

연구논문

온라인 여론조사의 편향 성격과 보정방법 탐색: 2014년 서울시장 선거 사례

Sources of Bias in Online Survey and Their Corrections:
A Case of 2014 City-Mayor Election in Seoul

임정재^{a)} · 강정환^{b)}

Jeong-jae Lim · Jeong-han Kang

본 연구는 최근 사용이 확대되고 있는 온라인 조사를 선거여론조사에 활용했을 때 생길 수 있는 편향의 성격을 파악하고 보정방안을 모색하고자 온라인 조사를 이용하여 선거결과 예측을 실시하였다. 이 과정에서 2014년 서울시장 선거 전후에 걸쳐 실시된 다섯 차례의 패널조사와 선거 후 실시된 유선전화조사의 결과를 실제 결과와 비교·분석하여, 온라인 조사의 횡단적·종단적 편향의 특성을 살펴보았다. 특히 온라인 조사에서 의심되는 응답자의 진보편향을 확인하기 위해 2012년 대통령선거 투표행태를 설문에 포함시켰다. 분석의 결과 대통령선거시의 응답자의 연령별 분포는 실제 서울시내 투표결과와 거의 일치하였고, 다섯 차례 조사에 걸친 응답자 탈락에도 불구하고 이 분포는 종단적으로 큰 변화가 없었다. 이처럼 온라인 응답자의 진보편향이 횡적, 종적으로 크지 않았음에도 불구하고, 서울시장 후보

1) 이 논문은 2013년도 정부재원(교육부 인문사회연구역량강화사업비)으로 한국연구재단의 지원을 받아 연구된 것으로서(NRF-2013S1A3A2055285), 2014년 한국정치학회 하계 학술회의에서 발표된 논문을 수정 및 보완한 것이다. 당시 건설적인 제안을 해주신 숭실대학교 이상신 교수님 그리고 유익한 논평을 해주신 세 분의 익명의 심사자 분들에게 진심으로 감사를 표한다.

a) 교신저자(corresponding author): 연세대학교 사회학과 박사과정 임정재.

E-mail: hello.jj.lim@gmail.com

b) 연세대학교 사회학과 부교수

중에는 진보적인 박원순 후보에 대한 지지율이 실제보다 높게 나타났다. 이러한 자료를 이용하여 서울시장 선거결과를 예측하기 위해 연령, 성별, 그리고 2012년 대선 투표율 분포 등을 활용하여 가중치를 적용한 결과 지지율이 실제 후보별 득표율에 가까워졌으나, 뚜렷이 과소표집되어 높은 가중치가 적용된 고연령층의 박원순 후보 편향이 특히 심했기 때문에 보정 정도에 한계가 있었다. 이러한 편향은 온라인 응답자 섭외가 특히 힘든 고연령층을 확대 및 대체 표집하는 과정에서 심화된 것으로 추론되며, 고연령층 섭외가 쉽고 정몽준 후보 지지율이 특히 높았던 유선전화조사의 편향과 대조되었다. 본 연구는 온라인 조사의 편향에는 기존에 주장되어 온 일반적인 진보편향보다는 해당 선거에 대한 관심편향(자기선택오차)이 특정 후보에 대한 지지와 연관될 위험이 고연령층처럼 응답협조가 어려운 집단에서 더 두드러짐을 보였다. 이러한 연구결과를 바탕으로 아직은 모집단 정의가 불가능하고 비확률표집에 의존할 수밖에 없는 온라인 조사가 향후 선거 여론조사에 활용할 수 있는 가능성 및 해결해야 할 과제를 토의하였다.

주제어 : 조사편향, 온라인 조사, 가중치, 선거 여론조사

This study investigates the characteristics of cross-sectional and longitudinal biases from online panel survey for election polls and tries correcting these biases. It particularly focuses on possible bias toward liberal candidates in online survey. For empirical analysis, we conducted a panel survey of 5 waves for 2014 City-Mayor Election in Seoul during 4 months, using a pool of online respondents in a survey company. We also conducted a 1-time telephone survey by the same survey company to compare with online panel. In both surveys, we ask respondents' choice in 2012 Presidential Election in order to examine and correct for ideological bias. Analysis revealed that our online respondents did not show bias in the last presidential election, which was also largely robust against panel attritions throughout 5 waves. We, however, observed bias toward a liberal candidate in a mayor election, particularly for the group of 60s and older. The bias was moderately corrected by weights on

age and gender, but remained notably. At the same time, bias in online survey was lesser than that of telephone survey, except for respondents in their 60s and older. The poor contacts and cooperations of older people in online survey led to replacement sampling and presumably sorted into selective respondents with higher interest in the mayor election even if those respondents cannot be characterized as generally liberal. We discuss possibilities and remaining problems for online election poll in the future. The fast increasing access to online device among older people is one of those possibilities whereas inability to define a population and consequent nonprobability sampling are problems.

Key words: survey bias, online survey, weight, election poll

I. 문제제기

여론조사라는 단어를 접하게 되면, 우리는 흔히 선거여론조사를 떠올리기 쉽다. 이는 여론조사의 발전이 선거와 밀접하게 연결되어 있을 뿐만 아니라, 선거철이 되면 대중들이 가장 쉽고 빈번하게 접할 수 있기 때문일 것이다. 더욱이 선거여론조사의 중요성은 점점 증대되어, 이제는 실제 선거와 정치에 많은 영향력을 미치게 되었다. 16대 대통령선거 당시 노무현 후보와 정몽준 후보의 단일화 과정이 여론조사로 진행되었을 뿐만 아니라, 17대 대통령선거 당시 한나라당 당내경선에서도 이명박 후보가 여론조사 결과에서 박근혜 후보보다 높은 점수를 얻음으로서 후보경선에서 승리할 수 있었다. 이렇게 여론조사에 의해 선택되었다고 볼 수 있는 노무현·이명박 후보가 실제 대선에서도 모두 대통령으로 당선되면서, 여론조사는 선거승리의 도구로서 각인되기도 하였다(홍명신 2009).

한편 사회적 환경이 변화함에 따라 선거여론조사의 방법은 많은 변화를

꺾어 왔다. 전통적으로 여론조사는 유선전화조사와 대인면접조사 방법으로 수행되어져 왔으나, 각 조사가 가지고 있는 문제점들을 보완하기 위해, 온라인 조사나 휴대전화조사와 같은 대안적 여론조사방법들이 대두되었고(강남준·백영민 2004) 이러한 맥락에서 온라인 조사에 대한 관심은 점점 증대되었다.

그러나 온라인 조사는 인터넷 접근성에 따른 포함오차(coverage error)와 자기선택오차(self-selection error)등의 단점으로 인해, 근본적으로 모집단의 정의가 모호하고 확률표집을 어렵게 하며 대중을 상대로 하는 여론조사에서는 활발히 사용되지 않는 경향이 있다(Dillman et al. 2008). 즉, 고연령층과 저소득층, 저학력층의 경우 인터넷 이용률과 접근성이 낮은 경향이 있으며(Couper 2000; Taylor 2000; Dillman et al. 2008), 동일한 연령대라고 하더라도 이용여부와 학력이나 직업 등의 차이가 존재함을 고려할 때(이계오·장덕현 2009; 이계오 외 2012) 이러한 사회인구학적 특성의 차이가 정치적 성향의 편향을 발생시킬 수 있다는 점이 명확한 한계로 지적되고 있다(이계오·장덕현 2009; 이계오 외 2012).

실제로 중앙선거여론조사 공정심의위원회 등록현황(중앙선거여론조사공정심의위원회 2014a)을 살펴보면, 2014년 제6회 전국동시지방선거와 관련된 814건의 선거여론조사 중에서 온라인을 통해 실시한 조사(웹조사)는 단 1건에 불과하였으며, 그마저도 정당지지도 관련 조사에 그쳐 온라인 조사가 실제 선거여론조사에서 활발히 사용되지 않음을 확인할 수 있었다. 이에 본 연구에서는 온라인 조사를 통해 수집한 자료를 바탕으로 선거결과 예측을 시도해 보고, 정치적 성향의 편향을 포함하여 추정을 왜곡하는 원인들을 살펴보고자 한다.

한편, 앞서 논의된 편향이 횡단적 조사에서 발생 가능한 편향이라면, 일정 응답자들을 반복적으로 접촉하는 종단적 조사가 진행될 경우 발생하는 편향에도 주의하여야 한다. 기존 유선전화조사의 경우 선거여론조사를 패널조사로 실시하였을 때 20대·60대 연령층(허명회 외 2009), 저학력, 저소득층의 탈락률이 높았으며, 정치에 관심이 높은 표본이 과대표집되는 경향을 보였다(김장수 2007). 온라인 조사는 강제력이 적어 재응답률이 상대적으로 낮기 때

문에 이렇게 특정한 집단이 계속해서 탈락하게 될 경우 그 편향을 증폭시킬 가능성이 높아질 것이다. 결국 온라인 패널조사를 통해 선거여론조사를 실시할 때 자기선택오차가 반복됨으로써 발생 가능한 종단적 편향 또한 함께 관심을 가져야 할 것이다.

이러한 문제의식에서 출발하여 본 연구는 2014년 제6회 지방선거 중 서울시장 선거와 관련하여 다섯 차례에 걸친 온라인 패널조사를 설계하고 자료를 수집하였다. 본 연구의 목적은 향후 선거여론조사에서 온라인 조사의 실제적 활용을 돕는 것이며, 따라서 기존 온라인 조사업체의 온라인 조사패널을 활용하여 조사를 진행하였다. 즉, 응답자의 추출은 확률표집이 아닌 할당표집이며, 이 한계 내에서 조사의 정확성을 높이기 위한 진단을 실시하려 한다. 이와 같이 수집된 자료를 바탕으로 선거기간 전·후에 실시된 조사자료 간의 결과 비교는 물론 유선전화조사 자료와의 비교를 통해 온라인 조사에서 발생 가능한 횡단적·종단적 편향이 어떠한 특성을 보이는지 파악하고자 한다.

II. 선행연구 검토

1. 선거여론조사에서의 온라인 조사

온라인 조사는 과거 선거여론조사에서 널리 사용되어 온 유선전화조사의 단점을 보완할 가능성이 높은 조사방법으로 평가되어 왔다. 예컨대 온라인 조사는 유선전화조사에 비해 1인 가구, 젊은 층, 일과시간 중에 가정에 머무르는 가구원이 없는 경우에도 상대적으로 접근이 용이하다(이계오·장덕현 2009; 조성겸 외 2011). 또한 온라인 조사는 응답자가 직접 접속하여 조사에 응하는 형태이므로 응답자 선정을 위한 노력과 비용이 적을 뿐만 아니라, 짧은 시간 동안 많은 양의 자료를 수집할 수 있으며, 시·공간적인 제약을 받지 않는다는 장점이 있다(이계오 외 2012).

실제로 온라인 조사를 통해 수집한 결과와 다른 조사방법을 통해 수집한 결과들을 비교한 경우, 온라인 조사의 긍정적인 면을 살펴볼 수 있었다. 온라

인 조사와 유선전화조사를 이용하여 정당지지도 관련 선거여론조사를 실시한 후 비교한 경우, 각각의 결과가 매우 유사하게 나타났는데, 특히 연령별 투표율을 고려하면 두 조사 간의 차이가 더욱더 감소함을 확인할 수 있었다(조성검 외 2005). 또한 온라인 패널조사와 면접조사를 동일한 표본추출방법과 동일한 형태의 설문문항을 이용하여 비교·분석한 결과, 대등한 수준의 데이터 품질을 가지고 있는 것으로 나타났고, 예측타당도와 내적 일관성, 신뢰도 측면에서는 오히려 더 우수한 데이터 품질을 가지고 있는 것으로 나타나기도 하였다(이지영·김진교 2006).

이와 같은 장점을 가지고 있는 것으로 평가받는 온라인 조사지만, 반대로 표본의 대표성 문제라는 한계를 가지고 있기도 하다. 대표성 문제는 첫째, 포함오차의 측면에서 접근할 수 있다. 인터넷 이용자와 비이용자 간의 사회인구학적 특성의 차이는 비교적 큰 편인데, 특히 고연령층은 젊은층에 비해 인터넷 이용률이 낮으므로 상대적으로 온라인 조사에 참여하기 어렵다(Dillman et al. 2008; 이계오·장덕현 2009). 뿐만 아니라 국내 조사 회사들이 응답자를 자회사의 온라인 응답자패널에서 추출·섭외하는 관행 때문에 추정치의 이론적 기초인 확률표집이 가장 지켜지지 않는 것으로 의심되기도 한다(한신갑 2012).

온라인 조사의 대표성 문제와 관련된 두 번째 측면은 자기선택오차에 따른 편향의 발생가능성이다. 대부분의 온라인 조사의 경우 자발적인 의사에 따라 패널(회원 DB)을 구축하고 이 패널에서 연구의 목적에 따라 조사 대상자를 추출하는 비확률 표본추출 방법을 기초로 하고 있으므로, 자발적 참여 및 무응답 등으로 인한 편향의 발생이 가능하다(김영원·배예영 2010). 실제 인터넷 응답자들의 경우 자발성이 기본적으로 높은 것으로 나타났으며, 자발성에 따라 정치성향에도 차이가 있는 것으로 나타났다(조성검 외 2005).

특히, 이와 더불어 종단적 연구를 위해 패널조사를 실시할 경우, 조사가 반복되는 과정에서 자기선택오차로 인해 탈락되는 응답자들의 특성에 주목해볼 필요가 있다. 선거여론조사를 패널조사 형식으로 실시한 선행연구들을 살펴보면 상대적으로 저학력·저소득층의 탈락률이 높았고, 상대적으로 정치에 대한 지식과 관심이 많은 층이 과대표집되는 정도가 높았던 것으로 나타났다

(김장수 2007). 또한 연령대의 경우 20대와 60대의 탈락률이 상대적으로 높았다(허명희 외 2009). 이러한 특성들이 정치적 성향과 관련 높은 변수인 것을 고려하면, 패널조사가 진행되는 과정에서 정치적 편향이 심화될 가능성도 존재할 것이다.

그러나 종단적 조사에 대한 선행연구들의 경우 유선전화조사를 중심으로 실시되었고, 반복 조사에서 탈락률과 관계된 특성들이 온라인 패널조사에서는 어떻게 나타나는지 활발히 연구되지 못 했다. 온라인을 통한 패널조사가 선거여론조사에서 활발히 사용가능하다면, 온라인 조사의 조사비용이 상대적으로 저렴하다는 장점을 이용하여 패널조사가 가지고 있는 단점 즉, 대표성 및 응답률 유지를 위한 초기패널 구축과 여러 번의 조사과정에서 발생하는 비용부담의 문제(김장수 2007)를 해결할 수 있을 것이다.

이러한 온라인 조사의 대표성 문제가 선거여론조사에서 특히 문제가 되는 이유는, 온라인 조사에서 대표성의 결여가 정치적 성향의 편향을 일으킬 가능성이 높기 때문이다. 일반적으로 고연령층의 정치적 성향이 더 보수적인 것으로 알려진 것을 고려하면, 온라인 조사에 고연령층이 과소표집될 경우 조사결과에서는 진보적인 편향이 발생할 가능성이 높다. 또한 동일 연령대 내에서도 이용자와 비이용자 간의 학력수준이나 직업 등에서 차이가 존재하는 것으로 알려져 있으며, 특히 온라인 조사 응답자들의 경우 상대적으로 고학력과 화이트칼라 등으로 편중되어 있다(이게오·장덕현 2009). 일반적으로 교육수준이 높을수록 진보정당을 지지하는 성향을 보이며(조성겸 외 2005; 이현출 2004; 장상철 2008), 더 나아가 정당일치감을 통해 특정후보에 대한 투표결정에 영향을 미치는 것으로 알려져 있는데(장상철 2008), 특히 이러한 정치적 성향 차이는 최근 온라인을 이용한 SNS 사용자들에서도 유사하게 나타남(장덕진 2011; 이세희·이종균 2013)을 고려하면 온라인 조사를 통해 수집한 자료는 진보적 편향이 존재할 가능성이 있다. 이는 유권자의 지지도를 정교하게 추정하여야 하는 선거여론조사에서 기존 조사들의 단점을 보완할 수 있는 장점을 가지고 있음에도 불구하고 온라인 조사가 활발히 사용되지 못하는 이유 중 하나로 볼 수 있을 것이다.

지금까지의 논의들을 정리하면, 온라인 조사의 경우 첫째, 표본추출과정에

서 포함오차와 자기선택오차로 인한 대표성의 문제가 발생할 수 있다. 둘째, 이는 응답자들의 사회인구학적 특성과 연관되어있으므로 정치적 성향의 편향이 발생할 수 있다. 마지막으로 이와 더불어 패널조사의 특성상 자발성에 의거한 자기선택오차와 조사 진행과정에서의 탈락으로 그 편향이 심화될 가능성도 있다. 따라서 본 연구에서는 온라인 조사를 통한 선거예측을 실시해 보고 이 과정에서 이러한 편향들이 어느 정도로 발생하는지 확인함으로써 선거 여론조사에서 온라인 조사의 활용 가능성을 증대시키고자 한다.

한편 유선전화를 이용한 전통적인 전화면접조사도 온라인 조사와 동일하게 포함오차의 위험성이 큰 조사방법 중의 하나이다. 과거에는 전화번호부를 표본틀로 사용함으로써 발생하는 포함오차와 이에 따른 사회경제적 편향이 문제였으나(강현철 외 2008), RDD(Random Digit Dialing: 임의번호걸기)가 보편화됨으로써 이에 대한 우려는 감소하였다. 그러나 최근 20~30대의 젊은 세대를 중심으로 유선전화를 보유한 가구가 점차 줄어드는 경향(정용찬 2014)으로 인해 RDD를 이용하더라도 유선전화만을 이용한 표집은 젊은 층을 충분히 표집하기 어려워 다시 포함오차의 위험성이 대두되었다. 특히, 앞서 논의한 연령과 정치적 성향의 관계를 고려했을 때 이 경우에는 정치적 성향의 편향이 발생할 가능성이 있다.

실제로 본 연구의 연구시점과 동일한 2014년 지방선거에 대한 150건의 전화조사의 정확성을 평가한 결과, 유선전화조사보다 유·무선전화 혼합조사의 결과가 더 정확한 것으로 나타나(김영원·황다솜 2014) 유선전화조사의 포함오차 가능성을 뒷받침하는 것으로 볼 수 있다. 이러한 맥락에서 유무선전화 혼합조사가 지배적인 조사방법이 되었으며, 최근에는 유선전화와 스마트폰앱 패널을 이용한 혼합조사 또한 실시되고 있다. 그러나 2014년 지방선거 광역단체장 선거여론조사에서 유선RDD를 통한 피조사자 선정방법은 여전히 적지 않게 사용되는 방법(약 34.7%)으로 나타났다(중앙선거여론조사공정심의위원회 2014a).

이렇게 유선전화조사도 온라인 조사와 마찬가지로 포함오차의 가능성이 가장 큰 조사방법 중의 하나이다. 그러나 온라인 조사가 고연령층을 포함하지 못해 편향이 발생할 가능성이 있는 것에 반해, 유선전화조사의 경우 젊은

연령층을 포함하지 못해 편향이 발생할 가능성이 높다. 따라서 동일하게 포함오차가 발생하더라도, 그로 인해 발생하는 정치적 성향의 편향은 반대로 나타날 확률이 크다. 이를 고려할 때 RDD를 이용한 유선전화조사와 온라인 조사를 비교한다면, 온라인 조사가 가지고 있는 편향의 특성과 방향을 더욱 더 뚜렷하게 살펴볼 수 있을 뿐만 아니라, 유선전화조사가 가지고 있는 편향의 성격 또한 함께 살펴보고 각각에 대한 시사점을 도출해 볼 수 있을 것이다.

2. 가중치를 활용한 예측결과 보정

일반적으로 실시되고 있는 선거여론조사들은 대부분 수집한 원자료에서 자체적으로 산출한 가중치를 적용하여 결과를 보정한 후 발표하고 있다. 따라서 본 연구에서도 다른 선거여론조사들과 동일하게 가중치를 이용하여 보정을 실시한 뒤 그 결과를 살펴보고자 한다. 엄격히 말하자면 가중치는 개별 응답자가 표집될 확률에 기반을 두어야 하나 실제 조사에서는 비확률표집에 해당하는 할당표집 후 할당을 맞추지 못했을 때의 보정용으로 사용된다.

일반적으로 선거여론조사에 사용되고 있는 가중치들은 대부분 연령, 성별, 지역과 같은 사회인구학적 변수들을 이용하여 산출 및 적용되는 것으로 확인된다. 2014년 제6회 지방선거와 관련된 선거여론조사들의 경우(중앙선거여론조사공정심의위원회 2014a), 전체 816개 조사 중에서 약 94%에 달하는 770개의 조사가 ‘연령’, ‘성별·연령’, ‘연령·지역’, ‘성별·연령·지역’의 조합으로 진행된 것으로 나타났다. 한편, 학력이 정치적 성향에 영향을 미칠 가능성(조성겸 외 2005)을 고려할 때, 학력 변수가 가중치 적용에 활발히 사용될 것으로 예상하였으나 이를 고려한 조사결과는 매우 소수에 불과했다¹⁾. 이는 학력의 경우 정확한 모수치의 측정이 어렵기 때문에 가중치 산출에 적용하기 어려우며, 특히 젊은 세대의 경우 대학진학률이 현격하게 높아진 상황에서 이를 가중치 산출에 유용한 변수로 간주하기에는 어려움(조성겸 외 2011)이

1) 등록결과를 살펴보면, 성·연령·지역·기타(학력 등)으로 제시되고 있으며 전체 816개의 조사결과 중에서 8건에 불과하였다(중앙선거여론조사공정심의위원회 2014a).

있기 때문으로 보인다. 실제로 센서스 결과로 교육수준을 보정한 경우 오히려 정확도가 떨어지는 사례도 나타났다(조성겸 외 2005).

최근 들어서는 일반적인 가중치 적용방법에서 더 나아가 성향점수 가중방법이 온라인 조사의 결과보정에 꾸준히 시도되어 왔다(김영원·배예영 2010; 김원용·이홍철 2003; 이계오·장덕현 2009; 조성겸 외 2011; 허명희·조성겸 2010; Taylor et al. 2001). 온라인 자료의 표본분포를 확률표집한 준거자료(reference data)의 표본분포와 비슷하게 조정하고자 가중치를 부여하는 방법으로서, 비확률표집에서 자연적으로 발생하는 선택편의(selection bias)를 최소화하고자 하는 방법이다(Dillman et al. 2008). 여기서 응답자별 성향점수는 준거자료와 온라인 자료에 공통으로 있는 변수들을 활용하여 로짓모형으로 구한, 그 응답자가 온라인 자료에 속할 확률을 뜻하며 그 성향점수의 역수를 응답자별 가중치로 활용한다. 결국 가중치의 기준이 되는 여러 변수들을 성향점수라는 하나의 차원으로 축소하면서 정보의 손실이 발생하지만, 기존의 가중치 방식에서 정확한 보정을 위해 가중치를 산출하는 기준변수들의 수를 증가시킬수록 변수 간의 조합 또한 증가하게 되어 관찰사례 수가 적은 조합의 정확성이 저하된다는 문제를 보완할 수 있다. 특히 해리스 인터랙티브(Harris Interactive)라는 여론조사업체가 2000년 미국 대통령선거 예측조사에서 이 방법을 사용하여 실제결과와 매우 유사한 결과를 발표(Taylor et al. 2001)한 이후부터 많은 온라인 조사의 추정치 보정에 활발히 사용되기 시작하였다.

이와 더불어 좀 더 최근에는 성향점수 가중방법 대신, 성향점수를 이용한 매칭방법을 통해 모바일 조사결과의 보정을 시도하는 연구도 발표되었다(강정한 외 2014). 성향점수를 이용한 매칭방법은 성향점수를 이용한 가중방법과 논리적으로 동일하나, 구체적인 매칭방식에 따라 그 결과에 차이가 있다. 온라인 자료의 사례들 중 준거자료의 사례와 성향점수가 같거나 가장 비슷한 사례들을 뽑아 준거자료에 매칭시킨 후 관심 모수를 추정한다는 점에서 보정의 원리에 직관적으로 가장 충실한 장점이 있다. 이러한 방식은 또한 여러 보정변수를 하나의 성향점수로 축소함에 따라 생기는 문제를 보완할 수 있다. 두 자료 간 성향점수가 겹치는 구간(region of common support)에 한정시켜

매칭을 시도하거나, 가중치 기준변수별로 가중 과정에서 자료의 왜곡이 심해 지지는 않는지 검토가 용이하다(Guo & Fraser 2010).

본 연구에서는 온라인 조사를 통한 선거결과 예측과 보정을 위해 전통적인 가중치 방법 및 성향점수 매칭방법을 모두 사용하고자 하며, 가중치를 적용하는 기준변수로 연령과 성별 외에도 2012년 대통령선거 투표행태를 사용하고자 한다. 우리나라의 경우 대통령에 대한 호감도가 지방선거에서의 후보 선택에 유의미한 영향력을 미치기 때문이다(강신구 2014). 특히 2012년 대통령선거의 경우 2000년대 이후 가장 높은 전국투표율을 보일 만큼 대중의 큰 관심과 참여가 있었기에 응답자들의 정치행태를 잘 반영할 것으로 보았다.

지방선거와 대통령선거의 투표율(박경미 2014)과 후보선택기준(류재성 2011)에 영향을 미치는 사회적 요인은 상이하다는 주장에 따르면, 지난 제5회 지방선거 투표행태를 이용하여 예측 및 보정을 실시하는 방법을 고려할 수 있다. 그러나 같은 지방선거라고 하더라도 정권 중·후반기에 실시되는 경우에는 지방선거가 정부 및 여당에 대한 평가의 성격을 보이며, 제5회 지방선거도 유사한 결과를 보였다는 점에서(류재성 2011; 황아란 2014) 제6회 지방선거와 상이한 투표행태를 보였을 가능성이 높아 이를 이용하여 가중치를 구하기에는 무리가 있다. 뿐만 아니라 대통령선거와 제6회 지방선거 간의 시간차는 약 1년 반이었던 것에 비해 제5회·6회 지방선거 간에는 4년이라는 상대적으로 큰 시간적 공백이 존재하였으며, 특히 서울시장 선거의 경우 무상급식 지원범위에 관한 서울특별시 주민투표, 또 이와 관련된 오세훈 전 시장의 사퇴 및 보궐선거 등의 많은 정치적 이슈들이 존재하였기 때문에 그간 투표행태가 변화되었을 가능성이 높다. 따라서 이보다는 2012년 대선에서의 투표행태가 이번 서울시장 선거 유권자들의 성향을 잘 반영할 것으로 보았다.

한편, 성향점수를 이용하여 온라인 조사 및 유선전화조사 결과 간의 비교를 시도한 국내 선행연구들의 경우, 선거여론조사를 주제로 진행하였지만 각기 다른 조사업체를 통해 개별적으로 수집된 자료를 이용하여 비교·분석하였거나(김영원·배예영 2010), 동일한 조사기관을 이용하였으나 선거여론조사를 주제로 진행하지 않았다는(이계오·장덕현 2009) 한계가 있다. 따라서

본 연구에서는 온라인 조사결과와 유선전화조사결과 그리고 실제 투표결과를 서로 비교할 뿐만 아니라, 이 과정에서 동일한 조사업체와 설문지를 이용하여 조사업체 간에 상이할 수 있는 표집 및 조사방식 및 설문지 구성에서 발생하는 차이를 통제하고자 한다.

정리하자면, 본 연구는 선거여론조사로서 온라인 조사의 가능성을 살펴보기 위해 서울시장 선거 전에 수집된 온라인 조사자료를 이용하여 선거예측을 실시하고 실제 선거결과와 비교한다. 이러한 선거예측 과정에서는 성별, 지역, 2012년 대선 투표후보 등의 정보를 사용한 가중치 적용 및 성향점수 매칭을 실시한다. 또한 서울시장 선거 후에는 같은 온라인 응답자패널로부터 수집된 실제 투표행태 및 동일한 조사회사의 유선전화조사 결과를 투표결과와 비교한다. 이 과정의 보정에는 성별, 연령별 실제 투표율을 활용한다. 이처럼 사전 예측 및 사후 보정결과의 비교평가를 모두 실시하여 온라인 조사가 가지고 있는 횡단적·종단적 편향의 특성을 면밀히 살피고자 한다.

III. 연구방법

1. 자료수집 설계

분석을 위해 수집한 온라인 패널자료는 2014년 제6회 전국동시지방선거 기간 전후 총 다섯 차례에 걸쳐 조사되었으며, 서울시장 선거를 포함한 지방선거 및 대통령선거 관련 여론조사 경험이 다수 있는 한 조사회사의 온라인 패널 데이터베이스(DB)에 속한 응답자 풀(pool)을 활용하였다. 이 패널DB에 속한 응답자 풀의 정확한 선정방식은 연구자에게 공개되지 않았으나, 이메일 확인이 가능한 사람에게 한정되었으며 앞에서 논한 포함오차나 자기선택오차의 한계를 갖고 있음은 분명하다. 이러한 한계 내에서 서울시장 선거구의 유권자를 대상으로 서울시내 권역별, 연령대별 인구비율을 고려한 할당표본추출(quota sampling)을 실시하였다. 1차부터 4차 조사까지는 모두 선거 전에 실시되었으며, 주로 선거 관련 이슈들이 발생한 직후에 조사하였다. 5차 조사

는 선거 직후 실시하였다.

한편 유선전화조사의 경우 온라인 패널조사에서 사용된 문항과 동일하게 구성된 문항들을 바탕으로 선거 직후 함께 실시하였다. 이는 유선전화 사용자를 대상으로 전화면접(CATI: Computer Assisted Telephone Interviewing) 방법을 통해 진행되었으며, 표본추출 과정에서는 성별과 연령을 기준으로 할당하여 RDD방식을 이용하여 추출하였다.²⁾ 온라인 조사와 유선전화조사 모두 조사회사에서 실제로 사용되어 온 표본추출, 협조요청, 조사방법 등을 이용함으로써 현실에서 빠른 속도로 수행되고 있는 선거여론조사의 문제점 및 개선 방안을 탐색하고자 하였다.

선거 전 예측을 위한 성별·연령별 가중치나 성향점수를 구할 때는 선거의 성별·연령별 실제 투표율을 사전에 알 수 없다는 한계가 있다. 따라서 본 연구에서는 2012년 대통령선거 당시 서울시의 성별·연령별 투표율 및 투표 후보 등을 활용하여 예측을 시도하였다. 선거 후 사후보정을 위해서는 지난 2014년 9월 2일, 즉 서울시장 선거 후 두 달 반 후에 중앙선거관리위원회(2014)에서 발표한 10% 표본조사를 이용하여 성별·연령별 서울시장 투표율을 산출하고 가중치에 활용하였다(중앙선거관리위원회 2014).

2. 변수

선거 전 예측을 위한 종속변수는 4차 조사의 ‘서울시장 지지후보’를, 투표 결과 보정을 위한 종속변수는 5차 조사와 유선전화조사의 ‘서울시장 선택후보’를 이용하였다. 이는 응답자들에게 서울시장선거에서 어떠한 후보를 지지하였는지 혹은 실제로 어떠한 후보에게 투표하였는지를 측정한 것이다.

온라인 조사에서 나타날 수 있는 진보편향을 판단하기 위한 변수로는

2) 그러나 시간과 비용상의 문제로 인해 할당이 엄격히 지켜지지 못 하였으며, 응답률 또한 측정되지 못하였다는 한계가 존재한다. 단, 지방선거 기간 중 후보자등록 신청 개시일(5월 15일) 이후 실시된 광역단체장선거관련 조사결과를 살펴보면, 유선RDD를 이용한 유선전화 면접조사들(총 30개)의 응답률 평균은 14.7%(최소 6.6, 최대 29.3, 표준편차 4.53)으로 나타났음을 밝힌다(중앙선거여론조사공정심의위원회 2014b).

‘2012년 대선 투표행태’를 활용하였다. 이 변수는 조사시점으로부터 약 1년 반 전인 2012년 대통령선거 당시 어떤 후보에게 투표하였는지를 질문한 것이며, 대선 당시 서울시에서 실제로 보인 후보별 득표율과 비교하여 온라인 및 유선전화 응답자의 정치적 편향 정도를 살펴보고자 한다. 또한 1차~5차 조사 간 대선 후보별 득표율도 비교하여, 패널조사 과정에서의 탈락(attrition)에 따라 정치적 편향이 심화되지는 않는지 알아보려 한다.

가중치 산출을 위한 연령구간은 20대 이하, 30대, 40대, 50대, 60대 이상 등 총 5가지 범주로 구분하였다. 이와 함께 ‘권역’, ‘학력’, ‘정치적 성향’, ‘투표 여부’ 등을 수집하였다. ‘권역’은 응답자의 거주지역에 따라 강북동(중랑구, 동대문구, 성동구, 광진구, 도봉구, 노원구, 강북구, 성북구, 강동구), 강북서(은평구, 종로구, 중구, 용산구, 마포구, 서대문구), 강남서(영등포구, 양천구, 구로구, 금천구, 강서구, 관악구, 동작구), 강남동(강남구, 서초구, 송파구) 총 4가지로 구분하였다. ‘서울시장 선거 투표 참여여부’는 서울시장 선거 투표에 참여하였는지를 측정하는 것이다.

3. 분석방법

우선 온라인 조사를 이용한 선거결과 예측을 위해 4차 조사에서 수집한 ‘서울시장 지지후보’ 변수를 바탕으로 지난 대선의 투표행태를 활용하여 각 후보별 득표율을 예측해 본다. 이러한 예측의 오차에 관여할 수 있는 요인은 다양하다. 표집단계부터 나타나는 포함오차와 자기선택오차, 특히 정치적 성향과 관련된 포함오차 및 패널조사 중 탈락에 따른 자기선택의 강화 등이 관여할 수 있다. 또한 지난 대선과 이번 서울시장 선거의 성별·연령별 투표율 차이에 따른 가중치의 오차, 부동산의 표심을 사전에 정확히 반영할 수 없는 한계에서 나타나는 오차 등이 존재할 수 있다. 이러한 다양한 오차의 정도를 실제 선거결과와 비교·분석할 것이다.

다음으로 선거 후 5차 조사와 유선전화조사 결과도 보정하여 실제 선거결과와 비교·분석한다. 우선 1차~4차 조사의 지지후보와 5차 조사의 투표후보 분포를 비교하며 자기선택오차의 강화여부를 판단한다. 또한 4차 조사에

서 부동층의 실제 투표행태를 관찰함으로써 부동층이 예측오차에 미친 영향도 추적한다. 한편 선거 후 실시된 5차 조사와 유선전화조사의 가중치 산출에 중앙선거관리위원회의 서울시장 선거 성별·연령대별 투표율을 사용하기 때문에, 예측단계에서 사용한 대선 투표율의 오차 정도도 평가할 수 있다. 이러한 오차들을 고려하고도 남는 오차는 패널조사의 응답자 표집 당시부터 존재하는 자기선택오차나 포함오차로 평가할 수 있을 것이며 이러한 오차의 특성은 특히 유선전화조사 결과와의 비교를 통해 알아본다.

예측 및 사후보정은 두 단계로 실시하였다. 첫 번째 단계에서는 성별·연령대별 가중치를 적용하는 방법을 사용하였다. 이는 성별·연령대에 따라 서울시장 선거 (예상) 투표비율을 계산한 후, 조사자료의 분포와 비교하여 표집 확률의 역수를 가중치로 적용시킨 방법이다. 선행연구에서 가중치로서의 유용성이 의심되는 학력변수의 경우(조성겸 외 2005), 본 연구에서도 연령변수와 함께 사용할 경우 효과가 거의 없어 최종분석에서는 제외되었다.

두 번째 단계에서는 첫 번째 단계에 2012년 대선 투표후보 변수를 추가하여 응답자의 정치적 성향을 고려한 보정을 실시하였다. 그런데 성별과 연령대에 대선 투표후보를 추가하면서 가중치의 기준이 되는 변수 간 조합의 수가 급격히 늘어남에 따라 빈도 수가 0이거나 0에 가까운 조합들이 다수 발생한다. 결과적으로 가중치 방식이 불가능해지거나 타당성이 떨어지기 쉬운데, 본 연구는 이에 대한 대안으로서 앞서 검토한 성향점수 매칭방법을 이용한다. 본 연구는 성별, 연령, 정치적 성향의 조합별로 실제 투표한 모집단의 분포와 동일한 가상의 준거집단을 구성한 후, 이 집단의 개별 응답자와 동일하거나 가장 가까운 성향점수 값을 갖는 표본을 조사 자료에서 추출 혹은 매칭하였다.³⁾ 이렇게 매칭하면 실제 투표자보다 과소표집되었던 특성의 응답자는

3) 우선 성별, 연령대, 대선 투표후보 조합별로 최소한의 사례수를 확보하기 위해 1만 명의 가상의 준거자료를 생성했다. 매칭방법을 위해서는 STATA 11.0의 `psmatch2` 명령어와 `ties` 옵션을 이용하였다. 이는 통제집단(즉 응답자) 중 실험집단(즉 가상 1만명 집단)과 성향점수가 같은 혹은 가장 비슷한 “모든” 케이스들을 매칭하는 방법으로서, 연령대와 성별만 보정변수로 이용한 경우는 보통의 가중치 방법과 동일한 결과를 얻을 수 있다.

그만큼 중복매칭이 되기 때문에 높은 가중치를 부여하는 것과 같은 논리이다.

그러나 중복매칭이 생기면서 실제보다 사례수가 늘어나고 추정치의 표준오차가 부정확해지는 문제가 발생하며, 앞서 언급했듯이 매칭에 사용하는 변수가 많아짐에 따라 차원을 하나로 축소한다는 점에서 회귀분석에서 통제변수를 투입하는 것과 본질적으로 다르지 않다는 한계가 존재한다(Pischke & Angrist 2009). 따라서 본 연구는 추정치의 표준오차에 의존하는 신뢰구간은 고려하지 않으며 성향점수 매칭을 가중치 적용이 어려운 경우의 보조적 역할로 활용한다.

IV. 분석결과

1. 응답자의 특성

<표 1>은 각 조사별로 조사기간, 응답자 수, 탈락률을 간략히 정리한 것이다. 표에 제시된 바와 같이 온라인 1차~4차 조사의 경우 선거일(2014년 6월 4일) 이전, 온라인 5차 조사와 유선전화조사의 경우 선거일 이후에 수집된 것이다.

<표 1> 각 조사별 조사기간과 응답자 수

조사 구분	조사기간	응답자 수	탈락률
온라인 패널조사	1 차	3 월 26 일 ~ 4 월 1 일	1,860
	2 차	4 월 17 일 ~ 4 월 24 일	1,601
	3 차	5 월 13 일 ~ 5 월 20 일	1,314
	4 차	5 월 28 일 ~ 6 월 3 일	1,165
	5 차	6 월 10 일 ~ 6 월 17 일	1,066
유선전화조사	6 월 12 일 ~ 6 월 13 일	503	-

<표 2>는 각 자료별 응답자들의 특성을 요약하여 제시한 표이다. 가장 우측열에 제시되고 있는 실제결과의 경우, 성별, 연령, 권역은 중앙선거관리위원회(2014)에서 발표한 10% 표본조사 결과를 기준으로 제시하였으며, 지난 2012년 대선 결과의 경우 서울시 결과를 기준으로 한 것이다.

결과를 살펴보면, 온라인 조사의 경우 전체적으로 남성응답자의 비율이 실제 투표율보다 높으며, 60대 이상의 응답자가 심각하게 과소표집되어 있어 온라인을 통한 고연령층 접근의 어려움을 알 수 있다. 반면 권역의 경우 실제 결과와 유사한 비율을 보여 비교적 할당에 충실하였다. 유선전화조사의 경우 고연령층의 압도적 투표율을 고려한다면 고연령층이 의도했던 할당보다 과대표집되었음을 알 수 있다. 성별은 여성이 과대표집되어 역시 온라인 조사와 대조를 이루는데, 이메일의 경우는 남성이, 가정 유선전화의 경우는 여성이 더 접근이 용이하기 때문으로 보인다. 유선전화조사의 경우 이틀만에 조사를 마치면서 권역의 비율을 잘 지키지 못했다.

<표 2> 조사별 응답자의 특성

구분		온라인(1차) N=1,860		온라인(4차) N=1,165		온라인(5차) N=1,066		유선전화조사 N=503		실제 결과
		빈도	%	빈도	%	빈도	%	빈도	%	%
성별	남	1,010	54.30	677	58.11	628	58.91	225	44.73	48.70
	여	850	45.70	488	41.89	438	41.09	278	55.27	51.30
연령	20 대	394	21.18	195	16.74	174	16.32	91	18.09	18.07
	30 대	474	25.48	312	26.78	285	26.74	89	17.69	18.40
	40 대	453	24.35	311	26.70	295	27.67	98	19.48	19.26
	50 대	389	20.91	267	22.92	240	22.51	95	18.89	19.96
	60 대 이상	150	8.06	80	6.87	72	6.75	130	25.84	24.31
권역	강북동	663	35.65	400	34.33	359	33.68	162	32.21	36.31
	강북서	313	16.83	195	16.74	182	17.07	111	22.07	16.79
	강남서	578	31.08	391	33.56	360	33.77	122	24.25	30.48
	강남동	306	16.45	179	15.36	165	15.48	108	21.47	16.41

(계속)

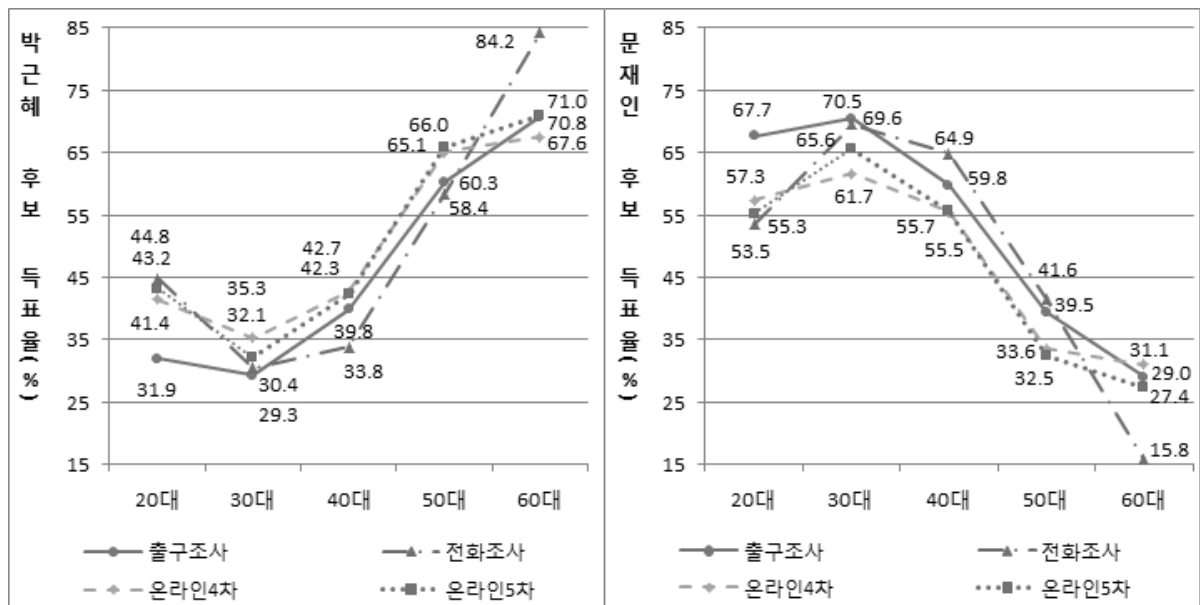
(계속)

		온라인(1차) N=1,860		온라인(4차) N=1,165		온라인(5차) N=1,066		유선전화조사 N=503		실제 결과
		빈도	%	빈도	%	빈도	%	빈도	%	%
학력	초졸 이하	4	0.22	3	0.26	1	0.09	34	6.76	
	중졸	17	0.91	7	0.60	7	0.66	22	4.37	-
	고졸	257	13.82	156	13.39	135	12.66	114	22.66	-
	대학교 재학	146	7.85	76	6.52	72	6.75	20	3.98	-
	대학교 졸	1,218	65.48	785	67.38	720	67.54	266	52.88	-
	대학원 이상	218	11.72	138	11.85	131	12.29	43	8.55	-
	모름/무응답	0	0	0	0	0	0	4	0.80	-
정치적 성향	매우 진보적	67	3.60	43	3.69	37	3.47	21	4.17	-
	다소 진보적	472	25.38	264	22.66	243	22.80	108	21.47	-
	중도	824	44.30	527	45.24	482	45.22	203	40.36	-
	다소 보수적	452	24.30	299	25.67	273	25.61	106	21.07	-
	매우 보수적	45	2.42	32	2.75	31	2.91	39	7.75	-
	모름/무응답	0	0	0	0	0	0	26	5.17	-
2012 년 대통령선거 선택후보 (투표행태)	박근혜	744	40.00	482	41.37	443	41.56	249	49.50	48.18
	문재인	822	44.19	511	43.86	461	43.25	184	36.58	51.42
	이정희	11	0.59	8	0.69	8	0.75	0	0	-
	박종선	3	0.16	1	0.09	1	0.09	0	0	0.05
	김소연	5	0.27	3	0.26	3	0.28	0	0	0.06
	강지원	7	0.38	5	0.43	5	0.47	0	0	0.18
	김순자	4	0.22	2	0.17	2	0.19	2	0.40	0.08
	기억나지 않음	85	4.57	50	4.29	46	4.32	13	2.58	-
	투표하지 않음	145	7.80	92	7.90	86	8.07	50	9.94	-
	투표권 없었음	34	1.83	11	0.94	11	1.03	5	0.99	-
서울시장투표 참여유무	투표함	-	-	-	-	950	89.12	424	84.29	58.60
	투표 않음	-	-	-	-	116	10.88	79	15.71	41.40
서울시장 지지후보 (1 차, 4 차) 선택후보 (5 차, 전화)	정몽준	391	21.02	268	23.00	287	30.21	169	39.86	43.02
	박원순	942	50.65	566	51.16	643	67.68	247	58.25	56.12
	정태홍	-	-	2	0.17	11	1.16	1	0.24	0.48
	홍정식	-	-	0	0	9	0.95	3	0.71	0.36
	지지후보 없음	391	21.02	108	9.27	-	-	-	-	-
	후보결정 못 함	8	0.43	220	18.88	-	-	-	-	-
	모름/무응답	-	-	1	0.09	0	0	4	0.94	-
	기타후보	128	6.88	-	-	-	-	-	-	-

주: 온라인 조사 설문지 구성 시 오류로 인해 2012년 대통령선거 선택후보 문항에서 '이정희 후보'가 선택 가능하였음.

학력의 경우 선행연구들에서 제시한 바와 같이(김장수 2007; 이계오·장덕현 2009) 온라인 조사에서는 비교적 높은 학력을 가진 응답자의 비율이 높은 것으로 나타났으며, 대졸 이상의 비율이 유선전화조사보다 17% 포인트나 높다. 이는 온라인 조사에서 고연령층이 훨씬 과소표집된 것과 일부 연관되어 있는 것으로 보인다. 이처럼 두 조사 간에 학력의 차이는 뚜렷하였으나, 응답자 본인이 평가하는 주관적 정치적 성향의 경우에는 온라인 조사와 유선전화조사 모두 고르게 분포되었다.⁴⁾

그렇다면 주관적 평가가 아닌, 지난 대통령선거에서 객관적으로 드러난 정치적 성향은 어떠한지 살펴보았다. 다음 <그림 1>은 지상파 방송3사(KBS·MBC·SBS)의 서울시 출구조사 결과와 비교한 <표 2>의 자료를 그래프로 구성한 것으로서 제시되는 득표율은 <표 2>의 값에서 ‘기억나지 않음’, ‘투표하지 않음’, ‘투표권 없었음’을 제외한 후 산출한 비율이다.



<그림 1> 2012년 대통령선거 득표율 결과 비교

4) *T*검정을 통해 확인해 본 결과에서도 두 집단 간에 통계적으로 유의미한 차이가 나타나지 않았다(4차와 비교 $t=1.1$, 5차와 비교 $t=1.2$).

온라인 조사의 경우 4차, 5차 모두 실제 결과와 큰 차이를 보이지 않았다. 60대를 제외하면 오히려 실제보다 박근혜 후보에게 투표한 비율이 약간 높아 온라인 응답자들이 일반적으로 진보편향이 있다고 보긴 곤란하다. 그러나 60대가 다른 연령대에 비해 박근혜 후보 편향을 보이지 않아 온라인 응답자 중 고령층은 다소의 진보편향이 추측된다. 한편 유선전화조사는 60대의 박근혜 후보 선호도가 분명해서 최고령층이 다소 과대표집되었을 뿐 아니라 보수적이라 볼 수 있다. 흥미로운 점은 유선전화조사와 온라인 조사 모두 공통적으로 20대에서의 박근혜 후보 편향이 나타났다는 점이다. 이는 조사 매체와 응답자의 정치적 편향의 관계 측면에서 최저연령층과 최고연령층이 가장 문제시될 수 있음을 보인다.

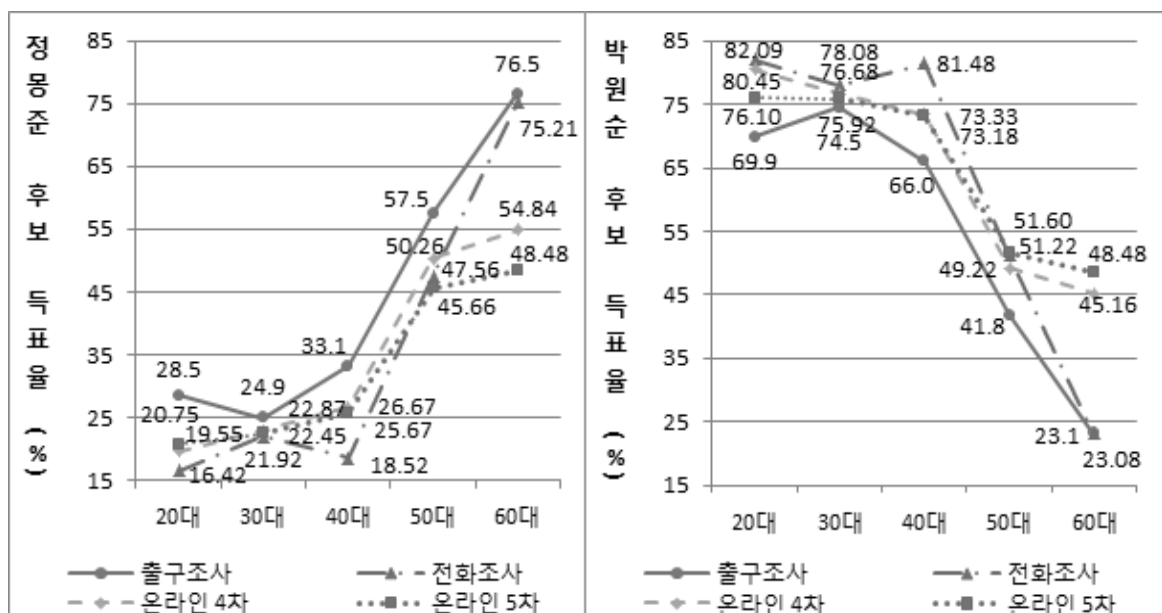
한편 1차 조사와 5차 조사까지의 탈락자 수를 확인하였을 때, 탈락자로 인해 성별과 연령대에서 다소 큰 폭의 비율변화를 보이는 것을 확인할 수 있었다. 그러나 2012년 대통령선거 선택후보와 정치적 성향에서는 비율의 차이가 비교적 크지 않음을 확인할 수 있었다. 특히 온라인 조사의 1차부터 5차 사이 지난 대통령선거 투표후보의 분포가 거의 동일하다는 점은, 패널조사가 반복됨에 따라 발생한 응답자의 탈락이 응답자의 정치적 성향과 연관이 없음을 보여준다. 이를 좀 더 자세히 검정하기 위해 1차~5차 조사 간 투표 행태의 분포를 통계적으로 비교한 결과, 차이가 전혀 없었으며(Pearson $\chi^2(36) = 11.3254$, p -값 = 1.000), 박근혜 후보와 문재인 후보 간의 상대적 득표율에 집중해 검정한 결과도 동일하였다(Pearson $\chi^2(4) = 1.1358$, p -값 = 0.889).

정리하자면, 본 연구에서 활용한 온라인 조사의 경우 일반적인 진보편향이 나타난다고 볼 수 없고, 서울시장 투표에 참여한 최종 응답자의 지난 대선 투표행태 분포의 경우 서울시의 실제 대선 결과와 거의 동일하였다. 즉, 온라인 조사의 진보편향은 최고령층을 제외하고는 횡단적, 종단적 편향 모두 발견되지 않았으며, 앞으로 살펴보겠지만 실제 서울시장 선거 결과보다 온라인 조사에서 박원순 후보가 압도적으로 많은 표를 얻은 것이 온라인 조사의 진보편향 때문이라고 설명하기 어려움을 뜻한다.

이처럼 진보적 편향이 나타나지 않은 것과는 대조적으로 온라인 조사의 응답자들은 조사대상인 서울시장 선거에 대한 매우 높은 관심을 보였다. 5차 조사 응답자들의 투표율 89.12%는 실제 투표율 58.6%보다 훨씬 높으며 유선 전화조사 응답자의 84.29%보다도 높다. 한편, 이러한 선거에 대한 관심편향도 종단적 차이는 보이지 않아, 1차~4차까지 알아본 서울시장 투표의지는 약 89.4%~91.5%를 보여 5차 조사의 실제 투표율 89.1%와 거의 일치하였다. 즉, 조사대상 선거에 대한 관심편향은 1차 조사 응답자부터 나타났으며, 최초 표집단계부터 자발성에 의존한 자기선택오차가 심했을 가능성을 암시한다. 아래에서 자세히 살펴보겠지만 이러한 편향은 고연령층에서 특히 심한 것으로 추론된다.

2. 조사별 서울시장 선거 투표결과

실제 투표에서 가장 높은 득표율을 보였던 상위 2인의 후보를 대상으로 <표 2>의 자료를 <그림 2>와 같이 지상파 방송 3사(KBS · MBC · SBS)의



<그림 2> 조사별·연령대별 득표율 결과 비교

출구조사와 비교·제시하였다. 선거 전 실시한 4차 조사 지지후보의 경우는 ‘지지후보 없음’과 ‘지지하는 후보를 정하지 못함’에 응답한 응답자들을 제외하고 지지율을 계산하였다. 즉, 이러한 부동층이 실제 투표에서는 각 주요후보(박원순 후보, 정몽준 후보)에게 다른 응답자들과 유사한 비율로 나뉠 것으로 가정하였다.

연령대별로 실제 투표결과와 유선전화조사 및 온라인 조사결과를 비교해보았을 때 주목할 만한 점은 유선전화조사의 경우 전체 득표율에서는 실제 투표결과와 유사한 결과를 보였지만, 각 연령대별로 나눠보았을 때는 60대를 제외하고는 박원순 후보를 선호한 편향이 오히려 더 강하게 나타난다는 점이다. 반면 온라인 조사의 경우 과소표집된 60대 이상에서의 박원순 후보에 대한 편향이 두드러지는데, 특히 4차에 비해 5차에서 다소 심화된 것으로 나타났다. 즉, 패널 응답자의 중도 탈락에 따른 편향이 전반적으로는 두드러지지 않았지만, 온라인 응답자 섭외가 특히 힘든 고연령층에서는 박원순 후보를 지지하는 사람들이 최초에도 더 잘 섭외되고 5차까지 남아 있었던 것으로 보인다. 20대의 경우 두 조사 모두 박원순 후보 선호가 비교적 뚜렷하여 지난 대선 보였던 박근혜 후보 선호와 대조되며 일반적 정치적 성향을 가늠하기 힘들다. 정리하자면 전체 득표율에서와는 달리 연령별 득표율에서는 온라인 조사가 유선전화조사보다 실제 결과에 근접하는 것으로 보이며, 온라인 조사의 경우 최고연령층의 박원순 후보 편향이 특히 심한 것으로 보인다.

투표 전과 후의 온라인 응답자 간 변심이 편향을 심화시키지 않는지 살펴보기 위해 온라인 4차 조사의 지지후보와 5차 조사의 투표후보 간 관계를 <표 3>을 통해 알아보았다. 정몽준 지지 응답자 중 박원순 후보에 투표한 응답자의 비율이 다소 높으나, 지지후보가 없거나 결정하지 못한 부동층의 투표는 특정 후보에 몰리지 않은 것으로 보인다. 4차 조사에서 정몽준 후보 지지(229명)와 박원순 후보 지지(489명)는 3.2 대 6.8의 비율을 보였는데, 4차의 부동층이 5차에서 두 후보로 갈린 비율도 이 비율과 비슷하다. 결과적으로 온라인 조사에서 보인 주요 편향이 부동층의 불확실성에 기인하거나 선거 전 응답자와 선거 후 실제 투표자 간 차이에서 기인한다고 보기는 힘들다.

<표 3> 온라인조사 4차 지지후보-5차 선택후보 간 관계

5차		4차					응답 거절
		정몽준	박원순	정태홍	아무도 지지 않음	아직 정하지 못 함	
정몽준	빈도	199	15	0	17	55	1
	%	86.9	3.07	0	27.42	32.93	100
박원순	빈도	28	474	0	36	105	0
	%	12.23	96.93	0	58.06	62.87	0
정태홍	빈도	2	0	2	5	2	0
	%	0.87	0	100	8.06	1.2	0
홍정식	빈도	0	0	0	4	5	0
	%	0	0	0	6.45	2.99	0
계	빈도	229	489	1	62	167	1
	%	100	100	100	100	100	100

3. 성별 · 연령 · 정치성향을 고려한 예측 및 보정 결과

우선 온라인 조사를 이용한 선거예측을 위해 4차 조사의 지지후보 분포에 성별 · 연령별 투표율을 고려한 가중치를 적용하였다. 단순히 성별 · 연령별 유권자 분포만으로 가중치를 줄 경우 투표율이 높은 고연령층의 영향력이 과소평가될 수 있기 때문이다. 그러나 선거가 실시되기 이전에는 투표율을 알 수 없으므로, 본 연구는 조사 시점에서 가장 최근의 주요 선거였던 2012년 대선 당시의 성별 · 연령대별 투표자 비율을 바탕으로 성별 · 연령대별 서울시장 선거 예상 투표자 분포를 계산하였고⁵⁾, 이를 4차 조사의 성별 · 연령대별 응답자 분포비율과 비교한 상대적 확률의 역수를 가중치로 산출하여 적용하였다. 가중치 산출의 과정은 <부록 1>에 정리하였으며 산출된 성별 · 연령별 가중치를 적용하여 예측을 실시한 결과는 <표 4>에 포함시키고 <그림 3>에 정리하였다.

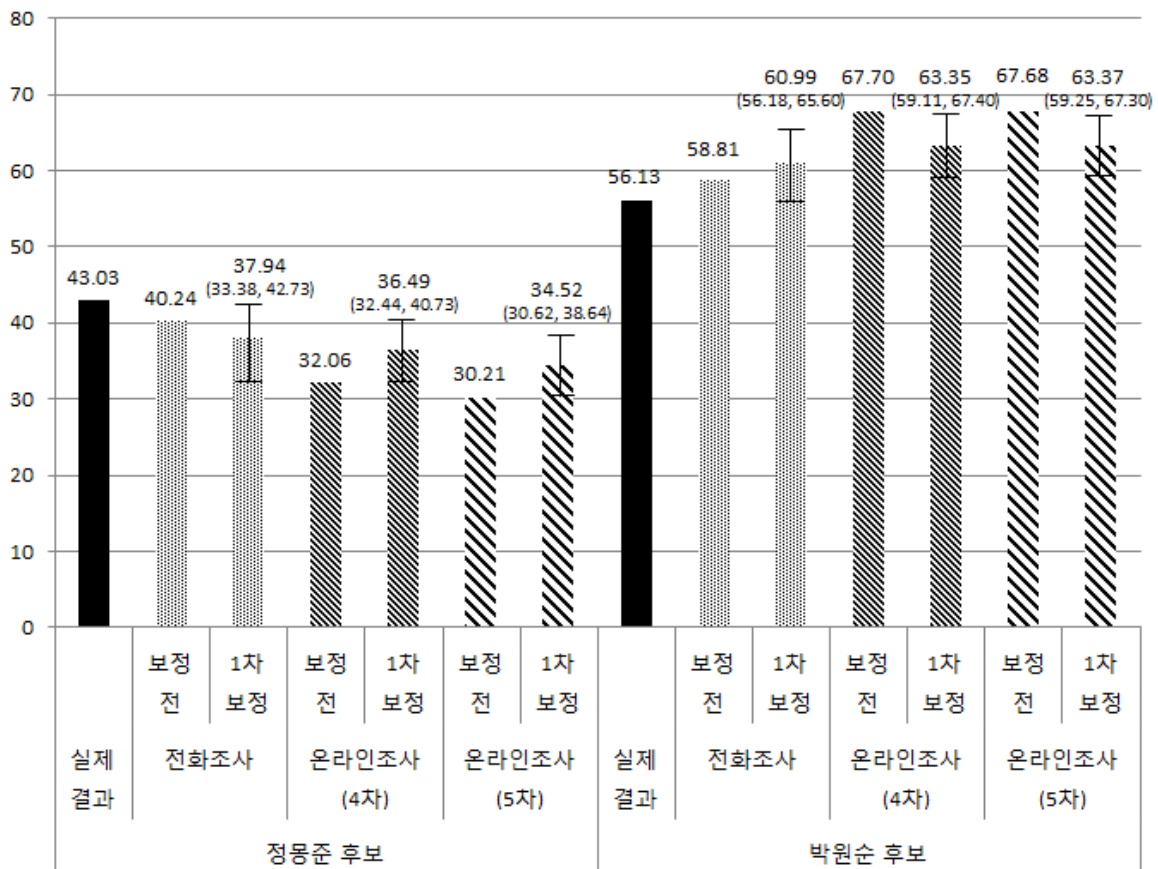
5) 성별 · 연령별 분포는 20대 · 30대의 경우는 서울특별시 거주주민의 비율을 바탕으로, 40대 이상은 전 국민의 비율을 바탕으로 하였다. 선거관리위원회의 자료가 40세 이상은 전국 단위에서만 10세 단위로 구분해 투표율을 제공하기 때문이다. 가중치에 사용한 분포들은 부록에 정리하였다.

<표 4> 각 조사별 보정 전·후 결과 비교(%)

구분		20대	30대	40대	50대	60대 이상	전체
정몽준 후보	출구조사	28.50	24.90	33.10	57.50	76.50	43.03
	유선 전화 조사	보정 전	16.42	21.92	18.52	47.56	75.21
			17.46	20.39	18.52	47.06	73.17
		1 차 보정	(09.85, 29.05)	(12.67, 31.13)	(11.46, 28.53)	(36.45, 57.95)	(63.97, 80.74)
	온라인 조사 (4 차)	보정 전	19.55	22.87	26.67	50.26	54.84
			20.12	22.72	26.53	50.04	55.46
		1 차 보정	(14.03, 27.99)	(17.63, 28.76)	(20.85, 33.10)	(42.73, 57.35)	(42.70, 67.54)
	온라인 조사 (5 차)	2 차 보정	18.82	23.06	27.30	51.51	55.70
		보정 전	20.75	22.45	25.67	45.66	48.48
			21.25	22.10	26.66	45.22	50.37
		1 차 보정	(15.51, 28.40)	(17.27, 27.82)	(21.30, 32.82)	(38.45, 52.17)	(38.25, 62.45)
	출구조사	69.90	74.50	66.00	41.80	23.10	56.13
박원순 후보	유선 전화 조사	보정 전	82.09	78.08	81.48	51.22	23.08
			80.68	79.61	81.48	51.85	24.68
		1 차 보정	(68.81, 88.77)	(68.87, 87.33)	(71.47, 88.54)	(41.00, 62.53)	(17.42, 33.72)
	온라인 조사 (4 차)	보정 전	80.45	76.68	73.33	49.22	45.16
			79.88	76.90	73.47	49.55	44.54
		1 차 보정	(72.01, 85.97)	(70.83, 82.02)	(66.90, 79.15)	(42.25, 56.87)	(32.46, 57.30)
	온라인 조사 (5 차)	2 차 보정	81.18	76.93	72.70	48.48	44.30
		보정 전	76.10	75.92	73.18	51.60	48.48
			75.74	76.38	72.19	52.39	47.13
		1 차 보정	(68.39, 81.83)	(70.57, 81.35)	(65.98, 77.65)	(45.46, 59.22)	(35.23, 59.36)
	출구조사	69.90	74.50	66.00	41.80	23.10	56.13
	유선 전화 조사	보정 전	82.09	78.08	81.48	51.22	23.08

주: 온라인 4차 조사는 선거 전 지지후보; 1차보정 결과 중 괄호 안은 신뢰구간을 의미

<그림 3>을 보면 온라인 4차 조사에서 이러한 가중치를 적용한 1차 보정의 결과, 정몽준 후보 지지율은 상향 조정, 박원순 후보 지지율은 하향 조정되어 실제 득표율에 가까워졌다. 그러나 여전히 실제 득표율과는 적지 않은 차이를 보이는데, 이는 성별 및 연령 가중치만으로는 고연령층의 박원순 후보 편향이 교정되지 않기 때문이다. 즉, 고연령층이 다른 연령층에 비해 정몽준 후보 지지도가 상대적으로 높으면서 뚜렷이 과소표집되었기 때문에 높은 가중치가 부여되었고 보정 전보다 실제 투표결과에 근접하였으나, 실제 투표자보다는 박원순 후보 편향이 심했기 때문에 교정 폭에 한계가 있다.



<그림 3> 성별 · 연령대별 가중치 적용 후 득표율 결과 비교

그렇다면 이러한 연령차별적 박원순 후보 편향을 보정하기 위해 성별·연령대별 가중치를 이용한 예측에서 더 나아가 2012년 대통령선거 투표행태를 바탕으로 심화된 예측을 실시하면 어떠할까? 예측결과를 확인하기 위해 성별·연령대별·투표한 대선 후보별 서울시장 선거 예상 투표자 분포를 구하였으며 <부록 2>와 같다. 이는 <부록 1>의 서울시장 선거 예상 투표율 분포값과 대선 후보별 득표율을 곱하여 산출한 값이다. 이 분포에서는 기타 후보까지 총 30개의 경우의 수가 나타나는데, 온라인 조사 응답자의 빈도가 거의 없는 경우들이 존재하기 때문에 앞에서 밝혔듯이 성향점수 매칭방법을 이용하여 보정을 실시하였다. 이렇게 보정을 실시한 결과는 <표 4>에서 ‘2차 보정’으로서 기존의 결과들과 비교·제시하였다.

2차 보정의 결과는 전체 예측값이 1차 예측값과 사실상 차이가 없었는데, 이러한 결과는 이미 <그림 1>에서 온라인 조사 참여자의 정치적 편향이 거의 없다는 관찰에서 예견된 것이라고 할 수 있다. 연령대별로 살펴보면 20대의 박근혜 후보 편향이 조정되어 정몽준 후보 지지율은 오히려 떨어지고 박원순 후보 지지율이 상승함을 확인할 수 있다. 이러한 경향은 전체적인 예측의 박원순 후보 편향을 심화시킬 수도 있었으나 다른 연령대에서의 보정에 의해 상쇄되었다. 본 연구에서 사용한 온라인 조사의 경우 권역별 할당을 지키기 위해 노력했고 최저와 최고 연령대를 제외한 대부분의 연령대에서 정치적 편향은 거의 없었던 것으로 보인다. 만일 응답자의 정치적 편향이 존재했다면 이러한 추가정보를 이용한 보정은 효과가 있을 것으로 기대한다.⁶⁾

선거 전 온라인 4차 조사의 지지후보를 이용한 투표예측에 이어 선거 후 5차 조사의 투표후보를 이용한 보정도 실시하여 <표 4>와 <그림 3>에 포함하였다. 선거 후 보정의 가중치는 실제 투표율을 사용하였고, 중앙선거관리위

6) 짧은 조사기간에 이루어져 권역별 할당이 지켜지지 않은 유선전화조사의 경우 특히 최고령층의 박근혜 후보 선호가 강했었고, 이러한 편향을 매칭방법으로 보정한 결과 본 논문에는 제시하지 않았지만 정몽준 후보 지지도는 더욱 하락, 박원순 후보 지지도는 더욱 상승하여 실제 서울시장 투표율에서 1차 보정보다도 멀어진다. 이러한 결과는 유선전화조사의 경우 1차 보정도 실제에서 더 떨어진 경향과 일관된다.

원회가 제공한 성별·연령별 투표율은 <부록 3>과 같다. 지난 대통령선거를 활용한 예상 투표율과 비교해 보면, 성별 비율은 거의 같으나 60세 이상 고연령층의 비율이 약간 늘어났음을 확인할 수 있다. 그러나 전체적으로 대선 당시의 투표율과 유권자 수를 곱해서 산출한 예상 투표자 분포와 큰 차이가 없기 때문에 보정의 결과도 4차 자료의 1차 보정과 큰 차이가 없다. 보정 전 정몽준 후보 득표율이 4차의 지지율에 비해 1.85% 포인트 낮은 것이 보정 후에도 거의 유지되었을 뿐이다.

유선전화조사의 경우 각 후보별 득표율이 보정 전에 비해 보정 후 실제결과와의 차이가 오히려 더 커짐을 관찰할 수 있다. 유선전화조사의 경우 비교적 정몽준 후보 지지경향이 높았던 60대 이상이 상대적으로 과대표집되어 있었고, 이를 보정하는 과정에서 정몽준 후보 득표율이 감소(더불어 박원순 후보 득표율은 증가)하게 되어 보정 전에 비해 실제결과와의 차이가 더 크게 나타난 것이다. 결국 보정 전 고령층의 과대표집에 의해 우연히 상쇄되어 있던 유선전화조사의 부정확성이 보정 후 드러났고, 유선전화조사로 쉽게 접근 가능했던 고연령층의 정치적 성향은 온라인 조사로 접근 가능했던 고연령층의 성향과 반대였다. 즉, 조사 매체에 따라 자기선택오차의 강도는 연령에 따른 매체 접근도에 따라 달라지고, 방향은 같은 연령대에서 반대로 나타났다.

V. 토의

본 연구는 선거여론조사에서 온라인 조사의 적용 가능성을 향상시키기 위해 2014년 서울시장 선거를 예로 실시한 온라인 조사의 편향과 그 특성을 이해하고자 하였다. 온라인 조사는 인터넷 사용이 가능한 응답자로 제한되고 응답자의 자발성에 의존하기 때문에 대표성 결여나 자기선택오차의 문제가 클 수 있다. 더구나 국내 조사회사들의 온라인 응답자가 추출 및 섭외되는 방식은 공개되지 않아 향상방안을 모색하기 쉽지 않다. 본 연구에서는 선거여론조사에서 중요하다고 판단되는 성별, 연령, 권역할당을 통해 이러한 한계를 우선적으로 줄이려 노력하였다. 또한 온라인 조사는 높은 탈락률이 문제가

될 수 있기 때문에 선거 전·후에 걸친 반복 패널조사를 실시하였으며, 정치적 편향을 보정하기 위해 지난 2012년 대선당시의 투표행태도 조사하였다. 이와 더불어 서울시장 선거 직후 유선전화조사도 병행하여 조사 매체별로 발생하는 편향의 특성을 비교·분석하려 하였다.

분석의 결과 온라인 조사 응답자들의 연령별 2012년 대선 투표행태의 경우, 보수후보 편향을 보인 유선전화조사와 달리 당시 서울시내 투표결과와 매우 유사한 결과를 나타냄으로써 일반적으로 알려진 온라인 조사의 진보편향이 존재한다고 보기 어려웠다. 서울시장 투표결과의 경우, 온라인 조사는 박원순 후보 쪽으로 강한 편향을 보였는데, 이는 대선 투표행태에서 보여준 정치적 성향이나 패널조사 기간 중 정몽준 후보 지지자의 탈락, 혹은 부동산의 박원순 투표 등으로 설명하기는 어려웠다. 그보다는 최초 표집 단계에서부터 온라인 응답자로 섭외하기 힘든 고연령층의 경우 박원순 후보 지지자가 특히 많이 섭외되었고, 5차 조사까지 이어진 자기선택오차가 주 원인으로 보인다. 이러한 경향은 고연령층에서 정몽준 후보 지지자가 많이 섭외된 유선전화조사와 대조를 보였다. 두 조사 모두 실제보다 매우 높은 투표율을 보여 서울시장 선거에 대해 높은 관심을 가진 사람들이 응답자로 섭외되었음을 알 수 있는데, 이러한 자기선택오차의 방향은 흥미롭게도 조사 매체에 따라 반대로 나타났다.

온라인 조사를 이용한 선거결과 예측을 위해서는 지난 대선의 성별·연령대별 투표율을 활용한 예상 투표율을 바탕으로 한 가중치를 선거 1주일 전 실시한 4차 조사의 지지후보에 적용하였다. 정몽준 후보 지지도가 다른 연령층보다 상대적으로 높으면서 뚜렷이 과소표집된 고연령층에게 높은 가중치가 부여되면서 보정 전보다 실제 투표결과에 근접하였으나, 고연령 온라인 응답자는 실제보다 박원순 후보 편향이 심했기 때문에 예측 정확도가 많이 향상되지는 않았다. 이러한 고연령 온라인 응답자의 박원순 후보 편향은 지난 대선 때의 문재인 후보 선호도로도 많이 설명되지 않아, 대선 지지후보를 추가 정보로 활용한 성향점수 매칭보정에서 정확성이 향상되지 않았다. 이처럼 선거전 4차 조사에서 지지후보와 예상 투표율을 활용한 예측과 선거 후 5차 조사에서 투표후보와 실제 투표율을 활용한 보정의 결과는 거의 비슷하여, 대

선을 활용한 성별 및 연령대별 예상 투표율은 유용한 것으로 보였다. 한편 유선전화조사는 보정 후 오히려 정확도가 떨어졌고 다소 과대표집되었던 고연령층의 보수적 성향의 문제점을 드러냈다. 비록 뚜렷하지 않은 보정효과는 아쉽지만, 반대로 보정과정에서 온라인 조사와 유선전화조사의 편향이 단순히 인터넷과 유선전화라는 조사 매체에서 발생하는 포함오차에 의한 것이 아니라, 온라인 조사는 고연령층의 자기선택오차로 인한 편향 그리고 유선전화에서는 고연령층 과대표집에 의한 편향이었음을 각각 밝혔다는 데 의의가 있다.

그렇다면 이러한 분석결과가 향후 온라인 선거여론조사에 갖는 함의는 무엇인가? 첫째, 온라인 여론조사는 근본적으로 확률표집이 불가능한 방식임을 분명히 해야 한다. 인터넷 혹은 이메일 사용가능자는 유권자 모집단을 대표할 수 없고, 확률추출을 위한 모집단 목록을 만들 수 없다. 이러한 한계는 본 연구가 실행한 것과 같은 할당표집으로 보완하기도 하고, 최근에는 사전매칭을 통해 대표성을 향상시키려 노력하기도 한다. 사전매칭은 본 연구처럼 조사 후 준거집단과 매칭하는 것이 아니라 조사 전 응답자 패널에서의 추출 단계에서부터 대표성 있는 준거자료와 매칭하는 방식으로, 조사 후 가중치 사용이나 사후보정 없이 자유롭게 분석을 할 수 있다. 조사회사가 자신들의 응답자 패널에 대한 정보가 많이 쌓일수록 유용한 방법이기 때문에 고려해볼 만하며, 실제로 이러한 사전매칭 방법을 이용하고 있는 폴리메트릭스(Polimetrix)라는 온라인 조사회사의 자료가 미국 주류 학계에서 인정받고 있기도 하다 (Gerber et al. 2010; Wright et al. 2012).

둘째, 할당이건 사전매칭이건 일단 추출해낸 접촉 대상자는 가능하면 응답자로 섭외에 성공해야 한다. 시간과 비용의 압박을 받는 국내 조사의 관행이 취약한 부분이라 생각하며 비확률표집의 문제를 최소화하기 위해서 가장 중요한 단계라 보인다. 본 연구의 온라인 응답자 섭외과정도 할당을 지키기 힘들어 계속 대체표본을 찾게 되었고, 결국 과소표집된 최고령층의 경우에 특정 후보에 대한 편향이 심했다. 그러나 전반적으로는 할당이 잘 지켜져서 응답자의 대선 선택 후보는 당시 서울시의 성별, 연령별 대선 선택 후보의 분포와 크게 다르지 않아 온라인 응답자가 일반적으로 진보편향을 보인다는 우려

는 해당되지 않음을 보였다. 반면 이틀 만에 조사가 이루어졌고 연구자가 조사과정에 관여하지 못했던 유선전화조사의 경우 권역별 할당이 잘 지켜지지 못했고 고연령층이 과대표집되기도 하였다. 다행히 고연령층의 투표율이 높아 편향 폭이 적었으나, 온라인 조사보다 더 뚜렷한 연령별 정치적 편향을 보였으며 보정 후 오히려 정확성이 떨어지기도 하였다. 서울시장 선거를 포함한 이번 2014년 지방선거의 여러 전화여론조사의 정확성을 분석한 연구(김영원·황다솜 2014)는 할당준수의 중요성을 관찰하였으며, 본 연구는 조사설계에 따라 최초 추출한 대상자와 조사를 성사시키는 것의 중요성은 조사방법에 상관없이 매우 크다는 점을 확인하였다.

셋째, 조사의 자기선택오차는 조사 매체에 대한 접근성에 따라 민감할 수 있으며, 온라인 조사의 경우 고연령층의 섭외 및 (패널조사의 경우) 중도탈락 방지에 특별히 신경써야 한다. 본 연구는 이러한 응답자들의 박원순 후보 편향이 적어도 1년 반 전 대통령선거의 투표에서 보인 선택과는 상당히 독립적임을 관찰했다. 본 온라인 조사가 실시되는 과정에서 60대 이상의 경우 일반적인 정치적 성향과 관계없이 박원순 후보를 지지하는 응답자들이 선택 표집된 것으로 보인다. 그렇다면 이러한 체계적인 자기선택오차가 왜 60대 이상에서 유독 심하게 나타났을까? 60대 이상의 경우 인터넷 접근성이 떨어져 접촉이 힘들고, 접촉에 성공하더라도 익숙하지 않은 온라인 조사를 더 꺼리는 경향이 있다. 즉, 고연령층의 경우에는 접촉률과 협조율 모두가 현격히 낮기 때문에, 표본대상이 부족하거나 협조거부에 따른 대체표본 물색이 자주 일어난다. 따라서 섭외과정을 반복하면서 고연령층일수록 자발성이 강한 응답자들이 섭외되며, 이는 서울시장 선거라는 조사주제에 더욱 관심이 있는 사람들로 응답자가 채워짐을 의미한다.

그러나 사실 선거에 대한 관심 자체가 선거결과 예측에 편향을 일으키지는 않는다. 선거결과에 어차피 투표할 만큼 관심이 있는 사람들에게 의해서 결정되기 때문이다. 문제는 고연령층임에도 온라인 조사에 협조할 만큼 선거에 대해 매우 높은 관심이 있는 응답자들이 왜 박원순 후보에 대한 지지와 연관되었는지는 선거마다의 특수한 맥락에 영향을 받기 때문에 정확히 파악하기 힘들다. 다만 한 가지 언급할 만한 점은 고연령층 부동층의 경우 이러한 투표

성향을 세월호 사건 발생에 따른 정권 심판의 욕구⁷⁾와 연관시키기는 힘들다는 점이다. 세월호 사건 직후 조사된 2차 조사에서 고연령층의 박원순 후보 지지도가 정몽준 후보 지지도에 비해 전혀 올라가지 않았다. 이는 여당에 대한 지지를 통해 박근혜 정부를 지켜달라는 호소가 효과를 보였기(가상준 2014) 때문으로도 볼 수 있을 것이다. 오히려 2차 조사와 3차 조사 사이 정몽준 후보 아들의 ‘국민 미개 발언’ 사건이 있었으나, 3차 조사에서는 정몽준 후보의 지지도가 뚜렷하게 증가하였는데, 이는 새누리당 경선을 통해 정식 후보로 선출된 효과로 볼 수 있을 것이다. 또한 정몽준 후보 지지 고연령 응답자들이 실제 선거에서 변심 투표를 한 경향도 발견되지 않았다. 결국 고령층 응답자의 박원순 편향이 발생한 까닭은 조사기간 중의 변화보다는 최초의 선택적 응답자 선별과정의 역할이 큰 것으로 보이며, 따라서 추출과 협조가 어려운 집단일수록 대체표본을 찾기보다는 반복접촉과 차별적으로 높은 조사보상 등을 통해서라도 응답협조를 얻어내는 것이 더욱더 중요하다. 그렇지 못할 경우 조사에 대한 강한 관심편향이 특정 후보에 대한 지지와 연관되어 있는 응답자가 대체된 표본 등을 통해 유입될 가능성이 높아질 수 있기 때문이다. 그리고 그러한 강한 관심이 모집단 평균에 비해 어떤 후보에 얼마나 쏠려 있는지는 선거 전에 미리 예측하기 힘든 요소이다.

그렇다면 온라인 조사는 향후 선거예측에 유용하게 쓰일 전망이 희박한가? 이에 대한 전망은 연령 간 인터넷 이용률 격차의 해소가 중요할 것으로 보이는데, 60대 이상 고연령층의 인터넷 이용률은 2008년 19%에서 매년 꾸준히 증가하여 2013년 26.8%까지 증가하였다(국가통계포털 2014). 특히 2013년을 기준으로 20~30대의 인터넷 이용률은 99%, 50대의 경우에도 80%가 넘는 것으로 나타나(국가통계포털 2014), 다른 연령대의 인터넷 이용률이 이미 포화상태에 있는 것을 감안하면 고연령층에서의 꾸준한 인터넷 이용률의

7) 청해진해운 소속 여객선 ‘세월호’가 2014년 4월 16일 전라남도 진도군 인근 해상에서 침몰하여 수백 명의 사상자를 낸 대형 사고이다. 특히 수학여행을 위해 단체로 탑승한 안산 단원고등학교 2학년 학생들의 다수가 실종 및 희생되어 많은 국민들을 비통에 빠뜨렸다. 이후, 사고발생 직후 정부의 적절하지 못한 대처와 혼선 등으로 인해 정부책임론이 대두되었다.

증가는 연령 간 격차를 감소시킬 것이다. 실제로 정보격차지수 및 실태조사 결과(한국정보화진흥원 2014)를 살펴보면, 60세 이상 노년층의 전체국민 대비 종합 정보화지수는 2007년 47.6%에서 2014년 56.9%로 꾸준히 증가하고 있음을 확인할 수 있었다. 또한 고연령층의 스마트폰 사용률 추이는 2012년 1월 13%에서 2013년 11월 27%까지 빠르게 증가하였으며(한국갤럽 2013), 스마트폰을 활용한 조사는 향후 고연령층에 대한 접근성을 높이는 방안이 될 수 있다. 55세 이상 스마트폰 사용자들을 대상으로 실시한 조사결과를 살펴보면, 최근 6개월 간 인터넷을 이용하기 위해 스마트폰을 사용한 시간이 이전보다 증가하였다고 응답한 인원이 2011년에는 12%였으나, 2013년에는 38%로 나타나(구글 2013) 이러한 가능성은 증대될 것이다.

이를 종합해 볼 때, 정보화 격차의 해소를 통해 포함오차의 문제는 자연스럽게 해결이 가능할 것으로 볼 수 있으나, 자기선택오차에 대한 해결이 더욱 더 중요해질 것이다. 따라서 온라인 조사뿐 아니라 모바일 조사에서 고연령층의 참여 혹은 표집을 어떻게 향상시킬 것인지에 대한 연구와 관심이 필요할 것으로 보인다. 이러한 노력이 지속된다면, 60대 이상을 제외하고 온라인 조사결과의 보정효과가 드러난 것을 고려할 때 다음 대통령선거(2017년 12월) 투표행태를 이용한 7회 지방선거(2018년) 결과예측을 시도해 볼 만할 것이다.

본 연구만으로 온라인 선거여론조사의 가능성에 대해 긍정적 혹은 부정적 평가를 내리기는 힘들다. 비확률표집과 자기선택오차의 문제, 가중치를 이용한 보정의 한계 등을 극복하지 못한 한편, 할당표집에 의한 정치적 성향 편향의 최소화나 성별 및 연령별 가중치 보정의 예측 유효성도 확인하였다. 또한 연령 간 온라인 및 모바일 보급 격차의 감소와 조사 전·후 매칭기법을 통해 비확률표집에 따른 편의를 줄일 수 있는 가능성은 여전히 열려 있다. 이러한 가능성들을 앞당기기 위한 중요한 요소로서 연구자들이 연구과정에서 깨달은 점은, 온라인 조사의 정확성을 누적적으로 향상시키기 위해서는 연구자가 조사과정 관련 자료를 요구할 권리를 사전에 조사회사와 주의 깊게 합의하고 과정을 면밀히 관찰해야 한다는 것이다. 빠른 속도로 진행되는 온라인 조사 혹은 유선전화조사 과정 속에서 낮은 조사단가의 문제점(한신잡 2012)이 겹

치면, 연구자는 때로 조사과정을 추적할 기회를 놓치기도 하고 조사회사는 연구자가 놓친 기회를 만회할 수 있는 정보의 제공에 비협조적이 되기도 한다.⁸⁾ 온라인 조사 응답자의 응답 행태와 답변 내용은 조사회사에 누적적으로 기록된다. 이러한 기록을 바탕으로 온라인 조사의 타당성을 누적적으로 향상시키는 것은 가능하다고 생각하며, 조사회사와 연구자 간 협업을 통해 앞당겨질 수 있다고 생각한다.

참고문헌

- 가상준. 2014. “6.4지방선거 결과 분석.” 《지방행정》 63(729): 16-19.
- 강남준·백영민. 2004. “대안적 여론조사의 표본편파(Sample Bias) 문제점과 가중치를 사용한 보정방법 : 성향점수 가중과 반복비례가중을 중심으로.” 《언론정보연구》 41(2): 43-78.
- 강신구. 2014. “지방선거: 지방정부의 구성을 위한 것인가? 중앙정치의 대리전인가?” 《OUGHTOPIA》 29(2): 65-95.
- 강정한·강예원·임정재. 2014. “응답자의 성격특성과 성향점수 매칭기법을 활용한 모바일조사 향상방안 연구.” 《조사연구》 15(2): 71-103.
- 강현철·한상태·김지연·정용찬·허명희. 2008. “RDD 전화조사와 주요결과.” 《조사연구》 9(1): 1-22.
- 김영원·배예영. 2010. “인터넷 선거여론조사 가중치보정을 위한 성향점수의 활용.” 《한국통계학회논문집》 17(1): 55-66.
- 김영원·황다솜. 2014. “2014년 지방선거 여론조사 전화조사방법에 따른 예측오차 및 편향.” 《조사연구》 15(4): 1-32.
- 김원용·이홍철. 2003. “웹조사의 모집단대표성 확보를 위한 성향가중모형의 적

8) 본 연구의 경우 온라인 조사의 연령별 표본대체율 등을 알 수 없었고 유선전화조사의 경우 응답률조차 제공받지 못했으며, 이러한 점은 편의의 원인을 파악하는데 걸림돌이 되었다. 온라인 응답자 패널 구성 및 응답자 표집방법 등은 회사의 영업상 비공개를 원칙으로 하고 있었다.

- 합성 검증.” 《방송연구》 56: 143-166.
- 김장수. 2007. “선거연구와 패널 여론조사 : 2006년 지방선거를 중심으로.” 《조사연구》 8(1): 81-104.
- 류재성. 2011. “지방 선거에서의 유권자의 특성: 총선 및 대선과의 비교 분석.” 《한국시민윤리학회보》 24(1): 47-73.
- 박경미. 2014. “선거별 투표율 결정 요인: 서울시 선거구의 집합자료 분석(1987~2010년).” 《한국정당학회보》 13(1): 95-122.
- 이계오·장덕현. 2009. “인터넷 선거조사에서 성향가중모형 적용사례.” 《조사연구》 10(3): 21-36.
- 이계오·홍영택·백정민. 2012. “인터넷 조사와 전화조사의 성향가중 결합추정 : 2007 대통령선거 예측조사 사례연구.” 《통계연구》 17(1): 1-18.
- 이세희·이종균. 2013. “SNS 사용에 따른 한국 정당지지의 변화 : 한나라당, 민주당, 민주노동당을 중심으로.” 《사회과학연구》 21(1): 186-222.
- 이지영·김진교. 2006. “온라인 패널 조사와 오프라인 대인면접 조사 간 데이터 품질 비교 : 신뢰도와 타당도를 중심으로.” 《마케팅연구》 21(4): 209-231.
- 이현출. 2004. “한국 유권자의 정당지지 구조와 안정성.” 《대한정치학회보》 12(2): 129-155.
- 장덕진. 2011. “트위터 공간의 한국정치 : 정치인 네트워크와 유권자 네트워크.” 《언론정보연구》 48(2): 80-107.
- 장상철. 2008. “누가 진보정당을 지지하는가 : 민주노동당과 진보신당 지지자들의 정당지지 결정요인.” 《한국사회학회 사회학대회 논문집》 12: 384-393.
- 정용찬. 2014. “유선전화 없는 가구의 방송매체 이용행태.” 《KISI STAT Report》 14-02-01. 정보통신정책연구원.
- 조성겸·박아현·허명희. 2011. “생활시간 조사를 이용한 가중치 부여방법 : 인터넷 조사에 대한 적용 가능성 검토.” 《조사연구》 12(2): 145-157.
- 조성겸·주영수·조은희. 2005. “인터넷 여론조사의 정확도 관련요인.” 《조사연구》 6(2): 51-74.
- 중앙선거관리위원회. 2014. 《제6회 전국동시지방선거 투표율분석》. 중앙선거관리위원회.
- 중앙선거여론조사공정심의위원회. 2014a. 《제6회 전국동시지방선거 선거여론조

- 사 심의백서》. 중앙선거여론조사공정심의위원회.
- 한국갤럽. 2013. “스마트폰 사용에 대한 조사(2012~2013년 월별 사용자 추이 포함).” 《한국갤럽 Gallup Report》 2013/13/10.
- 한국정보화진흥원. 2014. 《2014 정보격차지수 및 실태조사》. 한국정보화진흥원.
- 한신갑. 2012. “혼합식 조사와 웹패널의 (옅은) 빛과 (짙은) 그늘.” 《조사연구》 13(3): 1-31.
- 허명희 · 김지연 · 이양훈. 2009. “패널 이탈과 조건화의 영향 :KBS-MBMR의 2007년 대통령선거 패널조사 사례.”《조사연구》 10(1): 73-89.
- 허명희 · 조성겸. 2010. “자원자 패널에 의한 인터넷 조사의 성향조정 가중화.” 《조사연구》 11(2): 1-28.
- 홍명신. 2009. “대통령 후보 선출 과정의 변화와 여론조사의 역할 : 1987-2007.” 《정치커뮤니케이션 연구》 14: 277-315.
- 황아란. 2014. “2014년 지방선거의 특징과 변화.” 《21세기정치학회보》 24(3): 319-341.
- Couper, M.P. 2000. “Web Surveys: A Review of Issues and Approaches.” *The Public Opinion Quarterly* 64(4): 464-494.
- Dillman, D.A., J.D. Smyth, and L.M. Christian. 2008. *Internet, Mail and Mixed-Mode Surveys*(3rd ed.). Ney Jersey: John Wiley & Sons, INC.
- Gerber, A.S., G.A. Huber, D. Doherty, C.M. Dowling, and S.E. Ha. 2010. “Personality and Political Attitudes: Relationships across Issue Domains and Political Contexts.” *American Political Science Review* 104(1): 111-133.
- Guo, S. and M.W. Fraser. 2010. *Propensity Score Analysis - Statistical Methods and Applications*. California: Sage Publication, Inc.
- Pischke, J.S. and J.D. Angrist. 2009. *Mostly Harmless Econometrics*. Ney Jersey: Princeton University Press.
- Taylor, H. 2000. “Does Internet Research Work? Comparing Online Survey Results with Telephone Survey.” *International Journal of Market Research* 42(1): 51-63.
- Taylor, H., J. Bremer, C. Overmeyer., J.W. Siegel, and G. Terhanian. 2001. “Using Internet Polling to Forecast the 2000 Elections.” *Marketing*

Research 13(1): 26-30.

Wright, M., J. Citrin, and J. Wand. 2012. "Alternative Measures of American National Identity: Implications for the Civic-Ethnic Distinction." *Political Psychology* 33(4): 469-482.

구글(google). 2013. <http://think.withgoogle.com/mobileplanet/ko/>: Our Mobile Planet 연도별 연령별 모바일 행동 비교 검색.

국가통계포털. 2014. <http://kosis.kr/>: 인터넷이용률실태조사, 연령별인터넷이용률 검색.

중앙선거여론조사공정심의위원회. 2014b. <https://www.nesdc.go.kr>: 지방선거 여론조사 결과 및 득표율 비교정리표-조사기간별, 제6회 지방선거 여론조사 결과분석 자료 게시.

<접수 2014/10/30, 수정 2015/2/4, 게재확정 2015/4/29>

<부록 1> 성별 · 연령대별 시장선거 예상투표자 분포 비율 산출

구분		20대	30대	40대	50대	60대 이상	계
서울특별시 인구수	남	730,306	875,617	870,535	777,808	775,998	4,996,391
	여	731,747	862,435	869,583	822,651	943,986	5,132,832
	계	1,462,053	1,738,052	1,740,118	1,600,459	1,719,984	10,129,223
2012 년 대선투표율	남	70.00	68.00	73.50	81.00	85.90	-
	여	72.50	71.70	77.70	82.90	77.10	-
서울시장선거 예상투표자 수	남	511,214	595,420	639,843	630,024	666,582	3,043,084
	여	530,517	618,366	675,666	681,978	727,813	3,234,339
	계	1,041,731	1,213,785	1,315,509	1,312,002	1,394,395	6,277,423
서울시장선거 예상투표자 분포(%)	남	8.14	9.49	10.19	10.04	10.62	48.48
	여	8.45	9.85	10.76	10.86	11.59	51.52
	계	16.59	19.34	20.96	20.90	22.21	100.00
서울시장선거 전화조사 투표자 분포(%)	남	6.13	10.38	9.43	8.25	10.85	45.05
	여	9.67	7.31	9.67	11.08	17.22	54.95
	계	15.80	17.69	19.10	19.34	28.07	100.00
서울시장선거 온라인조사 투표자 분포(%)	남	7.37	14.95	18.21	14.63	4.11	59.26
	여	9.37	10.84	9.26	8.42	2.84	40.74
	계	16.74	25.79	27.47	23.05	6.95	100.00

주1: 2012년 대선투표율의 경우 19세는 고려하지 않고, 40대 이상은 자료 한계 상 전국투표자 비율 적용

주2: 서울시장선거 예상투표자수의 경우 2014년 현재 서울시 인구분포와 2012년 성별 · 연령별 대선 투표자 비율을 결합하여 예상 투표자수 계산

<부록 2> 2012년 대선 후보자별 득표 예상비율 분포(%)

구분		20대	30대	40대	50대	60대 이상
박근혜 후보	남	2.74	3.14	4.50	6.27	7.68
	여	2.85	3.26	4.75	6.79	8.38
문재인 후보	남	5.36	6.31	5.67	3.75	2.92
	여	5.56	6.55	5.98	4.06	3.19

주: 기타후보 생략

<부록 3> 성별 · 연령대별 서울시장선거 투표자 분포 비율

구분	20대	30대	40대	50대	60대 이상	전체
남	8.12	9.00	9.64	9.89	12.05	48.69
여	8.66	9.69	9.92	10.39	12.64	51.31
계	16.78	18.69	19.56	20.28	24.69	100.00

출처: 중앙선거관리위원회(2014)