MEASUREMENT ERROR AND HOUSE BIAS

IN 2004 PRESIDENTIAL CAMPAIGN POLLS

Mark Pickup
University of Oxford
&
University of Calgary

Richard Johnston
University of British Columbia

November 2005

An earlier version of this paper was presented to the 2005 Annual Meeting of the American Political Science Association, Washington, DC, September 2–5, 2005. We acknowledge the assistance and advice of J. Scott Matthews and Christopher Wlezien. We are also grateful for astute comments and advice from Herb Weisberg and David Kendall at the APSA session. No less helpful was the postdoctoral research seminar at Nuffield College, Oxford, where the discussion was led by Meredith Rolfe. We absolve all of the foregoing of responsibility for errors of fact or interpretation.

ABSTRACT

We consider alternative methods for extracting house bias and sampling error from estimates of voter preference, drawing upon polls reported during the 2004 US Presidential election. One method is sequential: house bias is extracted by a straightforward regression technique and the result house-adjusted series is filtered as a state-space model with Bayesian estimation. The other method, also a Bayesian state-space model, extracts house and sampling effects simultaneously. The latter technique also allows assessment, relative to the actual outcome, of the industry as a whole. House bias is made to seem stronger when estimation is sequential, and we explore the reasons for this. The burden of evidence suggests that bias in specific firms and polls is not a very serious problem. There might, however, be reason to worry about the industry as a whole. As a byproduct of the analysis, we argue that the vote intention series is a random walk. Moreover, even after house bias and sampling error are removed there remains real movement in public opinion during the election campaign. These findings suggest that candidates' campaigns cannot be dismissed as active forces in the final result.

MEASUREMENT ERROR AND HOUSE BIAS

IN 2004 PRESIDENTIAL CAMPAIGN POLLS

Polls published during a campaign both clarify the situation and confuse it. To the extent that movement in poll readings is driven by the evolution in preferences, it helps players orient to the future and, as required, coordinate on the core of truly viable candidates. But not all apparent movement reflects real underlying shifts. Some may be the product of measurement error, a reflection of inevitable limitations in sample size. Some may reflect differences among firms in sample design, question formulation, weighting, and screening. And if individual firms exhibit bias, so may the polling industry as a whole. This has important implications, as public opinion polls are not innocuous during the election campaign period. The media reports daily on the latest standing of the parties based on these polls, political strategists consult them to determine if there latest initiatives have been effective, and academics consult them to understand how public opinion responds to campaign dynamics.

Two approaches now exist for addressing sampling error and house bias. One is sequential: first extract house effects, on the example of Erikson and Wlezien (1999), and then reduce sampling error by filtering the resultant, house-controlled series. The other, exemplified by Jackman (2004), is simultaneous. The second technique can also be used to assess bias in the industry as a whole, by comparing estimates based only on pre-election polls to estimates anchored to the actual outcome. This paper ventures a controlled comparison of these approaches with official returns and published polls from the 2004 US Presidential campaign.

It matters immensely, it turns out, whether one extracts the effects sequentially or simultaneously. The sequential route—starting with house effects, then moving to sampling

error—yields much bigger estimates of house bias than does extracting the two kinds of error simultaneously. This in turn produces over-estimates of the magnitude of true shifts. Part of the problem lies on the sequential-estimation side: there is strong reason to believe that starting with house effects mistakenly assigns some combination of sampling error and real change to the house-bias component. But part of the problem may also reside on the simultaneous-estimation side: filtering for measurement error deploys information from adjacent observations and this may have the effect of smoothing away differences intrinsic to house practices.

Exploration of these possibilities reveals an implication for the reality of campaign effects. The time series of vote intention appears to be fully or fractionally integrated (interpretively, the two amount to the same thing). By implication, shocks to the series have their effect preserved, and the series is a random walk. This suggests that vote intentions, as they evolve, are dominated by on-line cognition (Lodge et al. 1995).[1] To the extent that this is true, strategic choices by the rival campaigns matter to the outcome. This and the finding of this paper that once we account for house bias and measurement error, there remains real movement in public opinion during the election campaign is strong evidence that campaigns do matter.

The sequential and the simultaneous approaches both hint that in 2004 the industry as a whole was biased. Although the bias appears small, it is worth contemplating.[2] This paper canvasses factors in polling practice that might pull estimates away from underlying reality. It also considers factors in play on Election Day that might pull true preferences away from where they were mere days before.

---

[1] The idea that on line cognition should produce a fully integrated series is argued by Wlezien and Erikson (2002). See also Johnston, Hagen, and Jamieson (2004).

[2] The bias is certainly much smaller than Jackman (2004) found for polls in Australia.

CONTRASTING TECHNIQUES

Erikson and Wlezien (1999) represent the starting point for removing house bias from estimates of vote intention. The essence of their approach is to represent each polling house but one with a dummy variable; the unidentified house serves as the reference category. Similarly, each day of the campaign is represented by a dummy variable. The coefficient for each day then becomes the basis for a house controlled estimate of vote-intention readings for the reference firm. There is no reason to believe that the arbitrarily chosen reference firm best represents the industry, however. The intuitively obvious gauge is the polling firm or firms at the median, as indicated by the full range of house coefficients. The coefficient associated with the median firm can then be added to each day's value to produce the best house-controlled estimate.

Extracting house bias in this way leaves a series that is still noisy with sampling error. But methods for extracting signal from noise are now well established. The obvious approach is the Kalman filter using the state-space approach.[3] Although conventional filtering models cannot account for house bias, we can apply the filter to the output of the house adjustment exercise described in the preceding paragraph. To do this, we set the house-adjusted series in state-space form, represent measured opinion as the product of two unobserved components, and recover the components by Bayesian analysis.[4] The first component is real public opinion and the second is noise produced by sampling error, thus:

$$kshare_t = \alpha_t + \nu_t$$
$$\alpha_t = \rho\alpha_{t-1} + \varepsilon_t^{\alpha} \qquad \textit{Equation 1}$$

[3] An example of this approach is now on the World Wide Web, Donald Green's "Samplemiser," accessible at www.samplemiser.com.

[4] For details on Bayesian estimation and the particular software used in this paper, see the Appendix.

where *kshare* is Kerry's measured share of two-party vote intention on day *t*;

α_t is Kerry's true share at *t*;

v_t is the sampling error component at *t*;

ρ is the first-order autoregressive component (about which more below); and

ε_t^{α} subsumes all unmeasured variation operating at *t*.

The variance of the sampling-error component each day can be estimated from the size of that day's sample.[5] Also, to the extent that public opinion contains memory, readings from surrounding days can be used to increase the accuracy of each day's measurement. This allows us to control for the noise produced by this error and to estimate the real public opinion component. It is important to note that in controlling for house bias so far the temporal relationship of each poll has not been taken into account but that relationship is considered in the removal of sampling error.

It is also possible to extract house bias and sampling error simultaneously, by modeling observed public opinion as a composite of three unobserved components, again in state-space form and estimated by Bayesian analysis. An example of such an approach is Jackman (2004), although our approach is not necessarily the same as his. This time, we deploy the state-space model as follows:

---

[5] The standard deviation of the estimated sampling error is calculated as $\sqrt{p_t(1-p_t)/N_t}$, where p_t is the proportion of valid respondents indicating a Kerry vote intention at time = *t* and N_t is the sample size based on the number of decided voters polled.

$$kshare_t = \alpha_t + \delta_t + \nu_t$$
$$\alpha_t = \rho\alpha_{t-1} + \varepsilon_t^{\alpha}$$
$$\delta_t = \sum_i \beta_i P_{it}$$

Equation 2

where *kshare*, α_t, ν_t, ρ, and ε_t^{α} are defined as above, and

δ_t captures bias from the firms in the field at *t*;

β_i is the bias for the ith firm; and

P_{it} is the ith firm's share of the total sample in the field at *t*.

Extended discussion is required for δ_t and $\sum_i \beta_i P_{it}$. The goal of controlling for house bias is to set the δ_t term equal to zero. In theory, this is accomplished by subtracting the δ_t series from the $kshare_t$ series. To do this we require estimates of β_i for each *i*, where β_i is the bias for house *i* relative to real public opinion. Before the fact, of course, we do not have a benchmark for real opinion. But we can, as in the sequential method, stipulate a benchmark firm, estimate the deviations of other houses from the benchmark, and then normalize the series to the median house.

This discussion underlines what we can only assume: that the industry as a whole converges on the truth. After the fact, we do know the truth, in a manner of speaking: the votes actually cast on Election Day. The election result can be added to the series, in effect as a poll with zero measurement error and zero house bias. Anchoring the series this way will drag the whole series toward the known value. Visual or arithmetic comparison of the results from each estimation with the results from late survey fieldwork gives an estimate of industry bias.

Also requiring comment is the presence of a first-order autoregressive parameter. The nature of the link, if any, between successive observations is highly consequential. Jackman (2004), for example, posits as a simplifying assumption that his Australian opinion series is fully integrated. If this assumption corresponds to the facts, then the time series has perfect memory: the effect of any shock is entirely preserved; only another shock will displace the system. In fact, this assumption is dispensable, and the degree of integration can be estimated directly. We can then treat the degree of integration as a matter for direct estimation rather than *a priori* assumption. We give *prima facie* evidence below that the memory in the series is indeed first-order autoregressive.

DATA ISSUES

The data comprise US presidential trial heats from July, 2004 to the November 2nd election, as registered in PollingReport.[6] Three issues in data preparation must be discharged before we can actually get to analyses. First is the disconnect between reports of results, on one hand, and the fieldwork that generated them, on the other. Second is the problem of pooled presentation of so-called "tracking polls." Third is the fact that a given poll may be reported in more than one way, so criteria are required for choice among the alternatives.

Firms almost never break down results by day in the field. Typically a program of fieldwork is integral, such that the day of interview within the total is affected by accessibility bias. At the same time, companies fear making claims on the small samples typically collected on any given day. So imputation from day of report to day of field is inescapable. One strategy is to assign all weight to the *midpoint* in the fieldwork period, the approach favored by Erikson and

[6] The site is www.pollingreport.com.

Wlezien (1999). This leaves gaps in the series and is, of course, highly arbitrary. An alternative is to treat each day of fieldwork as equally likely to have generated the total reading, impute the result to all those days, and divide each poll's sample size by the number of fieldwork days. We refer to this as the *disaggregated* approach. This approach is no less arbitrary, of course. We examine what difference, if any, the imputation strategy makes.

Late in the campaign some firms took to reporting poll data everyday, where the report pooled data from some number, typically three, of preceding days. The pooling was intended to offset sampling error. This poses a subtly different issue from the imputation issue of the preceding paragraph. In the midpoint treatment, the obvious thing for a tracking-poll report is to take the middle day of those pooled. But by construction, consecutive days of a tracking poll contain heavily overlapping information. For a three-day tracking for instance, roughly two-thirds of the data points in consecutive days' reports will be the same information. The solution is to take every k^{th} report, where k is the number of days pooled, and then make the midpoint imputations only for those polls. The disaggregated approach starts the same way, but then replicates the information for k days.

Tracking aside, a given poll was commonly reported more than once, so selection criteria are required. Duplication occurred for one or more of three reasons:

1. Some polls reported results with and without Ralph Nader's name on the menu. As Kerry-Bush prompts were more ubiquitous throughout, especially in earlier months, we made the two-candidate prompt the preferred one. Whenever a poll was reported both ways, we chose the two-candidate prompt, and discarded the one that mentioned Nader.

But if the only record included Nader—there was no parallel two-candidate prompt—we retained the observation.

2. Some reports contrast mention of the vice presidential candidate with no mention. As the no-mention condition was more ubiquitous, we always chose it if it was available. But if the only record referred to a vice-presidential candidate, we used it.
3. Finally, most reports claimed to report only "likely voters," identified by criteria that varied from firm to firm. Some firms compared "likely" voters with "registered" voters. In these situations we always chose "likely" voters. Similarly, some firms compared "likely" voters with all respondents indicating a candidate preference. Again, the relative ubiquity of the "likely" voter report made us choose that variant.[7] In the rare cases where a "likely" voter report was not made, we accepted the alternative. As registered voter samples outnumbered all-voter samples, we preferred the former when these were the only two options on offer.

The overall selection preference order was as follows:

$$L > LN > R > RN > A > AN > LV > LNV > RV > RNV > AV$$

where L = likely voter universe;

R = registered voter universe;

A = all voters universe;

N = mention of Nader; and

V = mention of vice presidential candidates.

---

[7] We did this even though we are skeptical of such doctoring of the sample. Erikson et al. (2004) show that the likely voter calculation damaged Gallup's 2000 tracking. See also the arguments in Johnston et al. (2004, chapter 2) for the National Annenberg Election Study.

SEQUENTIAL ESTIMATION

In this section we combine the Erikson-Wlezien (1999) strategy for extracting house bias with Bayesian filtering. The logic of the Erikson-Wlezien method implies that we start with house bias. Only when the series is purged of such bias should we proceed to removing sampling error.

Adjusting for House Bias

As already mentioned, the sequential method begins by regressing the measured Kerry share of the two-party vote on day dummy variables (every day, no constant) and house dummy variables (one polling house—in this analysis, "Zogby Tracking"—is reserved as the control). Table 1 reports house results for both midpoint and disaggregated imputations, including the computed comparison with the median house. Differences appear more significant in the disaggregated case than in the midpoint one, but this is an artifact of apparently greater degrees of freedom. In the midpoint setup, three houses stand out as significantly negatively biased; these three yield a one- to two-point lower Kerry share than the median house. In the disaggregated setup, ten houses lie significantly below the reference house and two lie significantly above it. The negative houses lie about one point below the median house and the positive houses lie two to four points above it.

Missing from the table are the coefficients on days of fieldwork. These coefficients yield the house-adjusted series, which can then be anchored to the level of the median house. Daily values purged of house bias and so anchored appear in Figures 1 and 2, for midpoint and disaggregated data respectively. In Figure 1, estimates appear only as points, as several days have no data. Disaggregated data, in contrast, cover every day and so appear in Figure 2 as a line plot. These figures also contain scatter plots of the original polling data.

It is obvious from the figures that the house-controlled series have less variance than the raw ones. Does this allow us to conclude that the difference is attributable to burning off house effects? It turns out that the matter is rather more complicated than that. Start with the more-or-less raw data. The midpoint data have a variance of 4.64 for 155 observations, where each observation is the imputation for a single published poll. Some days have no observations; others, especially late in the campaign, have more than one. For the disaggregated data, the variance is 4.48 for 524 observations, where each observation refers to that day's share of a given poll. In this case, every day has at least one observation, and most have more than one. The smaller variance for the disaggregated treatment reflects the fact that many observations are simply replications of others from the same multi-day poll. These, then, are two ways of representing the total variance in the data. Now consider variances for house-controlled data. In the midpoint treatment, the house-controlled variance is 4.41 for 76 observations. For the disaggregated data, the variance is 3.15 for 123 observations. The number of observations in each treatment flags that the reduction in variance may not be solely the result of removing house effects. In both treatments, and especially in the disaggregated one, adjustment for house effects also pools samples across houses to create a single value for each day. This alone will reduce sampling error.

It is conceptually impossible to partition this reduction into its pooling and house-adjustment components. But it is worth comparing the impact of house adjustment with that from a simpler strategy, taking a "poll of polls." The latter simply averages across all polls conducted on a given day, weighting by each poll's share of all respondents interviewed on that day. This obviously pools samples and so should reduce sampling variation. But it is also a crude house-

effects control. It controls only those houses in the field on the same day, however, and much scope remains for house-induced variation across days as firms check in and out of the field. The poll-of-polls variance in the midpoint case is 4.01 and in the disaggregated case, 3.34. With disaggregated data, controlling house effects makes for a slight improvement over the simple poll of polls. In the midpoint case, house adjustment makes things worse! This is a warning about house adjustment that we shall come back to.

In any case, some of the variance that remains is just noise. If there was no movement in public opinion and no house bias, sampling error alone would produce a variance of 1.89 in the midpoint treatment and 1.08 in the disaggregated treatment.[8] Thus even with the removal of house bias, each series contains more variance than can be attributed to measurement error alone. To produce a major reduction in noise, we need to take the next step and filter the data.

Bayesian Filtering

The intuition that house-adjusted data could be filtered at a second stage originates with Wlezien and Erikson (2002), although they ultimately argue against its utility.[9] As already mentioned, filtering involves accounting for the sampling error of each observation (a function of the sample size underlying the observation) and borrowing information from temporally adjacent observations. Borrowing is possible because public opinion contains memory. The magnitude of the memory must either be determined analytically or be assumed. We propose to estimate memory by taking advantage of a first-order autoregressive setup.

---

[8] This figure is calculated by estimating the sampling error variance for each day, based on the total combined number of people polled on that day and the value estimated for that day by the house bias adjusting method. The sampling error variance for the period is simply the average calculated sampling error variance over the period.

[9] Their estimation employed Samplemiser. Their preference to filtering is to model the MA process that sampling error theoretically produces in the vote intention timeseries and assume that it captures only sampling error, or acknowledge that it also captures real changes in preferences that are very short-lived.

Figure 3 strongly hints that we can do this and that the empirical recovery will largely vindicate the assumption of full integration. The figure presents autocorrelation (AC) and partial autocorrelation (PAC) functions for the house-adjusted vote estimated with midpoint data, the house-controlled series from Figure 1. As the midpoint series treats each day as an independent event,[10] AC and PAC functions should suggest the memory structure. Absolute levels in AC and PAC functions are likely to be underestimated where measurement error has not been accounted for.[11] So the true AC and PAC values are probably larger than Figure 3 suggests. For our purposes, however, what counts is the distribution—the *relative* sizes of values across the lag spectrum.

The slow decline of the AC function suggests that the process is integrated, at least fractionally so. The decline is somewhat unsteady but it is clearly not steep enough to suggest stationary autoregression. Also, the PAC function does not contain any indication that a moving average process needs to be modeled. The simplest interpretation of the AC and PAC functions is that memory in vote intention is fully integrated. A shift in opinion on a given day is completely incorporated, and opinion does not subsequently regress toward some pre-existing equilibrium level. The pattern in Figure 3 is also compatible with a fractionally integrated process, as Wlezien (2003) also shows for 2000 vote dynamics. Distinguishing degree of integration is beyond the scope of this analysis and in the long run the two are equivalent (Wlezien, 2000). Given this uncertainty, it seems wise to model opinion as a first order

[10] For disaggregated data, in contrast, a certain degree of memory is built in, as polls that run over several days have identical information spread over the days in question. Even the midpoint data contains a certain amount of memory because the polls from which these individual data points originate have overlapping periods in the field. Thus, some data points may be capturing similar information.

[11] Erikson and Wlezien (1999). Correction for this attenuation simply involves dividing the estimated correlations by the overall reliability of the data. This just increases all correlations by a constant multiplier.

autoregressive process. If estimation unearths an autoregression factor equal to one, we may conclude that the series is fully integrated.

Figures 4 and 5 contain filtered plots of house-controlled data, for midpoint and disaggregated treatments respectively. Unsurprisingly, filtering eliminates a large amount of variance. For midpoint data, the variance is 2.37; for disaggregated data, the value is 2.45. Compare these values with the house-controlled variance estimates above: for the midpoint series, this represents a 46% reduction in variance; for the disaggregated series, a 22% reduction. This leaves a fair degree of movement not explained by measurement error, suggesting that campaigns do matter in that they can move public opinion. Note that the two strategies yield almost identical true-variance estimates. Although disaggregating the data initially reduces the variance, relative to mid-point imputations, once measurement error is controlled and the structure of public opinion memory is incorporated, this difference is eliminated.

Regardless of imputation strategy, the resulting series appears to be fully integrated. The estimated first-order autoregressive term is estimated to be 0.9993 for midpoint data and 0.9997 for disaggregated data, each effectively indistinguishable from unity. Moreover, residuals from each estimation are white noise, as indicated by the Portmanteau (Q) test. For midpoint data:

$Q = 43.4286$

$\text{Prob} > \chi^2(40) = 0.3274$

For disaggregated data:

$Q = 43.8075$

$\text{Prob} > \chi^2(40) = 0.3132$

For both data types, the Q test suggests that we cannot reject the null hypothesis that the residuals are white noise. There appears to be no unmodelled moving average or further autoregressive process at work in the opinion data.

SIMULTANEOUS ESTIMATION

Having examined a method to remove house bias and a method to remove the sampling error, we now turn to the Bayesian state-space method of doing both simultaneously, as described by equation 2. Plots of opinion with house bias and sampling error extracted appear in Figures 6 (midpoint) and 7 (disaggregated). In addition to the sequentially estimated plots, each figure contains two simultaneously estimated plots, one anchored to the median polling firm (exactly parallel to the sequential case) and one anchored to the actual result. Two things are worth noting:

- One fairly leaps from the plots: the line drops 0.5 points when the series is fixed to the Election Day result. This is the value by which the smoothed tracking overpredicts Kerry's final share. Imputation to fieldwork makes no difference to the discrepancy.
- The series look broadly similar to those extracted sequentially. Indeed, the total variance for the simultaneous plot with midpoint data is 2.37, identical to that for sequential extraction. For disaggregated data, the variance is 2.74, somewhat larger than in the sequential case.

But the broad similarity of the plots masks important differences. The place to start is in estimates for house bias. The impact of technique appears in the contrast between Table 2, which gives house bias estimates from simultaneous estimation, and Table 1. Recall from Table 1 that by the midpoint imputation three houses stood out and that by the disaggregated imputation

twelve houses stood out. In Table 2, only one house—and a different one from before—stands out, and that one only in the midpoint setup. Summarily speaking, almost no house seems biased, although a few come close to the statistical threshold.

To sum up, the simultaneous approach reveals two things. First, it yields much smaller estimates for bias by individual houses. Perhaps house effects have been exaggerated. Second, it hints at bias for 2004 for the industry as a whole. The following sections consider each issue in turn.

DISCREPANCY IN HOUSE EFFECTS

Some of the discrepancy in estimated house effects must be laid at the feet of the sequential method. When we start with house effects and estimate them with simple regression, sampling error alone may lead us to overestimate their scale, either globally or in relation to a particular firm. The potential for random error to be assigned as bias to a particular house is most likely when the house has contributed only one or two polls. Table 3, which lists the number of polls each polling firm contributes to our dataset, illustrates the problem. For example, the disaggregated analysis in Table 1 assigns the largest house bias to National Public Radio (NPR). In fact, there is only one NPR poll, with an N of only 760. This suggests great potential for NPR bias, if any, to be conflated with sampling error. Two other houses contributed only one poll – New Democrat Network and Quinnipiac University. McLaughlin & Associates contribute only two.

The further danger is that real movement in opinion will be interpreted as house bias. This could occur if a polling firm's fieldwork is concentrated in a particular period. If this period happens to coincide with a particularly high or low point in Kerry's popularity, the accident of

timing will be misattributed as house bias. Table 2 presents a measure of the temporal concentration of each polling firm's polls described as the average absolute distance in days from that firm's (temporally) median poll. Using midpoint data, the polls estimated to have the largest bias are International Communications Research (ICR) and the George Washington University (GWU) Battleground polls. With the exception of one early August poll, these polls all fall within the August 29- November 1 period, the weeks following Kerry's late-August crash. Of the houses with more than a handful of polls, these two are among the three most temporally concentrated. Of course, other firms were polling through this period but only one other (Reuters/Zogby Tracking, which reported notably higher values for Kerry) was so concentrated in it.

Overestimation of house bias also produces overestimation of swings in preference. Assigning negative house bias to the ICR and GWU polls, just described, amplifies the estimated recovery of Kerry's popularity in late September and early October, when these polls were particularly concentrated. This can be seen by comparing the simultaneous and sequential lines in Figure 7. The sequential line surges more in apparent response to the first debate, in late September, such that Kerry is made to seem to draw nearly even with Bush. Then the advantage dissipates. Other curiosities inhabit the sequential series. Kerry's drop during the Republican convention seemed to reverse temporarily. There is no reason to believe this actually happened; more likely is that a low-frequency polling firm chose this point to be in the field, correctly captured the ongoing decline in Kerry's fortunes but had the reading register as negative house bias. Perhaps not coincidentally this is the period during which the single National Public Radio poll occurred, as did the one ICR poll that did not fall in the August 29-November 1 period.

Ideally, controlling for house bias should take into account opinion indications from other houses in the temporally proximate days. Doing so would minimize the potential for a house concentrated in a period of particularly high or low popularity from seeming as unusual and exhibiting bias. Simultaneous estimation arguably remedies this lack, as information is borrowed from adjacent observations. But this very strength may itself bias house estimates toward zero, and the culprit, if there is one, would be memory. The simultaneous estimation yields an integrated series, just as the sequential one did. But now the memory includes information from other polls, not information purged (and then some) of other poll indications. As a *per contra* experiment, we constrained the simultaneous method to have zero memory between consecutive days but otherwise continued to filter for sample error and house bias, and found statistically significant house effects. This is the message of the rightmost column in Table 2. With memory erased, four houses now seem distinct and a few more inhabit the border. Some of the polls that seemed distinct in Table 1 also seem so here, although the overlap is far from complete. Perhaps the simultaneous method throws out the baby with the bath water, so to speak. The answer may lie somewhere in between. But then, a series purged of memory does not capture reality. If the answer lies in between, it seems intuitively reasonable to suppose that it is closer to the target identified by simultaneous estimation.

INDUSTRY BIAS?

Even if house bias relative to the industry is truly small in the typical case, it seems premature to assume that the industry taken as a whole quite finds the correct location of opinion. It does not do to exaggerate the problem, of course. Figure 7 suggests that the industry may have been off by half a point. This is much less than one standard error in the sampling distribution of any

individual poll. And it is the bias estimated by one of the techniques, not the other. The median-house sequential-estimation line in Figure 7 ends up closer to the final result than the median-house simultaneous-estimation one (an overprediction of 0.32 percentage points rather than 0.53 percentage points in the disaggregated case). But this may be a fluke of timing. For reasons discussed above, we take the simultaneous estimation to be the better indicator of where the industry effectively sits. Either method produces a prediction closer to the raw results of the final poll conducted November 1, which over predicted Kerry's vote share by 1.75 percentage points.

Although relatively small, a half-point overestimate is a bigger gap than any consecutive-day shift in the late campaign.[12] If this is symptomatic of something real, four potential sources of discrepancy merit exploration. Two refer to industry practice and two, to factors operating outside industry control. Some firms weight data to a known—more to the point, supposed—distribution of party identification. To the extent that this practice is ubiquitous, the industry as a whole may have worked from obsolete data, as the underlying distribution of party identification appeared to shift away from the Democrats between 2000 and 2004.[13] Virtually all firms base their presentations on a "likely voter" calculation. For this reason, we privilege such presentation in our own assembly of the data (76.8% of all poll results included). The basis of likely voter calculation remains obscure for most firms, however. The small body of evidence on the question hints that they produce poorer predictions than simple reliance on the power of random selection (Crespi 1988; see also Johnston et al. 2004).

---

[12] There is also much reason to believe that the amplitude of shifts diminishes in the late campaign. See Wlezien and Erikson (2002) and Johnston et al. (2004), Chapter Two.

[13] We owe this idea to Herb Weisberg. See Weisberg (2005).

But then, maybe the industry did its best and was played false by facts beyond its control. One possibility in this realm is partisan asymmetry in willingness to declare an intention or reveal a behavior. This was hinted by the early exit polls in 2004, which suggested a Kerry victory. Likely voter computations or weighting for party identification ought not to have prejudiced these indications. Finally, it is entirely possible that the polls were collectively correct, or as correct as they could be under the circumstances. The predicted and the actual margins were, after all, razor thin, just as they had been in 2000. The difference between the prediction on the eve of the election and the result on Election Day could have reflected differential mobilization by the parties, to the advantage of the Republican party.[14]

DISCUSSION

Enduring house effects may be mostly illusory. One estimation strategy yields several houses that stand out from the rest, at least when we make a disaggregated imputation to fieldwork. Another estimation strategy suggests that house-specific bias is much weaker. The truth may lie in between: one method assigns some true and some sampling error to specific houses while the other method deploys time-series memory to wipe out house effects. The inclusion of memory certainly seems justified by independent tests but it probably blends house effects across days, obscuring the distinctive effect of each firm's practice.[15]

We went into the exercise worried about choice of fieldwork imputation. We come out concluding that the choice is not very consequential. The disaggregated imputation tends to yield

[14] This suggestion we owe to David Kendall.

[15] Perhaps the really important house effects do not lie in systematic bias, which is what the simple dummy variables used in all this paper's analyses assume. As Wlezien and Erikson (2002) acknowledge, variation *within* houses may be critical, as a result of, *inter alia*, weighting procedures and likely-voter screens. See also Erikson, Panagopoulos, and Wlezien (2004).

seemingly more robust results, but the robustness is an illusion of artificially multiplied degrees of freedom. When the two different data sets are put through the same trial by filtering, the end product is strikingly similar. Inescapably, neither can be wholly true to the facts.

Estimation of memory consistently produces a time series structure that seems fully integrated—as close as it is imaginable to get. This is true regardless of whether one estimates memory at the second stage (as in the sequential method) or right from the start. It is also true in spite of sizeable sampling error in the original data. Evidently, the system of candidate preference has a powerful memory, such that impulses are preserved, except to the extent that they are countered (or reinforced) by comparable shocks. Put another way, the system does not possess autonomic equilibrating forces that neutralize, sooner or later, the effect of shocks. Ironically, the fact that the system has memory suggests that many voters do not. If their opinions are displaced, internal cognitive or external conformity processes do *not* conspire to bring them back to their earlier position. This suggests that campaigns may play a strategic role, along the lines suggested by Johnston, Hagen, and Jamieson (2004) or Hillygus and Jackman (2004). Overall, the fact that campaigns matter is evident in the fact that variances (movement) in vote intentions remain even after house bias and sampling error are filtered out. Finally, the fact that either method of controlling for house bias and sampling error produces a final prediction for the outcome of the election which is substantially better than the raw data from the final poll on November 1 vindicates the need to take these sources of error seriously.

REFERENCES

Crespi, Irving. 1988. *Pre-Election Polling: Sources of Accuracy and Error*. New York: Russell Sage Foundation.

Erikson, Robert S., Costas Panagopoulos, and Christopher Wlezien. 2004. "Likely (and Unlikely) Voters and the Assessment of Campaign Dynamics." *Public Opinion Quarterly* 68: 588-601.

Erikson, Robert S. and Christopher Wlezien. 1999. "Presidential Polls As a Timeseries: The Case of 1996." *Public Opinion Quarterly* 63: 163-177.

Green, Donald P., Alan S. Gerber, and Suzanna L. DeBoef. 1999. "Track Opinion over Time: A Method for Reducing Sampling Error." *Public Opinion Quarterly* 63: 178-92.

Hillygus, D. Sunshine, and Simon Jackman. 2003 "Voter Decision-Making in Election 2000: Campaign Effects, Partisan Activation, and the Clinton Legacy." *American Journal of Political Science* 47: 583-96.

Jackman, Simon. 2004. "Polling and Smoothing the Polls over an Election Campaign." Stanford University: unpublished manuscript.

Johnston, Richard, Michael G. Hagen, and Kathleen Hall Jamieson. 2004. *The 2000 Presidential Election and the Foundations of Party Politics*. Cambridge, UK: Cambridge University Press.

Lodge, Milton, Marco R. Steenbergen, and Shawn Brau. 1995. "The Responsive Voter Campaign Information and the Dynamics of Candidate Evaluation." *American Political Science Review* 89: 309-26.

Weisberg, Herbert, and Dino Christenson. 2005. "Changing Horses in Wartime? The 2004 Presidential Election" Paper presented at the 2005 annual meeting of the American Political Science Association, Marriott Wardman Park, Omni Shoreham, Washington Hilton, Washington, DC.

Wlezien, Christopher. 2000. "An Essay on 'Combined' Time Series Processes." *Electoral Studies* 19: 77-93.

---------- . 2003. "Presidential Election Polls in 2000: A Study in Dynamics." *Presidential Studies Quarterly* 33:.

Wlezien, Christopher and Robert S. Erikson. 2002. "The Timeline of Presidential Campaigns." *Journal of Politics* 64: 969-93.

Table 1: Estimated House Bias – House Adjustment Method

	Fieldwork Imputation			
	Midpoint		Disaggregated	
	Relative to:			
Polling House or Sponsor	Control	Median House	Control	Median House
ABC/Washington Post	-.018* (.010)	-.011	-.016** (.004)	-.008
American Research Group	.001 (.011)	.008	.004 (.005)	.012
Associated Press-Ipsos	-.016 (.010)	-.009	-.008* (.005)	.000
CBS	-.011 (.010)	-.004	-.008* (.004)	.000
CBS/New York Times	.004 (.013)	.011	-.008 (.005)	.000
CNN/USA Today/Gallup	-.008 (.009)	-.001	-.017** (.004)	-.009
Democracy Corps	.015 (.010)	.022	.010** (.004)	.018
Fox News/Opinion Dynamics	-.001 (.010)	.006	-.005 (.005)	.003
George Washington University Battleground	-.026** (.011)	-.019	-.018** (.004)	-.010
International Communications Research	-.023* (.012)	-.016	-.024** (.004)	-.016
IBD/Christian Science Monitor/TIPP	.013 (.012)	.020	.010** (.004)	.018
Los Angeles Times	-.011 (.017)	-.004	-.006 (.005)	.002
Marist College	-.015 (.011)	-.008	-.013** (.005)	-.005
McLauchlan & Associates (R)	-.009 (.015)	-.002	-.018** (.008)	-.010
National Public Radio	.031 (.027)	.038	.031** (.008)	.039
NBC/Wall Street Journal	.002 (.010)	.009	-.008 (.005)	.000
New Democrat Network	-.007 (.024)	.000	-.010 (.008)	-.002
Newsweek	-.016 (.013)	-.009	-.007 (.005)	.001
Pew People & the Press	-.011 (.011)	-.004	-.009** (.004)	-.001
Quinnipiac University	.013 (.018)	.020	.003 (.007)	.011
Reuters/Zogby Tracking	-.005 (.010)	.002	-.006 (.005)	.002
Harris	-.002 (.014)	.005	.004 (.005)	.012
Time	-.007 (.010)	.000	-.012** (.004)	-.004
TIPP tracking	-.008 (.012)	-.001	-.006 (.005)	.002
Zogby America Poll (Control)	— —	.007	— —	.008

Note: values in parentheses are standard errors

p < 0.05 ; ** p < 0.01 ; *** p < 0.001

Table 2: Estimated House Bias – Bayesian Methods

	Midpoint		Disaggregated			
House or Sponsor			Memory		No memory	
ABC/Washington Post	.002	(.012)	.009	(.011)	.003	(.018)
American Research Group	.012	(.016)	.007	(.014)	.016	(.019)
Associated Press-Ipsos	.001	(.018)	-.013	(.017)	-.003	(.023)
CBS News	.000	(.019)	.006	(.015)	.013	(.023)
CBS News/New York Times	-.031	(.025)	-.006	(.019)	-.047*	(.025)
CNN/USA Today/Gallup	-.014	(.012)	-.011	(.012)	-.017	(.018)
Democracy Corps	.017	(.012)	.020	(.014)	-.009	(.022)
Fox News/Opinion Dynamics	.007	(.015)	.022	(.018)	-.001	(.033)
George Washington University Battleground	-.001	(.012)	.002	(.013)	-.022	(.019)
International Communications Research	-.008	(.016)	.020	(.016)	.005	(.022)
IBD/Christian Science Monitor/TIPP	.009	(.015)	.000	(.012)	.010	(.018)
Los Angeles Times	-.001	(.015)	-.015	(.017)	-.031	(.027)
Marist College	.012	(.014)	.000	(.019)	.063**	(.030)
McLauchlan & Associates (R)	-.016	(.025)	-.009	(.032)	-.006	(.054)
National Public Radio	.008	(.043)	.036	(.043)	.081	(.054)
NBC/Wall Street Journal	-.006	(.025)	.018	(.021)	-.006	(.031)
New Democrat Network	.006	(.032)	.049	(.049)	-.037	(.064)
Newsweek	.010	(.012)	.004	(.014)	-.002	(.022)
Pew People & the Press	-.001	(.012)	-.003	(.012)	.027*	(.017)
Quinnipiac University	-.003	(.033)	.009	(.024)	.086**	(.033)
Reuters/Zogby Tracking	-.005	(.015)	.005	(.021)	.016	(.028)
Harris	.034*	(.018)	.018	(.023)	-.0227	(.038)
Time	-.006	(.012)	-.005	(.016)	-.015	(.023)
TIPP tracking	-.009	(.023)	-.025	(.029)	-.032	(.050)
Zogby America	.018	(.013)	.016	(.010)	.020	(.017)

Note 1: control is estimated real public opinion based on actual electoral outcome

Note 2: values in parentheses are standard deviations of the estimated distributions of the house biases

Table 3: House Characteristics

Polling House	Average Number of Respondents	Number of Polls	Average Distance in Days from the Median Poll
ABC News/Washington Post	1280.6364	11	29
American Research Group	881.4000	5	37
Associated Press-Ipsos	889.5000	6	30
CBS News Poll	874.1429	7	26
CBS News/New York Times Poll	824.0000	3	27
CNN/USA Today/Gallup Poll	856.8333	12	29
Democracy Corps Poll	1010.4167	12	23
FOX News/Opinion Dynamics Poll	993.5556	9	30
George Washington University Battleground Poll	1031.2500	8	17
ICR Excel/International Communications Research poll	735.4286	7	21
Investor's Business Daily/Christian Science Monitor/TIPP	774.7143	7	24
Los Angeles Times Poll	1215.5000	4	33
Marist College Poll	807.8333	6	34
McLaughlin & Associates (R)	1000.0000	2	40
National Public Radio Poll	800.0000	1	0
NBC News/Wall Street Journal Poll	784.8000	5	35
New Democrat Network poll	800.0000	1	0
Newsweek Poll	949.6250	8	33
Pew Research Center for the People & the Press survey	1171.8571	7	28
Quinnipiac University	1551.0000	1	0
Reuters/Zogby Tracking	1200.0000	7	5
Harris	1044.0000	3	16
Time	880.0000	10	29
TIPP tracking	956.7500	4	6
Zogby America Poll	1060.1111	9	26

Figure 1: Removing House Bias, Midpoint Imputation

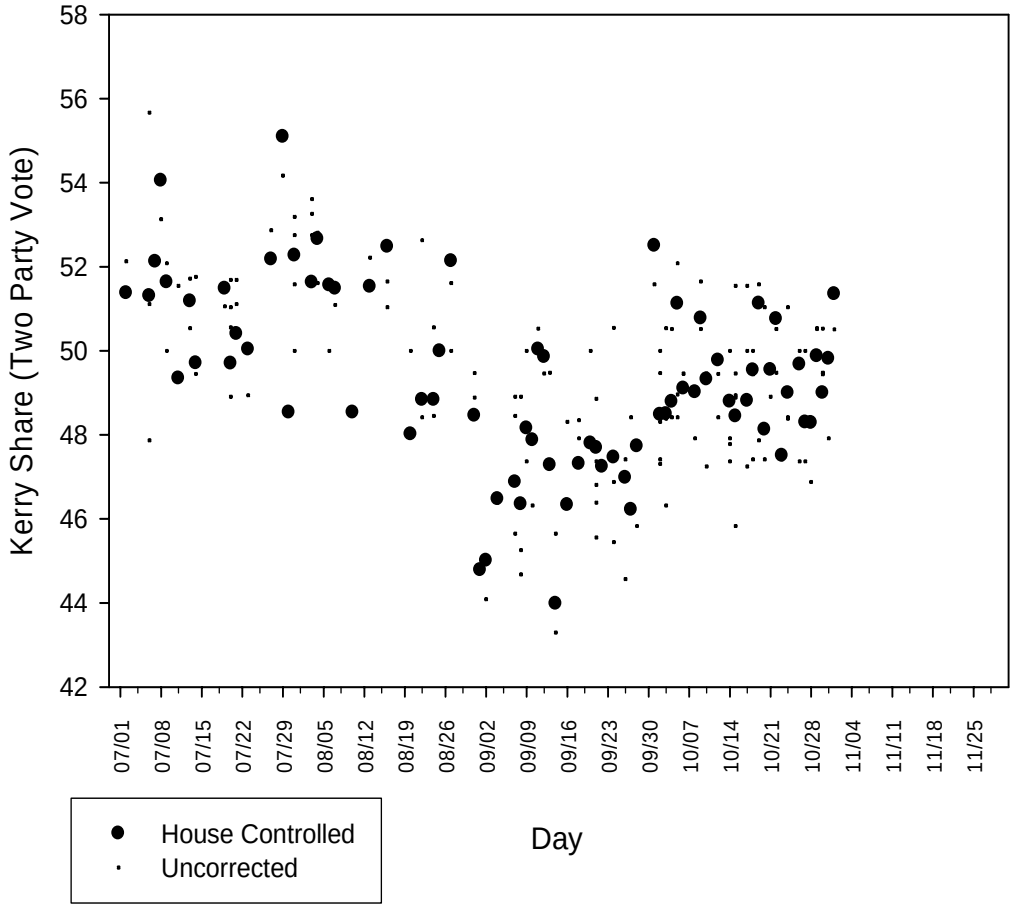


Figure 2: Removing House Bias, Disaggregated Imputation

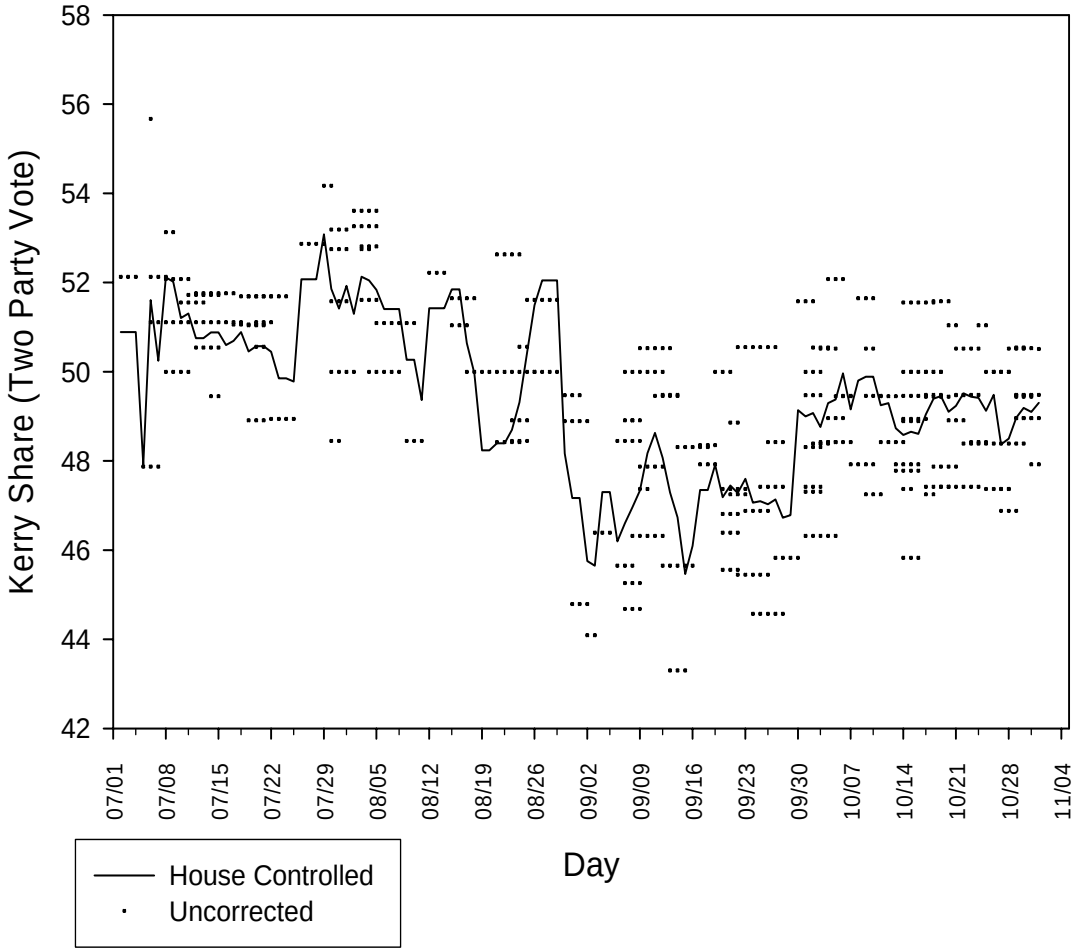
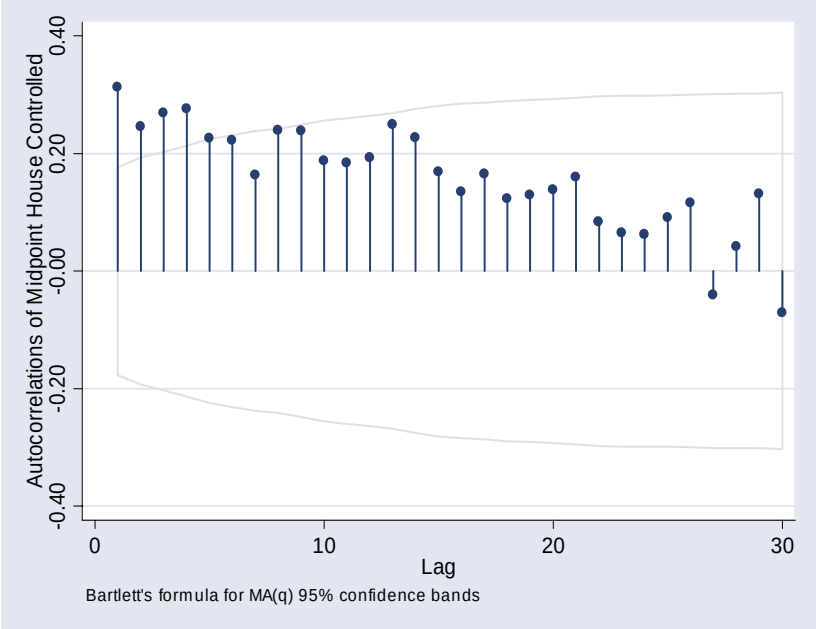


Figure 3: Autocorrelation and
Partial Autocorrelation Functions for House Adjusted Data



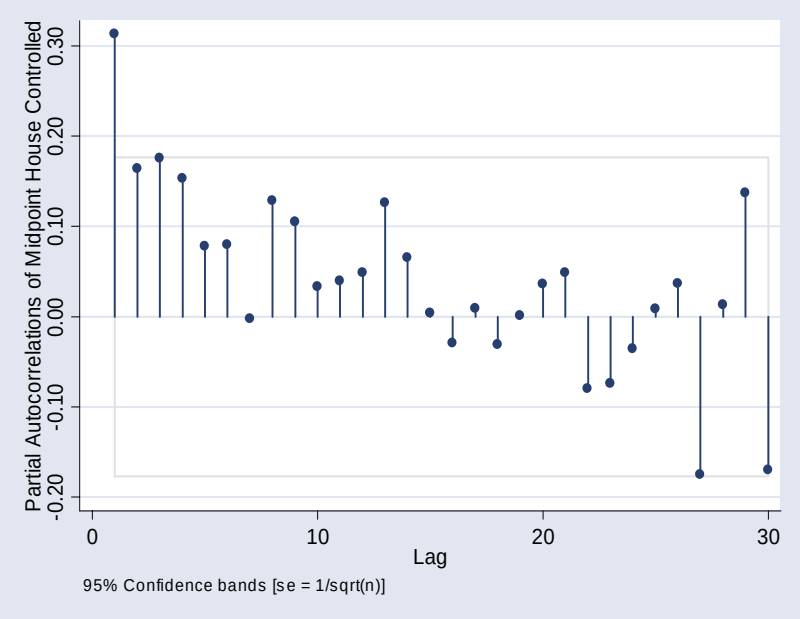


Figure 4: Filtered Data, Midpoint Imputation

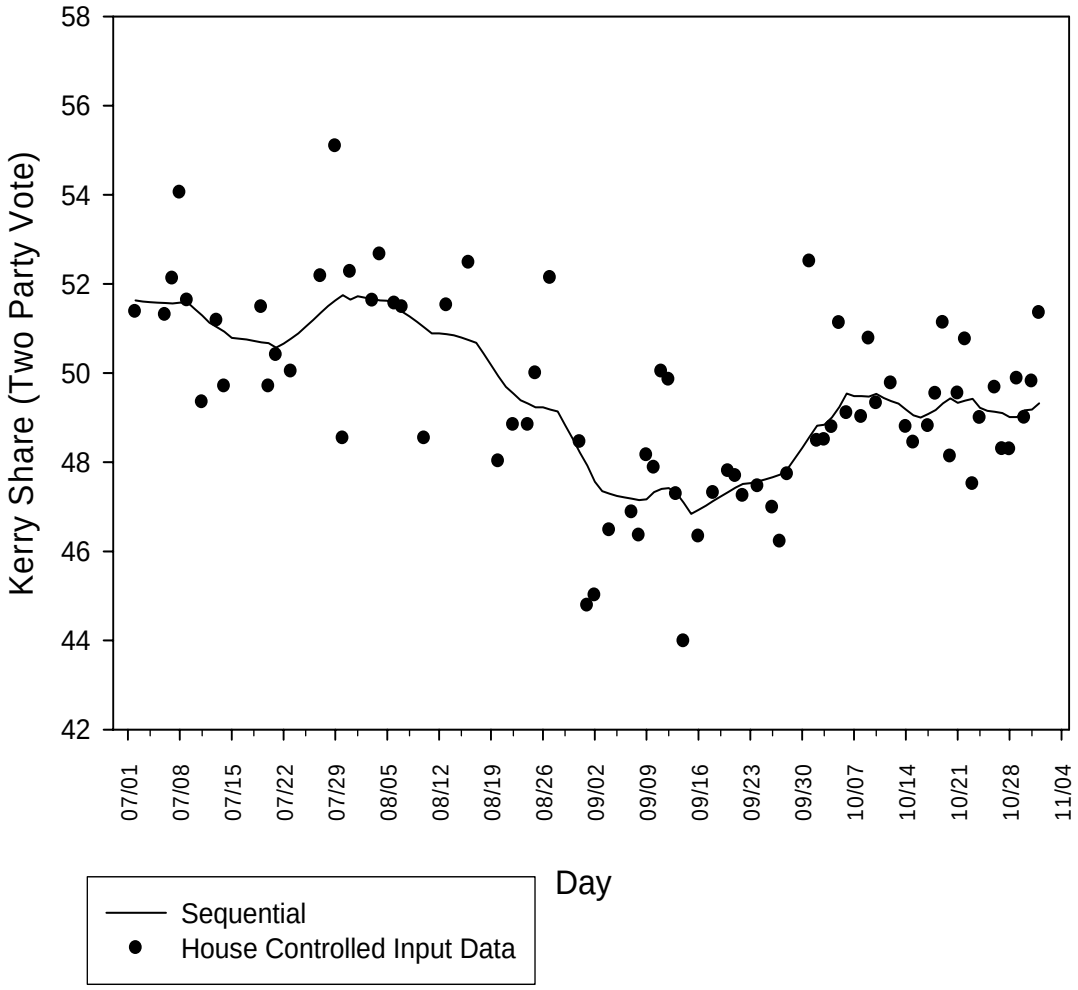


Figure 5: Filtered Data, Disaggregated Imputation

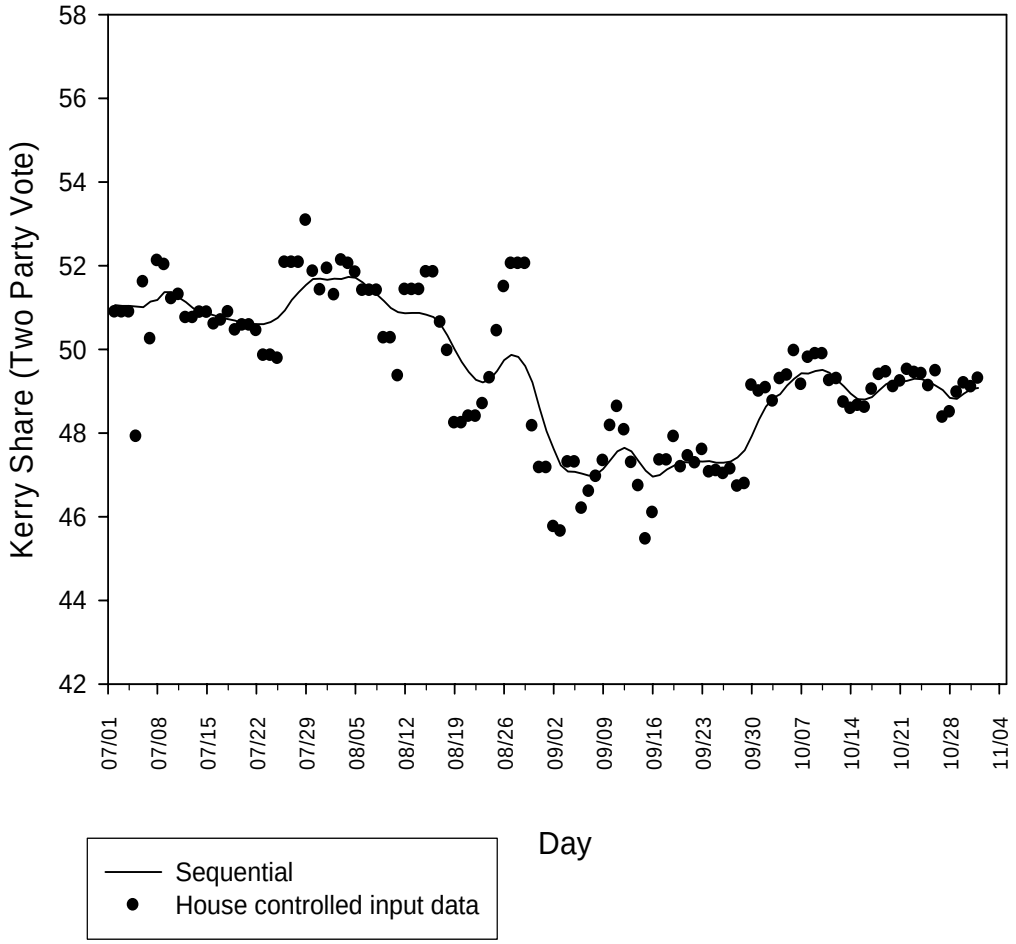


Figure 6: Simultaneous Estimation, Midpoint Imputation

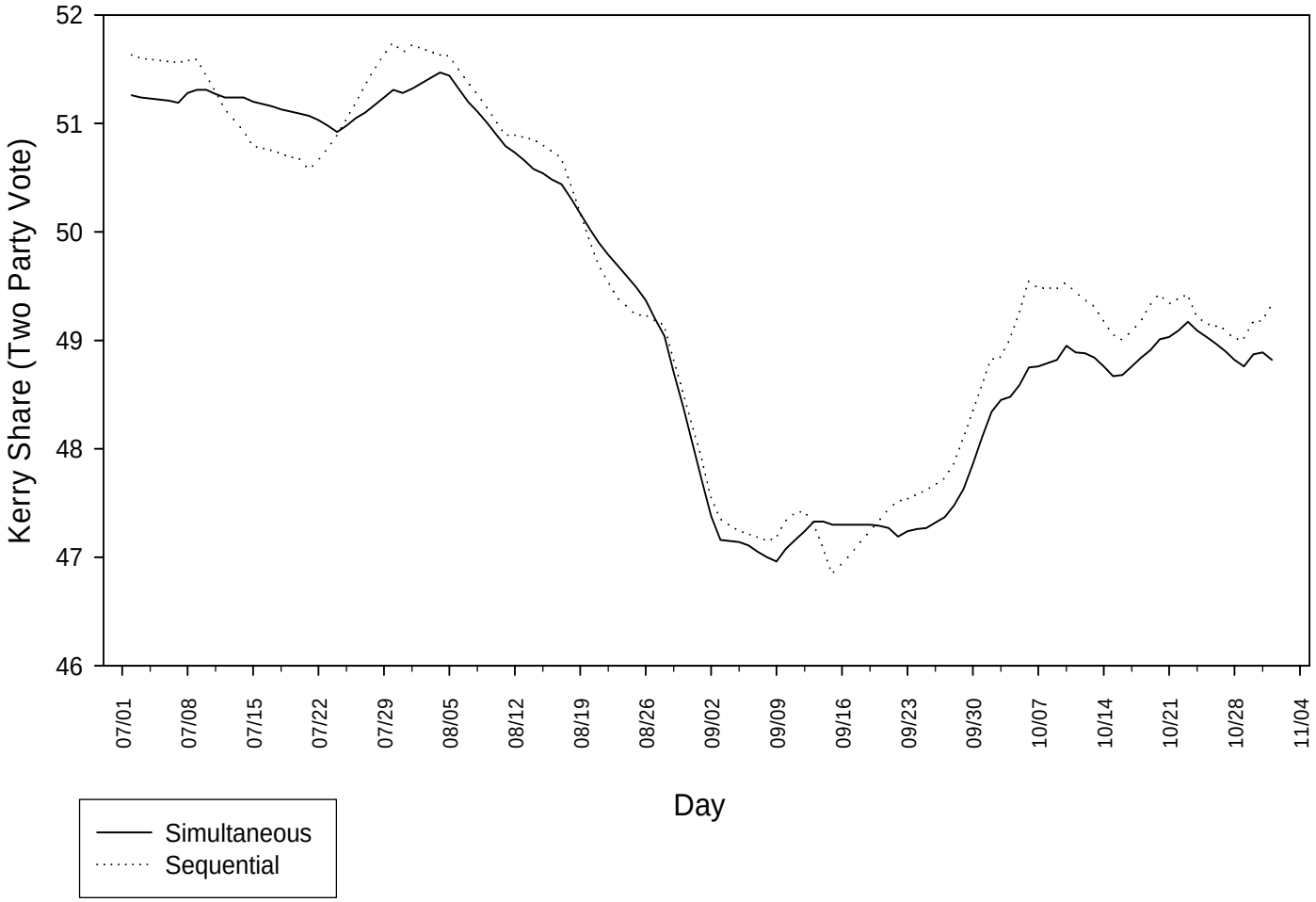
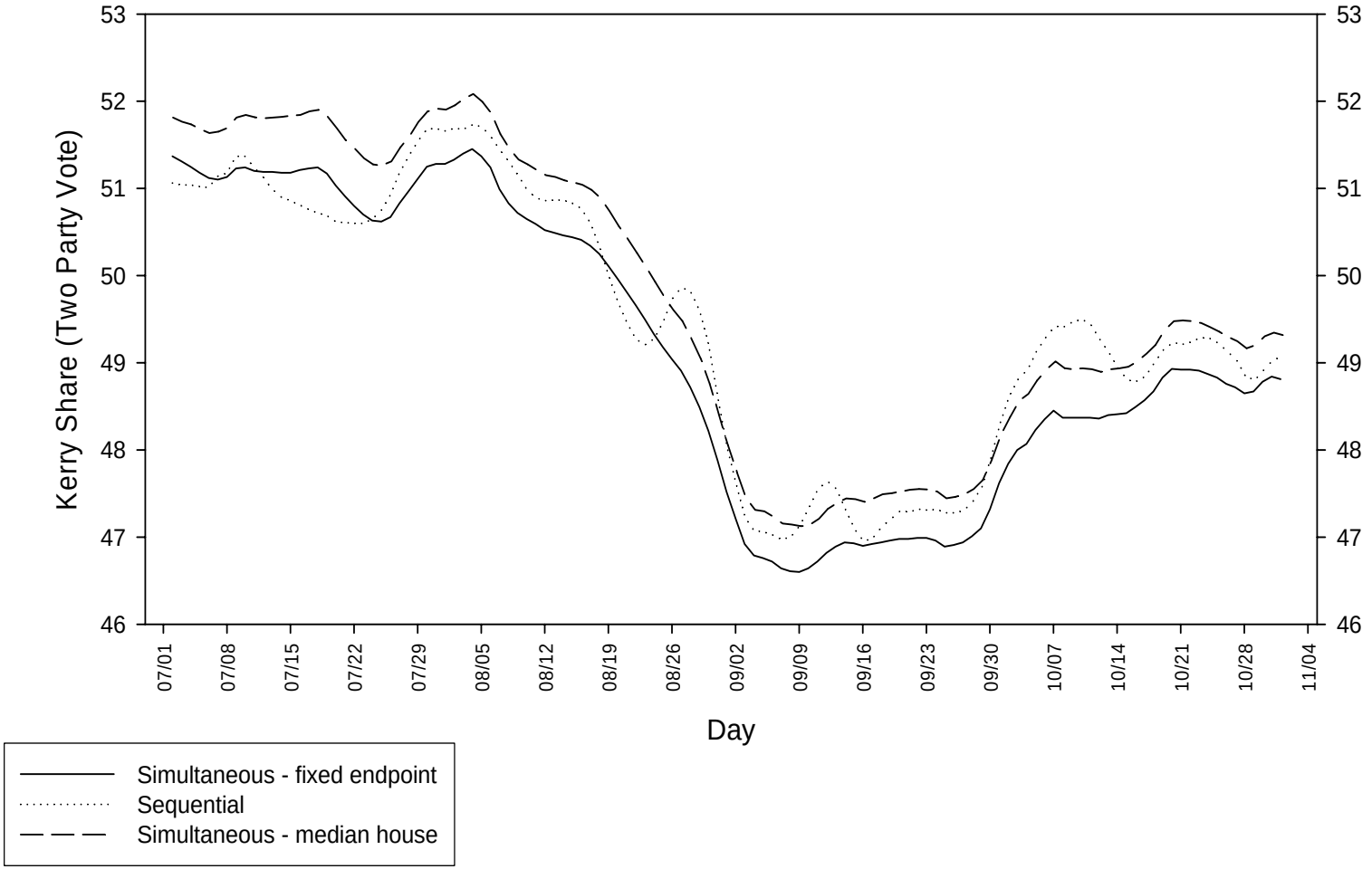


Figure 7: Simultaneous Estimation, Disaggregated Imputation



APPENDIX:

BAYESIAN ESTIMATED DISTRIBUTIONS OF THE STATE-SPACE MODEL PARAMETERS

Bayesian analysis of the state-space models was done using the program Winbugs. This program is designed to analyze statistical models using Markov chain Monte Carlo (MCMC) methods. In the estimation of the distributions of the parameters of our state-space models a single chain with 60,000 iterations was used. For each of the parameters estimated, a non-informative prior distribution was used. This means the parameters being estimated by Bayesian methods are equivalent to those that would be estimated using the principle of maximum likelihood. In each case, the parameters are estimated such that they that maximize the probability (likelihood) of observing the data that was sampled. However, a maximum likelihood estimation is a point estimation of the parameter (producing a single value), while Bayesian methods estimate the distribution of the parameter.

If the distribution of the parameter is roughly normal, the maximum likelihood estimation is essentially the mode (and ideally the mean and median) of the distribution determined through Bayesian methods. In this case, interpreting the statistical significance of the parameter estimated by maximum likelihood methods and the mode of the parameter distribution estimated by Bayesian methods will produce similar conclusions. In the case of maximum likelihood, the standard errors and confidence intervals are calculated based on large sample properties and significance is determined accordingly. In the case of Bayesian estimation, the appropriate percentiles are calculated from the distribution – typically 2.5 and 97.5 percentiles – and significance is determined by whether the null hypothesis for the mode (e.g., equal to zero) falls between these percentiles.