Forecasting Election Results: A Bayesian Frequentist Comparison

Erik Oldehed

Bachelor Thesis in Statistics, 15 ECTS

Lund University

2019

Supervisor: Jakob Bergman

LUND UNIVERSITY

STAH11, 1–8 (2019)

Forecasting Election Results: A Bayesian Frequentist Comparison

By ERIK OLDEHED*, *Department of Statistics, Lund University, Lund, Sweden*

(Received April 3, 2019)

ABSTRACT

We present a Bayesian and frequentist comparison when forecasting elections through polls. Our focus is on studying the differences of these approaches in forecasting elections. An evaluation of the fit is performed using the odds ratio. We propose a frequentist methodology for prediction horizons three months ahead while a Bayesian methodology may be slightly more accurate for shorter prediction horizons. The contribution of this paper lies in shedding light on the importance of the prediction horizon when choosing between a Bayesian or frequentist methodology to forecasting election results.

Keywords: Bayesian forecasting, frequentist forecasting, non-homogeneous hidden Markov models, autoregression, kernel smoothing.

1. Introduction

One of the hallmarks of modern democracy is that of competitive elections (Walther, 2015). Thus, it is not surprising that the practice of election forecasting is gaining an increased amount of attention in the field of political science. One of the core differences of this field in relation to others in political science is that of the research question. As Walther (2015) describe it; *"In that sense, the question 'how' rather than 'why' is in focus"*.

Historically, a more fundamental approach to the subject spurred a widespread literature regarding economics as the driving force behind electoral studies. Perhaps the words of Tufte, 1978 exemplify this best:

When you think economics, think elections;

When you think elections, think economics.

Opponents to these structural techniques argue that it is necessary to include polling data to accurately predict election outcomes in a multiparty context (Walther, 2015). Forecasting in this context has received scarce attention in the literature.

Although there are a few studies evaluating different forecasting methods, most of these focus on Bayesian modelling. A comparison of a Bayesian framework to a frequentist approach in election forecasting is not known to the author.

The fundamentals of Bayesian theory focus on deriving a probability distribution given information already at hand. This way of reasoning is particularly interesting in election forecasting since one can incorporate information about which party voters would vote for at a given time. The question, whether or not prior information holds any significant advantage over an analysis based on frequencies in election forecasting is not entirely clear. The contribution of this paper lies in shedding light on the importance of the prediction horizon when choosing between a Bayesian or frequentist methodology to forecasting election results.

This paper starts with a description of the data being used and methods for dealing with missing observations. Notation and basic concepts are introduced after which the actual models are specified. Furthermore, the fit of the models are presented after which some concluding remarks on the implications of the results are made.

2. Data

2.1. Swedish Polls

Data containing 62 opinion polls for the 2018 Swedish election are collected through Novus (Table 1). The mean date when these observations was collected ranges from 2018–01–05 to 2018–09–01. We split the data set into a training set and a test data set. The training set ranges from 2018–01–09 to 2018–06–07. The test data set span from 2018–06–11 to 2018–09–09 (election day). The polling organisations collecting these data are Demoskop, Ipsos, Novus, SCB, Sifo and Skop. The survey methods used for this data set are panel sample, random sample or a combination of these. We exclude self-selection web panels since these are not random. These polls together contain a total of 200667 observations of surveyed respondents out of which 95304 are random sample.

* Corresponding author.
e-mail: er0227ol-s@student.lu.se

Table 1. Description of data

Sampling Method	*Opinion Polls*
Panel	10
Random	12
Combination	40
Total	62

2.2. Missing Observations

An often encountered difficulty in time series analysis is that concerning missing observations. For some estimation methods it is not possible to derive parameter values with missing observations. As Anderson (1957) notes, maximum likelihood estimation of mean, variance and correlation is problematic to estimate since the likelihood function is conditional on the data. However, there are many settings in which the measurements are irregularly spaced (Sheather, 2009). One solution for this problem is to use kernel smoothing techniques or penalised linear regression splines. These methods often yield similar results (Hastie et al., 2009). For our purpose kernel smoothing is used to interpolate measurements into equally spaced time units.

3. Notation and Basic Concepts

3.1. Smoothing

Kernel smoothing can be described visually as a technique to derive a smooth line from a noisy one. It does so by estimating a weighted average of data points near a given point t_0. This procedure continues until the desired number of data points are estimated and results in a smooth estimate that has less variability. For our purpose we keep the same number of smoothed data points as points in the original series. This result in 62 observations for the smoothed series out of which 40 are used for the training data set.

Kernel smoothing is a regression technique where $\hat{f}(t)$ denotes the estimated regression function (Hastie et al., 2009). The kernel $K_\lambda(t_0, t_i)$ acts as a weighting function where the value at t_i is weighed according to its distance to t_0. For our purpose a Gaussian kernel is specified as follows by eq. 1 (Sheather, 2009).

$$K_\lambda(t_0, t_i) = \frac{1}{\lambda\sqrt{2\pi}} \exp\left(-\frac{(t_0 - t_i)^2}{2\lambda^2}\right) \tag{1}$$

To perform kernel smoothing one needs to choose a bandwidth. The choice of bandwidth λ representing the width of the kernel has a large impact on the fit (Sheather, 2009). A larger value of λ results in averages over more observations and a larger bias (Hastie et al., 2009). One commonly used method for obtaining an optimal bandwidth is the plug-in method (Sheather, 2009). We use the plug-in method for selecting bandwidth according to Ruppert et al. (1995) as well as graphical inspection for our kernel smoother.

Since a formal definition of the plugin method for selecting a bandwidth $\lambda = h$ as performed by Ruppert et al. (1995) is rather technical we merely give a general description of this methodology. This method revolves around choosing a λ such that the discrepancy between the estimated regression function $\hat{f}_h$ and the actual function f is minimised. However, the underlying curvature of the actual function f is not known and is estimated using kernel density estimates where a normality assumption on f is the basis for the bandwidth.

Once a value of λ is chosen a weighting method needs to be selected. There are many weighting methods for kernel estimation where Nadaraya–Watson is one of the most renown. Another widely used weighting method is local linear regression. This method derives these weights by, for each target point t_0, solving a weighted least square problem as follows by eq. 2, where T denotes the number of observed points (Hastie et al., 2009).

$$\min_{\alpha(t_0),\beta(t_0)} \sum_{i=1}^{T} K_\lambda(t_0, t_i)[y_i - \alpha(t_0) - \beta(t_0)t_i]^2 \tag{2}$$

By solving eq. 2 an estimate can be inferred such as $\hat{f}(t_0) = \hat{\alpha}(t_0) + \hat{\beta}(t_0)t_0$ (Hastie et al., 2009). This estimate is less prone to bias than the classic Nadaraya–Watson method since it automatically corrects for bias of the first order. By performing a Taylor expansion it can be shown that the bias only depends on quadratic or higher-order terms. Hence, a local linear regression approach with a Gaussian kernel is used to smooth the data. Smoothing is performed on all measurements used in our analysis in a similar manner as described in eq. 2.

3.2. Operationalization

We define the left block as Socialdemokraterna (Social Democratic Party), Vänsterpartiet (Left Party) and Miljöpartiet (Green Party). Four variables are used for our estimates of voters in favour of the left block throughout time (Table 2).

Our first measure is given by the combined percentage $\tilde{x}$ in favour of the left block as estimated by the polling organisations. It is worth noting however that these organisations weighs polling data such that $\tilde{x}$ is only an estimate of the relative number of individuals that would vote for the left block at a given time rather than the actual fraction of votes measured.

Our second measure is an estimate of the number of interviewed persons y that would vote for the left block at a certain time. The number of individuals in favour of the left block y is estimated by multiplying the percentage in favour of the left block $\tilde{x}$ by the total number of interviewed persons n. Our estimates of y are rounded to the nearest integer.

The time t at which a survey is conducted is measured such as the mean date of the sample period. The mean date when these observations was collected ranges from 2018–01–05 to 2018–09–01. The training set ranges from 2018–01–09 to 2018–06–07. Since smoothing did not result in a measurement for time

2018–06–09 we used 2018–06–07 as an endpoint for the training set. The test data set span from 2018–06–11 to 2018–09–09 (election day).

Table 2. Description of Variables

$Variable$	$Description$
y	The estimated voters favouring the left block
$\tilde{x}$	The estimated percentage favouring the left block
n	The number of interviews
t	The mean date at which data was collected

Note: Each of the variables y, $\tilde{x}$, and n are measured at over a time period with mean t. Smoothing is performed to achieve equal spacing between observations.

3.3. Stochastic Processes

A stochastic process may be defined as a sequence of random variables (Cryer and Chan, 2008). An autoregressive process of order one is in turn a stochastic process where its future state depends only upon its current state and an error term. Let $\{X_t\}$, $t = 1, 2, \ldots, T$ be a stochastic process for which the conditional probability distribution of future states depends only on the current state (Ching et al., 2006). The current state of X_t is a linear combination of the state at time $t-1$ and an innovation term e_t incorporating new information in the series at time t not explained by past values (Cryer and Chan, 2008). Since the error term e_t for every t is independent of $X_{t-1}, X_{t-2}, \ldots$ it is referred to as an innovation term. We define this process as expressed in eq. 3,

$$X_t = \beta X_{t-1} + e_t, \tag{3}$$

where e_t is assumed to be an independent and identically distributed (i.i.d.) white noise process with $\mathbb{E}(e_t) = 0$ and $\mathbb{V}(e_t) = \sigma_e^2$. The stochastic process X_t is said to be stationary if $|\beta| < 1$.

3.4. Markov Chains

A Markov chain may be defined as a sequence $X_1, X_2, \ldots$ of random elements of a set, if the conditional distribution of X_{t+1} given $X_1, \ldots, X_t$ depends on X_t only (Brooks et al., 2011). A state space is then defined such as the set M in which X_i takes values. In other words, the conditional distribution of any future state given the past states and present state is independent of the past states and depends on the present state only (Ching et al., 2006). A formal definition may be written such as,
$\mathbb{P}(X_{t+1} = i | X_t = j, X_{t-1} = i_{t-1}, \ldots, X_0 = i_0) = P_{ij}$ given that $t \geqslant 0$, where $i, j, i_0, i_1, \ldots, i_{t-1} \in M$. That is, the conditional probability of X at time period $t+1$ is only depending on its state at time period t. This is referred to as the Markov property. By this definition we can see that an autoregressive process of order one satisfies the conditions for a Markov chain.

A transition matrix can be defined as the probability $P_{ij} = \mathbb{P}(X_{t+1} = i | X_t = j)$ that a process in state j will be in state i after one transition (Ching et al., 2006). The n-step probability matrix is denoted $P_{ij}^{(n)}$. A stationary or time homogeneous transition matrix is one that does not depend on t. A non-homogeneous Markov chain has a transition matrix that is non-stationary in time (Meligkotsidou and Dellaportas, 2011).

3.5. Non-Homogeneous Hidden Markov Models

A hidden Markov model (HMM) is a model where a bivariate process $\{(Y_t, X_t)\}$ is the basis of a data generation mechanism (Meligkotsidou and Dellaportas, 2011). We let X_t be an unobservable, finite Markov chain that governs the distribution of the observable process $\{Y_t\}$. In a standard settings a HMM has one transition matrix

As described by Meligkotsidou and Dellaportas (2011), in the non-homogeneous case we assume that in addition to X_t, Y_t also depend on a set of exogenous covariates that are observable up to time $t-1$. Furthermore we assume that the hidden process $\{X_t\}$ is a non-homogeneous Markov chain with time-varying transition matrix. A more general explanation may be that the voting sentiment changes through time, where a model with a fixed parameter might be limited in its applicability.

3.6. Monte Carlo Sampling

Monte Carlo has progressed from being the most famous casino in the world in the 50s to a technical term for simulation of random processes (Brooks et al., 2011). Consider that one wants to estimate the expectation of a function $\mu = \mathbb{E}\{g(X)\}$ although an analytic solution is not possible. Simulating a number, n, of random i.i.d. variables possessing the same distribution as X one can perform a Monte Carlo approximation of μ such as $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$.

3.7. MCMC

Markov chain Monte Carlo (MCMC) is essentially merging Markov chain simulation with Monte Carlo methods. One can describe it as a special case of Markov chain simulation (Gelman and Hill, 2007). The fundamental idea is to draw values of θ from approximate distributions and then correct those draws to better approximate the target posterior distribution $\mathbb{P}(\theta|y)$. In that sense, these approximations are improving at each step in the simulation, converging to the target distribution.

Numerous algorithms have been proposed for MCMC sampling (Hoffman and Gelman, 2014). Some of the most commonly mentioned algorithms are Hamiltonian, Gibbs and the Metropolis-Hastings. What sets the Hamiltonian algorithm apart from Gibbs and Metropolis-Hastings is that it does not share their tendency to explore the parameter space via inefficient random walks. The downside of this method is that it is demanding to calculate and needs two parameters, a step size ϵ and a desired number of steps L to be specified in advance, where a poorly chosen value will cause a decreased efficiency.

The standard algorithm for performing sampling with RStan (an r package developed by Stan Development Team) is the No-U-Turn Sampler (NUTS) algorithm which is a development of the Hamiltonian algorithm (Hoffman and Gelman, 2014; R Core Team, 2013; Stan Development Team, 2018). Since this sampling method by automatic adaption requires no specified parameters in advance we rely on the NUTS algorithm.

4. Models

4.1. Frequentist Model

We refer to the frequentist model as Model I. An autoregressive model of order one is specified for comparison with the Bayesian model. We let $\tilde{X}_t$ be the smoothed time series obtained by the procedure described in section 3.1. where $\tilde{X}_t$ represents the percentage in favour of the left block at time period t. The index $t = 1, 2, \ldots, 40$ is increasing by approximately 3.92 days increments spanning from 2018–01–09 to 2018–06–07. We specify a centred model according to eq. 4 as follows,

$$\tilde{X}_t - \tilde{\mu} = \tilde{\beta}(\tilde{X}_{t-1} - \tilde{\mu}) + \tilde{e}_t, \tag{4}$$

where $\tilde{\mu}$ is the expectation of the process $\tilde{X}_t$ and $\tilde{e}_t$ is an iid process with zero mean and $\mathbb{V}(\tilde{e}_t) = {\sigma_{\tilde{e}}}^2$. Rewritten we have that,

$$\tilde{X}_t = \tilde{\alpha} + \tilde{\beta}\tilde{X}_{t-1} + \tilde{e}_t, \tag{5}$$

where $\tilde{\alpha} = \tilde{\mu}(1 - \tilde{\beta})$.

4.2. Bayesian Model

The Bayesian model is referred to as Model II. We let Y_t be an estimate of the observed number of people that would vote for the left block at time period t. The number of interviewed persons n_t is known and we also have an estimate of the proportion P_t in favour of the left block. The distribution of Y_t can then be derived as a consequence of the following stochastic relationship:

$$Y_t \sim Bin(n_t, P_t), \tag{6}$$

where t denotes the time at $t = 1, 2, \ldots 40$ equally spaced smoothed measurements from 2018–01–09 to 2018–06–07. The model can then be specified in terms of a latent unobservable process, a constant and an error term such as,

$$g(P_t) = \alpha + X_t + Z_t, \tag{7}$$

where α is a constant, X_t is the hidden process we would like to estimate during time period t, Z_t is a parameter controlling for over-dispersion and $g(P_t)$ is the logit link of P_t. Over-dispersion is a way of incorporating a larger variance than is accounted for by the model (McElreath, 2015). When variation in counts exceeds what would be expected from a binomial process one can control for this by allowing differences in constant terms for each observation. A time varying beta was chosen upon comparing the fit of this model with one where the coefficient was constant. Sigma is modelled using a standard half Cauchy prior, denoted HC. A Cauchy distribution constrained to positive values is referred to as a half Cauchy and is often preferred as a prior for scale parameters (Gelman, 2006). Priors are chosen as,
$\sigma_Z \sim HC(0, 1)$,
$\sigma_X \sim HC(0, 1)$,
$\alpha \sim N(0, 1)$,
$\beta \sim N(0, 1)$,
$Z \sim N(0, \sigma_Z)$,
$X_t \sim N(\beta_t X_{t-1}, \sigma_X)$,
where σ_Z is the standard deviation of the over-dispersion term Z_t, σ_X is the standard deviation of the hidden process X_t and β_t is a time varying coefficient.

Knowing that the log odds ratio of P_t is equal to $\alpha + X_t + Z_t$,

$$g(P_t) = \log\left(\frac{P_t}{1 - P_t}\right), \tag{8}$$

one can then traverse from $g(P_t)$ to P_t by logit link such that,

$$P_t = g^{-1}(\alpha + X_t + Z_t) = \frac{e^{\alpha + X_t + Z_t}}{1 + e^{\alpha + X_t + Z_t}}. \tag{9}$$

4.3. Measures of Predictive Accuracy

A reasonable way to evaluate a model is by the accuracy of its predictions. As noted by Gelman et al. (2014), this accuracy may be valued in its own right, such as when evaluating a forecast. In a different setting this accuracy may not be of interest in isolation but rather in relation to another model.

In this study, predictive accuracy is of interest for its own sake but more importantly for comparison of a Bayesian and frequentist methodology to election forecasting. A relatively recent approach is to use the odds ratio of a poll compared to the actual outcome (Jennings and Wlezien, 2018; Arzheimer and Evans, 2014; Martin et al., 2005). This ratio will be used both for evaluating the given model but also for comparisons. The odds-ratio of Model I is based on the predicted votes $\tilde{X}_t$ (eq. 10) at time period t while the odds of P_t (eq. 11) is the equivalent for Model II. These are compared to the actual votes v_t. Since the election day is two time units from the last smoothed t, we let $t = 64$ be the time at which we measure these models predictive accuracy.

$$A_t^{(1)} = \log\left(\frac{\tilde{X}_t}{1 - \tilde{X}_t}\frac{1 - v_t}{v_t}\right) \tag{10}$$

$$A_t^{(2)} = \log\left(\frac{P_t}{1 - P_t}\frac{1 - v_t}{v_t}\right) \tag{11}$$

5. Results

5.1. Smoothing

Upon smoothing the series using the plugin method for choosing bandwidth it appears that a large part of the variance of the original series disappeared (Figure 1). A value for λ of approximately 30.8 was obtained by this method. By choosing a lower value, setting λ to 4 a larger part of the original series was preserved (Figure 2). Since our aim is to model a latent process hidden in noise a series that is too smooth is not desirable. One could argue that the smooth series using $\lambda = 30.8$ already is an estimate of an underlying process and therefore nothing is left to model. When fitting the models to the smoothed series using the plugin method a rather poor fit was achieved which coincides with our suspicions. For this reason we rely on a $\lambda = 4$, chosen by eye instead.

5.2. Frequentist Model

An autoregressive model of order one was fitted to the smoothed series in Figure 2 using maximum likelihood estimation (Table 3). Both the coefficient, $\hat{\beta}$ and the mean of the process, $\tilde{\mu}$ are significant at a 0.05 level as suggested by the confidence intervals. The constant $\tilde{\alpha}$ is approximately 0.04. We see that $\tilde{\beta}$ is positive and takes on a large value although not close to one which implies that this process is stationary. Upon studying the auto-correlations of the residuals we found that these for most of the part are within the confidence interval which is another indication of a stationary process.

The predictions for Model I are presented in Figure 3. The smoothed series is represented by empty dots before the prediction period and crosses during the prediction period while the predicted values are marked by filled in dots. The election result is marked by an asterisk. We see that this model copes remarkably well in predicting the election outcome. Although lacking the alternating pattern of the smoothed series it seem to model the overall direction of the trend quite well.

The predicted value at the time of the election is 0.388 while the actual result is $v = 0.407$. The predictive accuracy measure amounts to $A_{64}^{(1)} = -0.0781$. Rounded to two decimals this model predict an outcome of 0.39.

5.3. Bayesian Model

A non-homogeneous HMM was fitted to the smoothed series in Figure 2 using MCMC sampling (Table 4). The posterior interval around the intercept α implies that this estimation is significant at a 0.05 level. The estimate of σ_Z is approximately one 3rd of the equivalent for σ_X. This suggests that the variability of the estimated voters favouring the left block throughout time is greater than allowed by a model without an over-dispersion term. The estimates of β_t along with 0.80 and 0.95 posterior intervals are given i Figure 4. We can see that these estimates are rather similar until the date 2018–05–26 where β_t starts to increase until it peaks at 2018–06–07.

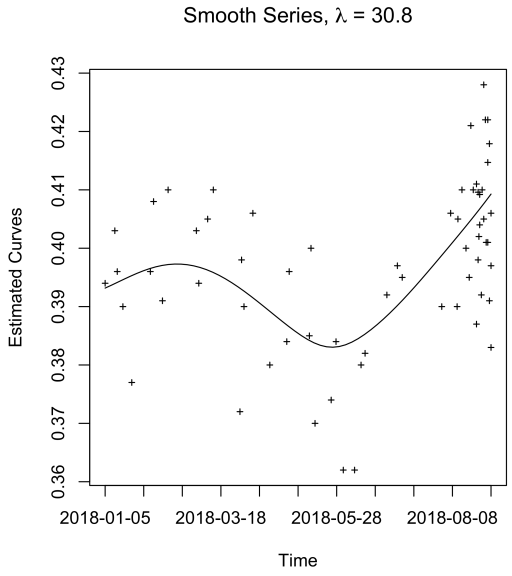

Fig. 1. Kernel smoothing using the plugin method, as applied to y_i, the estimated voters favouring the left block. An optimal bandwidth is obtained by minimising an approximation of the squared error of the true function and its estimator.

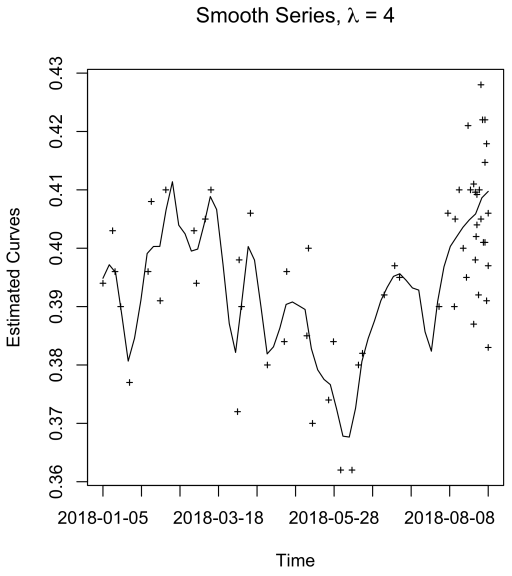

Fig. 2. Kernel smoothing by graphical inspection, as applied to y_i, the estimated voters favouring the left block. Bandwidth is selected by choosing the value preserving most variation in the original series while still yielding equally spaced data points.

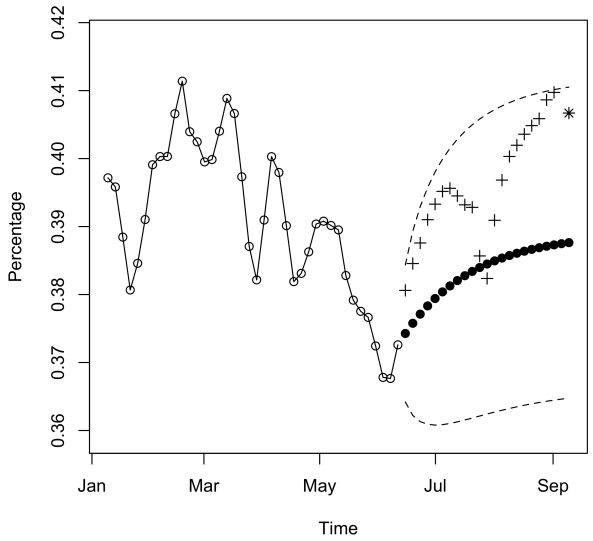

Fig. 3. Prediction, Model I. A prediction of the estimated voters favouring the left block using an autoregressive model of order one with a frequentist methodology.

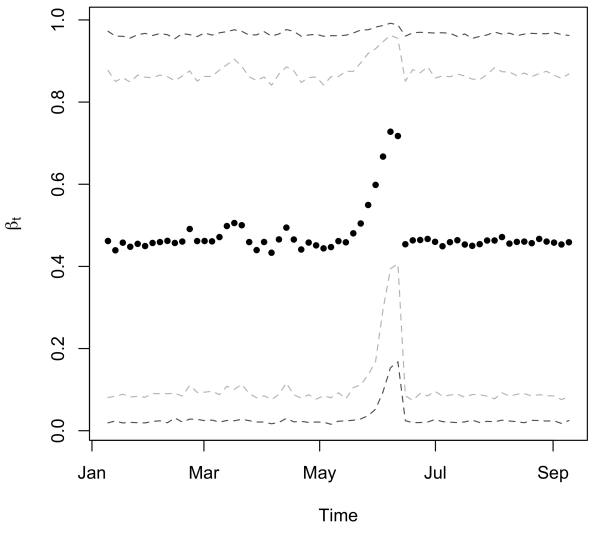

Fig. 4. Estimates of β_t, Model II. Each dot represent an estimate of the parameter β_t at time t with 0.80 (light grey) and 0.95 (dark grey) posterior intervals.

The predictions for Model II are presented in Figure 5. The smoothed series is represented by crosses while the latent trend is marked by empty dots before the prediction period and filled in dots during the prediction period. The election results is

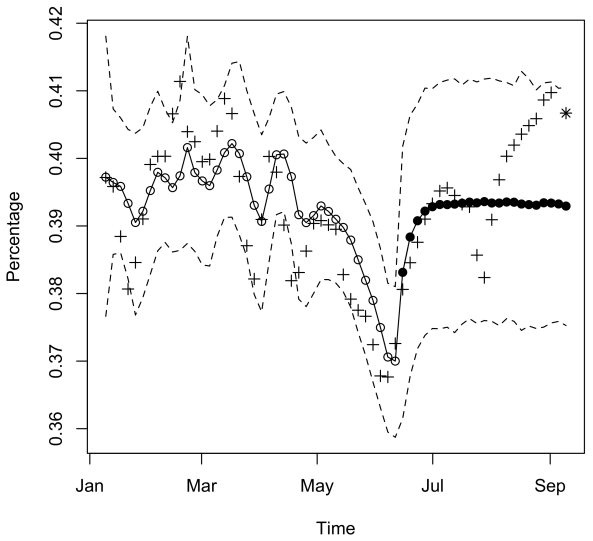

Fig. 5. Prediction, Model II. A prediction of the estimated votes favouring the left block using a non-homogeneous hidden Markov model with a Bayesian methodology.

Table 3. Estimates and standard errors for the constant term $\tilde{\mu}$ and the coefficient $\tilde{\beta}$ for Model I.

	$Estimate$	$S.E.$	LB	UB
$\tilde{\mu}$	0.389	0.007	0.375	0.403
$\tilde{\beta}$	0.898	0.074	0.769	1.028
$\sigma_{\tilde{e}}$	0.005			

Note: Estimates are presented with an, 0.95 confidence level, interval where LB denotes the lower bound and UB denotes the upper bound. The abbreviation $S.E.$ represent the standard error of the estimate.

Table 4. Estimates and standard errors of the estimators of the constant term α, σ_X and σ_Z for Model II.

	$Estimate$	$S.E.$	LB	UB
α	-0.433	0.000	-0.453	-0.415
σ_X	0.028	0.000	0.019	0.040
σ_Z	0.009	0.001	0.001	0.022

Note: Estimates are presented with a 0.95 posterior interval where LB denotes the lower bound and UB denotes the upper bound. The abbreviation $S.E.$ represent the standard error of the estimate.

marked by an asterisk. All of the smoothed observations fall within the 0.95 posterior interval during the prediction period. The first 9 estimates closely matches the smoothed series. Deviations become more substantial as time approaches election day. The predicted value at the time of the election is 0.393 while the actual result is $v = 0.407$. The predictive accuracy

measure amounts to $A_{64}^{(2)} = -0.0573$ which means that the more advanced Bayesian model does not underestimate the result as much as the simpler frequentist model does at the time of the election day. Rounded to two decimals this model predict an outcome of 0.39. Considering the magnitude of the predictions, these models do not differ substantially.

6. Conclusions

The goal of this paper has been to compare the predictive accuracy of Bayesian and frequentist modelling in a Swedish election setting.

First, a model was specified according to a frequentist framework. An autoregressive model of order one was fitted to a smoothed series of polling data stretching from January to June 2018. The model in question yielded a surprisingly good prediction three months ahead of the election day.

Secondly, a model was specified according to a Bayesian framework. A non-homogeneous hidden Markov model was fitted where a latent process was specified according to an autoregressive model of order one for a period between January and June 2018. An over-dispersion term was added to account for a larger variance than permitted by the model as well as a constant. A logit link was used for the overall model. This model produced an estimate closer to the election outcome than that of the frequentist model.

Comparing the simpler autoregressive model to a more advanced hidden Markov model we can conclude that differences are negligible when predicting election results three months ahead. These two models yield election day estimates that are almost identically. Although the Bayesian model underestimates the election outcome to a lesser extent than the frequentist counterpart, there is a small difference in the odds ratio between these models. Larger deviations from the smoothed values exist in the beginning of the prediction interval which implies that a more advanced Bayesian framework may be more accurate for shorter prediction horizons. Possibly a precision gain may be achieved for shorter prediction horizons while predictions far ahead in the future might not be very accurate regardless of methodology.

A valid question would be if there is a rationale for Bayesian modelling in election forecasting since these models are often more strenuous in terms of computational resources. Even though these results may not generalise to a wider setting they still spark an important controversy. Mainly, that of whether the effort in specifying a more advanced model considering prior beliefs is justified by the accuracy of the predictions. We make no such remark but merely note that prior information is not always a benefit. Perhaps, the main factor in deciding if a frequentist or Bayesian framework is preferred may be the prediction horizon. Further research may focus on comparing the effectiveness of a Bayesian and frequentist methodology in election forecasting for different prediction horizons.

References

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278):200–203.

Arzheimer, K. and Evans, J. (2014). A new multinomial accuracy measure for polling bias. *Political Analysis*, 22(1):31–44.

Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L., editors (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, Boca Raton, Florida.

Ching, W.-K., Ng, M. K., and Ching, W. (2006). *Markov Chains: Models, Algorithms and Applications (International Series in Operations Research & Management Science)*. Springer-Verlag, Berlin, Heidelberg.

Cryer, J. D. and Chan, K. S. (2008). *Time Series Analysis: With Applications in R*. Springer, New York.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.*, 1(3):515–534.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, New York.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2^{nd} edition.

Hoffman, M., D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1593–1623.

Jennings, W. and Wlezien, C. (2018). Election poll errors across time and space. *Nature Human Behaviour*, 2:276–283.

Martin, E. A., Traugott, M. W., and Kennedy, C. (2005). A Review and Proposal for a New Measure of Poll Accuracy. *Public Opinion Quarterly*, 69(3):342–369.

McElreath, R. (2015). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, Oakville, Canada.

Meligkotsidou, L. and Dellaportas, P. (2011). Forecasting with non-homogeneous hidden markov models. *Statistics and Computing*, 21(3):439–449.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruppert, D., Sheather, S., and P. Wand, M. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90:1257–1270.

Sheather, S. J. (2009). *A Modern Approach to Regression with R*. Springer, New York.

Stan Development Team (2018). RStan: the R interface to

Stan. R package version 2.18.2.

Tufte, E. R. (1978). *Political Control of the Economy*. Princeton University Press.

Walther, D. (2015). Picking the winner(s) : Forecasting elections in multiparty systems. *Electoral Studies*, 40:1–13.