한 개형이 나타나지만 개별 경기에서 투구 수에 대한 예측 RV를 그래프로 표현한 경우에는 추정이 불안정해 개형을 파악하기 어렵다. Figure 2.5는 2013/09/01의 경기에 대한 추정 RV를 나타내는데 불안정한 추정치로 인해 RV에 대한 경향성은 잘 드러나지 않는다.

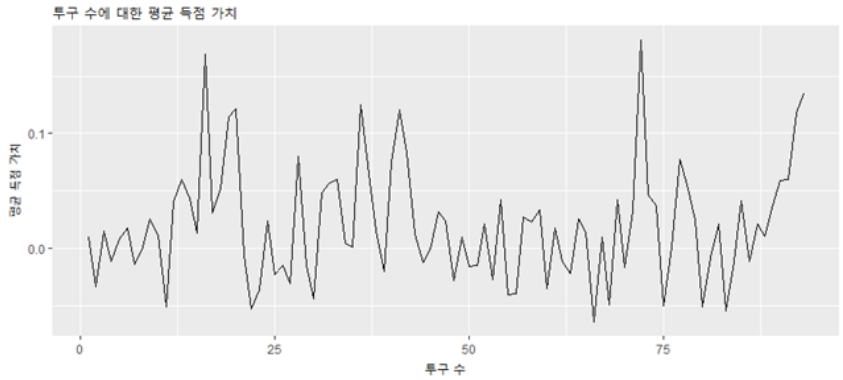

Figure 2.5. Run value of the Los Angeles Angels starting pitcher in the 2013/09/01 Los Angeles Angels vs. Milwaukee game.

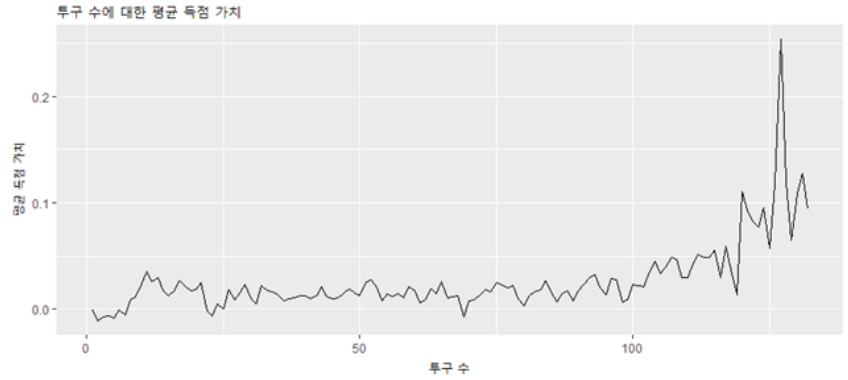

Figure 2.6. Clayton Kershaw's average run value from 2013 to 2015, from the random forests model trained only using Clayton Kershaw's data.

지금까지 적합한 Random Forest 모형은 각각의 투수들을 모두 동질적이고 대체 가능한 자원으로 간주했다. 그렇지만 실제로 각각의 투수는 동질적이지 않다. 투수별로 각각의 강점이 있고, 주력으로 사용하는 구질과 전략이 존재한다. 또한 투구의 질을 실제로 현장에 적용하게 된다면, 전 구단의 모든 데이터를 수집해 투구의 질을 평가하기보다는, 개별 팀에 대해 모형을 적합해 팀의 운영에 이용하게 될 것이다. 따라서 시즌 전체 자료가 아니라 투수 각 개인에 대한 PITCHf/x 자료를 개별적으로 구분해 모형을 적합해볼 필요성이 있다. 현재 MLB 최고의 투수 중 한 명으로 꼽히는 Clayton Kershaw의 경우 2013년부터 2015년까지 10,000개를 넘는 공을 던졌고, 그가 던진 기록만을 이용해 RV에 대한 개별 Random Forest 모형을 적합하고 투구의 질을 예측해보았다. Clayton Kershaw 선수에 대한 2013년부터 2015년까지 투구 기록이 10,284개로 모두 모형을 적합하는 학습용 데이터로 사용했다. 때문에, Clayton Kershaw 선수에 대한 투구의 질 예측은 학습용 데이터를 이용한 Out of Bag Prediction을 이용했다. Out of Bag Prediction이란, 각각의 학습용 데이터 관측치에 대해, 해당 관측이 모형을 형성하는데 사용되지 않았던 의사결정나무로 결과를 예측하는 방법이다. Figure 2.6은 Clayton Kershaw 선수에 대한 투구 수별 평균 RV의 그래프이다. 2013년 시즌으로 적합한 Random Forests 모형의 투구 수

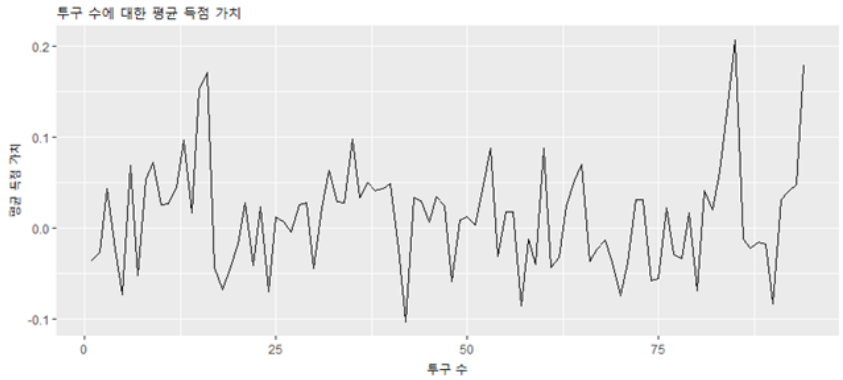

Figure 2.7. Run value of starting pitcher Clayton Kershaw in the 2013/04/01 Los Angeles Dodgers vs. San Francisco Giants game.

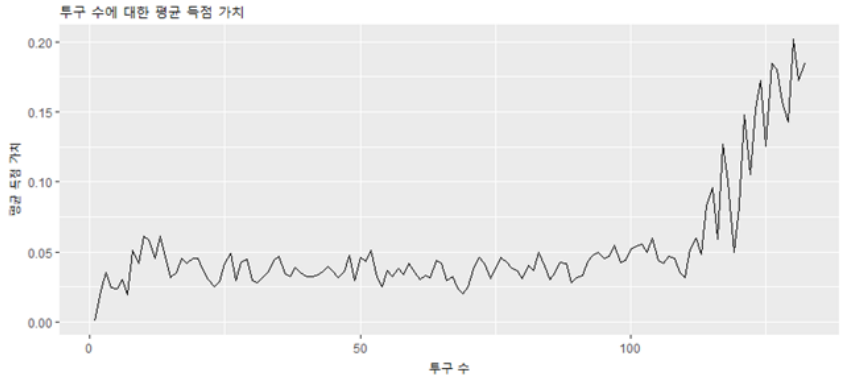

Figure 2.8. Clayton Kershaw's average run value from 2013 to 2015, from the random forests model trained using every starting pitcher data in 2013.

에 대한 평균 RV의 그래프와 비교하면 Clayton Kershaw 선수에 대한 Random Forests 모형에서도 평균 RV가 경기 초반에 상승한 뒤 안정세를 찾아가고, 110구가 넘어가면서 가파르게 증가하는 것을 확인할 수 있다.

Figure 2.7은 Clayton Kershaw 선수의 2013년 4월 1일에 San Francisco Giants를 상대로 선발 출전한 개별 경기에 대한 투구 수에 대한 RV 그래프이다. 매 시즌 각 팀의 구성 변화가 있기 때문에 시즌별 팀의 역량이라는 외부 요인이 존재하지만, 2013년부터 2015년까지 Clayton Kershaw는 일관적으로 좋은 성적을 보여주었기 때문에, 시즌별 개별 모형을 적합할 필요가 없다고 판단하였다. 또한 시즌별 모형을 적합하기에는 각 시즌의 투구 수가 목표 학습량인 10,000개 투구에 턱 없이 모자라기에 2013년부터 2015년까지 3개 시즌 투구 자료를 모두 사용하였다. Clayton Kershaw라는 개별 선수에 대해 적합한 그래프이기 때문에 Figure 2.5의 결과와 비교했을 때, 개별 경기에 대한 투구의 질 점수가 좀 더 안정적인 양상을 보일 것이라고 예측했지만, 개별 경기에 대한 투구의 질 추정은 여전히 불안정한 모습을 보였다.

Figure 2.8은 2013년부터 2015년까지 시즌 자료를 이용해 적합한 Random Forests 모형을 이용해

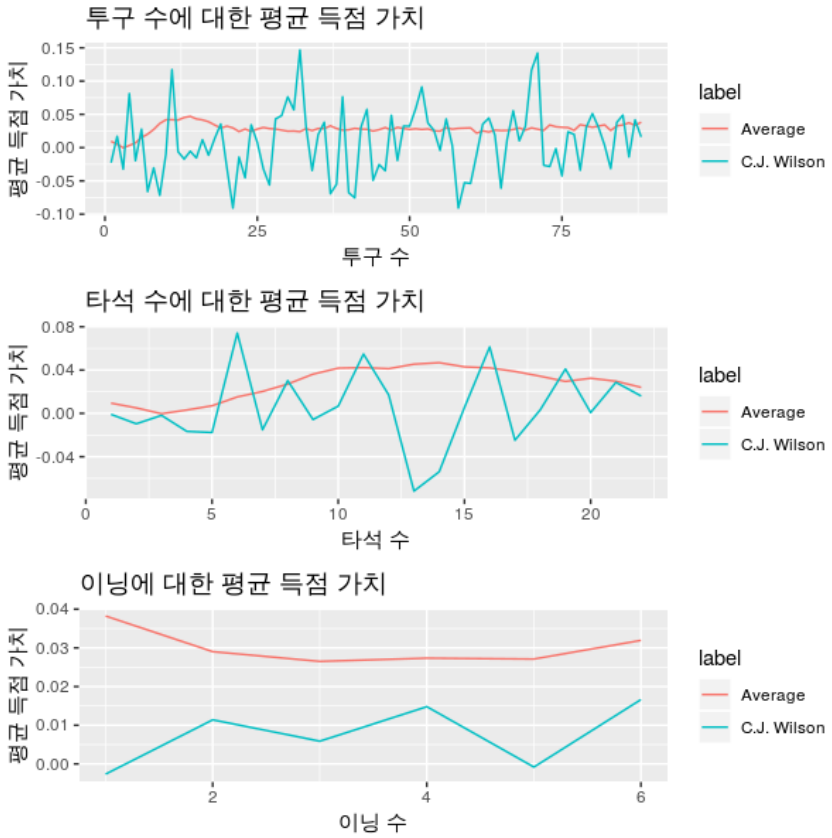

Figure 2.9. Run value of the Los Angeles Angels starting pitcher in the 2013/09/01 Los Angeles Angels vs. Milwaukee game, averaged by pitch count, atbat, and inning.

Clayton Kershaw 선수의 투구 수 별 평균 RV를 추정한 그래프이다. Figure 2.6은 Clayton Kershaw 선수에 대해 개별적으로 적합한 Random Forests 모형을 이용한 그래프로 둘을 비교해 보면, Figure 2.6이 Figure 2.8보다 안정적인 범위 내에서 투구의 질을 추정해주는 것을 확인할 수 있다. 시즌 전체 모형을 각 투수에게 적용하기보다는, 투구 결과가 어느 정도 누적된 선수에 대해서는 개별적으로 모형을 적용하는 것이 더 유리할 수도 있음을 시사한다.

2.3. Moving Average, 타석 평균, 이닝 평균

개별 경기에 대해 추정한 투구의 질 점수를 그래프로 그린 결과는 Figure 2.5과 Figure 2.7에서 살펴볼 수 있듯이 결과가 매우 불안정해 어떠한 경향성이나 특징을 발견하기 어렵다. 보다 시각적으로 명료한 시각화를 구현하기 위한 방법으로 매 투구별 추정한 투구의 질 점수 대신에 타석별, 이닝별 투구의 질 점수를 평균으로 요약한 다음 이에 대한 그래프를 그려볼 수 있다. 혹은 투구의 질 점수에 대해 이동평균을 계산하여 시각화를 하면, 투구의 질 변화 방향에 대한 경향성을 좀 더 쉽게 살펴볼 수 있다. Figure 2.9은 2013/09/01 Los Angeles Angels vs. Milwaukee 경기에서 Los Angeles Angels 선발 투수 C.J. Wilson의 추정 RV를 투구별, 타석별, 이닝별 평균 RV로 요약해 시각화한 그래프이며, Figure 2.10은

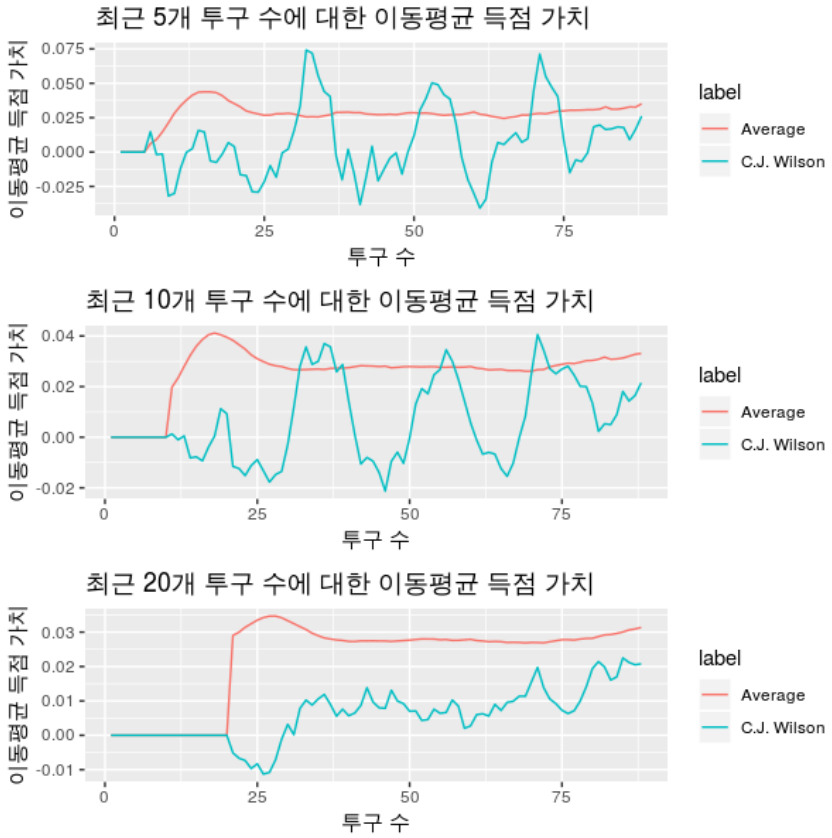

Figure 2.10. Moving average ($n = 5, 10, 20$) run value of the Los Angeles Angels starting pitcher in the 2013/09/01 Los Angeles Angels vs. Milwaukee game.

평균 RV의 이동평균(5개, 10개, 20개)으로 요약해 시각화한 그래프의 비교이다. C.J. Wilson 선수의 RV 추정에는 2013년 MLB 선발 투수의 투구 기록 10,000개를 이용한 Random Forests 모형을 이용하였다. 빨간색으로 나타난 추세선은 2013년 선발 투수 자원들의 평균 Run Value로 C.J. Wilson 선수와의 비교를 위해 제공하였다. 투구 수에 대한 RV 그래프에서는 경향이 쉽게 나타나지 않지만, 타석 수와 이닝에 대한 평균 Run Value를 살펴 보면 C.J. Wilson이 선발 투수 평균에 비해 낮은 RV를 기록한 좋은 성적을 냈음을 확인할 수 있다. 이동평균의 경우도 마찬가지로 10개와 20개 기록을 묶은 이동평균 그래프에서 C.J. Wilson 선수의 RV 기록이 선발 투수 평균보다 우수한 것을 확인할 수 있다. 해당 경기는 Los Angeles Angels가 5 : 3으로 승리했고, C.J. Wilson 선수는 승리 투수가 되었다.

3. 결론

본 연구는 경기를 실점의 위기에서 구해내는 투수의 능력을 평가하기 위해 개별 투구의 질을 Run Value (RV)를 이용해 추정하였다. 그 결과 경기 전반에 대한 선발 투수들의 투구의 질 변화 추이는 물론, 개별 투수에 대한 투구의 질 분석까지 다양한 결과를 확인할 수 있었다. 방대한 야구 지표와 기록이 축적되는

데이터 기반 야구 분석의 시대에 들어섰음에도 불구하고, 지금까지 투수 교체에 대한 자료 기반 진단 시도는 자주 이루어지지 않았다. 선발 투수 교체의 가장 대표적인 규범은 약 100개의 투구 후 교체하는 것인데, 본 연구에서 우리는 투구 수에 따라 투구의 질이 변화하는 각기 다른 양상을 확인하고 비교할 수 있었다. 투구의 질을 이용한다면 앞으로는 각 투수에 대한 맞춤형 진단 및 운용 방안에 대한 제안을 할 수 있을 것이다.

PITCHf/x는 투구를 분석하기 위한 혁신적이고, 방대한 자료를 제공했지만 동시에 주의해야 할 점들이 있다. 우선 아직, PITCHf/x 자료는 완성된 단계가 아니다. PITCHf/x 투구 관측 장비가 처음 도입된 지 약 10년의 시간이 흘렀지만, 아직도 PITCHf/x 방법론은 개선을 거듭하고 있다. 장비의 측정 방식 등이 매년 조금씩 조정되고 그에 따라 관측 값도 바뀌기 때문에 서로 다른 해의 데이터를 비교할 때는 주의해야 한다. 두 번째 주의해야 할 점은 측정 오차이다. MLB 경기장의 PITCHf/x 측정 장비마다 각각의 고유한 측정 오차 값이 존재한다. 물론 많은 경우에 이 측정 오차는 크지 않은 범위 내에 포함되지만, 고려하지 않으면 타당하지 않은 추론 결과를 야기할 수 있다. 세 번째 주의할 점은 PITCHf/x가 분류해서 제공하는 투구 구종은 100% 정확하지 않다. 이는 알고리즘이 아직 완벽하게 구종을 구분하지 못하기 때문이기도 하지만, 한 편으로는 투수에 따라서는 특정한 구종으로 분류하기 어려운 고유한 방식의 구종을 선사하기도 하기 때문이다. 여러 구종의 중간쯤에 위치하는 공을 던지는 투수에 대해서 정확히 구종을 구분하기란 앞으로도 어려운 분류 작업이 될 것이다. 물론 이러한 이유 때문에 type_confidence라는 변수로 분류한 투구에 대한 신뢰도를 표현하고 있지만, 분류 알고리즘도 새로운 시즌이 시작할 때마다 업데이트되기 때문에 주의를 해서 살펴보아야 한다.

야구 경기의 승패는 안타가 아닌 득점으로 결정된다. 선행 연구에서 투구의 질을 기대 피루타수로 추정했지만, 투구가 경기의 승패에 미치는 결과를 직접적으로 살펴보기 위해서 이 논문에서는 각 투구의 RV를 부여하는 방법을 고안했다. 본 모형은 개별 투구의 결과를 단순하게 계량하려는 것이 아니라, 실점이 가능한 상황을 극복해내는 투수의 위기관리능력을 투구의 질로 포착하려 했다. 따라서 피루타수로 투구의 질을 정의하는 선행모형과 RV로 피루타수를 정의하는 본 모형은 측정하고자 하는 반응 변수의 성질과 내용이 달라 일률적인 성능 비교를 하기 어렵다. 그러나 모든 투구에 대해 RV를 부여할 수 있는 한 가지 계산 방법을 제공했기 때문에 경기의 흐름을 확인할 수 있는 자료를 추가해 모형을 적합 한다면, 경기의 승리에 기여하는 투구의 질과 특성을 보다 정확하게 설명할 것으로 예상한다.

References

Fast, M. (2007). Glossary of the Gameday pitch fields, Retrieved from https://fastballs.wordpress.com/2007/08/02/glossary-of-the-gameday-pitch-fields/

Park, C., Kim, Y. Kim, J., Song, J., and Choi, H. (2014). *Datamining with R*, Kyowoo.

Swartz P., Grosskopf M., Bingham D., and Swartz, T. B. (2017). The quality of pitches in Major League Baseball, *The American Statistician*, **71**, 148–154.

Tango, T. M., Lichtman, M. G., Dolphin, A. E. (2007).*Playing the Percentages in Baseball*, Potomac Books, Inc.

PITCHf/x를 이용한 투구의 질 평가

박성민[a] · 장원철[a,1]

[a]서울대학교 통계학과

(2020년 2월 26일 접수, 2020년 3월 20일 수정, 2020년 3월 20일 채택)

요 약

미국 메이저리그 야구 경기는 야구공을 추적하는 3대의 고속 카메라를 통해 모든 투구에 대한 궤적 데이터 PITCHf/x를 수집하고 공개한다. 선행 연구에서는 PITCHf/x 데이터를 통해 각 투구의 기대 피루타수를 계산하고 이를 토대로 투구의 질을 평가했다. 다만 기대 피루타수는 경기 득점으로 매번 이어지지 않기 때문에 각 투구가 승리에 기여하는 영향을 직접적으로 평가하지 못한다. 이 논문에서는 득점 기댓값과 득점 가치의 개념을 조합해 투구에 대한 기대 득점 가치를 계산하고 이를 통해 투구의 질을 랜덤 포레스트 모형으로 평가한 뒤, 기대 피루타수를 이용한 투구의 질 평가와 비교 분석한다.

주요용어: 메이저리그, 투구의 질, 기대 피루타수, 득점 기댓값, 득점 가치, 랜덤 포레스트 모형

본 연구는 정부(과학기술정보통신부)의 재원 및 한국연구재단의 지원 보건복지부의 재원으로 한국연구재단 및 한국보건산업진흥원의 보건의료기술연구개발사업 지원에 의하여 이루어진 것임 (No. 2017R1A2B2012816, HI19C1234).

[1]교신저자: (08826) 서울특별시 관악구 관악로 1, 서울대학교 통계학과. E-mail: wcjang@snu.ac.kr