응용통계연구 제 **20**권 2호, *2007*년, *pp. 205–218*

BRADLEY-TERRY 모형을 이용한 2006 독일 월드컵 예측*

김도현[1] 이상인[2] 김용대[3]

요 약

2006년 독일 월드컵 개최전, 2002년 한.일 월드컵에서와 마찬가지로 한국이 16강에 올라 갈 수 있을까 하는 것은 전 국민의 관심사였다. 이러한 의문을 통계적인 관점에서 알아보고자 객관적인 자료만을 가지고, 독일 월드컵을 예측해 보았다.

경기결과를 예측하는 데는 과거 국가 간의 경기결과가 가장 중요한 자료가 되는데, 국가 간의 경기수가 매우 적고, 심지어는 2006년 독일 월드컵 본선 진출국 중, 특정 국가 간의 전적이 없는 것이 사실이다. 또한 우리가 2002년 월드컵에서 4강 신화를 이루는데는 홈 이점이 작용했다는 사실은 누구도 부인하지 않을 것이며, 기타 다른 요인들이 경기결과에 영향을 미쳤을 것이다.

이러한 점을 반영하여 경기결과를 예측하기 위해서는 전 세계 국가의 경기결과를 하나의 네트워크로 형성하고, 기타 다른 요인들을 고려해야 할 것이다. 우리는 수정된 Bradley-Terry 모형을 가지고 2006년 독일 월드컵 결과를 예측해 보았다.

주요용어: Bradley-Terry 모형, 월드컵 데이터.

1. 서 론

세계는 넓고, 축구를 하는 국가는 많다. 전 세계에 현재 200여개 국가가 있고, 이들 국가 거의 대부분 축구를 하고 있다. 하지만, 양국의 국가 대표팀과 겨루는 A매치는 1년에 국가당 게임수가 그다지 많지 않고, 대륙간 경기는 더욱더 적은 것이 사실이다. 이러한 상황에서 각 국가별로 누가 우세하다는 것을 예측하기란 쉽지 않다. 이러한 점을 극복하기 위해서는 축구의 네트워크를 형성하는 것이다. 즉, 전세계 국가의 경기 결과를 모두 하나로 묶어서 이를 통해서 국가간의 우세정도를 예측할 수 있는 것이다. 예를 들어, 한국과 같은 조인 토고의 경우를 살펴보기로 하자. 한국과 토고는 지금까지 경기를 갖은 적이 없다. 그러나 토고는 여러 아프리카 국가들과 경기를 하였고, 몇몇의 아시아 국가들과도 경기를 하였

* 본 논문은 2005년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2005-070-C00021).

1) (151-747) 서울시 관악구 신림동 서울대학교 통계학과, 박사수료
E-mail: stalove1@snu.ac.kr
2) (151-747) 서울시 관악구 신림동 서울대학교 통계학과, 석사과정
E-mail: lsi44@statcom.snu.ac.kr
3) (교신저자)(151-747) 서울시 관악구 신림동 서울대학교 통계학과, 부교수
E-mail: ydkim@stats.snu.ac.kr

다. 토고와 경기를 갖은 국가 중에 한국과 경기를 갖은 국가도 분명 포함되어 있다. 이를 바탕으로 한국과 토고의 경기결과를 예측할 수 있다.

2002년 한.일 월드컵에서 한국은 4강이라는 놀라운 성적을 거두었다. 이러한 성적을 거둔 결정적인 이유는 홈 이점 즉, 붉은 악마의 응원과 전 국민의 바램이라 할 수 있다. 축구뿐 아니라, 모든 스포츠에서 홈 이점은 경기 결과에 많은 영향을 미친다는 것은 모두들 잘 알고 있을 것이다. 이러한 사실은 국가간의 경기 결과를 경기 장소에 따라 살펴보면 간단하게 확인 할 수 있다. 예를 들어, 한국의 A매치 경기 결과를 경기 장소에 따라 살펴보면 다음과 같다.

표 1.1: 경기장소에 따른 A매치 결과(%)

경기 장소	승	무	패	합
홈	78 (54.9)	39 (27.5)	25 (17.6)	129 (40.1)
원 정	40 (46.5)	20 (23.3)	26 (30.2)	86 (22.7)
제 3국	66 (48.5)	28 (20.6)	42 (30.9)	136 (37.2)

경기 장소와 승,패에 대한 교차분석의 카이제곱 검정 결과, 경기 장소에 따라 승,패의 차이가 유의하게 나왔다($p = 0.041$). 따라서 경기가 어디에서 이루어지는가에 따라 승률이 달라짐을 알 수 있다. 즉,경기 결과에 홈 이점이 작용한다는 사실을 알 수 있다. 그리고 다른 국가의 결과를 살펴보아도 홈 이점이 작용한다는 사실을 알 수 있다. 그러나 월드컵은 개최국에서만 경기를 하기 때문에 개최국인 독일을 제외하고는 홈 이점을 고려할 수 없다. 그러나 분명 같은 대륙인 유럽국가들에게는 어느 정도의 이점이 있을 것이다. 따라서 본 연구에서는 홈 이점 이외에도 독일에서의 대륙간 거리를 변수로 고려하였고, 월드컵의 경기 결과를 예측하기 때문에 과거의 월드컵에서의 성적(참여수, 역대 최고성적)을 변수로 고려하였다.

본 연구에서는 국가간 경기 전적이 많은 상위 100개 국가의 A매치 결과와 각 국가간의 변수(대륙간 거리, 월드컵 참여수, 최고성적)를 고려하여 2006 독일 월드컵을 예측해 보았다. 분석 방법으로는 축구 경기 결과와 같은 쌍을 이룬 데이터(paired comparison data)를 분석하기에 적합한 Bradley-Terry 모형을 기초로하고, 위에서 언급한 변수를 고려하여 분석하였다. 또한 모형을 통해 얻은 결과와 2002년, 2006년 실제 월드컵 결과와 비교함으로써 여기에서 제시한 모형이 타당한가를 알아볼 것이다. 본 연구의 목적은 한국이 16강에 올라갈 확률을 구하는 것이므로, 한국과 같은 조인 국가를 중점으로 하여 내용을 전개해 나갈 것이다.

제 2절에서는 분석방법의 기초인 일반적인 Bradley-Terry 모형에 대해 살펴보고, 제 3절에서는 분석에서 사용한 데이터에 대한 설명과 분석을 위한 구체적인 모형을 살펴보고, 제 4절에서는 한국과 같은 조인 국가들의 기술통계와 분석을 통해 얻은 결과를 살펴보고, 제

5절에서는 분석결과와 실제결과를 비교함으로써 모형의 타당성을 알아볼 것이다.

2. Bradley-Terry 모형

본 연구에서 분석 방법의 기초로 고려한 쌍별비교(paired comparison) 모형은 Bradley-Terry 모형이다(Bradley와 Terry, 1952). p개의 팀들이 어느 대회에서 경기한다고 가정하자. 그리고 각 팀 i는 모수 π_i에 관련되어 있다고 가정한다. 가치 모수(worth parameter)라 불리는 이러한 모수는 각 팀의 능력치(ability), 또는 힘(strength)이라 할 수 있다. Bradley-Terry 모형은 어느 한팀이 다른 한팀을 이길 확률을 다음과 같이 나타낸다.

$$p_{ij} = \Pr(i \text{ defeats } j) = \frac{\pi_i}{\pi_i + \pi_j}. \tag{2.1}$$

여기에서 π_i와 π_j는 양수값을 갖는 모수(positive-value parameters)로써 각 팀의 능력치로 생각 할 수 있다. 위 모형은 무승부(tie)를 고려하지 않았으므로, $p_{ij} + p_{ji} = 1$ 이다. 또한 모형은 다음과 같이 재표현 할 수 있다.

$$p_{ij} = \Pr(i \text{ defeats } j) = \frac{e^{\gamma_i}}{e^{\gamma_i} + e^{\gamma_j}}. \tag{2.2}$$

여기에서 $\gamma_i = \log \pi_i, \ \forall i$.

Bradley-Terry 모형은 여러가지 방법으로 이끌어 낼 수 있다. 한가지 유도 방법(Davison, 1970)을 살펴보면, 팀 i가 경기를 할때, 그 팀은 아래와 같은 누적 분포 함수를 가지고, 상대팀과 독립적인 관측되지 않은 점수(unobserved score) S_i를 얻는다고 가정한다.

$$S_i \sim F_i(s) = \exp(-e^{-(s-\log \pi_i)}).$$

따라서 팀 i는 위치 모수(location parameter) $\log \pi_i$를 가진 극한 분포로부터 랜덤 점수를 얻는다. 이것은 점수차 $S_i - S_j$의 분포는 평균이 $\log \pi_i - \log \pi_j$인 로지스틱 분포를 따른다. 즉,

$$S_i - S_j \sim F_{ij}(s) = \frac{1}{1 + e^{-(s-(\log \pi_i - \log \pi_j))}}.$$

이것은 (2.1)에서와 같은 결과를 의미한다.

$$\Pr(S_i > S_j) = \Pr(S_i - S_j > 0) = 1 - \frac{1}{1 + e^{\log \pi_i - \log \pi_j}} = \frac{\pi_i}{\pi_i + \pi_j}. \tag{2.3}$$

이러한 모형 아래에서 쌍을 이룬 데이터(paired comparison data)를 분석하기 위해서, 팀 i와 j는 n_{ij}번 경기를 하였고, 이 중에서 팀 i는 y_{ij}번 경기에서 이겼고, $n_{ij} - y_{ij} = y_{ji}$ 번 경기에서 졌다고 가정한다. 이때, $\boldsymbol{y} = (y_{ij}, \ \ i, j = 1, 2, \ldots, p)$의 분포는 다음과 같다.

$$f(\boldsymbol{y}|\boldsymbol{\pi}) = \prod_{i<j} \binom{n_{ij}}{y_{ij}} \left(\frac{\pi_i}{\pi_i+\pi_j}\right)^{y_{ij}} \left(\frac{\pi_j}{\pi_i+\pi_j}\right)^{y_{ji}}. \tag{2.4}$$

여기에서 $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$는 Bradley-Terry 가치모수(worth parameter)이다.

$\boldsymbol{\pi}$에 대한 우도함수(likelihood)는 다음과 같이 표현할 수 있다.

$$\text{Lik}(\boldsymbol{y}|\boldsymbol{\pi}) \propto \frac{\prod_{i=1}^{p} \pi_i^{y_i}}{\prod_{i<j}(\pi_i+\pi_j)^{n_{ij}}}, \tag{2.5}$$

(여기에서 $y_i = \sum_{j=1}^{p} y_{ij}$는 i팀이 이긴 전체 경기수)

여기에서 $\boldsymbol{\pi}$에 대한 최대 우도 추정량(maximum likelihood estimates)은 보통 Newton-Raphson algorithm에 의해 얻어질 수 있다.

위에서 언급한 Bradley-Terry 모형은 단지, 경기 결과를 승과 패만을 고려한 것이다. 그러나 많은 스포츠에서 경기 결과는 승패 이외에도 무승부가 될 수도 있다. 무승부를 고려한 모형은 Davison(1970)에 의해서 제안되었다.

이러한 모형은 한팀이 다른 한팀을 이길 확률과 질 확률, 비길 확률을 다음과 같이 가정한다.

$$\begin{aligned}
p_{ij} &= \Pr(i \text{ defeats } j) = \frac{e^{\gamma_i}}{e^{\gamma_i} + e^{\gamma_j} + e^{\lambda+(\gamma_i+\gamma_j)/2}}, \\
p_{ji} &= \Pr(j \text{ defeats } i) = \frac{e^{\gamma_j}}{e^{\gamma_i} + e^{\gamma_j} + e^{\lambda+(\gamma_i+\gamma_j)/2}}, \\
p_{ij,0} &= \Pr(i \text{ ties } j) = \frac{e^{\lambda+(\gamma_i+\gamma_j)/2}}{e^{\gamma_i} + e^{\gamma_j} + e^{\lambda+(\gamma_i+\gamma_j)/2}}.
\end{aligned} \tag{2.6}$$

여기에서 모수 λ는 무승부와 관련된 모수이다. 큰 양수 λ값은 무승부의 확률이 높다는 것을 의미한다. 식 (2.7)은 하나의 모수 λ를 추가함으로써, 무승부를 고려한 모형이 충분하다고 가정하고 있다. 더 나아가서 팀별로 각각 모수 λ_{ij}를 추가함으로써, 무승부를 고려한 위 모형을 일반화할 수 있다(Singh와 Gupta, 1978).

이러한 무승부를 고려한 모형에서 쌍을 이룬 데이터(paired comparison data)를 분석하기 위해서 팀 i와 팀 j는 n_{ij}번 경기를 하였고, 이 중에서 팀 i는 W_{ij}번 이겼고, D_{ij}번 비겼고 L_{ij}번 졌다고 가정한다($n_{ij} = W_{ij} + D_{ij} + L_{ij}$). 이때, 우도함수는 다음과 같이 표현할 수 있다.

$$L = \prod_{i<j} (p_{ij})^{W_{ij}} (p_{ji})^{L_{ij}} (p_{ij,0})^{D_{ij}}. \tag{2.7}$$

3. 분석 모형

이 절에서는 월드컵 분석을 위한 모형을 제시할 것이다. 이 모형은 2절에서 설명한 Bradley-Terry 모형에 기초를 두고 있다. 분석 모형을 제시하기전, 3.1절에서 데이터에 대한 간략한 설명을 할 것이다.

3.1. 데이터 설명

분석에 사용한 데이터는 100개 국가에 대한 최근 20년간(1985년 ~ 2005년) A매치 경기 결과이다. 과거의 경기 결과는 최근 국가의 능력치에 거의 영향을 주지 않으므로 분석에서 제외하였다. 데이터는 전체 11,958 A매치 경기 결과로 구성되어 있다. 경기 날짜, 경기 장소(매 경기마다 각 팀의 홈팀여부) 또한 포함하고 있다. 그리고 고려된 변수로는 각 팀별로 최근 5회 월드컵 대회에서 참여수, 최고 성적, 홈팀인 독일에서의 대륙간 거리가 있다. 피파랭킹 또한 변수로 고려하려고 하였으나, 이것은 경기결과에 너무 많은 영향을 미치므로 변수에서 제외하였다. 여기에서 100개 국가의 선정 기준은 전체 경기의 승점과 승률을 기준으로 상위 100개 국가를 선정하였다. 아래표는 데이터의 이해를 돕기 위한 데이터의 일부를 나타낸 것이다.

team A	team B	HOME	dist A	dist B	cham A	cham B	high A	high B	win	loss	draw
Korea Republic	Japan	-1	0	0	5	2	3	1	1	0	0
Korea Republic	Japan	-1	0	0	5	2	3	1	0	0	1
Korea Republic	Japan	-1	0	0	5	2	3	1	1	0	0
Korea Republic	Japan	-1	0	0	5	2	3	1	0	0	1
Korea Republic	Korea DPR	0	0	0	5	0	3	2	1	0	0
Korea Republic	Korea DPR	1	0	0	5	0	3	2	1	0	0
Korea Republic	Korea DPR	0	0	0	5	0	3	2	1	0	0
Korea Republic	Korea DPR	0	1	1	5	0	3	2	1	0	0
Korea Republic	Kuwait	0	0	1	5	0	3	0	1	0	0
Korea Republic	Kuwait	0	1	0	5	0	3	0	0	1	0
Korea Republic	Kuwait	0	0	1	5	0	3	0	0	1	0
Korea Republic	Kuwait	0	0	1	5	0	3	0	0	1	0
Korea Republic	Kuwait	-1	1	0	5	0	3	0	0	1	0
Korea Republic	Kuwait	0	0	1	5	0	3	0	1	0	0
Korea Republic	Malaysia	0	0	0	5	0	3	0	1	0	0
Korea Republic	Malaysia	0	0	0	5	0	3	0	0	1	0
Korea Republic	Malaysia	1	0	0	5	0	3	0	1	0	0
Korea Republic	Malaysia	1	0	0	5	0	3	0	1	0	0
Korea Republic	Malaysia	-1	0	0	5	0	3	0	0	1	0
Korea Republic	Malaysia	-1	0	0	5	0	3	0	0	1	0
Korea Republic	Malaysia	0	0	0	5	0	3	0	1	0	0
Korea Republic	Mexico	0	2	0	5	4	3	2	0	1	0
Korea Republic	Mexico	-1	2	0	5	4	3	2	0	1	0
Korea Republic	Mexico	1	0	2	5	4	3	2	0	0	1
Korea Republic	Mexico	1	0	2	5	4	3	2	1	0	0
Korea Republic	Mexico	0	2	1	5	4	3	2	0	1	0
Korea Republic	Morocco	0	1	1	5	3	3	1	0	0	1
Korea Republic	Netherlands	0	2	0	5	3	3	3	0	1	0
Korea Republic	New Zealand	0	0	2	5	0	3	0	1	0	0
Korea Republic	New Zealand	0	2	2	5	0	3	0	0	0	1
Korea Republic	New Zealand	-1	2	0	5	0	3	0	1	0	0
Korea Republic	New Zealand	0	2	0	5	0	3	0	1	0	0
Korea Republic	Nigeria	1	0	2	5	3	3	1	0	0	1
Korea Republic	Nigeria	1	0	2	5	3	3	1	1	0	0
Korea Republic	Norway	0	2	2	5	2	3	2	1	0	0
Korea Republic	Oman	0	0	1	5	0	3	0	1	0	0
Korea Republic	Qatar	0	0	1	5	0	3	0	0	0	1
Korea Republic	Qatar	-1	1	0	5	0	3	0	1	0	0
Korea Republic	Romania	1	0	2	5	3	3	2	0	1	0
Korea Republic	Russia	0	0	2	5	4	3	3	0	0	1
Korea Republic	Saudi Arabia	1	0	1	5	3	3	1	0	0	1

그림 3.1. 데이터 일부

i번째 경기에서 team A가 홈이면, $h_i = 1$, 팀 team B가 홈이면 $h_i = -1$, 월드컵이나 세계대회 같이 다른 국가에서 경기를 했다면, $h_i = 0$. dist A와 dist B는 각 국가에서 독일과의 거리를 0, 1, 2로 환산한 값을 나타낸다. 만약 A국가가 유럽에 있는 국가라면 dist A = 0이다. 왜냐하면, 경기가 유럽국가인 독일에서 열리므로 유럽국가들은 시차적응, 응원단 등등 여러가지 이유로 다른 대륙의 국가들보다 이점이 있다고 볼 수 있기 때문이다. 또한 cham A, cham B는 각 국가의 최근 5개 월드컵에서의 참여수이고, high A, high B는 역대 월드컵 최고 성적을 나타내는 변수로써 4점은 최고성적이 우승 또는 준우승, 3점은 4강, 2점은 8강, 1점은 16강인 것이다. Win, loss, draw는 A국가를 기준으로 A국가가 이기면 win=1, B가 이기면 loss=1, 비기면 draw=1을 나타내고 있다.

3.2. 분석 모형

이 절에서는 3.1절에서 언급한 데이터를 분석하기 위한 구체적인 모형을 제시할 것이다. A_i는 i번째 경기에서의 한 국가, B_i는 i번째 경기에서의 다른 한 국가, x_{j,A_i}는 i번째 경기에서 A국가의 j번째 변수라 하자. 예를 들면 x_{2,A_1}는 첫번째 경기에서 team A의 월드컵 참여수를 나타낸 것이다. 그리고 국가 A와 B는 각각 가치모수(worth parameter) $\gamma_{A_i}, \gamma_{B_i}$와 관련되어 있고, 또한 경기 결과는 고려한 변수에 영향을 받는다고 가정한다. 모형을 간략하게 표현하기 위해, 다음과 같은 기호를 사용한다.

- $\mathrm{I}_i = \exp\{\gamma_{A_i} + \frac{\delta}{2} h_i + \sum_j \alpha_j x_{j,A_i}\}$,
- $\mathrm{II}_i = \exp\{\gamma_{B_i} - \frac{\delta}{2} h_i + \sum_j \alpha_j x_{j,B_i}\}$,
- $\mathrm{III}_i = \exp\left\{\frac{\gamma_{A_i}+\gamma_{B_i}+\sum_j \alpha_j (x_{j,A_i}+x_{j,B_i})}{2} + \lambda\right\}$,
- $\Delta_i = \mathrm{I}_i + \mathrm{II}_i + \mathrm{III}_i$.

본 연구에서 제시하는 분석모형은 한 국가가 다른 한 국가를 이길 확률을 다음과 같이 나타낼 수 있다.

$$P_{AB} = P(A \ \text{defeats} \ B) = \frac{\mathrm{I}}{\mathrm{I} + \mathrm{II} + \mathrm{III}}. \tag{3.1}$$

여기에서 모수 δ는 홈 이점, λ는 무승부, α_j는 j번째 변수에 관련된 모수이다. 즉, 큰 δ 값은 홈 이점이 경기 결과에 많은 영향을 미치고, 마찬가지로 α_j가 크다는 것은 j번째 변수가 결과에 많은 영향을 미친다는 것을 의미한다. 또한 모수의 추정을 위해서 우도함수를 다음과 같이 표현할 수 있다.

$$L = \prod_{i=1}^{N} \left(\frac{\mathrm{I}_i}{\Delta_i}\right)^{W_i} \left(\frac{\mathrm{II}_i}{\Delta_i}\right)^{L_i} \left(\frac{\mathrm{III}_i}{\Delta_i}\right)^{D_i}. \tag{3.2}$$

여기에서 N은 총 경기수이고, A_i 국가가 승리하면 $W_i = 1, L_i = D_i = 0$, B_i 국가가 승리하면 $L_i = 1, W_i = D_i = 0$, 두 국가가 비기면 $D_i = 1, W_i = L_i = 0$이다.

모수 $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{100}, \delta, \lambda, \alpha_1, \alpha_2, \alpha_3)$에 대한 최대우도 추정량(maximum likelihood estimates)은 Newton-Raphson algorithm으로 구하였다. Newton-Raphson algorithm은 최대우도 추정량을 구하는 널리 알려진 방법으로써, 임의의 초기값에 대하여 식 (3.3)과 같은 갱신규칙(update rule)을 수렴(converge)할 때까지 적용하여 최대우도 추정량을 구할 수 있다.

$$\hat{\gamma}_{t+1} = \hat{\gamma}_t - \left\{ \left(\frac{\partial^2 L(\gamma)}{\partial \gamma^2} \right)^{-1} \frac{\partial L(\gamma)}{\partial \gamma} \right\} \Big|_{\gamma = \hat{\gamma}_t}. \tag{3.3}$$

여기에서 최대우도 추정량은 공집합이 아닌 부분집합(non-empty subset)에 국가의 모든 부분(partition)이 있다면 유일(unique)하다는 것이 알려져 있다(Ford, 1957). 이 조건은 어떤 국가가 모든 경기를 진다면, 최대우도 추정량은 존재하지 않는다는 것을 의미한다.

4. 분석 결과

이 절에서는 3절에서 설명한 모형을 토대로 분석한 결과를 제시 할 것이다. 우선, 3.1절에서는 한국의 기술통계량과 16강 티켓을 놓고 싸울 스위스의 기술통계량을 제시함으로써 한국과 스위스를 비교할 것이다.

4.1. 한국의 기술 통계량

한국은 1954년 일본과의 A매치 첫 경기 이후, 최근까지 총 364경기의 A 매치를 치루었다. 총 364경기 중, 184승으로 전적만 보면 한국이 축구 강국인것처럼 보이지만, 현실은 그렇지 않다. 다음 표 4.1은 한국의 A매치 결과를 대륙별로 살펴본 것이다.

표 4.1: 한국의 대륙별 A매치 결과(%)

대 륙	승	무	패	합
아시아	131 (64.9)	38 (18.8)	33 (16.3)	202 (40.1)
유 럽	19 (27.9)	20 (29.4)	29 (42.6)	68 (22.7)
아프리카	13 (44.8)	10 (34.5)	6 (20.7)	29 (37.2)
북 미	8 (33.3)	8 (33.3)	8 (33.3)	24 (37.2)
오세아니아	9 (45.0)	6 (30.0)	5 (25.0)	20 (37.2)
남 미	4 (19.0)	5 (23.8)	12 (57.1)	21 (37.2)

표 4.1에서 알 수 있듯이, 한국은 A매치 전 경기의 반 이상을 아시아 국가와 하였고, 승리의 대부분을 아시아 국가들과의 경기에서 한 것임을 알 수 있다. 그러나 다른 대륙들과의

경기에서는 좋지 못한 성적을 거두었고, 특히 유럽과 남미 국가들을 상대로는 많은 경기를 하지는 않았지만, 약하다는 사실을 알 수 있다. 즉, 한국은 약팀과 많은 경기에서 승리하여 A매치 전적이 좋다는 결과를 얻을 수 있다. 이에 반해, 스위스는 총 266경기 중, 105승으로 전적만 보면 한국 축구가 더 우위에 있는 것으로 보이지만, 스위스의 결과도 대륙별로 살펴보면 아니라는 것을 쉽게 알 수 있다. 표 4.2는 스위스의 A매치 결과를 대륙별로 나타낸 것이다.

표 4.2: 스위스의 대륙별 A매치 결과(%)

대 륙	승	무	패	합
아시아	7 (87.5)	0 (0.0)	1 (12.5)	8 (40.1)
유 럽	88 (37.4)	60 (25.5)	87 (37.0)	235 (22.7)
아프리카	3 (60.0)	1 (20.0)	1 (20.0)	5 (37.2)
북 미	5 (55.6)	3 (33.3)	1 (11.1)	9 (37.2)
남 미	2 (22.2)	4 (44.4)	3 (33.3)	9 (37.2)

스위스는 총 266경기중, 235경기를 같은 대륙인 유럽국가와 한 것을 알 수 있다. 승률은 한국에 다소 미치지 못하지만, 대부분 아시아 보다 수준이 높은 유럽국가와 한 것으로 보아 한국보다 실력이 조금 높다거나 비슷하다고 평가할 수 있다. 다음 절에서는 이러한 경기 결과를 모형에 적합시켜 국가들의 이길 확률을 추정할 것이다.

4.2. 월드컵 예측

1952년에 Bradley와 Terry에 의해 제안된 Bradley-Terry 모형은 네트워크가 형성되어 있는 경기에서 팀간의 이길 확률을 추정할 수 있도록 제안되었다. 처음에 제안된 모형은 무승부는 고려할 수 없는 모형이었으나 이후 여러 통계학자에 의해 무승부를 다룰 수 있고, 다양한 변수 또한 고려할 수 있는 모형으로 발전하였다. 3절에서는 이러한 모형을 기초로 해서 우리의 데이터를 분석하기에 적합한 모형을 제시하였다. 모형을 통해서 각 국가의 능력치와 연관된 모수 γ_i $(i = 1, \ldots, 100)$를 추정하고, 무승부, 변수, 홈 이점에 관련된 모수 λ, δ, α_i $(i = 1, 2, 3)$를 추정함으로써 각 국가가 이길 확률, 비길 확률, 질 확률을 추정할 수 있다. 표4는 모수의 추정값의 일부를 나타낸 것이다. 위 모형으로 구한 가치모수의 추정값은 알파벳 순서인 앙골라 0을 기준으로 상대적인 것이다. 즉, 0보다 크면 기준인 앙골라 보다 강하다는 것을 의미하고, 0보다 작으면 약하다는 것을 의미한다. 또한 값이 크면 클수록 그 국가의 능력치가 높다는 것을 의미한다. 그리고 다른 모수들에 비해 홈이점에 관한 모수의 추정값이 높게 나왔다. 이것은 각 국가의 능력치 이외에도 경기 결과에 홈 이점이 많

표 4.3: 각 국가의 가치모수(Worth parameter)와 그 외 모수에 대한 추정값

한 국	독 일	폴란드	코스타리카	에쿠아도르
3.4093	5.2833	4.0971	2.9959	3.8511
프랑스	잉글랜드	스웨덴	파라과이	토바고
5.5284	5.0993	4.6437	4.3814	2.0266
스위스	네덜란드	아르헨티나	세르비아	코트디 부아르
3.8609	5.0615	5.4345	4.7117	1.4998
토 고	포르투칼	멕시코	이 란	앙골라
1.3111	4.7365	4.2297	3.1584	0
스페인	이탈리아	체 코	미 국	가 나
5.0502	5.2664	5.0129	3.3646	2.3974
튀니지	브라질	크로아티아	호 주	일 본
2.5561	5.6979	4.7924	3.2213	2.9594
우크라이나	사우디아리비아	홈이점	무승부	대륙간거리
3.5609	2.8743	0.8850	0.1219	0.0111
참여수	최고성적			
0.0999	0.1173			

은 영향을 미친다는 것을 의미한다.

본 연구의 목적은 각 국가가 16강에 올라 갈 확률을 추정하는 것인데, 우리가 제시한 모형으로는 이러한 확률을 구할 수가 없다. 따라서 16강에 올라 갈 확률은 모형으로 구한 각 국가의 확률을 토대로 시뮬레이션을 통하여 구하였다. 다음은 시뮬레이션의 과정을 간단히 나타낸 것이다.

1. 추정한 모수를 통해 식 (3.1)을 이용하여 국가간 승,무,패 확률을 구한다.
2. 위에서 구한 국가간 승,무,패 확률을 토대로 다항 확률변수 발생(Multinomial random number generate)을 통해 각 조마다 16강 예선전 모든 경기결과(6경기)를 얻는다.
3. 승점 산출 방식을 적용하여 각 조별로 4개 국가의 순위(ranking)를 구한다.
4. 단계 2,3을 10000번 반복한다.
5. 10000번의 모의 실험 중, 한 국가의 순위가 2위 안에 있는 실험의 수를 센다.
6. 한 국가가 16강에 올라 갈 확률은 「2위 안에 있는 실험의 수/10000」으로 구할 수 있다.

이러한 과정을 통하여 얻은 각 조의 각 국가가 16강에 올라 갈 확률은 아래의 표와 같다. 표 4.4의 결과를 살펴보면, G조에서는 시드를 받은 프랑스가 다른 국가들에 비해 현저

하게 강하다는 사실을 알 수 있다. 그리고 한국이 스위스에 비해 다소 미치지 못하지만, 비슷하다는 사실을 알 수 있다. 이에 반해 월드컵에 첫 출전한 토고는 상대적으로 아주 약하다는 결론을 이끌어 낼 수 있다. 16강에 올라 갈 확률을 살펴보면, 프랑스가 조 1위로 무난히 16강에 올라 갈 것으로 보이고, 스위스가 다소 높지만, 한국과 2위자리를 놓고 치열한 경쟁을 할 것으로 생각되어진다.

표 4.4: G조 각 국가의 16강에 올라 갈 확률(%)

국 가	한 국	프랑스	스위스	토 고
확 률	48.5	98.9	51.1	1.5

표 4.5: G조 각 국가의 승패에 관한 확률(%)

국 가	앞팀 승	뒷팀 승	무승부
한 국 대 프랑스	8.1	70.7	21.2
한 국 대 스위스	32.9	36.4	30.7
한 국 대 토 고	77.2	5.1	17.6
프랑스 대 스위스	69.3	8.8	21.9
프랑스 대 토 고	92.2	0.7	7.1
스위스 대 토 고	78.3	4.7	17.0

표 4.6: 각 조 각 국가의 16강에 올라 갈 확률(%)

A 조	독 일	폴란드	코스타리카	에쿠아도르
확 률	99.7	51.6	15.8	32.9
B 조	잉글랜드	스웨덴	파라과이	토바고
확 률	79.9	69.8	47.0	3.3
C 조	네덜란드	아르헨티나	세르비아	코트디 부아르
확 률	67.3	75.1	57.4	0.3
D 조	포르투칼	멕시코	이 란	앙골라
확 률	88.7	72.9	37.8	0.6
E 조	이탈리아	체 코	미 국	가 나
확 률	92.9	82.9	18.1	6.1
F 조	브라질	크로아티아	호 주	일 본
확 률	92.3	77.8	20.1	9.9
H 조	스페인	우크라이나	튀니지	사우디아라비아
확 률	95.0	60.9	14.0	30.0

5. 모형의 타당성

우리의 모형이 타당한가를 알아보기 위해서 분석 결과와 실제 결과를 비교할 것이다. 우선, 2006 독일 월드컵을 비교하기 전에 2002 한·일 월드컵을 비교할 것이다. 2002년에 한국은 4강이라는 놀라운 성적을 거두었다. 이것은 이변이었을까?

다음은 2002년 월드컵을 모형을 통해 분석한 결과와 실제 결과를 비교한 것이다. 데이터는 2002년을 기준으로 지난 20년간 자료를 사용하였고, 변수 또한 본 연구에서 고려했던 것처럼 그대로 사용하였다. 여기서 알아야 할 것은 한국과 일본이 홈팀이라는 것이다. 모수 추정 결과, 2006년 결과와 마찬가지로 홈 이점에 해당하는 모수 δ는 다른 모수들에 비해 크게 나왔다. 즉, 홈 이점의 영향을 많이 받아 한국과 일본이 좋은 성적을 거둘 수 있었던 것이다.

표 5.1: 모형의 타당성(2002 한·일 월드컵)

조	국가	모형을 통해 예측	실제 올라간 국가
A조	덴마크, 세네갈, 우루과이, 프랑스	프랑스, 덴마크	덴마크, 세네갈
B조	스페인, 파라과이, 남아프리카 , 솔로베니아	스페인, 파라과이	스페인, 파라과이
C조	브라질, 터키, 코스타리카, 중국	브라질, 터키	브라질, 터키
D조	한국, 미국, 포르투칼, 폴란드	한국, 포르투칼	한국, 미국
E조	독일 아일랜드, 카메룬, 사우디아라비아	독일, 아일랜드	독일, 아일랜드
F조	스웨덴, 잉글랜드, 아르헨티나, 나이지라아	잉글랜드, 아르헨티나	스웨덴, 잉글랜드
G조	멕시코, 이탈리아, 크로아티아, 에쿠아도르	이탈리아, 크로아티아	멕시코, 이탈리아
H조	일본, 벨기에, 러시아, 튀니지	일본, 러시아	일본, 벨기에

표 5.1의 결과를 살펴보면, 모형으로 통해 예측한 결과와 실제 결과와는 차이가 크지 않았다. 16국가 중 프랑스, 포르투칼, 아르헨티나, 크로아티아, 러시아를 제외하고 11개 국가는 모형을 통해 예측할 수 있었다. 모형으로 예측할 수 없었던 국가는 모두 월드컵 이후, 이변이라고 말하는 국가였다. 아마도 2002년 월드컵 최대의 이변, 전 대회 우승국인 프랑스, 남미 강호 아르헨티나의 예선 탈락은 축구 전문가들도 상상할 수 없는 결과였다. 즉, 이변이 아닌 대부분의 경우를 모형이 잘 설명하고 있음을 알 수 있다. 어떠한 모형을 적합시킨다하더라도 객관적인 자료만으로는 이러한 이변을 찾아낼 수 없을 것이다. 따라서 본 연구에서 제시한 모형은 타당하다고 할 수 있다. 그리하여 이 모형을 적합시켜 2006 독일 월드컵을 예측한 것이다.

월드컵이 끝난 지금, 우리가 예측한 독일 월드컵 결과를 실제 결과와 비교할 수 있게 되었다. 2006년 독일 월드컵의 예측 결과와 실제 결과를 살펴보면 다음과 같다(표 5.2).

위의 2002년 결과와 마찬가지로, 예측 결과와 실제 결과에는 차이가 크지 않았다. A조의 폴란드, E조의 체코, F조의 크로아티아를 제외하고는 모형을 통해 예측할 수 있었다.

표 5.2: 모형의 타당성(2006 독일 월드컵)

조	국가	모형을 통해 예측	실제 올라간 국가
A조	독일, 폴란드, 코스타리카, 에쿠아도르	독일, 폴란드	독일, 에쿠아도르
B조	잉글랜드, 스웨덴, 파라과이, 토바고	잉글랜드, 스웨덴	잉글랜드, 스웨덴
C조	네덜란드, 아르헨티나, 세르비아, 코트디	네덜란드, 아르헨티나	네덜란드, 아르헨티나
D조	포르투칼, 멕시코, 이란, 앙골라	포르투칼, 멕시코	포르투칼, 멕시코
E조	이탈리아, 체코, 미국, 가나	이탈리아, 체코	이탈리아, 가나
F조	브라질, 크로아티아, 호주, 일본	브라질, 크로아티아	브라질, 호주
G조	한국, 프랑스, 스위스, 토고	프랑스, 스위스	프랑스, 스위스
H조	스페인, 우크라이나, 튀니지, 사우디	스페인, 우크라이나	스페인, 우크라이나

6. 결론 및 토의

2002년 한·일 월드컵에 한국은 4강이라는 놀라운 성적을 이루었다. 이러한 결과는 홈이점과 한국 축구의 성장이라 할 수 있다. 그렇다면 2006년 독일 월드컵에서도 한국이 좋은 성적을 이룰 수 있을까하는 의문에서 본 연구가 시작되었다. 월드컵은 4년에 한번 열리는 대회이므로 1930년에 우루과이에서 시작한 월드컵은 이번 독일 월드컵이 18회째이다. 따라서 지난 월드컵 자료는 독일 월드컵을 예측하기에는 매우 적고, 또한 수십년전의 자료는 현재 축구를 예측하기에 적당하지 못하다. 그리하여 우리는 최근 20년간의 A매치 데이터만을 사용하였다. 과거의 경기 결과외에 미치는 변수도 여러가지가 있겠지만, 정량화의 어려움으로 그 중 몇가지만 고려하였다. 분석결과, 한국은 같은 조인 프랑스와 스위스보다 16강에 올라 갈 확률이 적게 나왔다. 그리고 개최국인 독일이 객관적인 전력으로는 브라질, 아르헨티나등 몇개국보다 밀리지만, 홈 이점의 영향을 많이 받아 우승할 확률이 가장 높게 나왔다. 실제 결과로는 한국은 예측결과와 같게 아쉽게도 스위스에 밀려 16강에 탈락하였고, 16강에 올라 간 다른 조의 대부분 국가들은 예측을 통해 맞출 수 있었다. 또한 우승은 이탈리아, 준우승은 프랑스, 독일은 3위에 머물렀다.

본 논문은 축구 전문가와 축구팬들이 생각할 수 있는 일반적인 결과를 A매치 자료를 통해 확률을 제시한 점에서 의미있는 연구라 할 수 있다. 본 연구에서 사용한 자료외에도 추가적으로 축구경기 결과에 영향을 줄 수 있는 변수도 고려할 수 있을 것이다. 그리고 Bradley-Terry 모형에서 능력치(ability)가 시간에 따라 변할 수 있기 때문에 시간을 고려한 모형도 이러한 분석을 하는데 고려해 볼 수도 있다. 본 연구에서 제시한 모형은 월드컵 분석뿐만 아니라, 다른 스포츠에도 적용 할 수 있을 것이다.

참고문헌

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs : The method of paired comparisons, *Biometrica,* **39**, 324–345.

Davison, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments, *Journal of the American Statistical Association,* **65**, 317–328.

Ford, L. R. (1957). Solution of a ranking problem from binary comparisons, *The American Mathematical,* **64**, 28–33.

Singh, J. and Gupta, R. S. (1978). A paired comparison model allowing for ties, *Scandinavian Journal of Statistics,* **5**, 65–68.

[2006년 12월 접수, 2007년 2월 채택]

Prediction for 2006 Germany World Cup using Bradley-Terry Model*

Dohyun Kim[1] Sangin Lee[2] Yongdai Kim[3]

ABSTRACT

It is our greatest concern of Korean team to enter round of 16. The past football results are the most important data for making a prediction. And we know that the home advantage is also considerable factor and there are many unobservable factors. However, there are few matches between the participants and even not the results for some nations. To overcome this difficulty, we model the network of results and consider other factors. We predict 2006 Germany World Cup results using modified the Bradley-Terry model.

Keywords: Bradley-Terry model, World cup data.

* This work was supported by the Korea Research Foundation Grant funded by the Korean Goverment(MOEHRD) (KRF-2005-070-C00021).

1) Graduate Student, Department of Statistics, Seoul National University, Seoul 151-747, Korea
E-mail: stalove1@snu.ac.kr

2) Graduate Student, Department of Statistics, Seoul National University, Seoul 151-747, Korea
E-mail: lsi44@statcom.snu.ac.kr

3) (Corresponding author) Associate Professor, Department of Statistics, Seoul National University, Seoul 151-747, Korea
E-mail: ydkim@stats.snu.ac.kr