

기획특집

2006년 지방선거 출구조사 현황 및 예측오차*

An Overview of Exit Polls for the 2006 Local Elections

김영원** · 김지현***

Young-Won Kim · Ji-Hyeon Kim

본 연구에서는 2006년 5월에 KBS/SBS에서 공동으로 실시한 제4회 전국동시지방선거 출구조사 과정을 소개하는 동시에 출구조사 자료를 통해 투표자의 특성(성별 및 연령대)에 따른 투표성향과 시간대별 투표자 특성을 정리하고, 출구조사 선거구별 득표율 예측치와 실제 개표결과를 비교하여 출구조사의 예측오차를 살펴본다. 이를 위해 이번 출구조사에서 적용된 전반적인 표본추출방법 및 실사과정을 정리하고, 출구조사에서 발생한 전체 오차를 편향과 표본추출오차로 분리하여 오차의 특성을 분석한다. 편향 발생원인 중 하나로 볼 수 있는 표본의 대표성을 검토하기 위해 출구조사 표본을 선관위에서 집계한 투표자 투표율 분석결과와 비교하여 평가한다. 아울러 이번 지방선거 출구조사에서 적용된 '층화 후 정렬계통추출법'에 따른 분산추정법을 적용하여 출구조사의 정확성을 표본추출오차 관점에서 살펴본다.

주제어: 2006년 지방선거, 선거예측, 출구조사, 투표소 표본추출

This article attempts to provide an overview of the exit polls for the 2006 local elections in Korea. The sampling method, sampling error, non-response rate, and prediction error of the exit polls are reviewed. Also, we explore the fact that the propensity to vote varies according to age and gender of voters. In terms of age and gender, the representativeness of the sample is investigated by comparing to the data released by the National Election Commission. Through this empirical research, we show that the exit poll samples are unbalanced in terms of age and this unbalance may be one of the causes of bias occurred in the prediction of the 2006 local election results. The design effects

* 본 연구는 숙명여자대학교 2006년도 교내연구비 지원에 의해 수행되었음.

** 교신저자(corresponding author): 숙명여자대학교 통계학과 교수 김영원.

E-mail: ywkim@sookmyung.ac.kr

*** 숙명여자대학교 통계학과 대학원생.

of the sample design implemented for the exit polls are also examined.

Key words: 2006 local elections, vote predictions, exit poll, voting place sampling

I. 서론

방송사 출구조사(exit poll)는 실제 참값과 표본조사를 통해 얻어진 추정값 간의 차이를 확인할 수 있는 표본조사이다. 개표가 진행되기 이전에 잠정적인 선거결과를 가늠해 볼 수 있다는 점은 널리 알려진 출구조사의 기능이다. 또한, 선거의 특성상 개표 후에도 확인할 수 없는 투표자 개인의 특성(성별 및 연령)에 따른 투표성향을 분석할 수 있다는 점에서도 큰 의미가 있다.

우리나라에서는 1999년 3.30재보선에서 최초로 출구조사가 도입된 이후, 1999년 6.3재보선, 2000년 16대 총선, 2002년 3회 지방선거, 2002년 8.8재보선, 2002년 16대 대선, 2004년 17대 총선, 그리고 이번 2006년 4회 지방선거까지 총 8번의 출구조사가 이루어졌다. 이 중 16대와 17대 총선에서 출구조사의 정확도가 떨어지는 것으로 나타났고, 다른 선거에서는 비교적 정확한 예측이 이루어졌다(조성겸·김지연 2004).

우리나라 출구조사와 관련해서 류제복(2003)은 우리나라에서 사용되는 출구조사의 문제점과 전반적인 개선방향을 정리한 바 있고, 김정훈(2003)은 16대 대선 출구조사와 관련해 출구조사의 방법론적 특성과 예측력을 좌우하는 요인에 대해 출구조사의 기획과 실행과정을 중심으로 상세히 정리하고 있다. 한편 이준웅(2004)은 17대 총선 예측조사에서 발생한 예측오차와 편향을 조사거절 및 응답불성실 관점에서 구체적으로 평가하고, 이에 대한 대처방안을 조사방법론 차원과 조사 외적인 요인의 차원으로 구분하여 논의하고 있다. 특히 출구조사에서는 어떤 방식의 투표소 추출방법을 적용하는지에 따라 예측결과에 많은 영향을 주게 된다. 홍내리·허명희(2001)는 16대 총선 예측조사와 관련해 전화조사와 출구조사 현황을 정리하고, 표본 투표소 수의 결정과 판단표본추출에 따른 투표소 추출방법 등과 같은 선거예측조사의 문제점을 지적하고 이

에 대한 개선방안을 제시하고 있으며, 조성겸·김지연(2004)은 기존에 흔히 사용했던 ‘대표구 추출법’과 ‘계통추출법’의 효율성을 비교함으로써 투표소 추출법의 개선 가능성을 보여 주었다. 김영원·임윤희(2005)는 2004년 17대 총선거 표결결과를 이용하여 계통추출법을 근간으로 한 다양한 확률추출법에 의한 투표소 추출방법의 효율성과 추출방법에 따른 표본추출오차를 제시하고 있다.

최근 출구조사 투표소 추출방법에 대한 연구가 주요 관심대상이 되고 있지만, 실제 출구조사를 통한 전반적인 선거예측 과정을 살펴보면 단순히 투표소 추출방법뿐만 아니라 조사원 관리 및 현장 실사방법, 조사거절과 거짓 응답, 표본 투표소에서 투표자 선정 방법, 무응답 및 가중치 적용방법 등 다양한 원인에 의해 예측오차가 발생하고 있다는 것을 알 수 있다. 따라서 출구조사의 정확성 제고를 위한 방안을 모색하기 위해서는 먼저 출구조사의 특성을 정확히 이해하고 이를 기반으로 현실적인 한계를 감안한 출구조사 개선방안이 모색될 필요가 있다.

이런 관점에서 본 연구에서는 2006년 제4회 지방선거를 위해 KBS/SBS에서 공동으로 실시한 6개 광역단체장 선거구에 대한 출구조사 수행과정을 소개하고, 이들 지역에서 얻어진 출구조사 자료분석을 통해 투표자의 성별, 연령대별 특성에 따른 투표성향과 투표 시간대별 투표자 성향을 살펴본다. 출구조사의 정확성을 살펴보기 위해 출구조사에서 얻어진 예측결과와 실제 개표결과와의 비교를 통해 출구조사의 오차가 어떤 패턴을 보이고 있는지 정리하고, 출구조사에서의 오차를 편향(bias)과 표본추출오차(sampling error)로 구분하여 오차의 발생원인 및 수준을 평가해 본다. 출구조사에서 발생한 편향의 원인을 알아보기 위해 중앙선거관리위원회(이하‘선관위’)에서 선거 후에 작성한 투표율 분석자료를 토대로 성별/연령별 구성비율 관점에서 출구조사 표본의 대표성을 평가한다. 아울러 출구조사의 표본설계를 반영한 분산추정법을 통해 표본추출오차를 산출한다.

참고로 2006년 지방선거 KBS/SBS 공동예측조사의 기획을 비롯해 표본추출방법, 전화조사방법, 현장실사 및 관리, 예측치 및 오차 산출 등에 대한 자문과 검증에 대해 방송사와 조사회사 관련자 및 자문교수단이 수차례 걸친 회의를 가졌으며, 본 연구내용 중에는 관련 자문회의에서 논의된 사항이 일부 포

함되어 있다는 점을 밝혀 둔다¹⁾.

II. 2006년 지방선거 출구조사 개요

1. 투표소 추출방법

우리나라에 선거예측을 위한 출구조사가 도입된 이후 어떤 방식의 투표소 추출방법이 효율적인 방법인가에 대해 끊임없는 논란이 있어 왔다. 최근 우리나라에서는 출구조사 투표소 추출방법으로 대체적으로 다음 두 가지 방법이 사용되고 있다.

첫 번째 방법은 대표구 추출법(bellwether polling place sampling)으로, 과거부터 대부분의 조사회사에서 가장 많이 사용하였던 방법이다. 이 방법은 과거 선거결과를 기준으로 선거구 전체의 투표결과를 가장 잘 대표할 수 있을 것으로 판단되는 투표소를 표본으로 추출하는 방법으로서, 경험적인 판단에 의존한 비확률표본추출법이기에 때문에 이론적으로 표본추출오차 등을 산출할 수 없다는 한계를 갖고 있다. 또한 과거 선거와 비교하여 투표소의 변동이 심하게 발생하는 경우 실제 적용이 어렵다는 적용상의 문제도 갖고 있다. 두 번째 방법은 층화(stratification) 및 계통추출법(systematic sampling)을 근간으로 한 확률추출법을 적용하는 것이다. 최근 우리나라에서 대표구 추출법의 대안으로 이에 대한 연구가 이루어져 왔다. 조성겸·김지연(2004)과 김영원·엄윤희(2005)는 이와 같은 확률추출법의 타당성 및 효율성에 대한 연구결과를 제시한 바 있다.

광역자치단체장 선거예측을 위해 KBS/SBS에서 실시한 2006년 지방선거 출구조사에서는 조사 준비과정에서 다양한 투표소 추출방법이 과거 선거자료를 이용한 시뮬레이션 등을 통해 충분히 검토되었으며, 종합검토 결과 '시·군·구 층화 후 정렬계통추출법'이 가장 효율적인 방안인 것으로 판단되어 표본 투표소 추출을 위해 다음과 같은 투표소 추출법을 적용했다.

1) 본 연구가 가능하도록 2006년 지방선거 KBS/SBS 공동예측조사 출구조사 자료의 사용을 허락해 준 KBS와 SBS 방송사 및 조사회사 관계자들에게 감사드린다.

먼저, 출구조사 대상 선거구(광역시·도)별로 시·군·구를 지역 인접성을 기준으로 총화하고, 층 내의 투표소를 과거 선거에서의 특정 정당 지지율을 기준으로 정렬했다. 예를 들어, 2006년 지방선거 출구조사 투표소 추출을 위해 각 투표소의 2002년 지방선거에서의 특정 정당 후보자의 지지율을 기준으로 층 내 투표소를 내림차순으로 정렬하는 방식이다. 각 선거구별로 추출 투표소의 수에 따라 추출간격을 결정하고, 이를 기초로 투표소를 계통추출했다. 이렇게 추출된 표본 투표소를 대상으로 사전 답사과정을 통해 출구의 개수 등 각 투표소의 실사 환경 등을 고려하여, 현실적으로 출구조사 실사관리에 부적합한 투표소가 있을 경우 정렬된 순서상 바로 인접한 투표소로 대체하여 최종 출구조사 투표소를 확정했다.

표본으로 추출된 투표소에서는 선거당일 투표시간 순으로 투표자를 5명 간격으로 계통추출하고, 조사원들은 표본으로 선정된 투표자(목표 응답자)를 대상으로 지지한 후보와 함께 응답자의 성별과 연령대를 조사했다.

2. 투표소 표본추출 결과

2006년 전국동시지방선거는 광역단체장, 기초단체장, 광역의회의원, 기초의회의원에 대한 선거였으며, 출구조사는 이 중 광역단체장에 한하여 이루어졌다. 광역단체장은 전국 16개 광역시·도에서 선출되며, 출구조사는 이 중 경합 지역에 해당하는 서울, 대전, 제주, 경기, 광주, 충남 등 6개 광역선거구에 대해 이루어졌다. 출구조사에서 각 선거구별로 구성한 층의 수와 추출된 투표소의 수는 <표 1>과 같으며, 추출률은 모집단 대비 표본 투표소의 수를 백분율로 나타낸 것이다.

<표 1> 시·도별 층의 개수 및 투표소의 개수

구분		서울	대전	제주	경기	광주	충남
층의 개수		8	5	6	12	5	7
투표소 개수	모집단	2,201	326	228	2,511	328	723
	표본	64	40	46	71	29	50
	추출률(%)	2.91	12.27	20.18	2.83	8.84	6.92

지역별로 시·군·구 상황이 많이 다르고 전체 투표소 수도 차이가 크기 때문에 투표소 추출과정에서 층을 나누는 기준은 선거구별로 다소 차이가 있다. 대전, 광주 는 시·군·구 수가 많지 않기 때문에 각 시·군·구를 한 개의 층으로 구분했다. 제주의 경우 지역 특성상 북제주군과 남제주군을 동·서로 나누어 각각 두 개의 층으로 구분하고 나머지 시를 각각 하나의 층으로 구성했다. 상대적으로 시·군·구의 수가 많은 서울, 경기, 충남의 경우는 크기가 큰 시·군·구는 한 개의 층으로 구분하고, 그렇지 않을 경우 지리적 인접성과 시·군·구 특성을 고려해 2개 이상의 시·군·구를 묶어 하나의 층으로 구성했다.

투표소 추출을 위해 우선 각 층에서 2002년 지방선거의 특정 정당 후보의 지지율을 기준으로 전체 투표소를 정렬하고 표본 투표소를 계통추출했다. 단, 제주지역의 경우 2002년 이후 보궐선거가 있었기 때문에 보궐선거에서의 지지율을 사용해 투표소를 정렬했다.

3. 정렬 계통추출법의 효율성 검토

투표소 표본추출을 위해 정렬 계통추출법을 적용하는 경우 정렬기준으로 사용된 과거 선거결과와 이번 선거에서 나타난 실제 투표결과가 투표소별로 상관관계가 높을수록 적용된 계통추출법의 효율성이 높아진다.

출구조사에서 적용한 계통추출법의 효과를 살펴보기 위해 이번 선거 당선자의 투표소별 지지율과 정렬기준으로 사용된 2002년 지방선거에서 같은 정당 후보자의 투표소별 지지율에 대한 상관계수를 구하면 <표 2>와 같다. 선거구별로 다소 차이가 있지만 상관관계가 가장 낮은 광주의 경우가 0.60으로 나타나, 전체적으로 상관계수가 상당히 높은 것을 볼 수 있다. 이를 통해 이번 출구조사에서 투표소를 계통추출하는 과정에서 먼저 과거 선거결과를 이용해 투

<표 2> 정렬기준자료와 개표결과 상관계수

지역	서울	대전	경기	광주	충남
상관계수	0.71	0.67	0.66	0.60	0.71

표소를 정렬한 후 계통추출법을 적용한 것이 상당히 효과적이었음을 확인할 수 있다. 여기서 제주의 경우는 정렬기준으로 사용된 보궐선거 자료를 선관위에서 공개하고 있지 않아 편의상 분석대상에서 제외했다.

4. 출구조사 무응답 현황

KBS/SBS 방송사의 출구조사는 2개 조사회사에서 수행했다. 한 조사회사(이하 A사)는 서울, 대전, 제주 지역을, 다른 조사회사(이하 B사)는 경기, 광주, 충남 지역을 담당했다. 실제 2개 조사회사에 의해 출구조사가 수행되었지만 투표소 추출방법 등 핵심적인 사항들은 자문교수단과의 심층적인 논의를 거쳐 결정된 방식에 따랐기 때문에 실제 출구조사 수행방법에 있어서 두 조사회사간에 큰 차이가 없다고 볼 수 있다. 하지만 현장 실사에 있어서는 조사회사별로 그 동안 사용하던 방법에 있어서 차이가 있고, 기존의 방법을 무리해서 변경하는 것은 조사회사 입장에서는 상당히 큰 부담이 될 수 있기 때문에 조사원 교육 및 현장 실사관리는 각 조사회사에서 사용하던 기존의 방법을 그대로 사용하기로 했다. 이에 따라 조사회사별로 실사규칙상 몇 가지 차이가 있다.

두 조사회사의 실사진행상의 차이 때문에 다음에 제시된 무응답률을 지역간에 비교하는 데 있어서 유의할 필요가 있다. 두 조사회사 모두 자신이 지지한 후보자에 대한 응답은 투표자가 직접 조사표에 기록한 후 모의 투표함에 넣도록 하여 최대한 응답자의 비밀을 보장해 주었다. 하지만 서울 등을 담당한 A사의 경우는 투표자를 대상으로 성별과 연령대를 면접원이 묻고 이를 면접원이 기입한 후 조사표를 건네주는 방식을 사용한 반면, 경기 등을 담당한 B사의 경우는 투표자 스스로가 지지 후보자뿐만 아니라 자신의 성별과 연령대를 조사표에 직접 기입하는 방식을 사용했다. 이에 따라 B사의 경우 조사에 참여한 투표자에 있어서도 성별과 연령대 항목에 무응답이 생길 수 있고, 상대적으로 다소 응답 부담이 더 클 수 있다. 또 다른 차이점은 만약 목표 응답자인 5번째 투표자를 놓쳤을 경우 대처방법에 관한 것이다. A사의 경우는 놓친 목표 응답자를 무시하고 그 다음 5번째 투표자가 조사대상이 되는 목표 응답자로 간주했다. 하지만 B사의 경우에는 놓친 목표 응답자를 무응답자로 분류했다. 따

라서 B사가 실사를 수행한 경기, 광주, 충남의 경우 무응답률이 상대적으로 높게 나타날 수밖에 없다는 점에 유의하기 바란다.

출구조사에서 무응답에는 조사거절과 조사에 참여했지만 조사표에 지지 후보자를 표기하지 않는 경우가 모두 포함되며, 실제 조사에서 무응답의 대부분은 조사거절 때문에 발생한다. 예를 들어, 이번 지방선거에서 경기도의 경우 선거예측에 활용 가능한 성별과 연령대가 파악된 목표 응답자 22,359명 중 무응답자는 2,648명이었고, 이들 중 조사에 참여하였지만 지지 후보자를 표기하지 않은 무응답자는 소수이고, 대부분은 조사거절자에 해당했다. 실사과정에서 조사거절자에 대해서는 면접원들이 성별과 연령대를 관측하여 조사표에 기입하여 투표함에 넣는 방식으로 조사가 진행된다. 출구조사에서 조사거절에 의한 무응답은 흔히 연령대가 높을수록 더 빈번히 발생하는 것으로 알려져 있다. <표 3>은 이번 2006년 지방선거 출구조사가 수행된 6개 선거구들에 대한 연령대별 무응답률을 정리한 것이다.

<표 3> 선거구별 무응답률

(단위: 명, %)

구분		서울			대전			제주		
		전체	남	여	전체	남	여	전체	남	여
연령별	20대	4.06	3.63	4.34	4.76	3.24	6.13	4.77	3.98	5.45
	30대	4.69	4.57	4.81	5.32	5.93	4.73	6.40	6.73	6.06
	40대	5.48	5.21	5.73	6.95	6.08	7.84	8.70	8.23	9.20
	50대	7.42	6.82	8.01	9.62	9.71	9.52	10.54	9.55	11.64
	60대 이상	12.18	8.20	15.94	15.88	13.49	18.23	16.15	13.05	18.59
전체	무응답수	1555	622	933	1130	518	612	1082	464	618
	무응답률	7.23	6.04	8.33	8.99	8.21	9.77	10.22	8.95	11.44
구분		경기			광주			충남		
		전체	남	여	전체	남	여	전체	남	여
연령별	20대	6.08	4.55	7.22	6.00	5.41	6.53	8.20	6.50	9.71
	30대	7.67	6.92	8.36	6.86	5.47	8.13	11.45	9.88	12.93
	40대	8.55	7.51	9.79	8.59	6.76	10.43	18.13	15.19	21.32
	50대	13.81	13.63	14.21	14.51	13.38	15.92	24.40	22.31	26.99
	60대 이상	22.08	18.19	26.38	22.97	19.47	26.81	38.20	30.88	45.81
전체	무응답수	2648	1191	1457	852	360	492	3129	1298	1831
	무응답률	11.84	10.52	13.20	12.44	10.65	14.18	23.75	19.89	27.55

각 선거구별 무응답률을 살펴보면 모든 지역에서 연령대가 높을수록 무응답률이 뚜렷하게 증가하는 것을 확인할 수 있으며, 특히 60대 이상의 무응답률은 20대 무응답률의 3배를 넘고 있다. 또한, 무응답률이 가장 높은 충남지역의 경우 무응답률이 가장 낮은 서울지역의 3배가 넘는 등 선거구에 따라 무응답률 차이가 크다.

III. 출구조사 응답자 투표성향

1. 성별·연령대별 투표 성향

출구조사에서는 투표자를 대상으로 자신이 지지한 후보자뿐만 아니라 응답자의 성별과 연령대를 조사하고, 조사된 성별, 연령대에 관한 정보는 출구조사 예측치를 산출하는 데 중요한 변수로 활용되고 있다. 따라서 출구조사 자료를 이용하면 실제 개표자료에서는 확인할 수 없는 후보자들에 대한 투표자의 성별·연령대별 지지율을 파악하는 것이 가능하다. 투표자의 성별, 연령대별로 각 후보자에 대한 지지율에 어떤 차이가 있는지 알아보기 위해 선거구별로 성별, 연령대에 따른 1위(당선자)와 2위 후보자의 단순 지지율을 살펴보면 <표 4>~<표 9>과 같다.

<표 4> 성별, 연령대에 따른 후보자 지지율(서울) (단위: %)

연령대	한나라당 오세훈			열린우리당 강금실		
	전체	남	여	전체	남	여
20대	54.51	59.61	51.01	35.47	29.62	39.47
30대	46.51	44.50	48.45	41.99	40.21	43.71
40대	62.34	61.46	63.18	26.72	25.68	27.70
50대	71.14	69.44	72.84	17.28	17.20	17.36
60대 이상	75.06	73.79	76.37	15.18	15.06	15.31
전체	63.28	62.90	63.63	25.93	24.32	27.45

〈표 5〉 성별, 연령대에 따른 후보자 지지율(대전) (단위: %)

연령대	한나라당 박성효			열린우리당 엄홍철		
	전체	남	여	전체	남	여
20대	35.39	34.55	36.17	48.46	48.71	48.23
30대	30.27	29.61	30.89	51.81	52.44	51.22
40대	46.28	47.89	44.60	41.70	40.23	43.23
50대	52.27	53.31	51.10	36.06	35.39	36.81
60대 이상	50.40	54.04	46.63	37.73	34.82	40.73
전체	44.10	45.56	42.60	42.38	41.19	43.61

〈표 6〉 성별, 연령대에 따른 후보자 지지율(제주) (단위: %)

연령대	무소속 김태환			한나라당 현명관		
	전체	남	여	전체	남	여
20대	40.38	40.24	40.51	42.30	43.79	41.02
30대	45.17	46.15	44.15	37.84	37.60	38.10
40대	43.73	43.61	43.86	43.31	43.07	43.57
50대	40.3	40.81	39.88	47.02	46.53	47.58
60대 이상	41.61	41.00	42.13	43.02	45.83	40.65
전체	42.49	42.64	42.35	42.61	43.27	41.95

〈표 7〉 성별, 연령대에 따른 후보자 지지율(경기) (단위: %)

연령대	한나라당 김문수			열린우리당 진대제		
	전체	남	여	전체	남	여
20대	50.37	47.86	52.33	39.76	41.71	38.22
30대	43.39	42.05	44.61	45.23	46.84	43.81
40대	62.26	61.10	63.56	29.86	30.20	29.44
50대	72.60	71.49	73.86	21.01	22.07	19.73
60대 이상	75.35	76.34	74.25	18.85	18.78	18.90
전체	60.77	60.60	60.88	30.96	31.17	30.78

〈표 8〉 성별, 연령대에 따른 후보자 지지율(광주) (단위: %)

연령대	민주당 박광태			열린우리당 조영택		
	전체	남	여	전체	남	여
20대	47.37	45.48	49.15	36.95	37.86	36.02
30대	41.89	40.16	43.51	35.42	35.00	35.81
40대	55.40	55.95	55.02	31.77	27.63	35.95
50대	64.88	62.99	66.92	24.93	25.13	24.86
60대 이상	63.63	63.96	63.19	29.90	29.49	30.42
전체	54.67	54.22	55.10	31.69	30.52	32.87

〈표 9〉 성별, 연령대에 따른 후보자 지지율(충남) (단위: %)

연령대	한나라당 이원구			국민중심당 이명수		
	전체	남	여	전체	남	여
20대	42.69	42.13	43.21	23.55	23.72	23.40
30대	38.88	38.27	39.44	19.73	20.91	18.51
40대	50.92	49.59	52.56	22.30	22.13	22.54
50대	54.65	56.03	53.06	26.77	25.42	28.46
60대 이상	51.34	54.32	47.67	27.67	27.27	27.79
전체	48.49	49.26	47.64	24.12	24.06	24.09

대체적으로 남녀간의 차이는 크지 않으나, 연령대에 따른 지지율 격차는 큰 것으로 보인다. 특히, 투표자의 연령대가 높은 경우 전반적으로 한나라당 후보자의 지지율이 열린우리당 후보자의 지지율보다 높은 것을 확인할 수 있다. 서울과 경기의 경우 특히 이런 연령대별 차이가 상대적으로 심한 것으로 나타났으며, 제주의 경우 무소속 후보가 강세를 보임에 따라 양당 대립구도를 벗어나 다른 지역과 달리 연령대별 차이가 발생하고 있지 않다.

2. 시간대별 투표성향

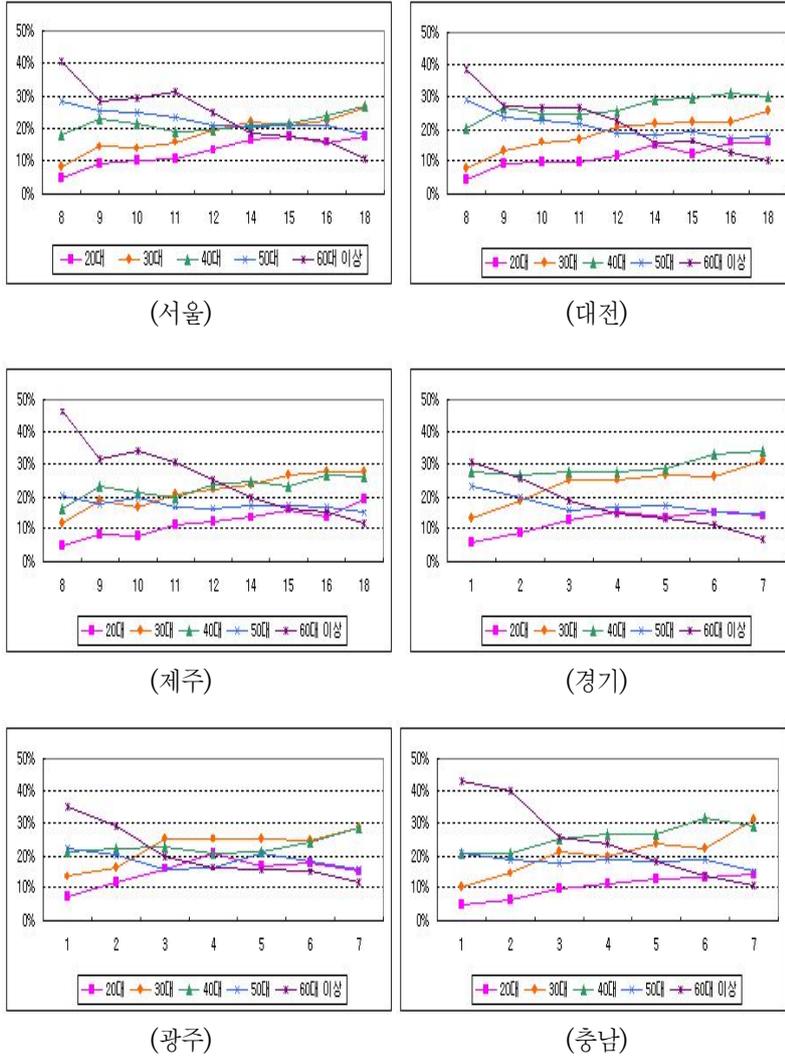
지방선거는 오전 6시부터 오후 6시까지 이루어지며 출구조사 조사원들은

조사결과를 취합하여 일정시간 간격으로 본사로 자료를 전송하게 된다. 참고로 A사의 경우 8시부터 16시까지 매 1시간 간격으로 그리고 투표가 종료되는 18시를 마지막으로 총 9회에 걸쳐 자료를 전송하고, B사의 경우 평균 2시간 간격으로 총 7회에 걸쳐 자료를 전송하게 된다. 실제 선거예측방송에 사용되는 출구조사 선거예측치는 17시 이전에 본사로 전송된 자료를 기초로 산출하게 되며, 각 조사회사에서는 사후분석을 위해 투표가 종료되는 18시까지 조사를 계속 진행한다. 본 연구에 사용된 자료는 18시까지 진행된 전체 출구조사 자료를 사용한 것이기 때문에 실제 선거예측방송에 사용된 결과와는 차이가 있을 수 있다.

이번 출구조사 관련 자문회의에서는 실제 예측치 산출에 사용되는 17시 이전까지의 투표자와 그 이후에 투표에 참여한 투표자들과 특성상 유의한 차이가 있는 경우 오차를 발생시킬 소지가 있다는 점이 집중적으로 논의되었으며, 과거 선거자료에 대한 검토과정을 거쳐 조사회사에서 사용하는 예측치 산출기법을 적용하는 경우 그 차이가 크지 않다는 결론을 내렸다. 그럼에도 불구하고 시간대별 투표자 연령대 분포의 변동 추이를 파악하는 것은 의미가 있다고 생각된다. 각 시간대별 전체 투표자를 100%로 보았을 때 시간대별 투표자 연령대별 분포를 선거구별로 정리해 보면 <그림 1>과 같다. 참고로 A사가 주관한 서울, 대전, 제주의 경우 가로축은 전송시간을 나타내고, B사가 주관한 경기, 광주, 충남의 경우는 전송 차수(총 7회)를 나타낸다.

모든 선거구에서 공통적으로 이른 시간에는 연령대가 높은 투표자가 상대적으로 더 많고, 시간이 지날수록 반대로 연령대가 낮은 층의 투표자들이 늘어나는 현상을 뚜렷하게 볼 수 있다.

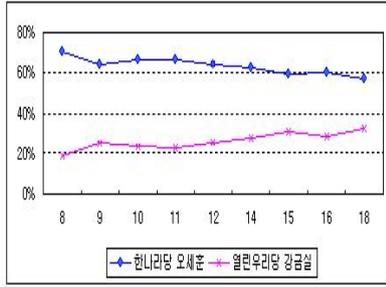
한편, 출구조사 자료를 통해 시간이 흐름에 따라 1위(당선자)와 2위 후보자의 지지율 추이를 살펴볼 수 있다. <그림 2>는 시간대별로 각 선거구의 1위와 2위 후보자의 지지율을 나타낸 것이다. 서울, 경기, 광주의 경우 1위 후보자가 계속 우세를 유지하면서 미미한 차이지만 시간이 지날수록 지지율의 격차가 줄어들어 가는 것을 볼 수 있다. 반면에 대전의 경우 2위 후보자의 지지율이 시간이 흐름에 따라 상승하여 늦은 시간대에서는 지지율이 오히려 앞섰지만 초반 열세를 뒤집지는 못한 것을 알 수 있고, 제주의 경우 두 후보자가 계속 근소한



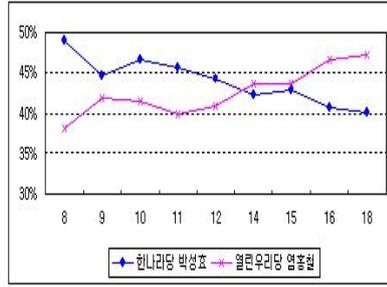
<그림 1> 시간대별 투표자 연령대 분포

차이로 접전을 보였다.

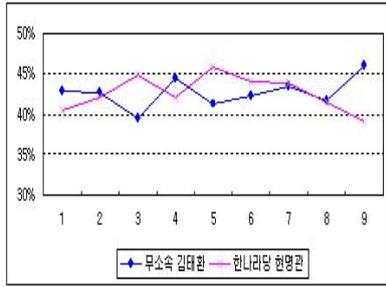
이와 같은 시간대별 지지율 변동 추이는 앞에서 살펴본 각 선거구별로 나타난 연령대별 후보자 지지율과 투표자의 시간대별 연령 분포와 연관하여 그 원



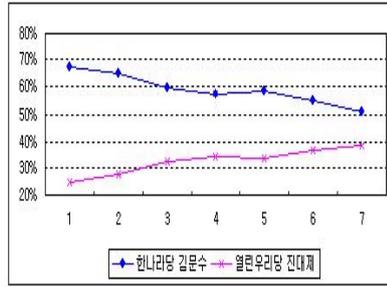
(서울)



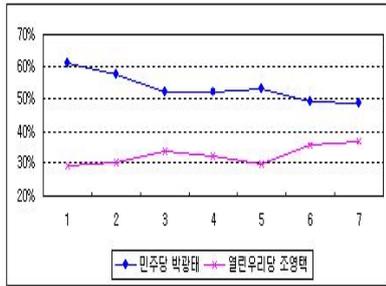
(대전)



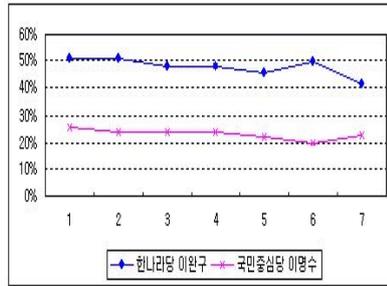
(제주)



(경기)



(광주)



(충남)

<그림 2> 시간대별 후보자 지지율 추이

인을 설명할 수 있다. 예를 들어 서울이나 경기의 경우 이른 시간대에 60세 이상을 포함한 고연령대의 투표자가 많았고, 이 연령층에서 상대적으로 한나라당 후보자에 대한 지지율이 높아 <그림 2>에서와 같은 시간대별 지지율이 나타

난 것으로 볼 수 있다. 다시 말해 투표 시간대별 지지율 차이는 각 시간대별 투표자들의 성별과 연령대 특성으로 설명할 수 있다.

III. 출구조사 예측결과 평가

1. 출구조사 지지율 추정치의 정확도

출구조사 결과의 정확성을 검증해 보기 위해 우선 출구조사 자료에서 산출된 1위와 2위 후보자 지지율 추정치와 실제 개표결과를 비교해 보기로 한다. 개표결과 후보자 지지율은 선관위 홈페이지에서 제공하는 5.31지방선거 투개표 조희시스템 자료를 이용하였으며, 출구조사 지지율 추정치는 오후 6시 개표가 종료될 때까지 수집된 출구조사 자료를 활용해 산출했다. 출구조사에서 무응답 처리가 예측치 산출에 많은 영향을 주기 때문에 출구조사 추정치 산출을 위해 무응답을 무시하는 경우와 무응답을 고려해 추정치를 보정하는 경우를 모두 고려하기로 한다. 참고로 여기에 제시된 지지율 추정치와 달리 실제 방송사에서 사용한 예측치는 오후 5시 전까지 취합된 출구조사 자료만을 사용한 것이며, 조사회사 고유의 다양한 무응답 처리기법을 적용해 산출하기 때문에 여기에 제시된 추정치와 약간 차이가 있다는 점에 유의하기 바란다.

무응답을 고려하지 않고 지지후보를 응답한 투표자들에 대한 조사결과만을 사용한 지지율 예측치는 일반적인 증화추출의 추정식에 따라 다음과 같이 간단히 산출될 수 있다.

$$\hat{p} = \sum_{h=1}^H \left(\frac{N_h}{N} \right) \hat{p}_h$$

여기서 \hat{p}_h 는 h 번째 층에서 특정후보에 대한 단순 지지율을 나타내고 N_h 는 h 번째 층의 총 투표소 수이며, N 은 선거구 전체 투표소의 수를 나타낸다.

한편, 무응답을 고려한 경우 다양한 방식의 무응답 보정에 따른 예측치 산출 방법이 있을 수 있다. 여기서는 우선 응답 자료를 통해 산출된 성별·연령대별 지지율을 산출한 후, 무응답자의 경우(출구조사의 특성상 무응답의 경우에도 성별과 연령이 조사되어 있음)에는 해당 성별·연령대별 지지율이 그대로 유지

된다는 것을 전제로 해당 범주의 지지율로 무응답을 처리했다. 구체적으로 정리하면 성별·연령대를 나타내는 범주를 c 라고 하면 선거구별로 각 범주의 특정 후보 지지율의 예측치 ${}_c\hat{p}$ 는 다음과 같은 층화추출 추정식으로 산출하고

$${}_c\hat{p} = \sum_{h=1}^H \left(\frac{N_h}{N} \right) {}_c\hat{p}_h$$

무응답을 고려한 최종 후보 지지율 추정치 \hat{p}^* 는 다음과 같이 구했다.

$$\hat{p}^* = \frac{\sum_c {}_c m^* {}_c\hat{p}}{M^*}$$

여기서, ${}_c m^*$ 는 무응답을 포함한 c 범주(성별/연령대)에 해당하는 조사대상 투표자 수이며, M^* 는 무응답을 포함한 모든 범주에 속하는 전체 조사대상 투표자 수이다.

〈표 10〉은 각 선거구별로 1위와 2위 후보자의 실제 개표결과 득표율과 출구 조사를 통한 지지율 추정치를 정리한 것으로, b_1 은 무응답을 고려하지 않은

〈표 10〉 실제 득표율과 출구조사 추정결과 비교 (단위 : %, %p)

구분	서울		대전		제주		
	1위	2위	1위	2위	1위	2위	
실제득표율(a)	61.06	27.31	43.84	41.15	42.73	41.11	
출구조사	b_1	63.02	26.10	44.16	42.30	41.81	42.31
	b_2	63.58	25.66	44.32	42.22	42.47	42.64
	$b_1 - b_2$	-0.55	0.44	-0.16	0.08	-0.65	-0.33
차이	$a - b_1$	-1.96	1.21	-0.32	-1.15	0.92	-1.20
	$a - b_2$	-2.52	1.65	-0.48	-1.07	0.26	-1.53
구분	경기		광주		충남		
	1위	2위	1위	2위	1위	2위	
실제득표율(a)	59.68	30.76	51.62	33.95	46.31	25.54	
출구조사	b_1	61.00	32.00	54.41	32.00	48.96	23.07
	b_2	61.37	30.45	55.21	31.53	48.88	24.47
	$b_1 - b_2$	-0.37	0.38	-0.79	0.47	0.08	-1.40
차이	$a - b_1$	-1.32	-0.07	-2.79	1.95	-2.65	2.47
	$a - b_2$	-1.69	0.31	-3.59	2.42	-2.57	1.07

경우, b_2 는 무응답을 고려한 추정치를 나타낸다.

출구조사 추정결과를 살펴보면 우선 제주를 제외하고 모든 선거구에서 한나라당 소속의 1위 후보자를 과대 추정하는 경향을 볼 수가 있다. 이런 현상은 투표자가 자신이 지지하는 후보자가 열세로 판단되는 경우 응답을 거절하는 경향이 있기 때문인 것으로 해석할 수 있다. 이런 특정 정당에 대한 응답편향은 17대 총선에서도 나타났는데, 17대 총선에서는 당선자 예측에 실패한 모든 선거구에서 한나라당 후보자가 과소 추정되는 편향이 발생했다. 이준웅(2004)은 제17대 총선 예측조사에서 나타난 이와 같은 특정 정당 후보자들에 대한 편향의 발생 원인을 ‘조사거절자 편파’ 이른바 ‘침묵의 나선효과’로 설명하고, 우리나라 선거예측조사에서 흔히 발생하는 이런 특정 정당에 대한 응답편향 문제를 집중적으로 다루었다.

한편, 무응답을 고려하지 않은 경우(b_1)와 무응답을 고려한 경우(b_2)의 추정치에 다소 차이가 있는 것을 볼 수 있다. 두 추정치와 실제 득표율(a)과의 차이를 보면, 대체적으로 무응답을 무시하고 구한 추정치가 무응답 처리를 한 추정치에 비해 실제 득표율과의 차이가 적은 것으로 나타난다. 여기서 무응답을 보정했을 때 더 큰 편향이 생기는 것은 동일한 성별과 연령대를 갖는 투표자의 경우에도 응답자 그룹과 무응답자 그룹 간에 실제 지지율에 상당한 차이가 있기 때문에 발생한 것으로 보인다. 예를 들어 무응답률이 높은 고연령층의 경우 응답자 기준으로 계산한 한나라당 후보에 대한 지지율이 실제보다 상당히 높게 추정된 것일 수 있으며, 무응답 보정은 응답자 기준으로 계산된 지지율을 그대로 무응답자 그룹에도 적용해 추정치를 산출하기 때문에 실제 무응답자 그룹의 지지율이 응답자 그룹의 지지율보다 낮을 경우 실제보다 출구조사 예측치는 지지율을 과대 추정하는 결과를 가져오게 된다.

한편, 방송사 입장에서 중요하게 생각하는 당선자 예측 관점에서 보면 당선자(주로 한나라당 후보) 지지율이 대체적으로 과대 추정된 경향이 있으나 제주를 제외한 나머지 5개 선거구에서 당선자 예측에 성공했다. 제주에서 당선자 예측에 실패한 것은 1위와 2위 후보자의 득표율 차이가 불과 1.63%일 정도로 경합이 치열했고, 사후 자체 분석에 의하면(A사가 주관한 선거구 중) 무응

답률이 상대적으로 높았으며, 다른 선거구에 비해 현장 실사과정에서 선관위와의 마찰이 심해 일부 투표소에서 조사가 중단되는 상황이 발생했다는 점 등이 복합적으로 작용한 것으로 추측된다.

일반적으로 출구조사와 같은 표본조사의 오차는 표본추출오차(sampling error)와 편향(bias)으로 구분할 수 있다. 현재 출구조사의 경우 표본크기가 상당히 크기 때문에 표본추출오차보다 편향에 의한 오차가 출구조사에서 중요하게 고려될 필요가 있다. 실제 편향은 조사거절, 거짓응답, 실사 수행상의 오류 등 다양한 요인에 의해 발생하고 편향 발생 요인을 파악하는 것은 현실적으로 거의 불가능하다. 하지만 가능하다면 편향 발생 원인을 규명하기 위해 주요 변수에 대한 표본의 대표성을 점검해 보는 것이 상당히 의미가 있을 수 있다. 특히 출구조사의 경우 무응답자에 대해서도 성별과 연령대를 관측한다는 점과 투표자의 성별과 연령대가 투표성향에 많은 영향을 준다는 점을 고려할 때, 만약 출구조사 표본의 성별·연령대별 구성 비율과 전체 투표자의 성별·연령대별 구성 비율이 일치하는지 여부를 확인할 수 있다면 이는 편향 발생 요인을 파악하는 데 중요한 단서가 될 수 있다. 따라서 표본의 대표성 관점에서 편향의 발생 가능성에 대해 다음 절에서 살펴보고자 한다.

2. 출구조사 표본의 대표성 검토

실제 선거 개표결과 자료를 이용하여 투표자의 성별·연령대별 구성비를 정확히 파악할 수는 없다. 하지만 선관위에서는 사후에 성별·연령대별 투표율을 산출하기 위해 표본조사를 실시하고 있으며 이렇게 작성된 투표율 분석 자료를 이용하면 출구조사 표본의 대표성을 검토해 볼 수 있다. 선관위의 투표율 분석자료는 선거가 종료된 후 전체 투표소 중 약 10%에 해당하는 투표소를 표본으로 추출하고, 이들 투표소의 선거대장을 일일이 확인하여 투표자의 성별과 연령을 조사한 자료를 토대로 선거구별 성별·연령대별 투표율을 산출한 것이다. 선관위에서 적용한 표본추출 및 자료수집 과정을 살펴보면 다음과 같다(중앙선거관리위원회 2006).

전국 13,106개 투표소 중 약 10%인 전국 1,350개 투표소의 3,740,333명의

선거인을 대상으로 조사가 수행되었다. 이를 위해 먼저 시·군·구 선관위별로 투표소를 이번 지방선거에서 나타난 투표율의 내림차순으로 정렬한 후, 전체 투표소 중 10%를 조사 대상으로 계통추출했다. 계통추출과정에서 추출간격은 해당 시·군·구 투표소 중 10%에 해당하는 투표소가 표본으로 추출되도록 결정했다. 이는 출구조사에서 투표구를 산출할 때 과거 선거결과를 기준으로 지지율을 내림차순으로 정렬하고 계통추출하는 것과 유사한 방식이다. 선관위 투표율 분석자료의 경우 직접 선거인 대장을 확인하여 조사하기 때문에 관측과정에서 편향이나 오류가 발생하지 않고, 표본크기도 상당히 크기 때문에 매우 정확한 성별·연령대별 투표자 구성비를 산출하는 것이 가능하다. 선관위 자료를 통해 산출한 성별·연령대별 구성비와 출구조사의 성별·연령대별 구성비를 비교해 보면 <표 11>과 같다.

<표 11> 선관위와 출구조사 표본의 성별·연령대별 구성비 (단위: %)

구분	서울		대전		제주		
	선관위	출구조사	선관위	출구조사	선관위	출구조사	
남	20대	8.73	5.25	8.45	5.40	8.80	4.99
	30대	9.40	8.96	9.66	8.85	11.14	10.53
	40대	10.91	10.45	13.87	13.47	12.01	11.59
	50대	9.90	11.32	9.52	11.30	8.13	9.49
	60대 이상	9.94	11.91	9.14	11.14	9.34	12.38
여	20대	8.44	7.71	6.98	5.97	7.02	5.90
	30대	10.10	9.29	10.98	9.24	10.76	10.14
	40대	11.83	11.04	13.31	13.19	11.02	10.68
	50대	10.27	11.50	8.55	10.03	8.08	8.61
	60대 이상	10.48	12.58	9.55	11.39	13.69	15.70
구분	경기		광주		충남		
	선관위	출구조사	선관위	출구조사	선관위	출구조사	
남	20대	7.52	4.72	9.52	6.48	6.68	3.97
	30대	10.28	10.34	9.41	9.88	7.55	8.30
	40대	14.13	15.07	11.82	11.44	11.65	12.09
	50대	9.07	10.21	8.66	9.33	10.19	9.80
	60대 이상	9.41	10.30	10.08	11.62	14.28	15.39
여	20대	6.70	6.32	7.03	7.37	4.75	4.46
	30대	11.67	11.51	11.28	11.14	8.10	9.04
	40대	12.94	13.39	11.87	11.20	11.42	11.86
	50대	8.22	8.22	8.82	9.08	9.50	9.17
	60대 이상	10.05	9.93	11.50	11.87	15.88	15.93

〈표 11〉을 보면, 선거구별로 차이는 있지만, 선관위 표본과 출구조사 표본의 성별·연령대별 구성비에 상당한 차이가 있다는 것을 확인할 수 있다. 특히 20대 남성의 경우 모든 선거구에서 출구조사의 구성비가 선관위 자료의 구성비에 비해 적고 그 차이가 두드러진다. 이는 그 원인을 쉽게 찾을 수는 없지만 출구조사에서 20대 남성이 실제보다 적게 추출되고 있다는 사실을 보여준다. 선관위 표본과 출구조사 표본의 성별·연령대별 10개 범주의 구성비가 통계적으로 유의한 차이가 있는지 확인하기 위해 각 선거구별로 카이제곱 검정을 해본 결과는 〈표 12〉와 같다. 결과를 보면 p값이 모두 0.01미만의 아주 작은 값으로 나타났으며, 따라서 선관위 표본과 출구조사 표본의 성별·연령대별 구성비는 통계적으로 유의한 차이가 있다고 할 수 있다.

〈표 12〉 성별·연령대별 구성비에 대한 카이제곱 검정결과

구분	서울	대전	제주	경기	광주	충남
검정통계량	568.5	304.8	277.3	294.4	2318.1	171.5
p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

그렇다면, 출구조사의 표본 구성비가 선관위의 표본 구성비와 같았다면 지지율 예측결과에 어떠한 영향을 미쳤을까? 만약 출구조사의 표본 구성비가 선관위 표본 구성비와 같이 유지됐을 경우 예측결과가 더 좋아진다면, 출구조사 표본의 대표성 부재를 출구조사 예측 편향의 발생원인 중 하나로 볼 수 있다. 이를 확인해 보기 위해 출구조사 자료의 성별·연령대별 지지율은 그대로 사용하고 각 선거구별로 선관위의 성별·연령대별 구성비를 가중치로 사용해 1위와 2위 후보자의 지지율을 추정하면 그 결과는 〈표 13〉과 같다.

〈표 13〉의 결과를 살펴보면, 선관위의 성별·연령대별 구성비를 반영하는 경우 출구조사에서 나타났던 한나라당 후보에 대한 과대 추정 편향을 전반적으로 감소시키는 효과가 있다는 것을 확인할 수 있다. 이론적으로 큰 의미가 있다고 볼 수는 없지만 흥미로운 사실은 당선자 예측에 실패한 제주 지역의 경우 선관위 구성비를 적용하는 경우 실제 득표율에 가까운 예측치를 산출하게 됨으로써 근소한 차이로 당선자와 2위를 정확하게 가려낸 것을 볼 수 있다. 따라서

성별·연령대 관점에서의 출구조사 표본의 대표성 부재가 편향을 발생시키는 하나의 원인이 될 수 있다고 판단된다. 출구조사에서 특히 20대 남성이 과소 추출되어 대표성에 문제가 생기는 이유는 쉽게 규명할 수 없지만, 투표소 추출상의 문제보다는 실사과정에서 목표 응답자 선정 및 추적조사 과정과 무응답자의 연령을 추정하는 과정 등에서 발생했을 가능성이 높은 것으로 보인다.

〈표 13〉 가중치 적용에 따른 지지율 추정결과 (단위: %, %p)

구분	서울		대전		제주		
	1위	2위	1위	2위	1위	2위	
실제득표율(a)	61.06	27.31	43.84	41.15	42.73	41.11	
출구조사(b)	63.02	26.10	44.16	42.30	41.81	42.31	
가중치 적용(c)	62.46	26.78	43.11	43.09	42.49	42.47	
차이	a - b	-1.96	1.21	-0.32	-1.15	0.92	-1.20
	a - c	-1.40	0.53	0.73	-1.94	0.24	-1.36
구분	경기		광주		충남		
	1위	2위	1위	2위	1위	2위	
실제득표율(a)	59.68	30.76	51.62	33.95	46.31	25.54	
출구조사(b)	61.00	30.83	54.41	32.00	48.96	23.07	
가중치 적용(c)	60.70	30.99	54.69	31.84	48.81	24.53	
차이	a - b	-1.32	-0.07	-2.79	1.95	-2.65	2.47
	a - c	-1.02	-0.23	-3.07	2.11	-2.50	1.01

IV. 표본추출오차

1. 출구조사의 표준오차

출구조사의 추출과정은 층화 2단 집락추출법에 해당하고, 실제 투표소 추출과정에서의 정렬계통추출을 단순확률추출로 간주하는 경우 각 후보자 지지율 추정량에 대한 분산은 다음 추정식으로 구할 수 있다.

$$\widehat{V} = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \widehat{V}_h(\widehat{p})$$

위 식에서 사용된 층별 분산 추정식은 다음과 같고,

$$\widehat{V}_h(\widehat{p}) = \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h M_h^2} s_h^2 + \frac{1}{n_h N_h M_h^2} \sum_{i=1}^{n_h} M_{hi}^2 \left(1 - \frac{m_{hi}}{M_{hi}}\right) \left(\frac{\widehat{p}_{hi} - \widehat{p}_h}{m_{hi} - 1}\right),$$

여기서 $s_h^2 = \frac{\sum_{i=1}^{n_h} M_{hi}^2 (\widehat{p}_{hi} - \widehat{p}_h)^2}{n_h - 1}$, $M_h = \sum_{i=1}^{N_h} M_{hi}$, $\overline{M}_h = \frac{M_h}{N_h}$ 이고, \widehat{p}_{hi} 는 h 번째 층의 i 번째 투표소에서 지지율, N_h 는 h 번째 층의 전체 투표소의 수, n_h 는 표본 투표소 수, M_{hi} 는 h 번째 층의 i 번째 투표소 투표자 수, m_{hi} 는 표본으로 추출된 투표자의 수, M_h 는 h 번째 층의 전체 투표자 수, \overline{M}_h 는 h 번째 층의 투표소별 평균 투표자 수를 나타낸다.

제시된 분산 추정식을 적용해 1위와 2위 후보자의 지지율에 대한 각 선거구별 표준오차와 층별 표준오차를 구한 결과는 <표 14>와 같다. 대체적으로 출구조사에서 각 후보자에 대한 지지율 추정량에 대한 표준오차는 1%p 미만인 것으로 나타났다. 따라서 지방선거 출구조사에서 실제적으로 예측오차를 줄이기

[표 14] 출구조사 표준오차 추정결과($\sqrt{\widehat{V}_h(\widehat{p})}$)

층	서울		대전		제주		경기		광주		충남	
	1위	2위										
1	1.00	0.78	1.29	1.62	1.00	0.74	1.99	1.67	2.67	2.59	2.01	1.09
2	1.36	0.89	1.35	1.37	1.56	1.73	3.14	3.24	2.00	1.21	1.43	0.98
3	1.08	1.01	1.35	0.81	3.33	2.41	2.05	2.29	1.14	1.35	1.63	2.58
4	4.48	2.96	3.25	1.67	1.08	1.64	1.62	1.52	1.62	1.31	4.66	3.53
5	1.76	1.61	1.71	1.90	1.99	1.04	1.74	1.28	2.31	1.40	1.87	4.86
6	1.69	1.28			4.58	5.42	3.49	1.88			2.11	3.07
7	2.61	2.00					3.26	3.33			2.18	3.70
8	1.74	1.17					4.67	3.70				
9							2.22	1.61				
10							5.60	4.33				
11							4.38	4.81				
12							1.59	1.47				
$\sqrt{\widehat{V}_h(\widehat{p})}$	0.69	0.53	0.79	0.62	0.91	0.98	0.89	0.80	0.90	0.67	0.90	1.04

위해서는 표본크기에 따라 조정이 가능한 표본추출오차보다는 편향을 줄일 수 있는 방안을 적극적으로 모색하는 것이 필요하다.

2. 출구조사 설계효과

출구조사의 경우 단순확률추출법과 다른 방법으로 표본을 추출하기 때문에 출구조사를 실시하기 전에 목표 정도를 만족하는 표본크기를 산정하기 위해서는 출구조사에서 사용하는 표본설계의 설계효과(design effect : Deff)를 파악하는 것이 도움이 된다(Lohr 1999: 241). 출구조사에서의 표본크기와 동일한 표본크기를 갖는 단순확률표본에서의 분산 추정결과와 앞에 제시된 출구조사 분산 추정결과를 이용하여 이번 출구조사의 설계효과를 계산하면 <표 15>와 같다.

<표 15> 출구조사 설계효과(Deff)

구분	서울		대전		제주		경기		광주		충남	
	1위	2위										
Deff	4.06	2.94	2.94	1.86	3.32	3.85	6.73	6.01	2.04	1.29	3.36	6.21

설계효과를 살펴보면 선거구에 따라 또한 1, 2위 후보자에 따라 많은 차이가 있지만 이번 출구조사의 설계효과는 대체적으로 3~4 수준인 것으로 정리될 수 있다. 이는 투표자 3,000명을 직접 단순확률추출하여 얻은 예측결과의 정확도를 그대로 유지하기 위해서는 출구조사에서는 투표자를 9,000~12,000명 정도 조사하는 것이 필요하다는 것을 의미한다. 이런 설계효과는 차후 출구조사의 표본설계 단계에서 목표오차를 만족하는 표본크기(표본 투표소의 수)의 결정과정에 유용하게 활용될 수 있을 것이다.

V. 결론

본 연구에서는 2006년에 실시된 제4회 전국동시지방선거의 출구조사 자료

를 이용하여 우선 투표자의 성별과 연령대에 따른 투표성향과 시간대별 투표자 특성을 살펴보고, 이번 출구조사의 정확성을 편향과 표본추출오차로 구분하여 살펴보았다. 특히 출구조사 자료를 선관위의 투표율 분석 자료와 비교하여 성별·연령대 관점에서 출구조사 표본의 대표성을 점검해 보았다.

지방선거나 대선을 위한 출구조사의 경우 표본크기가 충분히 크기 때문에 출구조사의 정확도 향상을 위해서는 다양한 원인에서 비롯되는 편향을 줄이는 방안이 초점을 둘 필요가 있다. 물론 출구조사에서 통계적인 모형 등을 도입해 사전에 편향을 조정한 예측치를 산출하는 것은 현실적으로 거의 불가능한 것으로 보인다. 하지만 최소한 편향이 어느 정도 발생하는지를 파악하여 신뢰구간 또는 당선자 예측과정에 반영하는 것은 출구조사 방법론 관점에서 의미가 크다고 판단된다. 참고로 금번 지방선거 관련 자문회의에서는 이와 관련해 상당 부분 논의가 이루어졌지만 이와 관련해 향후 좀 더 심층적인 연구가 수행될 필요가 있다.

아울러 다른 나라에서는 출구조사 예측치 산출을 위해 과거 선거에서 나타난 투표소별 특성을 반영한 비(ratio) 또는 회귀(regression) 추정량 형식의 다양한 통계적인 예측모형을 개발해 활용하고 있다는 점을 고려할 때(Pavia-Miralles 2005), 우리나라 출구조사에서도 투표소 추출과정 뿐만 아니라 추정과정에서도 과거 선거 정보를 활용하는 방안을 검토해 볼 필요가 있다.

본 연구에서는 출구조사를 개선할 수 있는 방향을 제시하는 것보다는 현재 우리나라 출구조사 현황과 특성을 소개하고, 이번 선거에서의 오차발생 원인을 정리하기 위해 노력했다. 마지막으로 본 연구가 다양한 분야의 연구자들이 출구조사 방법론에 관심을 갖게 하고 우리나라 출구조사를 한 단계 발전시키는 실용적인 연구를 촉진하는 계기가 되기를 기대해 본다.

참고문헌

- 김영원·엄윤희. 2005. “출구조사를 위한 투표소 확률추출 방법.” 《조사연구》 6(2): 1-32.
- 김정훈. 2003. “선거예측과 출구조사 : 16대 대선을 중심으로.” 《조사연구》 4(2): 87-102.
- 류제복. 2003. “출구조사의 역사와 개선방향.” 《조사연구》 4(1): 31-48.
- 이준용. 2004. “제17대 총선 예측조사의 문제: 조사거절자와 응답불성실자 편향을 중심으로.” 《언론정보연구》 41(1): 110-136.
- 조성겸·김지연. 2004. “출구조사의 투표소 표집방안 비교.” 《조사연구》 5(2): 3-29.
- 홍내리·허명희. 2001. “제16대 국회의원 선거의 예측조사에 대한 사후적 검증.” 《조사연구》 2(1): 1-35.
- 중앙선거관리위원회. 2006. “제4회 전국동시지방선거 투표율 분석 자료.” 중앙선거관리위원회 보도자료.
- Lohr, S. L. 1999. *Sampling: Design and Analysis*. Duxbury Press.
- Pavia-Miralles, J. M. 2005. “Forecasts from Nonrandom Samples: The Election Night Case.” *Journal of the American Statistical Association* 100(147): 1113-1122.