

Abschlussarbeit für das Modul Datenanalyse

27. Juli 2020

Inhaltsverzeichnis

1	Einleitung	2
2	Allgemeines zum Modul	2
2.1	Kurzer Überblick	2
2.2	Was ist R?	3
3	Projektidee	4
3.1	die Datensätze	4
3.2.1	Aktuelle Relevanz	4
3.2.2	Persönliche Motivation	5
3.3	Ziel geplantes Ergebnis	5
3.4	Hypothese	6
4	Praktische Durchführung	7
4.1	Daten	7
4.1.1	Datenkorrektur	7
4.2	Import	7
4.3	Dokumentation der Arbeitsschritte	7
4.3.1	Allgemeine Betrachtung der Umfrageergebnisse	7
4.3.2	Fehler der Umfrageergebnisse	8
4.3.3	Miteinbeziehen ältere Wahlergebnisse	15
5	Auswertung	19
5.1	Abweichung der letzten Umfragen und Wahlergebnisse	19
5.1.1	Wie ist die Abweichung im Bezug auf die Umfragemethoden? . . .	19
5.1.2	Wie ist die Abweichung im Bezug auf das letzte Umfragedatum? .	20
5.2	Wie ist die Abweichung im Bezug auf die Teilnehmerzahl?	21

5.3	Verbesserungsmöglichkeiten	21
5.3.1	ältere Umfragewerte	21
5.3.2	gleichzeitige Betrachtung	22
6	Kritik	22
6.1	Daten	22
6.1.1	Umfragemethoden	22
6.2	Durchführung	23
6.2.1	Unsere Auswertungsmethode na.approx	24
7	Quellen	24

1 Einleitung

Die folgende Abschlussarbeit soll die im Seminar "Datenanalyse mit R" erworbenen Kenntnisse und Erkenntnisse am Fallbeispiel politischer Wahlumfragen testen und anwenden. Hierfür werden im Folgenden die Datensätze der Umfrage- und Wahlergebnisse zu den Bundestagswahlen 2013 und 2017 bzw. dem Erhebungszeitraum zwischen den beiden Wahlen analysiert. Konkret soll dabei das zu testende Erkenntnisinteresse sein, welche Umfragemethodik sich zur Prognose von Wahlergebnissen am besten eignen könnte. Es soll untersucht werden, welche Unterschiede sich zwischen den verschiedenen Umfragen und ihren zugrundeliegenden Methoden in Bezug auf die Abweichung vom Wahlergebnis, auf die Relevanz des Erhebungsdatums und der Teilnehmerzahl feststellen lassen. Darüber hinaus wird danach gefragt, auf welche Weise sich die Vorhersageergebnisse womöglich optimieren lassen, indem entweder mehrere Umfrageergebnisse über einen längeren Zeitraum betrachtet bzw. die Umfragewerte verschiedener Meinungsforschungsinstitute zusammengefasst werden.

2 Allgemeines zum Modul

2.1 Kurzer Überblick

Statistik als wissenschaftliche Disziplin etablierte sich in Europa zwischen dem 17. und 19. Jahrhundert; ihre mathematischen Wurzeln reichen bis ins antike Griechenland bzw. Rom zurück. Ihre Relevanz entwickelte die Statistik insbesondere durch ihre zunehmende Bedeutung für staats-, bevölkerungs- und makropolitische Entscheidungen: Sie erlaubte es historisch, Übersichten über Sterberaten, Rentenpreise und Erbschaften zu entwickeln. Im Laufe des 19. Jahrhunderts wurde diese Art der statistischen Erhebung zunehmend professionalisiert und institutionalisiert, statistische Gesellschaften unter der Prämisse gegründet, "neutrales und objektives" Wissen über die sich abbildenden Verhältnisse zu erheben. Mit Entstehung des modernen, bürgerlichen Rechtsstaats und der parlamentarischen Demokratie verschob sich der Blick auf Statistik kontinuierlich: Sie wird heute zuweilen als eine Grundlage des regulativen Sozialstaatsprinzips definiert, da nur die

erhobene Information über den Zustand der Gesellschaft und des Staates eine soziale Ausgestaltung eben dieser Gesellschaft ermöglicht.

Die Teildisziplin der deskriptiven Statistik befasst sich mit der Darstellung und Zusammenfassung von Daten sowie deren Auswertung. Ziel ist es, die Eigenschaften und Merkmalsarten der verschiedenen Einheiten und Daten der beschriebenen Grundgesamtheit darzustellen, um Aussagen etwa über die Häufigkeitsverteilung oder die Abhängigkeiten der Daten untereinander treffen zu können. Zur Beschreibung (Deskription) und Analyse von Daten bedient sich die Statistik standarmäßig Stichproben einer Grundgesamtheit. Diese machen Statistik zeitlich und finanziell überhaupt erst möglich und praktikabel, da eine Totalerhebung in den meisten Fällen mit enormem zeitlich-finanziellen Aufwand verbunden ist und die Praktikabilität statistischer Erhebungen bzw. Beschreibungen stark einschränkt. Ziel von Stichproben ist es daher, die Grundgesamtheit, die durch eine Totalerhebung erhoben werden würde, bestmöglich in ihrer Zusammensetzung zu repräsentieren bzw. grobe Stichprobenfehler (Verzerrungen) zu vermeiden. Standardmäßig interessiert bei einer Datenanalyse der Zusammenhang unter den einzelnen Daten. Dieser kann beispielsweise mithilfe der Kovarianz oder des Korrelationskoeffizienten ermittelt werden. Während über die Kovarianz bereits angezeigt werden kann, ob und wie (negativ oder positiv) Variablen zueinander in Abhängigkeit stehen (hier aber nicht weiter interpretiert werden kann, in welchem Ausmaß bzw. welcher Richtung die Variablen voneinander abhängig sind), kann mithilfe der Korrelationskoeffizienten darüber auch der Zusammenhang selbst interpretiert werden, also ausgesagt werden, wie stark/schwach ein Zusammenhang zwischen Variablen ist bzw. in welche Richtung sich die Abhängigkeit richtet, z.B. ob Variable X die Variable Y beeinflusst oder vice versa. Ist solch ein Zusammenhang zwischen mindestens zwei Variablen ermittelt, ist es weiter grundlegender Teil einer Datenanalyse, diesen Zusammenhang mittels einer Regressionsanalyse weiter zu differenzieren: Die Regressionsfunktion etwa kann den Zusammenhang von mindestens zwei Variablen in Form einer Geraden exakt beschreiben und so Prognosen über die abhängige Variable ermöglichen, wenn etwa der Wert der unabhängigen Variable bekannt ist oder gesetzt wird.

2.2 Was ist R?

R ist eine Programmiersprache, die vor allem für statistische Datenanalyse und Grafikerstellung verwendet wird. Die Sprache ist Open-Source und somit ist die gesamte Sprache frei zugänglich, der gesamte Code frei einsehbar, änderbar und natürlich kostenlos nutzbar, kann aber trotzdem kommerziell eingesetzt werden. Dazu ist R plattformunabhängig, weshalb diese unter Linux, Unix, Windows und MacOS läuft. Im Gegensatz zu Programmen (wie z.B. Excel oder LibreOffice Calc), die statistische Auswertungen machen, bietet R den Vorteil, dass vom Anbinden an unterschiedliche Datenquellen, dem dadurch Erhalten immer aktuellen Daten, des Auswertens und Nutzens dieser Auswertungen (z.B. durch maschinelles Lernen), bis hin zur Visualisierung, alles möglich ist. Dies kann durch das Automatisieren von wiederkehrenden Tätigkeiten durch z.B. eine neue Datenlage die Effizienz drastisch steigern.

3 Projektidee

3.1 die Datensätze

Für dieses Projekt haben wir die Daten von der Webseite wahlrecht.de [1] verwendet. Diese zeigen Wahlumfragen von verschiedenen Meinungsforschungsinstituten.

Zur Auswahl der Datensätze der verschiedenen Umfragen haben wir folgende Kriterien aufgestellt:

- i Es müssen getrennte Werte von den Parteien CDU/CSU, SPD, die Grünen, FDP, die Linke und der AfD vorhanden sein. Die restlichen dürfen in Sonstige zusammengefasst werden.
- ii Es müssen Umfragewerte vom Zeitraum der Bundestagswahlen vom 22.09.2013 und dem 24.09.2017 vorhanden sein.
- iii Es muss einsehbar sein, wie die Daten erfasst werden. Die Erfassung muss dabei durch eines dieser Erfassungsverfahren durchgeführt worden sein: (Kriterium von wahlrecht.de übernommen)
 - (F) Face-to-Face-Befragung: persönlich-mündliche Befragung von nach Quotenvorgaben ausgewählten Personen
 - (T) Telefon-Befragung: telefonische Befragung von zufällig ausgewählten Personen
 - (O) Internetbasierte Befragung von nach Quotenvorgaben ausgewählten Mitgliedern eines Befragten-Pools
- iv Die Teilnehmeranzahl muss einsehbar sein.

Daraus ergaben sich die Umfragen folgender 5 Marktforschungsinstitute: Allensbach (F), Infratest dimap (T), Kantar (Emnid)(T), Forsa (T) und Insa (O).

Zu den Datensätzen mit den Ergebnissen der Wahlforschungsinstitute, gibt es noch einen mit den Ergebnissen der Bundestagswahlen 2013 und 2017, mit den aus den Kriterien aufgelisteten Parteien und einen Datensatz, in dem die Erfassungsmethodik des jeweiligen Marktforschungsinstituts gespeichert ist.

3.2

3.2.1 Aktuelle Relevanz

Wahlforschung und politische Umfragen sind zu einem einflussreichen und fundamentalen Teil von politischen Entscheidungsprozessen geworden: Sowohl für politische Entscheidungsträger*innen als auch für Wähler*innen sind Umfragewerte zu wichtigen Entscheidungs- und Orientierungswerten geworden. Politische Parteien und Politiker*innen selbst richten ihre Forderungen und Programme nicht zuletzt danach aus, welche Stimmung in

der Bevölkerung messbar ist, welche Themen für besonders wichtig erachtet werden und schließlich spielt es auch eine Rolle im wahltaktischen Verhalten von Wähler*innen, wie die Umfragewerte von Parteien und Politiker*innen zum Zeitpunkt einer Wahl sind. Beim Stimmensplitting etwa taktieren Wähler*innen aktiv und oftmals auf Basis von Umfragewerten: In Deutschland ist es z.B. durch die Teilung in Erst- und Zweitstimme möglich, bei der Bundestagswahl die Zweitstimme an Partei X zu geben - mit der Erststimme, die über das Wahlkreismandat für den Bundestag entscheidet, dann aber Partei Y zu wählen, weil Partei X womöglich zu klein ist, um das Direktmandat zu gewinnen und um das Mandat einer eindeutig abgelehnten Partei Z zu verhindern, geben Wähler*innen dann Partei Y lieber ihre Stimme. In diesem Fall wird also die Wahlentscheidung grundlegend von Vorwahlumfragen beeinflusst. Auch im Kontext weiterer internationaler Wahlen und Referenden wurden Umfragewerte in den vergangenen Jahren als ein einflussreicher Faktor sichtbar: So sagten die meisten Umfragen (2016) einen klaren Wahlsieg Hillary Clintons über Donald Trump voraus, sahen viele Statistiken das "RemainLager im Brexit Wahlkampf (2016) vorne und sahen zuletzt im Vorwahlkampf der amerikanischen Demokraten (2020) Joe Biden teilweise weit abgeschlagen und quasi chancenlos hinter seiner innerparteilichen Konkurrenz - Szenarien von weiterreichendem, teils weltpolitisch maßgeblichem Maßstab, die allesamt statistisch falsch oder fragwürdig erhoben und prognostiziert wurden. Politische (Wahl-)Umfragen sind für politische Systeme wie das der BRD aber auch generell gesprochen also zunehmend selbst von politischer Tragweite, können Einfluss auf politische Bündnisse, Abgrenzungen und Wahlverhalten haben und selbst Wirtschafts- und Börsenlagen beeinflussen. Die Untersuchung der Umfragemethodik selbst wird somit zu einem wichtigen wissenschaftlichen Forschungsthema.

3.2.2 Persönliche Motivation

Natürlich ist das Thema Wahlumfragen und vor allem Wahlvorhersagen viel zu komplex, um es in seiner Gänze betrachten zu können. Jedoch finden wir sowohl den gesellschaftlichen bzw. politischen Aspekt als auch den mathematischen Aspekt dahinter sehr spannend, und freuen uns, einen kleinen Ausschnitt daraus betrachten zu können. Das Thema verbindet somit auch unsere beiden Studienfächer der Mathematik und Politikwissenschaft.

3.3 Ziel geplantes Ergebnis

Mit diesem Projekt möchten wir folgende Punkte untersuchen:

- i Die Abweichungen der letzten Umfragen und Wahlergebnisse untersuchen
 - (1) Wie ist die Abweichung im Bezug auf die Umfragemethoden?
 - (2) Wie ist die Abweichung im Bezug auf das Umfragedatum der letzten Umfrage?
 - (3) Wie ist die Abweichung im Bezug auf die Teilnehmerzahl?

- ii Welche möglichen Verbesserungsmöglichkeiten gibt es, die man hätte mit den vorhandenen Daten hätte umsetzen können?
 - (1) Was passiert, wenn nicht nur die aktuelle Umfrage mit einbezogen wird? (Betrachtung unter Miteinbeziehen der letzten Wahlergebnisse und älteren Umfrageergebnisse)
 - (2) Was passiert, wenn die vorhandenen Daten der unterschiedlichen Marktforschungsinstitute gleichzeitig betrachtet werden?

Durch Untersuchen dieser Punkte erhoffen wir uns ein klareres Bild über die Auswirkungen der Methodik der Umfrageergebnisse und eine Idee, wie man die Differenz zu den eigentlichen Wahlergebnis gegebenenfalls reduzieren kann.

3.4 Hypothese

- (i 1) Aufgrund der Vor- und Nachteile der jeweiligen Umfragemethoden (s.h. Kritik Unsere Auswertungsmethode na.approx), sehen wir aktuell keinen besonderen Vorteil für eine dieser Methoden. Allerdings lassen sich diese auch in zwei Gruppen unterteilen, wo die eine Quotenvorgaben hat und die andere nicht. Auch wenn wir die Repräsentativität der Quotenvorgaben leider nicht beurteilen können, halten wir da eine höhere Genauigkeit gerade bei einer geringeren Anzahl an Umfrageteilnehmern für wahrscheinlicher.
- (i 2) Aufgrund von politischen Ereignissen die in der Zeit zwischen der letzten Umfrage und der Wahl geschehen können, sehe ich eine höhere Genauigkeit für die Umfragen, die näher an der Wahl sind.
- (i 3) Auch wenn es natürlich wichtig ist, eine repräsentative Menge zu befragen, und diese auch bestimmt einen gewissen Einfluss auf die Genauigkeit der Umfrageergebnisse nehmen wird, halten wir diese (sobald eine bestimmte Grundmenge an Personen erreicht ist) für eher weniger relevant. Was dies betrifft würden wir repräsentative Quotenvorgaben für relevanter halten.
- (ii 1) Da in den vorherigen Wahlergebnissen die ganz aktuelle politische Stimmung nicht enthalten sein mag, genauso wenig wie in den älteren Umfrageergebnissen, würden wir dadurch kein Verbesserungspotenzial sehen. Falls die Umfragen aber in einer höheren Frequenz stattgefunden haben, sehen wir unter der Voraussetzung, dass in der Zwischenzeit es kein stark veränderndes gesellschaftlich-politisches Ereignis gab die Möglichkeit, dass durch die Vergrößerung der Teilnehmeranzahl ein genaueres Ergebnis zustande kommen könnte. Bei den Wahlergebnissen sehen wir die Verbesserung auch darin, dass Daten mit einer extrem hohen Teilnehmeranzahl miteinbezogen werden.
- (ii 2) In dieser Auswertungsstrategie würden wir die größte Verbesserung sehen, da nicht nur aktuelle Daten verwendet werden und dies mit einer hohen Teilnehmeranzahl, sondern auch noch verschiedene Umfragemethoden verwendet werden, die zwar

alle ihre jeweiligen Vor- und Nachteile haben, diese aber in ihrer Kombination ausgleichen könnten.

4 Praktische Durchführung

4.1 Daten

4.1.1 Datenkorrektur

Da wir die Daten von der Webseite [1] kopiert haben, waren einige Korrekturen notwendig um sie einlesbar zu machen. Dafür haben wir diese in eine csv Datei kopiert. Da der Text-Editor Emacs einige nützliche Tools zur Korrektur bietet haben wir diesen verwendet. Die Spaltenerkennung funktioniert automatisch, sowie das Einfügen von Trennsymbolen.

Jedoch haben wir um Probleme zu vermeiden alle %-Zeichen entfernt und alle ", die als Zeichen für das nicht-vorhanden-Sein von Daten verwendet wurden, durch NA's ersetzt (Im Nachhinein ist uns aufgefallen, dass R das automatisch tut). Falls mehr als die in den Kriterien aufgelisteten Parteien in den Dateien der jeweiligen Marktforschungsinstitute aufgelistet waren, haben wir die übrigen zu Sonstige dazu addiert. Außerdem wollte wir, falls gegeben, das Startdatum und Enddatum der Umfragen betrachten können. Um die Daten getrennt einlesen zu können, haben wir alle " durch ein Trennzeichen ersetzen lassen und (damit keine Komplikationen auftreten) alle Kommas durch Punkte.

5 Umfragen von Allensbach haben nur ungefähre Teilnehmeranzahlen geboten, weshalb diese mit ca. markiert waren. Diese haben wir entfernt.

4.2 Import

```
1 allensbach <- read.csv("Allensbach.csv", sep=';')
2 dimap <- read.csv("dimap.csv", sep=';')
3 forsa <- read.csv("Forsa.csv", sep=';')
4 insa <- read.csv("insa.csv", sep=';')
5 kantar <- read.csv("Kantar.csv", sep=';')
6
7 bundw <- read.csv("Bundestagswahl.csv", sep=';')
8 inst <- read.csv("Institut.csv", sep=';')
```

4.3 Dokumentation der Arbeitsschritte

4.3.1 Allgemeine Betrachtung der Umfrageergebnisse

Als erstes wird eine Tabelle "last.Survey" mit den letzten Umfrageergebnissen erstellt:

```
1 INSTITUT <- c('Allensbach', 'Dimap', 'Forsa', 'Insa', 'Kantar')
2 CDU.CSU <- c(allensbach$CDU.CSU[1], dimap$CDU.CSU[1], forsa$CDU.CSU[1],
3             insa$CDU.CSU[1], kantar$CDU.CSU[1])
4 SPD <- c(allensbach$SPD[1], dimap$SPD[1], forsa$SPD[1], insa$SPD[1],
5          kantar$SPD[1])
```

```

4 GRUENE <- c(allensbach$GRUENE[1], dimap$GRUENE[1], forsa$GRUENE[1], insa$
  GRUENE[1], kantar$GRUENE[1])
5 FDP <- c(allensbach$FDP[1], dimap$FDP[1], forsa$FDP[1], insa$FDP[1],
  kantar$FDP[1])
6 LINKE <- c(allensbach$LINKE[1], dimap$LINKE[1], forsa$LINKE[1], insa$LINKE
  [1], kantar$LINKE[1])
7 AFD <- c(allensbach$AFD[1], dimap$AFD[1], forsa$AFD[1], insa$AFD[1],
  kantar$AFD[1])
8 OTHER <- c(allensbach$OTHER[1], dimap$OTHER[1], forsa$OTHER[1], insa$OTHER
  [1], kantar$OTHER[1])
9
10 last.Survey <- data.frame(INSTITUT, CDU.CSU, SPD, GRUENE, FDP, LINKE, AFD,
  OTHER)

```

Auf Seite 9 in der ersten Spalte sind die Umfrageergebnisse nach Parteien einmal graphisch dargestellt. Der schwarze Strich zeigt das Ergebnis der Bundestagswahlen 2017 und der graue den Mittelwert der Umfrageergebnisse. Da aber später auch die Umfrageergebnisse nach Umfragemethodik betrachtet werden sollen, wird ein Methodenmittelwert "methmean" berechnet, wo erst der Durchschnitt der verschiedenen Umfragemethoden berechnet wird und daraus dann erneut der Mittelwert genommen wird.

```

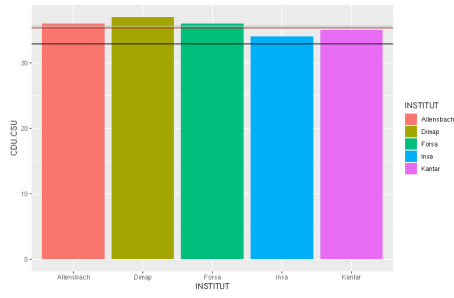
1 CDU.CSU.methmean <- mean(c(last.Survey$CDU.CSU[INSTITUT=='Allensbach'],
  last.Survey$CDU.CSU[INSTITUT=='Insa'], mean(last.Survey$CDU.CSU[
  INSTITUT=='Kantar' | INSTITUT=='Dimap' | INSTITUT=='Forsa'])))
2 SPD.methmean <- mean(c(last.Survey$SPD[INSTITUT=='Allensbach'], last.
  Survey$SPD[INSTITUT=='Insa'], mean(last.Survey$SPD[INSTITUT=='Kantar' |
  INSTITUT=='Dimap' | INSTITUT=='Forsa'])))
3 GRUENE.methmean <- mean(c(last.Survey$GRUENE[INSTITUT=='Allensbach'], last
  .Survey$GRUENE[INSTITUT=='Insa'], mean(last.Survey$GRUENE[INSTITUT=='
  Kantar' | INSTITUT=='Dimap' | INSTITUT=='Forsa'])))
4 FDP.methmean <- mean(c(last.Survey$FDP[INSTITUT=='Allensbach'], last.
  Survey$FDP[INSTITUT=='Insa'], mean(last.Survey$FDP[INSTITUT=='Kantar' |
  INSTITUT=='Dimap' | INSTITUT=='Forsa'])))
5 LINKE.methmean <- mean(c(last.Survey$LINKE[INSTITUT=='Allensbach'], last.
  Survey$LINKE[INSTITUT=='Insa'], mean(last.Survey$LINKE[INSTITUT=='
  Kantar' | INSTITUT=='Dimap' | INSTITUT=='Forsa'])))
6 AFD.methmean <- mean(c(last.Survey$AFD[INSTITUT=='Allensbach'], last.
  Survey$AFD[INSTITUT=='Insa'], mean(last.Survey$AFD[INSTITUT=='Kantar' |
  INSTITUT=='Dimap' | INSTITUT=='Forsa'])))
7 OTHER.methmean <- mean(c(last.Survey$OTHER[INSTITUT=='Allensbach'], last.
  Survey$OTHER[INSTITUT=='Insa'], mean(last.Survey$OTHER[INSTITUT=='
  Kantar' | INSTITUT=='Dimap' | INSTITUT=='Forsa'])))

```

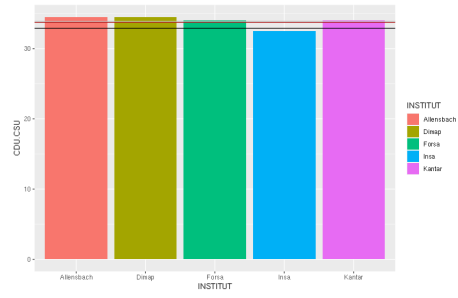
Dieser wird in den Grafiken auf Seite 9 durch die Rote Linie gezeigt.

4.3.2 Fehler der Umfrageergebnisse

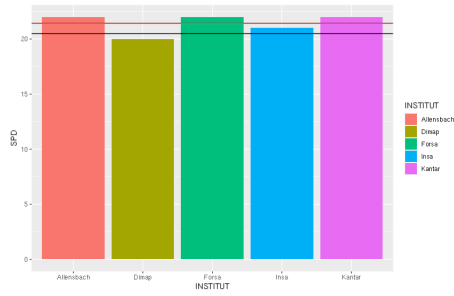
Als nächstes möchten wir den (absoluten) Fehler der letzten Umfrageergebnisse betrachten, welche wir als den Betrag vom Abstand zu den Bundestagswahlen betrachten. Diesen Fehler möchten wir einmal nach den Parteien und einmal nach den Umfrageinstituten



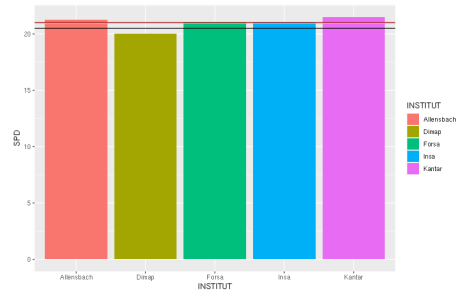
(a) CDU/CSU



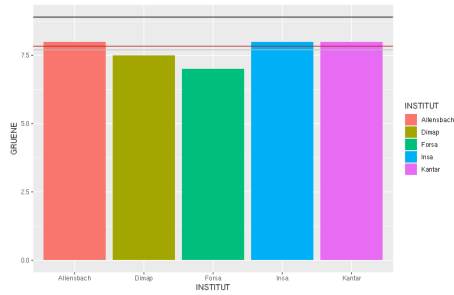
(b) CDU/CSU mit einebziehen älterer Umfragen



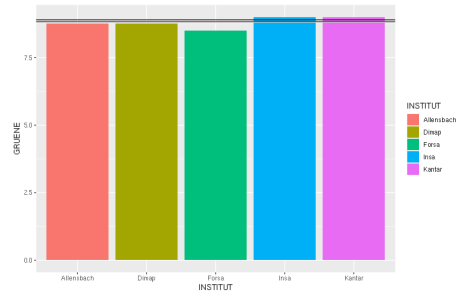
(c) SPD



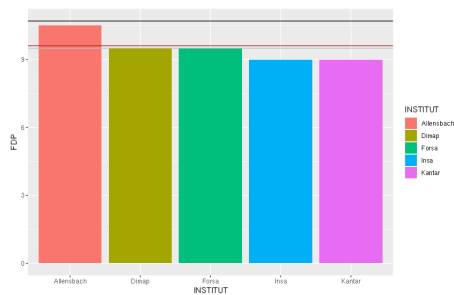
(d) SPD mit einebziehen älterer Umfragen



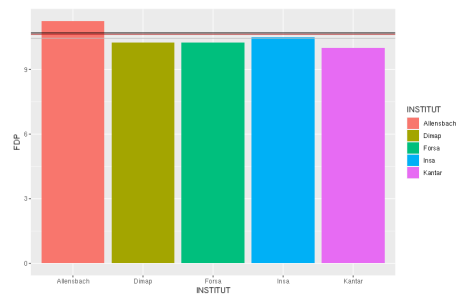
(e) GRÜNE



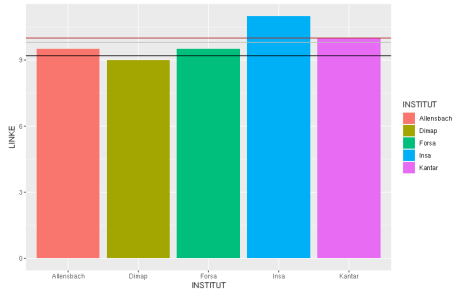
(f) GRÜNE mit einebziehen älterer Umfragen



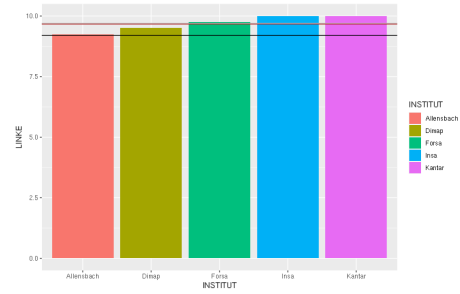
(g) FDP



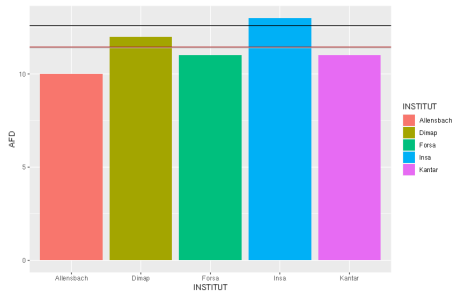
(h) FDP mit einebziehen älterer Umfragen



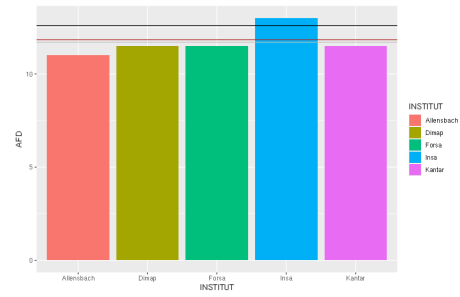
(a) LINKE



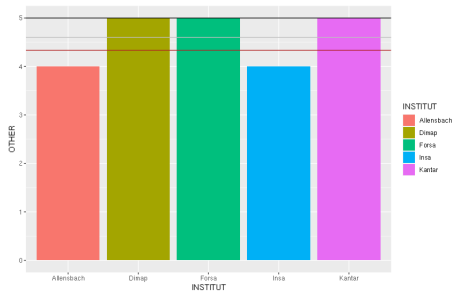
(b) LINKE mit einbezogen älterer Umfragen



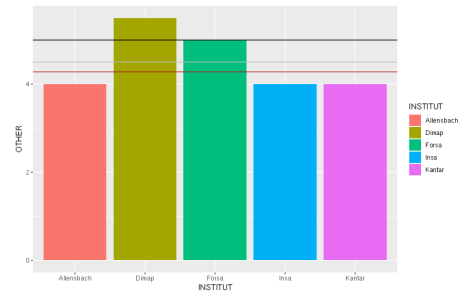
(c) AFD



(d) AFD mit einbezogen älterer Umfragen



(e) SONSTIGE



(f) SONSTIGE mit einbezogen älterer Umfragen

Abbildung 1: Umfrageergebnisse nach Parteien

betrachten.

$$absErr = |\text{Umfrageergebnis} - \text{Bundestagswahlergebnis}| \quad (1)$$

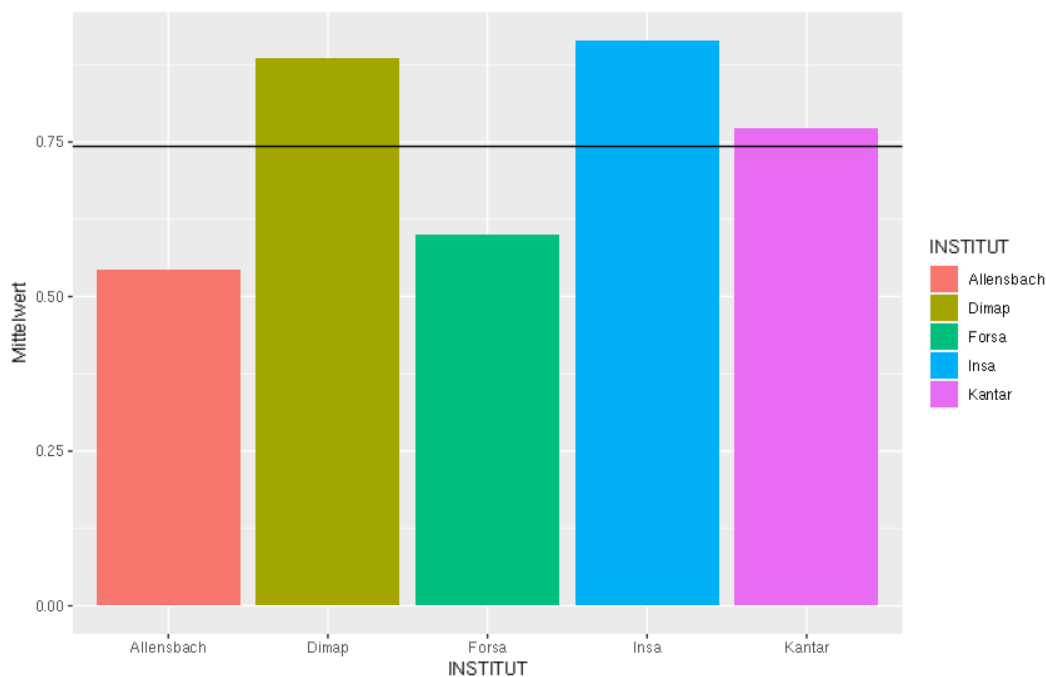
Dafür wurde ein Data Frame `last.survey.error` erstellt:

```

1 CDU.CSU.Error <- abs(CDU.CSU - bundw$CDU.CSU[1])
2 SPD.Error <- abs(SPD - bundw$SPD[1])
3 GRUENE.Error <- abs(GRUENE - bundw$GRUENE[1])
4 FDP.Error <- abs(FDP - bundw$FDP[1])
5 LINKE.Error <- abs(LINKE - bundw$LINKE[1])
6 AFD.Error <- abs(AFD - bundw$AFD[1])
7 OTHER.Error <- abs(OTHER - bundw$OTHER[1])
8
9 last.Survey.Error <- data.frame(INSTITUT, CDU.CSU.Error, SPD.Error, GRUENE
  .Error, FDP.Error, LINKE.Error, AFD.Error, OTHER.Error)

```

INSTITUT	CDU/CSU	SPD	GRUENE	FDP	LINKE	AFD	OTHER
Allensbach	3.1	1.5	0.9	0.2	0.3	2.6	1
Dimap	4.1	0.5	1.4	1.2	0.2	0.6	0
Forsa	3.1	1.5	1.9	1.2	0.3	1.6	0
Insa	1.1	0.5	0.9	1.7	1.8	0.4	1
Kantar	2.1	1.5	0.9	1.7	0.8	1.6	0



Dieser Plot zeigt den Mittelwert des absoluten Fehler nach Umfrageinstitut.

Um den Fehler der Umfrageinstitute für die einzelnen Parteien zu vergleichen, eignet sich dieser absolute Fehler, den wir da berechnen, ausreichend. Für den Vergleich der

Parteien untereinander kann dieser Fehler allerdings irreführend sein, da zum Beispiel eine Partei, die eher eine kleinere Wählerschaft hat, auch einen geringeren absoluten Fehler hat.

Deshalb ist es sinnvoll, auch den relativen Fehler zu betrachten.

$$relErr = \left| \frac{\text{Umfrageergebnis} - \text{Bundestagswahlergebnis}}{\text{Bundestagswahlergebnis}} \right| * 100 \quad (2)$$

Die Multiplikation mit 100 gehört eigentlich nicht zum relativen Fehler dazu. Wir tun dies allerdings trotzdem, um wieder Prozentualwerte rauszubekommen.

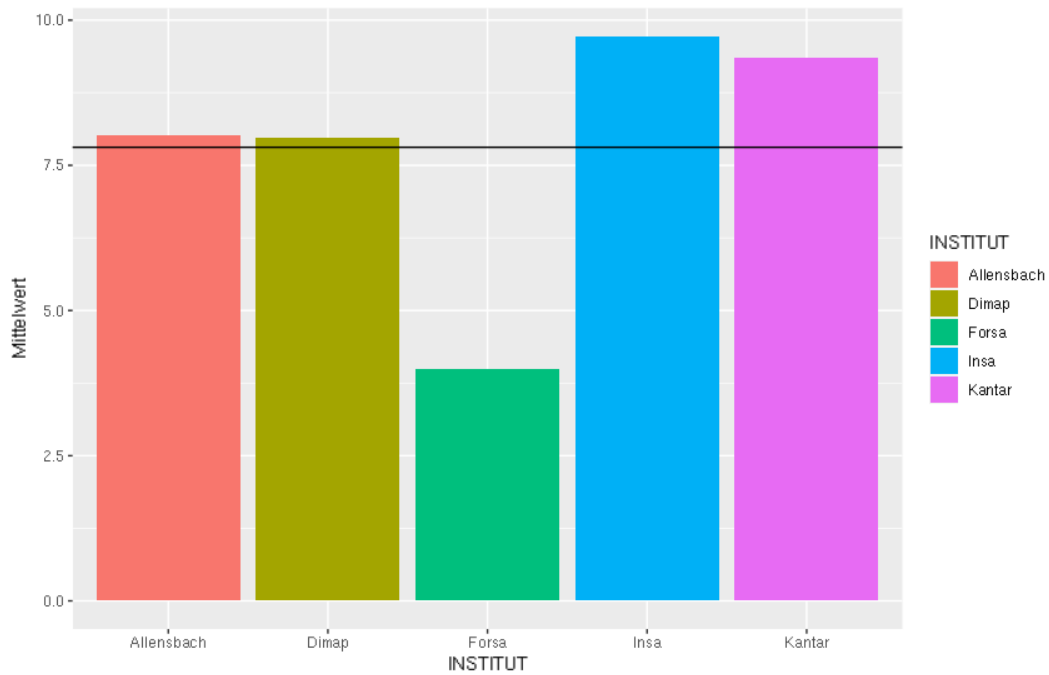
```

1 CDU.CSU.relError <- CDU.CSU.Error / bundw$CDU.CSU[1] * 100
2 SPD.relError <- SPD.Error / bundw$SPD[1] * 100
3 GRUENE.relError <- GRUENE.Error / bundw$GRUENE[1] * 100
4 FDP.relError <- FDP.Error / bundw$FDP[1] * 100
5 LINKE.relError <- FDP.Error / bundw$FDP[1] * 100
6 AFD.relError <- AFD.Error / bundw$AFD[1] * 100
7 OTHER.relError <- OTHER.Error / bundw$OTHER[1] * 100
8
9 last.Survey.relError <- data.frame(INSTITUT, CDU.CSU.relError, SPD.
  relError, GRUENE.relError, FDP.relError, LINKE.relError, AFD.relError,
  OTHER.relError)

1 last.Survey.relError <- data.frame(INSTITUT, CDU.CSU.relError, SPD.
  relError, GRUENE.relError, FDP.relError, LINKE.relError, AFD.relError,
  OTHER.relError)
2
3 last.Survey.Error.Inst.Mittel <- data.frame(INSTITUT=last.Survey.Error
  [,1], Mittelwert=rowMeans(last.Survey.Error[, -1]))
4 last.Survey.relError.Inst.Mittel <- data.frame(INSTITUT=last.Survey.
  relError[,1], Mittelwert=rowMeans(last.Survey.relError[, -1]))
5
6 ggplot(data=last.Survey.relError.Inst, aes(x=INSTITUT, y=Mittelwert, fill=
  INSTITUT)) +
7   geom_bar(stat="identity")
8
9 ggplot(data=last.Survey.Error.Inst, aes(x=INSTITUT, y=Mittelwert, fill=
  INSTITUT)) +
10  geom_bar(stat="identity")

```

INSTITUT	CDU/CSU	SPD	GRUENE	FDP	LINKE	AFD	OTHER
Allensbach	9.422492	7.317073	10.1123	1.869159	1.869159	20.634921	20
Dimap	12.462006	2.439024	15.73034	11.214953	11.214953	4.761905	0
Forsa	9.422492	7.317073	21.34831	11.214953	11.214953	12.698413	0
Insa	3.343465	2.439024	10.11236	15.887850	15.887850	3.174603	20
Kantar	6.382979	7.317073	10.11236	15.887850	15.887850	12.698413	0



Dieser Plot zeigt den Mittelwert des relativen Fehlers der Umfrageinstitute.

```

1 last.Survey.relError <- data.frame(INSTITUT, CDU.CSU.relError, SPD.
  relError, GRUENE.relError, FDP.relError, LINKE.relError, AFD.relError,
  OTHER.relError)
2
3 last.Survey.Error.Inst.Mittel <- data.frame(INSTITUT=last.Survey.Error
  [,1], Mittelwert=rowMeans(last.Survey.Error[, -1]))
4 last.Survey.relError.Inst.Mittel <- data.frame(INSTITUT=last.Survey.
  relError[,1], Mittelwert=rowMeans(last.Survey.relError[, -1]))
5
6 ggplot(data=last.Survey.relError.Inst, aes(x=INSTITUT, y=Mittelwert, fill=
  INSTITUT)) +
7   geom_bar(stat="identity")
8
9 ggplot(data=last.Survey.Error.Inst, aes(x=INSTITUT, y=Mittelwert, fill=
  INSTITUT)) +
10  geom_bar(stat="identity")

```

Da aber vor allem in der Spalte der sonstigen Parteien Extremwerte zustande kommen, muss in dem Fall der Mittelwert nicht unbedingt repräsentativ für die Aussagekraft der Umfrageinstitute oder Methoden sein, weshalb der Median auch betrachtet wird:

```

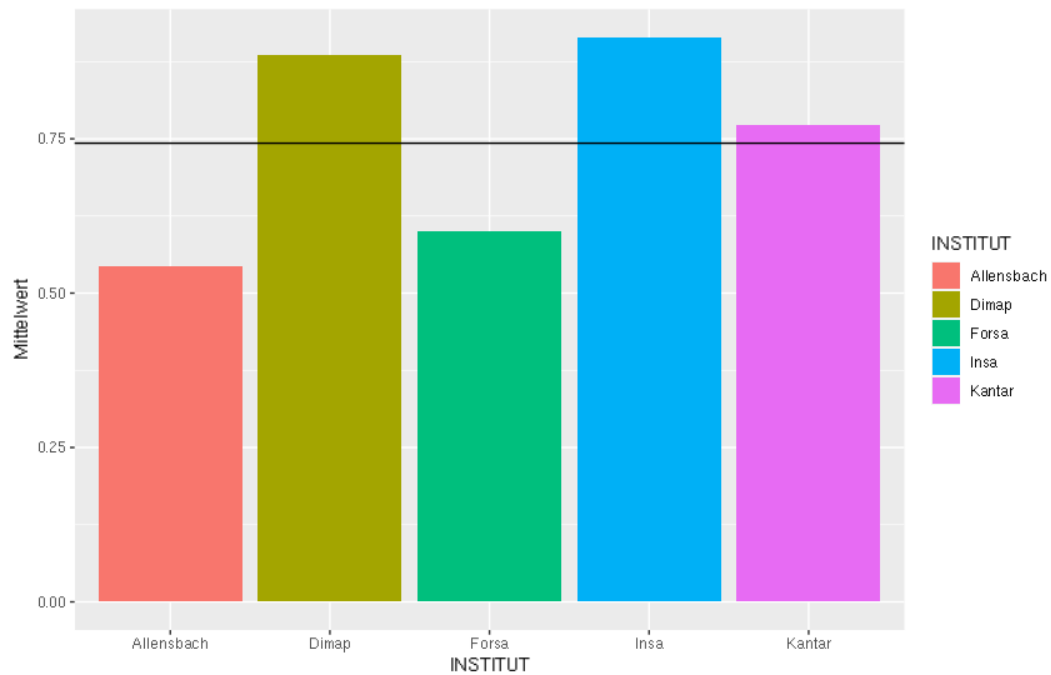
1 last.Survey.Error.Inst.Median <- data.frame(INSTITUT=last.Survey.Error
  [,1], Median=rowMedians(last.Survey.Error[, -1]))
2 last.Survey.relError.Inst.Median <- data.frame(INSTITUT=last.Survey.
  relError[,1], Median=rowMedians(last.Survey.relError[, -1]))
3

```

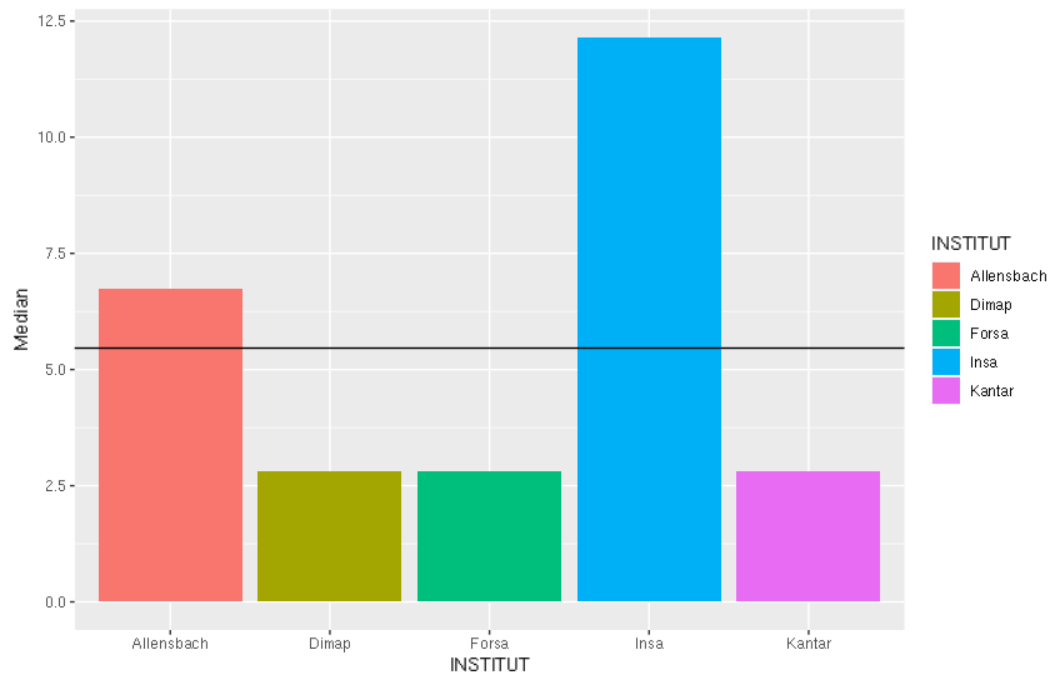
```

4 | ggplot(data=last.Survey.relError.Inst.Median, aes(x=INSTITUT, y=Median,
5 |   fill=INSTITUT)) +
6 |   geom_bar(stat="identity")
7 | ggplot(data=last.Survey.Error.Inst.Median, aes(x=INSTITUT, y=Median, fill=
8 |   INSTITUT)) +
9 |   geom_bar(stat="identity")

```



Dieser Plot zeigt den Median des absoluten Fehlers der Umfrageinstitute.

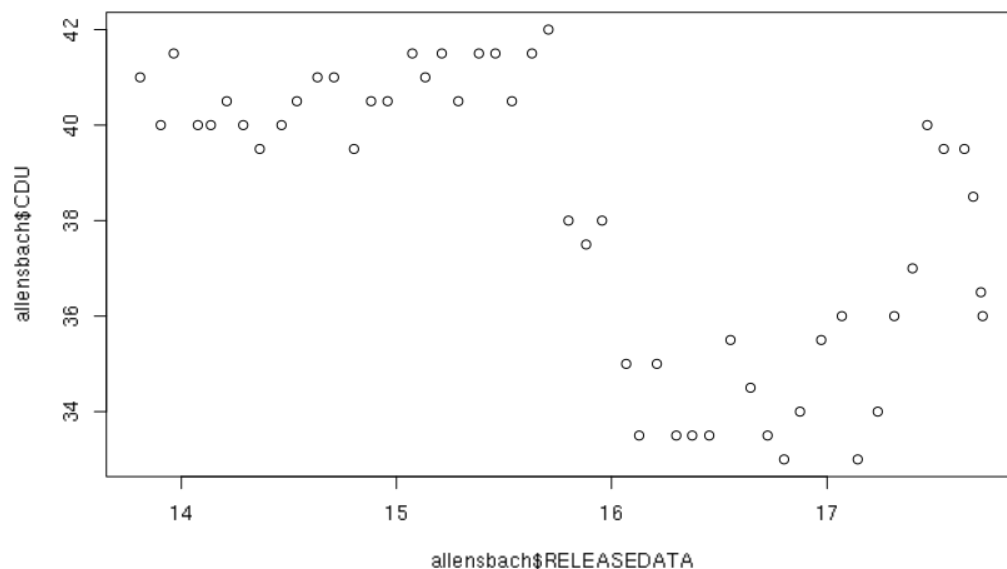


Dieser Plot zeigt den Median des relativen Fehlers der Umfrageinstitute.

4.3.3 Miteinbeziehen ältere Wahlergebnisse

In diesen Teil war unsere Idee, mithilfe einer Funktion, die durch alle vorhandenen Wahlumfragen des jeweiligen Instituts approximiert werden soll, ein mögliches Wahlergebnis zu berechnen.

Dies stellte sich als viel schwieriger als gedacht heraus, trotz der vielen vorderfinierten Approximationsfunktionen die es in den verschiedenen R libraries gibt, da das Verfahren zum einen ein Polynom approximieren sollte, da eine lineare Funktion als Approximationsfunktion mit einem Blick auf die Umfragedaten uns einfach nicht sinnvoll erschien, wie bei dieser Grafik von den CDU/CSU-Umfragewerten von Allersbach:



Jedoch bietet R auch einige Funktionen die interpolieren. Jedoch gab es bei der Funktion `interpolate` die Problematik, dass sie nicht mit einem Datum als x wert umgehen konnte und wenn das Problem durch das ersetzen eines Vektors der länge der Umfragenanzahlen umgangen ist, wurde immer noch ein Polynom eines Grades mit der Anzahl der Umfragen berechnet, was so lange dauerte, dass wir es garnicht hätten für alle Daten schaffen können.

Es gibt auch mehrere Funktionen, die unter anderem mithilfe von verschiedenen Interpolationsverfahren die NA-Werte approximieren. 2 von 3 konnten allerdings keine NAs am Anfang bzw Ende eines Datensatzes berechnen. Also blieb uns nur die Funktion `na.approx` aus der library `zoo`, wo dies funktionierte, wenn man 2 NA Zeilen oben einfügte.

Wir fügten nun am Anfang jeder csv-Datei von den Instituten das Datum der Bundestagswahl mit NA-Werten ein.

```

1 allensbach$CDU.CSU <- na.approx(allensbach$CDU.CSU)
2 allensbach$SPD <- na.approx(allensbach$SPD)
3 allensbach$GRUENE <- na.approx(allensbach$GRUENE)
4 allensbach$FDP <- na.approx(allensbach$FDP)
5 allensbach$LINKE <- na.approx(allensbach$LINKE)
6 allensbach$AFD <- na.approx(allensbach$AFD)
7 allensbach$OTHER <- na.approx(allensbach$OTHER)
8
9 allensbach <- allensbach[-c(1),] #Löschen der NA Zeile
10
11 dimap$CDU.CSU <- na.approx(dimap$CDU.CSU)
12 dimap$SPD <- na.approx(dimap$SPD)
13 dimap$GRUENE <- na.approx(dimap$GRUENE)
14 dimap$FDP <- na.approx(dimap$FDP)
15 dimap$LINKE <- na.approx(dimap$LINKE)
16 dimap$AFD <- na.approx(dimap$AFD)
17 dimap$OTHER <- na.approx(dimap$OTHER)
18
19 dimap <- dimap[-c(1),] #Löschen der NA Zeile
20
21 forsa$CDU.CSU <- na.approx(forsa$CDU.CSU)
22 forsa$SPD <- na.approx(forsa$SPD)
23 forsa$GRUENE <- na.approx(forsa$GRUENE)
24 forsa$FDP <- na.approx(forsa$FDP)
25 forsa$LINKE <- na.approx(forsa$LINKE)
26 forsa$AFD <- na.approx(forsa$AFD)
27 forsa$OTHER <- na.approx(forsa$OTHER)
28
29 forsa <- forsa[-c(1),] #Löschen der NA Zeile
30
31 insa$CDU.CSU <- na.approx(insa$CDU.CSU)
32 insa$SPD <- na.approx(insa$SPD)
33 insa$GRUENE <- na.approx(insa$GRUENE)
34 insa$FDP <- na.approx(insa$FDP)
35 insa$LINKE <- na.approx(insa$LINKE)
36 insa$AFD <- na.approx(insa$AFD)
37 insa$OTHER <- na.approx(insa$OTHER)
38
39 insa <- insa[-c(1),] #Löschen der NA Zeile
40
41 kantar$CDU.CSU <- na.approx(kantar$CDU.CSU)
42 kantar$SPD <- na.approx(kantar$SPD)
43 kantar$GRUENE <- na.approx(kantar$GRUENE)
44 kantar$FDP <- na.approx(kantar$FDP)
45 kantar$LINKE <- na.approx(kantar$LINKE)
46 kantar$AFD <- na.approx(kantar$AFD)
47 kantar$OTHER <- na.approx(kantar$OTHER)
48
49 kantar <- kantar[-c(1),] #Löschen der NA Zeile

```

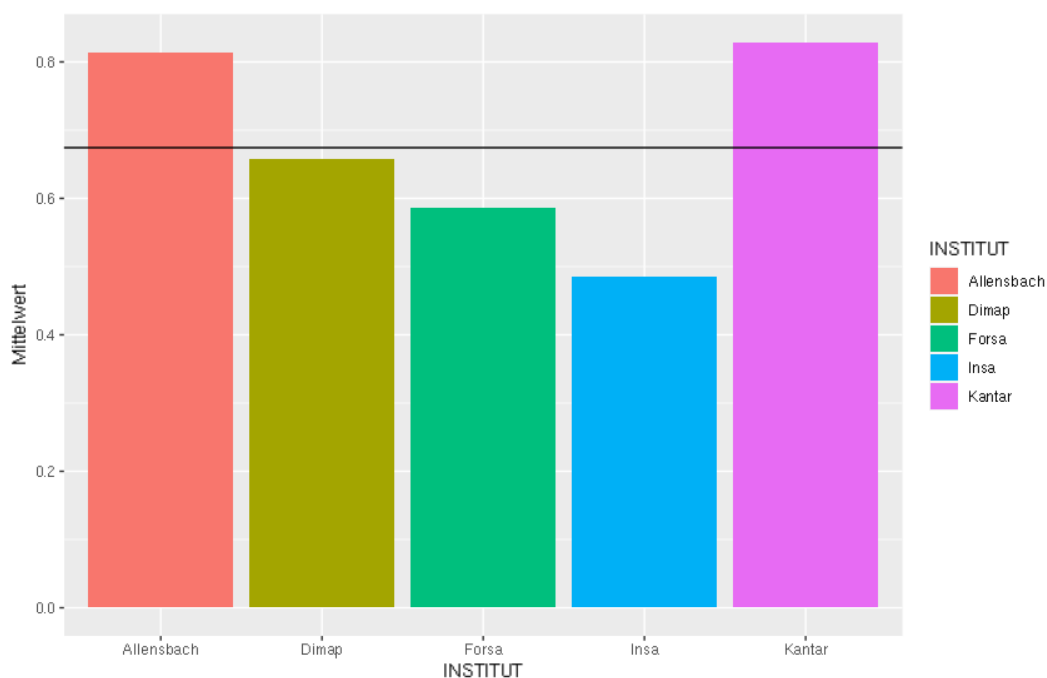
Danach führten wir nochmal die gleiche Analyse wie in 4.3.1 durch. Die Ergebnisse findet man in der rechten Spalte auf 9. Um den Fehler genauer betrachten zu können, haben wir anschließend den Fehler angeschaut genauso wie in 4.3.2. Die Tabelle des absoluten

Fehlers sieht jetzt folgendermaßen aus:

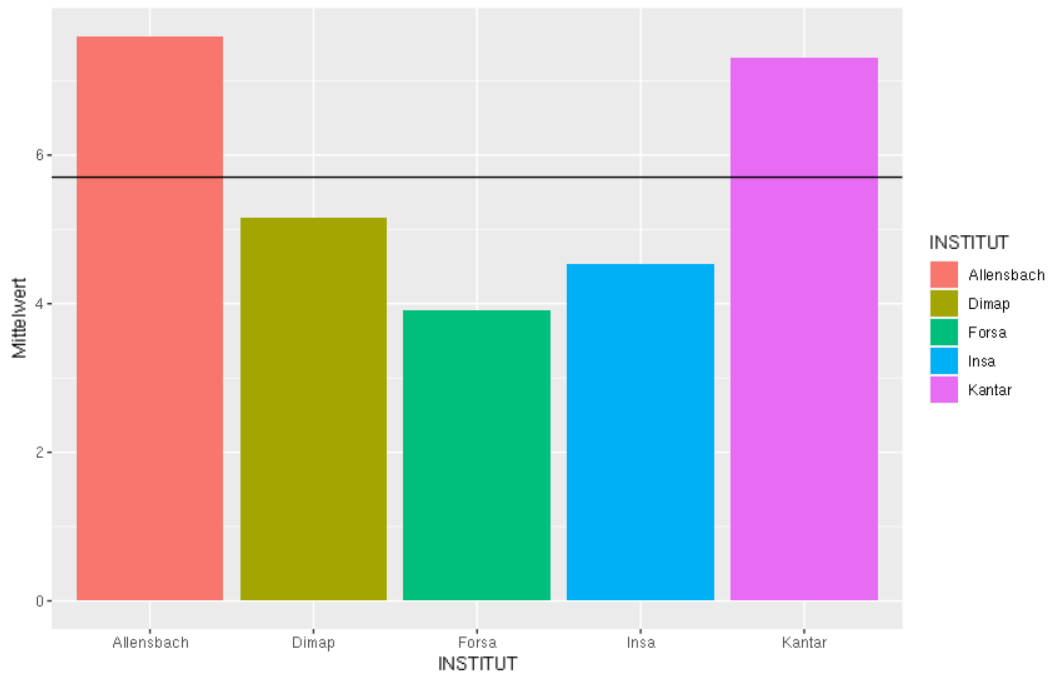
INSTITUT	CDU/CSU	SPD	GRUENE	FDP	LINKE	AFD	OTHER
Allensbach	1.6	0.75	0.15	0.55	0.05	1.6	1.0
Dimap	1.6	0.50	0.15	0.45	0.30	1.1	0.5
Forsa	1.1	0.50	0.40	0.45	0.55	1.1	0.0
Insa	0.4	0.50	0.10	0.20	0.80	0.4	1.0
Kantar	1.1	1.00	0.10	0.70	0.80	1.1	1.0

Und die vom relativen Fehler so:

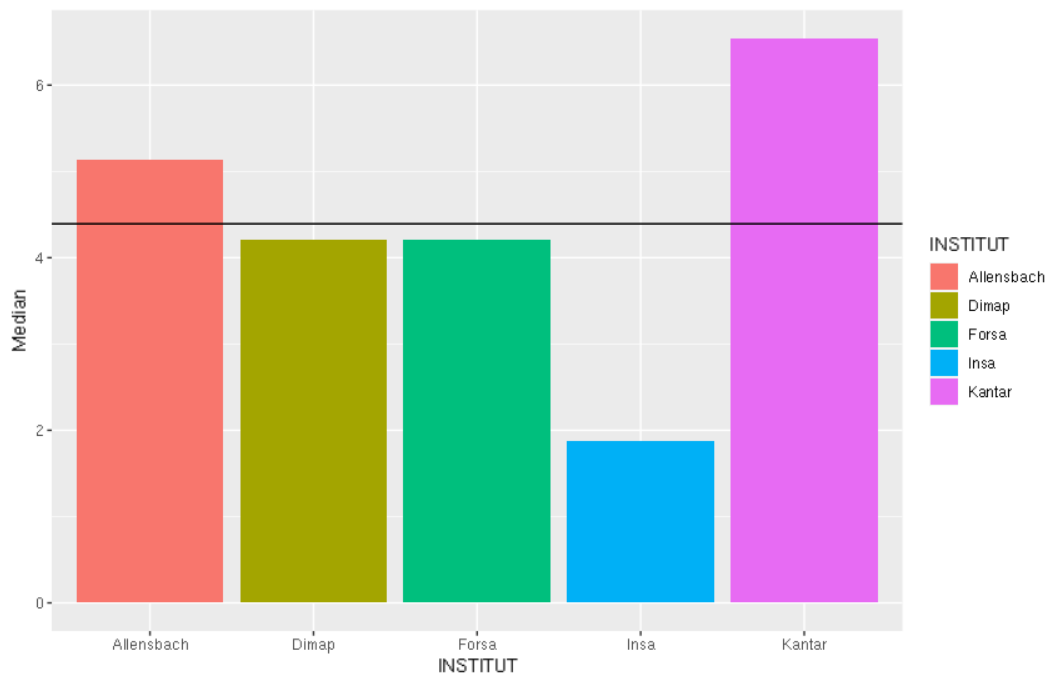
INSTITUT	CDU/CSU	SPD	GRUENE	FDP	LINKE	AFD	OTHER
Allensbach	4.863222	3.658537	1.685393	5.140187	5.140187	12.698413	20
Dimap	4.863222	2.439024	1.685393	4.205607	4.205607	8.730159	10
Forsa	3.343465	2.439024	4.494382	4.205607	4.205607	8.730159	0
Insa	1.215805	2.439024	1.123596	1.869159	1.869159	3.174603	20
Kantar	3.343465	4.878049	1.123596	6.542056	6.542056	8.730159	20



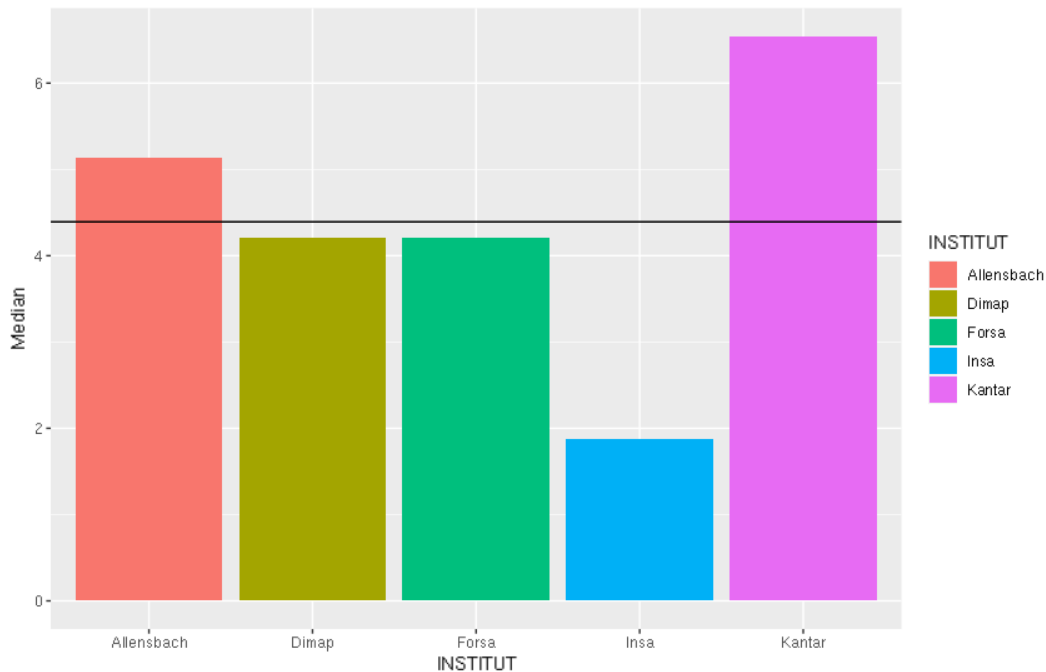
Dieser Plot zeigt den absoluten Fehler des Mittelwertes der Umfrageergebnisse nach Umfrageinstitut.



Dieser Plot zeigt den relativen Fehler des Mittelwertes der Umfrageergebnisse nach Umfrageinstitut.



Dieser Plot zeigt den absoluten Fehler des Medians der Umfrageergebnisse nach Umfrageinstitut.



Dieser Plot zeigt den relativen Fehler des Medians der Umfrageergebnisse nach Umfrageinstitut.

5 Auswertung

5.1 Abweichung der letzten Umfragen und Wahlergebnisse

Um die Umfrageinstitute übersichtlicher und im Vergleich zueinander bewerten zu können, betrachten wir die Mittelwerte und Medians der absoluten und relativen Fehler und vergeben Punkte, für jede Grafik. Das Umfrageinstitut mit dem jeweiligen kleinsten Fehler bekommt 4 Punkte, der zweitkleinste 3 usw. am Ende haben wir folgende Tabelle:

Fehler	Allensbach	Dimap	Forsa	Insa	Kantar
abs. Fehler Mittel	4	1	3	0	2
rel. Fehler Mittel	2	3	4	0	1
abs. Fehler Median	4	1	3	0	2
rel. Fehler Median	2	4	4	0	4
Summe	12	9	14	0	9

5.1.1 Wie ist die Abweichung im Bezug auf die Umfragemethoden?

Wie in 3.1 erwähnt, verwenden die Meinungsforschungsinstitute Dimap, Kantar und Forsa Telefonumfragen, Allensbach Face-to-Face Umfragen und Insa Online-Umfragen.

Auf die Umfragemethode bezogen heißt das, die Telefonumfragen haben durchschnittlich 10,6 Punkte, die Face-to-Face-Umfrage 12 und die Online-Umfrage 0. Da wir um die

Face-to-Face-Umfragemethode und Online-Umfragemethode zu wenige Umfrageinstitute haben, lässt sich für diese dadurch kein Ergebnis erschließen. Bei den Telefonumfrageinstituten kann man aber sagen, dass sie zumindest alle im durchschnittlichen bis überdurchschnittlichen Bereich liegen und somit auf jeden Fall erstmal kein besonderer Nachteil gegenüber anderen Methoden erkennbar ist.

Vor allem zeigt dies, dass zufällige Umfragemethoden wie die Telefonumfrage keinen zwangsläufigen Nachteil gegenüber Umfragen mit Quotenvorgaben zu haben scheinen, auch wenn das aufgrund der geringen Anzahl an miteinbezogenen Umfrageinstitute wieder mit Vorsicht zu genießen ist.

5.1.2 Wie ist die Abweichung im Bezug auf das letzte Umfragedatum?

Institut	letztes Veröffentlichungsdatum	Umfragebeginn	Umfrageende	Dauer
Allensbach	22.09.2017	13.09	20.09	8
Dimap	14.09.2017	12.09	13.09	2
Forsa	22.09.2017	18.09	21.09	4
Insa	22.09.2017	21.09	22.09	2
Kantar	22.09.2017	14.09	21.09	8

Da 4/5 Umfrageinstitute am selben Tag die letzte Umfrage veröffentlicht haben, lässt sich auch da kein Zusammenhang erkennen. Was allerdings ausreichend variiert, ist die Dauer der jeweiligen Umfragen. Um die Korrelation zu testen, berechnen wir den Korrelationskoeffizient nach Pearson.

```

1 > Punktzahl <- c(12,9,14,0,9)
2 > Dauer <- c(8,2,4,2,1)
3 > cor.test(Punktzahl,Dauer,method="pearson")
4
5 Pearson's product-moment correlation
6
7 data: Punktzahl and Dauer
8 t = 0.97686, df = 3, p-value = 0.4007
9 alternative hypothesis: true correlation is not equal to 0
10 95 percent confidence interval:
11 -0.6901292 0.9582133
12 sample estimates:
13 cor
14 0.4912461

```

Und auch wenn die Korrelation nicht allzu hoch ist, scheint ein gewisser Zusammenhang zwischen einer höheren Dauer und eines unterdurchschnittlich hohen Fehlers einer Umfrage da zu sein.

5.2 Wie ist die Abweichung im Bezug auf die Teilnehmerzahl?

Institut	Teilnehmerzahl	Dauer
Allensbach	1074	8
Dimap	1503	2
Forsa	2006	4
Insa	2000	2
Kantar	1994	8

Um dies zu betrachten, verwenden wir auch den Korrelationskoeffizienten nach Pearson:

```
1 > Teilnehmer <- c(1074, 1503, 2006, 2000, 1994)
2 > cor.test(Punktzahl, Teilnehmer, method="pearson")
3
4 Pearson's product-moment correlation
5
6 data: Punktzahl and Teilnehmer
7 t = -0.62287, df = 3, p-value = 0.5775
8 alternative hypothesis: true correlation is not equal to 0
9 95 percent confidence interval:
10 -0.9400160  0.7753566
11 sample estimates:
12 cor
13 -0.3383983
```

Entgegen unserer Hypothese, die besagte, dass eine höhere Teilnehmeranzahl, mit einem geringeren Fehler einhergehen könnte, scheint da ein wenn auch negativer Zusammenhang da zu sein. Da spricht dann doch eher dafür, dass andere Faktoren wie z.B. Quotenvorgaben entscheidender sein könnten.

5.3 Verbesserungsmöglichkeiten

5.3.1 ältere Umfragewerte

Um zu betrachten, ob unsere Methode die Umfragewerte besser zu nutzen funktioniert hat, vergleichen wir zuerst die Umfragewerte jeder Partei nach Institut mit den neuen Berechnungen nach Institut (Seite 7). Den Vergleich ziehen wir dabei auf die Ergebnisse der Bundestagswahl 2017, welche durch die schwarze Linie zu erkennen ist.

CDU/CSU: Im Gegensatz zu vorher, wo alle Umfrageergebnisse über die der Bundestagswahl lagen, sind jetzt alle Institute näher an dem Ergebnis dran. Für die Voraussage der Wahlergebnisse der CDU würde ich unsere Methode also als gelungen betrachten.

SPD: Wenn auch nicht so stark wie bei der CDU, ist bei der SPD auch eine Verbesserung zu sehen

GRÜNE: Bei den Grünen ist eine klare Verbesserung bei jedem Umfrageinstitut zu sehen.

FDP: Eine Verbesserung ist auch zu erkennen, auch wenn gerade bei Allensbach und Kantar immer noch klare Abweichungen zu sehen sind.

LINKE: Vor allem für Forsa ist eine sehr genaue Vorraussage zu erkennen, aber auch Dimap, die vorher recht stark daneben lagen, hatten eine ordentliche Verbesserung.

AFD: Bei der AFD ist nur eine geringfügige Verbesserung zu sehen.

SONSTIGE: In dem Fall ist sogar eine Verschlechterung für einige Umfrageinstitute zu sehen und für die anderen hat sich auch nichts verbessert.

Das Miteinbeziehen der Umfragewerte hat bei einigen Parteien bis auf der AFD und den sonstigen Parteien eine Verbesserung erwirkt.

Wenn man sich dann aber die vorherigen Umfragen anguckt, sind auch schon klare Tendenzen zu erkennen. Beispielsweise sind Umfragewerte der CDU von Allensbach in den letzten Umfragen am sinken, da ist es simpel vorrauszusagen, dass diese weiter sinken werden. Bei der AFD dagegen gehen die Umfragewerte immer wieder rauf und runter, allerdings in einem bestimmten Bereich. Es bleibt dann nur schwieriger, ein Tendenz zu erkennen.

5.3.2 gleichzeitige Betrachtung

Um die gleichzeitige Betrachtung bzw. in unserem Fall den Mittelwert der Umfrageinstitute zu betrachten ist bei den Umfrageergebnissen nach Partei (Seite 7) die graue Linie da.

Da ist natürlich klar, dass dies die Genauigkeit im Verhältnis zu den Umfrageinstituten senken und steigern wird und je nach dem, welche Umfragewerte wann betrachtet werden, dies eine Verbesserung oder Verschlechterung darstellt.

Deshalb schauen wir uns vor allem an, wie genau der Mittelwert unserer Methode an den der Bundestagswahl dran sind. Dies ist für uns vorallem praktisch, da es aktuell kein Umfrageinstitut gibt, welches am das beste Ergebnis mit unserer Methode hat.

Und bis auf bei den sonstigen Parteien und der AFD, scheint dieser Mittelwert doch ziemlich nah an den Wahlergebnissen dran zu sein und würde dann schon ein signifikant genaueres Ergebnis liefern.

6 Kritik

6.1 Daten

Generell lässt sich fragen, inwiefern die Ergebnisse der Datenanalyse verallgemeinerbar sind, da bei der Anzahl von fünf Umfrageinstituten zwei Umfragemethoden von jeweils nur einem Meinungsforschungsinstitut repräsentiert wurden.

6.1.1 Umfragemethoden

Zu den Fehlern, die durch die Datenlage entstehen können, kommen außerdem die jeweiligen Vor- und Nachteile der unterschiedlichen Umfragemethoden hinzu, die wiederum Fehler produzieren können. Bei Face-to-Face-Interviews (Allensbach-Daten) bringen

grundsätzlich eine recht hohe Rücklaufquote mit, ermöglichen unmittelbare Hilfestellungen der Interviewer bei Verständnisproblemen und zugleich explizite Nachfragen durch den Interviewer, um Aspekte klarzustellen und es ist möglich, recht viele weitere Informationen über die Befragten und somit die Stichprobe zu erfragen. Dagegen sind Face-to-Face-Umfragen sehr kosten- und durchführungsaufwendig und vor allem bringen sie die Gefahr des Intervieweffektes bzw. der sozialen Erwünschtheit mit sich und können so zu immanenten Verzerrungen führen. Telefon-Umfragen (Dimap, Kantar, Forsa) sind hingegen kostengünstiger und zeitlich effizienter in der Erhebung, schützen eher vor Intervieweffekten, da sie anonym sind. Hingegen ist die Interviewsituation am Telefon erfahrungsgemäß wesentlich sensibler und schwieriger aufzubauen, muss ohne Visualisierung auskommen und erzeugt standardmäßig eine wesentlich höhere Ablehnungsquote und lässt somit die Frage nach der Repräsentativität entstehen (da auch die Festnetzanschlüsse zwischen z.B. Alterskohorten sehr unterschiedlich verbreitet sein können). Gleiches Problem bergen Online-Umfragen (Insa): Sie sind zwar sehr schnell und kostengünstig durchführbar und erlauben eine schnelle Datengenerierung - jedoch stellt sich auch hier die Frage des Zugangs zur jeweiligen Umfrage und es gibt vergleichsweise hohe Risiken, dass Umfragen fehlerhaft oder mehrfach ausgefüllt werden. Dazu kommt dass diese meistens auf einer bestimmten Plattform, im Fall von insa sogar auf einer Social-Media-Plattform durchgeführt werden, die häufig eine Zielgruppe anspricht, wo es auch tendenziell eine andere Verteilung der politischen Orientierungen gibt, als es für Deutschland repräsentativ wäre.

6.2 Durchführung

Bei der Berechnung der Wahlprognosen über den Zeitraum von vier Jahren hinweg können zwar viele Umfragedaten einbezogen werden und so über den abgesteckten Zeitraum eine recht genaue Vorhersage über Zeit treffen. Jedoch ist hier die Generalisierbarkeit dieser Methode infragezustellen: Die Umfragewerte verschiedener Parteien spiegeln oftmals in Echtzeit die Stimmung in der Wählerschaft zu einem bestimmten Zeitpunkt wider, die wiederum extrem von temporären Ereignissen bestimmt werden können – sich aber nicht über vier Jahre, also eine Legislaturperiode halten und entsprechend nicht verallgemeinerbar sind. Die Statistik mag in diesen Fällen zwar womöglich die Wählerstimmung recht genau wiedergeben, jedoch hat man es in diesem Fall nicht statistischen, sondern politischen bzw. zeithistorischen Ausreißern zu tun. Da Wählerumfragen die Stimmung nur abbilden, jedoch derartige extreme politische Konjunkturphänomene nicht kontextualisieren können, stellt sich bei der Methode der Langzeitprognose also ein Verallgemeinerungsproblem. Wählerumfragen, kristallisiert in Umfragewerten, ist in der Zahlenform der Prozentangaben nicht mehr unmittelbar zu entnehmen, weshalb etwa die Grünen zu Hochzeiten der Fridays for Future-Debatte mit der CDU/CSU auf Augenhöhe waren und weshalb 2020 im Zuge der Coronakrise die Schere zwischen den beiden Parteien wieder beträchtlich auseinander ging. Ob Klimadebatte, Coronakrise, Migrationskrise, Finanzkrise, Black Lives Matter-Bewegung: Statistik, auch in der Langzeitbetrachtung, kann letztlich in dieser Form lediglich abbilden – gesellschaftliche Umschwünge und Umbrüche wie die oben genannten aber kaum vorhersagen. Diese Feststellung gilt jedoch

nicht nur für den Versuch, über einen längeren Zeitraum vorherzusagen. Gesellschaftliche und politische Extremereignisse, die die politische Stimmung einer Wählerschaft massiv beeinflussen, lassen sich kaum authentisch vorhersagen: Versuche der Wahlprognosen, gerade mit zeitlichem Abstand zur eigentlichen Wahl, birgen neben dem intra-statistischen Fehler also immer das Risiko, "dem Lauf der Dinge außerhalb der Statistik ausgesetzt zu sein. Um diesem Phänomen noch besser zu entgehen, wäre an dieser Stelle zum Beispiel ein internationaler Vergleich interessant, in dem so wie in der Analyse oben, mit den Daten aus verschiedenen Umfragemethoden gearbeitet wird. Da Wahlumfragen doch oftmals von nationalen Gegebenheiten maßgeblich und primär beeinflusst werden, ließe sich so noch eher die reelle gesellschaftliche und politische Kontextsituation der einzelnen Umfragen herauskürzen - zumindest in "normalen Zeiten, in denen nationale Themen die Politik bestimmen.

6.2.1 Unsere Auswertungsmethode `na.approx`

Die Funktion `na.approx` lieferte uns zum Teil zwar genauere Ergebnisse, hat aber zwei Nachteile, die durch eine eigene Approximationsfunktion ausgebessert werden könnten:

- (1) Wir wissen nicht genau, wie die Funktion die NA Werte approximiert. Wenn man selber eine Funktion schreibt, könnte man verschiedene Verfahren ausprobieren und vielleicht auch kombinieren und gucken, welche am besten dafür funktioniert.
- (2) Es können keine Daten als x-Achse angegeben werden. Da die meisten Umfragen in unregelmäßigen Zeiträumen stattfinden (also vor den Wahlen häufiger als sonst), wäre es sinnvoll ein Approximationsverfahren zu verwenden, welches nicht allen gegebenen y-Werten automatisch äquidistante Stützstellen zuordnet.

Dazu kommt, dass man, wenn man sich die letzten Umfrageergebnisse vor der Wahl anguckt, erkennt, dass sich die Ergebnisse nur verbessern, wenn sich in den letzten Umfrageergebnissen eine monotone Tendenz befindet.

Wenn diese vor einer Wahl nicht so eindeutig sein sollte, wie bei dieser, dann würde unsere Methode wahrscheinlich gar nicht funktionieren.

Außerdem haben wir den Fehler gemacht, dass wir zwar das Wahlergebnis versucht haben zu approximieren. Dabei aber letztendlich keine 100% rauskommen, da der sonstige Teil auch approximiert wird, und nicht aus den restlichen Prozent berechnet wird. Dabei wäre es interessant, ob dies gerade dabei eine Verbesserung bewirken würde, da dabei unsere Methode am wenigsten funktioniert hat.

7 Quellen

[1] <https://www.wahlrecht.de/umfragen/>