[6.13 지방선거 특집기사 기고]

베이지안을 활용한 2018년 6월 13일 선거 예측(경기지사 및 경남지사)

신뢰도가 낮은 여론조사로만 선거예측을 하신다고요?

지방선거가 일주일도 남지 않게 되면서 여론조사 결과 공표가 전면 금지되었다. 지난 6일 6일 발표된 마지막 여론조사 결과를 바탕으로 정당이나 후보자들은 지지층의 결집에 총력을 기울이거나 선거당일까지 결과를 굳히기 위한 집중 공세를 벌이고 있다. 언론사 및 정치평론가들 역시 여론조사 결과를 인용하거나 분석하며 다양한 선거 결과를 예측하고 있다.

선거 승리를 위한 다양한 공세 속에서 애꿎은 여론조사는 신뢰성을 의심받기도 한다. 실제 유권자들의 투표심리에 미치는 영향을 여론조사가 담지 못하고 있으며, 여론조사의 표본 역시 특정 정당에 치우쳐져 있다는 지적이다. 이처럼 여론조사에 대한 신뢰성이 지적되고 있는 상황에서 정당과 후보자, 정치평론가들이 여론조사 결과를 바탕으로 예측하고 있는 선거결과는 신뢰할 수 있을까? 이런 질문에서 시작된 우리의 분석은 한국탐사저널리즘센터 뉴스타파의 '여론조사 정확성 평가'기사(https://newstapa.org/39527)를 통해 구체화되었고, 베이지안 통계(Bayesian statistics, 이하 베이지안)를 적용하게 되었다.

베이지안을 활용한 선거예측은 선거 및 정치 분석 웹사이트 FiveThirtyEight 운영자이자 '신호와 소음'의 저자로 국내에 이름을 알린 Nate Silver가 2008년과 2012년 미국 대선을 정확하게 예측하면서 더욱 관심을 받기 시작했다. 물론 2016년 공화당 경선과 2016년 미국 대통령 선거 결과 예측을 번번이 틀리면서 그의 모델에 대한 물음표가 들긴 했지만, 베이지안을 활용한 선거예측 분석이 여론조사에 대한 신뢰도가 낮은 우리나라에서는 새로운 인사이트를 제공할 것이라고 판단되었다. 과거 우리나라에서도 베이지안을 활용해 2014년 서울시장 선거 결과를 비교한 전희원님의 분석도 있었다(http://freesearch.pe.kr/archives/4086). 여기서 아이디어를 얻어 우리도 베이지안을 바탕으로 2018 6.13 지방선거 결과를 예측했다.

이번 분석의 의뢰 및 기관은 데이터 분석모임 Foresight의 최정윤, 박희경, 원인재이다. 베이지안에서는 새로운 자료가 없는 상태에서 어떤 사건이 일어날 확률에 대한 가정이 필요한데, 이를 사전 확률(prior probability, 이하 prior)이라고 한다. 우리는 이러한 prior를 사건이 일어날 것에 대한 믿음이라고 정의한다. 즉, 여론조사 결과를 보기 전에 후보자가 선거에서 당선될 확률이 얼마나 되는지에 대한 믿음이다! prior과 선거예측 모형의 가중치 설정을 위해 우선 2014년 서울시장 지방선거와 부산시장 지방선거 데이터를 활용했다. 2014 지방선거 결과를 바탕으로 설정한 prior과 가중치를 선거예측 모델에 넣어 2018 6.13 지방선거 결과를 예측해 보았다.

Test: 2014년 지방선거를 통해 연습하기

2014년 서울시장 선거 결과와 부산시장 선거 결과를 분석을 통해 prior과 가중치를 어떻게 설정하는 것이 선거예측 결과에 더 정확할지 살펴보았다. 2014년 서울시장 선거 분석 자료는 전희원님의 깃허브를 통해서 얻었으며, 2014년 부산시장 지방선거 분석 자료는 중앙선거여론조사위원회에 등록된 여론조사결과에서 수집했다. 서울시장 분석데이터는 2014년 3월 24일부터 5월 28일까지 총 31건의 여론조사 결과이고, 부산시장 분석데이터는 2014년 5월 17일부터 5월 28일까지 총 22건의 여론조사 결과이다. 분석 프로그램은 R을 사용했으며, 디리슐레 분포의 몬테칼로 시뮬레이션은 MCMCpack의 함수(MCmultinomdirichlet)를 분석 패키지로 활용했다.

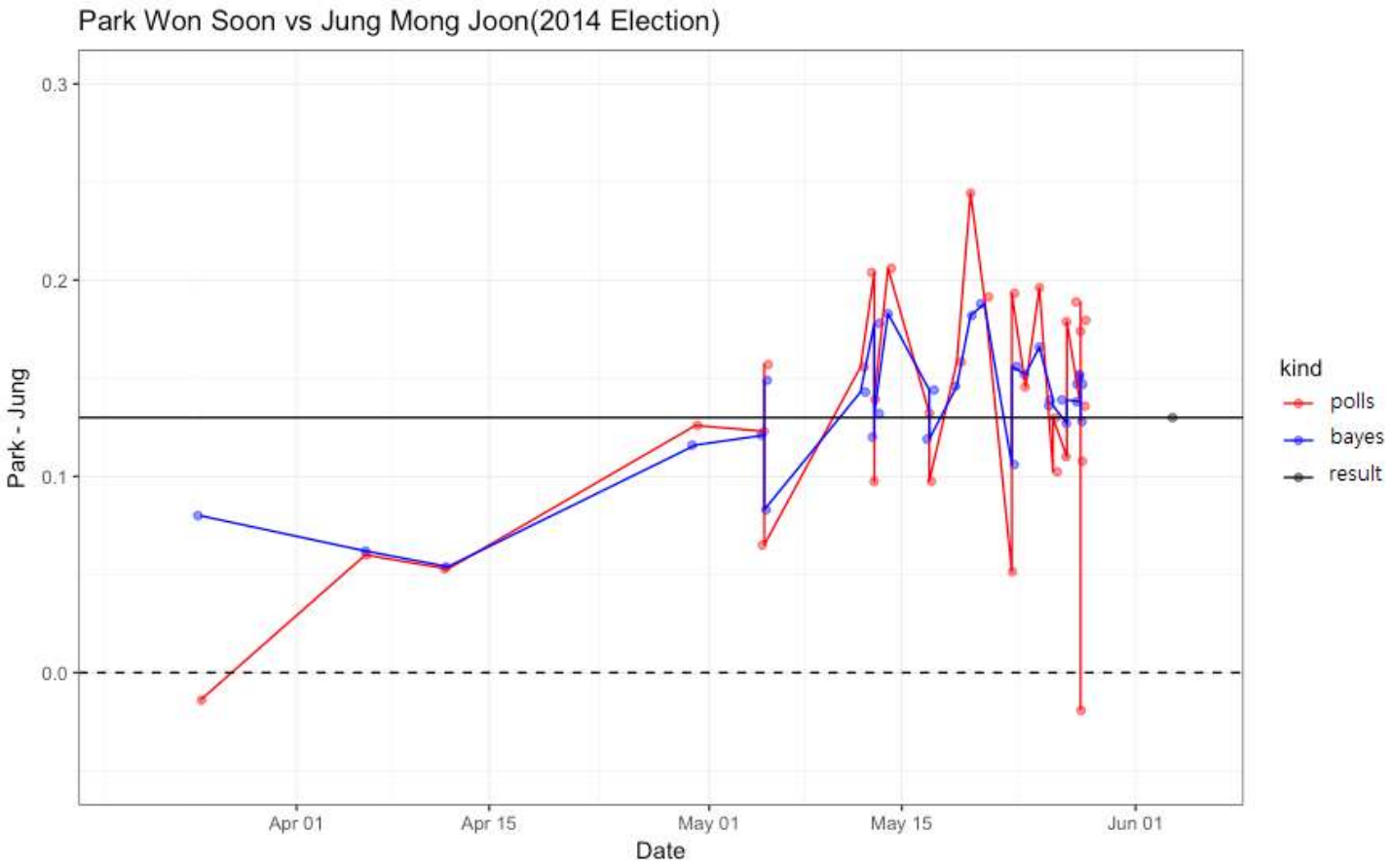
국내 지방선거는 정권을 표(票)로 심판하는 선거라는 인식이 있어 대통령의 국정운영 지지도와 정당 지지도와 강한 상관관계를 가진다. 정당의 지지도에 따라 지방선거 결과가 달라진다는 말이다. 실제로 대통령 취임 후 1년 이내에 진행된 지방선거에서는 여당이 승리하는 경향이 강하게 나타난다. 그러나 대통령 지지율은 긍정, 부정, 기타(무응답, 모름) 등의 형태로 조사되기 때문에 이를 여당후보, 야당후보, 기타 후보에 몰아서 prior로 설정하는데 있어 한계가 있다고 보았다. 이러한 경험과 이유를 근거로 우리는 prior를 선거 여론조사 시작 전 '정당 지지도'로 설정했다. 따라서 2014년 분석에서 prior은 2014년 5월 1주차 정당지지도이다. 정당지지도는 한국갤럽이 의뢰하고 자체 조사한 여론조사로 여당 39%, 제1야당 23%, 기타 3%, 없음(의견유보) 33% 순이었다. 조사인원은 808명, 표본오차 3.4%, 응답률 22%로 나타났다.

그리고 최근에 실시된 여론조사 결과가 실제 선거 결과와 가장 근접하다고 판단해 가장 최근에 실시한 여론조사에 대해 가중치를 주었다. 가중치의 경우 선거일을 기준으로 해 7/(선거일 - 현재일수)로 계산했다. 즉, 맨 마지막 선거일을 1로 하고, 앞에 실시된 여론조사의 가중치는 7/10, 7/12 등의 형식이다.

우리는 설정한 prior과 가중치가 실제 선거결과와 어떤 영향을 미치는지 살펴보기 위해 총 4분의 분석을 실시했다. ① prior X, 가중치 X, ② prior X, 가중치 O, ③ prior O, 가중치 X, ④ prior O, 가중치 O이다. 분석결과 prior과 최근 가중치를 모델에 넣은 ④ 모형이 가장 적합한 것으로 나타났다. 자세한 분석 과정은 블로그에서 확인 가능하다.

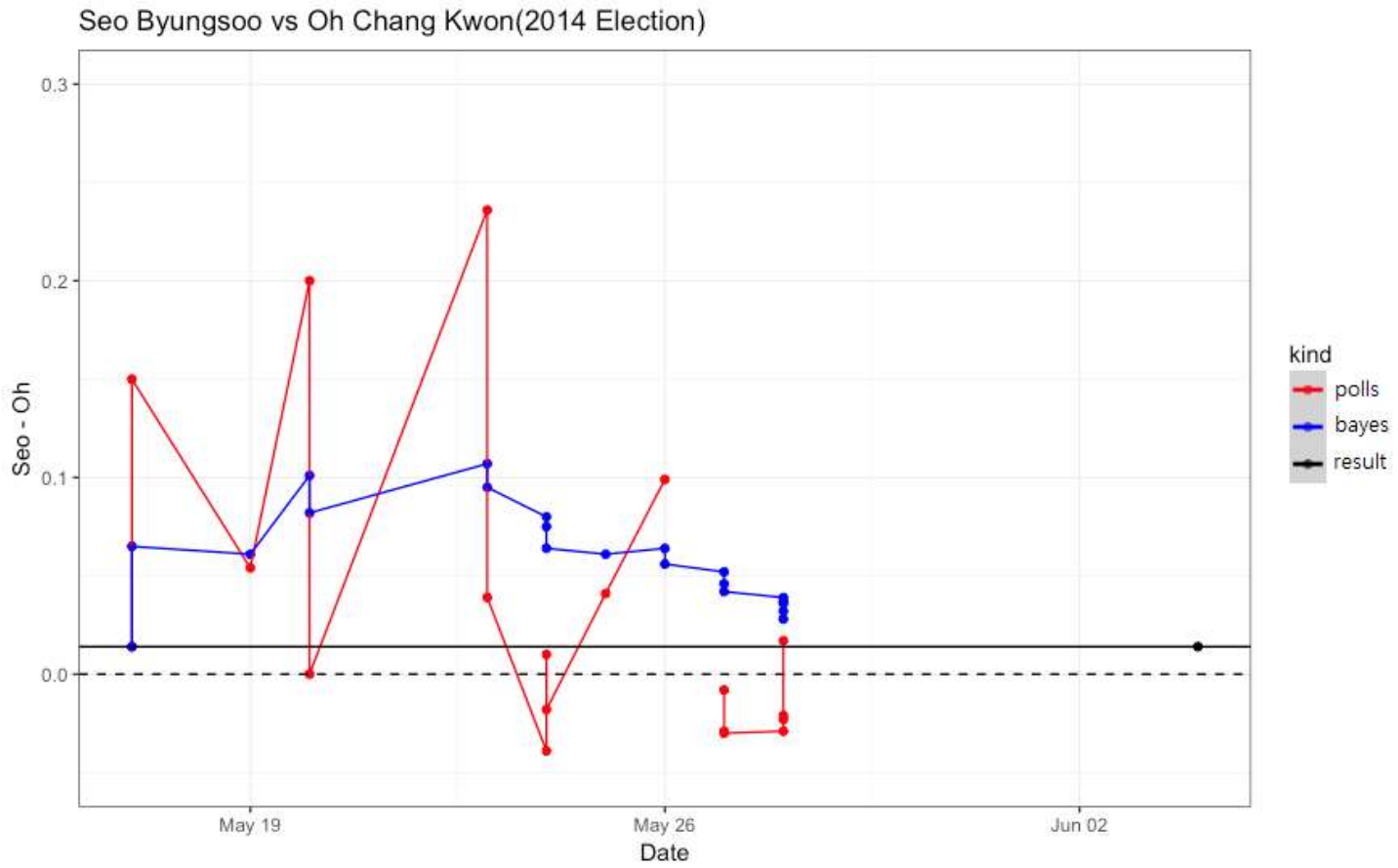
2014 서울시장 선거결과를 바탕으로 한 <그림1>은 여론조사와 베이지안을 활용했을 때 확인할 수 있는 지지율 차이 추이다. 빨간색 선은 각 실시된 여론조사의 결과로 계산된 지지율 차이이며, 파란색 선은 여론조사 결과 데이터를 바탕으로 베이지안을 활용한 지지율 차이이다. 검정색 선은 박원순 시장과 정몽준 후보의 지지율 차이인 13%p를 의미한다. 선거일이 다가올수록 베이지안을 이용한 파란색 점들이 빨간색 점들보다 더 검정색 선에 근접하다는 것을 확인할 수 있다.

<그림1 2014 서울시장 선거 분석결과>



<그림2>은 새누리당 서병수 시장과 무소속 오거돈 후보가 참여한 2014년 부산시장 선거를 분석한 결과이다. 오거돈 후보의 경우 표면적으로는 무소속 후보이었기 때문에 정당이 없다고 고려해 이번 모델에서는 prior를 넣지 않았다. <그림2>의 검정색 선 역시 실제 서병수 시장과 오거돈 후보의 2014년 지방선거 지지율 차이를 의미하며, 0.6%p이다. 분석결과 앞선 그래프와 동일하게 베이지안을 의미하는 파란색 점들이 여론조사 결과인 빨간색 선들보다 실제 지지율 차이에 근접하다는 것을 알 수 있다. 이러한 결과를 바탕으로 이제 2018년 6.13 지방선거 결과를 예측해보자!

<그림2 2014 부산시장 선거 분석결과>



Application: 6.13 지방선거의 승자는 누가 될 것인가?

6.13 지방선거는 전국적으로 열리지만, 화재성이 높은 '경기도지사'와 '경남지사'의 여론조사 데이터를 수집해 여론조사 지지율이 높은 두 명의 후보자들만 분석했다. 분석 데이터는 중앙선거여론조사위원회에 등록된 여론조사결과에서 수집했으며, 각 여론조사 내용은 중앙선거여론조사심의위원회 홈페이지를 참조했다. 경기지사 분석데이터는 공천확정 전 3월 30일부터 6월 2일까지 총 18건이었으며, 경남지사 분석데이터는 4월 13일부터 6월 4일까지 진행된 총 27건이다. 가장 최근에 발표된 6월 6일 여론조사 결과는 분석에서 제외되었다.

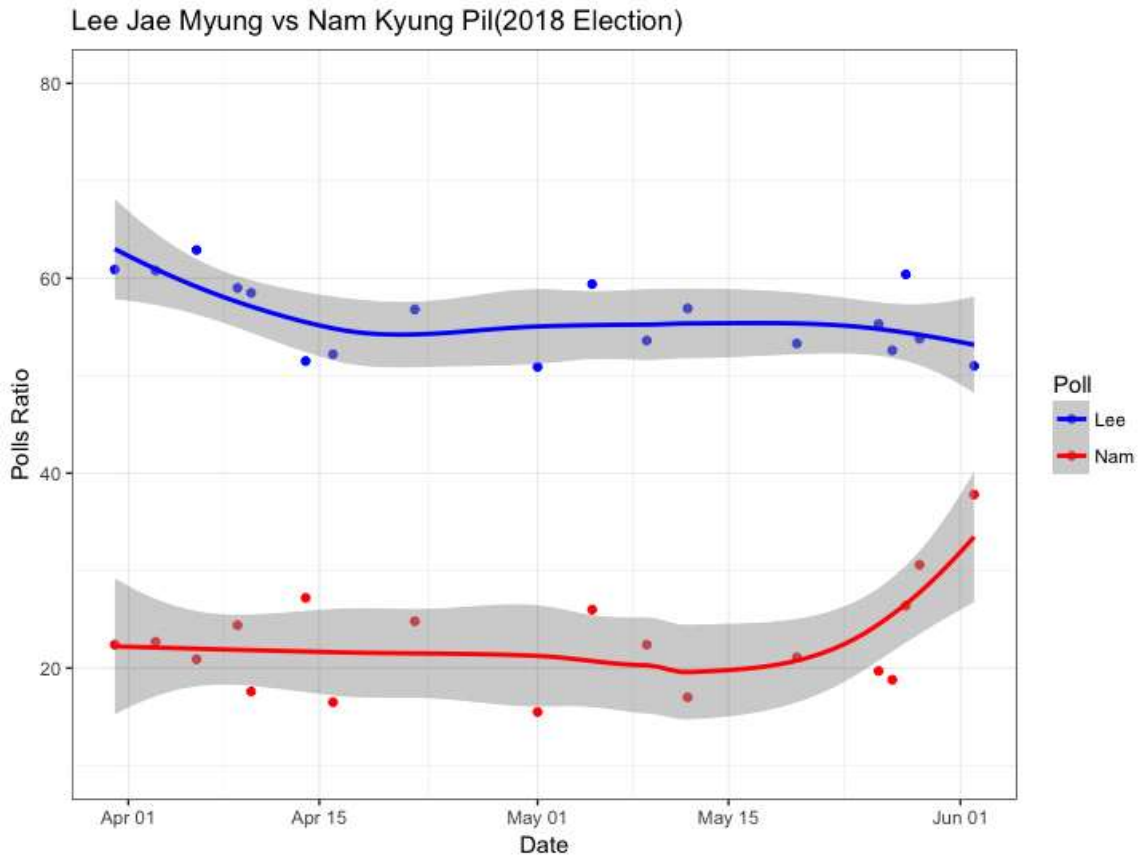
분석은 베이지안 통계(Bayesian statistics)를 활용했으며, 분석 프로그램은 R을 사용했다. 디리슐레 분포의 몬테칼로 시뮬레이션은 MCMCpack의 함수(MCmultinomdirichlet)를 분석 패키지로 활용했다.

분석 모델의 prior은 2018년 5월 1주차 정당지지도로 했다. 한국갤럽이 의뢰하고 자체 조사한 2018년 5월 1주차 정당지지도 여론조사는 더불어민주당 49%, 자유한국당 13%, 바른미래당 8%, 정의당 6%, 민주평화당 0.3%, 기타 0%, 모름 24% 순으로 결과를 보였다. 이때 조사 인원은 1,004명, 표본오차 3.1%, 응답률 16%로 나타났다. 최근 여론조사에 대한 가중치는 선거일을 기준으로 해 7/(선거일 - 현재일수)로 계산했다.

1. 경기도지사: 남경필 vs 이재명

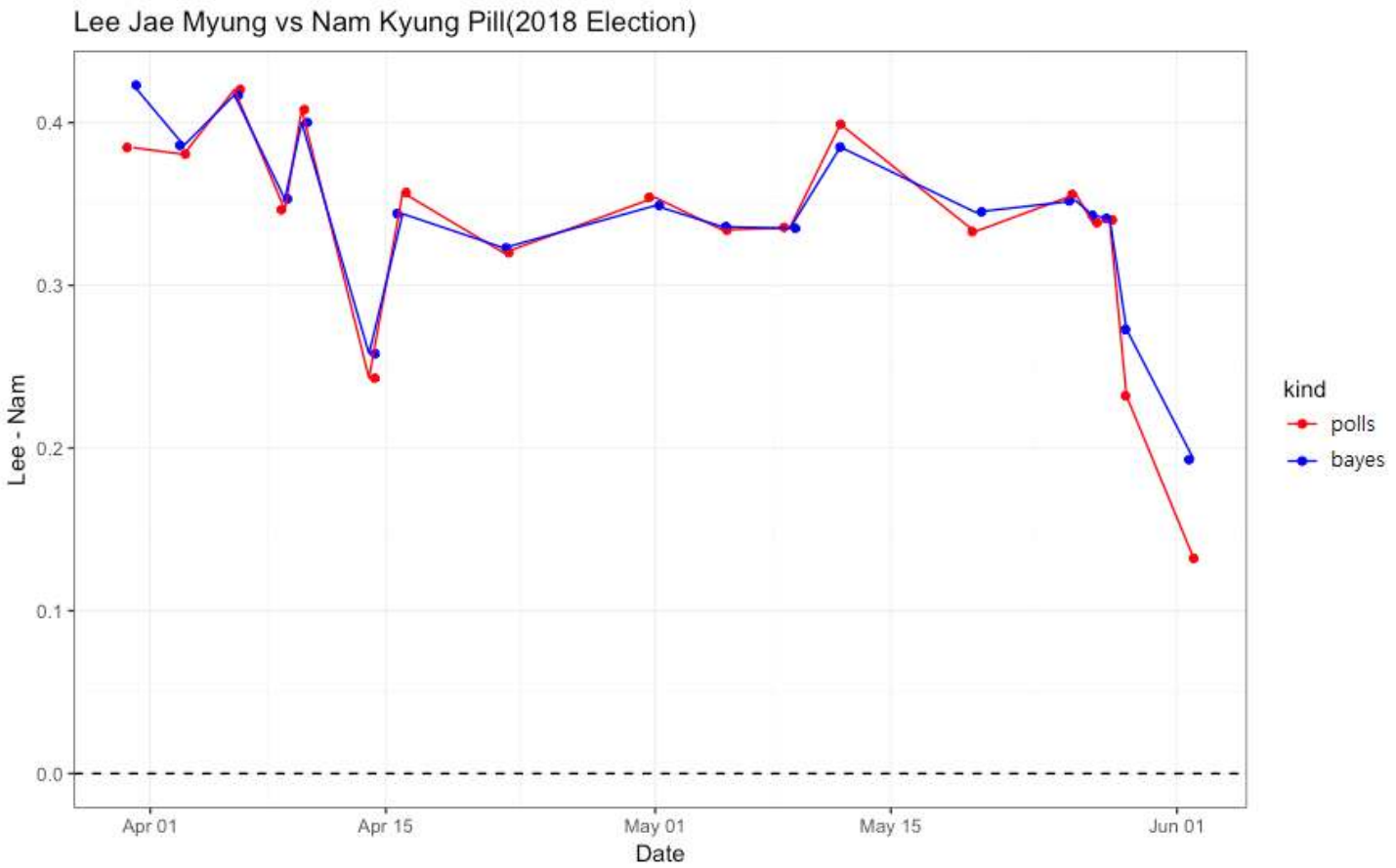
현재 경기도지사 선거에서 여론조사 지지율이 높은 후보자는 남경필 후보와 이재명 후보이었다. <그림3>은 더불어민주당 후보 공천이 확정되기 전인 3월 30일 여론조사의 가상대결에서부터 시간 흐름에 다른 두 후보자의 여론조사 결과를 보여주는 그래프이다. 두 후보자의 여론조사 지지율 격차는 약 38.5%p에서 시작했으나, 선거일로 가까워질수록 그 격차가 좁혀지고 있었다. 베이지안 모델은 두 후보자의 지지율 격차를 어떻게 분석할까? 결과는 <그림4>에서 확인할 수 있다.

<그림3 경기도지사 여론조사 트렌드>

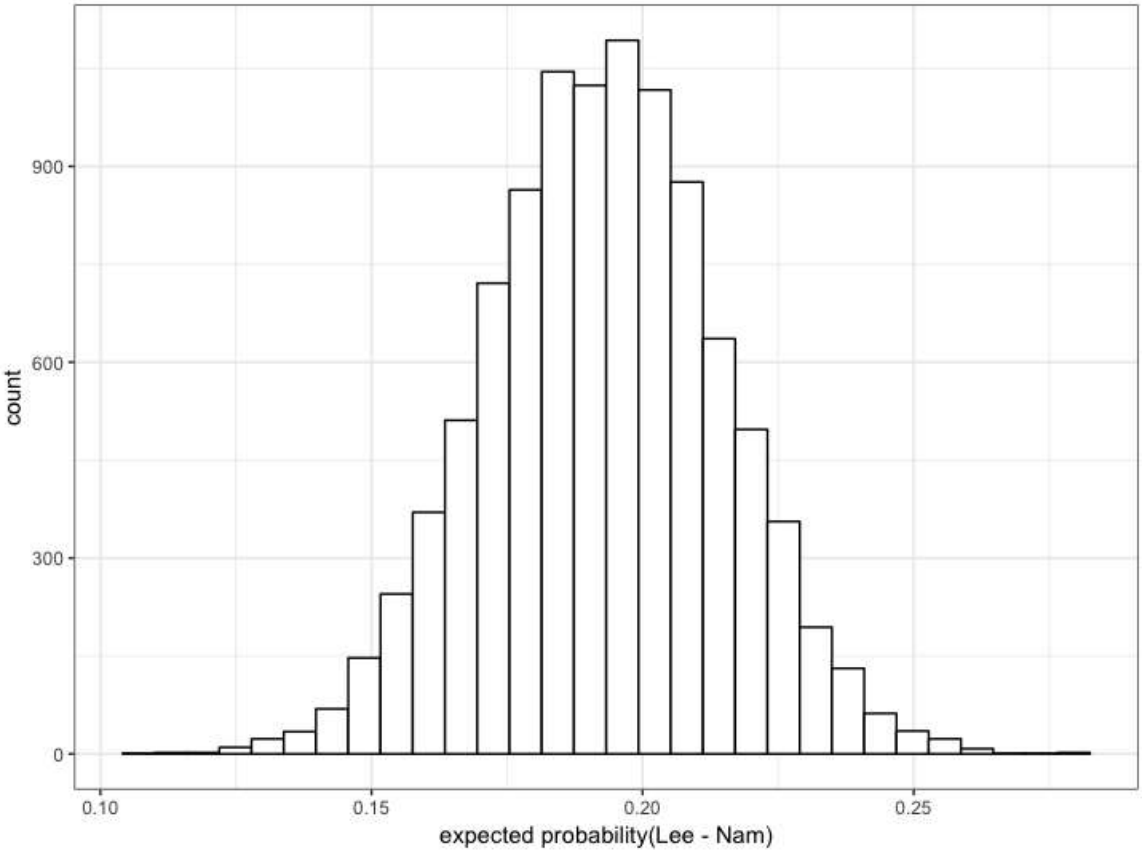


<그림4>는 이재명 후보와 남경필 후보에 대해 베이지안을 적용시킨 결과값으로 두 후보자 간 지지율 차이를 보여준다. 빨간색 선은 각 실시된 여론조사의 지지율 차이로 시간이 지나면서 이재명 후보와 남경필 후보의 지지율 격차가 좁혀지는 것을 알 수 있다. 앞선 여론조사 정보를 반영하는 베이지안의 지지율 차이 결과를 보여주는 파란색 선은 시간이 지나도 완만한 기울기를 보이고 있다. 베이지안 분석결과 두 후보자 간의 선거예측 결과는 <그림5>과 <그림6>에서 보는바와 같다. <그림6>의 결과는 prior를 반영하지 않은 히스토그램이다.

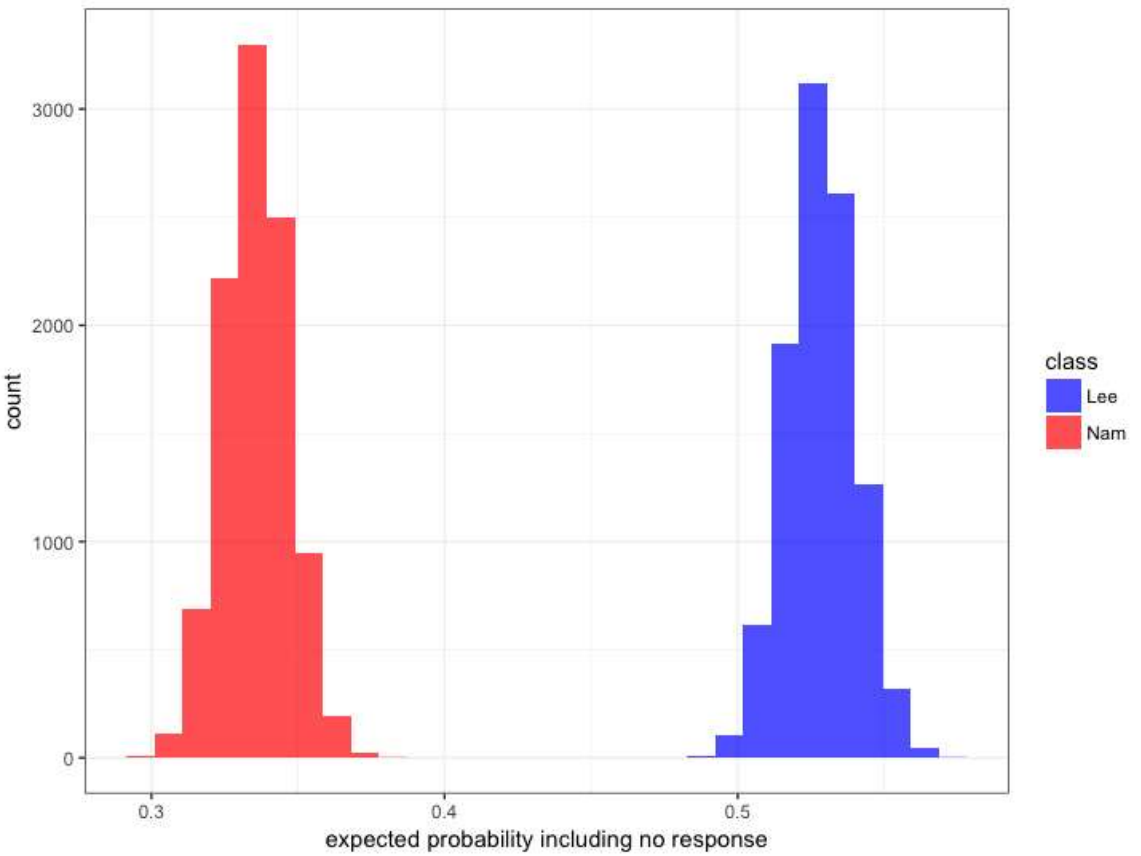
<그림4 베이지안 활용 경기도지사 선거예측 분석 결과>



<그림5 2018 경기도지사 선거 예측 Mean of Posterior>



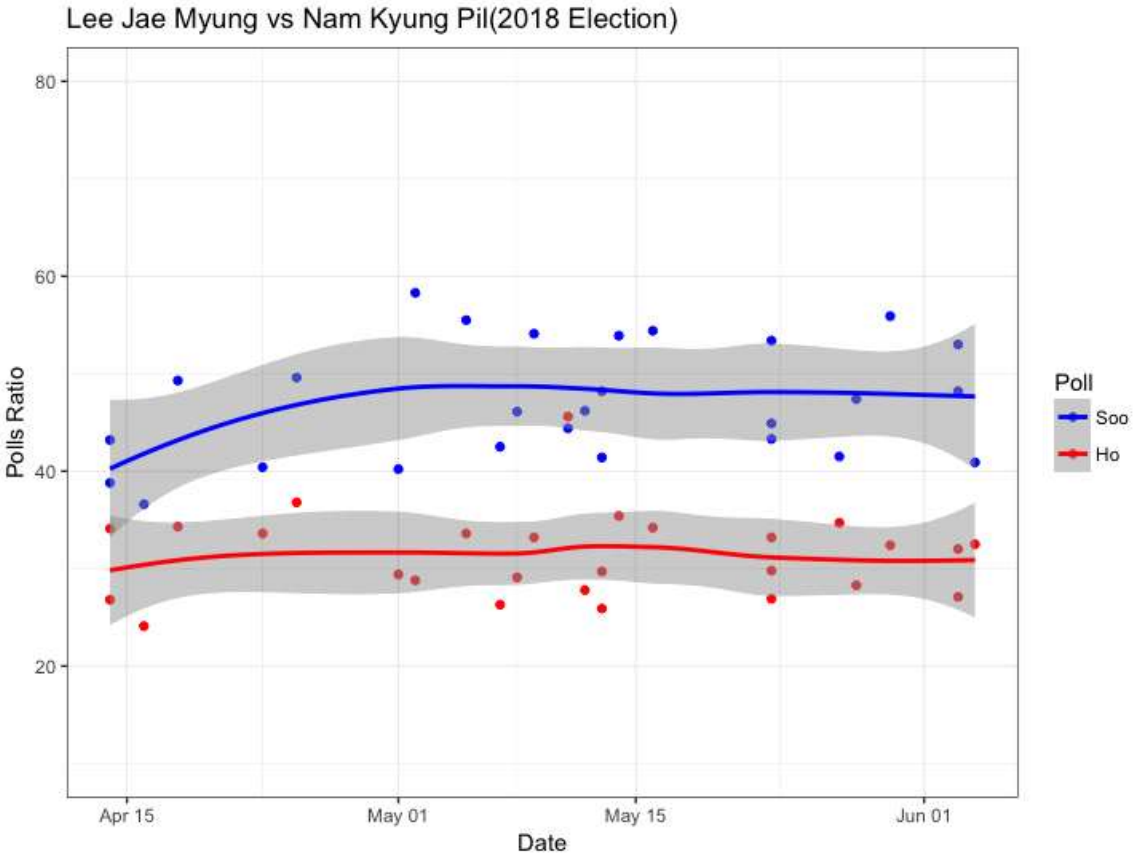
<그림6 2018 경기도지사 선거예측 결과 Histogram>



2. 경남지사: 김경수 vs 김태호

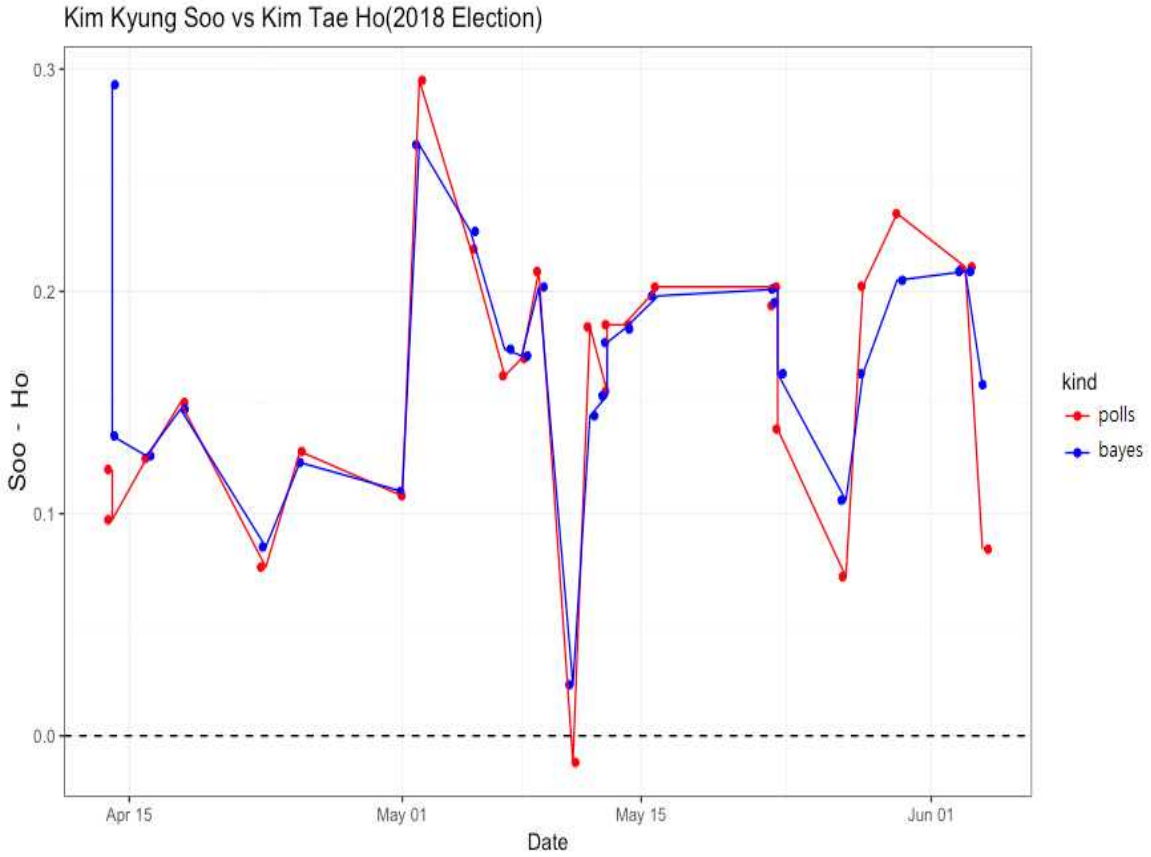
그렇다면 경남지사는 어떨까? 현재 경남지사에서는 더불어민주당 김경수 후보와 자유한국당 김태호 후보의 여론조사 지지율이 높은 지역이다. 두 후보자의 여론조사 지지율 격차는 <그림7>와 같다. 파란색 선은 김경수 후보의 여론조사 결과를 의미하고, 빨간색 선은 김태호 후보의 여론조사 결과를 의미한다. 여론조사가 실시되기 시작한 초기에는 그 격차가 크지 않았지만, 시간이 흐를수록 두 후보자의 여론조사 지지율 격차는 벌어지고 있다는 것을 확인할 수 있다.

<그림7 경남지사 여론조사 트렌드>

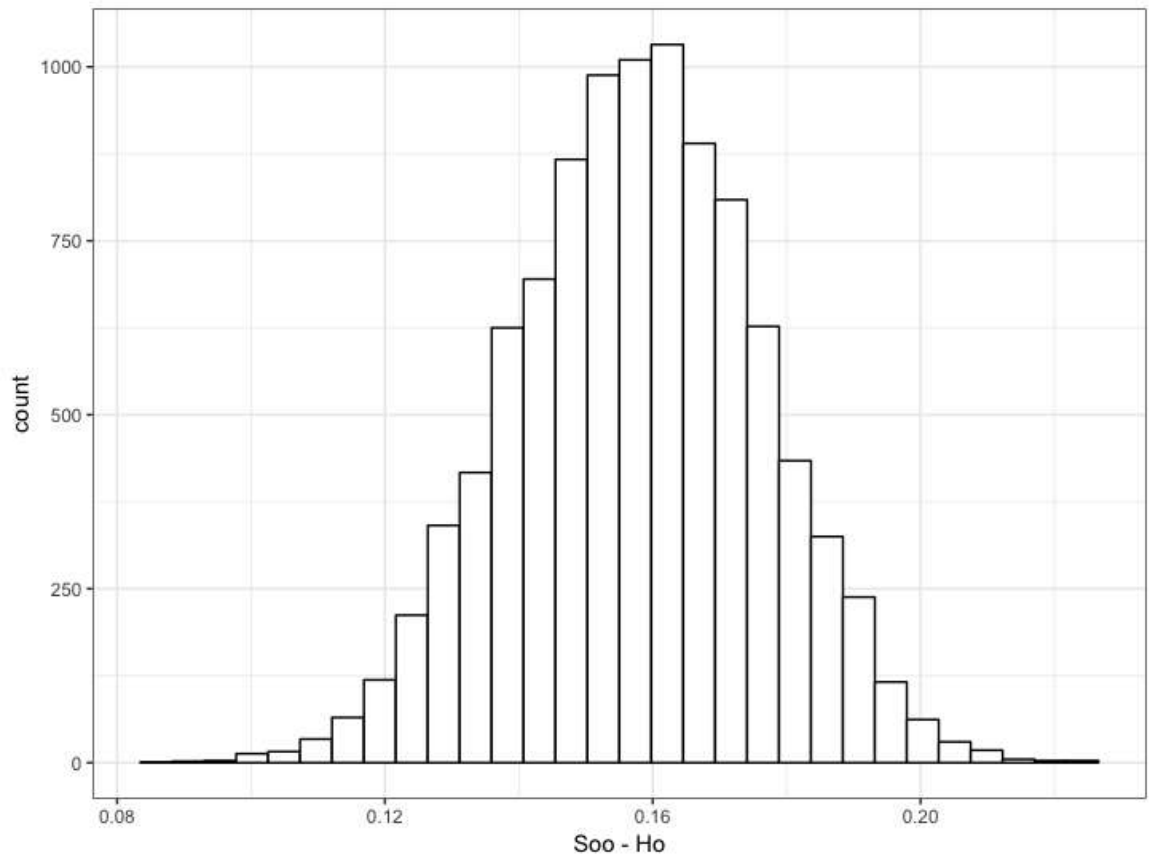


두 후보자의 지지율을 베이지안으로 분석한 결과는 <그림8>과 같다. 여론조사 지지율 격차를 보여주는 빨간색 선과 베이지안의 지지율 격차를 보여주는 파란색 선의 추이는 비슷한 형태를 가진다. 하지만 6월 4일 조사가 종료된 조원씨앤아이가 조사한 최근 여론조사에서는 두 후보자의 지지율 격차는 8.4%p이었다. 반면 베이지안은 두 후보자의 지지율 격차가 과거 여론조사 지지율 격차와 큰 차이가 없음을 보여준다. 이러한 베이지안 분석을 바탕으로 하는 두 후보자 간의 선거예측 결과는 <그림9>와 <그림10>과 같다. 이때도 <그림10>의 결과 역시 prior를 반영하지 않은 히스토그램이다.

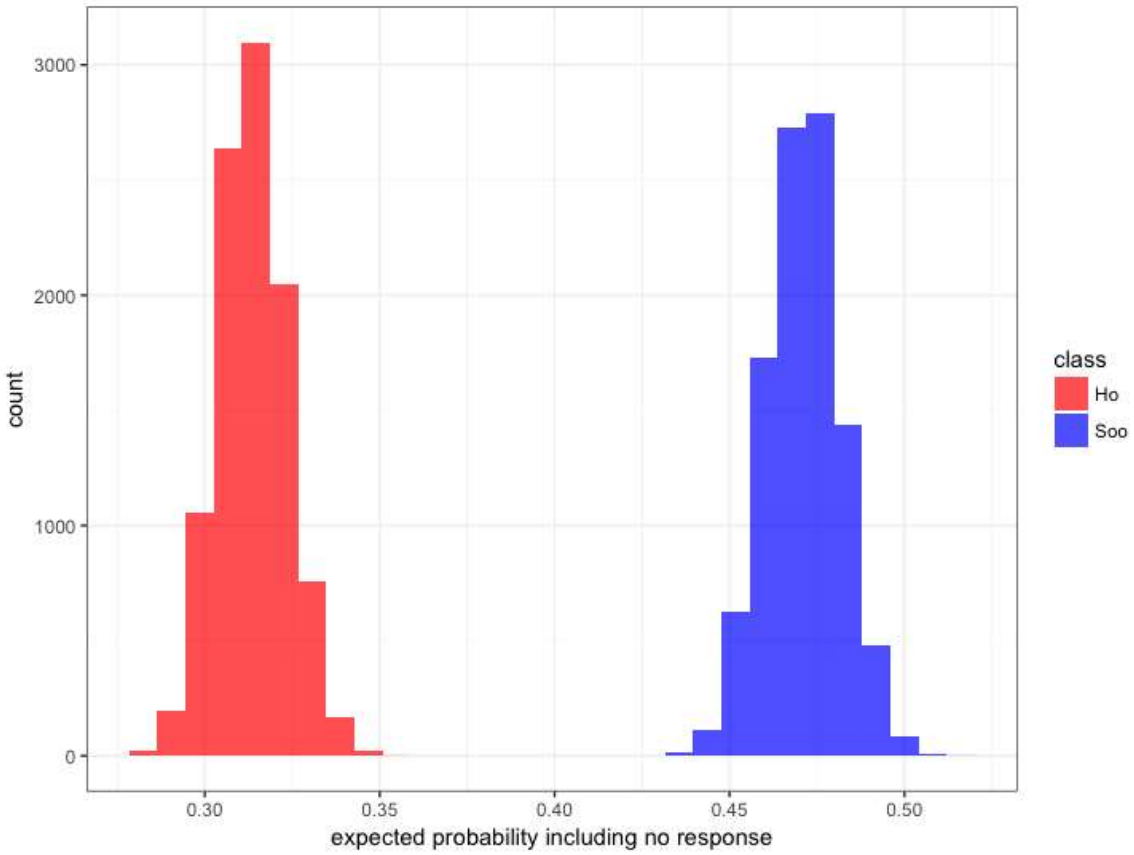
<그림8 베이지안 활용 경남지사 선거예측 분석 결과>



<그림9 2018 경남지사 선거 예측 Mean of Posterior>



<그림10 경남지사 선거예측 결과 Histogram>



“오늘 할 수 있는 최선의 예측을 했다”

신뢰성에 대한 문제가 제기되고 있는 여론조사 결과가 아니라 여론조사의 데이터를 활용한 베이지안(Bayesian)을 통해 ‘2018 경기도지사 지방선거’와 ‘2018 경남지사 지방선거’ 결과를 예측해 보았다. 두 지역의 후보들은 여론조사에서도 큰 차이를 보이고 있어 이번 선거예측 분석 결과가 새로운 인사이트를 제공하지 않을 수 있다. 우리는 이번 선거예측 모델에서 여론조사에 참여한 사람들의 수를 바탕으로 베이지안 방법을 적용했지만 각각의 여론조사 결과의 지지율을 기준으로 하는 베이지안 추정 또한 시도해볼 예정이다.

우리는 이번 분석을 통해 베이지안(Bayesian)이 여론조사들의 정보들을 종합적으로 반영한다는 것을 알게 되었다. 베이지안 방법론을 적용하면 개별 여론조사 결과뿐 아니라 앞서 실시된 여론조사들을 누적해 실시간으로 종합된 여론조사 결과를 파악할 수 있다. 사전 여론조사 정보를 베이지안이 통합한다는 함의는 지난 4년 전 서울시장 선거와 부산시장 선거에서도 나타났다. 이러한 여론조사 결과를 하나하나 독립 된 것으로 보지 않고 여러 가지 여론조사의 결과들을 반영해서 업데이트를 하는 베이지안의 특성은 현재 논란이 되고 있는 ‘여론조사’들의 편향성 문제를 완화시키는 효과가 있을 것이다.

우리들의 선거예측이 얼마나 정확할 것인가? 미래는 언제나 불투명하고 선거와 정치 역시 어떤 돌발적 상황이 발생할지 우리는 알지 못한다. 오늘도 후보자들을 대상으로 하는 다양한 스캔들이 쏟아지고 있으며, 후보자들의 단일화가 진행될 수도 있다. 국내외 정세도 급변하고 있다. 따라서 6.13지방선거가 다가올수록 처음 여론조사의 예측도 그리고 우리의 예측도 분명 바뀔 것이다. 네이트 실버의 지적처럼 예측과 그 확률은 관측한 시점에 따라 얼마든지 바뀔 수 있다. 그래도 우리는 그저 우리가 오늘 할 수 있는 최선의 예측을 할 뿐이다.

*** 이번 분석은 뉴스타파의 데이터 제공 및 조언을 바탕으로 진행되었습니다. 분석결과 및 분석과정에 대한 자세한 내용은 https://foresighters.github.io/에서 확인 가능합니다.**