Pak. J. Statist.
1999 Vol. 15(1) pp 47- 63

IMPROVED SAMPLE SURVEY ESTIMATES USING PAST VALUES

Robert Shavelle and David Strauss

Department of Statistics University of California
Riverside, CA 92521-0138 (909) 787-4631
shavelle@stat.ucr.edu
strauss@citrus.ucr.edu

Received: June, 1998 *Accepted: September, 1998*

Abstract

We consider a time sequence of sample survey estimates of a slowly changing population quantity, such as the proportion in favor of capital punishment. The current observation may not be the best estimate of the population value, as it ignores the information in previous observations. Improved estimates may be obtained with the use of a Kalman filtering scheme that incorporates previous values. A problem, however, is that the method has required a rather lengthy series of observations in order to estimate a key parameter. We develop a simple procedure that can be used with very short series. The method is surprisingly easy to use, and in many circumstances is superior to the simple reliance on the current observation. This is demonstrated both under a theoretical model and with real survey data.

Key words

Repeated Surveys; Kalman Filter; Exponential Smoothing.

1. INTRODUCTION

We begin with an example. Figure 1 shows the percentage of U.S. homes with exactly two residents. The survey was done by the National Opinion Research Center (NORC, presented by Wood, 1989) and included approximately 1500 people each year. Also shown is a smoothed version (dotted line) resulting from a method to be introduced in this paper. It can be seen that the smoothed values are generally closer to the true values (solid line) than are the observed values.

The use of smoothing rather than the raw (current) estimate is a possibility whenever sample surveys are conducted repeatedly over time. Familiar examples are

electoral races and long-term studies of attitudes or demographics. The potential advantage of smoothing is that information from previous surveys is "borrowed" to improve the accuracy of the current survey. Heuristically, smoothing will be most beneficial when the true quantity of interest changes slowly over time, or when the individual sample surveys have large standard errors. Intuitively one might expect that it is the ratio of these variabilities that matters. This intuition turns out to be correct. We will formalize it in this paper, and show how the ratio determines the right kind of smoothing to use.

The idea of smoothing a series of estimates is, of course, not new. One of the most familiar methods is exponential smoothing (Brown 1959), in which the current estimate is a weighted average of the previous estimate and current information. Exponential smoothing is less than ideal for series of surveys, however, because it does not reflect knowledge of sampling error. For example, if we have a random sample of size 100 and the true proportion is about 0.5 then the sample standard deviation is 5%, and this can help us choose optimal weights. In addition, exponential smoothing ignores the spacing between the observations (the longer the time gap, the more opportunity for the parameter of interest to change); in contrast, the method to be presented allows the weights to change accordingly.

The time series approach to the repeated surveys problem has a long history. An extensive literature has developed, built around the assumption that the population parameter of interest varies slowly in time. Key early references here are Blight & Scott (1973) and Scott & Smith (1974). Their work has been further developed by, among others, Harrison & Stevens (1976), Scott, Smith & Jones (1977), Smith (1978), Tam (1987), Binder & Dick (1989), Bell & Hillmer (1990), Pfeffermann (1991), Abraham & Vijayan (1992), and Tiller (1992).

There is a close analogy between the method we describe here and the systems used by engineers to track the movements of projectiles whose current position and velocity are measured with error. Despite this, the proposed method is remarkably simple and straightforward to use; it can, in fact, be carried out on a pocket calculator. The method is based on the well known Kalman filter (Kalman 1960). In the so-called state space form this has two components: (1) an evolution equation for the changes over time of the true quantity of interest (such as the percentage of U.S. homes with exactly two residents); and (2) an observation equation, expressing the current observation as the sum of the true quantity and random sampling error. The Kalman filter provides a smoothing for the series of observations, together with estimated standard errors of the smoothed values. It is optimal in the sense of generating estimates with the smallest mean squared error.

Previous researchers have noted the connection between exponential smoothing and the Kalman filter (Bunn 1980, 1981; Enns 1982; Gardner 1985; Harvey 1984, 1989) but were unable to utilize the power of the Kalman filter due to difficulties in estimating unknown parameters. Others have used time series long enough to take advantage of the asymptotic properties of the maximum likelihood estimators (Bell & Hillmer 1990;

Pfeffermann 1991; Scott, Smith & Jones 1977; Abraham & Vijayan 1992; and Tiller 1992). It has been found that estimators based on short to moderate length series (lengths less than 50) are highly unstable, lead to poor estimates, and that more robust procedures need be investigated (Bunn 1980, 1981; Tam, 1987; Binder & Dick 1989; and Pfeffermann 1991). Methods for working with shorter series have been developed (Singh, Mantel, and Thomas, 1994), but require modelling of supplemental data that may not always be available. We will see that the problem can generally be bypassed for our sample survey applications.

The approach developed here is to work with a user-supplied value of the variance parameter, rather than a data-based estimate of it. This choice determines the signal to noise ratio, which dictates the proper weights to be applied. We show that the resulting Kalman filter estimates are often surprisingly insensitive to the value selected. We suggest some simple guidelines for selecting the value. Throughout the paper we refer to the current estimate, which ignores all previous information, as "the survey estimate." We compare the Kalman filter estimates to the survey estimates both in theory under a known model, and in practice using real data for which the true values are essentially known. We close with a discussion of critical issues to consider when applying the method.

2. THE KALMAN FILTER ESTIMATOR

We wish to estimate a true population quantity, which is changing over time, by taking repeated polls at regular or irregular intervals. The population quantity, whose value at time t will be denoted by u(t), will typically be a percentage or an average. According to the simplest version of the Kalman filter model, u(t) changes over time according to a random walk model specified by the evolution equation:

$$u(t) = u(t-1) + \eta(t). \qquad (2.1$$

Here the η's are assumed to be independent "shocks," with mean zero and variance $\sigma_{ev}^2(t)$. In some applications, the variance will be taken as a constant, σ_{ev}^2. The observed percentage, y(t), is the true percentage corrupted by sampling error, and can be expressed as:

$$y(t) = u(t) + \varepsilon(t) \qquad (2.2$$

Here, the ε's are assumed to be independent with mean zero and variance $\sigma_o^2(t)$. The equations (1) and (2) form the well-known random walk plus error, steady, or local-level model. See Diderrich (1985) for a more general version of the filter. A key feature of the model is the so-called signal-to-noise ratio, q(t), given by

$$q(t) = \frac{\sigma_{ev}^2(t)}{\sigma_0^2(t)}. \tag{2.3}$$

It turns out that q(t) suffices to determine the relative weights applied to current and past estimates.

According to Kalman filter theory (Harvey, 1989, p. 107) the best linear unbiased estimator of the population quantity u(t) is the weighted average, m(t), where m(t) is computed recursively as

$$m(t) = [1 - k(t)]m(t-1) + k(t)y(t). \tag{2.4}$$

The weight k(t) applied to the current observation, which is known as the *gain*, lies in (0,1]. It is itself computed *recursively* by

$$k(t) = \frac{q(t) + k(t-1)[\sigma_0^2(t-1)/\sigma_0^2(t)]}{q(t) + k(t-1)[\sigma_0^2(t-1)/\sigma_0^2(t)] + 1}. \tag{2.5}$$

If the measurement errors are constant, equation (2.5) reduces to

$$k(t) = \frac{q(t) + k(t-1)}{q(t) + k(t-1) + 1}. \tag{2.6}$$

The variance of the estimator, m(t), is given by

$$P(t) \equiv var[m(t)] = \sigma_0^2(t)^* k(t). \tag{2.7}$$

Note that the above development assumes that the random walk plus error model holds. If it does not, then the specified estimates are not necessarily unbiased, and equation (2.7) is only a lower bound for the variance.

Equations (2.4)-(2.7) may appear unfamiliar. They derive, however, from a straightforward application of the usual Kalman filter equations to the random walk plus error model (see, for example, Harvey, 1989, pp. 100-101, 105-106). Our formulas are algebraically equivalent to those of Harvey (1989, p. 107).

According to equation (2.7), the advantage in using this method will be greatest when the gain, k(t), is small. Equation (2.5) shows that small gains will be realized only when q(t) is small. Conversely, when q(t) is very large, the gain is near one and the method reduces to the survey estimate (i.e., use of the current observation alone).

Normally one lets m(1)=y(1), which implicitly sets k(1)=1, and makes the variance of m(1) equal to $\sigma_o^2(1)$ (Harvey, 1989, p. 108). The estimation algorithm then starts at time t=2. A special case is q=0, which corresponds to an unchanging evolutionary parameter. If in addition the sampling variance is constant it is then easy to verify that k(t)=1/t, and hence m(t)=[y(1)+y(2)+ ... +y(t)]/t, the sample mean at time t, as expected.

At this point, an example may be helpful. We consider the first six years of data from Figure 1, the percentage of U.S. homes with exactly two residents. Table 1 shows the observed values from the NORC poll in column 3. The sample sizes for the polls are given in column 4. We now show how to obtain the remaining columns. Following convention we let m(1)=y(1)=.270 and k(1)=1. The next step (t=2) requires the

computation of q(2), k(2), and m(2). To compute q(2) we need to estimate the measurement error variance, $\sigma_o^2(2)$, and the evolutionary error variance, $\sigma_{ev}^2(2)$. For simplicity we assume that the polling procedure was random sampling so that we may use the familiar p(1-p)/n formula for the variance of the observed value. We have

$\sigma_o^2(2) = u(2)[1-u(2)]/n(2) \approx y(2)[1-y(2)]/n(2) = (0.30)(0.70)/1503 = 0.0001397$

We will use $\sigma_{ev}^2(2) = (.01)^2$ for the evolutionary variance at each step; we discuss how to choose σ_{ev}^2 shortly. Then, we may find q(2) as

$$q(2) = \frac{\sigma_{ev}^2(2)}{\sigma_0^2(2)} \approx \frac{(0.01)^2}{0.00014} = 0.716$$

Next, we compute the gain, k(2), using equation (5)

$$k(2) = \frac{0.716+1}{0.716+1+1} = 0.632$$

Finally,

$$m(2) = [1-k(2)]m(1) + k(2)y(2) = (1-.632)y(1) + (.632)y(2) = .289$$

In the same way one may generate the remaining entries of table 1.

Table 1. Kalman filter estimates of the percentage of U.S. homes with exactly two residents.

Year	t	y(t)	n(t)	$\sigma_o^2(t)$	$\sigma_{ev}^2(t)$	q(t)	K(t)	m(t)
1972	1	.270	----	-----	----	----	1.00	.270
1973	2	.300	1503	.00014	$(.01)^2$	.716	.632	.289
1974	3	.300	1482	.00014	$(.01)^2$	.706	.572	.295
1975	4	.300	1490	.00014	$(.01)^2$	.710	.562	.298
1976	5	.320	1497	.00015	$(.01)^2$	.688	.555	.310
1977	6	.310	1530	.00014	$(.01)^2$	.715	.560	.310

If $\sigma_o^2(t)$ is constant the signal-to-noise ratio will be constant if, in addition, the time gaps between polls are equal. In such case the gain converges to a *steady-state* value (Chatfield, 1989, p. 188; Harvey, 1989, pp. 118-119). This steady-state gain may be found by setting k(t)=k(t-1)=k and then solving for k in (4), and is given by

$$k = \frac{\sqrt{q^2+4q}-q}{2} = \frac{q+\sqrt{q^2+4q}}{2+q+\sqrt{q^2+4q}}. \qquad (2.8)$$

The rate of convergence depends on the value of q. One might be tempted to use this steady-state value for the gain as early as step 2. However, doing so would place too

little weight on the current observation, and result in suboptimal estimates. If we are willing to accept this handicap for the first few steps, we may use the *steady-state* value for the gain at *all* steps. Then, for any non-zero value of q, the resulting recursion for the estimator will be

$$m(t) = (1-k)m(t-1) + ky(t) \; for \; t = 2,3,\ldots \; and \; m(1) = y(1). \tag{2.9}$$

This makes the updating equation particularly simple. Exponential smoothing is a special case of this scheme, with the smoothing parameter set equal to k.

The key input to use the Kalman filter method is the value of q(t), which determines the gain k(t). We reiterate that non-constant q(t) is required if either the time intervals between polls varies, or if the sampling error is not constant. In such cases the smoothing parameter will change from one step to the next.

As other investigators have found (Bunn 1980, 1981; Enns 1982; Tam 1987; Binder & Dick 1989; and Pfeffermann 1991), it is difficult to estimate q(t) directly from the data when we have a short time series. Therefore our recommendation is to compute q(t) from the ratio of a suitably *chosen* value of $\sigma_{ev}^2(t)$ and the *known* value of $\sigma_o^2(t)$ from survey sampling theory. As we will see, the estimate of $\sigma_{ev}^2(t)$ does not need to be very precise; the only possible danger is in using a gross *under*-estimate of $\sigma_{ev}^2(t)$. This will be taken up in the next section.

3. PERFORMANCE OF THE KALMAN FILTER ESTIMATOR UNDER THE RANDOM WALK PLUS ERROR MODEL

In this section we assume that the true model is the random walk plus error, and compare the performance of the Kalman filter estimates with that of the survey estimates. Naturally, the performance of the Kalman filter will depend upon the value of q which is used (recall that we must use an estimate of this since the true q would not be known). It will be seen that our method outperforms the survey estimate for a wide range of q values used.

We consider two questions. Firstly, by how much does the Kalman filter outperform the survey estimate when the true value of q, q(true), is actually used in the filter? Secondly, how do the two methods compare when the value of q used in the filter, q(used), is not equal to q(true)? Our basis for comparison is the ratio MSE(Kalman)/σ_o^2. Here, for simplicity, we have assumed that the measurement error variance is constant, and that the observed values are equally spaced in time, so that the evolutionary variance may also be taken as constant. Thus, we may use the steady state value for the gain, k.

It is shown in the Appendix that

$$\frac{MSE(Kalman)}{\sigma_0^2} \approx \frac{k}{2-k} + \frac{(1-k)^2}{k(2-k)} q(true), \tag{3.1}$$

where k is the steady state gain computed from (2.8), with q = q(used). It follows from (2.7) that if we work with a known q(true), the MSE ratio in (3.1) would be equal to k. Thus the filter reduces the mean squared estimation error by a factor of 1-k. (In this sense, the

description of k as the "gain" appears to be a misnomer.)

The theoretical results are given in Figure 2. The vertical scale is the ratio of the mean squared error for the Kalman estimate, MSE(Kalman), to the mean squared error for the survey estimate, σ_o^2. The horizontal scale is the ratio of q(used) to q(true), henceforth referred to as the q ratio. The four curves correspond to q(true)=1/20, 1/4, 1, and 2. The horizontal line with ordinate one is the break-even point, where the two methods have equal mean squared errors.

Notice that if the q ratio is larger than 1, then for *all* values of q(true) the Kalman method always has smaller MSE than the survey estimate. The vertical axis corresponds to the case when the q ratio is equal to one (i.e. we are using the correct value of q in the filter) and the improvement achieved by the Kalman estimator is greatest here. The graph shows that when q(true)=2 the MSE is reduced by 23%; when q(true)=1 by 40%; when q(true)=1/4 by 60%; and when q(true)=1/20 by 76%. As the true value of q approaches zero, the advantage in using the Kalman filter increases dramatically. Conversely, for q(true) much greater than 2, the advantage in using the filter, even when the true value of q is employed, is minimal. Of course, as long as we do not under-estimate q(true), there is no loss in using the Kalman filter. The graph reminds that as q(used) increases the Kalman filter reduces to the survey estimate. It appears that if we believe q(true) to be much greater than 2 we might reasonably use the survey estimate alone.

Next consider the case when we do *not* use the correct value of q in the filter. When q(true)=1, i.e. $\sigma_{ev}^2=\sigma_o^2$, the curve is well below the break even point for most values of the q ratio. In fact, the filter is more accurate than the survey estimate except when the q ratio is 1/5 or less. When q(true)=1/4 the curve flattens out, showing that an accurate choice of q(used) is even less important than, say, when q(true)=1. For the entire range of the q ratio shown, the MSE ratio is 0.60 or less, meaning that use of the Kalman filter has led to an at least 40% reduction in the MSE (or 23% in the standard error).

The most extreme case shown is q(true)=1/20. For example, if $\sigma_{ev}^2=(0.01)^2$ and a random sample of size 125 is used, then σ_o=4.5% and q=1/20. In this instance, the curve lies well below the break even line for a very wide range of values of the q ratio. And for the range shown in Figure 2, if q(used) is within a factor of 2 of the correct value of q, MSE(Kalman) is less than 1/4 of σ_o^2. For comparison, this means that with the survey estimate alone, one would need four times the sample size to achieve the same precision as the Kalman filter.

We again emphasize that the only danger in using the Kalman filter is in using a gross *under*-estimate of the value of q. For all four curves shown, the q ratio must be less than 0.25 for MSE(Kalman) to be greater than σ_o^2. We feel that the user will rarely have such poor knowledge of σ_o^2 and σ_{ev}^2 that q(used) would be under-estimated to this extent. Further, since there is little penalty in overestimating q one can always err on the side of caution.

Lastly, we caution that the results of this section assume that the random walk plus noise model holds. It remains to investigate the efficacy of the Kalman filter estimator (derived using the random walk plus noise model) when applied to data generated by an alternative model. For example, the true series might have a fixed mean (a special case where $\sigma_{ev}^2 = 0$),

or itself be either a random walk plus error or an AR(1) model. In addition, the development of more elaborate models for more general error structures remains to be investigated; see note 2 below.

4. PERFORMANCE OF THE KALMAN FILTER ESTIMATORS IN PRACTICE

We have seen that the Kalman filter does well in theory when the underlying model is true. In this section we examine how it fares in practice. Testing the method on real data sets is not as straightforward as it may appear at first sight, as it is essential to have a true series against which to compare the various methods.[1]

Our "truth" data comes from the U.S. Bureau of the Census Current Population Reports (Current Population Survey or CPS) for the years 1972-1989, which rely on samples of approximately 59,000 homes. The error in such a survey, if reasonably sound procedures have been followed, is negligible. Our observed data comes from the National Opinion Research Center (NORC) as presented by Wood (1989). Comparability of the two data sets may be resolved by consulting the original data sources (NORC is based on the General Social Surveys, Davis and Smith, 1972-1994), though we have not done so here. We now consider three examples.

The first example, as shown in Figure 1, concerns the percentage of U.S. homes with exactly two residents. The NORC poll was based on about 1500 people. If for simplicity we act as if the NORC used simple random sampling,[2] then

$$\sigma_o \approx \sqrt{y(1)[1-y(1)]/n(1)} = \sqrt{(.27)(.73)/1500} \approx 1\%.$$

---

[1] It may seem that one could choose any sequence of sample surveys, draw small random samples at each time point and compute both the simple estimate and the smoothed Kalman estimate based on these sequences of small samples. One could then compare how close each sequence is to the original (larger) data set. The problem here is one of potential bias. For example, if the original data is subject to a one-time error due to poor sampling technique then the Kalman filter will only partially respond to the error. As a result the survey estimate would be incorrectly judged to outperform the filter.

[2] The assumption of a simple random sample will often lead to an underestimate of $\square_o^2$ (Schaeffer et al., 1990, p. 258) and thus an overestimate of q. According to the results of the previous section and equation (5), this produces an estimate that is conservative, in the sense of applying a (possibly) larger than optimal weight to the current observation. The precise variances of the NORC estimates are available (c.f. Davis and Smith 1994) and may be used to compare alternative models and methods. Even if these were used, however, one would still need to take account of the overlapping sampling procedure used by NORC, which induces autocorrelation in the sampling errors. We have chosen to make the simplifying assumption of a simple random sample in order to highlight the power of the Kalman filter, and the insensitivity to the results of the computation of q.

We might wish to improve upon the accuracy of the survey estimate by employing the Kalman filter. To compute q we need to impute a value for σ_{ev}^2. Our experience with the true proportions for demographic data suggests that σ_{ev}=0.01 is generally a reasonable choice, and is conservative in the sense of being unlikely to be an under-estimate. [We calculated the actual values of σ_{ev} for 20 series which were demographic in nature and found that the values ranged from 0.002 to 0.01, with most being about 0.005.] Using the usual formula for sampling error from a random sample and equations (2.8) and (2.9), we obtained all the estimates as shown in the graph. The mean squared error (MSE) for the Kalman estimates is 0.00011, compared to 0.00020 for the survey estimates. This means that the standard error is reduced by about 25%, corresponding to a 45% reduction in the sample size required to achieve the same accuracy if one had used the survey estimate alone. This example shows that the properly employed Kalman filter yields a worthwhile improvement in estimation accuracy.

Since we know the values for the true series, we may compute a crude estimate of the true value of q. The estimate is crude in that it comes from a short series, is obtained under the assumption of the random walk plus noise model, and ignores the overlapping sampling structure of the Current Population Survey. We find $q \approx 0.08$, so that according to theoretical results in the previous section the method should perform well.

The second example, Figure 3, shows the true, observed, and filtered values for the percentage of black Americans who have finished at least one year of a college education. Under the same assumptions and following the same technique as before, the MSE for Kalman is 0.0017 compared to 0.0050 for the survey estimates.

Our last example shows that the method may perform well even if there is a linear trend in the data. Figure 4 shows the true, observed, and filtered values for the percentage of Black respondents whose homes have seven or more residents. The MSE for Kalman is 0.00005 compared to 0.00017 for the survey estimates. Notice that although the observed values seem not to be severely biased, and do tend to mirror the decline in true values, they nevertheless vary about the true values much more than does the smoothed version. Once again, the Kalman method is clearly superior.

5. DISCUSSION

The method presented here may be summarized as follows:

1. Our first estimate is taken to be the first observed value.
2. The new estimate is a weighted average of the new observed value and the most recent estimate.
3. The weight applied to the current observation, known as the gain, is a simple function of the signal-to-noise ratio q(t), i.e., the ratio of $\sigma_{ev}^2(t)$ to $\sigma_o^2(t)$.
4. The estimated variance of the current estimate is equal to the sampling error variance multiplied by the gain.

We have seen that the technique is simple to use and is superior to the survey estimate in many situations. The method is also very flexible. For example the weight placed on the new

observation is a function of q(t), and thus can be made to reflect unequal time intervals between polls and/or unequal sample sizes. The former is easily modelled by taking the evolution variance as proportional to the time interval, as is appropriate for a random walk. Also, sample size differences are reflected in the error variances $\sigma_o^2(t)$, as computed from sampling error formulas. This flexibility is not possible with traditional exponential smoothing methods. In general we expect the Kalman filter to perform well compared to the survey estimate when (2.1) the sample is not subject to consistent bias; (2.2) the evolutionary error is relatively small in comparison to the measurement error, so that q is small (1/10 or less is desirable); and, (2.3) the true series is roughly a random walk. We examine each of these three issues in turn.

Firstly, to assess a series of polls for possible bias one must either have a true series for comparison, which is rarely possible, or have definite convictions about the sampling procedure which generated the observed values. In our perusal of real data sets we encountered a few instances where the observed series of values was *consistently* above or below the underlying true values. For example, one NORC survey (not shown) gave percentages below the true values for all 16 polls taken. When this happens the Kalman filtered estimates are similarly biased and are of little value.

Secondly, we have seen (Figure 2) that if q is not much larger than unity, say, the Kalman approach may provide a substantial improvement over the survey estimate. In most applications, the user can easily decide whether the method is worthwhile by considering the ratio of the variances. For example, as noted previously, actual demographic proportions often vary by less than 1% annually (σ_{ev}=0.01). A random sample of size 250, say, results in a standard error of about 3% (provided that the unknown percentage is not close to 0 or 1), and thus $q=(0.01/0.03)^2$, or about 1/10. For a true q in this range, Figure 2 shows that the Kalman method offers a substantial improvement over the survey estimate (provided that the q value used is correct within an order of magnitude).

Application to cases where the value of q is much larger than 1 would be straightforward if an accurate estimate of q were available. However, in the absence of such information, and for series of modest length, use of the filter is not recommended. In particular, the method is perhaps not to be recommended when the underlying true value could change as much as 5% between polls (for example, the swing in the percentage of Americans who would vote Republican after their national party convention). Since polling organizations typically use sample sizes of 1000 or more, resulting in a standard error of 2% or less, an evolution standard deviation of 5% would give rise to a very large q.

Thirdly, we comment on two types of departures from the random walk plus error model assumptions. One possibility is that the true series may have a *systematic* trend; for example the percentage of Black homes with seven or more residents, as shown in Figure 4, is steadily decreasing. As we saw in that example, the Kalman method may well outperform the survey estimate even in the presence of a trend. Of course, if we observe a trend in the data and expect it to continue then we should use a forecasting method that explicitly models it. See, for example, Tiller (1992). A second problem is that of abrupt changes in the "level" of the process. For example, Americans' opinion about Middle Eastern military involvement

changed suddenly after the Iraqi invasion of Kuwait. Like any smoothing method, the Kalman filter will show an undesirable lag in response to such "step" changes. If these changes are believed likely then the Kalman method may be modified by the addition of an appropriate explanatory indicator variable, or the inclusion of a trend constant–a non-zero mean for $\sigma(t)$ (Box and Tiao 1975).

Finally, the natural extension to the tracking of multiple, possibly dependent, time series is of interest. A familiar example is that of tracking the percentage of persons who would vote for one of a number of rival political candidates. The Kalman filter theory can support such work, as it has been developed in matrix form.

This study raises a number of issues for further research. As mentioned in section 3, other underlying true models might be assumed to hold, and the performance of the Kalman filter estimator (which was derived under the random walk plus noise model) examined. Further, more general state-space estimators, possibly with more complex error structures, could be considered. Finally, a further comparison of various estimators (c.f., Makridakus et al. 1984) as applied to a number of empirical data sets is warranted.

6. APPENDIX: RELATIVE MSE'S OF KALMAN AND SURVEY ESTIMATORS

If the true values u(t) change over time according to a random walk

$$u(t) = u(t-1) + \eta(t),$$

then u(t) is related to the initial value u(1) by

$$u(t) = u(1) + \sum_{i=2}^{t} \eta(i).$$

The observed value, y(t), is the true value corrupted by measurement error, so that

$$y(t) = u(1) + \sum_{i=2}^{t} \eta(i) + \varepsilon(t).$$

We now assume that the evolutionary variance, $\sigma_{ev}^2(t)$, and measurement error, $\sigma_o^2(t)$, are constant, and that the time gaps between observations are equal. Thus, we may use the steady state value for the gain

$$k = \frac{q + \sqrt{q^2 + 4q}}{2 + q + \sqrt{q^2 + 4q}}.$$

Then the recursion defined in equation (9) leads to the Kalman filter estimator, m(t), of the true value, u(t), at time t

$$m(t) = k \sum_{i=1}^{t-2} (1-k)^i y(t-i) + (1-k)^{t-1} y(1).$$

Some simple algebra gives

$$m(t)-u(t)=(1-k)^{t-1}\varepsilon(1)+\sum_{i=0}^{t-2}k(1-k)^{i}\varepsilon(t-i)-\sum_{i=2}^{t}(1-k)^{t-i+1}\eta(i)$$

The mean squared error of the Kalman filter estimator, m(t), at step t is given by

$$\begin{aligned}
MSE(Kalman) &\equiv E[(m(t)-u(t))^2] \\
&= var(m(t)-u(t))+[E(m(t)-u(t))]^2 \\
&= var(\varepsilon)^*[(1-k)^{2(t-1)}+\sum_{i=0}^{t-2}k^2(1-k)^{2i}]+\mathrm{var}(\eta)^*\sum_{i=2}^{t}(1-k)^{2(t-i+1)} \\
&= \sigma_0^2[(1-k)^{2(t-1)}+\frac{k^2[1-(1-k)^{2(t-1)}]}{1-(1-k)^2}]+\sigma_{ev}^2[\frac{[1-(1-k)^{2(t-1)}](1-k)^2}{1-(1-k)^2}].
\end{aligned}$$

If t is of even moderate size, say 5 or more, the MSE may be approximated by a much simpler expression which does not depend upon t:

$$MSE(Kalman)\approx\sigma_0^2\frac{k}{2-k}+\sigma_{ev}^2\frac{(1-k)^2}{k(2-k)}.$$

Finally, the ratio of the mean squared error for the Kalman filter estimator, given above, to the mean squared error, σ_o^2, for the (unbiased) survey estimator is

$$R=\frac{k}{2-k}+\frac{(1-k)^2}{k(2-k)}q(true),$$

where k is the steady state gain computed from (8) by using q(used), and q(true) is the true value of q.

7. REFERENCES

1. Abraham, B., and Vijayan, K. (1992). Time series analysis for repeated surveys. *Communications in Statistics-Simulations*, 21, 893-908.
2. Bell, W.R., and Hillmer, S.C. (1990). The time series approach to estimation for repeated surveys. *Survey Methodology*, 16, 195-215.
3. Binder, D.A., and Dick, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.
4. Blight, B.J.N., and Scott, A.J. (1973). A stochastic model for repeated surveys. *Journal of Roy. Statis., Soc.*, Series B, 35, 3-8.

5. Box, G.E.P. and Tiao, G.C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70, 70-79.

6. Brown, R.G. (1959). *Statistical Forecasting for Inventory Control*. New York: McGraw-Hill.

7. Bunn, D.W. (1980). A comparison of several adaptive forecasting procedures. *Omega*, 8, 485-491.

8. Bunn, D.W. (1981). Adaptive forecasting using the Kalman filter. *Omega*, 9, 323-324

9. Chatfield, C. (1989). *The Analysis of Time Series, An Introduction*. New York: Chapman and Hall.

10. Davis, J.A. and Smith, T.W.: *General Social Surveys, 1972-1994*. [machine-readable data file]. Principal Investigator, James A. Davis; Director and Co-Principal Investigator, Tom W. Smith. NORC ed. Chicago: National Opinion Research Center, producer, 1994; Storrs, CT: The Roper Center for Public Opinion Research, University of Connecticut, distributor. 1 data file (32,380 logical records) and 1 codebook (1073 pp.).

11. Diderrich, G.T. (1985). The Kalman filter from the perspective of Goldberger-Theil estimators. *The American Statistician*, 39, 193-198.

12. Enns, P.G., Machak, J.A., Spivey, W.A., and Wrobleski, W.J. (1982). Forecasting applications of an adaptive multiple exponential smoothing model. *Management Science*, 28, 1035-1044.

13. Gardner, E.S. Jr. (1985). Exponential Smoothing: The State of the Art. *J. Forecasting*, 4, 1-28.

14. Harrison, P.J., and Stevens, C.F. (1976). Bayesian forecasting. *J. Roy. Statist. Soc.*, Series B 38, 205-247.

15. Harvey, A.C. (1984). A Unified View of Statistical Forecasting Procedures. *J. Forecasting*, 3, 245-275.

16. Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. New York: Cambridge University Press.

17. Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Engineering*, 82, 34-45.

18. Makridakus, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. (1984). *The Forecasting Accuracy of Major Time Series Methods*. New York: Wiley.

19. Pfeffermann, D. (1991). Estimation of seasonal adjustments of population means using data from repeated surveys (with discussion). *J. of Business and Economic Statistics*, 9, 163-177.

20. Schaeffer, R.L., Mendenhall, W., and Ott, L. (1990). *Elementary Survey Sampling*. Boston: PWS-Kent.

21. Scott, A.J., and Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *J. Amer. Statist.l Assoc.*, 69, 674-678.

22. Scott, A.J., Smith, T.M.F., and Jones, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *Inter. Statist. Rev.*, 45, 13-28.

23. Singh, A.C., Mantel, H.J., and Thomas, B.W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology*, 20, 33-43.

24. Smith, T.M.F. (1978). Principles and Problems in the Analysis of Repeated Surveys. In *Survey Sampling and Measurement*, Ed. N.K. Namboodiri, pp. 201-216. New York: Academic Press.

25. Tam, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. *Inter. Statist. Rev.*, 55, 63-73.

26. Tiller, R. (1992). Time series modeling of sample survey data from the U.S. current population survey. *J. Official Statistics*, 8, 149-166.

27. U.S. Bureau of the Census, *Current Population Reports*, Series P-20, "Marital Status & Living Arrangements, Educational Attainment in the U.S., Money Income of Households in the U.S., and Household and Family Characteristics: March 1972 through March 1989," U.S. Government Printing Office, Washington, D.C., 1973-1990.

28. Wood, F.W. ed., (1990). *An American Profile-Opinions and Behavior, 1972-1989*. Detroit: Gale Research.

Figure legends

Figure 1. The percentage of U.S. homes with exactly two residents. The dashed line is the observed data from a National Opinion Research Center (NORC) poll of approximately 1500 people. The dotted line represents the smoothed values obtained by using the Kalman filter. The solid line is the "true" data from Current Population Reports of the U.S. Bureau of the Census.

Figure 2. Plot of MSE ratio versus q ratio shows the effectiveness of the Kalman filter method for various choices of q values used in the filter when the true value of q is equal to 1/20, 1/4, 1, and 2. The vertical axis is the ratio of the mean squared error for the Kalman filtered estimate, MSE(Kalman), to the mean squared error for the survey estimate, $\square_o^2$. The horizontal axis is the ratio of the value of q used in the filter, q(used), to the true value of q, q(true). For discussion see text.

Figure 3. The percentage of Black Americans who have finished at least one year of a college education. The dashed line is the observed data from a National Opinion Research Center (NORC) poll of approximately 150 people. The dotted line represents the smoothed values obtained by using the Kalman filter. The solid line is the "true" data from Current Population Reports of the U.S. Bureau of the Census.

Figure 4. The percentage of Black homes with seven or more residents. The dashed line is the observed data from a National Opinion Research Center (NORC) poll of approximately 150 people. The dotted line represents the smoothed values obtained by using the Kalman filter. The solid line is the "true" data from Current Population Reports of the U.S. Bureau of the Census. Notice that for most years the smoothed values are closer to the "true" values than are the observed values.

Figure 1 Plot of MSE ratio versus q ratio

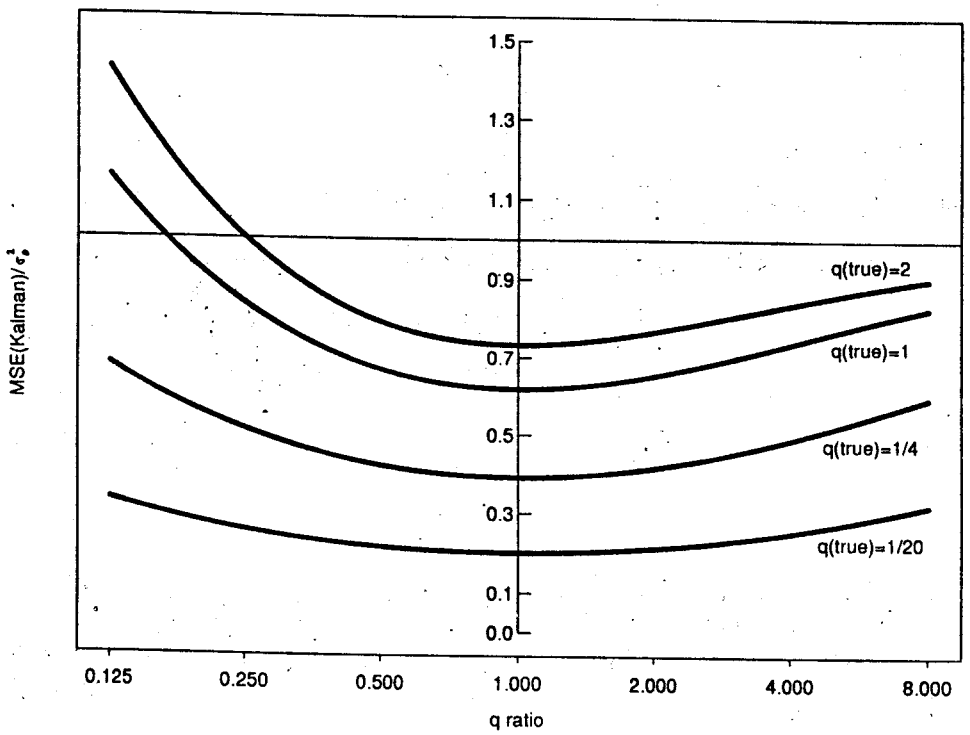

Figure 2 Percentage of U.S. homes with exacty two residents

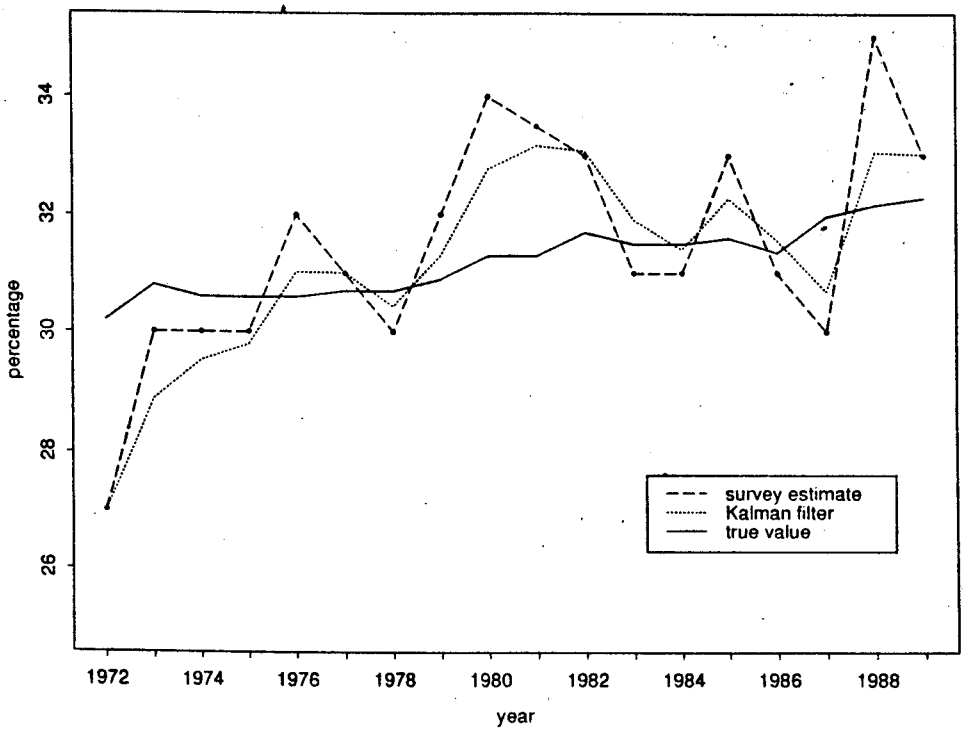

Figure 3 Percentage of black Americans who have finished at least one year of college

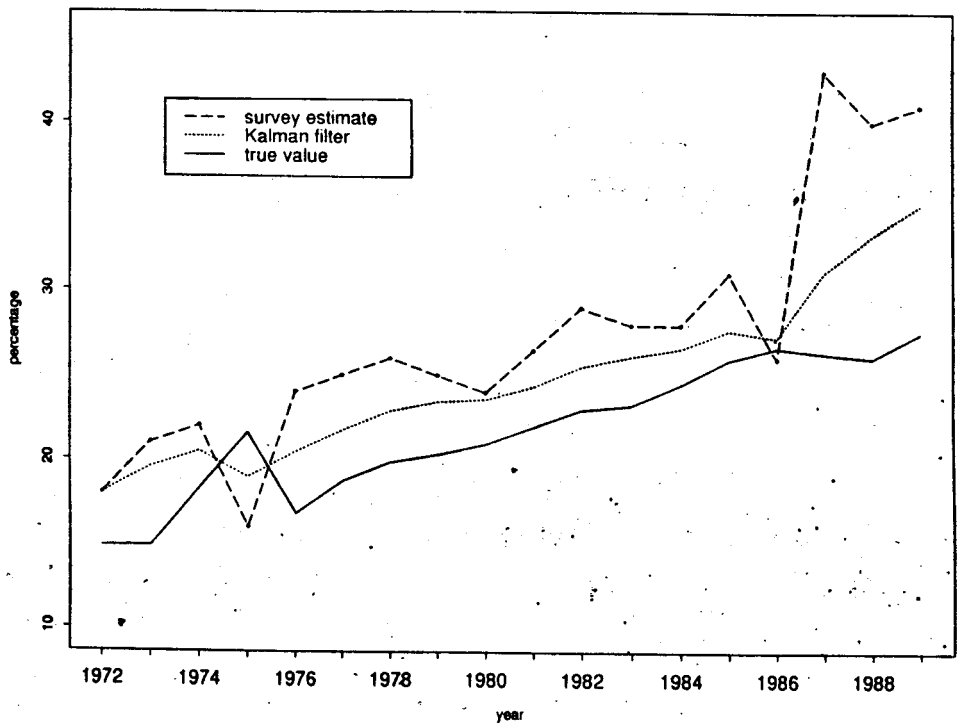

Figure 4 Percentage of black homes with seven or more residents