Charles University

Faculty of Social Sciences
Institute of Economic Studies

MASTER THESIS

Forecasting Election Results in the Czech Republic

Author: **Bc. Kateřina Doskočilová**

Supervisor: **doc. PhDr. Tomáš Havránek Ph.D.**

Academic Year: **2018/2019**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, July 30, 2019

Signature

Acknowledgements

This thesis is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 681228. I would like to thank here to Pavol Boško who helped me with the data for the analysis. Next, I would like to thank to my supervisor Tomáš Havránek for his patience and useful comments throughout the writing process. Finally, I am also grateful for my work colleagues, who encouraged me and allowed me to take some time off for writing the thesis. Last but not least, I would also like to thank to my sister, for taking the time to go through the thesis to check for any spelling and grammar mistakes.

Abstract

In this thesis, a forecasting model for the 2017 legislative election in the Czech Republic is built. As the Czech Republic has a multi-party system, the outcomes of the model are the expected vote shares for each party. There are two types of forecasts calculated. Firstly, a poll-based forecast using a dynamic linear model and Kalman filter to weigh the information in the polls. Secondly, the prices on betting markets are translated into probabilistic forecasts for the expected vote shares. This is a novel approach as prediction markets were previously used to forecasts only the probabilities of winning an election. Finally, the two types of forecasts are combined into one and weighed by their variance. Comparing the forecasts, we conclude that the betting market is able to predict the exact vote shares the most accurately right before the election.

JEL Classification	C53, D72, C32
Keywords	forecasting, elections, dynamic linear model

Author's e-mail	kacka.doskocilova@gmail.com
Supervisor's e-mail	tomas.havranek@ies-prague.org

Abstrakt

V této práci je vytvořen model pro prognózu výsledku voleb do Poslanecké sněmovny České republiky v roce 2017. Protože v České republice je několik menších stran, výsledkem tohoto modelu jsou procentuální zisky každé strany. Spočítané jsou dva typy předpovědí. Zaprvé, odhad založený na průzkumech veřejného mínění pomocí dynamického lineárního modelu a Kalmanova filtru, který váží informace obsažené v jednotlivých průzkumech. Zadruhé, sázkové kurzy jsou převedeny na pravděpodobnostní předpověd získaného podílu hlasů. To je originální přístup, protože sázkové kurzy byly zatím využívány pouze k předpovědi pravděpodobnosti výhry. Nakonec jsou oba typy předpovědi zkombinované do jedné, vážené rozptylem. Ze srovnání předpovědí můžeme soudit, že sázkové kurzy jsou schopné určit procentuální zisky hlasů těsně před volbami nejpřesněji.

Klasifikace JEL	C53, D72, C32
Klíčová slova	předpovídání, volby, dynamický lineární model

E-mail autora	`kacka.doskocilova@gmail.com`
E-mail vedoucího práce	`tomas.havranek@ies-prague.org`

Contents

List of Tables

List of Figures

Acronyms

DLM Dynamic linear model

EMH Efficient market hypothesis

KF Kalman filter

MAE Mean average error

RMSE Root mean squared error

Master's Thesis Proposal

Author	Bc. Kateřina Doskočilová
Supervisor	doc. PhDr. Tomáš Havránek Ph.D.
Proposed topic	Forecasting Election Results in the Czech Republic

Motivation When predicting the results of upcoming elections, we genrally rely on pre-election surveys. However, these surveys do not usually predict the outcome of the actual elections, but they show what the results would be on the day of the survey. Therefore, events happening between the survey and the election itself are not accounted for. Also, there may be substantial uncertainty because of the undecided people and depending on the share of the undecided voters, this uncertainty may be very important. Another source of information is the so-called 'prediction markets', where people can trade the outcome of events. Building on the efficient market hypothesis, the assumption is that the prices on these prediction markets reflect the beliefs of traders over an unknown outcome and thus show the underlying probabilities of these outcomes. If the prices did not correspond to the beliefs, there would be a potential for profit, which would be used by a rational market player.

The prediction markets on their own have an advantage against the pre-election surveys as they include the information from the polls, which is public, but also any relevant unpublished information not known to the general public. Rothschild (2009) compares poll-based forecast and prices from a prediction market for the 2008 US Presidential election and concludes that the market prices provide more accurate prediction than the aggregate polls. Nevertheless, these markets are susceptible to speculative bubbles and self-reinforcement. A similar mindset of users of prediction markets may lead to biased opinions and taking the market odds as correct and not updating for new information may worsen the forecasts.

The forecast does not necessarily have to predict the actual outcome of the election, but rather assign a probability for each possible outcome. This maintains the underlying uncertainty in the prediction and shows the most likely outcome. The uncertainty then decreases as the election approaches and the forecasts become more precise as more information is revealed.

Hypotheses

Hypothesis #1: Aggregating numerous pre-election surveys improves the results in terms of predicting the outcome of elections as compared to looking only at one survey. In other words, the aggregated average performs better in the long-term than any individual survey for predicting the election results.

Hypothesis #2: Including soft information in the form of prices on betting markets improves the performance of the election forecast.

The first two hypotheses will be tested on a few of the most recent elections in the Czech Republic. The final model will then be used to predict the results of municipal elections in Prague taking place in October 2018.

Methodology The first step for building the forecasting model is to collect the data for polls and pre-election surveys from previous elections in the Czech Republic and rating them based on their historical performance. This rating will then be used to calculate the aggregate forecast. Each poll has its corresponding statistical error and therefore aggregating them into an average improves the precision of the predicted outcome as it decreases the overall error of the forecast. The information from the betting markets will be added to the model. The methodology will follow models used by David Rothschild (predictwise.com) and Nate Silver (fivethirtyeight.com).

The forecast is created in five steps:

1. Collection of polling data. Determining weights for each poll based on the historical accuracy, sample size, recentness, etc. Then calculating the weighted average.

2. Adjusting polls when necessary, depending on the type of the elections, the methodology of the poll conducted. Adjusting for trends and possible biases.

3. Combining polls with demographic and economic data. Determining regional and demographic differences in voters' preferences and evaluating the current economic situation. The forecast assumes that better economic situation favours the incumbent. Allocating the undecided voters. The demographics have decreasing weight in the forecast as the actual election approaches and more and more voters are decided as well as more polls are available.

4. Calculating forecast form prices on the betting market, which should include all available information in a given time. Combining this forecast with the averaged polls-only forecast and averaged polls-plus demographic data.

5. Accounting for uncertainty and simulating the election to get the probabilities. The uncertainty decreases towards the elections. There are three types of

errors included in the forecast. Firstly, the national error, which accounts for systematic bias in the polls across the country, it is based on the time until the elections, the number of undecided voters. Secondly, there is a demographic error, which considers common demographic factors such as religion, race, education and accounts for bias in the polls for these groups. And thirdly, the region-specific error which accounts for bias in polls in a given region as the elections are conducted by regions or voting districts.

Expected Contribution The aim of this thesis is to provide a general method for forecasting election results in the Czech Republic, by aggregating the information obtained by pre-election surveys as well as using the information contained in betting market prices about the beliefs of betters on the market. It has been shown the publication of forecasts for election may influence the public opinion and therefore these predictions might be partly self-fulfilling (Rothschild & Malhotra, 2014), which suggests that providing as accurate forecasts as possible is crucial.

The developed forecasting model can then be used to predict the outcome of future elections in the Czech Republic with all information available at the given time.

Outline

1. Motivation
2. Literature Review: Survey of the relevant literature on forecasting and prediction markets..
3. Methodology: Theoretical description of the forecasting model.
4. Data
5. Results
6. Conclusion

Core bibliography

Armstrong, Jon Scott, ed. Principles of forecasting: a handbook for researchers and practitioners. Vol. 30. Springer Science & Business Media, 2001.

Leigh, Andrew, and Justin Wolfers. "Competing approaches to forecasting elections: Economic models, opinion polling and prediction markets." Economic Record 82.258 (2006): 325-340.

Lewis-Beck, Michael S. "Election forecasting: principles and practice." The British Journal of Politics & International Relations7.2 (2005): 145-164.

Lewis-Beck, Michael S., and Ruth Dassonneville. "Forecasting elections in Europe: Synthetic models." Research & Politics 2.1 (2015): 2053168014565128.

Malkiel, Burton, G.. 2003. "The Efficient Market Hypothesis and Its Critics ." Journal of Economic Perspectives, 17(1): 59-82.

Rothschild, David. "Forecasting elections: Comparing prediction markets, polls, and their biases." Public Opinion Quarterly 73.5 (2009): 895-916.

Rothschild, David, and Neil Malhotra. "Are public opinion polls self-fulfilling prophecies?." Research & Politics 1.2 (2014): 2053168014547667.

Rothschild, David M., and Justin Wolfers. "Forecasting elections: Voter intentions versus expectations." (2011). Available at SSRN: https://ssrn.com/abstract=1884644 or http://dx.doi.org/10.2139/ssrn.1884644

Silver, Nate. "How the FiveThirtyEight Senate Forecast Model Works." Internet, http://fivethirtyeight. com/features/how-the-fivethirtyeight-senate-forecast-model-works/, accessed on February 8 (2018): 2015https://www.overleaf.com/project/5c7b9ec06a

Silver, Nate. "How FiveThirtyEight Calculates Pollster Ratings." Internet, https://fivethirtyeight.com/features/how-fivethirtyeight-calculates-pollster-ratings/ accessed on February 8 (2018): 2015.

Williams, Leighton Vaughan, and J. James Reade. "Forecasting elections." Journal of Forecasting 35.4 (2016): 308-328.

Wolfers, Justin, and Eric Zitzewitz. 2004. "Prediction Markets." Journal of Economic Perspectives, 18(2): 107-126.

Author

Supervisor

Chapter 1

Introduction

Election forecasting has become popular in the academic sphere as well as in general life. Similarly to weather forecasting, election forecasts can be judged ex-post and their accuracy evaluated. However, whereas weather forecasters aim to predict the evolution of a dynamic chaotic system, election forecasters predict changes in public opinion. Election forecasting models can be based, among others, on economic and political measures, such as inflation, unemployment, or the duration of previous governance, on opinion polls, voters' intentions or voter's expectations, or on prediction markets. One thing that should be pointed out, is that the election forecasts can be partially self-fulfilling. Research suggests, that polls can influence voters' behaviour on the election day, whether they will vote, but also who they will vote for (Sudman (1986), Morwitz & Pluzinski (1996), Rothschild & Malhotra (2014)), which has important implications for public policy as well as for survey methodology of the polls.

In two-party systems, the outcome of the forecast models is usually the probability to win. This thesis focuses on elections in the Czech Republic, specifically the legislative election which took place in October 2017. Multiparty representative systems, such as the one in the Czech Republic, pose a unique challenge for forecasters, as there might not be a clear winner, and even distinguishing the party with the highest vote share as the winner is not enough to assess the results of the election as a whole. From that follows that in this thesis, the aim is to predict the exact vote shares gained by each party.

Traditionally, election forecasting models were most often built on two types of data - vote intention polls and structural models. More recently, researchers also started using voters' expectations and prediction markets to expand the prediction models and attain new information (as examples for prediction mar-

kets we can mention Reade & Williams (2019) and Berg & Rietz (2019), for voters' expectation see Murr (2011) and Temporão *et al.* (2019)). Both prediction markets and voters' expectation work on the assumption that aggregating the beliefs over the whole population reveals the true probabilities of an event. There are some conditions which need to hold for that to be true, such as the independence of decisions and diversity of information. Amongst other issues connected with these two types of forecasting models is that participants on the prediction markets might not behave in a risk-neutral manner, which results in the prices not reflecting the true mean belief (Wolfers & Zitzewitz, 2004). Furthermore, for voters' expectations in a multi-party system, the task to simultaneously predict vote shares for multiple parties can prove to be too difficult for the respondents (Ganser & Riordan, 2015). On the other hand, voters' expectations tend to be precise at the local level (Murr (2011) and Temporão *et al.* (2019)).

As structural models are not suitable for multi-party systems for multiple reasons (Walther, 2015), in this thesis poll data and betting market prices are used to form forecasts and then combined into one forecast weighing it by variance. This is a novel approach, as to my knowledge, the betting market prices have not been used for predicting the exact vote shares in the academic literature before.

For the polls, we model the development of parties' preferences throughout the election cycle using dynamic state-space model. Then the Kalman filter algorithm is used to distinguish between the true movements in the preferences and simple inaccuracies caused by sampling errors. The Kalman filter weighs the individual polls, creating a weighted average after each new poll, which comprises all the information from the previous ones. We also test for differences between the polling agencies, with the results suggesting the need for adjusting for the so-called *house effects*.

After comparing the polls, odds and the combined forecasts, the results suggest, that the betting market is able to predict the vote shares more accurately than the polls right before the election. However, as the betting market prices are very flexible, they can be adjusted close to the election in reaction to last-minute changes in preferences, whereas polls cannot be published by law three days before the election starts in order not to influence the voters' decision.

The structure of this thesis is as follows, the next chapter offers some background on forecasting in general and then presents an overview of the literature

on election forecasting, focusing mainly on opinion polls. We also briefly discuss the prediction markets and outline the political system in the Czech Republic. The third chapter outlines the methodology, first broadly for election forecasting models, followed by a more detailed description of the methods used in this thesis. Chapter 4 describes the data with the results being presented in chapter 5 together with a discussion on possible further expansions of the model. Finally, chapter 6 summarizes and concludes this thesis.

Chapter 2

Background and Literature Review

2.1 Forecasting

A vast number of forecasts, some formal and some only informal, is done every day in various areas of interest, from the weather and earthquakes, through sports to economic activity and political events. However, measuring the performance of a given probability forecast is quite complicated as one needs many observations to be able to correctly asses if the forecasted probability corresponds to the actual frequency in which the outcome is happening. A well-calibrated forecast will assign a 20 % probability to an outcome which occurs once in every five times after the event. An overconfident forecast will assign, for example, a 20 % probability to an outcome which will occur only once in twenty times and an underconfident outcome, on the other hand, will assign the 20 % probability to an outcome occurring once in three times.

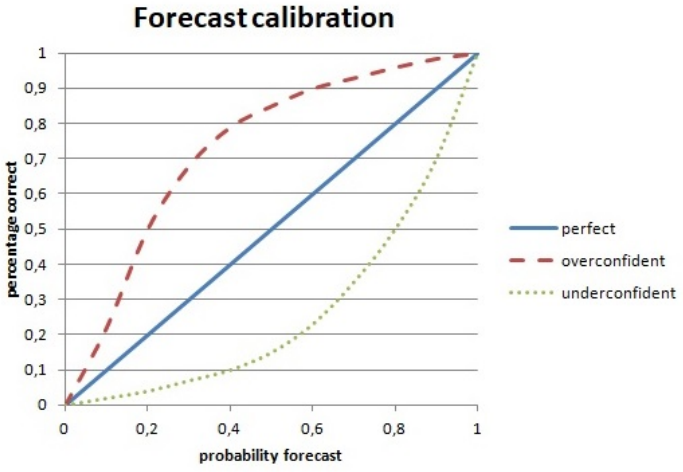

Figure 2.1: Forecast calibration

To be able to measure the performance of a forecast, one needs to either forecast an event that happens regularly or make enough forecasts on many different events and keep track of all the forecasts' accuracy. Weather forecasting, for example, is one where the accuracy can be measured quite easily, weather forecasts are made every day, therefore one can easily observe if it was raining once in five times after the forecast said there is a 20 % chance of rain. However, when a rare event is forecasted, such as a major earthquake happening once every two thousand years, evaluating the accuracy of this forecast in such a way is not possible.

Forecasts of political events are one example, where the accuracy is hard to measure and if the forecasts are not evaluated after the event, they might lose proper meaning as the perception of probability is very subjective. Many informal forecasts are done only in the way that an outcome is said to be probable, almost certain or unlikely, but every person might understand these words differently in terms of how often is an unlikely outcome actually realized. People tend to have a binary viewing of the world. If a forecaster gives a 10 % probability to an outcome, they do not expect this outcome to occur and if it does, they will say that the particular forecast was incorrect. Nevertheless, the 10 % forecasting probability said that the outcome will happen once in every ten realizations of the event. Therefore, if the event was repeated under the same conditions ten times and the particular outcome happened once, this forecast would be actually correct. Precisely because of that, we need multiple realizations of the events to be able to correctly asses the accuracy.

Any good forecast has to be probabilistic, providing a range of possible outcomes rather than just one number. This is because many events cannot be forecasted precisely, such as the example with weather forecasts. Weather is the result of a dynamic non-linear system, which means that the so-called 'chaos theory' applies. The behaviour of the system at one point influences how it will behave in the future, which brings uncertainty and makes it rather complicated to predict. Therefore, long-term weather forecasts might be less accurate than just taking the historical records and calculating the average temperatures in a given season, and that would not be caused by bad forecasting, but by the nature of the system. Similar limitations will apply to forecasts of election results, as they might be influenced by many different factors, that are simultaneously very hard or even impossible to measure.

With uncertainty about the accuracy of the forecasts come doubts whether the forecasts are conclusions of some meaningful expert analysis or just wild

guesses, with no expertise behind them. And furthermore, there are also doubts whether or not outcomes of political events in a dynamic world can even be predicted. As a result of that in the mid-1980s, Philip Tetlock, a professor of psychology, decided to run an experiment with the aim to measure the accuracy of forecasts about the outcomes of political events made by experts. He recruited almost three hundred experts for the experiment and asked them to make predictions about outcomes during more than ten years, naming the experiment "Expert Political Judgment" project and publishing the final result in 2005 in a book of the same name. The results of this experiment showed that on average the experts' predictions were actually as accurate as random guesses. However, a group of experts was determined, who were consistently more accurate than the average forecaster, meaning that their forecasts were well-calibrated. [1]

In the next section, we move on to election forecasting specifically, which has a long tradition as well. Election forecasts can have an important economic impact, by influencing stock prices, for example, as well as possibly affecting the behaviour of voters. For this reason, it is not only an academic interest be to able to determine which forecasts consistently perform the best.

2.2 Opinion polls

Surveys are a useful tool for researchers to get important information about population by using specifically designed samples of respondents, which represent the entire population, as surveying the whole population is not possible. Surveys are used in many fields such as market research, sociology, psychology, etc. Opinion polls are a type of surveys, which serves to asses the distribution of opinion about various social and political issues among the public. Political polling, as one category of opinion polls, including parties' preferences, has the advantage of being able to be regularly compared to the actual state during elections. Whereas in other fields, how well the poll results reflect the public opinion about a certain issue can be hardly verified, while the accuracy of election polls can be quantified and measured quite easily. And thus political polling is in the interest of researchers and has been studied in the literature thoroughly.

[1] Tetlock labeled this group "Superforecasters" and in a follow-up book, published in 2016, he tries to define the attributes that make a regular forecaster into a superforecaster as well as providing guidelines to accurate forecasting obtained during the experiment.

2.2.1 Getting information from polls

Political polls are conducted regularly during the political cycle by multiple polling agencies. As the preferences are not constant, each poll represents the distribution of preferences in the population at the time at which the poll was conducted. Naturally, there will be differences between each poll, which may come from numerous sources. The variation between the polls is widely discussed in the literature on survey methodology as well as in political science in the case of election polling (see for example Zukin (2004)). We will now go through the factors, that make the polls different one from another.

Firstly, each poll has its corresponding sampling error, which describes the difference between the sample and the population, since no sample can perfectly represent the whole population. If the sampling error is random, it can be reduced by aggregating the polls, as it should sum up to zero over the polls, we will discuss this later on.

Secondly, there are different types of polls depending on if all eligible voters are included, only likely voters or only decided voters. Also in some polls, respondents choose multiple parties that they are considering voting for. Furthermore, it depends on how the polls are conducted and how the samples are created. There can be polls conducted by phone or online which will exclude all households with potential voters who do not have access to phones or the Internet, etc. These are collectively called *design effects* and are associated with each specific poll.

Thirdly, there are the so-called *house effects*, which relate to the differences between each polling agency. These might be correlated with the design effects, as typically one agency uses a similar design for their polls. However, the agencies usually provide multiple types of polls as well. The house effects then represent, for example, the inherent bias in polls, when a given agency favours a specific party on a long-term basis as compared to other polls. Historically, some polling agencies tend to be leaning towards the left or the right of the political spectrum (Wang *et al.*, 2015).

Finally, we have the time effect. Therefore, if we account for all the differences in the polls mentioned above, the only difference left among the polls should represent the true opinion change in the population in time. Some shifts in public opinion may be permanent, but others are only temporary. Wlezien & Erikson (2003) model the preferences as a time series using a basic ARIMA model, where the opinion changes are represented by shocks. The ARIMA

model allows for persistent as well as decaying 'shocks' to the preferences.

To be able to study the changes in preferences in the population, it is important to distinguish between the poll errors and the true underlying preferences. Green *et al.* (1999) demonstrate the use of Kalman filtering on polling data to distinguish between actual shifts in preferences and movements caused by sampling error. Due to sampling error in polls, we cannot directly observe the true preferences in the population. Kalman filter calculates a weighted average from the time series with weights based on the uncertainty about each observation. We can calculate the filtered value for each observation, using the information from previous periods. In the first period, we will take the filtered value equal to the observed value $F_1 = X_1$ and we calculate the uncertainty as

$$P_1 = \sigma_u^2 + \sigma_{e_1}^2,$$

where $\sigma_{e_1}^2$ is the sampling error and σ_u^2 is the variance in preferences. Now the weight to calculate the filtered value in the next period is set as

$$W_2 = \frac{P_1}{P_1 + \sigma_{e_2}^2}. \tag{2.1}$$

Finally, the filtered value for the second observation will be equal to

$$F_2 = W_2 X_2 + (1 - W_2) F_1 \tag{2.2}$$

and the uncertainty for calculating the weight for the next observation is

$$P_2 = P_1(1 - W_2) + \sigma_u^2. \tag{2.3}$$

The advantage of Kalman filtering is that once we calculate the correspondent filtered value for a given observation, it contains all the information from previous observations. As we can see from 2.1 if the sampling error is high, the weight given to the observations will be lower, whereas the weighted information from previous polls will be given larger weight. On the other hand, if the true variance in preferences is high, the observation from the most recent poll will be given higher weight as it gives the most information about the true preferences. Generally, the variability of opinion is important, the larger it is, the less the previously conducted polls tell about the current state of the public opinion (Green *et al.*, 1999).

Thus Kalman filtering offers a useful method to study the changes in public

opinion over time. Polls are conducted throughout the election cycle, more frequently close to the election as would be anticipated. The polls get more accurate the closer the election is, as more voters decide who to vote for. However, by analyzing the movements in opinion during the election cycles, Erikson & Wlezien (1999) show that the electoral decision is in place before the start of the election campaign. Jackman (2005) also uses a dynamic linear model to study the changes in preferences during an election campaign. He corrects for variation in the polls caused by the house effects to identify the true changes in voters' preferences.

Measuring poll accuracy

Each election presents an opportunity to look back at the pre-election polls, to compare them with the actual results and determine, how accurate the polls were. Researchers have dealt with this topic, Mosteller *et al.* (1949) first proposed measures to evaluate the accuracy of election polls which became prevalent in the following years. Out of the eight proposed measures, six were based on the estimated proportion of votes received by the leading candidate or the estimated margin between the leaders. Later on, the appropriateness of these measures has been discussed as well. Mitofsky (1998) responded to criticism of polls after US presidential election in 1996, when the majority of polls overstated Clinton's lead. Comparing the 1996 election to the presidential election in 1948, when the polls actually predicted the wrong winner, they show that by most of the measures of poll accuracy the 1996 polls were more accurate then polls in 1948. Martin *et al.* (2005) review the measures and propose a new one, serving as a summary measure. As the pre-election poll forecasts can be evaluated directly after each election, they can considerably influence the public opinion on polls' and surveys' accuracy in general, which makes measuring the accuracy important not only for election forecasting but for other research fields as well.

2.2.2 From polls to forecasts

For a long time, raw polling data were used as the sole base for predicting the results of any upcoming election. In early 2000s, poll aggregation became popular, especially with the use of the Internet. However, the aggregated polling data, while increasing stability and accuracy of the prediction, still shows only what the result would be, if the election was held on the day of the polls.

Therefore, the prediction models were gradually transformed into probabilistic forecasts with the predicted outcome changed from the share of votes gained to the probability of victory.

The limitations of using only polling data are numerous and have been discussed in literature thoroughly. Wlezien & Erikson (2003) show what the polls tell about the preferences of voters over time. Using the polling data as a forecast, we implicitly assume that all previous changes in preferences will persists, which is usually not true as many movements in the preferences may be only temporary. The authors model the voters' preferences using a basic ARIMA model, which allows for persistent as well as decaying 'shocks' to the preferences.

Aggregating the polls might improve the accuracy, still it does not solve all the issues. It will decrease the variance of the forecast, but if there is some implicit bias present in multiple polls, the performance will not improve (Wright & Wright, 2018). Aggregating works on the assumption that the sample error is random and therefore should sum to zero across the polls, however it does not improve the performance if there are non-random errors.

Generally, representative samples are taken as a necessary condition for opinion polls to be able to reflect the population as a whole. However, conducting polls on representative samples may be costly and time ineffective. Wang *et al.* (2015) show that using non-representative samples for election forecasting can produce similar or even more accurate forecasts with appropriate adjustments, including poststratification and multilevel regression. These findings are promising even for other fields relying on public opinion polls, as collecting data from non-representative samples, e.g. with online surveys or on social media, is much cheaper and less time consuming.

2.3 Forecasting models

Another approach for predicting the outcome of an election is the so called *citizen forecasting* (Lewis-Beck & Skalaban, 1989). This method uses the wisdom of crowds by getting information from the general public. Whereas traditional election polls ask the respondents who they will, or consider to, vote for, in citizen forecasting respondents are asked what they think the result will be, which makes this approach similar in some way to prediction markets. If we assume that the expectations are randomly distributed, by aggregating them we cancel out the random errors and therefore the aggregated prediction should be more

accurate. Nevertheless, same as with the vote intention polls, this approach cancels out only random errors, whereas non-random errors will not be canceled and can still offset the forecast. Citizen forecasting in theory is an extension of Condorcet's jury theorem originally stated in 1785 (Condorcet, 1785). The theorem states that if individual group members have probability of predicting the correct outcome higher than 0.5, the probability of correct group prediction will increase rapidly towards infinity with increasing the group size. The theorem was derived under three main assumptions: group members' expectations are independent, the probability of correct prediction is equal for all members, and the concerning decision is binary. Later on, researchers have generalized the theorem and showed that it still holds without these assumptions (see for example Grofman *et al.* (1983) and List & Goodin (2001)).

Lewis-Beck & Skalaban (1989) tested citizen forecasting on US elections and showed that voters are generally able to predict the winner in advance of the election itself. They also tested factors which may affect the ability to forecast such as political involvement, partisanship, or vote intention. The results show that vote intention often makes the prediction biased towards the party the respondent is going to vote for. Further on, Lewis-Beck & Stegmaier (2011) show that the citizens were able to predict not only the election winner, but also the resulting seat share in the UK General Election through the aggregated citizen forecast.

Murr (2011) focuses on British elections and by applying the citizen forecasting at local level, asking the respondents to predict the results in their corresponding constituencies, he directly predicts the actual seat share as the election outcome. He argues that the citizens have more information about the local political situation and therefore are able to form more precise forecasts. Temporão *et al.* (2019) also apply citizen forecasting at constituency level to predict the election in Canada with positive results.

Ganser & Riordan (2015) extend the research on citizen forecasting on the case of German federal election in 2013. As Germany has a multi-party system, rather then predicting the winner, they ask the respondents to predict exact vote shares for each party. However, this task proved to be too complex as the voters' expectations were less accurate than vote intention polls. They also identify higher volatility in the election results and weaker partisan preferences in recent years, which makes election prediction more complicated.

Furthermore, Rothschild & Wolfers (2011) show that in the case of the two-party long established system in the United States, voters' forecasts are

more precise than voters' intention polls. Regardless of the fact, that voters' expectations are to some extent influenced by their party affiliation or their own decision of who they are intending to vote for.

There are also forecasting models based on fundamentals, which may be useful especially in more stable political system with long established parties with strong political programme. In multi-party political system, where the governing body is formed from a coalition of parties, structural models are much harder to use, as it is difficult to assign responsibility for economic and political issues to individual parties. It is unclear if good economic situation helps all the parties in the current coalition similarly, or how it affects the parties in opposition (Walther, 2015).

Brown & Chappell Jr (1999) combine poll data with a fundamental model and form continuous forecasts throughout the election cycle. They show that including the polling data significantly improves the forecast based on historical fundamentals only and that, in fact, with an appropriate approach, the polling data leads the forecast. Rothschild (2015) builds forecasts based on polling data, fundamental data and prediction market data. He shows that all three data types have a significant effect on improving the efficiency of the forecasts. He adjusts the weights of the different data types in the forecasts over time and shows that combining them improves the outcome especially early on in the cycle, as with the election approaching, the prediction market and polling data get very close. He also explores the translation of raw polling data and prediction market data into the forecast and gives further insight into those issues.

Generally, many other variables may be included in the forecasting models with the aim to improve their performance. Lewis-Beck & Tien (2018) use candidate-specific variables as predictors in their model and show that it increases the accuracy, however the effects are minimal. Nevertheless, as elections are not that frequent and therefore the number of observations is very limited, including multiple variables in the model may very often result in overfitting.

2.3.1 Multi-party systems

The stable two-party system, such as in the USA, makes an ideal target for forecasters. The elections always have one clear winner, which is not the case in countries with multi-party parliamentary systems with proportional representation. To forecast those types of elections, the models have to be adjusted,

among other things, the need for at least some polling data is clear as pure structural models are hard to apply (Walther, 2015). Alternatively, researchers can change the outcome variable from the election winner to the combined votes share of the governing parties (e.g. Norpoth & Gschwend (2010), Aichholzer & Willmann (2014)).

Walther (2015) show that the dynamic linear model can be applied to multi-party systems as well, to predict the election outcome from polling data, testing it on election data from Germany and Sweden. This approach also works well for newly established parties, whereas models based on fundamentals cannot predict the result for parties with no political history.

The issues arising in multi-party systems can be overcome by changing the predicted outcome. Norpoth & Gschwend (2010) build a structural forecast model, similar to models developed for US presidential elections, predicting the combined vote share for the currently governing parties in German Bundestag elections. Their model also includes data from opinion surveys with the main predictors being the long-term partisanship, chancellor approval and declining incumbent support. Aichholzer & Willmann (2014) forecast the combined vote share of the 'grand coalition', comprising of two mainstream parties, in the 2013 Austrian parliamentary election. They use unemployment rates, incumbency and dealignment over time as the predictors in their fundamentals forecast model. Using the combined vote share as the outcome variable, which is to be predicted, allows building the model based on economic performance, where it is assumed that the voters hold all governing parties accountable for the economic performance of the country during their tenure.

2.3.2 Model comparison

Further research has been done to evaluate the performance of forecasting models using different types of information. Rothschild (2009) showed that for predicting the 2008 US presidential election, prediction markets based models were more accurate early on in the cycle as compared to the poll-based forecast. However, both types of forecasts suffer from inherent biases, therefore the model predictions are de-biased and the results might be different depending on the method used for minimizing the biases. In Rothschild (2015) the author argues, that combining various data sources into one forecast brings advantage mainly early on in the election cycle, with time coming closer to the election the individual forecasts converge together. Another forecast comparison was

done by Reade & Williams (2019), who combine polls with multiple prediction markets. They transform opinion polls into probabilistic vote share forecast, correcting for known and unknown bias, to be able to compare it with prediction markets. They conclude that the polls, after correcting, exhibit only little bias, whereas the prediction markets are more precise.

2.4 Election forecasting in the United States

The main idea behind the methodology of this thesis is built mostly on forecasting models for elections in the United States. Election forecasting has a long tradition in the US, as the stable two-party political system allows for building forecasting models on a lot of available data.

FiveThirtyEight.com

FiveThirtyEight is run by Nate Silver and started as a poll-aggregation website in 2008. In the same year, he correctly predicted the outcome of the US 2008 presidential election in all but two states. The primary forecasting model is based on polling average, with weights assigned to different polls based on their historic track record, recentness, and the sample size. FiveThirtyEight have developed an elaborate five-step pollster rating system, collecting polls for presidential general elections, presidential primaries, senate elections, gubernatorial elections and US House elections dating from 1998 and assessing their accuracy (Silver, 2014), which is then used to appropriately calculate the polling average. The website provides a detailed description of their methodology as well as the data collected to the public.

The polling average is then combined with a regression-based model analyzing the demographic data. As shown in figure 2.2, the polls are given more weight in the forecast getting close to the election date itself, as there are more polls available and they are more accurate, because more people have decided who they are going to vote for (Silver, 2016). During the pre-election period, the authors of the website usually present three types of models - the poll average, polls-plus forecast and a nowcast. The poll average weighs all available polls, the polls-plus forecast incorporates also the fundamental data into the model, the nowcast predicts what the election result would be if the election was held today.

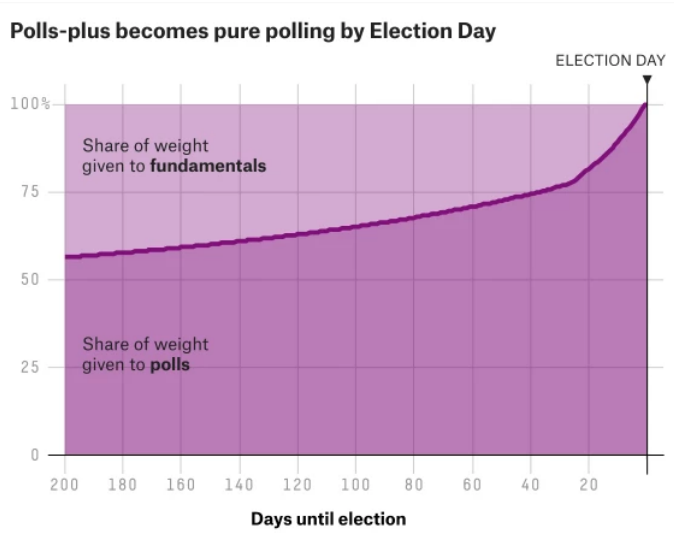

Figure 2.2: Shares of polls and the regression model in the forecast

Source: FiveThirtyEight

PredictWise.com

PredictWise, on the other hand, is a website which provides forecasts by combining polls and market data from prediction markets, it was founded by the economist David Rothschild. Prediction markets aggregate different beliefs over an unknown future outcome by trading binary options. They can be seen as crowdsourcing all the relevant information available about an event. According to the Efficient market hypothesis (EMH), which will be discussed in more detail in the next section, if the beliefs are formed independently and information is diversely distributed, aggregating the beliefs on prediction markets should result in revealing the true probability of an event. PredictWise collects information from six different prediction markets, the collected prices are then de-biased and normalized. The website publishes predictions on politics, sports, entertainment and finance.

Pollster.com

Pollster.com is a website that focuses on aggregating various polls, it does not provide forecasts about outcomes, but the polling average is calculated using the locally weighted moving average or local regression (LOESS) method. It has the longest tradition out of the three websites.[2] They aggregate all avail-

[2]The original Pollster.com has been since renamed as HuffPost Pollster, now available at `https://elections.huffingtonpost.com/pollster`

able polls conducted on representative samples. Similarly to FiveThirtyEight, Pollster publishes and regularly updates topical statistics, such as the current presidential approval rating.

Apart from these online services, US elections are also in the interest of many academic researchers. Campbell & Lewis-Beck (2008) provide a lucid overview of various forecasting models for the US presidential elections in the academic literature. They include fundamental models based on economic measures (e.g. Lewis-Beck & Tien (1996), Erikson & Wlezien (1999)) as well as poll-oriented forecasts (such as Pickup & Johnston (2005)).

2.5 Efficient market hypothesis

The idea to use betting markets as a forecasting tool for elections is in theory based on the efficient market hypothesis. The EMH, as discussed in finance as an investment theory, says that the market prices of assets are efficient, in the sense that they reflect all available information and therefore investors cannot consistently beat the market as the price movements represent only random shocks which cannot be predicted. The EMH requires utility-maximizing agents with rational expectations, who update their expectations after any new information is revealed to them, which is then fully reflected in the prices.

The hypothesis has three different forms:

- weak-form efficiency - for the weak-form efficiency, past prices do not bear any information that could be used for predicting the future prices, as the current price reflects all available information
- semi-strong-form efficiency - the semi-strong-form contains an instantaneous reaction of prices to new information
- strong-form efficiency - with the strong-form efficiency the prices do not reflect only the public information, but any private information as well

As it is assumed that the agents on the market are maximizing their utility, whenever the current stock price does not reflect the actual value of the asset, people will buy or sell stock and thus change its price until it corresponds to the beliefs based on all available information. Prediction markets work on a similar idea, but instead of trading stocks, they let people bet on the outcomes of various events. Betting exchange is such a place, where agents can trade

in real-time on the outcome of a discrete event. Each agent forms his own beliefs about the probabilities of the different outcomes and compares them to the betting odds. If they believe that the probability of the outcome is higher than what the odds on the market represent, and if the agent is rational and maximizing his utility, they should bet on the outcome to maximize his profit. On the other hand, if he believes the probability to be lower, he should sell to minimize the possible expected losses. In this way, whenever there is any new relevant information, which might influence the event, the agents on the market will update their beliefs and buy or sell accordingly, driving the odds to the real probability.

Prediction markets reflect the so-called "wisdom of crowds", people will have different beliefs about the probabilities, but on average, the beliefs should be correct and demonstrate the real probability of the outcome. The market odds should contain all available information, public and private as well, as agents who have access to any private information can also bet. Therefore the prices on prediction markets predict the future events in a similar sense as futures markets. The EMH is only a theory, which will hold under specific circumstances, Manski (2006) analyzes the price determination on prediction markets. He concludes that the prices do not reveal the mean belief, but rather just a boundary for the belief, with the market participants often behaving as risk-loving. The prices are also influenced by players' budget, which means that if we assume that the market price represents the average belief of the market participants, the average should be weighed by their respective budgets. Similar findings are also reached by Wolfers & Zitzewitz (2004) and Gjerstad & Hall (2005). Wolfers & Zitzewitz (2004) summarizes that the prediction markets have three important roles - they provide an incentive to seek information as well as an incentive to truthfully reveal the information, and last but not least, they act as an aggregator of diverse opinion.

2.6 Electoral system in the Czech Republic

The legislature in the Czech Republic is elected at a national level. Since 2013, the president is elected directly every five years using a two-round runoff voting. Members of the Chamber of Deputies are elected by proportional representation with a four-year mandate, there are 200 deputies in total. The Senate has 81 members, each representing one constituency, they are elected by similar two-round runoff voting as the president in their respective constituencies with a

six-year mandate. One third of the seats are re-elected every other year. For local governance, regional and municipal elections are held every four years.

The forecasting model has to be adapted to each election type. Generally, the polling average with different weights for each kind of poll may be similar, however the polls are conducted mainly for presidential and legislative elections, for other types, the number of polls will be limited. The same applies to the betting markets as bookmakers do not open betting for all elections.

The focus of this thesis is the legislative election held in October 2017, where members of the Chamber of Deputies were elected. There is a minimum limit set at 5 % of the popular vote that a party needs to gain in order to be voted into the Chamber of Deputies. We will form forecasts for all parties, that were above this minimal line in the election, which was nine parties altogether. In appendix A a list of the political parties with their respective acronyms used throughout this thesis is available. As often no one party obtains the majority of the seats, the party with the largest seat share can either form a coalition with other parties to have the majority or act as a minority government.

Chapter 3

Methodology

The main focus of this thesis is to build a general forecasting model for predicting the outcomes of elections held in the Czech Republic. The methodology partly follows Nate Silver's `fivethirtyeight.com` and David Rothschild's `predictwise.com` who forecast, among others, the results of elections in the United States. However, it is adjusted to fit the different electoral system in the Czech Republic. This section first describes a general forecasting model, such as is most commonly used in the US, then the approach used in this thesis is discussed in more details. As the aim of this thesis is to forecast the results of a legislative election in a multi-party system, the outcome to be predicted is the exact vote share gained by each party. The final model uses polling data and betting odds.

A general forecasting model has three main components.

- polling average
- regression model
- betting market prices

Polling average

The first step for building the model is to collect, weigh and average polling data. Before each election, there are numerous opinion polls conducted by polling agencies and other institutions. If all of them are done with random samples, then intuitively, averaging multiple polls should provide more precise results, as the size of the sample is increasing. Also given that the measurement

error of each poll is normally distributed, averaging the polls should eliminate it, or at least decrease it. Nevertheless, some polling companies tend to be more precise than others, depending on the methodology of the poll (Wright & Wright, 2018). For example, the polls can be done on a sample of all eligible voters, on decided voters, or on likely voters, which will all give different results. The precision also depends on the time of the poll, with the election approaching the polls should get more precise as pre-election campaigns reach their peak and most voters have decided who they are going to vote for. For this reason, the polling average is not just a simple average of all the polls, but each poll is given a different weight based on its historical accuracy, on the sample size, and on the timing in regards to the election itself.

Generally, polls give information about what would the results be if the election happened on the day when the poll was conducted, which means that they do not account for uncertainty coming from events that will possibly influence the voters' decision after the polling data are collected. Therefore, the polls themselves do not form a forecast, but they can be used as a Bayesian prior probability for the resulting forecast.

One option to weigh the polls is to use dynamic state-space model (e.g. Jackman (2005), Green *et al.* (1999), Pickup & Johnston (2005)). The model extracts measurement and other poll-specific errors, to get the true changes in preferences and weighs the polls by variance. If the variance is high, the long-term average is given more weight, whereas if the variance is low, each new observation should reflect the state of preferences at the given time better. It is also useful if there is not much information to properly assess the historic accuracy of individual polling companies or in multi-party systems, where there might be newly established parties and the political scene is more dynamic in general.

Adjusting polls

There are some factors that suggest the need to adjust the raw polling numbers. Typically, polls tend to have an anti-incumbency bias, which increases when the current economic situation is not optimal (i.e. the bias is stronger during economic recessions and weaker when the economy is going through an expansion, for a more detailed discussion on incumbency bias, see e.g. Fowler (2018)). Also, we might need to control for polls' party-biases, historically, some polling companies may tend to favour the candidates of a given party

(or may traditionally be leaning towards the left or the right, (Wang *et al.*, 2015)). This would apply especially to polls conducted by the political parties themselves, or by companies or media affiliated to a party or some politician.

Another option for improving the polling results is the trend line adjustment. If several consecutive polls show a steady growth of support for one candidate, we might expect this trend to continue towards the election and therefore increase their chances given by the current poll numbers.

By aggregating the polls we decrease the variance. If we assume that the measurement error is randomly distributed, it should sum to zero across all the polls. Nevertheless, aggregating does not improve the prediction if there are some non-random errors, such above mentioned party-bias or anti-incumbency bias (Wright & Wright, 2018).

Regression model

The regression model includes information about fundamentals, such as the current economic situation (inflation, unemployment,...), demographics, etc. It is given more weight in the final forecast early on in the election cycle, when not that many polls are available, however, the share of weight in the forecast is decreasing in time. Regression models often suffer from overfitting, especially if there are not enough observations as will be the case with Czech elections. Therefore, such models cannot provide accurate forecasts for future events (Babyak, 2004). The lack of observations and the resulting overfitting is a serious problem for building a regression-based model for forecasting. For example, when trying to forecast the outcome of the latest presidential election in the Czech Republic, there is only one previous election that could be used to build the model. Also with the traditional multiple-party system, it is more difficult to forecast the results as compared to the two-party system in the United States, where the two main parties have a strong consistent historical track record of their views and policies on various topics and it is clear who the voters take accountable for the present economic performance. In multi-party systems, it is unclear how the economic performance will influence the voters' preferences for a given a party, either in a governing coalition or in opposition, as they might perceive each party accountable at a different level. Nevertheless, to overcome this problem, the forecasting model can be built to predict the combined vote share of the governing parties, viewing them accountable as a whole (e.g. Aichholzer & Willmann (2014), Norpoth & Gschwend (2010)).

Betting market prices

The idea behind including the prices on betting markets into the model for forecasting comes from the efficient market hypothesis. Put simply, the prices should reflect all available information at any given time. If someone had any additional information which would make them believe that the current prices are off, it would be rational for them to participate in the market (in the case of betting market to place a bet) to gain a profit. This action would then move the price to its fair value again comprising all available information.

Prediction markets

Prediction markets allow people to place bets on the outcome of different events. They indicate what the public thinks is the probability of an event, following the efficient market hypothesis, the prices on the market should reflect all available information. There are three conditions which should hold, for that to be true. Firstly, the information should be distributed diversely amongst the public, secondly, the decisions should be made independently and thirdly, the market organization should be decentralized.

The prediction markets have two main advantages. They aggregate vast amounts of information, beliefs and data. Moreover, they are able to provide truthful and relevant information through financial and other incentives. The marginal trader hypothesis says that "there will always be someone seeking out places when the crowd is wrong", i.e. whenever there is a possibility of profit coming from prices not reflecting the actual probabilities, a rational agent will use it and drive the prices to the true probabilities. However, there are also factors which may offset the prices.

Firstly, the prediction markets are susceptible to speculative bubbles, similarly to any other investment-based financial markets. Speculative bubbles occur when there are exaggerated expectations increasing the trading volume and bringing more buyers than there are sellers pushing the price higher without it actually having the underlying value.

Secondly, prediction markets can be used as a form of insurance against a possible negative outcome of a particular event. By betting on the negative outcome (negative meaning that the outcome leads to an expected financial loss), even if the outcome has a low probability, the agents can insure themselves. Then in case the event results in the negative outcome and thus the

expected financial loss occurs, it will be balanced by the winning bet. This sort of self-insurance was, for example, common before the Brexit vote held in United Kingdom in 2016 or during 2016 US presidential elections.

Thirdly, the prediction markets suffer from self-reinforcement. Generally, the users of prediction markets have similar mindsets. When the players on the market see that others have similar beliefs about the probabilities, they might be persuaded to take the market odds as correct and not update for new information. Therefore, they tend to get overconfident about their own beliefs and if there is some information not supporting them they tend to not account for it and only focus on the evidence supporting their convictions. This kind of self-reinforcement could be observed before the Brexit vote, when the prediction markets failed to anticipate the win of the 'Leave' side.

Finally, there is the so-called 'favourite longshot bias', which drives the prices away from the actual probabilities at the tails. If an outcome has a very high probability, the prices will not reflect it accurately, especially if it is a long time before the event. It is because betting on the prediction markets comes with transaction costs as well as opportunity costs from not holding the money until the event occurs and the outcomes are revealed.

3.1 Model

The three components are then combined into one forecast and uncertainty is introduced into the equation. The forecast does not give a precise prediction of the actual results, but rather the probabilities of the most possible outcomes. Bayesian inference is used to calculate the probabilities.

As mentioned above, the polling average gives the Bayesian prior probability. The regression model and betting prices together form a signal, which is used to update the prior beliefs to get the posterior probability.

$$E(posterior) = (1 - \theta) \cdot (value\ of\ the\ signal) + \theta \cdot E(prior),$$

$$\text{where } \theta = \frac{Var(noise\ of\ the\ signal)}{Var(noise\ of\ the\ signal) + Var(prior)}$$

The parameter θ weighs the information coming from the signal and the prior based on their variance. If the signal is noisy (the variance is large) more weight will be given to the prior, if the prior has large variance, more weight will be given to the signal (Rothschild, 2015). For example, as the

election approaches, the polls get more precise and their variance goes down. Therefore, the closer to the election, the more weight will be given to the prior, whereas if there is a long time to the election, there are generally only a few polls available, as well as a lot of undecided voters, which increases the variance of the polls and in the forecast more weight will be given to the signal. Figure 3.1 illustrates how the initial belief, i.e. the prior probability, is updated using the information coming from the signal to form the posterior probability distribution.

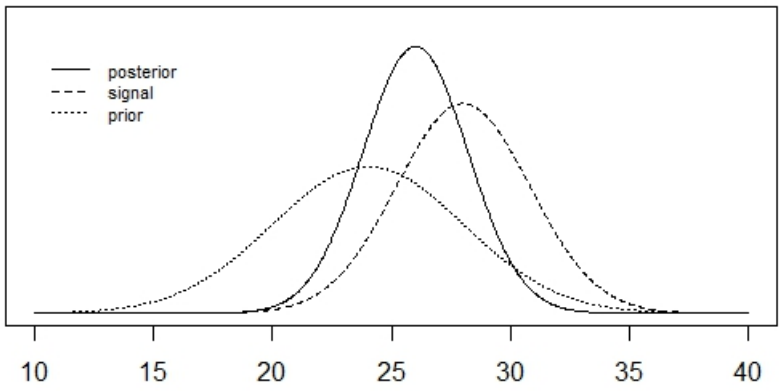

Figure 3.1: Bayesian posterior distribution

If we assume that the measurement error of the polls is normally distributed across them with a zero mean, the prior should be unbiased with the expected value equal to the actual outcome.

The forecasting model in this thesis is an extended version of a model developed by `kdovyhrajevolby.cz` for the 2017 legislative election, the results of which were published online.

3.2 Hypotheses

The final model built following the methodology described in this section is used in this thesis to forecast the results of the legislative election in the Czech Republic held in October 2017. Subsequently, the model can be used for future elections as well with adjustments for the current situations.

Hypothesis #1 Aggregating numerous pre-election surveys improves the results in terms of predicting the outcome of elections as compared to looking only at one survey. In other words, the aggregated average performs better in the long-term than any individual survey for predicting the election results.

Hypothesis #2 Including soft information in the form of prices on betting markets improves the performance of the election forecast.

The performance of the model will be also evaluated with an out of sample forecast for the most recently held elections. Furthermore, the accuracy of each type of forecast will be evaluated on its own, as well as their different combinations, with the aim to find the method which gives consistently as accurate predictions as possible.

3.3 Dynamic linear model

Poll results are influenced by numerous factors, which may skew the polled preferences, from the true preferences in the population. As was discussed in section 2.2.1, these include the sampling error, but also house and design effects connected with the poll's structure and the polling agency. Figure 3.2 shows the polled vote share for ANO, during three year period before the 2017 legislative election together with the changes in the preferences between each poll. The poll results are very volatile and it is obvious that the movements in the raw polling data are not reflecting only the true shifts in the underlying preferences, but are affected by other factors. Therefore, to estimate the preferences from the polling data, Kalman filter method is used (according to Green *et al.* (1999)).

Traditionally, Kalman filtering is a statistical method used mainly in engineering to analyze state-space models. The method helps to extract the signal from the noise in the data. In this thesis, Kalman filtering is applied to polling data to distinguish between genuine movements in public opinion and random movements caused by sampling error.

We are interested in the true preferences which are not observed but are measured by polls. However, as polls are conducted only on a small portion of the population, the information they give about the true preferences is not perfect. We want to estimate the underlying preferences from the polling data, therefore, we need to remove the noise caused by sampling errors. The state-

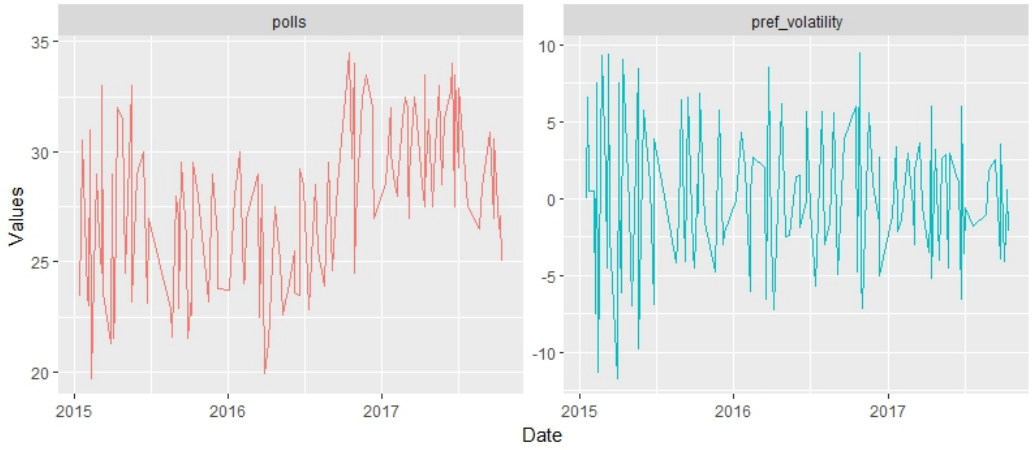

Figure 3.2: Vote share from polls for ANO

space model in a general form is described by a set of two equations - the measurement equation:

$$Y(t) = H \cdot X(t) + v(t), \quad v(t) \sim N(0, R) \tag{3.1}$$

Where $Y(t)$ are the observed values, $X(t)$ are the underlying preferences (or states in the state-space model), H is the transformation of states matrix and $v(t)$ is a vector for the measurement errors. Secondly, there is the transition equation, which represents the dynamic process driving the preferences, the current state is a modification of the state in the previous period, with $u(t)$ being the transition (state) error:

$$X(t) = A + F \cdot X(t-1) + u(t), \quad v(t) \sim N(0, Q), \tag{3.2}$$

We will assume that the measurement error and transition error are independent, i.e. $Cov(u(t), v(t)) = 0$.

The measurement equation 3.1 can be decomposed into trend and cycle, to better represent the cyclical changes in the preferences. The two equations will then have the following form. First the measurement equation is expressed as

$$Y(t) = [h_1 h_2 h_3] \cdot \begin{vmatrix} trend(t) \\ cycle(t) \\ cycle(t-1) \end{vmatrix} + v(t), \quad v(t) \sim N(0, sigmaR). \tag{3.3}$$

And the transition equation is

$$\begin{vmatrix} trend(t) \\ cycle(t) \\ cycle(t-1) \end{vmatrix} = \begin{vmatrix} a_1 \\ a_2 \\ a_3 \end{vmatrix} + \begin{vmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{vmatrix} \cdot \begin{vmatrix} trend(t-1) \\ cycle(t-1) \\ cycle(t-2) \end{vmatrix} + \begin{vmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{vmatrix},$$
$$u(t) \sim N(0, sigmaQ). \quad (3.4)$$

To be able to identify the model, we need to set some restrictions on the parameters, otherwise, the estimation will not converge. For the measurement equation, we restrict $h_1 = 1$, $h_2 = 1$ and $h_3 = 0$. For the transition equation, we need restrictions for each of the decomposed parts. For the trend equation, we set $f_{11} = 1$ and $f_{12} = f_{13} = 0$, for the cycle equation $f_{21} = a_2 = 0$ and the last equation, which is representing the lagged cycle identity, we have $f_{32} = 1$ and $f_{31} = f_{33} = a_3 = 0$. And finally, the variance matrix Q for the transition error is diagonal. The final model now has the following form:

- measurement equation

$$Y(t) = [110] \cdot \begin{vmatrix} trend(t) \\ cycle(t) \\ cycle(t-1) \end{vmatrix} + v(t), \quad v(t) \sim N(0, sigmaR) \quad (3.5)$$

- transition equation

$$\begin{vmatrix} trend(t) \\ cycle(t) \\ cycle(t-1) \end{vmatrix} = \begin{vmatrix} a_1 \\ 0 \\ 0 \end{vmatrix} + \begin{vmatrix} 1 & 0 & 0 \\ 0 & f_{22} & f_{23} \\ 0 & 1 & 0 \end{vmatrix} \cdot \begin{vmatrix} trend(t-1) \\ cycle(t-1) \\ cycle(t-2) \end{vmatrix} + \begin{vmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{vmatrix},$$
$$u(t) \sim N(0, \begin{vmatrix} sigmaQ_1 & 0 & 0 \\ 0 & sigmaQ_2 & 0 \\ 0 & 0 & 0 \end{vmatrix}). \quad (3.6)$$

The variance matrix R for the measurement error will be represented by sample errors from each poll. The parameters left to be estimated by the model are then $a_1, f_{22}, f_{23}, sigmaQ_1$ and $sigmaQ_2$. The Kalman filter algorithm uses the maximum likelihood method to estimate the parameters of the model. Polls with larger sample error are given less weight, for example, polls with larger sample size should have lower sample error and therefore will be given

higher weight. The model is first calculated with guessed parameters which are optimized in the next step using the maximum likelihood estimation. Each party's preferences are taken as a separate time series, therefore the model parameters are estimated individually for each of the parties.

The final model with estimated parameters can then be used to forecast the changes in preferences and thus the election results. The forecast is created using Monte Carlo simulation and calculating the prediction from the estimated Kalman filter model going into the next period. 95 % confidence interval is then calculated from the simulations for the forecast.

Kalman filter uses only the previously observed values for the estimation, which is why it is useful for forecasting. Nevertheless, for ex-post estimation we can also use Kalman Smoother, which uses also observations from the following periods for the maximum likelihood estimation of the parameters and therefore should be more precise as it benefits from more information in the data.

3.4 Translating odds into probabilities

The next part of the model is based on betting market odds, we need to make some transformations to be able to properly translate the odds into probabilities, which are described in the following section.

3.4.1 Estimating the mean

The odds offered on the probabilities to win will not give us much information in the case of a multi-party system, as we are interested in the actual vote share gained by a given party, rather than the winner of the election. For the forecast, we can use the odds offered on exact vote share for each party. They are set in a way, in which the agency determines a probable vote share level for the given party and sets the probabilities that the actual vote share will be below or above this level based on their beliefs. Early in the election cycle, when the odds are set for the first time, the probabilities are often symmetric around the determined vote share level. Such was the case for the party ANO, for which the first odds offered on the 30th of May were 1 to 1.85 that the vote share will be between 0 and 26.99 % and 1 to 1.85 that it will be 27 % or higher. Translating these into probabilities and scaling to sum to 1, we get a 50 % probability, that the vote share will be lower than 27 % and 50 % probability that it will be higher than 27 %. We assume that the actual vote share can

be represented by a random variable following normal distribution around the determined level of 27 %, which illustrates the underlying uncertainty. This is represented in figure 3.3, where the mean is set at 27 % of the vote share and we assume, that the actual vote share will fall, with the probability of 95 %, somewhere in the shaded area, which is calculated from the standard deviation of the assumed distribution.

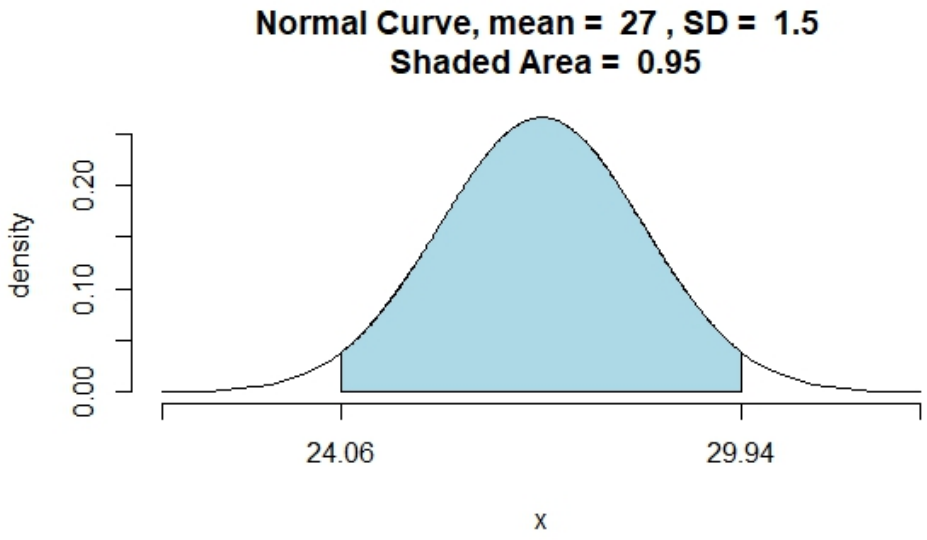

Figure 3.3: Normal distribution - probability distribution function

Nevertheless, this works well only when the odds are symmetric, and in most cases the probabilities given by the odds that the actual vote share will be above or below the determined level differ. Figure 3.4 illustrates that case, in the figure the threshold, at which the odds are set, is equal to 27 % of the votes. However, the probability that the vote share will be lower than these 27 %, represented by the shaded area, is equal to 84.13 %. Consequently, the true mean in this example would be equal to 25.5 % as the expected vote share. Therefore, we need to adjust the mean to form the actual forecast. To do so, I first generate a sample from normal distribution as if the probabilities were symmetric (i.e. the mean of the normal distribution is the boundary of the odds), but then resample again with probabilities assigned according to whether the observation is below or above the determined vote share level. By resampling, a new sample is created, which corresponds perfectly with the probabilities given by the odds. From this sample, we calculate the mean as well as the 95% confidence interval, which creates the forecast.

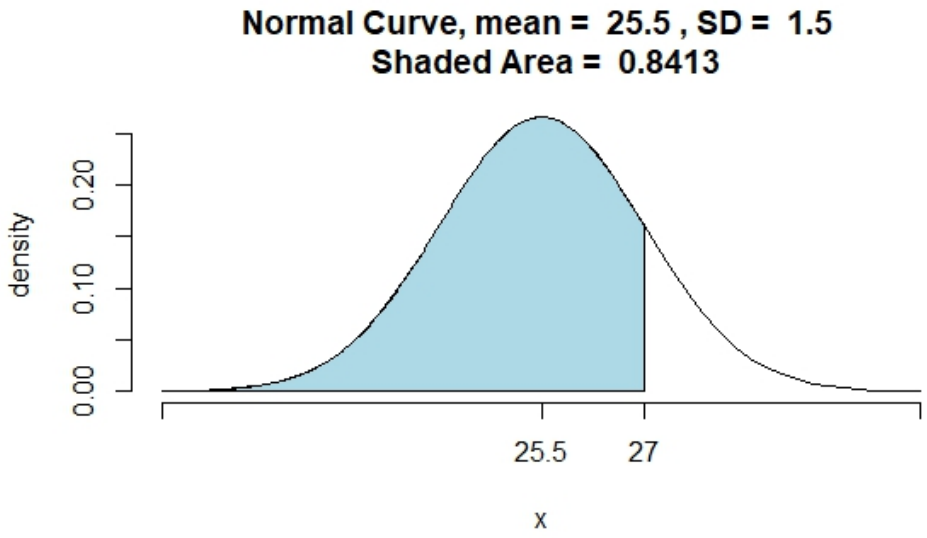

Figure 3.4: Normal distribution - probability distribution function - adjusting the mean

3.4.2 Estimating the standard deviation

So far, we have focused on the mean of the underlying normal distribution for the vote share. However, we also need to estimate the second parameter, the standard deviation, to know how steep the curve plotting the probability density is. I will approximate the standard deviation, using the case when two different sets of odds with different vote share level are set for a given party at the same time. This gives us two points on the curve, and by adjusting the parameters to fit the odds, we can approximate the standard deviation. As we assume, that the vote share is normally distributed, the odds represent points on the cumulative distribution function given as:

$$P(X < x) = F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}} \,\mathrm{d}t$$

.

Table 3.1 shows two sets of odds offered for ANO at the same time, from them we get the two points on the distribution function $P(X < 25.5) = F(25.5) = 0.62$ and $P(X < 27) = F(27) = 0.9$. Using these two points I have approximated the distribution by adjusting the parameters and the standard deviation is roughly equal to 1.5.

We will assume that this is the basic standard deviation that the betting agency is calculating with. However, it is still only an approximation of the true

	odds	probability
0 - 25.49 %	1 : 1.75	62%
25.5 % and more	1 : 1.95	38%
0 - 26.99 %	1 : 1.2	90%
27 % and more	1 : 4	10%

Table 3.1: Odds - ANO, 20.9.2017 10:32

distribution representing the attached uncertainty. Therefore we will also calculate forecasts and corresponding confidence interval with increased standard deviation.

Chapter 4

Data

The forecasting model in this thesis combines two types of data, firstly, the voters' intention polls and secondly, the betting market odds.

4.1 Polling data

There are several different types of polls depending on the structure of the questions and on which respondents are included. Generally, we can distinguish five main types:

- Electoral model - an estimate of the election results, included are only the answers of respondents, who intend to vote. Undecided voters are not included.
- Party preferences - an estimate of the party's popularity. Contains also respondents who do not intend to vote.
- Election potential - how many votes would the party gain if everyone who considers voting for it would do so. Respondents can state multiple parties, included are all respondents who do not exclude their participation in the election.
- Electoral core - voters who are firmly decided to vote for a given party.
- Election prognosis - forecast of the election results made by polling agencies. The actual methodology for creating the forecast is usually not published.

The most accurate representation of the situation on the election day, that can be used for building a forecasting model is, therefore, the electoral model (excluding the election prognoses, which are forecasts in themselves). The other types of polls indicate movements and changes in public opinion. Polling agencies publish all types of polls regularly during the whole election cycle. We start the model for the 2017 election with polls from January 2015 onwards, up until the election (the Czech law prohibits to publish election polls in the period starting three days before the election until the voting closes, in order not to influence voters' decision). In those almost three years, 118 electoral models were published by 9 different polling agencies. Figure 4.1 shows the number of polls conducted each month during the period. Depending on the agency's methodology, some polls can be conducted over multiple days, some are taken over a month, some collect data only one day. There can also be a delay between when the poll was conducted and when it was published. For each poll, we take the last day that the data was collected for it as its date.

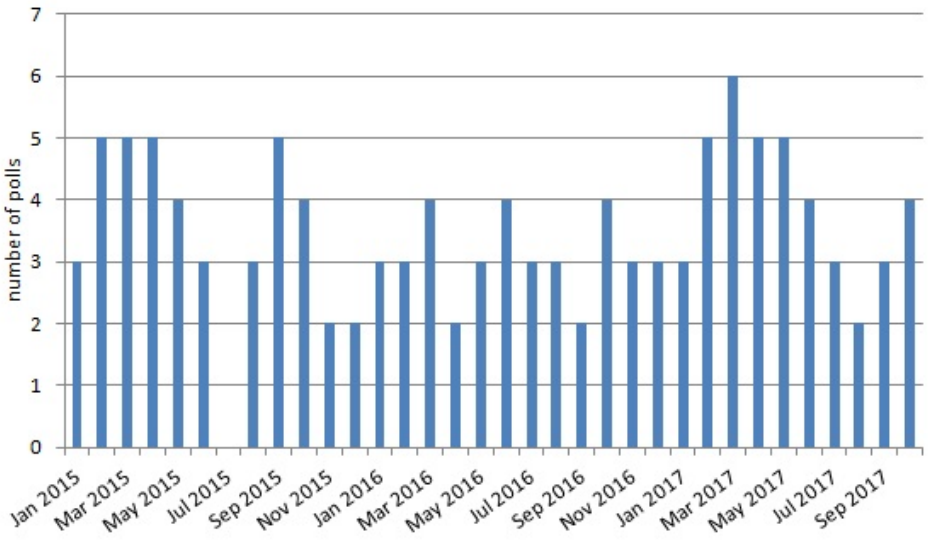

Figure 4.1: Number of polls conducted each month

Table 4.1 summarizes the number of polls published by each agency together with their average standard errors. We calculate the sampling errors from the polls' sample sizes, most often the polling agency reports an approximate sampling error with a range of around 2 percentage points. The calculated sample errors are usually slightly above the average of the reported range. Some of the polls include coalitions of two or more parties, in that case, the polled vote share is divided equally between the parties included in the coalition, as

in the case of the 2017 legislative election all main parties stood separately as candidates. Two parties announced their separate candidature later on in the election cycle, namely SPD, which was established only in June 2015 and announced its candidature in November 2016, and STAN, which stood as a separate movement for the first time in this election. For this reason, the polls started including these parties only later on.

	number of polls	average standard error
SANEP	24	1.52%
CVVM	24	2.92%
TNS Aisa	19	2.62%
MEDIAN	21	2.63%
PPF Factum	6	2.6%
STEM	11	2.82%
Médea Research	4	2.94%
Kantar TNS	5	2.72%
Focus	4	3.23%
total	118	2.51 %

Table 4.1: Polls by polling agencies - number of polls and standard errors

Each party's preferences are treated as a separate time series in the model. However, the polls have their associated standard errors, which apply to all parties in the given poll. This causes some issues for the parties with low expected vote share, which will also be discussed later, as the poll' standard error is relatively higher for parties with lower vote share as compared to parties with higher vote share. We estimate the preferences of all parties, whose vote share on the election day was higher than the minimum threshold to get in the Chamber of Deputies set at 5 % of the public vote. Altogether, nine parties obtained at least those five per cent and were therefore elected into the Chamber of Deputies.

4.2 Betting odds

The odds are primarily set by betting offices, who offer bets on different events. They set the initial odds based on their beliefs, which are then continuously updated with new information. That can be either events connected with the

election campaigns, including television and online debates, the publication of new polls or the behaviour of betters, which can reveal some private information not reflected in the betting prices so far.

Most often, when the prices on prediction markets are used to predict elections, they are set as probabilities to win in two-party (or two-candidates) systems, such as the case of US elections. However, as the aim here is to forecast the exact vote shares for each party, the probabilities of winning do not give sufficient information. Therefore, odds that are set on exact vote shares are used in the model. The odds and how they are translated into probabilities have been already described in the methodology, section 3.4.

As the election campaign officially started in May 2017, the first odds were offered by the polling agencies around the same time. We mainly focus on three parties, for which we have the complete data on the changes in the odds up until the time of the election itself. In figure 4.2 the number of updates in the odds in the months before the election for each party is shown. We can see that close to the election, the odds are updated more frequently, as there is new information coming, more people place their bets and the expectations are formed more precisely.

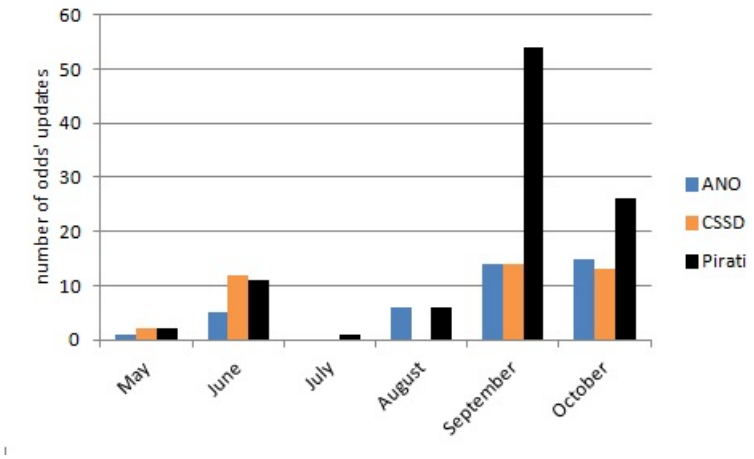

Figure 4.2: Number of changes in odds

For the actual election forecast, we use the last odds that were available on the betting market right before the start of the voting. This gives an advantage to the odds forecast as the polls cannot be published before the election, but betting market prices can change up until the election itself, reacting to last-minute changes in the preferences.

Chapter 5

Results

5.1 Preferences from polls

To get the underlying preferences from the polling data, the Kalman filter method was used to separate the signal from the noise in the data. The algorithm updates the model parameters with each new observation, i.e. with each published poll. Figure 5.1 plots the polled vote share for ANO together with the filtered values representing the true preferences. From the graph, we can see that the Kalman filter decreases the variance and computes a smoothed average from the polls. Figure 5.2 shows the filtered poll results for all parties, which were voted into the Chamber of Deputies (i.e. obtained at least 5 % of the popular vote) in the 2017 legislative election, starting from January 2015.

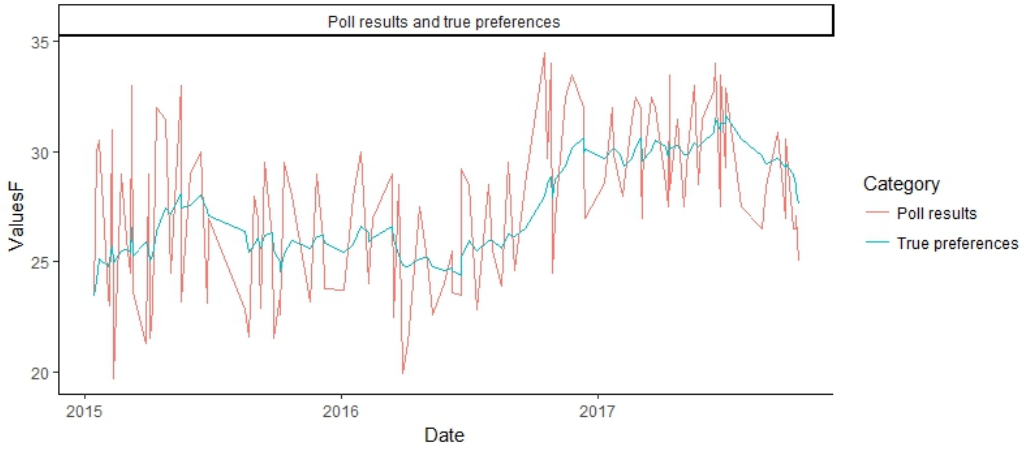

Figure 5.1: Estimated preferences for ANO from polls using Kalman filter

Each poll has its specific sampling error, which is taken as the measurement

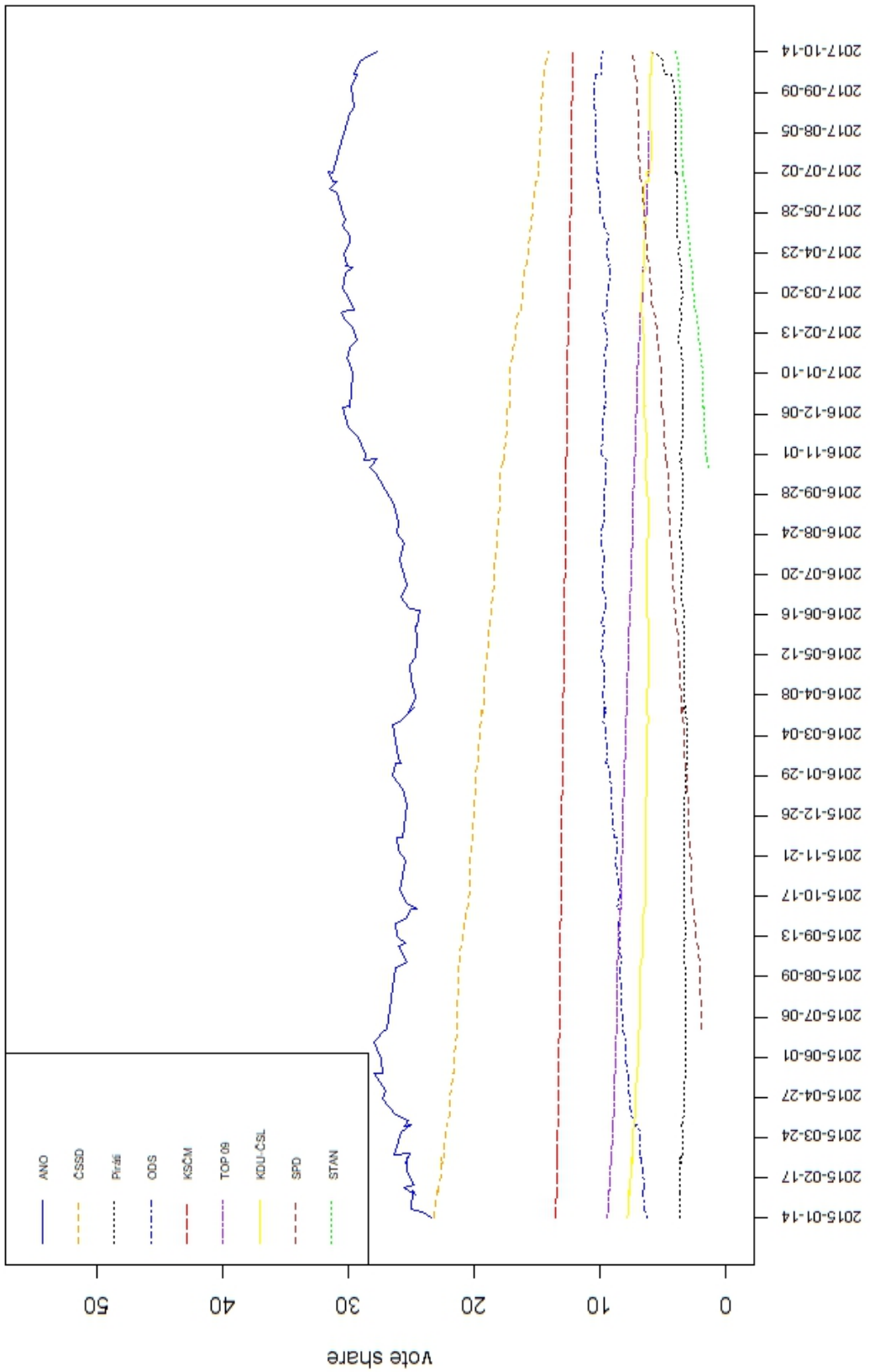

Figure 5.2: Filtered polls

error in the state space model. The average sampling error across all polls is around 2.5 percentage points. This is rather high for parties with low expected vote share, because for parties with estimated preferences at around 5 % of the public vote, this regular sampling error means possible differences of up to 50 % between the polled value and the actual one. This means that the Kalman filter puts most of the weight on the long-term average as the variance in the individual polls is relatively high and the resulting estimated true preferences are almost perfectly smoothed.

5.2 Poll forecast

Using the maximum likelihood method, I have estimated the parameters of the underlying dynamic linear model, which describes the movements of preferences for each party. The model with the estimated parameters is then used to forecast the vote shares gained in the election by each party using the Monte Carlo Simulation. Figure 5.3 shows the 95 % confidence interval for each party's vote share calculated from the simulations. The red dots represent the actual results of the election.

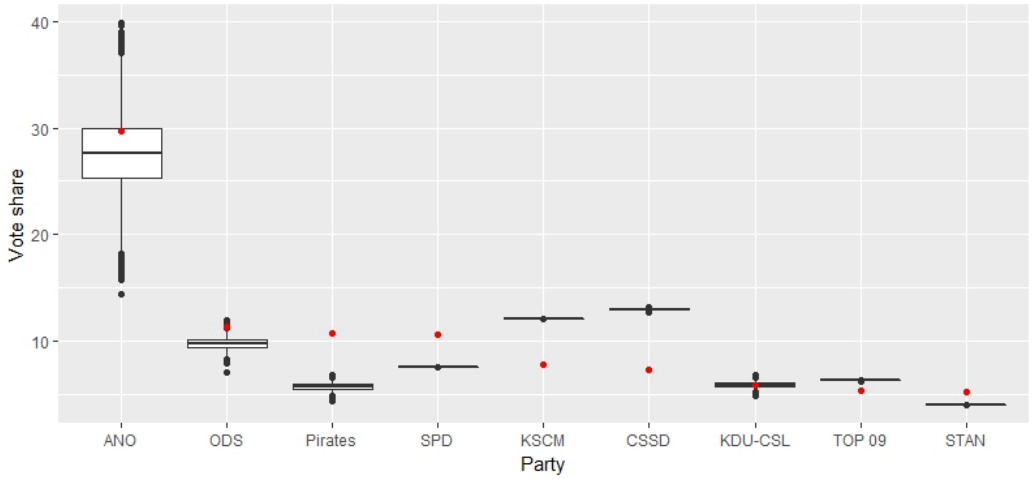

Figure 5.3: Forecasts with confidence intervals

As was mentioned above, for parties with a lower vote share the sampling error in the polls is very high as compared to the vote share. This results in Kalman filter putting most of the weight on the long-term average, which means that the final polls' forecasts then have very low variance, as the algorithm predicts that the vote share will follow the long-term average after accounting for the measurement error. This is also caused by the fact that we calculate the

forecast right before the election, therefore the model assumes that the latest poll-average will be a close representation of the true preferences. If we create the forecast more in advance, some additional variance should be included in the poll forecast accounting for the future possible changes in the preferences. The actual vote shares are within the confidence intervals only for two parties - ANO and KDU-CSL, three other parties are relatively close - namely ODS, TOP 09 and STAN. For the remaining four parties the forecasts are more than three percentage points off.

5.3 Betting odds

As was described in section 3.4, the odds from the betting market are translated into estimated vote shares, which creates the forecast. First, taking the odds on gained vote share for each party, a normal distribution with the boundary of the odds being the mean is generated and then a sample is drawn from this distribution based on the probabilities calculated from the offered odds. The first odds on percentages of the votes were offered at the start of the official election campaign in May 2017. They are then regularly updated, based on new information, such as newly published polls, important events in the election campaign, or realized bets, which may reveal previously unknown information. Figure 5.4 shows the development of the odds forecast, i.e. the estimated mean for the vote share, for the three selected parties, for which we have the complete data.

Taking the final odds offered right before the start of the election, we will calculate the predicted vote share and its confidence interval for each party that was above the minimal level of 5 % of the popular vote to be voted into the Chamber of Deputies. Figure 5.5 shows the forecasts with their 95 % confidence intervals, calculated from the samples drawn based on the probabilities. The red dots represent the actual result of the party in the election. The results of two parties, SPD and KDU-CSL, are within the confidence intervals, next three parties, ODS, TOP 09 and STAN, are just on the edge of their confidence interval. For the remaining four parties, namely ANO, Pirates, KSCM, and CSSD, the actual election results are outside the forecasted confidence intervals for the vote shares.

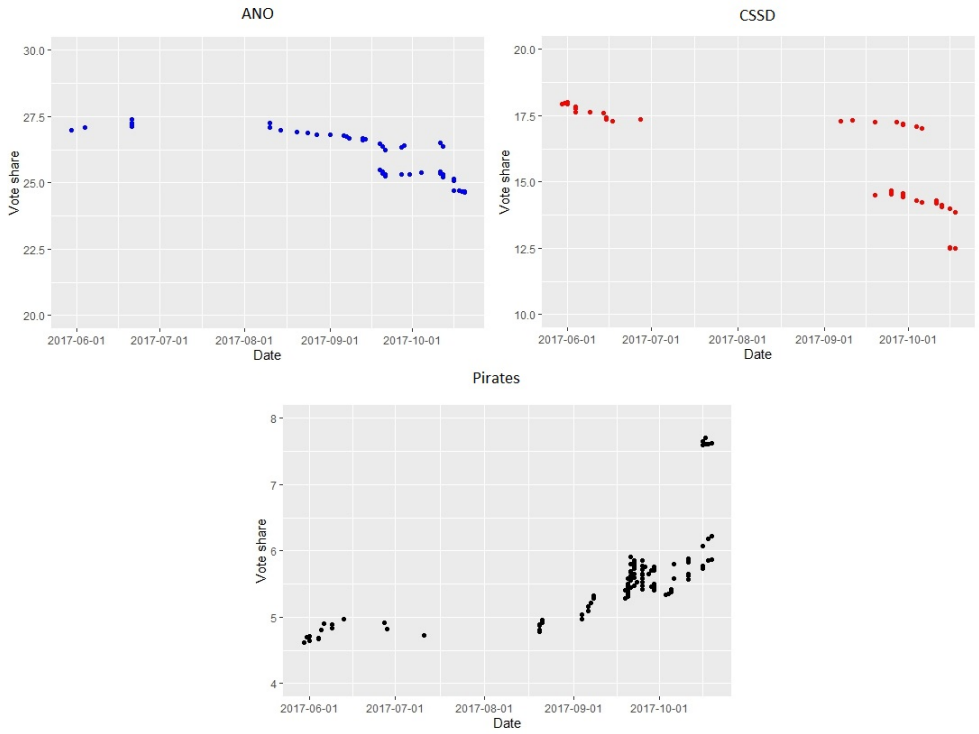

Figure 5.4: Development of betting odds forecast - ANO, CSSD and Pirates

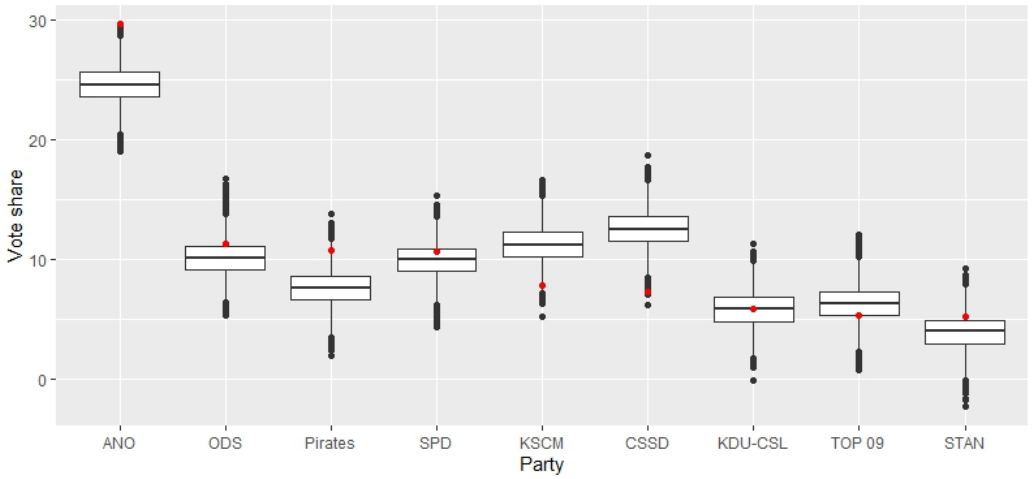

Figure 5.5: Odds forecasts with confidence intervals, sd = 1.5

As the parameters for the odds forecasts are only an approximation, we also create forecasts with increased standard deviation. The confidence intervals are again calculated from samples drawn based on the probabilities from the odds from the generated normal distributions. The resulting confidence intervals are plotted in figure 5.6 with red dots depicting the election results.

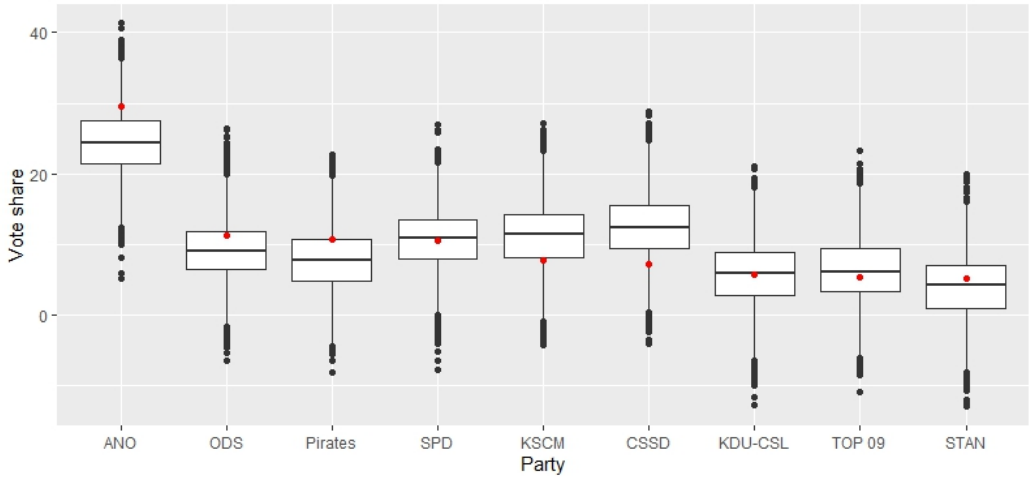

Figure 5.6: Odds forecasts with confidence intervals, sd = 4.5

Now we have the results of five of the parties within their confidence intervals. Two others are on the edges and two are still outside of their respective confidence intervals. Nevertheless, by increasing the standard deviation and thus allowing for more uncertainty in the forecast, we were able to create more accurate forecasts.

5.4 Combined forecast

Next, we want to combine the information from the polls with the information from the betting odds to form one forecast. The polls are taken as a Bayesian prior distribution, which is updated by the betting odds, i.e. the signal, to get the posterior probability. The signal represents some observed data, based on which we can update our prior beliefs about the probabilities. The two are weighed by variance. If there is high variance in the polls, they will be given less weight, if the variance in the polls is low, the information from polls will have more weight. As was already discussed, the Kalman filter puts most of the weight on the long-term average for parties with low vote share and does not allow for much variance in the poll forecasts, as it treats most of the fluctuations in preferences as pure measurement errors. We, therefore, restrict the variance at a minimal value equal to 0.10, in order for both of the data types being still at least somehow accounted for in the combined forecast, as otherwise, the variance of the poll forecast converges to zero for some of the parties and the combined forecast would put all of the weight only on the polls. With the exception of ANO, the combined forecast still puts most of the weight

on the polls and only very little on the odds. Table 5.1 summarizes the odds and polls forecasts for each party, together with the weights used to calculate the combined forecast.

	signal - odds			*prior - polls*			*posterior*
	mean	variance	weight	mean	variance	weight	mean
ANO	24.63	2.27	0.85	27.66	12.49	0.15	25.10
ODS	10.12	2.17	0.12	9.75	0.30	0.88	9.80
Pirates	7.58	2.22	0.04	5.70	0.10	0.96	5.78
SPD	9.86	2.09	0.05	7.52	0.10	0.95	7.63
KSCM	11.21	2.26	0.04	12.06	0.10	0.96	12.03
CSSD	12.50	2.27	0.04	12.99	0.10	0.96	12.97
KDU-CSL	5.80	2.28	0.04	5.87	0.10	0.96	5.87
TOP 09	6.29	2.25	0.04	6.30	0.10	0.96	6.30
STAN	3.87	2.28	0.04	4.04	0.10	0.96	4.03

Table 5.1: Combined forecasts

Nevertheless, similar results are presented by Rothschild (2015), who weighs three types of data in his forecast - fundamental model, the polls, and prediction markets. Towards the election day, the weight on polls in the forecasts is over 80 %, with most of the remaining weight being put on the fundamental data (which we do not use here as it is not suitable for multi-party systems) and only a very little weight put on the prediction markets. In this sense the model for the combined forecast in this thesis is analogous.

5.5 Comparing forecasts

Now, we can compare the forecasts and determine, which was the most accurate. First, we look at the development of the polls' and odds' forecast in the months coming to the election. The filtered poll data comprise information from all previous polls, as the algorithm is updated with each new observation, whereas the odds are taken as a new forecast each time they were updated by the polling agency. Figure 5.7 shows the filtered poll results and odds forecast for ANO. The dashed line represents the actual vote share obtained by ANO in the election in October. We can see that the polls were relatively closer to the actual vote share during the whole time period, however, both polls and

odds, predicted a decline in the preferences close to the election, which did not turn out to be true.

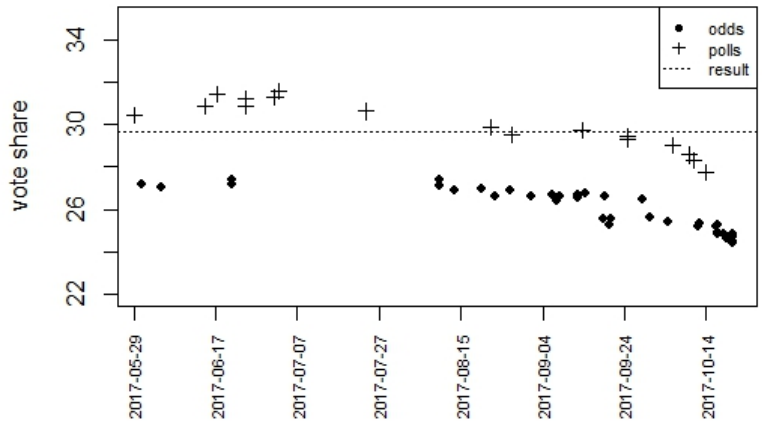

Figure 5.7: Comparison of polls and odds - ANO

For CSSD, the odds started higher than the polls, however, they got close over time and the odds forecast also detected a decline in preferences before the election. The filtered polls show only a slow steady decline over time. The comparison is shown in figure 5.8.

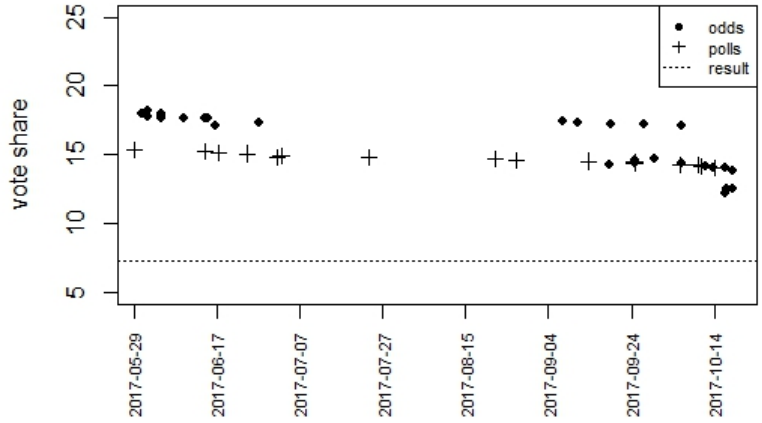

Figure 5.8: Comparison of polls and odds - CSSD

We also looked at the development of polled expected vote shares and odds forecasts for the Pirate Party shown in figure 5.9. The odds were predicting

higher vote share throughout the election cycle, than what was expected based on the conducted polls. Even though the odds' predicted vote share increases right before the election, it still underestimates the actual result. Especially for the Pirate party, we can see that the odds follow the polls quite closely. For ANO, on the other hand, the betting agencies might have been calculating with some bias in the polls, as they were expecting the resulting vote share to be lower, although it was in the end actually higher, than the polls suggested.

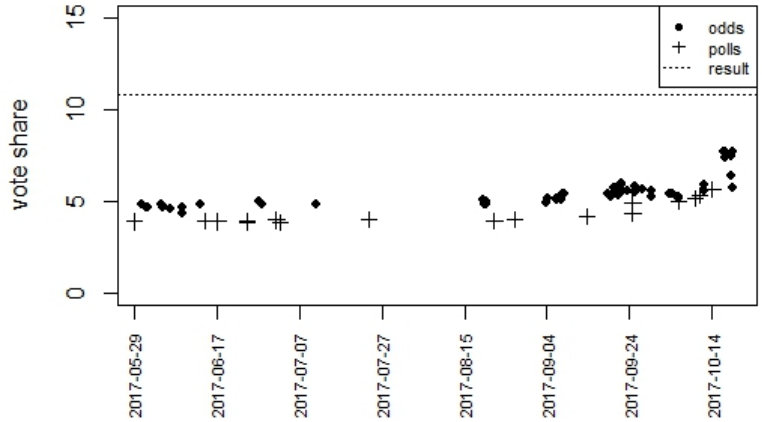

Figure 5.9: Comparison of polls and odds - Pirate party

To assess the accuracy of the forecasts, we will use two measures. The mean absolute error, calculated as

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

,

is the average absolute difference. We also calculate the root mean square error, which is the square root of averaged square errors.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}}$$

.

Both measures are scale-dependent, with the RMSE being much more sensitive to outliers. Table 5.2 shows the individual forecasts and their respective MAEs and RMSEs.

Comparing the results by the RMSE and MAE, the odds' forecast is in total

	polls	odds	combined	election results
ANO	27.66	24.63	25.10	29.64
ODS	9.75	10.12	9.80	11.32
Pirates	5.7	7.58	5.78	10.79
SPD	7.52	9.89	7.63	10.64
KSCM	12.06	11.21	12.03	7.76
CSSD	12.99	12.50	12.97	7.27
KDU-CSL	5.87	5.80	5.87	5.8
TOP 09	6.30	6.29	6.30	5.31
STAN	4.04	3.87	4.03	5.18
RMSE	10.61	8.83	12.23	
MAE	2.66	2.35	2.92	

Table 5.2: Forecasts' comparison

the closest to the actual results. This is mainly due to the fact, that the odds were able to predict increase in preferences for two parties, which were newly elected into the Chamber of Deputies, namely the Pirate party and SPD. On the other hand, polls were closer to the actual result for the biggest party ANO. The combined forecast performs the worst as would be expected based on the weights as it is mostly similar to the polls only with the exception of ANO, where it puts most of the weight on the odds, which were however further from the final results. As Rothschild (2015) notes, combining the forecasts brings the most advantage early on in the election cycle. The more time between the polls and the election the more uncertainty that brings to the forecasts. We did not deal with thoroughly with the timing of the polls and the increasing uncertainty, as we made forecasts based on the data available at the time just before the election. The time aspect should be further acknowledged when applying the model on any upcoming elections.

5.6 Discussion

From previous research, both types of data used in the forecasting model offer room for adjustments to make the forecasts more precise (Rothschild, 2015). The prices on betting markets are skewed by the margin of the betting agency, which needs to earn some profit. The prices may also be distorted due to the players participating behaving in a risk-loving manner (Wolfers & Zitzewitz, 2004). For polls, we need to properly weigh the individual polls to calculate the average (Silver, 2014). Some researchers also adjust the polls for present bias, e.g. Jackman (2005) who pools the polls, which reduces the sampling errors and adjusts them, distinguishing between the random sampling errors and non-

sampling errors, namely house effects, at which we will look more closely in the next section.

5.6.1 House effects

House effects as a term cover the differences between polls caused by the fact that polls are conducted by multiple different agencies. As was discussed in section 2.2.1, these can include different methodologies or bias of a polling agency towards one party. To analyze this effect, a set of factor variables was created, representing each of the polling agencies. Then, using the simple linear model, each party preferences were regressed on the set of the agencies' factors. The results shown in table 5.3 display significant effects for most agencies' polled party vote shares. This implies that polls from the given agency systematically tend to favour or disfavour certain parties with their vote share being higher or lower than in polls from other agencies. However, this does not mean that the agency might be purposely biased, but rather that the effects can be influenced by the timing of the polls, the sample selection, method in which the polls are conducted, or how the agency treats undecided voters.

The results suggest the need for further analysis of the house effects in the polling data. However, as compared to the United States, for example, where there are two major long-established parties and it can be shown that some polling agencies historically tend to favour e.g. the Republican candidates, the political situation in the Czech Republic is much more complex. With nine parties currently in the Chamber of Deputies, no clear division between the left and the right, and the biggest party with the most seats being established only five years before the last legislative election, the house effects are more difficult to estimate as there are not enough data. Nevertheless, in Jackman (2005), the author finds that up to 40 % of the variation between the polls is due to the house-to-house differences between the polling agencies, which supports the need to adjust the polls with regards to the polling agencies.

Table 5.3: House effects

	Dependent variable:							
	ANO	ODS	ČSSD	KSČM	Piráti	TOP 09	KDU-ČSL	SPD
Focus	−2.508**	2.421***	−6.283***	−0.121	2.075***	1.100	−1.829***	3.481***
	(1.243)	(0.694)	(1.733)	(0.698)	(0.595)	(0.784)	(0.527)	(1.032)
Kantar TNS	1.417	2.096***	−10.583***	−2.196***	1.175**	0.475	−0.479	3.406***
	(1.131)	(0.632)	(1.577)	(0.636)	(0.541)	(0.714)	(0.479)	(0.944)
Médea Research	−0.388	−0.489	−7.641***	−3.441***	3.347***	−0.388	−0.982*	2.296**
	(1.243)	(0.694)	(1.733)	(0.698)	(0.595)	(0.784)	(0.527)	(1.032)
MEDIAN	−4.850***	0.377	−3.507***	0.871**	0.775**	2.061***	−0.732**	0.911
	(0.688)	(0.384)	(0.959)	(0.386)	(0.329)	(0.434)	(0.291)	(0.631)
PPM Factum	−8.383***	−1.871***	−2.550*	2.904***	1.042**	2.642***	−0.263	0.756
	(1.050)	(0.587)	(1.465)	(0.590)	(0.503)	(0.663)	(0.445)	(1.391)
SANEP	−6.671***	1.121***	−2.142**	0.854**	0.725**	0.629	−0.517*	0.617
	(0.664)	(0.371)	(0.926)	(0.373)	(0.318)	(0.419)	(0.282)	(0.622)
STEM	−1.802**	−0.231	−7.474***	0.850*	1.802***	−0.743	−0.229	1.587**
	(0.838)	(0.468)	(1.168)	(0.471)	(0.401)	(0.529)	(0.355)	(0.714)
TNS Asia	−1.683**	0.338	−3.773***	−2.164***	0.331	1.696***	−0.792***	−0.187
	(0.707)	(0.395)	(0.985)	(0.397)	(0.343)	(0.446)	(0.299)	(0.665)
Constant	30.683***	8.504***	22.483***	12.796***	2.725***	6.725***	7.029***	3.794***
	(0.470)	(0.262)	(0.655)	(0.264)	(0.225)	(0.296)	(0.199)	(0.440)
Observations	118	118	118	118	117	118	118	93
R^2	0.624	0.322	0.436	0.581	0.328	0.338	0.147	0.262
Adjusted R^2	0.596	0.272	0.395	0.550	0.279	0.289	0.084	0.191
F Statistic	22.616*** (df = 8; 109)	6.473*** (df = 8; 109)	10.542*** (df = 8; 109)	18.871*** (df = 8; 109)	6.598*** (df = 8; 108)	6.946*** (df = 8; 109)	2.350** (df = 8; 109)	3.719*** (df = 8; 84)

Note: This table was created using the Stargazer package for R v.5.2.2. (Hlavac, 2018) *p<0.1; **p<0.05; ***p<0.01

Chapter 6

Conclusion

This thesis describes forecasting models for predicting election results. The focus is the 2017 legislative election in the Czech Republic, which has a multi-party political system, and for that reason the model outcome is a forecast of the exact vote shares each party will gain in the election. Election forecasting has a long tradition, however, it is always necessary to take into account the country's specific situation, such as the political system, the political stability, or the quality of polls. Unlike most forecasting models, which are built for two-party systems, this thesis adds to the literature on predicting election results in countries with multi-party systems, which pose some unique challenges for the researchers. Alongside with the polling data, prices on the betting market are used to predict the vote shares for each party which, to my knowledge, have not been done in academic literature before.

The model has three main parts. Firstly, the development of parties' preferences during the election cycle is analyzed based on vote intention polls. Using the dynamic linear model and Kalman filter, the underlying true preferences are retrieved from the observed polled data. The state-space model is then used to form a forecast for the election result. Secondly, odds from the betting market are transformed into probabilistic forecasts with the underlying uncertainty illustrated by the normal distribution. And thirdly, the polls and odds are combined into one forecast, treating the poll average as a Bayesian prior and updating it with the information gained from the odds. In the combined forecast, most of the weight is put on the long-term poll average, as its variance is low, especially for parties with low vote share. This originates from the nature of the Kalman Filter method, which smooths the fluctuations in the preferences. We calculate the forecasts right before the election, which means

that the latest polls should be a close representation of the preferences in the population at that time. If we were to calculate the forecasts more in advance, we would need to reflect that by increasing the variance of the poll forecast, because the polls always say only what the results would be, if the election was held in the time of the poll.

Averaging the polls helps to improve the performance supposing that there are random sampling errors. However, it will not help with non-random errors (Wright & Wright, 2018). According to previous findings in the literature, we, therefore, test the polls for the presence of the so-called *house effects* and find significant differences between polls conducted by different agencies. This suggests the need to further adjust the polls, for example using a similar method as Jackman (2005).

When comparing the forecasts' accuracy by calculating the mean average and root mean square errors, the betting odds were the closest to the actual election results. While the long-term poll average may be more precise in the long run, it reflects changes in the preferences short before the election with difficulties. On the other hand, the odds are much more flexible and can be updated right up to the election, which gives them an advantage. Nevertheless, we were comparing the final forecasts before the election and as the previous research has shown, combining different types of data in the forecasting model can bring advantage especially early on in the election cycle, whereas close to the election the forecast is usually mostly based on the polling data (Rothschild, 2015).

Even though Rothschild & Wolfers (2011) show that in the case of the United States, voters' expectations are more precise than the more generally used voters' intention polls, Ganser & Riordan (2015) conclude in their application of citizen forecasting in Germany, that simultaneously predicting eight interdependent vote shares is too complex an exercise for respondents of voters' expectations surveys. The results of this thesis offer a solution for countries, with such a dynamic political situation, in the form of betting markets, which aggregate the beliefs across the population and do not ask each individual to predict the outcome as a whole. As Wolfers & Zitzewitz (2004) have shown, betting markets incentivize the participants to gather and subsequently reveal truthful information. Even though the prices should be adjusted when translated into aggregated beliefs, depending on the market design and its participants' characteristics.

This thesis represents the first attempt to build a comprehensive forecasting

model based on polling data and betting markets prices in the Czech Republic. As for any good probabilistic forecast, the consistency is very important, so it would be recommended that the results should be tested on other upcoming elections. As the analysis suggests, the poll-average should be adjusted for house effects and other possible factors as well. Further research is also needed into transforming the odds into exact vote share forecasts. Finally, the timing of the forecasts should be addressed, as the more in advance of the election itself we are able to form a precise forecast the better.

Bibliography

Aichholzer, J. & J. Willmann (2014): "Forecasting austrian national elections: The grand coalition model." *International journal of forecasting* **30(1)**: pp. 55–64.

Babyak, M. A. (2004): "What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models." *Psychosomatic medicine* **66(3)**: pp. 411–421.

Berg, J. E. & T. A. Rietz (2019): "Longshots, overconfidence and efficiency on the iowa electronic market." *International Journal of Forecasting* **35(1)**: pp. 271–287.

Brown, L. B. & H. W. Chappell Jr (1999): "Forecasting presidential elections using history and polls." *International Journal of Forecasting* **15(2)**: pp. 127–135.

Campbell, J. E. & M. S. Lewis-Beck (2008): "Us presidential election forecasting: An introduction." *International Journal of Forecasting* **24(2)**: pp. 189–192.

Condorcet, M. d. (1785): "Essay on the application of analysis to the probability of majority decisions." *Paris: Imprimerie Royale* .

Erikson, R. S. & C. Wlezien (1999): "Presidential polls as a time series: the case of 1996." *Public opinion quarterly* pp. 163–177.

Fowler, A. (2018): "A bayesian explanation for the effect of incumbency." *Electoral Studies* **53**: pp. 66–78.

Ganser, C. & P. Riordan (2015): "Vote expectations at the next level. trying to predict vote shares in the 2013 german federal election by polling expectations." *Electoral Studies* **40**: pp. 115–126.

Gjerstad, S. & M. Hall (2005): "Risk aversion, beliefs, and prediction market equilibrium." *Economic Science Laboratory, University of Arizona* .

Green, D. P., A. S. Gerber, & S. L. De Boef (1999): "Tracking opinion over time: A method for reducing sampling error." *Public Opinion Quarterly* pp. 178–192.

Grofman, B., G. Owen, & S. L. Feld (1983): "Thirteen theorems in search of the truth." *Theory and decision* **15(3)**: pp. 261–278.

Hlavac, M. (2018): "stargazer: Well-formatted regression and summary statistics tables." *R package version 5.2.2* .

Jackman, S. (2005): "Pooling the polls over an election campaign." *Australian Journal of Political Science* **40(4)**: pp. 499–517.

Lewis-Beck, M. S. & A. Skalaban (1989): "Citizen forecasting: can voters see into the future?" *British Journal of Political Science* **19(1)**: pp. 146–153.

Lewis-Beck, M. S. & M. Stegmaier (2011): "Citizen forecasting: Can uk voters see the future?" *Electoral Studies* **30(2)**: pp. 264–268.

Lewis-Beck, M. S. & C. Tien (1996): "The future in forecasting: Prospective presidential models." *American Politics Quarterly* **24(4)**: pp. 468–491.

Lewis-Beck, M. S. & C. Tien (2018): "Candidates and campaigns: How they alter election forecasts." *Electoral Studies* **54**: pp. 303–308.

List, C. & R. E. Goodin (2001): "Epistemic democracy: Generalizing the condorcet jury theorem." *Journal of political philosophy* **9(3)**: pp. 277–306.

Manski, C. F. (2006): "Interpreting the predictions of prediction markets." *economics letters* **91(3)**: pp. 425–429.

Martin, E. A., M. W. Traugott, & C. Kennedy (2005): "A review and proposal for a new measure of poll accuracy." *Public Opinion Quarterly* **69(3)**: pp. 342–369.

Mitofsky, W. J. (1998): "Was 1996 a worse year for polls than 1948?" *The Public Opinion Quarterly* **62(2)**: pp. 230–249.

Morwitz, V. G. & C. Pluzinski (1996): "Do polls reflect opinions or do opinions reflect polls? the impact of political polling on voters' expectations, preferences, and behavior." *Journal of Consumer Research* **23(1)**: pp. 53–67.

Mosteller, F. *et al.* (1949): "The pre-election polls of 1948; report to the committee on analysis of pre-election polls and forecasts.(bull. 60)." .

Murr, A. E. (2011): ""wisdom of crowds"? a decentralised election forecasting model that uses citizens' local expectations." *Electoral Studies* **30(4)**: pp. 771–783.

Norpoth, H. & T. Gschwend (2010): "The chancellor model: Forecasting german elections." *International Journal of Forecasting* **26(1)**: pp. 42–53.

Pickup, M. & R. Johnston (2005): "Measurement error and house bias in 2004 presidential campaign polls." In "Annual Meeting of the American Political Science Association, Washington, DC, September," pp. 2–5.

Reade, J. J. & L. V. Williams (2019): "Polls to probabilities: Comparing prediction markets and opinion polls." *International Journal of Forecasting* **35(1)**: pp. 336–350.

Rothschild, D. (2009): "Forecasting elections: Comparing prediction markets, polls, and their biases." *Public Opinion Quarterly* **73(5)**: pp. 895–916.

Rothschild, D. (2015): "Combining forecasts for elections: Accurate, relevant, and timely." *International Journal of Forecasting* **31(3)**: pp. 952–964.

Rothschild, D. & N. Malhotra (2014): "Are public opinion polls self-fulfilling prophecies?" *Research & Politics* **1(2)**.

Rothschild, D. & J. Wolfers (2011): "Forecasting elections: Voter intentions versus expectations." *Available at SSRN 1884644* .

Silver, N. (2014): "How fivethirtyeight calculates pollster ratings." `https://fivethirtyeight.com/features/how-fivethirtyeight-calculates-pollster-ratings/`. Accessed: 2019-07-27.

Silver, N. (2016): "A user's guide to fivethirtyeight's 2016 general election forecast." `https://fivethirtyeight.com/features/a-users-guide-to-fivethirtyeights-2016-general-election-forecast/`. Accessed: 2019-07-27.

Sudman, S. (1986): "Do exit polls influence voting behavior?" *Public Opinion Quarterly* **50(3)**: pp. 331–339.

Temporão, M., Y. Dufresne, J. Savoie, & C. van der Linden (2019): "Crowdsourcing the vote: New horizons in citizen forecasting." *International Journal of Forecasting* **35(1)**: pp. 1–10.

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

Wang, W., D. Rothschild, S. Goel, & A. Gelman (2015): "Forecasting elections with non-representative polls." *International Journal of Forecasting* **31(3)**: pp. 980–991.

Wlezien, C. & R. S. Erikson (2003): "The evolution of electoral preferences: What the polls reveal as the campaign unfolds." .

Wolfers, J. & E. Zitzewitz (2004): "Prediction markets." *Journal of economic perspectives* **18(2)**: pp. 107–126.

Wright, F. A. & A. A. Wright (2018): "How surprising was trump's victory? evaluations of the 2016 us presidential election and a new poll aggregation model." *Electoral Studies* **54**: pp. 81–89.

Zukin, C. (2004): "Sources of variation in published election polling: A primer." In "ideas," .

Appendix A

List of the political parties

List of the acronyms used in the thesis for each party, together with their Czech official name.

- ANO - ANO 2011
- ODS - Občanská demokratická strana
- Pirates - Česká pirátská strana
- SPD - Svoboda a přímá demokracie - Tomio Okamura
- KSCM - Komunistická strana Čech a Moravy
- CSSD - Česká strana sociálně demokratická
- KDU-CSL - Křestanská a demokratická unie - československá strana lidová
- TOP 09
- STAN - Starostové a nezávislí