Journal of the Korean Data & Information Science Society
2015, **26**(4), 865–874

http://dx.doi.org/10.7465/jkdi.2015.26.4.865
한국데이터정보과학회지

시뮬레이션을 통한 프로야구 타자들의 공격능력의 종합적인 평가[†]

김남기[1] · 김선호[2]

[12]전남대학교 산업공학과

접수 2015년 5월 29일, 수정 2015년 6월 9일, 게재확정 2015년 6월 30일

요 약

본 연구에서는 시뮬레이션을 활용하여 타자의 공격능력, 즉 타자로서의 타격능력과 주자로서의 주루능력을 포괄하는 득점생산능력을 종합적으로 평가한다. 이를 위하여, 각 타자의 스코어링 인덱스를 구하는데, 여기서 스코어링 인덱스란 한 팀의 모든 타자가 동일한, 한 선수로만으로 구성되었을 때, 기대되는 경기당 득점이다. 시뮬레이션 입력으로는 2014시즌 한국 프로야구 데이터를 사용하였는데, 주요 출력결과로서 상위 10명의 타자들의 스코어링 인덱스 및 9개 구단과 2014시즌 한국 프로야구의 스코어링 인덱스를 제시한다. 이렇게 구한 스코어링 인덱스는 타자 및 팀의 공격능력의 종합적인 평가뿐만 아니라, 대표선수 및 선발타자의 선정, 선수들의 연봉의 책정 등에도 활용될 수 있을 것이다.

주요용어: 공격능력, 득점, 세이버메트릭스, 시뮬레이션, 프로야구.

1. 머리말

야구에서 타자는 기본적으로 잘 치고, 잘 달려서, 팀이 많은 득점을 올릴 수 있게 기여해야 한다. 2014시즌 한국 프로야구에서 넥센의 서건창은 타율 3할7푼으로 1위, 도루 48개로 3위를 기록한 잘 치고, 잘 달리는 호타준족의 대표적인 선수이다. 한편 같은 시즌, 넥센의 박병호는 홈런 52개로 1위, 타점 124개로 1위를 기록하며, 팀의 득점에 많은 기여를 한 강타자 (slugger)이다. 두 선수 중 어느 선수가 종합적인 공격능력, 즉 타격 및 주루능력을 포괄하는 종합적인 득점생산능력이 더 우수한 타자인가? 라는 질문이 이 연구의 발단이 되었다.

야구에서 타자의 공격능력을 종합적으로 평가하면서, 동시에 간단하게 계산 할 수 있으며, 일반 팬들이 직관적으로 이해하기 쉬운 지표를 개발하기 위한 연구는 미국 프로야구의 세이버메트릭스 (sabermetrics)를 필두로 활발하게 진행되어 왔다. 이들 지표들 중, 가장 쉽고 친숙한 지표는 타수에 대한 안타수의 비율인 타율이다. 하지만, 타율은 안타의 질 (1루타, 2루타, 3루타, 홈런)을 고려하지 않아 타자의 공격능력을 포괄하는 척도라고 보기 어렵다.

이를 보완하기 위해 최근에 많이 활용되고 있는 지표로 OPS (On-base Plus Slugging)와 GPA (Gross Production Average)가 있다. OPS는 출루율과 장타율을 합한 값이고, GPA는 출루율에 1.8을 곱한 후 장타율과 합한 값이다. OPS에서 장타율보다 크기가 작은 출루율이 과소평가 되는 점을 보완

[†] 이 논문은 2014년도 전남대학교의 학술연구비 지원에 의하여 연구되었음.
[1] 교신저자: (500-757) 광주광역시 북구 용봉로 77, 전남대학교 산업공학과, 교수.
E-mail: freedom@jnu.ac.kr
[2] (500-757) 광주광역시 북구 용봉로 77, 전남대학교 산업공학과, 학부생.

하기 위한 지표가 GPA이다. 같은 맥락에서, 한국 프로야구의 실정에 맞는 적절한 가중치를 갖는 가중 OPS를 찾기 위해, Kim (2012)는 한국 프로야구의 경기당 득점과 가중 OPS와의 상관관계를 분석하였다. 2007년부터 2011년 한국 프로야구 데이터의 경우, Kim (2012)는 출루율에 장타율보다 1.33배의 가중치를 두는 가중 OPS가 경기당 득점과의 상관관계가 가장 높다고 발표하였다. 추가로, Lee (2014b)는 1990년부터 2013년 한국 프로야구 데이터의 경우, 출루율에 1.56배의 가중치를 둔 가중 OPS가 경기당 득점과의 상관관계가 가장 높다고 발표하였다. 나아가, Kim과 Kim (2014)는 OPS를 계산할 때, 출루율 대신 희생플라이의 기여도를 보다 정확히 평가하는 수정된 출루율을 사용하는 것을 제안하였다. 이를 바탕으로, Kim과 Kim (2014)는 1982년부터 2013년 한국 프로야구 데이터로부터, 수정된 출루율에 1.5배의 가중치를 둔 가중수정 OPS가 득점 생산성을 잘 설명한다고 발표하였다.

OPS와 OPS의 변형들은 단순한 산술계산을 통하여 누구나 쉽게 계산할 수 있다. 이 보다 계산이 다소 복잡하지만, 여전히 산술계산만을 통해 타자의 득점 생산성을 설명하기 위한 대표적인 세이버메트릭스 지표들로는 RC (Runs Created), XR (eXtrapolated Runs), wOBA (weighted On Base Average)와 이들의 파생지표들이 있다. Lee와 Kim (2005)는 1983년부터 2004년까지의 한국 프로야구 데이터로부터, 세이버메트릭스 주요 지표들 중, XR 추정량과 선형변환된 XR유형 추정량이 한국 프로야구의 득점을 추정하는데 가장 타당한 것으로 발표하였다.

야구경기를 모델링하여 분석하는 방법은 야구경기의 진행을 마르코프체인으로 모델링하는 방법과 시뮬레이션으로 모델링하는 방법 등이다. 마르코프체인으로 모델링한 연구로 가장 두드러진 연구는 국외의 경우는 Bukiet 등 (1997)의 연구이며, 국내의 경우는 Moon 등 (2013)의 연구이다. 야구경기를 마르코프 체인으로 모델링하면 타자의 타격으로 인한 주자상태와 아웃상태의 변화를 세밀하게 묘사할 수 있다. 하지만, 상태변화가 타자에 의해서 전적으로 이루어지다 보니, 주자의 능력으로 인한 진루를 모델링할 수 없다는 단점이 있다. 예를 들어, 타자가 1루타를 쳤을 때, 2루에 있는 주자가 홈까지 진루할 수도 있고, 3루까지만 진루할 수도 있는데, 이것은 실제로 타자의 타격과 함께 주자의 주루능력으로 결정된다. 마르코프체인의 모델링에서는, 이러한 주자의 진루가 전적으로 타자에 의해서만 결정된다. 이것은 마르코프체인 모델링에서 루상에 주자가 있는지 없는지만 상태변수에 고려되며, 루상에 누가 있는지는 일반적으로 상태변수에 반영되지 않기 때문이다.

시뮬레이션을 통한 야구의 연구는 Freeze (1974)에 의해 본격화 되었다. 야구를 시뮬레이션으로 모델링하여 분석한 비교적 최근의 두드러진 연구는 다음과 같다. Sugano (2008)은 마르코프 체인에 기반한 시뮬레이션을 통하여 야구경기의 다양한 양상 (타자-투수 매치업, 경기 중 전략, 베팅전략 등)을 분석하였다. Baumer (2009)와 Beaudoin (2013)는 MLB (Major League Baseball)의 방대한 데이터를 활용하여 야구경기를 시뮬레이션을 통해 매우 세밀하게 묘사하였다. 이들 시뮬레이션들 간의 구체적인 비교는 2.3절을 참고하기 바란다.

본 연구에서는 타자의 공격능력을 시뮬레이션을 활용하여 평가한다. 이 시뮬레이션에서는 야구경기의 진행상황을 프로그램 언어로 묘사한 후, 컴퓨터 상에서 가상적으로 구현한다. 이후, 가상현실 상에서 9회의 야구경기를 수십만 번 진행시키고, 이러한 경기진행을 매우 빠른 속도로 재생한다. 이렇게 하면, 오랜 시간 경과하여야 알 수 있는 결과를 매우 짧은 시간 안에 얻을 수 있다.

본 연구에서는 타자의 공격능력, 즉 타자로서의 타격능력과 주자로서의 주루능력을 포괄하는 득점생산성의 종합적인 평가를 위해, 1번 타자부터 9번 타자까지 모두 동일한 선수로 이루어진 가상의 팀을 구성한다. 예를 들면, 1번 타자부터 9번 타자까지 모두 한화의 김태균으로만 구성된 김태균 팀을 구성한다. 이 후, 이 가상의 팀의 9회 경기를 50만 번 반복하여 경기당 득점 (또는 이닝당 득점)을 구한다. 이러한 방식으로 얻은 평균득점을 D'Esopo와 Lefkowitz (1960)는 스코어링 인덱스 (scoring index; SI)라고 불렀다. 이렇게 SI를 구하면, 타자의 타격능력과 주루능력을 포괄하는 종합적인 지표를 구할 수 있다. 또한, 이렇게 구한 SI는 타율처럼 직관적이고 이해하기 쉬운 장점도 있다. 예를 들어, 어떤 선수의

SI가 10이라고 하면, 이 선수로만 구성된 가상의 팀이 9회 경기 동안 평균적으로 10점을 낸다는 의미이다. 참고로, 타자의 공격능력을 평가하기 위한 최근의 연구로는, Lee (2014a)의 8가지 주요 세이버메트릭스 지표를 통합한 타자등급지표에 대한 연구와 Cho와 Lee (2015)의 타율에 기반된 타격능력의 베이지안 추정이 있다.

본 논문의 구성은 다음과 같다. 2절에서는 시뮬레이션에서 사용한 가정을 설명하고 기존연구들에서 사용한 가정과 비교한다. 3절에서는 시뮬레이션의 입력자료를 설명한다. 특히 주자의 주루능력을 반영하기 위해 세이버메트릭스 지표 중의 하나인 스피드 스코어를 시뮬레이션에서 어떻게 활용하였는지 설명한다. 4절에서는 시뮬레이션 결과를 실제 경기데이터와 비교하고, 시뮬레이션에서 구한 SI값과 기존의 지표들과의 상관관계를 검토한다. 2014시즌 한국 프로야구 데이터로부터, 상위 10명의 타자들의 SI값들과 각 구단별 SI값들, 시즌 전체의 SI 값도 제시한다. 5절에서는 본 연구의 의의와 한계점을 논의한다.

2. 시뮬레이션 모델링

2.1. 시뮬레이션 기본가정

야구경기를 시뮬레이션으로 모델링하는 과정에서 다음을 가정한다.

A.1 타자의 타격가정

(A.1.1) 각 타자의 타격결과는 1루타, 2루타, 3루타, 홈런, 사사구 (볼넷과 사구), 삼진, 땅볼 또는 뜬공아웃 중의 하나로 이루어진다.

(A.1.2) 각 타자의 타격결과는 독립 (independent)이고 동일 (identical)한 확률값을 따른다. 즉, 다른 여타상황 (주자상황, 아웃상황, 득점상황, 타순, 경기진행상황, 이전 타격결과, 상대투수, 상대수비능력, 홈/원정, 컨디션, 심리적 요소 등)에 영향 받지 않으며 (독립) 항상 동일한 확률값을 따라 (A.1.1)의 타격결과 중 하나가 발생한다.

A.2 주자의 진루가정

(A.2.1) 타자가 n루타를 친 경우, 주자는 n개의 루를 '기본' 진루하고, 주자상황 및 아웃 상황, 주자의 주루능력에 따라, 한 루 더 '추가' 진루할 수도 있다. 두 루 이상 추가진루 하거나, 주루중 아웃 되는 경우는 없다.

(A.2.2) 타자가 사사구를 얻은 경우, 주자는 떠밀리는 (forced 되는) 경우에만 진루한다.

(A.2.3) 삼진과 땅볼 또는 뜬공아웃 시, 타자만 아웃 되고, 주자는 진루하지 않는다 (따라서, 주자가 아웃되는 경우는 없다. 땅볼아웃 때의 더블플레이도, 플라이 아웃 때의 태그업도 없다).

A.3 경기 중에 일어나는 작전 (in-game strategy, 희생번트/강공작전, 스퀴즈 작전, 도루작전 등) 및 기타 경기 내외적 상황은 고려하지 않는다.

2.2. 주자의 진루가정 상세

가정 (A.2.1)에서 n루타 시, 주자가 n루 진루하는 것을 '기본진루'라 하고, 기본진루보다 한 루 더 진루하는 것을 '추가진루'라 하자. Fox (2005)가 2000년부터 2004년 동안의 MLB (Major League Baseball)자료를 정리한 표로부터, 추가진루 평균비율을 정리하면 Table 2.1과 같다. 이 표에서 주목할 것은 주자의 진루는 아웃상황에 따라 매우 다른 양상을 띤다는 것이다. 예를 들어 1루타 시, 2루주자가 홈까지 진루한 비율은 노아웃 (0.4217)과 원아웃 (0.5486) 상황보다 투아웃상황 (0.7708)에서 월등히 크다. 이는 투아웃 상황에서는 더블플레이에 대한 부담 없이 주자가 적극적으로 다음 루를 향해 스타트할 수

있기 때문이다. 이에 대한 통계적 유효성에 대한 논의는 Beaudoin (2013)을 참조하기 바란다. 시뮬레이션에서는 이 표의 추가진루 평균비율과 주자의 개인적인 주루능력을 평가한 스피드 스코어를 활용하여 추가진루 여부를 결정한다. 주자가 추가진루를 하지 못하면 기본진루만 하는 것으로 모델링하였다. 이와 관련된 자세한 내용은 3.2절에서 상술한다. 참고로, 주자가 여러 명 있는 경우, 가장 앞선 주자부터 차례대로 진루를 결정하도록 모델링하였다. 이 과정에서, 한 루에 두 명의 주자가 있을 수 없으므로, 뒤따르는 주자가 (추가진루를 통하여) 앞선 주자의 루까지 진루할 수 없도록 시뮬레이션이 구현되었다. 예를 들어, 노아웃 1, 2루 상황에서 1루타 시, 선행주자인 2루주자가 3루까지 밖에 진루하지 못하게 되었다면, 1루주자는 무조건 2루까지만 진루하게 구현되었다.

Table 2.1 Runner advancement (FOX, 2005)

Out	Single, Runner on 1st to 3rd	Single, Runner on 2nd to Home	Double, Runner on 1st to Home
0	0.2627	0.4217	0.3457
1	0.2796	0.5486	0.3636
2	0.3022	0.7708	0.5449

2.3. 시뮬레이션 가정과 기존연구와의 비교

타자의 타격가정인 (A.1)에 대한 기존연구의 가정은 대동소이하다. 땅볼아웃과 뜬공아웃을 구분하여, 보다 더 세밀하게 모델링하느냐 안 하느냐의 차이이다. 반면, 주자의 진루가정인 (A.2)에 대해 선행연구들의 가정은 조금씩 다르다. Bukiet 등 (1997)은 D'Esopo와 Lefkowitz (1960)의 가정을 따라, 1루타 시, 1루주자는 2루까지만 진루하고, 나머지 주자는 모두 득점하는 것으로 가정하였다. 같은 방식으로, 2루타 시, 1루주자는 3루까지만 진루하고, 나머지 주자는 모두 득점하는 것으로 가정하였다. Sugano (2008)은 본 연구에서처럼, 땅볼 또는 뜬공아웃 시, 주자가 진루할 수 없는 것으로 가정하였으나, 견제사, 도루, 희생타, 에러 등을 추가로 고려하여 섬세한 시뮬레이션을 구현하였다. Baumer (2009)는 추가진루 뿐만 아니라 도루와 태그업, 병살타 등을 개별 주자 별로 고려하여, 보다 더 섬세하게 주루상황을 묘사하였다. 한편, Sugano (2008)과 Baumer (2009)는 본 논문에서처럼 아웃상황에 따라 추가진루 확률이 달라지는 점은 고려하지 않았다. Beaudoin (2013)은 땅볼아웃과 뜬공 아웃에 따른 주루상황의 변화를 섬세하게 시뮬레이션으로 구현하였으나, 본 논문에서처럼 개별주자의 주루능력을 고려하지는 않았다. Moon 등 (2013)은 마르코프 모형을 활용하여 타자의 타격으로 인한 야구경기의 진행을 섬세하게 묘사하였다. 반면, 전술한 마르코프 모형의 한계로 인하여, 주자의 개별적인 주루능력을 반영하지 못했다.

2.4. 시뮬레이션 프로그램

시뮬레이션 프로그램은 크게 타자의 타격모듈과 주자의 진루모듈로 구성된다. 이렇게 구성된 시뮬레이션의 프로시져 (procudure)는 다음과 같다. 먼저 타격모듈에서 A.1의 가정에 따라 타자의 타격이 일어난다. 타격결과로 쓰리아웃이 되면 새로운 이닝이 시작되지만, 그렇지 않은 경우는 진루모듈에서 A.2의 가정에 따라 주자별 진루가 이루어진다. 이때 득점이 일어나면 득점을 기록한다. 다시 새로운 타자가 등장하고, 같은 프로시져가 9회까지 진행된다. 9회가 끝나면 경기가 종료되고, 새로운 경기를 시작한다. 이런 방식으로 원하는 경기수만큼 빠른 속도로 반복하면서 경기당 득점수의 평균을 계산한다.

시뮬레이션 프로그램은 C언어를 사용하여 코딩 되었으며, 난수는 C언어에서 제공하는 서브루틴을 활용하여 생성하였다. 시뮬레이션의 반복회수는 50만 경기 (약 3906시즌)로 설정하였다. 한 선수의 SI를 구하기 위한 시뮬레이션 실행시간은 컴퓨터의 성능에 따라 다르지만, 보통의 데스크탑의 경우, 50만경기당 1~2초 내외 정도였다.

3. 시뮬레이션의 입력

3.1. 시뮬레이션 주요 입력자료

2014시즌 한국 프로야구, 각 팀당 128 경기 중, 100타석 이상을 기록한 타자 125명에 대한 KBO 타자자료를 기본적인 입력데이터로 사용하였다 (KBO, 2015). 사용한 타격자료는 타수 (AB), 1루타 (1B), 2루타 (2B), 3루타 (3B), 홈런 (HR), 사사구 (BB+HBP), 삼진 (SO), 뜬공아웃 (AO), 땅볼아웃 (GO)이다. 이들 125명의 자료의 평균값을 갖는 가상의 선수를 '평균'선수라고 하면 이 평균선수의 입력자료는 Table 3.1과 같다.

Table 3.1 Mr. Average's input for Scoring Index (SI)

	AB	1B	2B	3B	HR	BB+HBP	SO+GO+AO
Mr. Avg	295.2	60.0	16.3	1.8	9.0	37.4	208.1

3.2. 스피드 스코어

주자의 개별적인 주루능력을 반영하기 위해서 James (1987)가 제안한 스피드 스코어 (speed score; SS)를 사용하였다. 스피드 스코어는 주자의 주루능력을 평가하는 10점 만점의 척도이다. 이는 세이버메트릭스 지표 중의 하나로서, 상당히 강건한 척도로 알려져 있다 (Sugano, 2008). 상술하면, James의 스피드 스코어는 주루능력과 관련된 여섯 가지 범주 (도루 성공률, 도루 시도율, 3루타 비율, 출루 시 득점비율, 병살타 회피비율, 수비위치)를 0과 10사이의 값으로 평가 한 후, 이들 중 가장 낮은 범주값을 제외한 나머지 다섯 범주값들의 평균값으로 계산한다. 최근에는 수비위치에 대한 범주값을 제외한 나머지 다섯 가지 범주값들만을 고려하여 스피드 스코어를 계산하는데, 이 논문에서도 이 다섯 범주값들의 평균으로 개별 주자의 스피드 스코어를 계산하였다. 각 범주에 대한 계산식은 James (1987)를 참고하기 바란다.

스피드 스코어 계산식에 대입하기 위해 필요한 125명의 KBO 자료는 타수 (AB), 안타 (H), 득점 (R), 1루타 (1B), 3루타 (3B), 홈런 (HR), 사사구 (BB+HBP), 삼진 (SO), 도루 (SB), 도루실패 (CS), 병살타 (GDP)이다. 특별히, '평균'선수의 경우의 입력자료는 Table 3.2와 같다. 이를, James의 계산식에 대입하여 전술한 바와 같이 계산하면 평균선수의 스피드 스코어는 4.74이다. 125명의 스피드 스코어의 평균은 4.19였으며, 이 중 상위 10명을 제시하면 Table 3.3과 같다. 넥센의 서건창의 스피드 스코어가 8.8로 가장 높았다.

Table 3.2 Mr. Average's input for Speed Score (SS)

	AB	H	R	1B	3B	HR	BB+HBP	SO	SB	CS	GDP	SS
Mr. Avg	295.2	87.2	48.0	60.0	1.8	9.0	37.4	54.9	7.4	3.2	6.7	4.7

Table 3.3 Speed Score Top 10 players

	name	AB	H	R	1B	3B	HR	BB+HBP	SO	SB	CS	GDP	SS
1	SGC	543	201	135	136	17	7	67	47	48	17	1	8.80
2	PMW	416	124	87	93	9	1	67	89	50	10	5	8.66
3	KSS	427	123	74	90	8	5	48	77	53	6	7	8.63
4	PHM	310	92	65	76	4	1	43	45	36	8	4	8.08
5	KJH	317	83	58	66	4	2	25	55	22	7	4	7.78
6	CSB	431	132	79	101	7	6	52	68	32	5	8	7.64
7	CDW	443	116	74	97	5	2	47	72	37	7	8	7.59
8	KHW	208	55	32	46	4	0	9	44	4	2	1	7.45
9	SJK	360	105	60	66	6	9	31	92	20	7	6	7.40
10	OJW	397	104	72	68	8	8	58	102	28	12	8	7.31

시뮬레이션에서는 스피드 스코어가 높은 주자의 추가진루확률은 Table 2.1의 추가진루 평균비율보다 높게, 스피드 스코어가 낮은 주자의 추가진루확률은 Table 2.1의 추가진루 평균비율보다 낮게 조정하여, 주루능력이 뛰어난 주자가 더 자주 추가진루하도록 모델링하였다. 구체적인 방법은 Sugano (2008)의 방식을 따라 다음과 같이 하였다. 먼저, 각 선수들의 스피드 스코어 값으로부터 스피드 스코어 백분위 (speed score percentile; SSP)를 구한다. 백분위는 집단에서 주어진 자료의 크기에 따른 상대적인 위치를 나타내는 값이다. 본 논문에서는, 125명의 스피드 스코어를 크기 순서로 배열한 후, 스피드 스코어 n위의 SSP를 $(126-n)/125$로 계산하였다. 그러면 스피드 스코어 1위의 SSP는 125/125=1, 125위인 SSP는 1/125이다. 이후, SSP가 0.85 이상인 경우는 모두 0.85로 처리하고, 0.15이하인 경우는 모두 0.15로 처리하였다.

시뮬레이션에서 사용한 개별 주자의 추가진루 확률 계산방식 다음과 같다 (Sugano, 2008). Table 2.1의 추가진루 평균비율표에서, 해당되는 추가진루 평균비율을 p라고 하자. 그러면, 어떤 주자의 추가진루 확률은, 평균비율 p에 그 주자의 $SSP/(1-SSP)$ 만큼 가중치를 준 값으로 설정한다. 이렇게 하면, SSP가 높은 선수는 추가진루 확률이 평균비율보다 커지게 된다. 즉, 어떤 주자의 추가진루확률은

$$\left[p\frac{SSP}{1-SSP}\right] \Big/ \left[p\frac{SSP}{1-SSP}+(1-p)\right] \tag{3.1}$$

로 설정하고, 기본진루확률은 확률합이 1이 되도록

$$[1-p] \Big/ \left[p\frac{SSP}{1-SSP}+(1-p)\right] \tag{3.2}$$

로 설정한다. 예를 들어, 1사 2루에서, 1루타 시, 2루주자의 SSP가 0.75인 경우, 이 주자가 홈까지 진루할 확률의 계산은 다음과 같다. Table 2.1에 따라, 평균추가진루 확률 $p=0.5486$이고, $SSP/(1-SSP)=3$이므로, 이 주자가 홈까지 추가진루 확률은 $(0.5486)\cdot 3/(0.5486\cdot 3+0.4514)=0.7848$이고, 3루까지만 기본진루할 확률은 $1-0.7848=0.2152$이다.

4. 시뮬레이션의 출력

4.1. 시뮬레이션 결과와 실제 경기데이터와의 비교

시뮬레이션 결과가 실제 경기데이터와 일관된 값을 주는지를 다음과 같이 확인하였다. 먼저 2014시즌 한국 프로야구 9개 각 구단의 타격자료 (1루타, 2루타, 3루타, 홈런, 사사구, 삼진, 뜬공아웃 또는 땅볼아웃)와 주루자료 (타수, 안타, 득점, 1루타, 3루타, 홈런, 사사구, 삼진, 도루성공, 도루실패, 병살타)를 KBO 자료로부터 얻었다 (KBO, 2015). 이에 따르면, 2014 시즌 KIA는 1루타 885개, 2루타 246개, 3루타 27개 등을 기록했다 (Table 4.1). 이러한 자료를 갖는 가상의 선수를 김기아라고 하자. 그러면 1번타자부터 9번타자까지 모두 김기아로만 구성된 가상의 팀을 구성한 후, 전술한 방식으로 시뮬레이션하여 경기당 평균득점을 구하면, KIA의 스코어링 인덱스 (SI)를 구할 수 있다. 마찬가지 방식으로 다른 8개 구단의 SI를 구할 수 있고, 더 나아가 2014시즌 한국프로야구의 SI도 구할 수 있다. 이렇게 구한 SI 값을 2014 시즌 각 구단 및 한국프로야구의 실제 경기당 득점 (runs per game; RPG)과 비교하면 Table 4.1과 같다. SI의 95% 신뢰구간의 1/2길이는 모두 0.01정도였다. 특별히, 2014시즌 한국 프로야구의 경기당 득점 (5.62)에 대비한 SI (5.81)의 오차 (%)는 3.27% 정도였다.

참고로, 2014시즌 한국 프로야구와 각 구단의 스피드 스코어 백분위 (SSP)는, 시뮬레이션에서 고려한 125명의 선수 중, 2014시즌 한국 프로야구와 각 구단의 스피드 스코어 (SS)에 가장 가까운 선수의 스

피드 스코어 백분위를 할당하였다. 예를 들면, NC의 스피드 스코어가 6.19로 가장 높았는데, 이 값과 스피드 스코어가 가장 가까운 롯데의 신본기 타자 (6.20)의 SSP인 0.85를 NC의 SSP로 할당하였다.

Table 4.1 SI for 9 teams and KBO 2014

name	1B	2B	3B	HR	BB+HBP	SO+AO+GO	SS	SSP	SI	RPG	error (%)
DS	930	235	26	108	532	3171	5.18	0.75	5.71	5.37	6.32
HW	900	223	24	105	543	3223	4.24	0.59	5.11	4.84	5.47
KIA	885	246	27	121	478	3199	5.10	0.74	5.48	5.17	5.87
LG	889	213	22	90	587	3192	4.71	0.70	5.14	5.22	-1.46
LT	889	252	26	121	604	3240	4.60	0.66	5.73	5.59	2.35
NC	842	228	36	143	510	3228	6.19	0.85	5.63	5.76	-2.27
NX	818	275	31	199	624	3151	5.39	0.78	7.12	6.57	8.38
SK	903	245	22	114	521	3173	5.64	0.79	5.67	5.73	-0.60
SS	923	238	23	161	569	3162	5.90	0.82	6.59	6.34	3.92
KBO	7979	2155	237	1162	4968	28739	5.23	0.75	5.81	5.62	3.27

4.2. 2014시즌 선수들의 SI

2014시즌 중 100타석 이상을 기록한 타자 125명의 SI에 대해, 상위 10명을 제시하면 Table 4.2와 같다. SI의 95% 신뢰구간의 1/2 길이는 0.015내외였다. 특별히 125명 자료의 평균값을 갖는 (Table 3.1) 가상의 '평균'선수의 SI는 6.06±0.01이었다. 참고로, 평균선수의 스피드 스코어 백분위는 0.5로 설정하여, 평균선수가 Table 2.1의 평균적인 주루능력을 갖도록 하였다. 125명의 SI의 평균은 5.37이었다.

Table 4.2 SI Top 10 players

	name	1B	2B	3B	HR	BB+HBP	SO+AO+GO	SS	SSP	SI
1	KJH	71	36	2	40	81	271	3.80	0.43	12.46
2	PBH	69	16	2	52	108	324	3.99	0.53	10.79
3	Thames	79	30	6	37	65	297	6.39	0.85	10.60
4	SAS	129	25	3	18	85	310	5.28	0.78	10.49
5	KTK	106	30	0	18	81	273	0.93	0.15	9.77
6	SKC	136	41	17	7	67	344	8.80	0.85	9.73
7	CHW	89	33	0	31	57	283	3.23	0.30	9.66
8	Navarro	95	27	1	31	97	351	5.71	0.79	8.88
9	NSB	94	28	5	30	57	321	6.20	0.85	8.83
10	PSM	64	21	0	27	66	247	1.81	0.15	8.65

4.3. SI와 세이버메트릭스 타격지표들과의 상관관계

2014시즌 125명 각 선수들의 SI값들이 기존의 세이버메트릭스 지표값들과 일관된 값을 주는지를 파악하기 위해 상관관계를 분석하였다. 특별히, 타자의 공격능력을 평가하는 세이버메트릭스 지표들 중, Lee (2014a)가 고려한 8가지 지표들과의 상관관계를 분석하였다. 이들 지표들은 Lee (2014a)의 Table 2.1의 식들을 사용하여 계산하였다. 추가로, 가중 OPS의 일종인 GPA, Kim (2012)이 제안한 가중 OPS, Lee (2014b)가 제안한 가중 OPS, Kim과 Kim (2014)가 제안한 가중수정 OPS와의 상관관계도 분석하였다. 이 지표들 중, 8개의 지표들과의 상관계수를 표로 정리하면 Table 4.3과 같다. (상관계수는 미니탭 Release 14를 사용하여 계산하였다.) 출루율과 장타율에 적절한 가중치를 둔 GPA와 Kim (2012), Lee (2014b), Kim과 Kim (2014)의 지표들은 큰 차이 없이 모두 SI와의 상관계수가 높았다.

Table 4.3 Correlation coefficients

RC/27	GPA	Lee (2014b)	TA	wOBA	Kim and Kim (2014)	Kim (2012)	OPS
0.985	0.980	0.979	0.978	0.978	0.977	0.976	0.970

특별히, 타자의 타격능력을 나타내는 대표적인 지표인 OPS와 SI와의 산점도는 Figure 4.1과 같다. 여기서 스피드 스코어 백분위 (SSP)가 0.5보다 큰 경우와 그렇지 않은 경우를 구분하여 타점하였다. 이로부터, 비슷한 수준의 OPS인 경우, 스피드 스코어 (SSP)가 높으면 SI값이 더 높게 나오는 경향이 있다는 것을 확인할 수 있다. 이는 OPS가 타자로서의 타격능력인 출루율과 장타율만을 고려하는 반면, SI가 타격능력에 추가하여 주루능력까지를 포괄하여 고려하기 때문인 것으로 보인다. 즉, 주자의 주루능력이 좋아서 SSP가 높으면, 안타 시 주자의 추가진루 가능성이 높아져서 더 많은 득점을 올리는 것이 SI에 반영되기 때문이다.

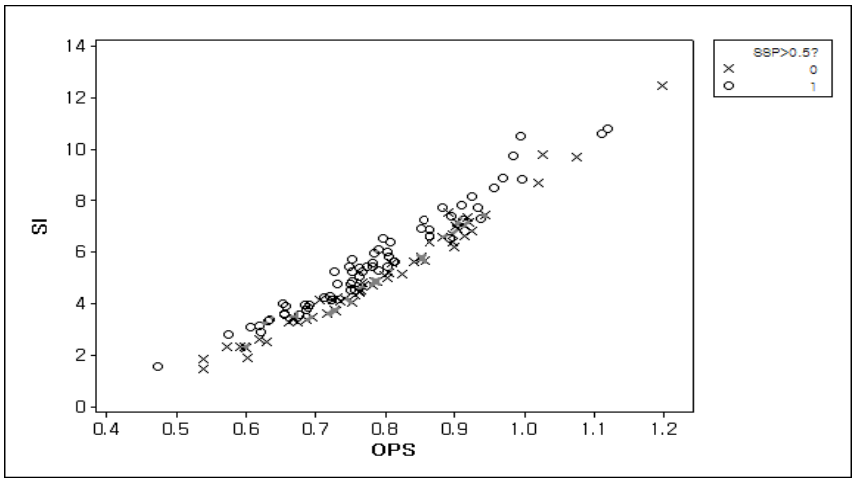

Figure 4.1 Scatter diagram: OPS vs SI

5. 결론

본 연구에서는 타자의 공격능력 즉, 타격능력과 주루능력을 포괄하는 득점생산능력을 종합적으로 평가하기 위하여, 시뮬레이션을 활용하여 각 타자의 스코어링 인덱스를 구하였다. 2014시즌 한국 프로야구 데이터를 사용하여 시뮬레이션 한 결과, 스코어링 인덱스는 실제 경기당 득점수와 5% 내외의 오차를 가졌으며, 기존의 세이버메트릭스 지표들과 비교적 높은 상관관계를 갖았다. 시뮬레이션 결과에 따르면, 서두에서 언급한 서건창의 SI는 9.73이고, 박병호의 SI는 10.79로 나타났다. SI를 통한, 혹은 다른 지표를 통한, 선수들의 우열평가는 매우 조심스럽게 접근 되어야 한다. 선수들은 타자로서의 역할 뿐만 아니라 수비수로서의 역할, 정량적으로 평가 할 수 없는 팀 구성원으로서의 독특한 역할 등이 있다. 따라서 SI를 비롯한 여러 지표들은 선수들의 획일적인 평가의 기준이 되어서는 안 될 것이다.

세이버메트릭스 지표들은 비교적 간단한 산술계산을 통하여 계산할 수 있다. 반면, 시뮬레이션으로 구한 스코어링 인덱스는 시뮬레이션을 구축하는 데 많은 시간과 노력이 들어간다. 하지만, 일단 구축이 되면, 보다 정밀하고 강건한 결과를 기대할 수 있다. 이렇게 구한 스코어링 인덱스는 타자 및 팀의 공격능력의 종합적인 평가 뿐만 아니라, 대표선수 및 선발타자의 선정, 선수들의 연봉의 책정 등에도 활용될 수 있을 것이다.

본 연구의 의의는 다음과 같다. 첫째, 한국 프로야구경기를 시뮬레이션으로 모델링한 점이다. 둘째, 각 주자들의 스피드 스코어를 계산하여 주자의 주루능력을 시뮬레이션 모델링에 반영한 점이다. 이를 바탕으로, 안타 시, 주자들의 추가진루 여부를 주자들의 주루능력과 아웃상황을 모두 고려하여 모델링하였다. 셋째, 2014년 한국 프로야구 타자들의 타격능력과 주자로서의 주루능력을 포괄적으로 고려하

여 공격능력을 종합적으로 평가하는 스코어링 인덱스를 제시한 점이다. 각 구단별 스코어링 인덱스와 2014시즌의 한국 프로야구의 스코어링 인덱스도 추가로 제시하였다. 마지막으로, KBO에서 공시된 자료만을 주로 가공하여 (Table 2.1 제외) 주요결과들을 제시한 점이다.

시뮬레이션을 통한 야구경기의 모델링과 분석의 한계는 주로 입력 데이터의 한계로 기인한다. 미국의 레트로시트 (retrosheet)의 경우처럼, 추후, 한국 프로야구 데이터들을 쉽게 접근할 수 있게 된다면, 보다 더 정교한 모델링이 가능할 것이며, 따라서 보다 더 정확한 분석을 기대할 수 있을 것이다.

6. 사사

논문의 완성도를 크게 높여 주신 심사위원님들게 깊이 감사 드립니다. 또한, 본 논문의 영감을 주고 자료정리를 도와준 한화 이글스의 열성팬, 대전 삼육초의 김태윤군에게도 감사의 뜻을 전합니다.

References

Baumer, B. S. (2009). Using simulation to estimate the impact of baserunning ability in baseball. *Journal of Quantitative Analysis in Sports*, **5**, Iss. 2, Article 8.

Beaudoin D. (2013). Various applications to a more realistic baseball simulator. *Journal of Quantitative Analysis in Sports*, **9**, 271-283.

Bukiet, B., Harold, E. R. and Palacios, J. L. (1997). A Markov chain approach to baseball. *Operations Research*, **45**, 14-23.

Cho, Y. J. and Lee, K. H. (2015). Bayesian estimation of the Korea professional baseball players' hitting ability based on the batting average. *Journal of the Korean Data & Information Science Society*, **26**, 197-207.

D'Esopo, D. A. and Lefkowitz, B. (1960). *The distribution of runs in the game of baseball*, SRI Internal Report, USA.

Fox, D. (2005). *Circle the wagons: running the bases part I*, The Hardball Times, USA.

Freeze, R. A. (1974). An analysis of baseball batting order by Monte Carlo simulation. *Operations Research*, **22**, 728-735.

James, B. (1987). *The Bill James handbook 2007*, 1st Ed., ACTA Publications Skokie, IL.

KBReport. (2015). `http://www.kbreport.com/leader/main`.

Kim, H. J. (2012). Effects of on-base and slugging ability on run productivity in Korean professional baseball. *Journal of the Korean Data & Information Science Society*, **23**, 1065-1074.

Kim, H. J. and Kim, Y. H. (2014). Explanation of run productivity using weighted adjusted OPS in Korean professional baseball. *The Korean Journal of Applied Statistics*, **27**, 731-741.

Korea Baseball Organization. (2015). `http://www.koreabaseball.com/Record/Main.aspx`.

Lee, J. T. (2014a). Measurements for hitting ability in the Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **25**, 349-356.

Lee, J. T. (2014b). Estimation of OBP coefficient in Korean professional baseball. *Journal of the Korean Data & Information Science Society*, **25**, 357-363.

Lee, J. T. and Kim Y. T. (2005). A study on runs evaluation measure for Korean pro-baseball player. *Journal of the Korean Data Analysis Society*, **7**, 2289-2302.

Moon, H. W., Woo, Y. T. and Shin, Y. W. (2013). Analysis of the Korean baseball league using a Markov chain model. *The Korean Journal of Applied Statistics*, **26**, 649-659.

Sugano, A. P. (2008). *A Player Based Approach to Baseball Simulation*, Ph. D. Thesis, University of California, Los Angeles.

Journal of the Korean Data & Information Science Society
2015, **26**(4), 865–874

http://dx.doi.org/10.7465/jkdi.2015.26.4.865
한국데이터정보과학회지

Comprehensive evaluation of baseball player's offensive ability by use of simulation[†]

Nam Ki Kim[1] · Sun Ho Kim[2]

[12]Department of Industrial Engineering, Chonnam National University

Received 29 May 2015, revised 9 June 2015, accepted 30 June 2015

Abstract

This research is to comprehensively evaluate offensive abilities of baseball players who are expected to produce as many runs as possible by their hitting and running. To this end, we establish a simulation program to obtain the so-called scoring index of an individual player. The scoring index of a player is defined as an expected number of runs scored by an imaginary team that is composed of nine copies of the player. As a simulation input, we use 2014 season data of Korean pro-baseball. As a result, we present the scoring indices of top 10 players, 9 Korean pro-baseball teams, and overall 2014 season. The scoring index can serve as a comprehensive evaluation of offensive ability of a player or a team, selection of players for a (national) team or for a starting line-up, estimation of player's worth, and so on.

Keywords: Baseball, offensive ability, runs, sabermetrics, simulation.

[†] This study was financially supported by Chonnam National University, 2014.
[1] Corresponding author: Professor, Department of Industrial Engineering, Chonnam National University, Gwang-Ju 500-757, Korea. E-mail: freedom@jnu.ac.kr
[2] Senior student, Department of Industrial Engineering, Chonnam National University, Gwang-Ju 500-757, Korea.