Zumofen Guillaume, 11-206-950
Route de la Crête Blanche 28 B
3977 Granges
guillaume.zumofen@ipw.unibe.ch

Github ZumofenG/SwissParliament_CAS2020

Data Science Project

Vote Prediction and Representation Quality in the Swiss Parliament

Conceptual Design Report

14 October 2020

Abstract

The objective of this Data science project is twofold. First, it develops a method to predict vote in the Swiss Parliament. It highlights the relevance -or not -of legislative bargaining. Second, it measures the quality of political representation in the Swiss Parliament. It evaluates congruence between citizens' preferences and legislators' decision. To follow these goals, I exploit the Open Data Parliamentary Web Services and combine it with survey data on citizens' preferences. I use a machine learning method to predict vote and hypothesis testing to measure political representation.

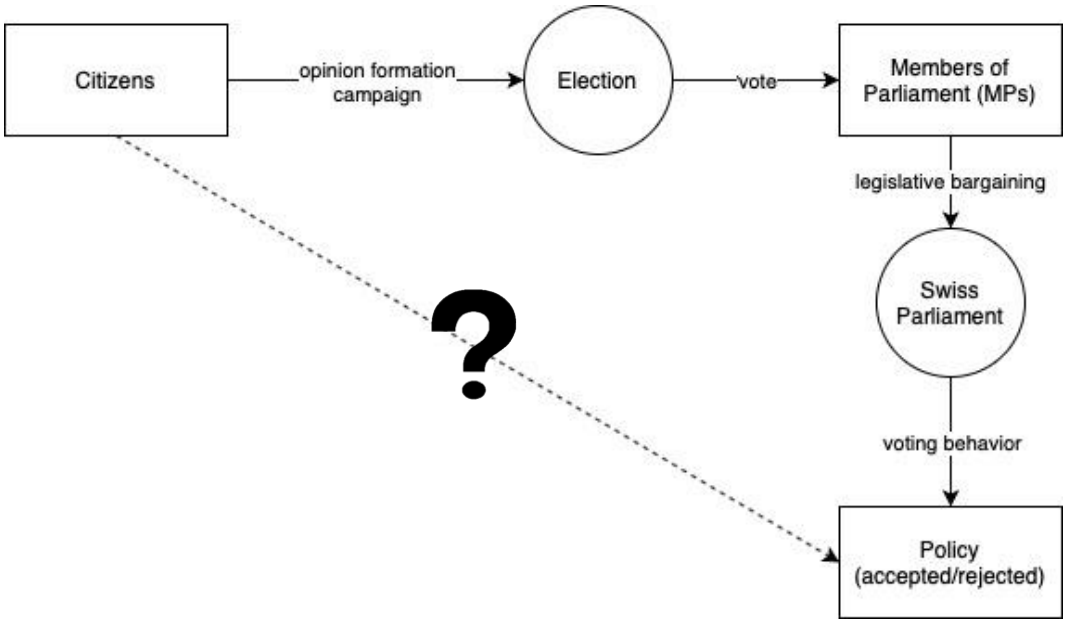
Table of Contents

1 Project Objectives

Measuring political representation, i.e., do politicians really follow their constituents' preferences, is of paramount importance in a representative democracy. In such political system, citizens elect Members of the Parliament (MPs) to represent their preferences and implement futures policies (prospective voting). The ideal of a representative democracy is a congruence between legislators' decisions (MPs) and constituents' preferences (citizens). In other words, citizens expect MPs to "do" what they "want". However, a persistent public perception is that a positional gap exists between MPs policies' decisions and citizens' preferences. Such perception is driven by a declining confidence in political institutions and actors, a recurrent trade-off between vote-, policy- and office-seeking for MPs, the growing importance of campaign with overbidding electoral pledges, an increasingly polarized party system, and a mediatization of politics.

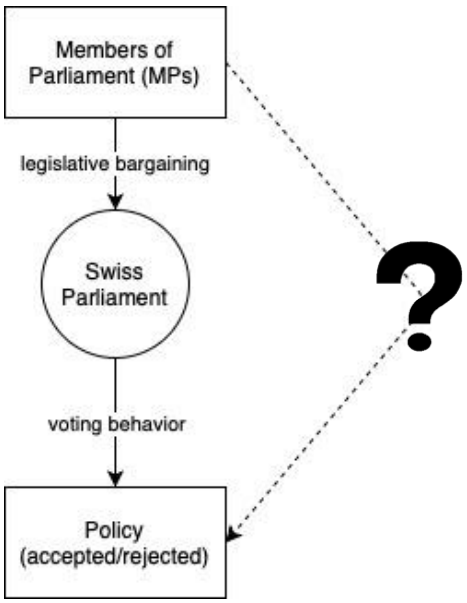
The major problem of any empirical analysis of political representation is the absence of a direct measure of constituents' preferences for each real policy proposals. This impede a comparison with legislators' decisions on real policies. Thus, literature uses indirect measures, e.g., ideology scores, but fails to measure accurately the quality of political representation. Figure 1 is a conceptual representation of a representative democracy. It emphasizes the process leading to policy implementation, and the lack of direct measure of the quality of political representation.

Figure 1. Conceptual representation of the Swiss representative democracy – Quality representation



During the Covid-19 pandemic crisis, the Parliament sat on an extraordinary session from 04th to 08th of May. This extraordinary session costed approximately 1,5 millions CHF. An intense debate upsurged in the society. The question which arises is: is legislative bargaining in the Parliament worth spending 1,5 millions CHF? For example, one might claim that MPs are only following party recommendations to vote. In other words, is it possible to easily predict MPs vote, which would make seating in the Parliament useless, or is legislative bargaining actually influencing vote in the Parliament.

Figure 2. Conceptual representation of the Swiss representative democracy – Vote Prediction



The goal of this project is then to assess not only the quality of representation but also to find a method to predict vote in the Swiss Parliament. The project is then twofold:

First, it benefits from the open data Parliamentary Services. Then, an R package (SwissParl) developed by Grünenfelder Zumbach GmbH provides an interface to access the most important data on parliamentary activities. This generates a dataframe with the most relevant variables driving voting behavior and legislative bargaining in the Parliament. Using this dataset, the project exploit machine learning to predict vote in the Parliament.

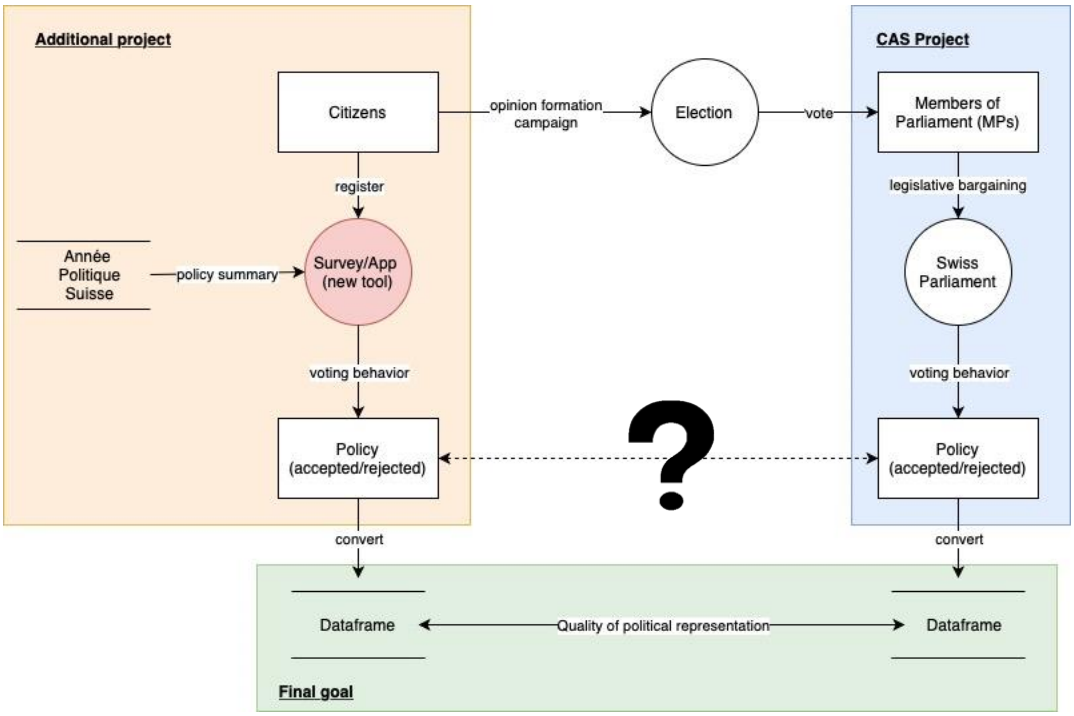
Second, the objective is to develop a survey/App (tool) providing a summary of all policies filed in the Parliament, and allowing users to vote directly on these policies. The policies summaries are extracted from the Année Politique Suisse database.[1] The newly-build tool provides a dataset with a direct measure of citizens' preferences. Merging this new dataset with the Swiss

[1] Année Politique Suisse offers, since 1965, a precise, factual and objective description of all political evolution in Switzerland, focusing specifically on legislative bargaining within the Swiss Parliament.

Parliament dataset (SwissParl), I am able to directly evaluate the quality of political representation in Switzerland.

Figure 3 conceptually demonstrate how the project exploit the two datasets to evaluate the quality of political representation.

Figure 3. Conceptual representation of the project



As part of the CAS project, I focus at first on the objective related to predicting vote in the Parliament (see Figure 3 blue part). Namely, the goal is to exploit machine learning to predict vote in the Swiss Parliament. To do so, I use the SwissParl R package to extract all relevant data from 1995 to 2018 about voting behavior and legislative bargaining in the Swiss Parliament.

→ **1st objective = Predict vote in the Parliament.**

Second, I develop a survey to run a pre-test on the second objective, i.e., measure the quality of political representation in the Swiss Parliament. This survey allows to obtain data on citizens' preferences. It focuses on two specific topics in the Parliament: Economy and Monetary Policies (Chapter 4a and 4b in Année Politique Suisse). Then, it uses both the Parliamentary Service (Curia Vista) and the Année Politique Suisse database to provide respondents with a summary of the text, and a summary of pro and contra arguments of each policies which will be considered in the next Parliamentary session. At the end, it records citizens' preferences regarding this future policies. This survey is a first step (pre-test) before developing an App able

to record citizens' preferences for all topics and all future policies. Such App would use machine learning and language natural processing combined with human proofreading to develop these summaries.

→ **2nd objective = Pre-test on quality of political representation.**

2 Methods

In order to be developed, the project benefit from a computer infrastructure (1), a data acquisition process (2) and specific statistical methods (3).

2.1 Infrastructure

The computation is done on a computer. The processing hardware is a MacBook Pro (2,7 GHz Intel Core i5) with 250 Go memory. As for the 1st part (Predict vote in the Parliament and Pre-test with a survey), it is assumed that the planned data storage is largely sufficient.[2] To be precise, the .csv file dataset used to Predict vote in the Parliament is 506 Ko.

The data analysis is run with Python 3 on the Jupyter Notebook environment (Version 6.0.3). The interface environment is launched via the Anaconda Navigator (1.9.12). Additionally, data acquisition is performed via the RStudio software (1.1.414) and data preparation was partially done with Stata software (15.1)

The libraries used in Python are: pandas, numpy, matplotlib, pyreadr, scipy and sklearn.

The R package used for data acquisition is: SwissParl.

The survey is run with the Qualtrics software.

2.2 Data acquisition[3]

For the first objective "Predict vote in the Parliament", data are extracted from the open data Parliamentary Services with the SwissParl R package. It provides a relational database on the period 2011 to 2018. I also extract .xlsx dataset directly from the parliament website which I merge with my Parliamentary dataset.

[2] If necessary, the data will be stored on the UBELIX servers (Bern University Service Provider) when the App will be developed.

[3] To be honest, data acquisition did not run as expected. I spent days working on the Open Data Parliamentary services, trying to directly extract data with Python. It was only half a success. Due to time restriction, I ended up stepping back to Stata and R to run some extraction. I hope that I will find a cleaner approach before February (M3), using only Python. This should provide a larger dataset with additional variables.

For the second objective "Pre-test on the quality of political representation", data will be collected with a survey which will be sent to a representative national sample with the Qualtrics software. At the end, it will provide a relational database which will then be merged with Parliamentary data.

2.3 Statistical methods

The project will use machine learning methods to predict vote in the Parliament (M3 CAS). Additionally, it will use descriptive statistics - and domain expertise - to identify variables of interest and the distribution of YES and NO vote regarding policies in the Parliament.

Then, it will use hypothesis testing to measure the quality of political representation. Hypothesis testing will be run at the individual-level and the level of the electorate. First, each survey respondent elected some specific MPs. Then, I test hypotheses between a respondent's preferences and the voting behavior on policies of the MPs he/she elected. Second, at the level of the electorate, I will test hypotheses between the aggregate respondents' preferences and the Parliament decision on policies.

3 Data

The dataset *data_swissparl* is extracted from the Open Data Parliamentary Services. I count N=1103 observations (rows) and 89 variables (columns). Due to time restrictions and difficulty in data acquisition (see point 2 above), I restricted my dataset to years 2011 to 2018 (*legislative periods* = 49 and 50), to only *motion* and *postulat* affairs, to only german-speaking affairs and to only the National council.

The main dependent variables are *yes_vote* and *no_vote*. To be accurate, I also consider *abstain_vote* and *missing_vote.* These dependent variables are continuous variables varying from 0 to 200.

3.1 Descriptive statistics

Table 1 and Figure 4 (Boxplot) display the main descriptive statistics of my dependent variables. On average, the probability for a motion or a postulat to be rejected is slightly higher (0.58) than the probability to be accepted (0.42). Mean Yes vote is equal to 81.40 and mean No vote is equal to 100.89. In other words, politics is mostly conservative. This is not surprising as the Swiss Parliament- on legislative periods 49th and 50th – was slightly leaning on the right-conservative seats, with the CVP and SVP obtaining a lot of seats. Furthermore, I detect no significant differences between motion and postulat. Finally, abstaining is not a very common

strategy (2% of MPs) (mean = 4.60= and on average 5% of MPs are missing votes (mean=9.75, with 200 seats).

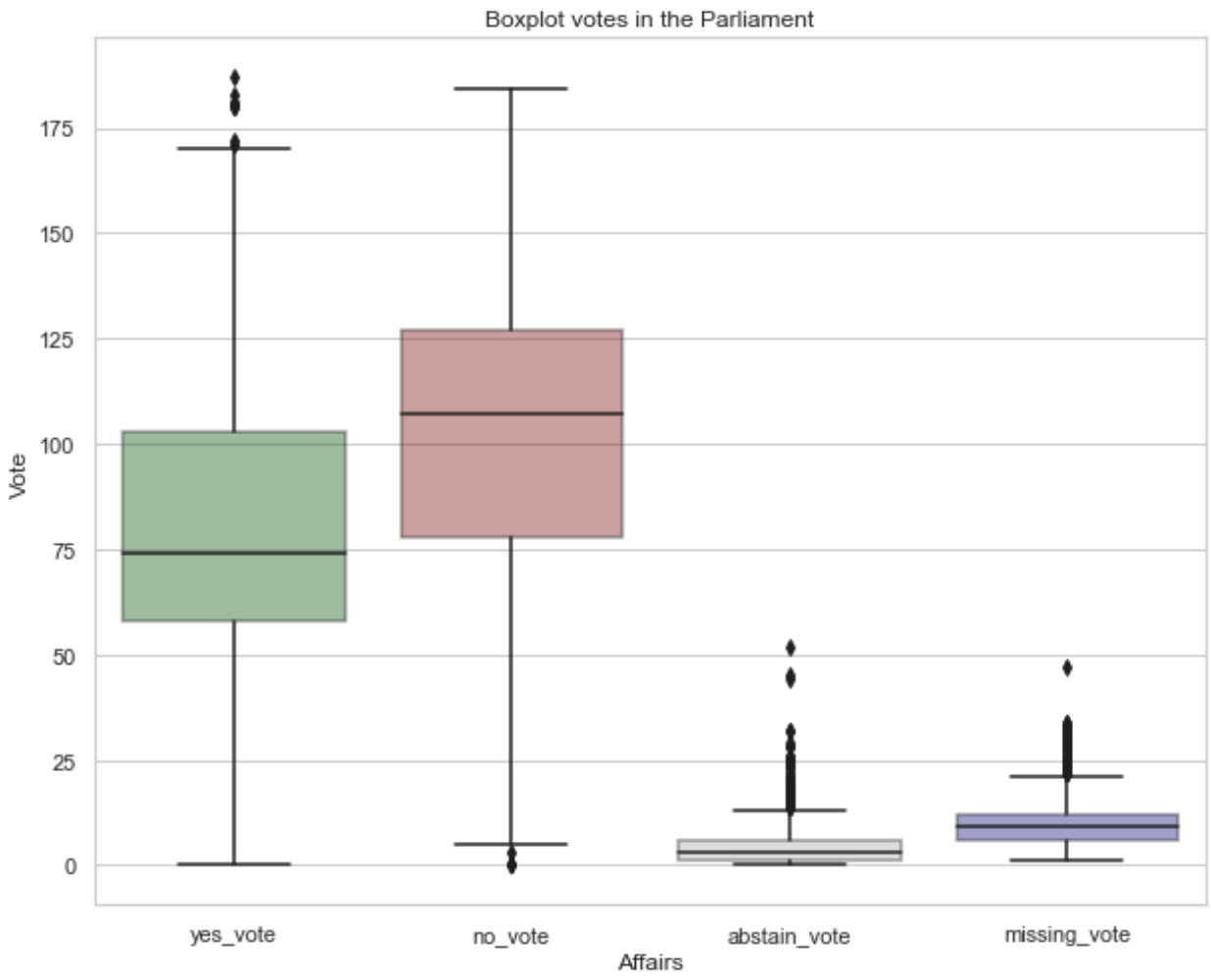
As independent variables, I consider: *sex_author* (categorical), *department* (categorical), *commission* (categorical), *vote_hour* (ordinal), *legislative_periods* (binary), *affair_type* (binary), *years_parliament* (ordinal), *party_author* (categorical), *canton_author* (categorical)., *author_affilation_party_representation* (continuous) Additionally, I measure the party representation in each legislative periods, counting seats: *CVP_parl* (continuous), *FDP_parl* (continuous), *SP_parl* (continuous), *SVP_parl* (continuous), *GPS_parl* (continuous), *GLP_parl* (continuous), *BDP_parl* (continuous) and *Other_parl* (continuous).[4]

Table 1. Descriptive statistics – Vote in the Parliament

	Mean	Median	Standard deviation	Min	Max
Yes	81.40	74.00	32.12	0.00	187.00
No	100.89	107.00	33.00	0.00	184.00
Abstain	4.60	3.00	5.44	0.00	52.00
Missing	9.75	9.00	5.89	1.00	47.00

Figure 4. Boxplot – Vote in the Parliament

[4] See Appendix for further details.



3.2 Distribution

Kurtosis and Skewness are two measures to evaluate the distribution of my dataset (see table 2).

Looking first at the 'tailedness' of the probability distribution, I conclude that the variables *yes_vote* and *no_vote* meet a normal distribution expectation (-1 < kurtosis < 1). To the contrary, the distribution of *abstain_vote* and *missing_vote* is too 'peaked' (kurtosis >1).

Then, I detect that *yes_vote* is slightly skewed on the right and *no_vote* is slightly skewed on the left. This is no surprise considering that the probability of a yes vote is smaller than probability of a no vote (conservative politics). In addition, the distribution of both *abstain_vote* and *missing_vote* is skewed on the right. Again, it meets our expectation as abstaining and missing a vote is rare in the Parliament.

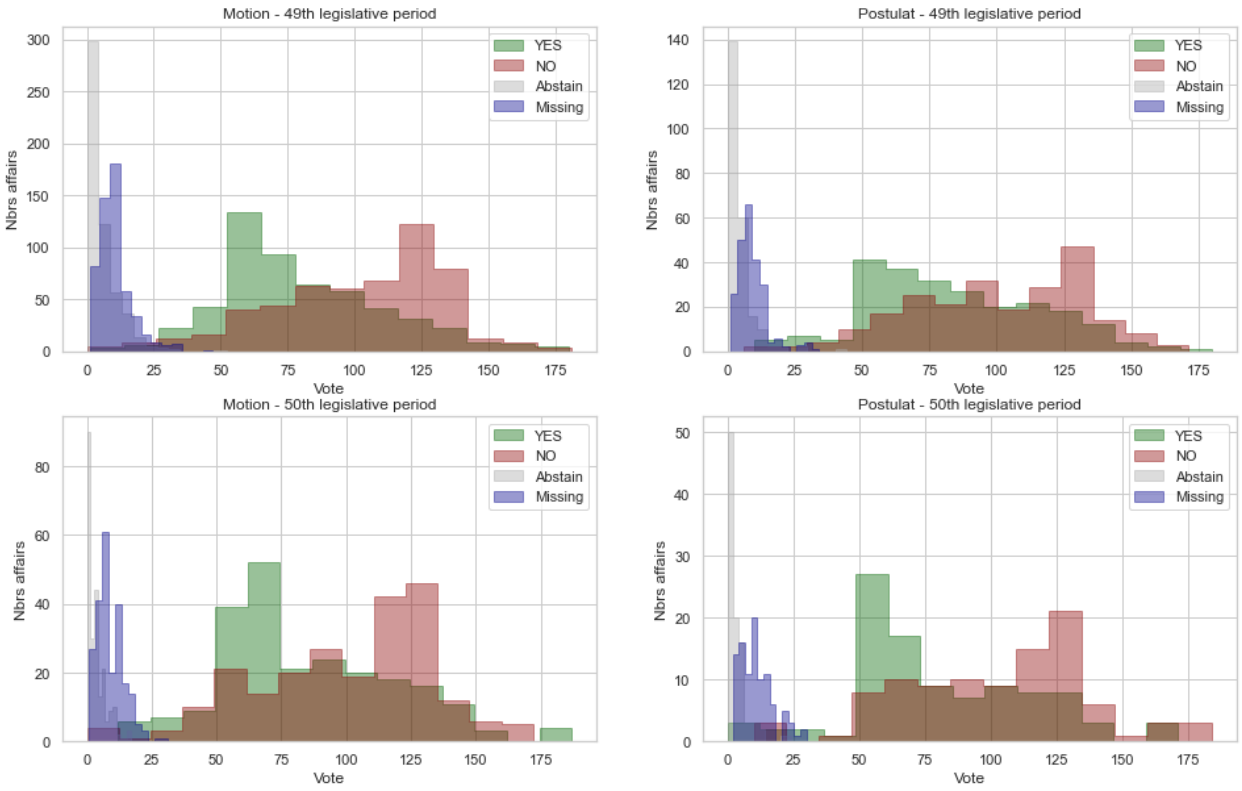
Table 2. Kurtosis and skewness – Vote in the Parliament

	Kurtosis	**Skewness**

Yes	0.07	0.54
No	-0.08	-0.53
Abstain	12.93	2.75
Missing	3.72	1.56

Figure 4 displays histograms of my dependent variables. It is divided by motion and postulat, and by legislative periods. I detect no meaningful differences between motion and postulat, and between legislative period 49th and 50th. I conclude that my dependent variables yes and no vote follow a normal distribution.

Figure 4. Histogram – Vote in the Parliament – By affair and by legislative period



3.3 Security

Data are extracted from the Open Data Parliamentary Services (Parliamentary Services of the Federal Assembly, Bern). They are publicly available. The Open Data Parliamentary Services provides a few guidelines:

- Publication should not give the impression of being an official Parliamentary publication
- Content cannot be altered
- Date of data extraction must be indicated
- The sole property of the content remains to the Parliamentary services or any other legal copyright holder.

Furthermore, the author accepts full liability for publication or distribution. Apart from these elements, no legal, ethical or security obstacles exist.

4 Metadata

The metadata required to reproduce the analysis are stored in the Stata do-file and R-script file. It provides all necessary information regarding the variables. These files are stored on Github repository ZumofenG.

Information concerning the swissparl R package are stored on the Github repository Zumbov2.

Then, the dataset are also stored on the Github repository ZumofenG.

5 Data Quality

The dataset is extracted from the Open Data Parliamentary Services. Considering the public relevance of such data, it can be assumed that quality and precision is met. Indeed, Parliamentary Services are under public scrutiny and is liable for all publication regarding legislative bargaining. Therefore, I consider that the measurement precision of my dependent variables (*yes_vote*, *no_vote*) is of high quality and is hardly questionable.

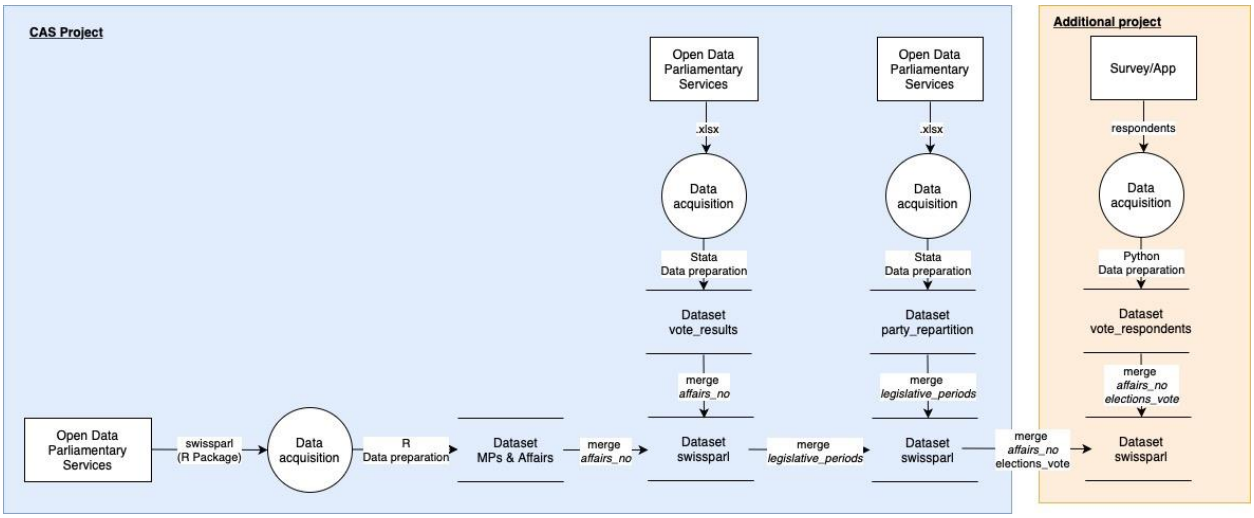
To meet my objectives, I need a dataset as large as possible. Currently, I restricted my focus on data going only from 2011 to 2018, to only motion and postulat, to only german-speaking affairs and to the National council only. This provides a comparatively small dataset (N=1450). It will be necessary to extend this dataset. First, I need to incorporate data going from 1996 to 2020. Second, it is necessary to also incorporate parliamentary initiative, cantonal initiatives and Federal council affairs. Third, french- and italian-speaking affairs have to also be included. Fourth, the inclusion of the Council of States is a must. This will strongly improve the quality of my dataset.

Additionally, it is worth including as many explanatory variables as possible. To begin with, it is necessary to include *topics* in my analysis. Second, the opinion of the Federal council is a variable that has to be included. Then, using natural language processing, it would be also beneficial to include legislative debates and affair summary. Regarding the incorporation of text-as-data variables, the quality of transcription, translation and the method used is of prime importance to meet the data quality requirements.

6 Data Flow

For the first objective "Predict vote in the Parliament", data are extracted from the open data Parliamentary Services with the SwissParl R package. It provides a relational database on the period 2011 to 2018. Then, I build a second dataset obtaining data on parliament vote results from the open data Parliamentary Services. These data are .xlsx files for each session since 2011 (Wintersession). The dataset is merged using the *affair_no* variables. I restrict my focus only on "motion" and "postulat" affairs. Finally, this dataset is merged with data on political party representation in the Parliament. This dataset is also extracted from the open data Parliamentary service (.xlsx file). It is merged with my master dataset using the variable *legislativePeriods*. Figure 5 describes the data acquisition process.

Figure 5. Data acquisition process



For the second objective "Pre-test on the quality of political representation", data will be collected with a survey which will be sent to a representative national sample with the Qualtrics software. The survey content will be elaborated with the Année Politique Suisse database and the Curia Vista website content (url request) with machine learning method and natural language processing, combined with human proofreading. At the end, it will provide a relational database which will then be merged with Parliamentary data.

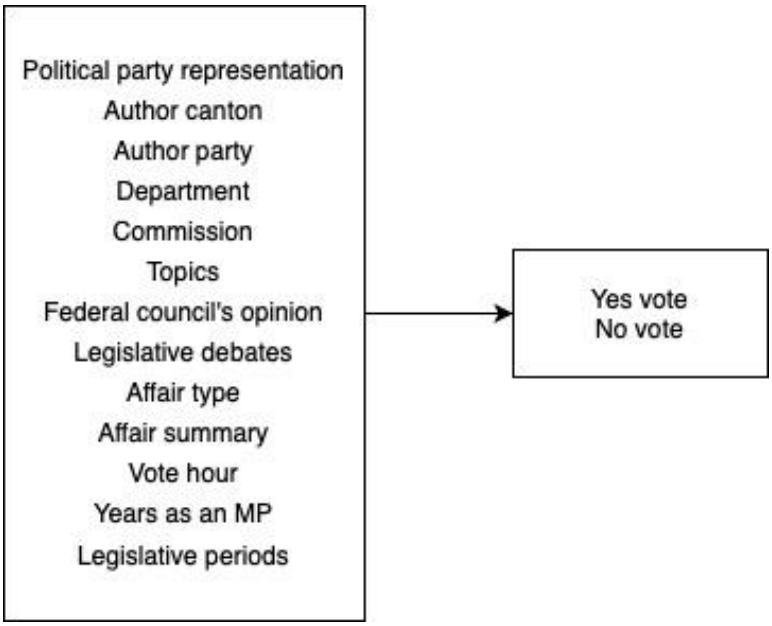
7 Data Model[5]

7.1 Conceptual model

The objective of the model is to predict MPs vote in the Parliament. My dependent variables are yes_vote and no_vote on the period 2011-2018, on Motion and Postulat, on German-speaking affairs and in the National council. I consider as explanatory variables Political party representation in the Parliament, Author canton, Author sex, Author party, Author's party seats in the Parliament, Department, Commission, Topics, Federal Council's opinion, Legislative debates, Affair type, Affair summary, Vote hour, Years as an MP and Legislative periods.

Figure 6. Conceptual model

[5] For this section, I consider only my 1st objective – Vote prediction in the Parliament.



7.2 Logical model

The first dataset (voten.rds) is extracted from the Open Data Parliamentary Services with the R package swissparl.

The second dataset is an .xlsx file extracted from the Parliament website (vote_parliament.xlsx), preprocessed with Stata and merged with the master dataset using the variable *affairs_no*.

The third dataset is also an .xlsx file extracted from the Parliament website (parliament_party.xlsx), preprocessed with Stata and merged with the master dataset using the variable *legislative_periods*. This provides a dataset with N=1450 observations (rows) and 55 variables (columns)

In a next step, a fourth dataset will be added with legislative bargaining texts and affair summary texts. These will be extracted from the Open Data Parliamentary Services using url request with Python.

7.3 Physical model[6]

[6] For the 2nd objective (Quality representation), the data will be stored on the UBELIX servers (Bern University Service Provider) when the App will be developed (if necessary).

Given the size of the dataset, a laptop is sufficient. The dataset will then be stored on my laptop.

Pprocessing hardware is a MacBook Pro (2,7 GHz Intel Core i5) with 250 Go memory. As for the 1st part (Predict vote in the Parliament), the .csv file dataset used to Predict vote in the Parliament is 506 Ko.

The data analysis is run with Python 3 on the Jupyter Notebook environment (Version 6.0.3).

The interface environment is launched via the Anaconda Navigator (1.9.12).

Additionally, data acquisition is performed via the RStudio software (1.1.414) and data preparation was partially done with Stata software (15.1).

8 Risks

The Open Data Parliamentary Services will be developed on a brand-new platform and provide additional data starting in Spring 2021. This future development is pivotal for my project. A delay in the implementation of this new platform would be detrimental to the project. To face this potential risk, I already extracted data from the http://ws-old.parlament.ch url to improve my understanding of the available data.

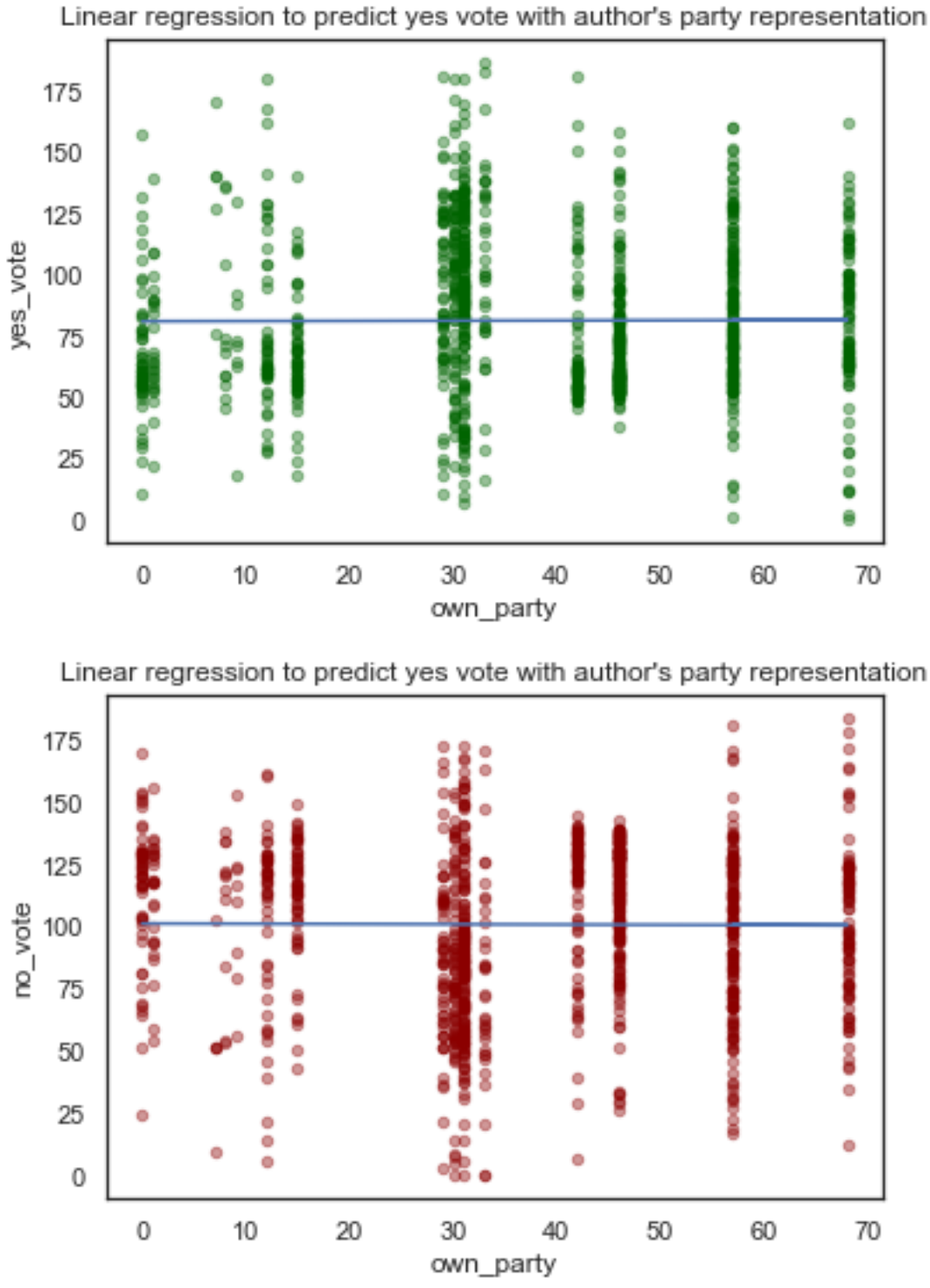
The development of a survey/App which measures citizen's true opinion on multiple parliamentary affairs is dependent on the insertion of the *Année Politique Suisse* database in my project. This *Année Politique Suisse* must be up-to-date to be relevant for the respondents. However, the accuracy and 'updateness' of this database is out of my control. To reduce risks, I plan on combining this dataset with the affair summaries provided via the Open Data Parliamentary service.

9 Preliminary Studies

First, I ran a linear regression to determine if the affair author's political party relative size within the Parliament (seats) influences vote in the Parliament. Figure 7 displays the results.

Based on the results, I detect no significant influence of an author's political party representation on the probability to obtain more vote. Such conclusion confirms that Swiss politics is about compromises and concordance. Being in the larger party is not sufficient to have affair adopted. This would mean that legislative bargaining is of prime importance to reach enough yes vote. This is good news for democracy.

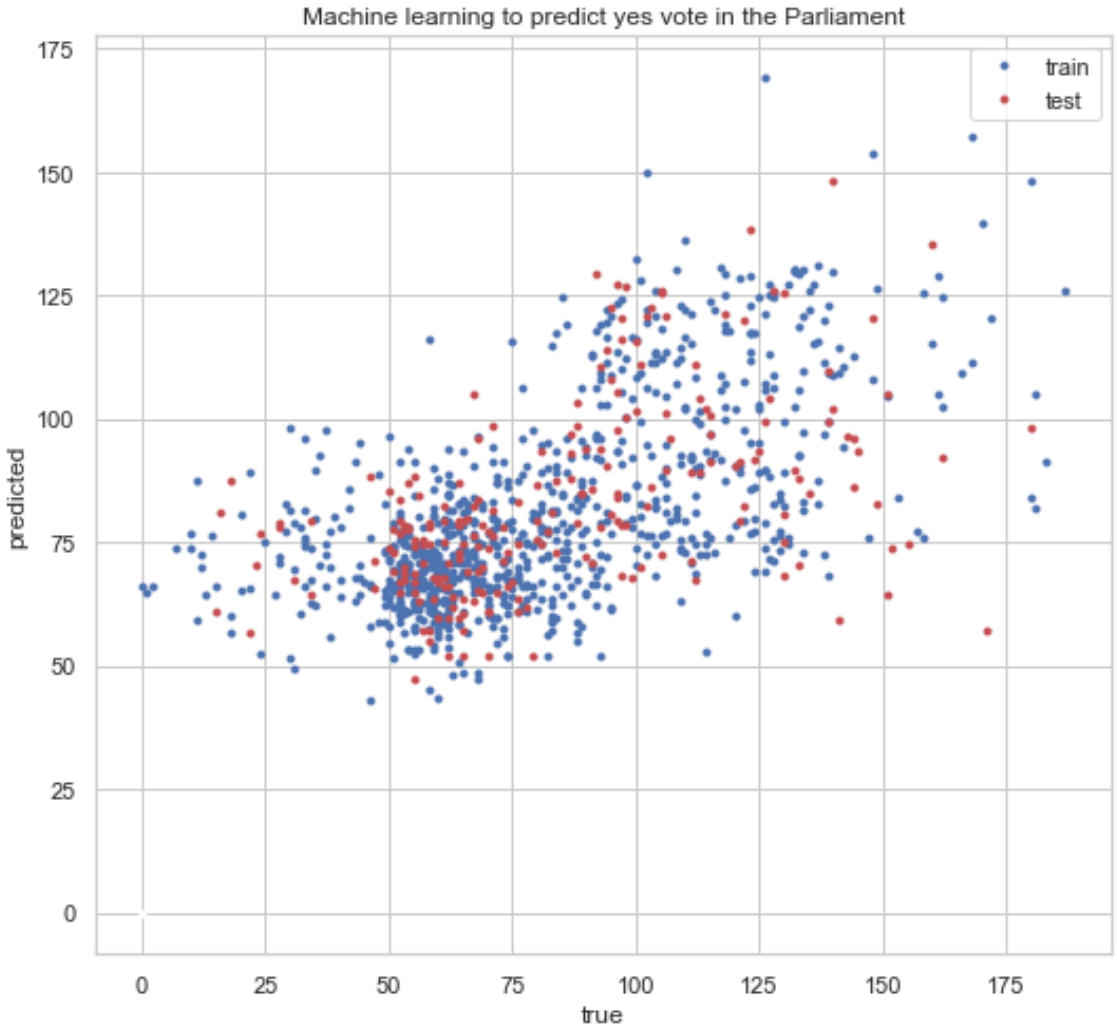
Figure 7. Linear regression – Yes/No vote and author's party representation in the Parliament

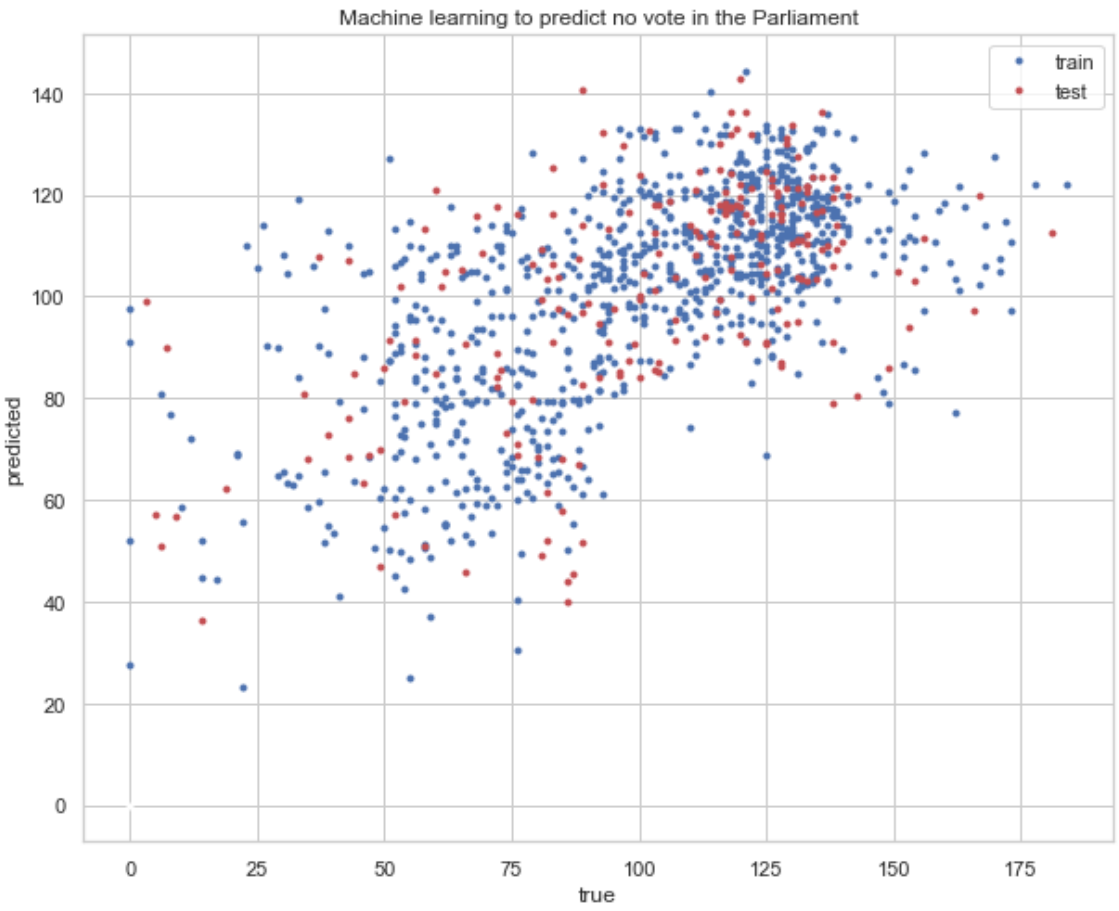


Second, I develop a first (and rather crude) machine learning model to predict yes and no vote in the Parliament. Figure 8 displays the graphical results of the machine learning model.

This first machine learning model is half a failure. It does not predict yes/no vote accurately. To be precise, for Yes vote I used 20% of my dataset as a test set and obtain mean squared error of the test and the train which are respectively equal to 28.4 and 25.0. To measure the accuracy of my model, I consider the rmse (root mean squared error). In this case, it is approximately equal to 5. Thus, my model predicts Yes votes within a range of 10 votes approximately (-5 to +5). Then, for No vote I used 20% of my dataset as a test set and obtain mean squared error of the test and the train which are respectively equal to 27.8 and 25.4. Similarly, I consider the rmse to measure accuracy. Thus, my model predicts No votes within a range of 10 votes approximately (-5 to +5).

Figure 8. Machine learning – Predict vote in the Parliament





At the end of the day, the goal is to determine if a policy is accepted or rejected, and not how many yes/no votes were casted. Therefore, my model is able to predict accurately vote decision (accepted/rejected) for vote who have a larger spread than 20 (10 yes variation + 10 no variation) regarding yes and no vote. This occurred in 905 out of 1103 votes in my dataset. It means that my model predicts an accurate decision (accepted or rejected) in 82.0% of the time. Though the mse is pretty large, this conclusion is possible because, in most cases, the spread between yes and no vote is large (bigger than 20). However, this could be improved by including additional independent variables such as Topics, Federal Council's opinion, Legislative debates and Affair summary.

10 Conclusions

The goal of this Data science project is not only to Predict vote in the Parliament but also to measure the quality of political representation. To do so, I benefit from the Open Data Parliamentary Web Services to extract data regarding legislative bargaining in the Swiss Parliament. In addition, I combine this dataset with survey data on citizen's preferences on policies.

I conclude that Yes and No votes in the Swiss Parliament follow a normal distribution. Then, I confirm that Swiss politics is conservative. Motion and Postulate obtain on average 85.5 Yes and 95.8 No votes.

My first attempt to predict vote in the Parliament reach an accurate conclusion for 82% of affairs. The mean squared error is approximately equal to 25 votes. First, this preliminary conclusion is occurring because most vote are decided with a spread between yes and no vote which is large. At the end, the goal is to predict if the affair is accepted or rejected, and not how many yes or no vote are casted. Second, the mean squared error remains pretty high regarding yes and no vote. It is then pivotal to incorporate additional independent variables such as Federal council's opinion, topics, legislative debates, affair summary.

The 2nd part of the data science project is necessary to answer the second question regarding quality of political representation, which is truly pivotal for democracy.

Acknowledgements

I acknowledge David Zumbach (Grünenfelder and Zumbach GmbH) who helped me with data acquisition, and more specifically with the swissparl R package.

Appendix

Figure A1. Bar plot – Affairs by department

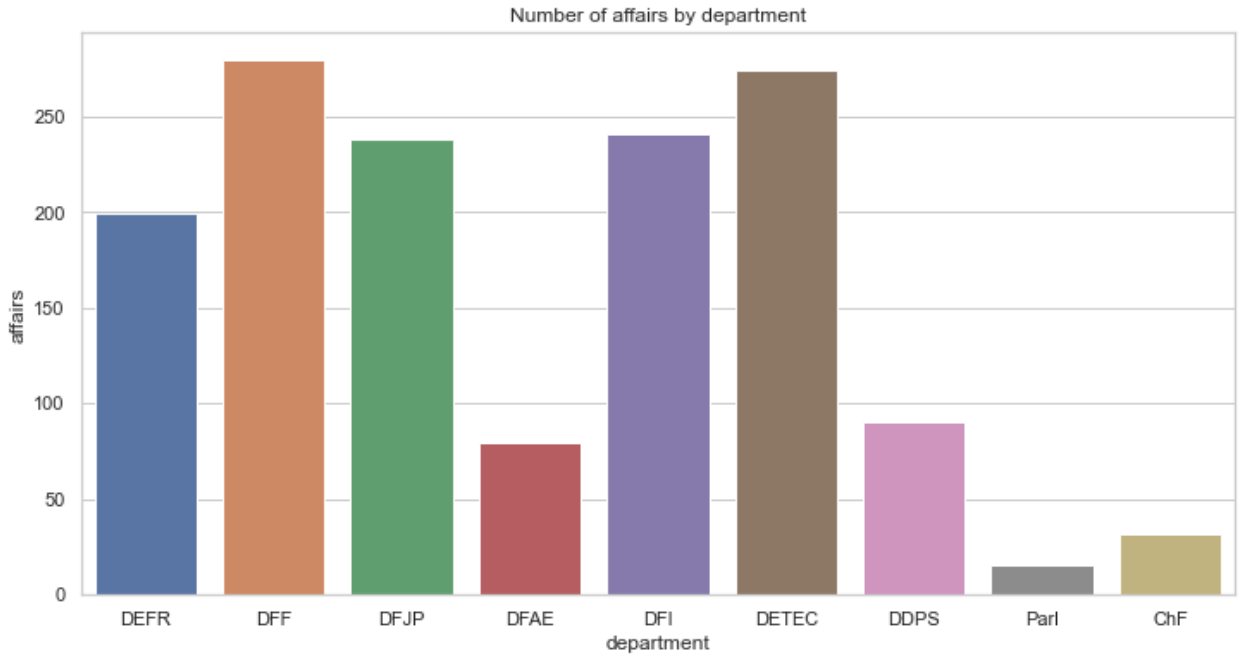


Figure A2. Bar plot – Affairs by commission

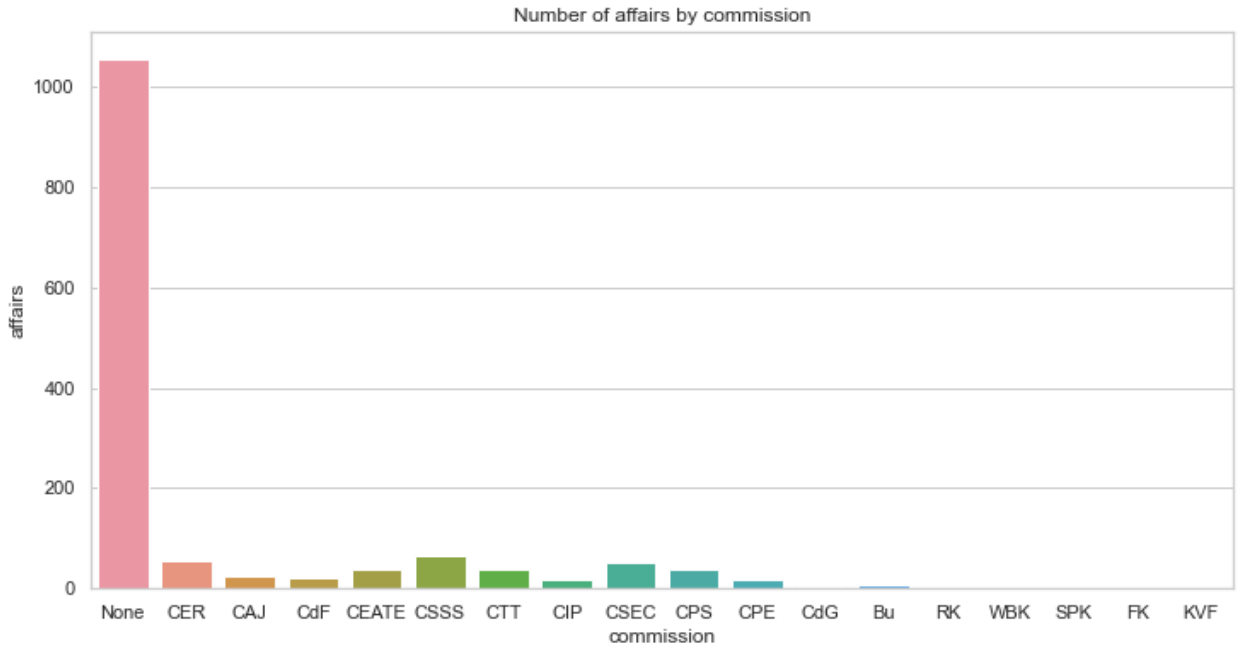


Figure A3. Bar plot – Affairs by canton

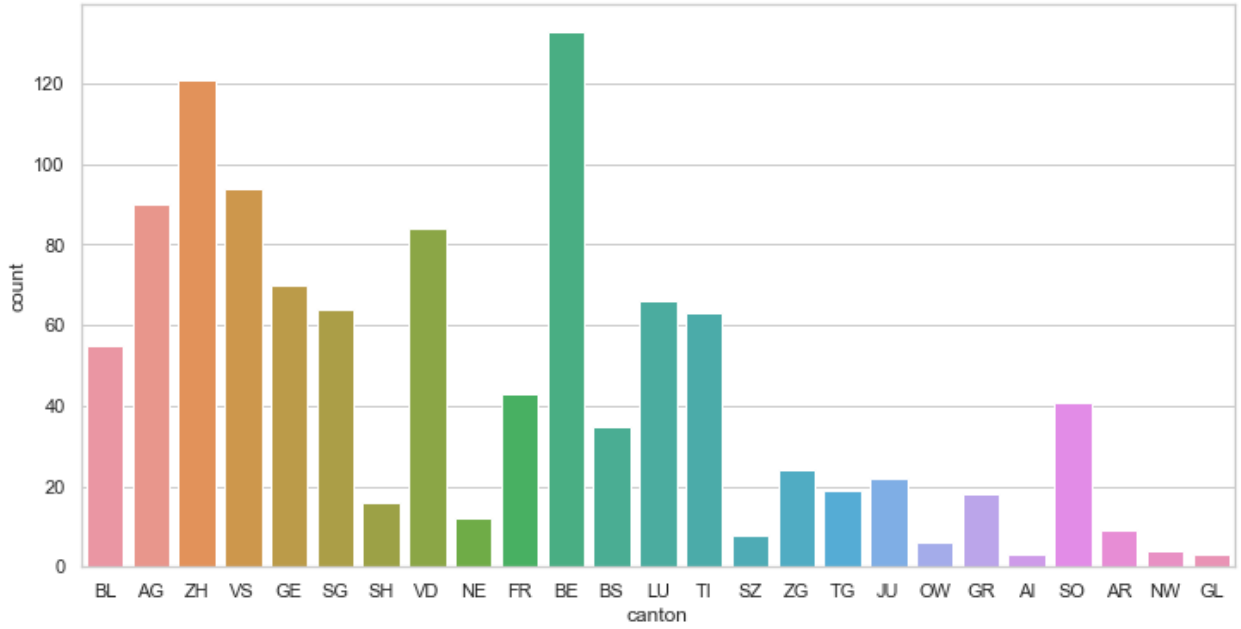
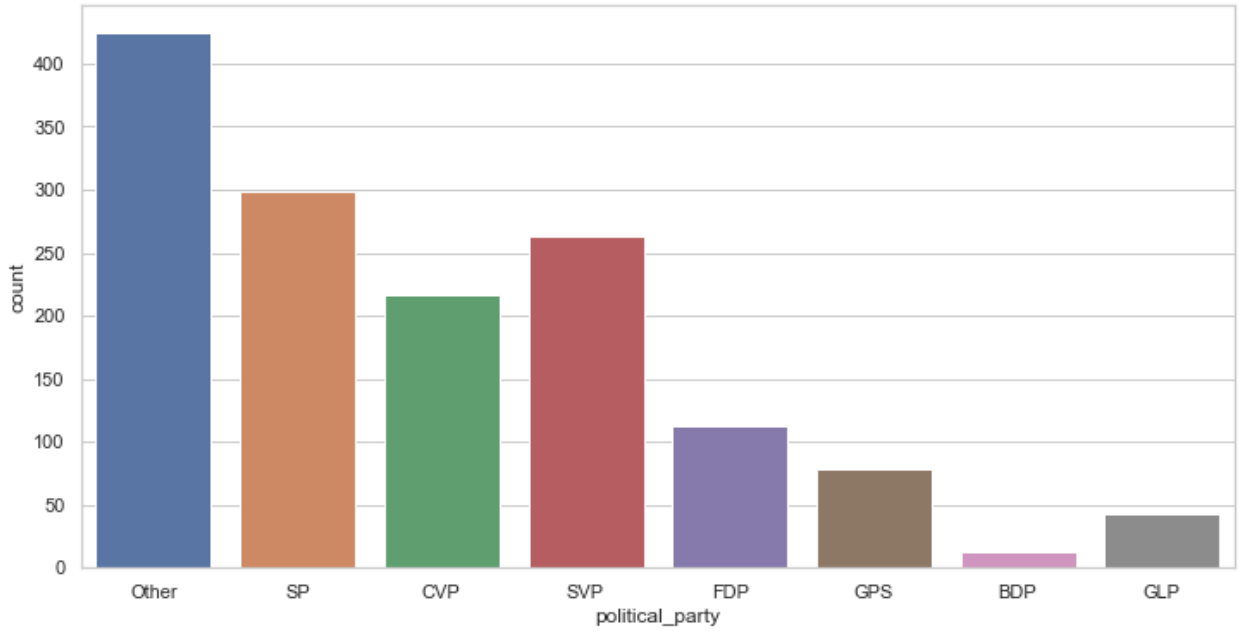


Figure A4. Bar plot – Affairs by political party



References and Bibliography

Please number any information source you used in the report with corresponding links here [1]:

[1] Zumbach, David (2019): Datensatz: Wortmeldungen im eidgenössischen Parliament (1999–2018), Grünenfelder Zumbach GmbH/Année Politique Suisse: Zürich/Bern.

[2] Parliamentary Services of the Federal Assembly (2020) Open Data, url: http://ws-old.parlament.ch, Bern, data acquired on the 05.10.2020.

[3] Zumbach, D. & Grünenfelder, B. (2020) swissparl, R package, version 0.2.1.