Combination of sample surveys or projections of political opinions

Daniel Thorburn[1] and Can Tongur[2]

[1]Stockholm University, e-mail: Daniel.Thorburn@stat.su.se
[2]Stockholm University and Statistics Sweden, e-mail Can.Tongur@scb.se

Abstract

In Sweden sample surveys of the party preferences are made almost every month by different institutes. The sample sizes are usually between 1000 and 2000 which means that the standard deviations are between 1 and 1.5 %. We study how these estimates can be combined to get better estimates taking the trends and voter mobility into account. Our model is a combination of a dynamic model based on Wienerprocesses and sampling theory with design effects. Since the party preferences are modelled as random processes it will be possible also to talk about the probability for events like a party (block) has more than 50 % of the preferences. Assuming that the same model and the same parameters will hold also in the future we can also give intervals for the future election results. short abstract is nice at the beginning of the text to describe the contents and main results of the paper.
Keywords: Dynamic models, Metaanalysis, Party preferences, Poll of polls, Sample surveys, Wiener processes

1 Introduction

In Sweden many different institutes make opinion polls almost every month . Four important private actors are SIFO, Temo, SKOP and Novus. The sample sizes are usually between 1000 and 2000. Statistics Sweden makes a poll twice a year with a sample size of about 7000. We study how these estimates can be combined to get better estimates taking the trends and the voter mobility into account. Our approach is a combination of a dynamic model based on Wienerprocesses (West & Harrison, 1997) and sampling theory with design effects. Assuming that the same parameters will hold also in the future it will be possible to give prediction intervals for the upcoming election results.

A recently presented method to weight together previously presented polls is "Poll of polls". A simple description is given by Salmond.(2012). A more detailed description is given by Silver (2008). His basic idea is to weight the previous polls in order to get the best estimate of the present opinion. Silver (2008) also discussed estimation of trends. He used a simple Gibbs sampling technique based on regression analysis to forecast the outcome of the next election. We suggest the use of a random walk process as the underlying model to avoid the rigidity of linear models.

2 Simple Basic Model

In our basic model the proportion of people supporting a party/party block preferences behaves like a random walk over time. This is a mathematical model and does not explicitly take all available extra information into account. For example, will the time when a party changes its leader be modeled as taking place at an unknown time in the future random which cannot be exactly predicted. In the same way all other important events like financial crises, sexual scandals, political debates or special campaigns are viewed as random events which cannot be predicted neither the time nor its effects. The model uses only the observed data and does not use subjective but generally held beliefs as the observation that the opinion usually swings against the sitting government in the middle of an parliamentary term to go back when the election gets closer.

In this random walk model it will be equally likely that a party decreases or increases from the present level. What happens will depend on future unknown events. A specialist in political science can very likely improve on the model by using his expert knowledge. Our aim, however, is making estimates and predictions using only observed polls. The proportion of a party/party block at time t is denoted P_t and its development is modelled by a random process given by the stochastic differential equation

$$dP_t = \gamma \, dW_2(t),$$

where $W_2(t)$ is white noise and contains all the small and large things that affect the voters' opinions. The solution to this equation says that if a party at a certain day t has the proportion P_t, than the distribution s days ahead will be normally distributed around P_t and with the variance γs (and thus standard error $\sqrt{(\gamma s)}$). The proportion is usually unknown but measured with a random error by opinion polls and exactly by general elections.

Suppose that the model says that P_t is normal with mean EP_t and variance CPP_t at a certain time. Suppose further that at time t the result X_t of an opinion poll becomes available measuring P_t with the variance V_t. Combining the prior belief with the observed proportion the best guess of the true proportion P_t is that it is normally distributed with a mean that is weighted between the prior mean and the observed poll. The weights should be inversely proportional to the variances

$$\frac{V_t * EP_t + CPP_t * X_t}{V_t + CPP_t}. \tag{1}$$

The new variance is given by the smaller value

$$\frac{V_t * CPP_t}{V_t + CPP_t}. \tag{2}$$

This model is very similar to the models used by financial experts modelling prices at the Stock Exchange. There is a lot of information around but most of the effects are already capitalised in the prices and it is difficult to predict the future development of stock prices. It may be possible but the fact is that only very few persons succeed in making a fortune on the stock market shows that most of the information is already reflected in the prices. It may also be viewed as a Dynamic Linear Model using Kalman filters (West & Harrison, 1997).

3 A Model with Trend, Forecasting s Periods Ahead

3.1 Background

Here we will introduce a trend into the simple model above. Thus the best predictor of the future given only the previous polls was earlier the same as the best estimate of the present level. One might argue that there can be a trend. If a party has been steadily increasing its share of the voters for the last period one could expect that it would continue to do so. To study this, a model with trend is introduced. But the trend cannot be going on forever so the chosen model says that the trend is expected to decrease gradually and be replaced by other trends. The model contains three parameters, γ, which measures the size of the short term fluctuations, α, which measure how fast the trend disappears (e.g. = 0.05 means half the trend will have disappeared after roughly 1/0.05=20 days) and β, which measure how much randomness is explained by the (changing) trend.

3.2 Formulas

Let P_t, as before, be the true level of the sympathy for a party (party block) and let T_t be the trend in P_t.

Assume that the trend behaves like Ornstein-Uhlenberck process, i.e. it follows the stohastic differential equation

$$dT_t = -\alpha T_t dt + \beta dW_1(t),$$

where dW_1 is white noise and α and β are positive constants. Solving and expressing T_{t+s} in terms of T_t gives

$$T_{t+s} = \exp(-\alpha s)T_t + \beta\int_t^{y+s} \exp(-\alpha(t+s-u))dw_1(U). \tag{3}$$

The party's share of the votes has a drift proportional to T_t and a constant noise term at time t, i.e. it follows the stochastic differential equation

$$dP_t = T_t + \gamma\, dW_2(t)$$

where dW_2 is another independent white noise and γ a positive constant. Solving and expressing the future party sympathy in terms of the situation at time t gives

$$P_{t+s} = P_t + \int_t^{t+s} T_u du + \gamma\int_t^{t+s} dW_2(u) \tag{4}$$

If the expression (3) for T_t is inserted we get that P_{t+s} equals

$$P_t + \int_t^{t+s} (\exp(-\alpha(u-t)T_t) + \beta\int_t^u \exp(-\alpha(u-v))dW_1(v))du + \int_t^{t+s} dW_2(u) \tag{4}$$

The first part of the middle term is easily computed. For the second part we change the order of integration and after that compute the (new) inner integral. The result is that P_{t+s} equals

$$P_t + ((1+\exp(-\alpha s)/\alpha)T_t + \beta\int_t^{t+s} (1-\exp(\alpha(t+s-v)))/\alpha + dW_1(v) + \gamma\int_t^{t+s} dW_2(u) \tag{5}$$

The next problem is to update the expected values and variances when a new opinion poll is observed. Before the observation at time t the expected levels are denoted EP_t and ET_t and the variances CPP_t,

CTT_t_and CPT_t. After observing X_t from a poll with precision $Var(X_t - P_t) = V_t$, the prior is updated by combining the observation and the prior getting a posterior. The expected levels are updated exactly as in Formulas (1) and (2).

$$E(P_t \mid X_t) = \frac{V_t * EP_t + CPP_t * X_t}{V_t + CPP_t}.$$

with the variance

$$Var(P_t \mid X_t) = \frac{V_t * CPP_t}{V_t + CPP_t}.$$

Updating the distribution of the trend is also fairly simple.

4 Other Features

4.1 Design Effects

Even though the four Swedish institutes use slightly different techniques, the methods are fairly similar. There are also studies based on web surveys in Sweden but we will not use any of them. Their results are not completely comparable to those based on probability sampling. The party preference study of Statistics Sweden is based on a simple random sample from the Swedish population register and is performed by telephone interviewing with about 30% non-response. About half of the non response consists of refusals and the other half are persons who are not found (e.g. not at home or no telephone number found). The studies by the private institutes are based on some version of RDD and telephone interviewing. All institutes use some sort of weighting to decrease the variance. Statistics Sweden calibrates with some known register variables. Most institutes ask about the voting at last election and weights to some extent with e.g age, sex and the outcome at that election. This means that there probably is a design effect and that the variance is smaller than what the binomial distribution formula says ($1/(nP_t (1-P_t))$).

We will assume that the design effects are the same for all institute gives

$$\begin{aligned} Var(X_t - P_t \mid P_t) &= V_t \\ &= \delta P_t(1-P_t)/n \end{aligned} \tag{4}$$

(The data does not allow us to efficiently estimate the different design effects without introducing an informative prior). The institutes have also different ways of formulating the interview questions. Thus we allow them to have different biases.

4.2 Parameter Estimation

The model described above contains many parameters, α, β, γ, δ and the institute effects. The model also contains starting values, i.e. the means EP_0 and ET_0 and the variances CPP_0,$CTT_0$and CPT_0. Since the data start at an election with the outcome X_0, the choice of starting values are natural. As we are going to study a longer period our solutions will be quite robust against miss-specified starting values.

In order to estimate the parameters owe maximise the likelihood function. Since each observation of X_t - EP_t given the history (i.e. all observations before time t) is approximately normal, the following likelihood can be used

$$L(\alpha,\beta,\gamma,\delta) = \prod_t (1/((2\pi(V_t+CPP_t))^{1/2})) \exp(-(((X_t-EP_t)^2)/(2(V_t+CPP_t)))),$$

where the product is over all opinion polls. Taking logarithms this becomes.

$$-2l(\alpha,\beta,\gamma,\delta) = \sum_t ((X_t - EP_t)^2)/(V_t + CPP_t) - \ln(V_t + CPP_t)$$

Note that the last term is important since it depends on the parameters. This expression can also be used for testing hypothesis on the parameters.

4.3 Variance Stabilising Transformation

Whenever the proportion of voters for a certain party or party block is relatively large, it may be reasonable to assume that its development can be described by the processes described above. However, when the proportions are close to zero (or one) the process may reach zero and the forecasted proportions may become negative. To cope with this situation, which is likely to occur in an election system having several smaller parties, we use a logistic transformation

$$Q_t = \ln(P_t) - \ln(1-P_t)$$

and assume that Q_t will follow the above processes. The logistic transformation has also the advantage that the process will never become larger than 1. Additionally, this model is symmetric in the sense that it can be used both for the party size and also the complimentary event of not voting for this party (with opposite sign).

If X_t is the result of a poll at time t with mean P_t and variance V_t, then $Y_t = \ln(X_t) - \ln(1-X_t)$ is a an estimator with approximate mean Q_t and approximate variance $V_t/(P_t(1-P_t))^2$. If the variance is given by (6) this simplifies to

$$Var(Y_t - Q_t) = \delta/(nP_t(1-P_t))$$

Thus we can use the same formulas as above for this process but use this variance expression in the innovation formulas.

5 Data

5.1 Parties in Sweden

The parties in the Swedish Parliament can be divided int0 two blocks. The Bourgeois block called the Alliance forms the government since 2006 and comprises the conservatives, "Moderaterna" (M), the christan democrats (Kd), the liberal party, "Folkpartiet" (Fp), and the centre party (C). The opposition consists of the three red or green parties; They are the green party, "Miljöpartiet" (Mp), Labour, "Socialdemokraterna" (S), and the left party, "Vänstern" (V). During half of the last parliamentary trem these three term formed a coalition with the object of winning the election in 2010 and forming the government together. However they did not gain power and now they act as three separate parties. However, it is still common among political commentators, to compare their total size to that of the alliance. These seven parties were represented in in the parliament in 2006. In 2010 a new populist party, Sverigedemokraterna (Sd), took seats in the parliament but does not belong to any of the two blocks. There exist also some other small parties in Sweden, which are not represented in the parliament.

5.2 The Data Set

For the analysis in this paper we have used all Swedish party preference studies from October 2010 until beginning of March 2012, altogether some 258 observations from five institutes. We have focused on the four established private institutes Novus, SIFO, Skop and Temo/Synovate/Ipsos but also Statistics Sweden's opinion polls as well, together with the election result of 2010. The data are from the public home page of the Novus group (2012). In this paper we focus on the proportion of Bourgeois voters of all parties in the parliament at that time.

The fit of the model is shown in the first graph. Here the results of all opinion studies are shown together with 95 % prediction intervals given all the previous studies. (When two studies have the same reference date only the interval for the last study is shown). The raggedness of the boundaries depends partly on the fact that the studies had different sizes and that the prediction intervals become shorter for large studies. Another reason is that each time that new information is added the predicted level changes with a small jump.

6 Estimation of the Party Preferences

The previous discussion concerned the prediction of the opinion polls given all previous data. The intervals were ragged due to the different sizes of the polls. The second graph shows the same period but now we estimate the true level for the period and use both old and future measurements.

Even though our goal with this project was meta-analysis, the model can be used for mechanical projections. The last graph shows forecasts for the election in 2010 at different times after the election in 2006, given the polls that have been published at that data.

References

Eklund J. & Järnbert, M. (2011), The Party Preference Study (PSU) – Description of the Statistics, ME0201, (in Swedish), SCB, Stockholm.

Novus Group, (2012), All Swedish Opinion Polls, to be found at http://www.novusgroup.se/vaeljaropinionen/samtliga-svenska-vaeljarbarometrar, (apr 2012)

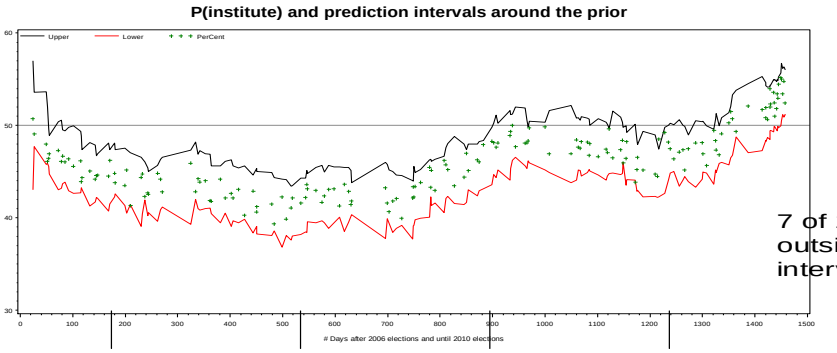
Salmond, R., (2012), Pundit Poll of Polls: How we do it, http://pundit.co.nz/content/poll-of-polls, (feb 2012), *Pundit*

Silver, N., (2008), Politics done right, http://www.fivethirtyeight.com/2008/03/frequently-asked-questions-last-revised.html , (feb 2012) *Fivethirtyeight*

West, M. & Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, 2nd edit, Springer

Figure 1 The alliance, prediction of polls

(based on data before the poll, interelection period)

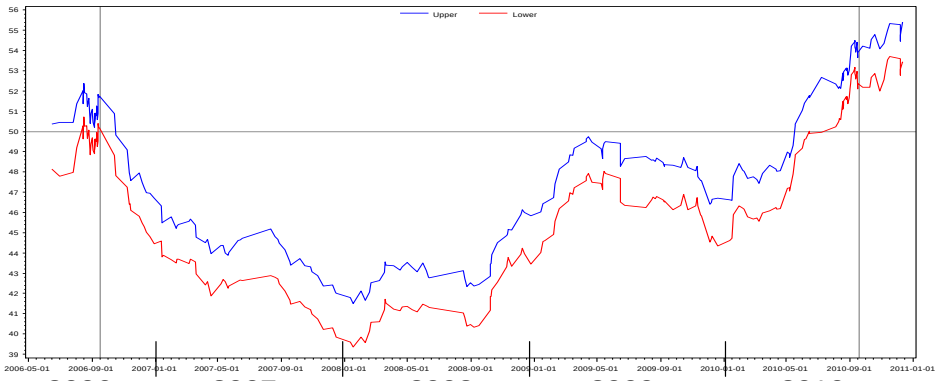


7 of 175 outside the interval

2006 2007 2008 2009 2010

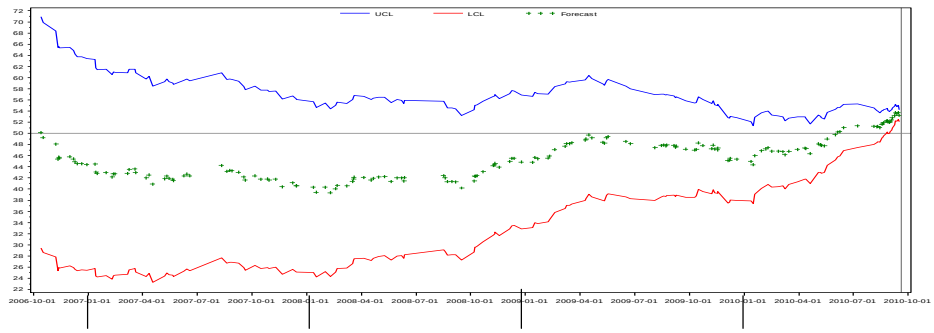
Estimates for the Alliance during last interelection period

All data, 95 % probability intervals, excl. minor parties and Sd



2006 2007 2008 2009 2010

Figure 3 Forecasts for the 2010 election, based on previous polls



2006 2007 2008 2009 2010