Die Nutzung von Daten im politischen Wahlkampf

Übertragbarkeit von Methoden aus den USA auf Deutschland und Modellierung des Wahlverhaltens mit einem Random Forest

Masterarbeit
Ulrike Janke

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Technische Universität Darmstadt

Fachbereich Rechts- und Wirtschaftswissenschaften

Fachgebiet Finanzwissenschaft und Wirtschaftspolitik

Prof. Dr. Michael Neugart

Betreuer: Prof. Dr. Michael Neugart

Fachbereich Informatik

Fachgebiet Knowledge Engineering Group

Prof. Dr. Johannes Fürnkranz

Betreuer: Prof. Dr. Johannes Fürnkranz

Masterarbeit zum Thema:

Die Nutzung von Daten im politischen Wahlkampf: Übertragbarkeit von Methoden aus den USA auf Deutschland und Modellierung des Wahlverhaltens mit einem Random Forest

Bearbeitet von: Ulrike Janke

Matrikelnummer: 2282198

Studiengang: Master of Science Wirtschaftsinformatik

Eingereicht am: 25.11.2016

Förmliche Erklärung

Hiermit erkläre ich, Ulrike Janke, geboren am 24.01.1991, an Eides statt, dass ich die vorliegende Masterarbeit ohne fremde Hilfe und nur unter Verwendung der zulässigen Mittel sowie der angegebenen Literatur angefertigt habe.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Darmstadt, den 25.11.2016

_____

Unterschrift

Inhaltsverzeichnis

Förmliche Erklärung 3

Inhaltsverzeichnis 4

Abkürzungsverzeichnis 6

Abbildungsverzeichnis 7

Tabellenverzeichnis 8

1. Einleitung 9

2. Theoretische Grundlagen 10

2.1. Parteien und Wahlen 10

2.2. Technische Grundlagen 14

2.2.1. Big Data 14

2.2.2. Data Mining 14

2.2.3. Maschinelles Lernen 16

3. Bedeutung von Daten im US-amerikanischen Präsidentschaftswahlkampf 19

3.1. Die US-amerikanische Präsidentschaftswahl 19

3.2. Die Entstehung des computergestützten Wahlkampfes 20

3.3. Wählerregister 22

3.4. Prädiktive Scores und Microtargeting 24

3.5. Web und soziale Medien 25

3.6. Experimente 26

3.7. Entstehung von spezialisierten Unternehmen 28

4. Übertragbarkeit auf den deutschen Bundestagswahlkampf 30

4.1. Grundlegende Rahmenbedingungen zu Wahlen in Deutschland 30

4.2. Wählerverzeichnisse 31

4.3. Datenschutz 33

4.4. Parteifinanzierung 35

4.5. Verfügbarkeit von Datenquellen 40
4.5.1. Parteimitglieder 40
4.5.2. Parteispenden 42
4.5.3. Angebote des Bundeswahlleiters 43
4.5.4. Statistikämter 43
4.5.5. Soziale Medien 44
4.5.6. Privatwirtschaftliche Unternehmen 45
4.5.7. Forschungsinstitute 46
4.5.8. Zusammenfassung der Datenquellen 48

4.6. Kommunikation mit dem Wähler 49
4.6.1. Offline-Kanäle 49
4.6.2. Online-Kanäle 50

4.7. Zusammenfassung des Vergleichs 53

5. Praktische Umsetzung eines Vorhersagemodells 54

5.1. Fachliche Ziele des Data-Mining-Projekts 54

5.2. Erstellung der Datenbasis 55
5.2.1. Zensusdaten 55
5.2.2. Wahlergebnisse 57
5.2.3. Datenvorverarbeitung 58

5.3. Modellierung des Klassifikationsproblems und Modellerstellung 60

5.4. Evaluierung der Modelle 67

6. Diskussion der Ergebnisse 79

7. Zusammenfassung und Ausblick 82

Literaturverzeichnis 83

Anhang A: Attribute im Datensatz 90

Anhang B: Rangfolgen des Wahlergebnisses 92

Abkürzungsverzeichnis

AfD	Alternative für Deutschland
ARFF	Attribute-relation file format
BDSG	Bundesdatenschutzgesetz
BMG	Bundesmeldegesetz
BWahlG	Bundeswahlgesetz
BWO	Bundeswahlordnung
CDU	Christlich Demokratische Union Deutschlands
CSU	Christlich-Soziale Union in Bayern e.V.
CRISP-DM	Cross Industry Standard Process for Data Mining
EStG	Einkommensteuergesetz
FDP	Freie Demokratische Partei
GG	Grundgesetz
Grüne	Bündnis 90/Die Grünen
PAC	Political Action Committee
PartG	Parteiengesetz
SPD	Sozialdemokratische Partei Deutschlands
WStatG	Wahlstatistikgesetz

Abbildungsverzeichnis

Abbildung 1 Verteilung der Stimmen bei der Bundestagswahl 2013 11

Abbildung 2 Phasen des Prozessmodells CRISP-DM 15

Abbildung 3 Abgeleitete Regel für die Wahlbeteiligung 17

Abbildung 4 Berechneter Entscheidungsbaum für die Wahlbeteiligung 18

Abbildung 5 Formular zur Registrierung für eine Wahl (United States Government 2006) 22

Abbildung 6 Soziale Nachricht auf Facebook zur Kongresswahl 2010 28

Abbildung 7 Wahlbeteiligung bei den Bundestags- und Präsidentschaftswahlen (Statista 2016) 30

Abbildung 8 Entwicklung der Parteimitgliederzahl seit dem Jahr 1990, erstellt auf Basis von (Niedermayer 2015) 41

Abbildung 9 Abstimmungsverhalten der Bundestagsabgeordneten am Beispiel von Brigitte Zypries (abgeordnetenwatch.de 2016) 51

Abbildung 10 Interaktion mit Politikern am Beispiel von Brigitte Zypries (abgeordnetenwatch.de 2016) 52

Abbildung 11 Transformation der Testdaten 61

Abbildung 12 Algorithmus zur Erstellung eines Random Forest nach (Hastie, Tibshirani und Friedman 2008, S. 588) 62

Abbildung 13 Verkürzte Arff-Datei der Trainingsmenge 65

Abbildung 14 Verkürzte Arff-Datei der Testmenge 65

Abbildung 15 Bewertung einer Klassifikation mit Konfusionsmatrizen 68

Abbildung 16 Konfusionsmatrix der Trainingsdaten bei Modell_10025 68

Abbildung 17 Rangfolge der Parteiwahl inklusive der Gemeindeanzahl 92

Tabellenverzeichnis

Tabelle 1 Beispieldatensatz zur Wahlteilnahme 16
Tabelle 2 Vermögen der deutschen Parteien in den Jahren 2010 bis 2014 35
Tabelle 3 Einnahmen und Ausgaben der Parteien in Tausend Euro im Jahr 2014 36
Tabelle 4 Mitgliedsbeiträge bei der Links-Partei 37
Tabelle 5 Datenerfassung bei Parteieintritt 40
Tabelle 6 Zusammenfassung der für deutsche Parteien verfügbare Datenquellen 48
Tabelle 7 Datenquellen für die Wahlergebnisse der Bundesländer 57
Tabelle 8 Aufbau des Datensatzes 59
Tabelle 9 Erstellte Modelle und ihre Parameter 66
Tabelle 10 Tatsächliches und vorhergesagtes Wahlergebnis für Berlin 71
Tabelle 11 Tatsächliches und vorhergesagtes Wahlergebnis für Birtlingen 71
Tabelle 12 Tatsächliches und vorhergesagtes Wahlergebnis für Nusbaum 72
Tabelle 13 Tatsächliches und vorhergesagtes Wahlergebnis für Saarbrücken 72
Tabelle 14 Tatsächliches und vorhergesagtes Wahlergebnis für Zweifelscheid 72
Tabelle 15 Tatsächliches und vorhergesagtes Wahlergebnis für Sonnerberg 72
Tabelle 16 Tatsächliches und vorhergesagtes Wahlergebnis für München 73
Tabelle 17 Tatsächliches und vorhergesagtes Wahlergebnis für Oberweser 73
Tabelle 18 Tatsächliches und vorhergesagtes Wahlergebnis für Konnersreuth 73
Tabelle 19 Fehlerwerte über alle Gemeinden 75
Tabelle 20 Absoluter Fehler nach Parteien 77
Tabelle 21 Verwendete Attribute 91

1. Einleitung

Im Jahr 2017 wird die Wahl zum 19. Deutschen Bundestag stattfinden. Als Verstärkung wird die SPD im Wahlkampf von Jim Messina unterstützt werden. Dieser hat bereits Wahlkampferfahrung aus den USA. Bei der Präsidentschaftswahl im Jahr 2012 verstärkte er das Wahlkampfteam des erneut antretenden Präsidenten Barack Obama, für den es in der Wahl um eine zweite Amtszeit ging. Um die Stimmen der Wähler zu gewinnen, wurde eine sehr große Menge an Daten verwendet. Diese ermöglichten eine persönlichere politische Kommunikation mit den Bürgern. Mittels der Daten wurden für eine Vielzahl an Wahlberechtigten Kennzahlen berechnet, welche die optimale Interaktion mit diesen Personen gewährleisten sollten. Den Wahlkampf-Verantwortlichen wurde dadurch ermöglicht, die mit dem Wahlkampf verbundenen Tätigkeiten gezielter zu steuern.

Für die Wahlberechtigten wurde zum Beispiel berechnet, wie hoch die Wahrscheinlichkeit ist, dass er als Freiwilliger zur Unterstützung des Obama-Wahlkampfs gewonnen werden kann. Außerdem wurde ermittelt, ob ein Wahlberechtigter zur Teilnahme an der Wahl bewegt werden konnte, vom demokratischen Präsidentschaftskandidaten überzeugt werden konnte oder ein potentieller Spender war, der den Wahlkampf finanziell unterstützen würde. (Nickerson und Rogers 2014)

Für diese Berechnungen wurden Daten über individuelle Bürger gesammelt und verarbeitet. Die Ausgangsbasis dafür bildeten die öffentlich zugänglichen Wählerverzeichnisse der einzelnen Bundesstaaten. Um seine Stimme abgeben zu dürfen, muss sich ein Bürger in das Wählerverzeichnis seines Staates eintragen. Diese Daten wurden mit weiteren Daten angereichert, um eine möglichst genaues Bild über individuelle Bürger und ihr wahlrelevanten Einstellungen und ihr voraussichtliches Verhalten zu zeichnen. Dieser datengetriebenen Strategie wurde häufig auch das Schlagwort Big Data zugeordnet. Damit werden Datenmengen beschrieben, deren Umfang so groß ist, dass die traditionelle Datenverarbeitung nicht mit ihnen umgehen kann. Eine interessante Aussage zum Thema Big Data stammt von der Bundeskanzlerin Angela Merkel. Ihrer Ansicht nach sind Daten die Rohstoffe des 21. Jahrhunderts, die für den Zugang zu Kunden sehr wichtig sind. Deutschland tue sich schwerer als andere Länder, diese Daten auszuwerten. Die Politik müsse dafür sorgen, dass der rechtliche Rahmen so gesetzt wird, „dass man das Big Data Management sehr gut machen kann". (Bundeskanzlerin 2016)

Im Lichte der bald anstehenden Bundestagswahl drängt sich die Frage auf, inwieweit das Vorgehen in den USA im deutschen Wahlkampf eine Rolle spielen könnte und generell überhaupt möglich ist. Diese Frage zu beantworten ist das Ziel der vorliegenden Masterarbeit.

Um die Übertragbarkeit von datengetriebenen Wahlkampfmethoden auf Deutschland zu überprüfen wird folgendermaßen vorgegangen: Zunächst wird im zweiten Kapitel auf die theoretischen Grundlagen eingegangen, die für die Bearbeitung des Themas wichtig sind. Dabei werden politische und technische Themen betrachtet. Daraufhin wird im dritten Kapitel das Vorgehen der Wahlkampagnen von Barack Obama näher behandelt. Anschließend wird die Übertragbarkeit der datengetriebenen Methoden im US-Wahlkampf auf Deutschland überprüft. Dabei wird neben den Grundlagen der deutschen Parteienfinanzierung und den verfügbaren Datenquellen auch auf die in Deutschland geltenden relevanten rechtlichen Bestimmungen eingegangen. Nach der theoretischen Betrachtung des Themas folgt im fünften Kapitel eine praktische Umsetzung einer Vorhersage mit einem Teil der in Deutschland verfügbaren Daten. Dafür werden Daten aus verschiedenen Quellen zusammengetragen, miteinander verknüpft und ausgewertet. Aus den Daten wird mit einem Random Forest ein Klassifikationsmodell erlernt. Danach folgt eine Diskussion der Ergebnisse, wobei auch die Limitationen der Arbeit aufgezeigt werden. Abgeschlossen wird die Arbeit durch eine Zusammenfassung des Themas und einen Ausblick.

2. Theoretische Grundlagen

Zunächst wird auf die politischen und technischen Grundlagen der Arbeit eingegangen. Nach einer Beschreibung des Parteibegriffs nach deutschem Recht wird der Ablauf einer Bundestagswahl erläutert. Im Zuge dessen wird auch auf die Bedeutung des Wahlkampfs für Wahlen eingegangen und es werden die derzeit bedeutendsten deutschen Parteien vorgestellt. Auf der technischen Ebene erfolgt eine kurze Behandlung der Themen Big Data, Data Mining und maschinelles Lernen.

2.1. Parteien und Wahlen

Parteien sind Vereinigungen von Bürgern, die entweder für eine längere Zeit oder dauernd für den Bereich des Bundes oder des Landes Einfluss auf die politische Willensbildung nehmen oder an der Vertretung des Volkes im Bundestag oder einem Landtag mitwirken wollen. Durch die Beeinflussung der politischen Willensbildung übernehmen Parteien eine öffentliche Aufgabe. Parteien müssen eine ausreichende Gewähr bieten, dass dieses Ziel ernsthaft ist. Die Ernsthaftigkeit wird beeinflusst von Umfang und Festigkeit der Organisation, der Mitgliederanzahl und ihrem Hervortreten in der Öffentlichkeit. Mitglieder einer Partei können nur natürliche Personen sein. Parteien sind in Gebietsverbände untergliedert und müssen über eine schriftliche Satzung und ein schriftliches Programm verfügen. Eine Partei setzt sich aus verschiedenen Organen zusammen. Die Mitgliederversammlung, die auch als Parteiversammlung oder Hauptversammlung bezeichnet wird, ist das oberste Organ eines Gebietsverbandes. Neben der Mitgliederversammlung gibt es einen Parteivorstand. Dieser setzt sich aus mindestens drei Mitgliedern zusammen und wird alle zwei Jahre gewählt. Er übernimmt die Aufgaben der Leitung des Gebietsverbandes und dessen Geschäftsführung. Zusätzlich zu diesen beiden Organen existieren Vertreterversammlungen, Parteischiedsgerichte, allgemeine Parteiausschüsse und ähnliche Einrichtungen. Parteiausschüsse haben gemäß der Satzung umfassende Zuständigkeiten für die Beratung oder Entscheidung politischer oder organisatorischer Fragen der Partei. (§§ 1, 6, 7, 9, 11-14 PartG)

In Deutschland finden Wahlen auf kommunaler, Landes-, Bundes- und auf europäischer Ebene statt. Wahlberechtigt und wählbar ist ein deutscher Bürger, der das 18. Lebensjahr vollendet hat. Durch Wahlen übt das deutsche Volk die von ihm ausgehende Staatsgewalt aus. Das Volk muss eine Vertretung haben, die aus einer Wahl hervorgeht. Dabei gelten die Grundsätze einer allgemeinen, unmittelbaren, freien, gleichen und geheimen Wahl. Die bei einer Bundestagswahl gewählten Abgeordneten des deutschen Bundestags sind Vertreter des ganzen Volkes. Sie sind an Aufträge und Weisungen nicht gebunden und nur ihrem Gewissen unterworfen. Sie werden für vier Jahre gewählt. (Artt. 29 Abs. 2, 38 Abs. 1-2, 39 Abs. 1 GG)

Das deutsche Bundesgebiet wird für die Bundestagswahl in 299 Wahlkreise unterteilt, welche wiederum in Wahlbezirke aufgeteilt sind. In der Regel bilden Gemeinden mit maximal 2500 Einwohnern einen Wahlbezirk. Bei größeren Gemeinden erfolgt eine Einteilung in mehrere Wahlbezirke. Diese werden von der Gemeindebehörde bestimmt. Darüber hinaus bilden bestimmten Einrichtungen Sonderwahlbezirke. Dabei handelt es sich zum Beispiel um Krankenhäuser, Altenheime und Pflegeheime. Dadurch wird Wahlberechtigten, die während des Wahltags an die Einrichtung gebunden sind, erleichtert bzw. ermöglicht. Zur Veranschaulichung wird die Gemeinde Konnersreuth betrachtet. Konnersreuth ist eine Gemeinde mit knapp 2000 Einwohnern. Da sie nicht mehr als 2500 Einwohner hat, ist sie nicht in mehrere Wahlbezirke aufgeteilt. Konnersreuth gehört zum 235. Wahlkreis Weiden. Mit der Erststimme können die Bewohner von Konnersreuth einen der Direktkandidaten für den Wahlkreis Weiden wählen. (§§ 1-2 BWahlG, §§ 12-13 BWO)

Wähler haben bei der Bundestagswahl zwei Stimmen, eine Erst- und eine Zweistimme. Diese Stimmen legen die Sitzverteilung der 598 Sitze im Bundestag fest. Mit der Erststimme wird ein Wahlkreisabgeordneter in einem der 299 Wahlkreise gewählt, in die das Bundesgebiet eingeteilt ist. Der Kandidat mit den meisten Stimmen in seinem Wahlkreis ist gewählt und erhält ein Direktmandat,

das ihm einen Sitz im Bundestag sichert. Die Erststimme gewährleistet, dass jeder der 299 Wahlkreise im Bundestag vertreten ist. Die Zweitstimme dient der Wahl der Landesliste einer Partei. In der Landesliste sind nummeriert die Kandidaten einer Partei aufgelistet, die von einem Bundesland in den Bundestag geschickt werden. Die Zweistimmen werden zusammengezählt und, basierend auf ihrer Verteilung auf die Parteien, die Sitzverhältnisse im Bundestag festgelegt. Die Sitze, die eine Partei erhält, werden zunächst mit den mit der Erststimme gewählten Wahlkreisabgeordneten besetzt. Die danach verbleibenden Plätze werden mit den Kandidaten der Landesliste von oben absteigend aufgefüllt. Wenn eine Partei aus den Erststimmen mehr Direktmandate erhält, als ihr Sitze gemäß der Zweitstimme zustehen, erhält die Partei sogenannte Überhangmandate, also zusätzliche Sitze. Damit die Sitzverteilung dennoch gemäß den Zweitstimmen aufrechterhalten wird, erhalten die anderen Parteien Ausgleichsmandate. Die Anzahl der Sitze im Bundestag wird so lange erhöht, bis das Sitzverhältnis wieder im richtigen Verhältnis ist. Ausgleichsmandate wurden erst im Jahr 2013 eingeführt. Der derzeitige Bundestag hat 630 Sitze. Bei der Wahl ziehen nur Parteien in den Bundestag ein, die mindestens 5% der Zweitstimmen erhalten oder in mindestens drei Wahlkreisen einen Sitz errungen haben. (§§ 1 Abs. 2, Abs. 4, 6 Abs. 3 BWahlG)

Die nachfolgende Abbildung zeigt, wie viel Prozent der Stimmen die einzelnen Parteien bei der Bundestagswahl 2013 erhalten haben.

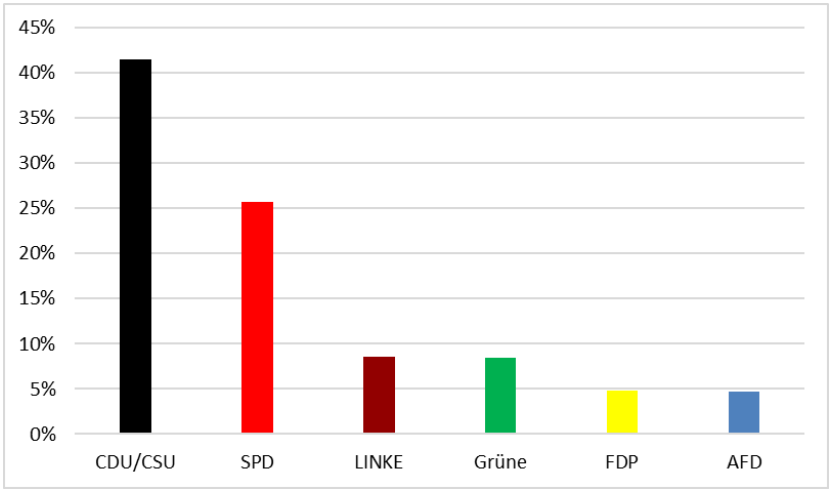


Abbildung 1 Verteilung der Stimmen bei der Bundestagswahl 2013

Die CDU und ihre Schwesternpartei CSU haben zusammen mehr als 40% der Stimmen erhalten. Danach folgte die SPD mit knapp 26% der Stimmen. Die Linke als nächststärkste Kraft konnte 8,6% der Stimmen auf sich vereinigen, knapp gefolgt von den Grünen mit 8,4%. Die Parteien FDP und AfD scheiterten mit 4,8% beziehungsweise 4,7% knapp an der 5%-Hürde.

Dem Wahltag geht ein Wahlkampf voraus. Dieser kann als eine politische Auseinandersetzung von Parteien zum Gewinnen der Zustimmung des Bürgers zu Personen und Programmen definiert werden. Manchmal wird auch von einem kontinuierlichen Wahlkampf gesprochen. Der eigentliche Wahlkampf aber findet zwischen der Auflösung des Parlaments und der Wahl des neuen Parlaments statt. Im Wahlkampf präsentieren die Parteien dem Bürger ihre Ziele, ihre Politiker und ihren Standpunkt zu wichtigen Themen. In dieser Zeit der intensivierten Wählerkommunikation wird der Wähler nicht nur sachbezogen, sondern auch emotional angesprochen. Ein Wahlkampf übt die drei Funktionen Information, Identifikation und Mobilisierung aus. Während des Wahlkampfs wird der Bürger verstärkt informiert. Dies geschieht unter anderem in Form von Wahlprogrammen, politischen Äußerungen der Kandidaten oder Parteiwerbung an Plakatwänden. Für jeden Wahlkampf wird von jeder Partei ein Wahlprogramm erstellt, das dem Wähler neben einer besseren Orientierung auch eine weitere Identifikation mit der Partei ermöglicht. Die politischen Botschaften sind dabei

vereinfacht und werden besonders einprägsam und öffentlichkeitswirksam vermittelt. Wichtig ist das Besetzen von Themen und das Aufzeigen eines Kompetenzvorsprungs in diesen. Die Identifizierung mit der Partei zielt vor allem auf die Mitglieder und Anhänger einer Partei. Während der verstärkten Außendarstellung der Parteien im Wahlkampf können sich Mitglieder und Anhänger leichter zur Partei bekennen und für sie werben. Die Identifizierung mit der Partei fördert die Motivierung und die Mobilisierung von Mitgliedern und der Partei nahestehenden Wählergruppen und Personen. (Woyke 1998, S.105-108)

Wählermobilisierung bedeutet, Wahlberechtigte dazu zu bewegen, zur Wahl zu gehen. Eine hohe Wahlbeteiligung bedeutet eine bessere demokratische Legitimation der Regierung. Aus demokratischer Sicht ist eine möglichst hohe Wahlbeteiligung also wünschenswert. Für die einzelnen Parteien trifft dies nicht unbedingt zu. Für das Wahlergebnis einer Partei ist es am besten, wenn möglichst viele Wähler, die die eigene Partei wählen, zur Wahl gehen. Gleichzeitig profitiert eine Partei davon, wenn Wahlberechtigte, die den anderen Parteien zugeneigt sind, nicht zur Wahl erscheinen. Parteien steht für ihre Wahlkampagnen ein begrenztes Budget zur Verfügung. Es wird versucht, mit den zur Verfügung stehenden Mitteln das bestmögliche Wahlergebnis zu erreichen. Dazu ist eine effiziente Zuordnung der vorhandenen Ressourcen nötig. Diese Zuordnung kann durch die Verwendung von Informationen verbessert werden.

Um einen Überblick über die aktuelle Parteienlandschaft zu erhalten, werden die aktuell in den Bundestag gewählten Parteien CDU, CSU, SPD, Grüne und Linke kurz vorgestellt. Außerdem wird auf die Parteien FDP und AfD eingegangen, welche im nächsten Jahr Umfragen zufolge in den Bundestag einziehen werden (wahlrecht.de 2016). Die Vorstellung erfolgt alphabetisch auf Basis der von den Parteien erstellten Grundsatzprogramme, in denen ihre Ausrichtung und ihre Werte festgeschrieben sind. Die Grundsatzprogramme aller Parteien werden auf der Seite des Bundeswahlleiters bereitgestellt. (Der Bundeswahlleiter 2016)

Die im Februar 2013 gegründete Alternative für Deutschland (AfD) bezeichnet sich als liberal, konservativ und demokratisch. Ihrer Ansicht nach wird die gegenwärtige Politik von einem Bruch von Recht und Gesetz und verantwortungslosem Handeln gegen die Prinzipien der Vernunft geprägt. Die Partei strebt die dauerhafte Erhaltung der Würde des Menschen, der Familie mit Kindern, der abendländischen christlichen Kultur, der Sprache und der Kultur in einem Nationalstaat des deutschen Volkes an. Die EU in Form von „Vereinigten Staaten von Europa“ wird abgelehnt und soll eine Wirtschafts- und Interessensgemeinschaft souveräner, lose verbundener Einzelstaaten sein. Die Einführung des Euro war ein rein politisches Projekt und führte zu Feindseligkeiten zwischen den europäischen Völkern. Die AfD tritt für die Abschaffung des Rundfunkbeitrags ein und sieht den Islam im Spannungsverhältnis zur deutschen Werteordnung.

Die Christlich Demokratische Union (CDU) bezeichnet sich als Volkspartei der Mitte mit konservativen, liberalen und christlich-sozialen Wurzeln. Sie hat ein christliches Menschenverständnis und sieht die Verantwortung des Menschen vor Gott. Die Eigenverantwortung der Bürger soll gefördert werden. Der Boden der Leitkultur in Deutschland ist durch die europäische und deutsche Geschichte mit ihren föderalen und konfessionellen Traditionen gegeben. Das Fundament der Gesellschaft bilden Ehe und Familie. Das Wirtschaftssystem soll eine soziale Marktwirtschaft mit ökologischer Ausrichtung sein. Der europäische Einigungsprozess muss fortgesetzt werden; die Nationalstaaten sollen dabei nicht aufgelöst werden. Die CDU tritt in allen Bundesländern mit Ausnahme von Bayern an. Dort tritt ihre Schwesterpartei, die Christlich-Soziale Union, an. Diese ist der CDU inhaltlich ähnlich. Die CSU steht für eine starke Leistungskultur. Sie sieht, dass viele Menschen mangelnde Chancen haben. Der Grund dafür ist nicht die Globalisierung, sondern politische Fehlsteuerungen. Der politische Irrweg des Versorgungsstaats schwächt die Eigeninitiative, untergräbt die soziale Verantwortung des Einzelnen und bringt die Menschen in eine falsche Abhängigkeit. Für die CSU gehören Weltoffenheit und Heimatliebe zusammen.

Die Freie Demokratische Partei (FDP) ist eine demokratische, liberale Partei, die das Zusammenleben in einer freien, offenen Bürgergesellschaft gestalten will. Ihrer Aussage nach schaffen es nur die Liberalen, dass Wachstum und die Grundlagen zu sichern, auf denen Frieden, Freiheit und Wohlstand gedeihen. Die Partei will die Voraussetzungen schaffen, dass jeder Mensch faire Chancen hat, seine Talente zu nutzen, von seiner Arbeit zu leben und auf seine Weise glücklich zu werden. Dabei ist auch Toleranz ein wichtiger Aspekt. Um die Freiheit der Menschen zu bewahren, legen der Rechtsstaat, die soziale Marktwirtschaft und die Demokratie fest, wo die Freiheit des Einzelnen endet. Liberale Bildungspolitik garantiert gleiche Startchancen, aber nicht gleiche Ergebnisse. Die Akzeptanz für die Wirtschaftsordnung soll zurückgewonnen werden. Die Staatsverschuldung soll von 80 Prozent auf 50 Prozent zurückgeführt werden. Dafür muss die Gefälligkeitspolitik aufhören, die unbezahlbare Ansprüche an den Staat fördert. Europa soll stärker zusammenwachsen.

Im Mittelpunkt der Partei Bündnis90/Die Grünen steht der Mensch mit seiner Würde und seiner Freiheit. Er kann als Teil der Natur nur leben, wenn er sie als Lebensgrundlage schützt. Die Ökologie bildet die Grenze des Industrialismus. Jeder Mensch ist einzigartig und verdient die gleiche Anerkennung. An der gleichen Behandlung von Menschen misst sich die Gerechtigkeit. Neoliberale Wirtschaftspolitik wird abgelehnt. Europa kann sich nicht als Wohlstandsinsel gegen die übrige Welt abschotten. Das Programm der Grünen wird durch zwölf Schlüsselprojekte geprägt. Diese sind der Aufbruch ins ökologische Zeitalter, Transparenz für Verbraucher, eine neue Landwirtschaft, eine Entwicklung von Ostdeutschland, eine Grundsicherung als Grundlage sozialer Sicherheit, Generationengerechtigkeit, Wissenszugang als Bürgerrecht, die Gleichstellung der Geschlechter, die Einwanderung als Chance, europäische Integration und auf globaler Ebene fairer Handel und internationale Standards.

Die Linke ist eine demokratische, sozialistische Partei. Sie strebt eine Abkehr des bestehenden Wirtschafts- und Gesellschaftssystems an und will einen demokratischen Sozialismus aufbauen. Der Kapitalismus soll überwunden werden. Die Partei kämpft für Menschenrechte und Emanzipation und gegen Faschismus, Rassismus, Imperialismus und Militarismus. Das Programm der Linken wird durch drei Grundideen geprägt, welche auf sozialen und ökologischen Kräften basieren. Die erste Idee beinhaltet die individuelle Freiheit und Entfaltung der Persönlichkeit für jeden durch eine sozial gleiche Teilhabe an den Bedingungen eines selbstbestimmten Lebens. Zweitens soll die Wirtschaft der solidarischen Entwicklung und dem Erhalt der Natur untergeordnet werden. An der Stelle profitorientierten Wachstums soll eine nachhaltige Entwicklung stehen. Durch die vorherigen beiden Ideen wird in einem länger dauernden Prozess die Vorherrschaft des Kapitals überwunden. Die EU soll vor allem eine soziale, ökologische Friedensunion mit demokratisch kontrollierter Wirtschaftspolitik sein.

Die Sozialdemokratische Partei Deutschlands (SPD) tritt für Fortschritt und Gerechtigkeit im 21. Jahrhundert ein. Entstanden als Teil der Arbeiterbewegung, vertritt sie die Grundwerte Freiheit, Gerechtigkeit und Solidarität. Die SPD will Politik für die solidarische Mehrheit machen. Wichtig für die Partei sind eine solidarische Bürgergesellschaft und ein demokratischer Staat, die Gleichstellung der Geschlechter, nachhaltiger Fortschritt und qualitatives Wachstum. Die SPD steht für einen vorsorgenden Sozialstaat ein, der Armut bekämpft, eine gleiche Chance auf ein selbstbestimmtes Leben eröffnet und die großen Lebensrisiken absichert. Der Mensch wird als vernunftbegabt, lernfähig, aber auch fehlbar angesehen. Gleiche Chancen bedeuten, Raum für die Entfaltung der individuellen Fähigkeiten zu geben, die nicht von der sozialen Herkunft abhängig sein soll. Für die EU werden verbindliche gesamtwirtschaftliche Vorgaben gebraucht und die europäische Sozialunion muss gleichrangig neben die Wirtschafts- und Währungsunion treten.

2.2. Technische Grundlagen

Um die technischen Grundlagen des Themas zu erläutern, erfolgt eine Einführung in die zusammenhängenden Themen Big Data, Data Mining und Maschinelles Lernen.

2.2.1. Big Data

Der Begriff Big Data entstand, um Datenmengen zu beschreiben, die so groß sind, dass für ihre Bearbeitung Supercomputer benötigt werden. Die Notwenigkeit der Verwendung eines Supercomputers ist vom aktuellen Stand der Technik abhängig. Aufgrund der beständigen Verbesserung von Rechenkapazitäten sind Datenmengen, für die zu einem früheren Zeitpunkt ein Supercomputer benötigt wurde, gegenwärtig mit einem normalen Desktop-Computer analysierbar. Daher ist die Bezeichnung Big Data ungünstig gewählt, da er suggeriert, dass alleine das Volumen der Daten von Bedeutung ist. Bei Big Data geht es weniger um große Datenmengen, sondern um die Fähigkeit, große Datenmengen zu durchsuchen, zu aggregieren und eine Beziehung zwischen den einzelnen Datenelementen herzustellen. Big Data wird dabei von vier Komponenten bestimmt. Diese sind Varietät, Umfang, Geschwindigkeit und Wert. Im englischspracheigen Raum sind diese vier Begriffe *variety*, *volume*, *velocity*, *value* und werden wegen ihres Anfangsbuchstabens auch als die vier Vs bezeichnet. Varietät bedeutet, dass Daten aus einer großen Quellenvielfalt stammen und entweder strukturiert, semi-strukturiert oder unstrukturiert vorliegen. Umfang meint die Größe der Daten, die die Größenordnung von Petabytes übersteigt und für die bisherige Speicher- und Analysemethoden nicht ausreichend sind. Geschwindigkeit betrifft den Umstand, dass der Datenstrom für zeitlimitierte Prozesse umgehend beim Erhalt der Daten genutzt werden sollte. Wert sagt aus, dass die Erkenntnisse, die sich aus den ausgewerteten Daten ergeben, für den Auswertenden von Bedeutung sind. (Boyd und Crawford 2012, Sagiroglu und Sinanc 2013)

2.2.2. Data Mining

Data Mining ist eng mit den Themengebieten Big Data und maschinellem Lernen verknüpft. Es geht dabei um die Anwendung von Algorithmen zur nicht-trivialen Extraktion von impliziten, unbekannten und möglicherweise nützlichen Informationen aus Daten. Die Methoden von Data Mining können für verschiedene Anwendungszwecke nützlich sein. Dabei kann zwischen prädiktiver und deskriptiver Induktion unterschieden werden. Das Ziel der prädiktiven Induktion ist die Entdeckung von Wissen für Klassifikation und Vorhersage. Dafür können unter anderem Verfahren der Klassifikation oder Regression verwendet werden. Bei der deskriptiven Induktion geht es um die Extraktion von interessantem Wissen aus den Daten. Methoden hierfür sind Assoziationsregeln oder die Entdeckung von charakteristischen Teilgruppen. Bei der Entdeckung von Teilgruppen wird von einer Population von Individuen ausgegangen und einer Eigenschaft dieser Individuen. Die Aufgabe besteht darin, die Teilgruppen der Population auszumachen, die hinsichtlich der Eigenschaft aus statistischer Sicht am interessantesten sind. (Fayyad, Piatetsky-Shapiro und Smyth 1996, Herrera, Carmona, González et al. 2011)

Ein Standard-Prozess-Modell für die Umsetzung eines Data-Mining-Projekts ist CRISP-DM, das im Rahmen eines EU-Projekts entwickelt und im Jahr 1999 veröffentlicht wurde. Der Name steht dabei für Cross Industry Standard Process for Data Mining. Es ist ein hierarchisches Prozessmodell, das aus einer Menge an Aufgaben besteht, die auf vier Abstraktionsebenen beschrieben werden. Auf der obersten Abstraktionsebene stehen die Phasen. Diesen Phasen ist auf der Abstraktionsebene darunter jeweils eine Menge an generischen Aufgaben zugeordnet. Auf der dritten Ebene sind die spezialisierten Aufgaben, wobei eine Menge an spezialisierten Aufgaben jeweils genau einer generischen Aufgabe zugeordnet ist. Auf der niedrigsten Ebene stehen die Prozessinstanzen, die den spezialisierten Aufgaben zugeordnet werden. Der Ablauf der Phasen ist in der nachfolgenden Abbildung dargestellt.

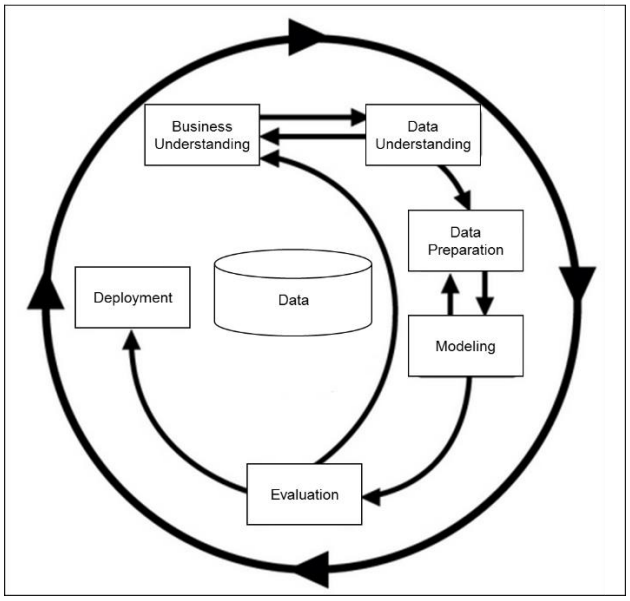


Abbildung 2 Phasen des Prozessmodells CRISP-DM

In der Phase Business Understanding geht es um das Bestimmen der Geschäftsziele, die erreicht werden sollen. Im Zuge dessen wird auch die geschäftliche Situation unter anderem hinsichtlich Kosten, Anforderungen und Risiken abgeschätzt. In dieser Phase werden auch die Ziele des Data Mining festgelegt. In der nachfolgenden Phase des Data Understanding werden die Daten aus den verschiedenen Datenquellen gesammelt, beschrieben, ihre Eigenschaften untersucht und die Datenqualität bewertet. Die Phase der Data Preparation besteht aus der Auswahl der Daten aus den zuvor gesammelten Daten, dem Reinigen der Daten, der Bearbeitung der Daten durch zum Beispiel Transformation oder Ableitung neuer Attribute, der Integration von Daten aus verschiedenen Datenquellen und der Formatierung der Daten. Die Phase Modeling beinhaltet die Auswahl einer Modellierungstechnik, das Erstellen eines Test-Designs und den Bau und die Bewertung des Modells. Anschließend werden in der Phase der Evaluation die Ergebnisse bewertet und der bisherige Prozess überprüft. Zuletzt werden im Schritt Deployment die Ergebnisse veröffentlicht. Zu beachten ist, dass die Phasen nicht streng sequentiell ablaufen. Die Erkenntnisse aus einer Phase können sich auf vor- oder nachgelagerte Phasen auswirken. (CRISP-DM 2016)

2.2.3. Maschinelles Lernen

Bei maschinellem Lernen geht es, neben dem Bau von Computersystemen, die sich durch Erfahrung verbessern, um die grundlegenden Mechanismen von Lernprozessen. Maschinelles Lernen verbindet Informatik, Statistik und zu einem geringeren Teil auch die Lernforschung in der Psychologie und Neurowissenschaft. Es geht darum, Muster, Regelmäßigkeiten und Modelle in Daten zu erkennen (Mitchell 2006). Methoden des maschinellen Lernens finden in der Phase *Modeling* des CRISP-DM-Prozessmodells aus Abbildung 2 statt. Die in den Daten gefundenen Muster werden verallgemeinert. Auf Basis dieser Verallgemeinerung wird ein Modell erstellt, das Objekte anhand ihrer Eigenschaften unterscheidet. Durch das Modell können die einzelnen Objekte einer von mehreren Klassen zugeordnet werden. Dies soll nun anhand eines Beispiels verdeutlicht werden. Dazu dient der im Folgenden dargestellte Beispieldatensatz.

Nr.	Geschlecht	Bundesland	Parteimitglied	Wahlteilnahme
1	Männlich	Sachsen	Ja	Ja
2	Weiblich	Bayern	Ja	Ja
3	Weiblich	Sachsen	Nein	Nein
4	Männlich	Bayern	Nein	Ja
5	Weiblich	Hessen	Ja	Ja
6	Männlich	Hessen	Ja	Ja
7	Männlich	Sachsen	Nein	Nein

Tabelle 1 Beispieldatensatz zur Wahlteilnahme

Die Tabelle besteht aus sieben Datenobjekten oder Instanzen, wobei jedes Datenobjekt eine Person darstellt. Jede Person hat vier verschiedene Attribute. Das Klassenattribut legt die Klasse des Datenobjekts fest und ist im vorliegenden Fall das Attribut *Wahlteilnahme*. Die Spalte mit dem Klassenattribut ist in der Tabelle hellgrau gekennzeichnet. Das Klassenattribut hat zwei verschiedene Ausprägungen, *Ja* und *Nein*, die als Klassenlabel bezeichnet werden. Ein Lernalgorithmus wird auf Basis der Attribute ein Modell erlernen, welches für ein unbekanntes Datenobjekt das Klassenattribut vorhersagt. Im Beispiel würde einem unbekannten Datenobjekt dann die Klasse ja oder nein zugeordnet werden. Mit dem gelernten Modell ist es möglich, neuen Datenobjekten, deren Klasse unbekannt ist, einer Klasse zuzuordnen. Da für jedes Objekt die Klasse bekannt ist, handelt es sich in diesem Fall um überwachtes Lernen.

Um die Qualität eines Modells zu bewerten, werden die Datenobjekte vor dem Lernen des Modells in drei disjunkte Mengen aufgeteilt. Diese sind die Trainingsmenge, die Validierungsmenge und die Testmenge. Mit der Trainingsmenge wird ein Klassifizierer gelernt. Die Validierungsmenge dient der Optimierung des Klassifizierers oder der seiner Auswahl, wenn zwischen mehreren gelernten Klassifizierern ausgewählt wird. Auf der Testmenge wird der Klassifizierer angewandt. Sie dient der Beurteilung seiner Qualität bei neuen Daten. Es können auch nur zwei der Mengen verwendet werden. Dann wird der Datensatz in eine Trainingsmenge und eine Testmenge aufgeteilt und die Validierungsmenge entfällt. Für einen Klassifizierer wird ein Fehlerwert berechnet. Dieser hängt davon ab, wie viele Objekte der Klassifizierer der richtigen Klasse zuordnen kann. Dieses Vorgehen soll das Auftreten einer Überanpassung an die Daten verhindern. Im einfachsten Fall würde ein Modell

sich alle Objekte und ihre Klassenzugehörigkeit einfach merken. Dann läge die Genauigkeit für den gelernten Datensatz bei 100%. Es ist aber der Fall, dass das auf diese Weise gelernte Modell eine Überanpassung, ein sogenanntes Overfitting, auf die Daten darstellt. Das Modell ist dann schlecht generalisierbar. Generalisierbarkeit gibt an, wie gut ein Modell neue Objekte, das heißt solche, die nicht für das Lernen verwendet wurden, klassifizieren kann. Ein überangepasstes Modell wird ein gutes Ergebnis auf der Trainingsmenge erzielen, aber ein deutlich schlechteres Ergebnis auf der Testmenge haben.

Klassifikationsverfahren lassen sich in zwei Gruppen einteilen. Zum einen gibt es symbolische Ansätze, die induktiv symbolische Beschreibungen lernen. Dieser Gruppe sind Regeln, Entscheidungsbäume und logische Repräsentationen zuzuordnen. Die andere Gruppe besteht aus statistischen Methoden oder Methoden der Mustererkennung. Darunter fallen instanzbasierte Methoden, der Bayes-Klassifizierer und neurale Netze. (Fürnkranz, Gamberger und Lavrač 2012, S. 1 f.)

Das durch ein Klassifikationsverfahren gelernte Modell muss für den Menschen nicht unbedingt verständlich sein. Die Nachvollziehbarkeit der Klassifikation hängt stark vom gewählten Klassifikationsverfahren ab. Gut für den Menschen interpretierbar sind Regeln und Entscheidungsbäume. Diese sollen kurz anhand des obigen Datenbeispiels dargestellt werden.

Eine Regel besteht aus einem Regelkörper und einem Regelkopf. Der Regelkörper beinhaltet eine Konjunktion von Bedingungen, die ein Datenobjekt erfüllen muss, um von der Regel abgedeckt zu werden. Der Regelkopf übernimmt die Vorhersage der Klasse für die Datenobjekte, die vom Regelkörper abgedeckt werden. Der Regelkörper wird mit dem Wort ‚IF' eingeleitet, der Regelkopf mit einem ‚THEN'. Für den Datensatz ergibt sich die in Abbildung 3 gezeigte Regel, die alle Beispiele mit der Klasse Nein abdeckt. Die Regel sagt aus, dass die Wahlteilnahme die Ausprägung *Nein* hat, falls das Bundesland Sachsen ist und keine Parteimitgliedschaft vorliegt. Zu beachten ist, dass diese Regel nicht alle Datenobjekte abdeckt. Das heißt, dass für die übrigen Datenobjekte weitere Regeln erstellt werden müssen. Bei mehreren Regeln wird von Regelmengen gesprochen. (Fürnkranz, Gamberger und Lavrač 2012, S. 25)

```
IF Bundesland = SACHSEN AND Parteimitglied = Nein
THEN Wahlteilnahme = Nein
```

Abbildung 3 Abgeleitete Regel für die Wahlbeteiligung

Bei einem Entscheidungsbaum wird eine Datenmenge so lange unterteilt, bis eine ausreichend gute Klassifizierung der Daten erreicht ist. Ein Baum besteht aus Knoten und Kanten. Ausgehend von einem Wurzelknoten wird ein Attribut ausgewählt, mit dem die Datenmenge getrennt wird. Dieser Vorgang wird solange durchgeführt, bis die Daten genau genug unterteilt sind. Knoten, die nicht weiter aufgeteilt werden, werden als Blattknoten bezeichnet. Die Blattknoten beinhalten die Klasse des Datenobjekts. Die Auswahl des Attributs, mit dem ein Knoten eine Menge in Untermenge aufteilt, erfolgt anhand einer Heuristik. Durch die Heuristik wir dasjenige Attribut ausgewählt, welches die Daten am besten trennt. Unterschiedliche Heuristiken können dabei zu unterschiedlichen Entscheidungen führen. Der Entscheidungsbaum ist fertiggestellt, wenn alle Datenobjekte in einem Blattknoten derselben Klasse angehören. Alternativ kann bestimmt werden, dass ein Knoten eine Mindestmenge an Datenobjekten enthalten muss, um weiter aufgeteilt zu werden, oder dass der Baum nur eine vorher festgelegte Tiefe haben darf. (Mitchell 1997, S. 52 ff.)

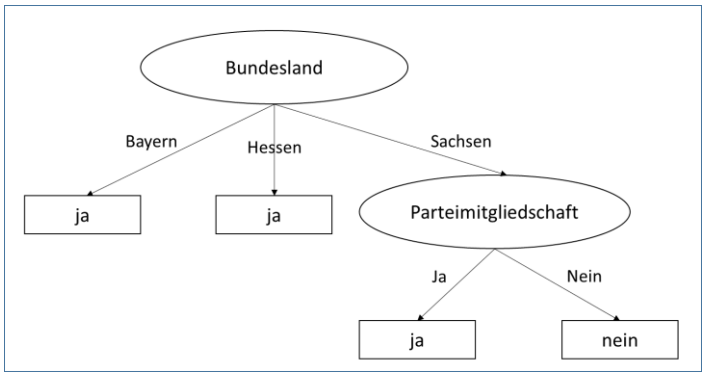


Abbildung 4 Berechneter Entscheidungsbaum für die Wahlbeteiligung

Der obige Entscheidungsbaum zeigt, dass das Bundesland und die Parteimitgliedschaft die diskriminierenden Attribute sind. Wie bei den Regeln ergibt sich, dass die Wahlteilnahme das Attribut Nein hat, wenn eine Person aus Sachsen stammt und kein Parteimitglied ist. Anders als bei Regeln werden durch einen Entscheidungsbaum alle Datenobjekte abgedeckt und klassifiziert. Anhand des erstellten Modells, entweder Regel oder Entscheidungsbaum, würden nun neue unbekannte Beispiele klassifiziert werden.

3. Bedeutung von Daten im US-amerikanischen Präsidentschaftswahlkampf

In diesem Kapitel wird auf die Rolle von Daten im US-amerikanischen Wahlkampf eingegangen. Dieses Thema stieß nicht nur in den USA, sondern unter anderem auch in Deutschland auf ein breites Medieninteresse. Im Fokus standen dabei die Wahlkampagnen von Barack Obama in den Jahren 2008 und 2012. Die Datennutzung der Partei der Republikaner erfuhr, wahrscheinlich auch aufgrund ihres ausbleibenden Wahlerfolges, nur wenig Aufmerksamkeit. Nach einer knappen Behandlung der wichtigsten Punkte der Präsidentschaftswahl wird die Entstehung des computergestützten Wahlkampfs thematisiert. Danach wird auf die verschiedenen Faktoren eingegangen, die den Präsidentschaftswahlkampf geprägt haben. Diese sind voneinander abhängig und können daher nicht vollkommen getrennt voneinander behandelt werden.

3.1. Die US-amerikanische Präsidentschaftswahl

Die Wahl eines Präsidenten folgt einem komplizierten und langwierigen Prozess. Die USA haben ein präsidentielles Regierungssystem, das im Gegensatz zu einem parlamentarischen Regierungssystem wie in Deutschland steht. Der Ablauf einer Präsidentschaftswahl ist in der Verfassung der Vereinigten Staaten von Amerika festgelegt. Präsident kann werden, wer mindestens 35 Jahre alt ist, in den USA geboren wurde und dort in den letzten 14 Jahren gelebt hat. Die Amtszeit des Präsidenten beträgt vier Jahre und die Präsidentschaft ist auf zwei Amtszeiten, also insgesamt acht Jahre, beschränkt. Der Präsident bildet die Exekutive der Regierung. Er ist der Regierungschef, das Staatsoberhaupt und hat die Kontrolle über die Streitkräfte. Der Präsident wird nicht direkt von den wahlberechtigten Bürgern gewählt. Bei der Wahl des Präsidenten werden zunächst in den Vorwahlen für jede Partei Delegierte gewählt, die beim Nominierungsparteitag dann den Kandidaten wählen dürfen, der für ihre Partei antreten soll. Die Vorwahlen finden nicht zeitgleich statt, sondern es gibt verschiedene Termine für die unterschiedlichen Bundesstaaten. Es gibt zwei Arten von Vorwahlen, Primary und Caucus. Die Art der Vorwahl wie auch die genaue Ausgestaltung werden von den Bundesstaaten festgelegt. In einigen Bundesstaaten ist es erforderlich, als Parteimitglied registriert zu sein, um an der Vorwahl teilnehmen zu können. Bei einer Primary wird in einer Wahlkabine die Stimme geheim abgegeben. Ein Caucus ist eine lokale Versammlung der Parteianhänger. Die Teilnehmer diskutieren miteinander und versuchen dabei für den von ihnen bevorzugten Kandidaten zu werben. Nach den Diskussionen findet eine Abstimmung statt, bei der die Parteianhänger den Kandidaten festlegen. Die in den Vorwahlen festgelegten Delegierten, sowie weitere sogenannte Superdelegierte, die nicht an die Wahl eines bestimmten Kandidaten gebunden sind, wählen auf dem Nominierungsparteitag ihrer Partei dann den Kandidaten. Nachdem die Kandidaten für die Parteien nach diesem Monate dauernden Prozess bestimmt sind, findet die Hauptwahl des Präsidenten statt. Der Präsident wird nicht von der wählenden Bevölkerung gewählt, sondern die Wahl erfolgt indirekt über 538 Wahlmänner. Die Anzahl der Wahlmänner pro Staat wird durch die Einwohner des Staates festgelegt, wobei mehr Einwohner die Zuteilung einer höheren Anzahl an Wahlmännern bedeutet. Die Wahl der Wahlmänner findet für alle Bundesstaaten am gleichen Termin statt. Die Wahlmänner werden mit Ausnahme von zwei Staaten mit einer einfachen Mehrheitswahl bestimmt. Aus der einfachen Mehrheitswahl folgt ein Winner-takes-it-all-Prinzip. Alle Wahlmänner eines Staates gehen an denjenigen Kandidaten, der die einfache Mehrheit der Stimmen auf sich vereinigt. Durch dieses Wahlsystem kann es dazu kommen, dass der Kandidat mit den meisten Stimmen die Wahl verliert. Im Jahr 2000 bekam der Präsidentschaftskandidat Al Gore 600.000 Stimmen mehr als sein Konkurrent George Bush, verlor aber die Wahl. Zudem sorgt das Wahlsystem dafür, dass kleine Parteien kaum eine Chance auf den Wahlsieg haben. Schließlich bringt es mit sich, dass im Wahlkampf bestimmten Staaten wie Texas oder Alabama keine besondere Beachtung geschenkt wird, da in diesen Staaten erfahrungsgemäß fast sicher eine der beiden Parteien den Großteil der Stimmen erhält. Hingegen ist eine kleine Anzahl an Staaten hart umkämpft, da der Wahlausgang dort offen ist. Diese Staaten werden Swing States genannt. Nachdem die 538 Wahlmänner gewählt wurden, geben diese 41 Tage später

ihre Stimme auf einem Stimmzettel ab. Präsident wird, wer die absolute Mehrheit, also mindestens 270 Stimmen erhält. Neben der alle vier Jahre stattfindenden Präsidentschaftswahl gibt es eine zweijährliche Kongresswahl. Der Kongress ist für die Kontrolle des Präsidenten, die Gesetzgebung und den Haushalt zuständig. Er besteht aus dem Repräsentantenhaus und dem Senat. Das Repräsentantenhaus hat 435 Sitze. Die Sitze werden entsprechend der Einwohnerzahl auf die Bundesstaaten verteilt. Der Senat hat 100 Mitglieder, wobei jeder der 50 Bundesstaaten zwei Senatoren entsendet. Die Senatoren sind für sechs Jahre im Amt und bei jeder Kongresswahl wird eine Drittel der Senatoren ausgetauscht. Dieses System ermöglicht, dass die Partei des Präsidenten nicht unbedingt mit der Partei der Mehrheit der Kongressmitglieder übereinstimmen muss. Die Finanzierung der Parteien ist in den USA nicht staatlich geregelt. Die Geldmittel, die den Kandidaten zur Verfügung stehen stammen von individuellen Spendern, Political Action Committees, den privaten Vermögen, den Parteien und der nur bei Präsidentschaftswahlen vorhandenen staatlichen Finanzierung. Den mit Abstand größten Anteil machen individuelle Spenden aus. Daher sind Kandidaten stark vom Sammeln von Spenden für den Wahlkampf abhängig. Pro Person beschränkt sich der Spendenbetrag auf 2600 US-Dollar. Bei einem Political Action Committee, kurz PAC, handelt es sich um eine Organisation, die Geldbeträge annimmt oder Ausgaben macht, um die Nominierung oder Wahl eines Individuums in ein politisches Amt zu beeinflussen. Sie werden von Unternehmen, Verbänden und anderen Organisationen genutzt, um Kandidaten finanziell zu unterstützen. Die Spenden, die ein PAC erhalten und an den Kandidaten weitergeben darf, sind begrenzt. Eine Besonderheit stellen hier sogenannte Super-PACS dar. Für diese gibt es keine Beschränkungen in der Höhe der Wahlausgaben, sofern die Ausgaben unabhängig, also nicht mit einer Partei oder einem Kandidaten koordiniert, gemacht werden. Dies wird durch das Recht auf freie Meinungsäußerung geschützt. (Weinmann 2016)

3.2. Die Entstehung des computergestützten Wahlkampfes

An der Entstehung computergestützter Politik haben nach (Tufekci 2014) mehrere zusammenhängende Faktoren mitgewirkt. Diese sind Big Data, individualisiertes Targeting, computergestützte Modellierung, die Anwendung von Verhaltensforschung zur Wählerüberzeugung, dynamische Echtzeitexperimente, die durch digitale Medien ermöglicht werden und das Entstehen von Datenbrokern, deren Geschäftsmodell die Bereitstellung von Daten ist. Zu teilweise gleichen Ergebnissen kommt (Bennett 2015), der vier Trends identifiziert, die die Wahlkampagnen in den USA beeinflusst haben. Auch er stellt fest, dass Nachrichten, die an die breite Masse gesendet werden, einem Micro-Targeting weichen, welches von kommerziellen Datenhändlern gekaufte Informationen verwendet. Zusätzlich identifiziert er drei weitere Trends. Aus technischer Sicht werden anstelle von Wählermanagement-Datenbanken integrierte Wählermanagement-Plattformen genutzt. Außerdem ist festzustellen, dass die Verwendung von sozialen Medien und des sogenannten sozialen Graphen zur Analyse genutzt werden. Schließlich ist eine Dezentralisierung von Daten mittels mobiler Anwendungen zu lokalen Kampagnen hin zu beobachten. Nach (Nickerson und Rogers 2014) konnte der jetzige computerbasierte Wahlkampf entstehen, als die technologischen und personellen Hemmnisse verschwunden waren. Auf technologischer Seite mangelte es an für Parteien bezahlbare Speicher- und Verarbeitungsverfahren von Daten. Außerdem waren die vorhandenen Daten weniger zuverlässig. Erst im Jahr 2002 war eine elektronische Speicherung des Wahlverhaltens von Bürgern bei den letzten vier Wahlen verbindlich. Auf personeller Seite fehlte es an Kompetenz in quantitativen Methoden auf Seiten der Politikberatung. Für Wahlkampfberater war ein breites technisches Wissen nicht erforderlich und die Politiker selbst haben, wie es auch in Deutschland der Regelfall ist, eine nicht-technische Ausbildung. Eine professionelle Datenanalyse war in dieser Zeit daher eher ein Nischenbereich. Als umfassende Datenverarbeitungsmethoden erschwinglich wurden und die Politikberatung starke quantitative Kompetenzen erwarb, waren die Hemmnisse beseitigt.

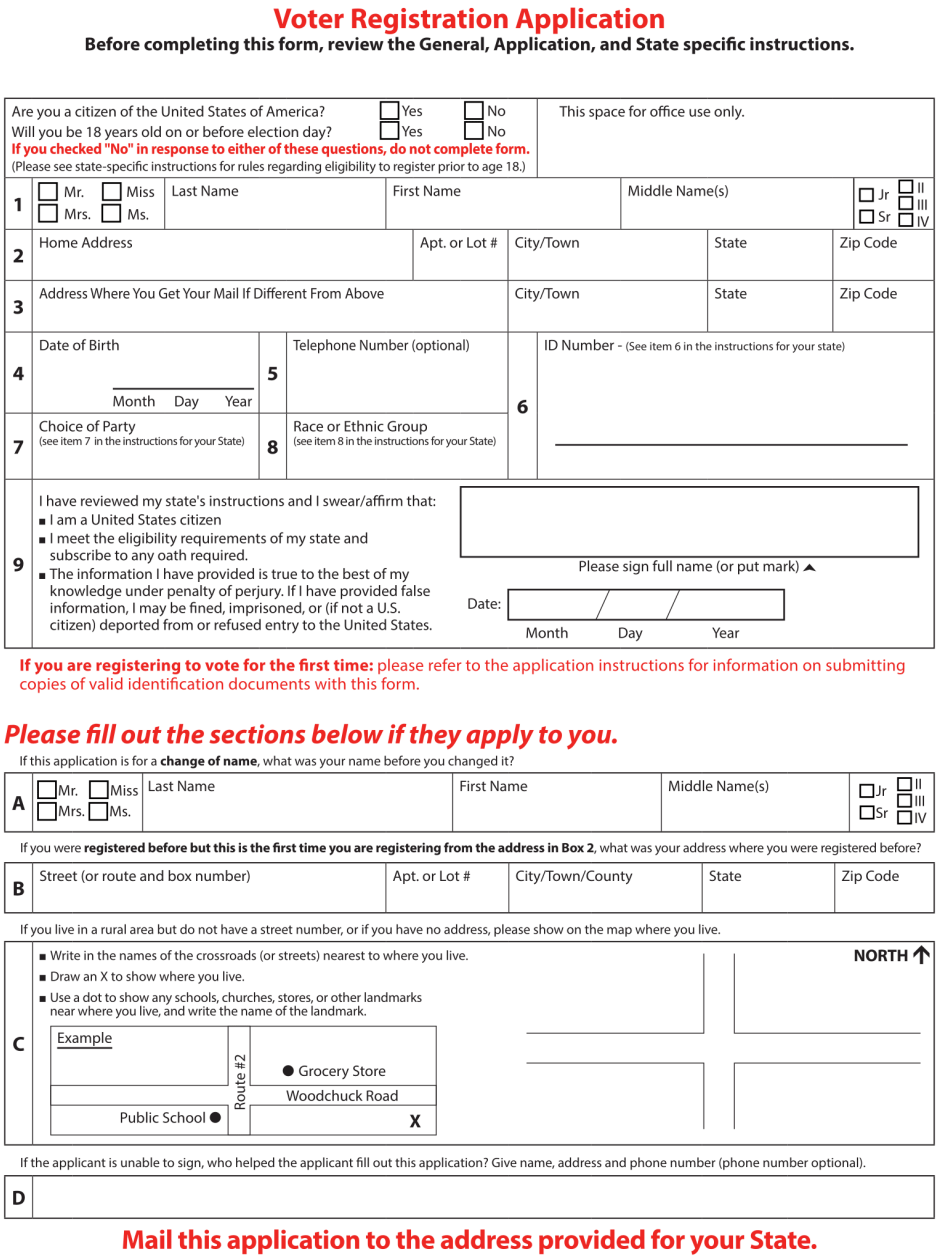
Vor der Verwendung anspruchsvoller Datenanalyse-Methoden beschränkte sich die Datenauswahl auf Parteizugehörigkeit, Charakteristika der Wahlbezirke, die Wahrscheinlichkeit eines Bürgers zur Wahl

zu gehen basierend auf den letzten vier zurückliegenden Wahlen, dem Kontaktieren von vorherigen Spendern und Umfragewerten. Zu diesem Zeitpunkt waren bereits Wählerregister auf Staatenebene und Zensusdaten vorhanden. Neben dem Wählerregister auf Staatenebene und Zensus-Daten wurden Daten von kommerziellen Verkäufern erworben und eigene Wählerdatenbanken betrieben. Diese Daten wurden genutzt, um statistische Methoden auf Kampagnen-Aktivitäten und Daten anzuwenden (Nickerson und Rogers 2014). Als aktuell verwendete Datenquellen werden in (Rubinstein 2014) vier verschiedene Arten von Wählerdaten beschrieben. Diese sind Wählerregister-Datenbanken auf Staatenebene, Daten über Spendengeber und die Reaktionen von Wählern auf verschiedene Maßnahmen, Daten von Kampagnenwebseiten und staatliche und nationale Wählerdateien.

Als Grund für eine Vorreiterposition der USA in der Wahlkampfführung kann der Umstand gesehen werden, dass das Fernsehen und die neuen Medien ihren Ursprung in den USA hatten, wodurch sich die am Wahlkampf Beteiligten dort früher zu einer Auseinandersetzung mit diesen Medien gezwungen sahen (Keim und Rosenthal 2016, S.308). Weiterhin gibt es in den Vereinigten Staaten begünstigende Faktoren für die Nutzung von Big Data und analytischen Methoden. Neben liberalen Gesetzen zur Wahlkampffinanzierung gibt ein dezentrales Parteiensystem viel lokale Autonomie. Das polarisierte politische System verschärft den Wettbewerb hin zu immer anspruchsvolleren Data Mining und Analysewerkzeugen. Zusätzlich existiert ein ausgedehnter kommerzieller Markt für private Daten ohne die Existenz von umfassenden Datenschutzgesetzen. Neben der reinen Machbarkeit aus technologischer Sicht gibt es aber auch eine Begründung, die aus einer gesellschaftlichen Entwicklung resultiert. Es ist eine Abwendung der Wähler von einzelnen Parteien zu beobachten. Weniger Wähler sind Parteimitglieder oder verspüren noch eine starke Zugehörigkeit zu einer Partei. Diese Entwicklung resultiert aus einem Vertrauensverlust in politische Institutionen. Aufgrund dieser Entwicklung besteht für Parteien die Notwendigkeit, neue Mittel zu finden, mit denen Geldgeber, Freiwillige und Mitglieder gefunden werden können. Diese Mittel sind diejenigen des computergestützten Wahlkampfs. (Bennett 2015)

3.3. Wählerregister

Um an der Wahl teilnehmen zu können, müssen sich die wahlberechtigten Bürger in das Wählerregister ihres Staats eintragen. Es gibt kein nationales Wählerverzeichnis. Um die Registrierung zu erleichtern, bietet die Regierung einen Service an, der einem Wahlberechtigten das nötige Vorgehen in seinem Staat erläutert beziehungsweise an die verantwortliche Stelle weiterleitet (United States Government 2016). Die folgende Abbildung zeigt ein generisches Formular, das ein Bürger zur Registrierung ausfüllen an die entsprechende staatliche Stelle weiterleiten kann.



Voter Registration Application

Before completing this form, review the General, Application, and State specific instructions.

Are you a citizen of the United States of America? ☐ Yes ☐ No
Will you be 18 years old on or before election day? ☐ Yes ☐ No
If you checked "No" in response to either of these questions, do not complete form.
(Please see state-specific instructions for rules regarding eligibility to register prior to age 18.)

This space for office use only.

1	☐ Mr. ☐ Miss ☐ Mrs. ☐ Ms.	Last Name	First Name	Middle Name(s)	☐ Jr ☐ II ☐ Sr ☐ III ☐ IV
2	Home Address	Apt. or Lot #	City/Town	State	Zip Code
3	Address Where You Get Your Mail If Different From Above		City/Town	State	Zip Code

4	Date of Birth Month Day Year	5	Telephone Number (optional)	6	ID Number - (See item 6 in the instructions for your state)
7	Choice of Party (see item 7 in the instructions for your State)	8	Race or Ethnic Group (see item 8 in the instructions for your State)		

9 I have reviewed my state's instructions and I swear/affirm that:
- I am a United States citizen
- I meet the eligibility requirements of my state and subscribe to any oath required.
- The information I have provided is true to the best of my knowledge under penalty of perjury. If I have provided false information, I may be fined, imprisoned, or (if not a U.S. citizen) deported from or refused entry to the United States.

Please sign full name (or put mark) ▲

Date: ___/___/___ Month Day Year

If you are registering to vote for the first time: please refer to the application instructions for information on submitting copies of valid identification documents with this form.

Please fill out the sections below if they apply to you.

If this application is for a **change of name**, what was your name before you changed it?

A	☐ Mr. ☐ Miss ☐ Mrs. ☐ Ms.	Last Name	First Name	Middle Name(s)	☐ Jr ☐ II ☐ Sr ☐ III ☐ IV

If you were **registered before but this is the first time you are registering from the address in Box 2**, what was your address where you were registered before?

B	Street (or route and box number)	Apt. or Lot #	City/Town/County	State	Zip Code

If you live in a rural area but do not have a street number, or if you have no address, please show on the map where you live.

C
- Write in the names of the crossroads (or streets) nearest to where you live.
- Draw an X to show where you live.
- Use a dot to show any schools, churches, stores, or other landmarks near where you live, and write the name of the landmark.

Example: ● Grocery Store; Route #2; Woodchuck Road; Public School ●; X

NORTH ↑

If the applicant is unable to sign, who helped the applicant fill out this application? Give name, address and phone number (phone number optional).

D

Mail this application to the address provided for your State.

Abbildung 5 Formular zur Registrierung für eine Wahl (United States Government 2006)

Abhängig davon, in welchem Bundesstaat eine Person lebt, müssen mehr oder weniger Felder ausgefüllt werden. Auf einige Unterschiede soll nun exemplarisch eingegangen werden. Im Staat Texas wird keine Angabe über die ethnische Zugehörigkeit gemacht, in Tennessee ist sie optional und in South Carolina verpflichtend. Dort kann eine Nichtangabe der Ethnie zur Ablehnung des Antrags führen. Unterschiedlich ist auch, wann die Registrierung abgeschlossen sein muss, um an der Wahl teilnehmen zu dürfen. In Michigan ist die Registrierung bis spätestens 30 Tage vor der Wahl nötig, in South Dakota sind es 15 Tage und in Vermont muss die Registrierung spätestens am letzten Mittwoch vor der Wahl vorliegen. Ein weiterer wichtiger Unterschied besteht in der Wahl einer Partei. In Alabama ist es nicht notwendig, eine Partei anzugeben, um an der Vorwahl teilnehmen zu können. Anders ist es in Colorado, wo für die Teilnahme an der Vorwahl bei der Registrierung eine Partei ausgewählt werden muss. Schließlich gibt es für die Wahlberechtigung zwischen den Staaten Unterschiede. Diese sind abhängig von begangenen Straftaten oder einer erklärten Unzurechnungsfähigkeit einer Person. (United States Government 2006)

Die Wählerregister sind öffentlich verfügbar, aber ein Staat kann Restriktionen bezüglich ihrer Verwendung festlegen. Alaska hat beispielsweise keinerlei Restriktionen für die Nutzung festgelegt. Verhältnismäßig streng ist die Verwendung der Daten in Kalifornien geregelt. Dort dürfen die Wählerregister nur für politische Zwecke genutzt werden, was eine kommerzielle Verwendung ausschließt. Außerdem dürfen die Informationen nicht von Personen außerhalb der USA verwendet werden. Dies bedeutet für die politische Nutzung der Daten für Parteien in den USA keinerlei Beschränkungen, da diese als nicht kommerziell gelten. (NationBuilder 2016)

3.4. Prädiktive Scores und Microtargeting

Das Wähler-Microtargeting stellt eine neue Form des politischen Direktmarketings dar. Es geht darum, individuelle Wähler zielgerichtet zu adressieren. Dies wird durch das Anwenden von Vorhersagemodellen auf Daten ermöglicht, die über einzelne Personen gesammelt werden. Durch Microtargeting kann umgangen werden, viel Geld für das Senden von Inhalten an ein weitgehend undifferenziertes Massenpublikum auszugeben. Dieses ineffektive Broadcasting weicht einem Narrowcasting, einem Kontaktieren von zuvor bestimmten Zielgruppen. Basierend auf den Ergebnissen von Vorhersagemodellen ist es möglich, einzelnen Personen maßgeschneiderte Nachrichten zu senden, die die für eine Person wichtigen Themen ansprechen. Denkbar ist, einer Person, die sich für Umweltschutz einsetzt, eine E-Mail zu senden, die die Leistungen und die Ziele des Kandidaten in diesem Bereich beinhaltet. Am Ende kann die Person gebeten werden, den Kandidaten zu wählen, damit die Aufrechterhaltung oder Verbesserung des Umweltschutzes gesichert werden kann. Eine Analyse ergab zum Beispiel, dass sich viele potentielle Obama-Wähler auf der sozialen Nachrichtenseite Reddit aufhielten. Daher wurde versucht, die Aufmerksamkeit dieser Personen für Barack Obama zu gewinnen. Um dies zu tun, wurde eine Aktion durchgeführt, bei der die Nutzer der Webseite eine halbe Stunde lang Fragen an Barack Obama stellen konnten, die dieser dort beantwortete. (reddit 2012, Rubinstein 2014)

Die prädiktive Modellierung folgt mehreren Schritten. Im ersten Schritt stellt ein analytisches Team eine Testgruppe basierend auf Informationen in einem Wählerregister zusammen. Zu diesen Daten werden Verbraucherdaten hinzugefügt, die Informationen über sozioökonomische Aspekte geben. Außerdem werden aktuelle Daten darüber hinzugefügt wie Personen auf bisherige Interaktionen reagiert haben. Diese geben gegebenenfalls Aufschluss über die Parteineigung, Sympathien für einen Kandidaten und ihre Interessen. Im zweiten Schritt werden analytische Algorithmen auf diese Daten angewendet, um Korrelationen und Muster zu finden, die die persönlichen Charakteristiken der Gruppenmitglieder mit ihren politischen Einstellungen verbinden. Nach der Feststellung der Muster baut das analytische Team ein Modell, das voraussagt, wie sich die Wähler außerhalb der Testgruppe wahrscheinlich verhalten werden. Dieses Modell wird dann auf eine größere Wählermenge angewendet. (Rubinstein 2014)

Mit den Modellen können für eine Person verschiedene Scores errechnet werden. Als Scores existieren Verhaltens-Scores, Unterstützungs-Scores und Antwort-Scores. Verhaltens-Scores nutzen früheres Verhalten und demographische Informationen, um Wahrscheinlichkeiten zu berechnen, mit denen Bürger bestimmte politische Aktivitäten ausführen. Diese Aktivitäten sind das Wählen am Wahltag, das Geben von Spenden und die Bereitschaft zur Freiwilligenarbeit. Die Wahrscheinlichkeit der Wahlteilnahme wird auch als Turnout-Score bezeichnet. Unterstützungs-Scores sagen die politischen Präferenzen eines Bürgers vorher. Die Antwort-Scores sagen voraus, wie Bürger auf bestimmte Aktionen der Kampagne reagieren werden. Die Voraussage von Reaktionen gestaltet sich aber schwierig. Als Grundlage können randomisierte Experimente dienen, die die durchschnittliche Reaktion auf eine Aktion messen. Das Ergebnis der Experimente dient der Modellierung der wahrscheinlichen Reaktionen auf Aktionen. (Nickerson und Rogers 2014)

Die Kontaktaufnahme zu Wahlberechtigten kostet Zeit und Geld. Diese Ressourcen werden durch die Verwendung von Scores klüger verwendet, was durch ein Beispiel verdeutlicht werden soll. Ein Freiwilligenteam hat den Auftrag in einer Tür-zu-Tür-Aktion die Hausbewohner in einem Stadtviertel zu überzeugen, an der Wahl teilzunehmen und sie vom eigenen Kandidaten zu überzeugen. Das Freiwilligenteam schafft es, in einer Stunde fünf Haushalte zu besuchen. Die Überzeugungsrate der Haushalte beträgt im Schnitt 10 Prozent. Nachdem ein Team acht Stunden gearbeitet und 40 Haushalte besucht hat, konnten insgesamt vier Haushalte überzeugt werden. Die Anzahl an überzeugten Haushalten pro Stunde kann durch die Ergebnisse aus den Scores erhöht werden. Dafür muss zunächst festgestellt werden, für welche Haushalte sich ein Besuch lohnt und für

welche nicht. Dazu können der Unterstützungs-Score und der Turnout-Score hinzugezogen werden. Der Unterstützungs-Score gibt an, wie wahrscheinlich eine Person den eigenen Kandidaten wählen wird. Der Turnout-Score steht für die Wahrscheinlichkeit, mit der eine Person zur Wahl geht. Besuche sind für zwei Arten von Haushalten sinnvoll: Zum einen sind dies Haushalte mit Personen, die einen hohen Unterstützungs-Score bei gleichzeitig niedrigem oder mittlerem Turnout-Score haben. Diese Haushalte bevorzugen den eigenen Kandidaten, haben aber eine geringe Wahlwahrscheinlichkeit. Bei diesen Haushalten wird durch das Kontaktieren versucht, die Wahlwahrscheinlichkeit zu erhöhen. Sinnvoll ist auch der Besuch von Personen, die wahrscheinlich zur Wahl gehen werden, aber noch unentschlossen sind, wen sie wählen werden. Ein Besuch bei Haushalten mit einem sehr niedrigen Unterstützungs-Score ist ein Besuch hingegen nicht sinnvoll, da diese den gegnerischen Kandidaten bevorzugen. Ebenso müssen Haushalte mit einem hohen Unterstützungs- und Turnout-Score nicht kontaktiert werden, da diese ohnehin den eigenen Kandidaten wählen werden. Wenn die Freiwilligenteams also die Haushalte gemäß den Scores besuchen, kann eine größere Anzahl an Haushalten in der gleichen Zeit überzeugt werden.

Ein ausschlaggebender Punkt für den Erfolg der Kampagne bestand in der Integration aller Daten. Zu diesem Zweck wurde ein Programm namens Narwhal entwickelt. Es führte die in verschiedenen Bereichen gesammelten Daten in einer einzigen Datenbank zusammen, die nach dem Ende der Wahl im Jahr 2012 größer als 50 Terrabyte war (Nickerson und Rogers 2014). So konnten die Informationen einer Person über ihr Online-Verhalten, ihre Daten aus dem Wählerverzeichnis und die von Drittanbietern zugekauften Informationen mit den Informationen über die Reaktion der Person bei einem Kontakt durch das Wahlkampfteam verbunden werden. Das Operieren auf dieser integrierten Datenbasis erhöhte die Qualität der berechneten Scores.

3.5. Web und soziale Medien

Beim Wahlkampf des Kandidaten Barack Obama waren eine Online-Plattform namens mybarackobama.com und soziale Medien von Bedeutung. Die Online-Plattform diente zur Koordination der Freiwilligenarbeit. Freiwillige Wahlkampfhelfer konnten sich dort registrieren, ein Profil erstellen, sich vernetzen und ihre Aktivitäten koordinieren. Eine zentrale Aktivität, die von Freiwilligen durchgeführt wurde, war der Haustürwahlkampf. Dieser ist Bestandteil der sogenannten Graswurzelbewegung, also einem Wahlkampf der durch einzelne Personen von unten geführt wird. Sie zogen von Tür zu Tür und leisteten durch ein Gespräch mit den Hausbewohnern auf Basis eines vorgefertigten Gespräch-Skripts Überzeugungsarbeit für ihren Kandidaten. Die Wirksamkeit des Gespräch-Skripts wurde zuvor experimentell getestet. Die Freiwilligen waren mit mobilen Geräten wie Mobiltelefonen und Tablets unterwegs, in die sie Angaben über den Verlauf und das Ergebnis der einzelnen Gesprächen machten. Diese wurden dann an die integrierte Datenbank gesendet.

Im Wahlkampf spielte das soziale Netzwerk Facebook eine wichtige Rolle. Die Registrierung auf mybarackobama.com war neben dem Anlegen eines neuen Nutzerkontos auch mit dem Facebook-Account möglich. Zudem wurde eine Facebook-Anwendung mit dem Namen OFA, Obama for America, entwickelt. Die Facebook-Anwendung ermöglichte es, auf verschiedene Daten von Facebook-Nutzern zuzugreifen. Der Zugriff ist aber nur durch eine vorherige Zustimmung des Nutzers möglich. Das soziale Netzwerk wurde für das sogenannte „Targeted Sharing“ genutzt, bei dem das Ziel war, Kontakte von Unterstützern von Barack Obama zu überzeugen. Die Anwendung griff dazu auf die Freundeslisten der Nutzer zu. Die Personen aus den Freundeslisten wurden von den Analysten daraufhin mit den dem Wahlkampfteam bekannten Personen verglichen und verbunden. Auf diese Weise konnten den Nutzerprofilen im sozialen Netzwerk die der Partei vorliegenden Personen mit den für sie errechneten Kennzahlen zugeordnet werden. Für jede Person, die in beiden Listen vorkam und für die die Kennzahlen eine Kontaktaufnahme nahelegten, wurden die optimale Nachricht ausgewählt, die dann an die Person gesendet wurde. Auf diese Weise kontaktierten über eine Millionen Obama-Unterstützer insgesamt etwa fünf Millionen Menschen. (Rubinstein 2014)

Soziale Medien können jedoch nicht nur dazu genutzt werden, damit Freiwillige auf Stimmenfang unter ihren Freunden gehen. Aus dem Verhalten von Nutzern in sozialen Netzwerken können viele Informationen über die jeweiligen Personen abgeleitet werden. In (Kosinski, Stillwell und Graepel 2013) wurden die Facebook-Likes von 58000 Nutzern analysiert. Die Like-Funktion ermöglicht Facebook-Nutzern, ihre positive Einstellung zu Online-Inhalten auszudrücken. Darunter fallen die von Facebook-Kontakten verfassten Statusnachrichten und hochgeladenen Fotos und Facebook-Seiten von Unternehmen, Restaurants, Webseiten, Medien und bekannten Persönlichkeiten. Die Like-Anzahl der Teilnehmer hatte einen Durchschnittswert von 170. Aus den Likes konnte zu 85% bestimmt werden, ob eine Person Demokrat oder Republikaner ist. Die höchste erreichte Genauigkeit war 95% und betraf die Unterscheidung zwischen hellhäutigen und afroamerikanischen Nutzern. Fast so gut funktioniert die Vorhersage mit 93% für das Geschlecht. Ob ein Nutzer hetero- oder homosexuell ist, wurde für Männer mit einer Wahrscheinlichkeit von 88% und für Frauen mit einer Wahrscheinlichkeit von 75% ermittelt. Die Unterscheidung zwischen Christen und Muslimen war in 82% der Fälle korrekt. Außerdem wurde vorhergesagt, ob ein Nutzer Zigaretten raucht (73%) oder Alkohol trinkt (70%). Wird die Tatsache berücksichtigt, dass den Wahlkampfteams umfassende Daten über die einzelnen Wähler zur Verfügung standen, zeigt sich die potentiell hohe Aussagekraft, die durch die Analyse von Daten erreicht werden können.

Auch unabhängig von sozialen Medien können Dritte im Web Informationen über eine Person sammeln, die dann zum Beispiel zum Schalten passender Werbung verwendet werden. Eine Methode, um das Webverhalten eines Nutzers zu tracken, ist die Verwendung von Cookies. Cookies erlauben einem Webserver eine kleine Menge an Daten auf dem Computer des Besuchers einer Webseite zu speichern, die auf Anfrage an den Webserver zurückgesendet werden. Mit dieser Information können Nutzer mittels Third-Party-Cookies wiedererkannt werden und es kann ein Profil über den Nutzer erstellt werden. Das Tracking des Surfverhaltes durch Cookies ist vielen Menschen bewusst und Cookies können ohne viel Aufwand gelöscht werden. Eine alternative Methode bietet die Geräterkennung durch einen Fingerabdruck. Zur Erstellung dieses Fingerabdrucks werden unter anderem die Informationen genutzt, ob der Browser Flash erlaubt, wie die Do-not-track-Option eingestellt ist, welche Schriftarten auf dem Computer des Nutzers installiert sind und die Bildschirmauflösung des Geräts. (Nikiforakis, Kapravelos, Joosen et al. 2013)

3.6. Experimente

Um die Verwendung von Experimenten auf das Wählerverhalten zu verstehen, ist es dienlich, die Perspektive von (Rogers, Fow und Gerber 2013) einzunehmen. Diese beschäftigen sich mit der Frage, warum Bürger überhaupt wählen. Sie distanzieren sich dabei von der vorherrschenden Ansicht in der Politikwissenschaft und Ökonomie. Diese geht davon aus, dass Wählen eine quasi-rationale Entscheidung von eigennützig handelnden Individuen darstellt. Bei der Entscheidung wird dabei der Aufwand, der der Wahl vorausgeht, abgewogen mit der erwarteten Wahrscheinlichkeit, dass ihre Stimme eine Verbesserung des Wahlausgangs bringt und wie groß diese Verbesserung sein wird. Die Autoren hingegen sehen Wählen als soziales Verhalten, mit dem sich eine Person ausdrücken kann. Das Verhalten wird dabei von Ereignissen beeinflusst, die vor und nach dem Moment der tatsächlichen Wahl eintreten. Mit Experimenten wird versucht herauszufinden, wie die vor der Wahl eintretenden Ereignisse zu gestalten sind, um das gewünschte Ergebnis, also möglichst viele Stimmen für den eigenen Kandidaten, zu erreichen.

Im Wählerverzeichnis ist sichtbar, an welchen der letzten vier Wahlen eine Person teilgenommen hat. Die Person muss dafür aber im Wählerverzeichnis eingetragen sein. Bei Personen, die bei den letzten vier Wahlen nicht teilgenommen haben, kann experimentell ermittelt werden, welchen Erfolg verschiedene Mobilisierungsmaßnahmen haben. Dafür wird für jede alternative Mobilisierungsmaßnahme eine Gruppe aus einer Menge an zufällig ausgewählten Personen gebildet. Zusätzlich werden einige zufällig ausgewählte Personen einer Kontrollgruppe zugeordnet, bei der keine

Maßnahme angewendet wird. Nach Anwendung der Maßnahmen kann nach der Wahl anhand des Wählerregisters festgestellt werden, ob eine Versuchsperson an der Wahl teilgenommen hat oder nicht. Aus dem Ergebnis kann ermittelt werden, welchen Effekt die einzelnen Maßnahmen hatten. Maßnahmen, die einen Erfolg bringen, können dann künftig in der Praxis verwendet werden. Studien dieser Art werden als randomisierte Feldexperimente bezeichnet.

Einige der in diesem Bereich durchgeführten und veröffentlichten Experimente sollen nun kurz vorgestellt werden. Ein wichtiges Experiment stammt von (Gerber und Green 2000). Sie erforschten den Erfolg von Nachrichten, die eine Person zur Wahl bewegen sollten. Es gab etwa 30.000 Testpersonen. Die Nachrichten wurden entweder durch eine Person, postalische Zusendung oder durch einen Telefonanruf überbracht. Die Wahlbeteiligung stieg am meisten bei den Personen, denen die Nachricht von einer Person überbracht wurde. Die Sendung per Post hatte einen geringen positiven Effekt und Telefonanrufe zeigten keine Wirkung. Dies zeigt die Wichtigkeit von Personenkontakt, der von den drei betrachteten Kontaktmöglichkeiten aber die aufwendigste ist.

In einem anderen Experiment wurde den Einfluss von sozialem Druck auf die Wahlteilnahme untersucht. Das Experiment wurde an circa 180.000 im staatlichen Wählerverzeichnis eingetragenen Haushalten durchgeführt. Getestet wurde die Wirkung von vier verschiedenen Nachrichten auf die Wahlteilnahme. Die erste Nachricht erinnerte den Haushalt daran, dass die Wahlteilnahme eine bürgerliche Pflicht ist und endete mit der Aufforderung, an der Wahl teilzunehmen. Die übrigen drei Nachrichten bestanden aus dem gleichen Text und jeweils einem anderen Zusatz. Die zweite Nachricht teilte dem Haushalt mit, dass er Teil einer Studie zum Wählerverhalten ist. Die dritte Nachricht führte die Wahlteilnahme der Haushaltsmitglieder an den beiden vorhergehenden Wahlen auf und gab an, nach der Wahl erneut einen Brief mit der Wahlteilnahme der Haushaltsmitglieder zu senden. Die vierte Nachricht funktionierte wie die dritte Nachricht, mit dem Unterschied, dass darüber hinaus die Wahlbeteiligung aller Nachbarn aufgeführt war. Es zeigte sich, dass die Nachrichten aufsteigend von eins bis vier eine jeweils größere Wirkung hatten. Die höchste Wahlbeteiligung hatten also die Haushalte, die eine Veröffentlichung ihrer Wahlteilnahme in der Nachbarschaft erwarteten. Im Vergleich zu einer Kontrollgruppe, die keine Nachricht erhalten hatte, lag die Wahlbeteiligung der Haushalte, die die vierte Nachricht erhalten hatten, um 8,1 Prozent höher. Das Ergebnis legt nahe, dass das Aufbauen von sozialem Druck einen positiven Effekt auf die Wahlbeteiligung hat. (Gerber, Green und Larimer 2008)

Ein Feldexperiment während der Präsidentschaftswahl im Jahr 2008 mit knapp 300.000 Personen kam zu dem Ergebnis, dass Personen eher an der Wahl teilnehmen, wenn nicht nur gefragt wird, ob sie wählen werden, sondern auch zu welcher Uhrzeit und wie sie planen, zum Wahllokal zu gelangen. Durch die Tatsache, dass sich die Befragten eine Umsetzung der Wahlhandlung konkret vorstellen mussten, stieg die Wahlbeteiligung um 4,1 Prozent. Die bloße Frage, ob eine Person zu Wahl gehen wird, hatte hingegen nur einen vernachlässigbaren positiven Effekt. (Nickerson und Rogers 2010)

Ein weiteres interessantes Experiment stützt sich auf die Online-Plattform Facebook. Es diente zur Erforschung der Wählerüberzeugung, und umfasst die enorm hohe Anzahl von 61 Millionen Testpersonen. Es fand im Rahmen der Kongresswahl in den USA im Jahr 2010 statt und hatte das Wahlverhalten von Facebook-Nutzern zum Gegenstand. Die Testpersonen wurden in drei Gruppen eingeteilt. Der ersten Gruppe wurde eine informationelle Nachricht zugesendet. Die zweite Gruppe erhielt eine informationelle Nachricht mit einer zusätzlichen sozialen Komponente. Die dritte Gruppe bildete die Kontrollgruppe, die keine Nachricht erhielt. Die nachfolgende Abbildung zeigt die soziale Nachricht. Die informationelle Nachricht war ausgenommen vom fehlenden unteren Teil, der die wählenden Freunde anzeigte, identisch.

Abbildung 6 Soziale Nachricht auf Facebook zur Kongresswahl 2010

Nutzer, die die soziale Nachricht erhalten hatten, klickten mit einer höheren Wahrscheinlichkeit von 2,08% auf den Button mit der Bezeichnung „I Voted". Insgesamt taten dies 20,04% beziehungsweise 17,96%. Die Aussage, dass eine Testperson gewählt hatte, wurde mit dem Eintrag im Wählerverzeichnis abgeglichen. Es zeigte sich, dass die Wahlbeteiligung der Gruppe mit der informationellen Nachricht und der Kontrollgruppe gleich hoch war. Die Wahlbeteiligung der Gruppe mit der sozialen Nachricht war um 0,39% höher. (Bond et al. 2012)

Ein Anstieg von 0,39% mag nicht viel erscheinen. Wird aber berücksichtigt, dass es viele Millionen Facebook-Nutzer gibt, macht eine soziale Nachricht einen Unterschied. Dieses Experiment zeigt damit auch die Macht zur politischen Einflussnahme von sozialen Netzwerken. In der Theorie könnte Facebook die Wahlbeteiligung erhöhende Nachricht nur denjenigen Nutzern zukommen lassen, die mit hoher Wahrscheinlichkeit den vom Unternehmen präferierten Kandidaten unterstützen.

Die Ausführungen zeigen, dass die bisherigen Experimente den Parteien gute Hinweise für die richtigen Kommunikationsmittel mit den Bürgern geben. Neben den veröffentlichten Ergebnissen existieren Experimente, die innerhalb der Wahlkampfteams durchgeführt werden. Bei diesen wird zum Beispiel geprüft, welche Formulierungen bei einem Spendenaufruf oder bei der Überzeugung vom eigenen Kandidaten für verschiedene Personengruppen erfolgreich sind. Da gewonnene Erkenntnisse aus den Experimenten einen Wissensvorteil gegenüber Dritten bedeuteten, werden diese Ergebnisse nicht bekannt gemacht.

3.7. Entstehung von spezialisierten Unternehmen

In den USA sind kommerzielle Datenhändler entstanden, die auch als Datenbroker bezeichnet werden. Diese sammeln Daten aus verschiedenen Quellen. Jede Quelle für sich bietet nur einen kleinen, abgetrennten Einblick in das Leben einer Person. Die Kombination ermöglicht aber eine detaillierte und umfassende Sicht auf das Leben einer Person. Pro Person sind bis zu 3000 Datenpunkte vorhanden (Tufekci 2014). Beispielsweise werden die Daten eines Wählerverzeichnisses mit den Daten von Facebook verknüpft. Die Verknüpfung der verschiedenen Datenquellen erfolgt anhand von identifizierenden Merkmalen. Identifizierende Merkmale sind der Name und die Adresse, Telefonnummern, E-Mailadressen und IP-Adressen. Der Vorteil von Datenbrokern besteht darin, dass sie auch Daten zu nicht-registrierten Wahlberechtigten gespeichert haben. Diese wären bei der alleinigen Verwendung von Wählerverzeichnissen nicht abgedeckt.

Auf einige bedeutende Datenbroker wird nun kurz eingegangen. Catalist ist eine Datenbank, die Information über mehr als 240 Millionen einzelne wahlberechtigte Bürger in den USA beinhaltet. Diese teilen sich in 185 Millionen registrierte und 55 Millionen unregistrierte Bürger auf. Sie verbindet Daten aus Quellen, die für die Wahl zuständig sind mit Verbraucherdaten und Daten aus dem Zensus. Die Datenbank hat hunderte Attribute und umfasst Informationen zu Haushalt, Kauf- und Investmentverhalten, Spenden, Beruf, Freizeit und Engagement in verschiedenen Gruppen. (Catalist 2016)

Das Unternehmen Cambridge Analytica charakterisiert Personen basierend auf den fünf Persönlichkeitsdimensionen des in der Psychologie bekannten Big-Five-Modells. Die Dimensionen sind Neurotizismus, Extraversion, Offenheit für Erfahrungen, Verträglichkeit und Gewissenhaftigkeit. Neben der geographischen und demographischen Sicht wird also zusätzlich eine psychologische Sicht auf die Wähler angeboten, die andere Unternehmen nicht bieten. (Cambridge Analytica 2016)

Aristotle bietet Technologie, Daten und Strategien für Kampagnen und Öffentlichkeitsarbeit an. Es bietet Zugang zu einer angepassten Wählerliste aus circa 190 Millionen registrierten Wählern. Insgesamt werden etwa 205 Millionen Konsumenten erfasst. Diese Liste ist angereichert mit Telefonnummern, demographischen Daten und Informationen zum Lebensstil. Außerdem hat es Daten zu über 80 Millionen Spenden, die in der Vergangenheit getätigt wurden. Zusätzlich zur Wählerliste wird ein sogenannter Campaign Manager angeboten, mit dem das Spendensammeln und Compliance-Konformität gesteuert werden kann. (Aritstotle 2016)

Neben Aristotle existieren noch andere Unternehmen, die Plattformen zur Unterstützung einer Wahlkampagne anbieten. Beispiele dafür sind Nation Builder und NGP Van. Die Software von Nation Builder kann durch ein Abonnement genutzt werden. Es wird angegeben, dass ein kostenloser Zugang zu allen Wählern in einem Bezirk geschaffen wird. Daneben wird ein vollständig integriertes Management von Wählern, Freiwilligen und Geldgebern genannt und es ist die Rede von einem fortgeschrittenen Targeting. Eingebaut in die Software ist ein A/B-Testen von E-Mails, das die Effektivität von verschiedenen Versionen misst. Zusätzlich können mehrere mobile Anwendungen genutzt werden, die zum Beispiel den Erfolg von persönlichen Wählerkontakten oder die Aktivitäten von Kampagnenunterstützern erfassen. Ziel ist das Anbieten einer zentralen Plattform, in der die Daten aus den verschiedenen Aktivitäten der Wahlkampagne integriert werden und auf dem aktuellen Stand sind. NGP Van bietet seine Dienste nur Politikern der Demokraten und nicht den Republikanern an. Es unterstützt die Spendensammlung durch ein Kontakt- und Geldgebermanagement und die Segmentierung von Spendergruppen. Durch das System wird auch die rechtlich notwendige Berichterstattung über die Spenden unterstützt. Daneben ist auch das Management der Freiwilligenarbeit möglich. Ebenso wie bei NationBuilder ist eine mobile Anwendung zum Erfassen von Wählerkontakten verfügbar. Zusätzlich werden Email-Templates, eine Möglichkeit, online Spenden zu sammeln und Telefondienste angeboten. Diese betreffen automatische Anrufe durch einen Computer, mit denen Kontakt zu potentiellen Freiwilligen, Wählern und Geldgebern hergestellt wird. Sowohl NGP Van als auch NationBuilder bieten ebenso wie die anderen Unternehmen eigene Wählerverzeichnisse an. (NationBuilder 2016, NGP VAN 2016)

Zusätzlich zu allgemeinen Anbietern von Konsumentendaten, die neben Wirtschaftsunternehmen auch von Politikern genutzt werden können, hat sich ein Markt für den informationsbasierten Wahlkampf entwickelt. Auf diesen werden die Methoden und Begrifflichkeiten der Privatwirtschaft übertragen. An die Stelle von Kunden-Microtargeting tritt Wähler-Microtargeting und aus dem Customer-Relationship-Management wird das Voter-Relationship-Managment. Der Umgang mit Wählerdaten und -beziehungen wird zu einer Managementaufgabe, die durch neu entwickelte Informationssysteme unterstützt wird.

4. Übertragbarkeit auf den deutschen Bundestagswahlkampf

In diesem Kapitel wird die Übertragbarkeit der zuvor beschriebenen Vorgehensweise im Wahlkampf auf die deutsche Bundestagswahl überprüft. Dabei werden, neben grundlegenden wahlbezogenen Unterschieden, Aspekte des Datenschutzes, der Parteifinanzierung und die Verfügbarkeit von Datenquellen für den Wahlkampf behandelt.

4.1. Grundlegende Rahmenbedingungen zu Wahlen in Deutschland

Zunächst wird auf grundlegende Unterschiede zwischen der Bundestagswahl und der Präsidentschaftswahl eingegangen. In Deutschland gilt ein personalisiertes Verhältniswahlrecht mit Sperrklausel. Es finden nicht wie in den USA Vorwahlen statt, in denen ein Kandidat für die Präsidentschaft gewählt wird. Der Bundeskanzler wird stattdessen vom deutschen Bundestag gewählt. Die beiden größten Parteien CDU und SPD bestimmen vor der Bundestagswahl bereits einen Kanzlerkandidaten, mit dem sie ihren Wahlkampf betreiben und der bei einem Sieg der Partei als Bundeskanzler zur Wahl gestellt wird. Im Gegensatz zu den USA geht es nicht darum, mit dem Winner-takes-it-all-Prinzip die Mehrheit der Stimmen in den jeweiligen Bundesländern zu erhalten, sondern das Stimmverhältnis aller Stimmen ist entscheidend. Auch die Stimmen, die nicht der Mehrheitsmeinung des Bundeslandes entsprechen, fallen ins Gewicht. Somit haben auch kleine Parteien eine große Chance, in den Bundestag einzuziehen. In den USA gibt es alle zwei Jahre Kongresswahlen und alle vier Jahre wird der Präsident gewählt. In Deutschland werden im Zuge der Bundestagswahl der Bundestag und der Bundeskanzler bestimmt, der wie der US-Präsident der Regierungschef des Landes ist. Der Bundeskanzler nicht wie der Präsident in den USA das Staatsoberhaupt. Das deutsche Staatsoberhaupt ist der Bundespräsident. Dieser wird nicht bei der Bundestagswahl, sondern alle fünf Jahre von der Bundesversammlung gewählt. Anders als in den USA wird der Wahlkampf nicht für jeden Kandidaten, wovon es vor den Vorwahlen mehr als einen pro Partei gibt, sondern für eine Partei geführt. Außerdem wird im Zuge der Bundestagswahl nicht nur der Regierungschef, sondern auch die gesetzgebende Gewalt bestimmt. Die Wahlbeteiligung in den USA ist im Vergleich zu Deutschland niedrig. Die prozentuale Wahlbeteiligung an den Bundestags- und Präsidentschaftswahlen wird in der folgenden Abbildung gezeigt.

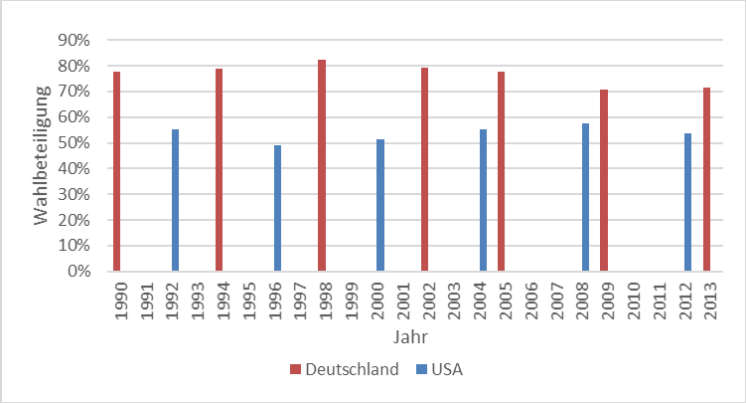


Abbildung 7 Wahlbeteiligung bei den Bundestags- und Präsidentschaftswahlen (Statista 2016)

Bei der letzten Bundestagswahl lag die Wahlbeteiligung in Deutschland bei 71,5%. In den USA wählten 2012 bei der Präsidentschaftswahl 53,6% der Wahlberechtigten. Fast jeder zweite wahlberechtigte US-Amerikaner ging also nicht zur Wahl. Bei den Kongresswahlen in den Jahren, in denen kein Präsident gewählt wird, ist die Wahlbeteiligung noch niedriger. Dieser Prozentsatz ist im Vergleich zum deutschen niedrig. Doch auch bei der deutschen Wahlbeteiligung besteht noch ausreichend Raum, die richtigen Wähler zu mobilisieren. Dies ist auch dadurch sichtbar, dass sich die

Wahlbeteiligung in den fünf Bundestagswahlen bis zum Jahr 2005 auf einem Niveau von circa 80% bewegte und seit 2009 auf einen Wert von etwa 70% abgesunken ist.

Die Präsidentschaftswahl zu gewinnen, ist aufgrund des Wahlsystems und der extremen Dominanz der zwei großen Parteien der Demokraten und Republikaner für andere Parteien in den USA wie die Grüne Partei, die Verfassungspartei oder die libertäre Partei nahezu unmöglich. In Deutschland sind derzeit vier Parteien im Bundestag vertreten. Die in den USA für viele Bürger berechneten Unterstützungs-Scores gaben die Wahrscheinlichkeit an, mit der eine Person entweder den demokratischen oder republikanischen Kandidaten wählen wird. Das heißt, die Entscheidung eines Wählers beschränkte sich in den meisten Fällen auf diese beiden Parteien. Bei der deutschen Bundestagswahl kann ein Wähler zwischen mehr als 30 Parteien auswählen, wobei voraussichtlich sechs Parteien in den Bundestag einziehen werden. Diese größere Anzahl an Parteien erschwert die Berechnung von Unterstützungs-Scores, da mehr als zwei Dimensionen berücksichtigt werden müssen. Es gestaltet auch die Überzeugung eines Wählers von der eigenen Partei schwieriger. In den USA bedeutet eine Überzeugung des Wählers von den Schwächen der anderen Partei automatisch, dass, abgesehen von einer Nichtwahl, die eigene Partei die einzige realistische Alternative darstellt. In Deutschland ist dies nicht der Fall. Beim Aufzeigen der Versäumnisse und Fehler einer einzelnen Partei, hat der Wähler immer noch ausreichend andere Optionen zur Stimmvergabe. Dabei bestehen unterschiedliche Distanzen zwischen Parteien hinsichtlich ihrer thematischen Punkte und politischen Positionen. Zwei Parteien können hinsichtlich bestimmter Themen eine geringe und bezüglich anderer Themen eine größere Distanz aufweisen. Es kann für eine Partei leichter sein, Wähler einer näheren Partei als Wähler einer weiter entfernteren Partei zu gewinnen. Dafür müssen die richtigen Themen zur Wähleransprache ausgewählt werden. Für eine Partei ist es lohnenswert, diese thematischen Unterschiede zu analysieren und die Themen herauszuarbeiten, mit denen bestimmte Wählerzielgruppen von der eigenen Partei überzeugt werden können.

4.2. Wählerverzeichnisse

Im Gegensatz zu den USA ist keine Registrierung in einem Wählerverzeichnis notwendig, um an der Wahl teilnehmen zu können. Die Führung eines Wählerverzeichnisses ist Aufgabe der Gemeindebehörden. Die Bestimmungen zum Wählerverzeichnis sind in der Bundeswahlordnung festgelegt. Die Gemeindebehörde legt vor jeder Wahl für jeden allgemeinen Wahlbezirk ein Verzeichnis der Wahlberechtigten mit Vornamen, Familiennamen, Geburtsdatum und Wohnung an. Außerdem gibt es je eine Spalte für Vermerke über die Stimmabgabe und für Bemerkungen. Eingetragen wird, wer bei der Meldebehörde gemeldet ist oder einen unter bestimmten Umständen möglichen Antrag zur Eintragung stellt. Anders als in den USA werden also alle bei einer Gemeinde gemeldeten, wahlberechtigten Personen automatisch von der Gemeinde in einem Wählerverzeichnis eingetragen. Die Informationen im Wählerverzeichnis fallen mit Name, Geburtsdatum und Adresse deutlich geringer aus. In den USA ist es in einigen Bundesstaaten erforderlich, eine Parteizugehörigkeit anzugeben, um an den Vorwahlen teilnehmen zu können. Dadurch ist es möglich, öffentlich einzusehen, für welchen Kandidaten eine Person mit erhöhter Wahrscheinlichkeit stimmen wird. Eine Kombination dieser politischen Neigung mit weiteren Informationen ermöglicht die Verbindung verschiedenster Informationen eines Individuums mit seiner politischen Ausrichtung. Dadurch können Vorhersagemodelle ermittelt werden. In Deutschland wird keine politische Präferenz in den Wählerregistern angegeben, daher ist eine Verbindung der Daten auf die beschriebene Weise nicht möglich. (§ 17 BWahlG , § 14 BWO)

Das Wählerverzeichnis kann eingesehen werden. Die Gemeindebehörde macht bekannt, von wem, zu welchen Zwecken und unter welchen Voraussetzungen dies möglich ist. Für Parteien ist nicht die Einsicht, sondern das Anfertigen eines Auszugs aus dem Verzeichnis interessant, da sie so alle Wahlberechtigten einer Gemeinde kontaktieren können. Das Anfertigen von Auszügen ist für eine Person nur für den Zweck zulässig, in dem es in Zusammenhang mit der Prüfung des Wahlrechts

einzelner, bestimmter Personen steht. Der Auszug darf nur zu diesem Zweck verwendet und unbeteiligten Dritten nicht zugänglich gemacht werden (§ 21 BWO). Das Anfertigen eines Auszugs durch einen Wahlberechtigten zur Unterstützung eines Parteienwahlkampfs ist daher nicht möglich. Zum einen wird, wenn das Vorgehen in den USA als Vorbild genommen wird, ein Auszug nicht nur von einzelnen Personen gemacht. Zum anderen steht der Auszug in diesem Fall nicht im Zusammenhang mit der Prüfung des Wahlrechts einzelner bestimmter Personen.

Amtliche Stellen des Wahlgebiets dürfen Auskünfte aus Wählerverzeichnissen erteilen, wenn sie für den Empfänger im Zusammenhang mit der Wahl erforderlich sind. Ein Erfordernis liegt insbesondere bei Verdacht von Wahlstraftaten, Wahlprüfungsangelegenheiten und bei wahlstatistischen Arbeiten vor (§ 89 BWO). Es muss geprüft werden, ob es sich bei der Nutzung von Wählerdaten zur Optimierung des Wahlkampfes um ein Erfordernis handelt. Die besonders hervorgehobenen Fälle des Verdachts auf Wahlstraftaten und der Wahlprüfungsangelegenheit treffen nicht zu. Ob die Auswertung ein Erfordernis bei wahlstatistischen Arbeiten ist, ist von den genannten Bedingungen noch am zutreffendsten, aber ist höchstwahrscheinlich nicht der Fall. Dass die Wählerdaten-auswertung mit Wählerverzeichnissen ein Erfordernis im Zusammenhang mit der Wahl darstellen, ist daher zu verneinen. Ein Erfordernis wäre für eine Partei höchstens gegeben, wenn einige Parteien Zugriff auf die Daten hätten und andere Parteien nicht. Da die bisherigen Wahlkämpfe auch ohne Auskünfte aus dem Wählerverzeichnis möglich waren, ist ein Erfordernis nicht gegeben. Dementsprechend ist es nicht erlaubt, dass Parteien die Wählerverzeichnisse der Gemeinden für ihren Wahlkampf nutzen können.

Unabhängig vom Wählerverzeichnis sind die Informationen, die von Meldebehörden ausgegeben werden dürfen, im Bundesmeldegesetz geregelt. Bei Auskünften kann zwischen einfachen Melderegisterauskünften und erweiterten Melderegisterauskünften unterschieden werden. Die Auskünfte sind kostenpflichtig. Bei der einfachen Melderegisterauskunft wird Auskunft über den Familiennamen, Vornamen, Doktorgrad, Anschriften und ob die Person noch lebt, gegeben. Die Daten dürfen nicht zum Zwecke der Werbung oder des Adresshandels verwendet werden, außer die betroffene Person hat eingewilligt. Sofern die Daten für gewerbliche Zwecke verwendet werden, dürfen sie nur für den vom Anfragensteller angegebenen Zweck verwendet werden (§ 44 BMG). Eine erweiterte Melderegisterauskunft kann erteilt werden, wenn ein berechtigtes Interesse glaubhaft gemacht wird. Dann wird Auskunft über frühere Namen, Geburtsdatum und Geburtsort, Familienstand, Staatsangehörigkeit, frühere Anschriften, Einzugs- und Auszugsdatum, Name und Anschrift des gesetzlichen Vertreters und des Lebenspartners und gegebenenfalls Daten zum Tod gegeben (§ 45 BMG). Die Zweckbindung und die nötige Einwilligung der Betroffenen machen eine Verwendung der Daten schwierig. Es ist also weder gegeben, dass für erfasste Wahlberechtigte eine Parteipräferenz angegeben ist, noch dürfen die Informationen eingesehen werden und auch die Auskünfte der Melderegister abseits der Wählerverzeichnisse sind eingeschränkt. Die Informationen, die in den USA die Grundlage für die Datenauswertung darstellen, sind demnach in Deutschland nicht vorhanden.

4.3. Datenschutz

In Deutschland gilt das Recht auf informationelle Selbstbestimmung. Dieses wurde vom Bundesverfassungsgericht im Jahr 1983 im sogenannten Volkszählungsurteil festgelegt. Es besagt, dass der Einzelne die Befugnis hat „grundsätzlich selbst zu entscheiden, wann und innerhalb welcher Grenzen persönliche Lebenssachverhalte offenbart werden“ (Bundesverfassungsgericht 1983). Die Bestimmungen zum Datenschutz sind im Bundesdatenschutzgesetz, dem BDSG, festgelegt. Es wurde im Jahr 1990 ausgefertigt und zuletzt im Februar 2015 geändert. Es unterscheidet zwischen der Datenverarbeitung durch öffentliche und nicht-öffentliche Stellen, für die jeweils andere Regelungen gelten. Parteien sind Personenvereinigungen und gehören damit zu den nicht-öffentlichen Stellen (§ 2 PartG, § 2 BDSG). Das Gesetz hebt einige Daten als besondere Arten personenbezogener Daten hervor. Dabei handelt es sich um die rassische und ethnische Herkunft, politische Meinungen und religiöse oder philosophische Überzeugungen. Außerdem fallen die Zugehörigkeit zu einer Gewerkschaft und Informationen zu Gesundheit oder Sexualleben darunter (§ 3 Abs. 9 BDSG). Für diese besonderen Arten personenbezogener Daten existieren Sondervorschriften, die ihre Verarbeitung erschweren. Diese Sondervorschriften beziehen sich auf die Datenerhebung und -speicherung für eigene Geschäftszwecke (§ 28 Abs. 6-9 BDSG), die geschäftsmäßige Datenerhebung und -speicherung zum Zweck der Übermittlung in nicht-anonymisierter und anonymisierter Form (§ 29 Abs. 5 BDSG) und die geschäftsmäßige Datenerhebung und -speicherung für Zwecke der Markt- und Meinungsforschung (§ 30 Abs. 5 BDSG). Beim Umgang mit personenbezogenen Daten müssen die Prinzipien der Datenvermeidung und Datensparsamkeit beachtet werden. Die Erhebung, Verarbeitung und Nutzung personenbezogener Daten und die Gestaltung von Datenverarbeitungssystemen soll so gestaltet sein, dass so wenig personenbezogene Daten wie möglich verwendet werden (§ 3a BDSG). Diese gesetzliche Bestimmung steht dem Prinzip von Big Data deutlich entgegen.

Die Verarbeitung oder Nutzung personenbezogener Daten ist zulässig, soweit es sich um zusammengefasste Daten über Angehörige einer Berufsgruppe handelt. Dabei müssen sich die Daten auf die Zugehörigkeit des Betroffenen zu dieser Personengruppe, seine Berufs-, Branchen- oder Geschäftsbezeichnung, seinen Namen, Titel, akademischen Grad, Anschrift und Geburtsjahr beschränken. Außerdem muss die Verarbeitung der Daten erforderlich sein für Zwecke der Werbung bei steuerbegünstigte Zwecken nach dem Einkommensteuergesetz (§ 28 BDSG). Diese steuerbegünstigten Sonderausgaben sind in (§ 10 EStG) aufgezählt. Diese Regelung ist bedeutend, da unter steuerbegünstigte Zwecke auch Parteispenden fallen. Diese sind bis 1650€ bzw. bei Zusammenveranlagung von Ehegatten bis 3300€ steuerbegünstigt. Gleiches gilt für Mitgliedsbeiträge an Parteien und weitere Zuwendungen. In diesem Fall müssen die Betroffenen, deren Daten verarbeitet werden, keine Einwilligung geben. Diese Ausnahmeregelung wird als Listenprivileg bezeichnet. Das heißt, dass eine Partei ohne Einwilligung der Betroffenen aggregierte Daten über Angehörige einer Berufsgruppe sammeln darf, sofern diese Daten für die Werbung von Spenden oder Mitgliedsbeiträgen verwendet werden. Hinsichtlich des Sammelns von Spenden könnten Personen kontaktiert werden, die aufgrund bestimmter Merkmale potentielle Spendengeber sein könnten.

Für Deutschland geltender Regelungen zum Datenschutz existieren nicht nur auf der nationalen, sondern auch auf der europäischen Ebene. Der Schutz personenbezogener Daten ist in Artikel 9 der Charta der Grundrechte der europäischen Union festgeschrieben. Der Gesetzeswortlaut ist dabei: „Jede Person hat das Recht auf Schutz der sie betreffenden personenbezogenen Daten. Diese Daten dürfen nur nach Treu und Glauben für festgelegte Zwecke und mit Einwilligung der betroffenen Person oder auf einer sonstigen gesetzlich geregelten legitimen Grundlage verarbeitet werden. Jede Person hat das Recht, Auskunft über die sie betreffenden erhobenen Daten zu erhalten und die Berichtigung der Daten zu erwirken. Die Einhaltung dieser Vorschriften wird von einer unabhängigen Stelle überwacht.“ (§ 8 GRCh)

Im deutschen und europäischen Recht wird die Zweckbindung der Datenverarbeitung betont. Daten dürfen nur zu dem Zweck verarbeitet werden, zu dem der Betroffene explizit eingewilligt hat. Falls die Daten zu einem anderen Zweck verwendet werden sollen, ist das Einholen einer Einwilligung des Betroffenen für diesen Zweck erforderlich. Es können also nicht die bereits gesammelten Daten für einen neuen Zweck verwendet werden, für den bei Erhebung der Daten nicht die Einwilligung gegeben wurde. Ein neuer Zweck wäre hier die Verwendung von Daten zum Erstellen prädiktiver Modelle. Auf europäischer Ebene wird die Datenschutz-Grundverordnung ab Mai 2018 voll wirksam sein. Sie schafft einen EU-weit einheitlichen Rahmen zum Datenschutz. Für Unternehmen aus dem EU-Ausland, die in der EU tätig sind, gilt mit der Datenschutz-Grundverordnung ein einheitlicher Standard. Es ist nicht mehr notwendig, sich für jedes Land gesondert in den Datenschutz einzuarbeiten. Sie wird ab 2018 das Bundesdatenschutzgesetz ablösen. In der Datenschutz-Grundverordnung ist ebenso wie im deutschen Bundesdatenschutzgesetz die Einwilligung des Betroffenen und die Zweckbindung verankert (EU-DSGVO). Für die nächste Bundestagswahl im Jahr 2017 hat sie noch keine Bedeutung.

4.4. Parteifinanzierung

Finanzen spielen im Wahlkampf eine bedeutende Rolle. Je mehr finanzielle Mittel einer Partei zur Verfügung stehen, desto umfangreicher kann sie im Wahlkampf agieren. Ein größeres Parteivermögen bedeutet mehr Handlungsspielraum und eine größere Kampagne. Daher ist es sinnvoll zu untersuchen, wie sich Parteien in Deutschland finanzieren können. Im Gegensatz zu den USA, wo die Wahlkampffinanzierung fast ausschließlich auf individuellen Spenden basiert, ist die Finanzierung deutscher Parteien deutlich mehr vom Staat unterstützt. Wie vermögend deutsche Parteien sind und aus welchen Quellen sich das Vermögen zusammensetzt, ist Gegenstand dieses Unterkapitels. Um die Größenordnung des Parteivermögens aufzuzeigen, ist in der nachfolgenden Tabelle das Vermögen der deutschen Parteien in den Jahren 2010 bis 2014 dargestellt. Die Informationen stammen aus den Rechenschaftsberichten, die die Parteien jedes Jahr erstellen müssen. Die Rechenschaftsberichte müssen beim Präsidenten des Bundestags eingereicht werden, der diese dann veröffentlicht. (Deutscher Bundestag 2016)

Partei	Vermögen 2010	Vermögen 2011	Vermögen 2012	Vermögen 2013	Vermögen 2014
SPD	174.756.623	188.907.634	206.984.768	185.388.041	178.995.674
CDU	111.936.338	120.530.150	135.366.179	135.062.211	127.384.761
B90	26.014.336	30.298.191	37.979.611	34.771.885	33.874.962
CSU	28.496.873	33.631.485	38.290.741	31.889.068	22.472.693
Linke	23.590.664	25.473.080	29.482.946	25.780.432	23.584.256
FDP	5.876.074	5.461.460	10.796.978	6.271.535	3.430.447
AfD	-	-	-	4.182.012	2.338.125

Tabelle 2 Vermögen der deutschen Parteien in den Jahren 2010 bis 2014

Das Vermögen der Parteien liegt im ein- bis dreistelligen Millionenbereich. Das höchste Vermögen im Jahr 2014 hatte die SPD mit knapp 179 Millionen Euro. Dies ist mehr als die CDU und CSU zusammen besitzen. Über circa 60 Millionen Euro weniger verfügte die CDU. Mit großem Abstand folgten die Grünen, die CSU und die Linke, deren Vermögen sich jeweils im Wertebereich zwischen 20 und 35 Millionen befand. Nach einem weiteren Abstand folgten FDP und AFD mit circa 3,4 bzw. 2,3 Millionen Euro an Parteivermögen. Die Finanzierung deutscher Partei setzt sich aus der Eigenfinanzierung und einer staatlichen Teilfinanzierung zusammen. Zur Eigenfinanzierung zählen unter anderem Mitgliedsbeiträge, Mandatsbeiträge und Spenden. Die staatliche Teilfinanzierung steht einer Partei zu, sofern sie die nötigen Voraussetzungen dafür erfüllt. Die folgende Tabelle stellt die Einnahmen und Ausgaben der betrachteten Parteien für das Jahr 2014 basierend auf ihren Rechenschaftsberichten dar.

	AfD	B90	CDU	CSU	FDP	Die Linke	SPD
Einnahmen in 1000 Euro							
Mitgliedsbeiträge	2.401	8.795	38.191	9.728	6.270	9.277	49.984
Mandatsträgerbeiträge und ähnliche regelmäßige Beiträge	53	9.150	18.885	3.689	1.827	3.763	24.459
Spenden von natürliche Personen	2.011	4.090	18.011	9.732	5.837	2.253	12.576
Spenden von juristischen Personen	50	657	7.910	3.891	1.966	35	2.532
Einnahmen aus Unternehmenstätigkeit und Beteiligungen	2.584	1	37	0	70	0	2.134
Einnahmen aus sonstigem Vermögen	3	129	2.429	176	481	138	7.707
Einnahmen aus Veranstaltungen, Vertrieb von Druckschriften und Veröffentlichungen und sonstiger mit Einnahmen verbundener Tätigkeit	40	632	12.398	6.638	1.208	244	12.792
Staatliche Mittel	5.411	14.818	47.889	12.697	9.201	10.715	48.649
Sonstige Einnahmen	19	1.396	1.380	219	96	727	994
Summe	12.573	39.669	147.131	46.771	26.957	27.151	161.827
Ausgaben in 1000 Euro							
Personalausgaben	1.021	14.116	42.580	10.764	4.327	10.177	47.441
Sachausgaben des laufenden Geschäftsbetriebs	898	6.264	28.064	6.811	7.849	4.823	27.941
Sachausgaben für allgemeine politische Arbeit	1.349	6.882	31.078	13.955	5.766	5.755	29.689
Sachausgaben für Wahlkämpfe	4.834	12.779	48.567	24.232	10.571	8.512	52.106
Sachausgaben für die Vermögensverwaltung einschließlich sich hieraus ergebender Zinsen	0	305	2.114	0	135	32	10.703
Sonstige Zinsen	7	7	329	274	1.020	4	206
Sonstige Ausgaben	2.619	213	2.078	152	130	45	133
Summe	10.729	40.565	154.809	56.187	29.798	29.348	168.219
Überschuss/Defizit	1.844	- 897	-7.677	-9.416	-2.841	- 2.196	-6.392

Tabelle 3 Einnahmen und Ausgaben der Parteien in Tausend Euro im Jahr 2014

Alle Parteien mit Ausnahme der AfD machten im Jahr 2014 Verlust. Interessant sind die Sachausgaben für Wahlkämpfe. Die höchsten Ausgaben hatte die SPD mit 52 Millionen Euro, gefolgt von der CDU mit 49 Millionen und der CSU mit 24 Millionen. Auffällig ist der hohe Anteil der Einnahmen aus Unternehmenstätigkeit und Beteiligungen bei der AfD, da diese Einnahmeform bei den anderen Parteien relativ niedrig bis sehr niedrig ist. Diese Einnahmen dürften zu einem großen Teil aus dem Handel mit Gold stammen, in Folge dessen eine Anpassung des Parteiengesetzes stattfand. Spenden von natürlichen und juristischen Personen, die in den USA den Hauptteil der Wahlkampffinanzierung

ausmachen, sind in Deutschland von untergeordneter Wichtigkeit, schwanken jedoch von Partei zu Partei. Der Großteil der Einnahmen ist durch Mitgliedsbeiträge, Mandatsbeiträge und die staatlichen Mittel gegeben. Auf einige Einnahmequellen wird im Folgenden näher eingegangen.

Parteien erheben Mitgliedsbeiträge. Diese sind bis zu einem Betrag von 1650 Euro beziehungsweise bis 3300 Euro bei Ehepartnern steuerlich begünstigt. Die Hälfte des Beitrags erhält ein Zahler bei Abgabe der Steuererklärung zurück. Die Höhe der Beiträge ist in der Finanzordnung in der jeweiligen Parteisatzung festgelegt. Die Satzungen der Parteien sind unter (Der Bundeswahlleiter 2016) einsehbar. Die AfD erhebt einen Mindestmitgliedsbeitrag von 120 Euro pro Kalenderjahr. Bei besonderen sozialen Härtefällen ist eine Reduzierung auf 30 Euro pro Kalenderjahr möglich. Empfohlen wird ein Mindestbeitrag von 1% des Jahresnettoeinkommens. Bei den Grünen beträgt der Mindestbetrag 1% vom Nettoeinkommen. Bei besonderen finanziellen Härtefällen ist eine Ausnahme von dieser Regelung möglich. Die CDU erhebt als Mindestbeitrag 6 Euro. Darüber hinaus werden Orientierungsbeiträge abhängig vom Bruttoeinkommen angegeben. So wird bei einem Bruttoeinkommen ab 2500, 4000 und 6000 Euro eine monatliche Zahlung von 15, 25 beziehungsweise 50 Euro empfohlen. Bei der CSU muss jährlich mindestens der Basisbeitrag in Höhe von 62 Euro bezahlt werden. Bei einem Bruttoeinkommen ab 40000 Euro pro Jahr kann freiwillig ein erhöhter Leistungsbeitrag von 120 Euro bezahlt werden und bei 60000 Euro ein freiwilliger Beitrag von 200 Euro. Optional kann ein beliebiger Jahresbeitrag gezahlt werden, der höher als 62 Euro ist. Falls das jährliche Einkommen nicht höher als der steuerliche Grundfreibetrag ist, gilt ein Beitrag von 50 Euro. Sind bereits Familienmitglieder in der Partei, sinkt der Mindestbeitrag auf 30 Euro. In begründeten Einzelfällen kann der Mitgliedsbeitrag verschoben, gesenkt oder erlassen werden. Bei den Linken beträgt der Mindestbeitrag 1,50 Euro pro Monat. Dieser gilt für Mitglieder ohne Einkommen und Transferleistungsbeziehende. In begründeten Ausnahmefällen kann ein Mitglied für ein Jahr von der Zahlung befreit werden. Darüber hinausgehende monatliche Beiträge sind in der Beitragstabelle der Linken festgelegt, die in der nachfolgenden Tabelle abgebildet ist.

Monatsnettoeinkommen in Euro	Monatsbeitrag in Euro
0	1,50
1 - 500	4,00
500 – 600	5,00
600 – 700	7,00
700 – 800	9,00
800 - 900	12,00
900 – 1000	15,00
1000 – 1100	20,00
1100 – 1300	25,00
1300 – 1500	35,00
1500 – 1700	45,00
1700 - 1900	55,00
1900 – 2100	65,00
2100 – 2300	75,00
2300 – 2500	85,00
> 2500	4% des Nettoeinkommens

Tabelle 4 Mitgliedsbeiträge bei der Links-Partei

Zusätzlich muss ein Mitglied einen jährlichen Beitrag an die Partei der Europäischen Linken zahlen. Dieser kann frei festgelegt werden, beträgt aber mindestens 6 Euro. Mitglieder mit einem Nettoeinkommen bis 700 Euro sind von dem Beitrag an die Europäische Linke befreit. Die FDP hat einen Mindestbeitrag, der an die Höhe der monatlichen Bruttoeinkünfte gekoppelt ist. Bei Einkünften bis 2600 Euro beträgt der Mindestbeitrag 9 Euro, bei 2601 bis 3600 Euro Bruttoeinkommen monatlich

12 Euro, bei 3601 bis 4600 Euro monatlich 18 Euro und bei einem Bruttoeinkommen von über 4600 Euro mindestens 24 Euro im Monat. Als Richtwert gelten 0,5% des monatlichen Bruttoeinkommens. Bei der SPD zahlen Mitglieder ohne Einkommen einen Beitrag von 2,50 Euro. Der Beitrag für Mitglieder mit Einkommen hängt von der Höhe des Monatsnettoeinkommens ab. Bis 1000 Euro gilt ein Beitrag von 5 Euro. Bis 2000 Euro kann ein Beitrag aus den Alternativen 7,50/15/20 Euro gewählt werden. Bis 3000 Euro besteht die Wahl zwischen 25/30/35 Euro. Bis 4000 Euro kann ein Beitrag von 45/60/75 Euro geleistet werden. Bei darüberliegendem Einkommen beträgt der Monatsbeitrag 100/150/250 oder mehr Euro. Generell kann zu Mitgliedsbeiträgen deutscher Parteien folgendes festgestellt werden: Es existiert ein monatlicher oder jährlicher Mindestbeitrag, der bei besonderen Härtefällen erlassen werden kann. Die Höhe des Mindestbeitrags ist für alle Parteimitglieder gleich oder errechnet sich aus dem Netto- oder Bruttoeinkommen. Zusätzlich zum für sie geltenden Mindestbeitrag steht es Parteimitgliedern frei, eine höhere Summe zu zahlen. Die Empfehlung liegt häufig bei einem bestimmten Prozentsatz des Einkommens, typischerweise 0,5% oder 1%. Ähnlich zum Konzept der Parteimitgliedschaft ist eine sogenannte Fördermitgliedschaft. Diese wird für Personen angeboten, die eine Partei zwar unterstützen wollen, sich aber nicht durch einen Parteieintritt zu stark an diese binden wollen. Fördermitglieder sind keine Parteimitglieder, zahlen aber einen monatlichen Betrag. Dafür erhalten sie regelmäßige Informationen zum Parteigeschehen und dürfen an Parteiveranstaltungen teilnehmen, haben dort allerdings kein Stimmrecht. Fördermitgliedschaftsbeiträge sind genau wie Parteispenden steuerlich vergünstigt. Mandatsbeiträge sind regelmäßige Geldleistungen, die ein Mandatsträger, also Inhaber eines öffentlichen Wahlamtes einer Partei, zusätzlich zu seinem Mitgliedsbeitrag leistet. (§27 Abs. 1 PartG)

Parteien erhalten Spenden von natürlichen und juristischen Personen. Diese sind nur teilweise außerhalb der Parteien sichtbar. Spenden über 10.000 Euro müssen im Rechenschaftsbericht der Partei veröffentlicht werden. Die Spender sind mit Name, Adresse und Höhe der Spende anzugeben. Spenden machten im Jahr 2014 zwischen 8% und 29% der Parteieinnahmen aus. Den geringsten Spendenanteil an den Gesamteinnahmen haben die Linken (8%) und die SPD (9%), gefolgt von den Grünen (12%), der AfD (16%) und der CDU (18%). Der höchste Spendenanteil haben CSU und FDP mit jeweils 29%. Als Spender sind bei den juristischen Personen die Bau-, Metall- und chemische Industrie stark vertreten. Zudem spenden Unternehmen aus den Branchen Finanzen, Versicherungen, Nahrungsmittelproduktion, Rüstung, Tabak und Glücksspiel. Die AfD und die Linke erhielten im Jahr 2014 keine meldepflichtigen Spenden von juristischen Personen. Das Spendenverhalten der juristischen Personen ist unterschiedlich. Teilweise spenden Unternehmen an alle der fünf Parteien, die Spenden erhalten haben, wobei die Summe von Partei zu Partei allerdings variiert. Dieses Vorgehen wählten unter anderem die Daimler AG und die Allianz SE. Andere juristische Personen spenden nur an bestimmte Parteien. Die höchsten Einzelspenden gingen an die CSU und FDP vom Verband der Bayerischen Metall- und Elektroindustrie mit einer Höhe von jeweils 330.000 Euro. Bei den Spenden von natürlichen Personen fällt nach einiger Recherche auf, dass die Spender häufig ein politisches Amt innehaben. Sie sind unter anderem Mitglieder des Bundestags, Landtags oder europäischen Parlaments oder hatten früher ein wichtiges politisches Amt inne. Unter den prominenten Spendern befinden sich neben vielen anderen der bayerische Ministerpräsident Horst Seehofer, Altkanzler Gerhard Schröder, der ehemalige Fraktionsvorsitzende der Grünen Hans-Christian Ströbele und der aktuelle linke Ministerpräsident des Landes Thüringen Bodo Ramelow. Einige Mandatsträger von Parteien leisten neben ihren Mitglieds- und Mandatsbeiträgen zusätzlich Parteispenden. Daneben leisten sonstige stark mit der Partei verbundene wohlhabende Privatpersonen Spenden von mehr als 10.000 Euro.

Parteien erhalten für ihre Tätigkeit eine staatliche Teilfinanzierung. Das Gesamtvolumen staatlicher Mittel, das allen Parteien höchstens ausgezahlt werden darf, beträgt für das Jahr 2016 160,5 Millionen Euro (Lammert 2016). Parteien, die bei Bundestags- und Europawahlen 0,5% bzw. bei Landtagswahlen 1% der Stimmen gewinnen, erhalten bis zur nächsten Wahl jährlich 83 Cent für jede für sie abgegebene, gültige Stimme. Sie erhalten außerdem 0,45 Euro für jeden Euro, den sie als Zuwendung

erhalten haben. Eine Zuwendung ist entweder ein eingezahlter Mitgliedsbeitrag, Mandatsbeitrag oder eine rechtmäßig erlangte Spende. Es werden nur Zuwendungen bis zu 3000€ pro natürliche Person berücksichtigt. Die Höhe der staatlichen Parteifinanzierung darf nicht höher sein als die Summe der Einnahmen der Partei. (§18 PartG). Dies wird anhand eines kleines Beispiels verdeutlicht. Angenommen einer Partei stünden nach dem Parteiengesetz 10 Millionen Euro an staatlichen Mitteln zu. Um die vollen 10 Millionen Euro zu erhalten, muss die Partei durch die anderen Einnahmearten ebenfalls 10 Millionen Euro erwirtschaften. Gelingt dies nicht und eine Partei kann auf diesen Wegen nur 9 Millionen Euro erwirtschaften, so erhält sie auch nur 9 Millionen Euro an staatlichen Mitteln. Aus diesem Grund können sich die Parteien nicht auf staatliche Einnahmen verlassen, sondern müssen aktiv Geld einnehmen. Die staatlichen Mittel machten bei allen Parteien außer der SPD die größte Einnahmequelle aus. Bei der mitgliedsstarken SPD liegen die Einnahmen aus Mitgliedsbeiträgen knapp davor.

Die weiteren Einnahmequellen sind Mandatsträgerbeiträge und ähnliche regelmäßige Beiträge, Einnahmen aus Unternehmertätigkeit und Beteiligungen, Einnahmen aus sonstigem Vermögen, Einnahmen aus Veranstaltungen, Vertrieb von Druckschriften und Veröffentlichungen und mit sonstiger mit Einnahmen verbundener Tätigkeit und schließlich sonstige Tätigkeit. Eine Einnahmeform, die auf den Webseiten aller Parteien zu finden ist, ist das Betreiben eines Parteishops. Dort werden Materialien für die Parteiarbeit wie politische Stände angeboten. Daneben gibt es auch zahlreiche andere Produkte mit und ohne Parteibezug wie Textilien, Computer- und Handyzubehör, Weihnachtskarten, Fußbälle, Tassen, Bierkrüge und sogar Badeenten. Die CSU bietet zahlreiche Produkte an, die nicht die Partei, sondern das Bundesland Bayern zum Thema haben. Zum einen werden auf diese Weise Anhänger der Partei mit Informations- und Werbematerial versorgt. Zum anderen können Personen ihre Identifikation mit der Partei nach außen tragen und bei der finanziellen Unterstützung der Partei auch einen Gegenwert erhalten.

4.5. Verfügbarkeit von Datenquellen

Deutschen Parteien stehen wahlkampfrelevante Informationen aus verschiedenen Datenquellen zur Verfügung. Diese werden im Folgenden beschrieben und bewertet.

4.5.1. Parteimitglieder

Parteien haben Parteimitglieder. Um Mitglied einer Partei zu werden, muss ein Bürger einen Mitgliedsantrag ausfüllen und absenden. Dies ist bei allen Parteien online möglich. Aus den Anträgen ergeben sich die Informationen, die ein Partei über ein neues Mitglied erhält. Die mit den Mitgliedschaftsanträgen abgefragten Daten sind in der nachfolgenden Tabelle zusammengefasst. Die erste Spalte zeigt dabei die gefragte Information. Die übrigen Spalten führen für alle Parteien auf, ob sie eine bestimmte Information verpflichtend oder optional erheben oder überhaupt nicht danach fragen. Wurde bei einem Mitgliedschaftsantrag nicht zwischen verpflichtenden und optionalen Angaben unterschieden, wurde angenommen, dass die Angabe verpflichtend ist. Informationen nach denen nicht gefragt wurde, sind mit einem Bindestrich gekennzeichnet.

Erhobene Information	CDU	CSU	SPD	Grüne	Linke	FDP	AFD
Vorname	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht
Nachname	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht
Bankdaten und Beitragshöhe	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht
Geburtsdatum	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht
Adresse	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht
Telefonnummer /Handynummer	Pflicht	Pflicht	Pflicht	Pflicht	Optional	Optional	Pflicht
E-Mail	Pflicht	Pflicht	Pflicht	Pflicht	Pflicht	Optional	Pflicht
Staatsange-hörigkeit	Pflicht	Pflicht	Pflicht	Pflicht	-	Pflicht	Pflicht
Geburtsort	Optional	-	-	Pflicht	-	Pflicht	-
Geschlecht	Pflicht	Pflicht	Pflicht	Optional	-	Optional	Pflicht
Familienstand	Optional	-	-	-	-	-	-
Konfession	Optional	Pflicht	-	-	-	-	-
Beruf	Optional	Pflicht	Pflicht	-	Optional	Optional	
Art der Beschäftigung	Optional	Pflicht	Pflicht	-	-	Optional	-
Arbeitgeber	-	-	Pflicht	-	-	-	-
Frühere Partei-mitgliedschaft	-	-	Pflicht	-	Pflicht	Optional	Pflicht
Mitgliedschaft in Gewerkschaft oder Vereinen	-	-	Pflicht	-	-	-	-
Mitgliedschaft in sozialen Netzwerken	-	-	-	-	Optional	-	-
Ehrenamtliche Tätigkeit	Optional	Pflicht	-	-	-	-	-
Werber (bei Werbung durch Dritten)	Pflicht	Pflicht	Pflicht	-	-	-	-
Politische Themenfelder	-	-	-	-	Optional	-	-

Tabelle 5 Datenerfassung bei Parteieintritt

Die erhobenen Informationen sind, abgesehen von den notwendigen Basisdaten, je nach Partei durchaus verschieden. Außerdem bestehen Unterschiede in der Genauigkeit, mit der nach Informationen gefragt wird. So wird bei der Kategorie „Art der Beschäftigung“ bei der FDP zwischen 13 verschiedenen Beschäftigungsarten unterschieden. Bei den anderen Parteien sind es deutlich weniger. Zu bemerken ist überdies, dass im Vergleich zu den Wählerregistern der USA nicht nach der Ethnie des Antragstellers gefragt wird. Dies könnte daran liegen, dass die Gesellschaft in Deutschland im Vergleich zu den USA homogener ist und die Erfassung der Ethnie als unnötig beziehungsweise irritierend angesehen würde. Stattdessen wird nach der Staatsangehörigkeit des Antragstellers gefragt und bei den Grünen und der FDP zusätzlich nach dem Geburtsort. In welchem Maß die Parteimitglieder auch optionale Daten angegeben haben, kann an dieser Stelle nicht bewertet werden.

Das Datenvolumen, auf das eine Partei bei der Analyse ihrer Parteimitglieder zurückgreifen kann, ist durch die Anzahl der Parteimitglieder festgelegt. Die Anzahl der Parteimitglieder ist in Deutschland gesunken. Die Entwicklung der Anzahl der Parteimitglieder nach Partei in den letzten 24 Jahren ist in der nachfolgenden Abbildung dargestellt.

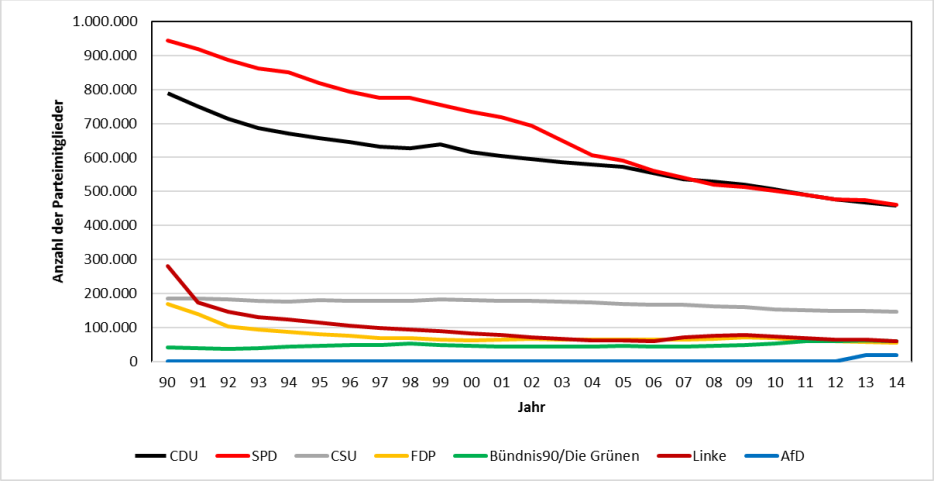


Abbildung 8 Entwicklung der Parteimitgliederzahl seit dem Jahr 1990, erstellt auf Basis von (Niedermayer 2015)

In den größten Parteien SPD und CDU ist die Mitgliederzahl seit dem Jahr 1990 deutlich gesunken. Die SPD, die im Jahr 2014 die meisten Parteimitglieder hatte, hat im Zeitraum von 24 Jahren mehr als die Hälfte ihrer ursprünglichen Mitglieder verloren. Der Mitgliederschwund muss nicht zwingend allein durch das Austreten aus einer Partei stattfinden, sondern kann auch durch Todesfälle bedingt sein. Durch den sehr großen Mitgliederverlust in den letzten 24 Jahren ist aber davon auszugehen, dass der Tod als alleinige Ursache ausscheidet. Die im Jahr 2013 gegründete AfD hat die wenigsten Mitglieder. Die einzige der älteren Parteien, die im betrachteten Zeitraum die Anzahl der Parteimitglieder steigern konnte, ist die Partei der Grünen. Der Mitgliederverlust fällt in relativen Zahlen besonders drastisch bei den Linken und der FDP aus. Die Linke hat seit dem Jahr 1990 etwa 220.000 Parteimitglieder und damit fast 80% der ursprünglichen Mitglieder verloren. Bei der FDP machte der Verlust der Parteimitglieder mit 115.000 knapp 70% aus. Aus der Analyse der Parteimitglieder lassen sich gegebenenfalls Bedarfe für bestimmtes politisches Handeln ableiten. Beispielsweise ist denkbar, dass die austretenden Mitglieder in einem bestimmten Zeitraum ähnliche Merkmale aufweisen. Die Partei kann also ableiten, dass politische Handlungen oder Versäumnisse in der Vergangenheit einen negativen Einfluss auf eine bestimmte Personengruppe hatten. Es ist auch möglich, dass Daten über Parteimitglieder an die Öffentlichkeit gelangen. Dies geschieht durch sogenannte Partei-Leaks, bei denen politisch motivierte Gruppen parteiinterne Daten veröffentlichen. Zwei bekannte Vorfälle dieser Art betrafen die Partei AfD in den Jahren 2015 und 2016. Im Jahr 2015 wurden knapp 3000 Namen inklusive Adresse, Mailadresse, und Mitgliedsnummer von AfD-

Mitgliedern veröffentlicht, welche für den Parteitag in Bremen angemeldet waren. Im Jahr darauf gab es einen Leak des Parteitags in Stuttgart. In diesem wurden die Namen von über 2000 Teilnehmern veröffentlicht. Dabei wurde im Vergleich zum Jahr zuvor zusätzlich die Telefonnummer und das Geburtsdatum der Mitglieder angegeben. Außerdem wurden die Mitgliederlisten für verschiedene Städte veröffentlicht. Die Daten wurden auf der Webseite linksunten.indymedia.org veröffentlicht, sind inzwischen aber nicht mehr aufrufbar. Daten dieser Art scheinen für Parteien interessant, da sie einen Einblick in die Mitgliederstruktur von andern Parteien geben. Zum Beispiel kann geprüft werden, ob die Parteimitglieder der AfD zuvor Mitglieder der eigenen Partei waren. Jedoch ist eine Nutzung dieser Daten aus rechtlichen und moralischen Gründen abzulehnen.

4.5.2. Parteispenden

Parteispenden haben neben der in Unterkapitel 4.4 aufgezeigten finanziellen Bedeutung auch einen Informationszweck. Die Spenden für deutsche Parteien erfolgen per Überweisung, Onlineformular oder SMS. Die Spenden stammen von natürlichen und juristischen Personen. Juristische Personen als neue Spender zu gewinnen, erscheint fragwürdig. Parteispenden von Unternehmen haben immer den Beigeschmack einer Einflussnahme. Falls Parteien bei juristischen Personen aktiv nach Spenden fragen würden, würde dies eine größere Nähe zwischen Partei und Geldgeber bedeuten als die bloße passive Annahme einer Unternehmensspende. Wegen der mangelnden Transparenz bei Parteispenden wurde Deutschland bereits mehrfach von der Staatengruppe gegen Korruption des Europarats gerügt, zuletzt in diesem Jahr. Die Spenden, die Parteien erhalten, können erst zwei Jahre später in den Rechenschaftsberichten eingesehen werden. Parteispenden über 50.000 Euro müssen jedoch unverzüglich veröffentlicht werden. Daher kann für die anderen Parteien schlecht abgeschätzt werden, wie viel Geld ihnen für den Wahlkampf zur Verfügung steht. Da auch nur Spenden über 10.000 Euro ausgewiesen werden müssen, sind Spendendaten überwiegend parteiinterne Daten. Wer die Spender sind, die einen Beitrag unter 10.000 Euro spenden, bleibt unklar. Eine mangelnde Transparenz bedeutet weniger beziehungsweise stark verzögerte Einblicke in die Spenden der anderen Parteien. (GRECO Staatengruppe gegen Korruption 2016)

Falls die Spender zugleich Parteimitglieder sind, kann versucht werden, in den zum Parteimitglied bekannten Informationen Muster zu finden, die einen Zusammenhang zum Spendenverhalten aufzeigen. Daraus können gegebenenfalls neue Spender identifiziert werden. Bisherige Spender können in einem wichtigen Wahljahr um eine erneute Spende gebeten werden. Dabei kann nach dem gleichen Beitrag wie bei der letzten Spende oder um den Beitrag mit einem kleinen prozentualen Aufschlag, zum Beispiel fünf Prozent gefragt werden. Außerdem ist es möglich experimentell zu prüfen, welche Themen und Textbausteine die höchsten Spenden erzeugen. Diese können dann verwendet werden. Der Erfolg derartiger Experimente kann sehr gut bewertet werden. Der Spender kann anhand seines Namens identifiziert werden. Bei einer Überweisung ist sein Name als Auftraggeber ersichtlich. Und bei einer Online-Spende trägt der Spender seinen Namen zusätzlich in das Online-Formular ein. Zufällig ausgewählten Testgruppen kann dann zum Beispiel fünf Varianten einer Nachricht mit der Bitte um eine Spende zugesendet werden. Die Nachrichten unterscheiden sich in der Wortwahl. Nach einem festgelegten Zeitraum nach Versenden der Nachricht kann anhand der eingegangenen Spenden überprüft werden, welche Nachricht am erfolgreichsten war. Dies kann anhand der Antwortrate pro Gruppe oder Höhe der eingegangenen Spenden bewertet werden. Eine textliche Variation wird die Spendenbereitschaft nicht sofort deutlich steigern können. Vielmehr geht es darum, durch eine kleine Verbesserung in der Masse ein besseres Ergebnis zu erzielen.

4.5.3. Angebote des Bundeswahlleiters

Vom Bundeswahlleiter können Informationen zu Wahlen und Parteien bezogen werden. Darunter befindet sich die repräsentative Wahlstatistik. Diese entsteht durch die statistische Auswertung des Ergebnisses zur Wahl des deutschen Bundestags (§ 1 WStatG). An der Statistik dürfen maximal 5% der Wahlbezirke und der Briefwahlbezirke des Bundesgebiets und maximal 10% der Wahlbezirke und der Briefwahlbezirke eines Landes teilnehmen. In dieser wird die Stimmenverteilung für verschiedene Altersgruppen und Geschlechter in verschiedenen Gebieten angegeben. Jedes Bundesland stellt ein Gebiet dar. Außerdem wird Deutschland als Gebiet betrachtet sowie die alten Bundesländer inklusive Berlin-West und die neuen Bundesländer inklusive Berlin-Ost. Die Altersgruppen sind unterteilt in die Altersgruppen 18-24 Jahre, 25-34 Jahre, 35-44 Jahre, 45-49 Jahre, 60-69 Jahre und über 70 Jahre. Aus diesen Statistiken ergibt sich ein grober Überblick, wie Alter und Geschlecht sich auf das Wahlverhalten der Wähler auswirkt. Beispielsweise ist es möglich anhand des Gini-Indexes zu berechnen, wie die „Parteikonzentration" der Altersgruppen in den einzelnen Bundesländern auf eine Partei ist. Für die Bundestagswahl 2013 zeigt sich dann, dass Frauen aus Bayern in der Altersgruppe „70 und mehr" den geringsten Gini-Index haben. Das heißt, dass in dieser Gruppe die Stimmen am konzentriertesten sind. Den höchsten Gini-Index haben Männer aus Berlin in der Altersgruppe 35-44. Hier verteilen sich die Stimmen also am homogensten auf alle Parteien. Dadurch lässt sich möglicherweise ableiten, bei welchen Gruppen ein Überzeugungsversuch von der eigenen Partei wahrscheinlicher ist als bei anderen. Aus der Statistik lässt sich nachvollziehen wie das Stimmverhalten der Bevölkerung von Geschlecht und Alter abhängt. So kann ein grober Überblick über das Wahlverhalten gewonnen werden. Neben diesem Angebot können die Parteiunterlagen aller Parteien heruntergeladen werden. Aus diesen kann die Beitragsordnung und damit die Höhe der Mitgliedsbeiträge bei den anderen Parteien eingesehen und nachvollzogen werden.

4.5.4. Statistikämter

In Deutschland gibt es verschiedene öffentliche Statistikämter. Auf Bundeseben ist dies das Statistische Bundesamt. Auf Landeseben gibt es in der Regel für jedes Bundesland ein statistisches Landesamt. Ausnahmen bilden dabei das Amt für Statistik Berlin-Brandenburg und das Statistikamt Nord, welche jeweils für zwei Bundesländer zuständig sind. Beim Statistikamt Nord sind dies Schleswig-Holstein und Hamburg. Unter Koordination der statischen Ämter fand am 09. Mai 2011 im Rahmen der ersten gemeinsamen Volkszählung der Staaten der europäischen Union die Durchführung des Zensus 2011 statt. Beim Zensus handelt es sich um eine Bevölkerungs-, Gebäude-, und Wohnungszählung. Die diesem Zensus vorhergehenden Volkszählungen fanden zuletzt im Jahr 1987 für die Bundesrepublik und im Jahr 1981 für die Deutsche Demokratische Republik statt. Zwischen dem letzten und dem aktuellen Zensus liegen also 24 beziehungsweise 30 Jahre. Der nächste Zensus ist für das Jahr 2020 geplant (Statistische Ämter des Bundes und der Länder 2015). Neben dem Zensus findet jährlich der Mikrozensus statt. Der Mikrozensus ist eine repräsentative Haushaltsbefragung. Dabei werden 1% der Bevölkerung, die zufällig ausgewählt werden, stellvertretend für die Gesamtbevölkerung zu ihren Lebensbedingungen befragt. Beim Mikrozensus wird eine Person über ihre Lebensverhältnisse in verschiedenen Bereichen befragt. Darunter fallen Angaben zur Person, zu Beruf und Bildung, zu Einkommen, Altersvorsorge und Wohnsitz. Alle vier Jahre werden beim Mikrozensus auch die Bereiche Wohnsituation, Migration, Versicherungen, Pendlerverhalten und Gesundheit abgefragt (Statistisches Bundesamt 2016). Neben dem Zensus gibt es zahlreiche, regional aufgeschlüsselte Publikationen über verschiedenste Themen. Veröffentlicht werden die Statistiken in der Genesis Online-Datenbank, die ein gemeinsam entwickelten Datenbanksystem des Bundesamts und der Landesämter ist. Die Themen, zu denen Daten bezogen werden können sind Gebiet, Bevölkerung, Arbeitsmarkt, Wahlen, Bildung, Sozialleistungen, Gesundheit, Recht, Wohnen, Umwelt, die Wirtschaftsbereiche, Außenhandel, Unternehmen, Handwerk, Preise, Verdienste, Einkommen und Verbrauch, öffentliche Finanzen, Steuern, Personal, Gesamtrechnungen und nationale und internationale Indikatorensysteme (Statistisches Bundesamt 2016). Bei einer Beachtung verschiedener

statistischer Daten für die einzelnen deutschen Gebiete auf unterschiedlich genauen Betrachtungsebenen, ergibt sich für jede Gemeinde somit ein charakteristisches Bild, bei dem versucht werden kann, es mit Wahlergebnissen und Ergebnissen aus Umfragen in Verbindung zu bringen. Die Integration der Daten bedeutet allerdings einigen Aufwand, da die Daten für unterschiedliche Gebiete vorliegen und von Zeit zu Zeit auch eine Neugliederung der Gebiete stattfindet. Zudem werden die Daten zu unterschiedlichen Zeitpunkten gesammelt und es sollte versucht werden, den Datenbestand auf dem aktuellen Stand zu halten. Zusammenfassend ist festzustellen, dass in Deutschland ein breiter Bestand an statistischen Daten zur Bevölkerung und zu wirtschaftlichen Themen besteht, der öffentlich für jeden zugänglich ist.

4.5.5. Soziale Medien

Deutschen Parteien sind in den sozialen Medien angekommen. Alle Parteien sind in den gängigen Netzwerken wie Twitter, Facebook und auch bei der Videoplattform Youtube vertreten. Neben den Parteiauftritten auf den Netzwerken haben auch viele deutsche Politiker eigene Facebook- und Twitter-Accounts, die sie zur politischen Kommunikation nutzen. Auf der Foto- und Videoplattform Instagram sind deutsche Parteien und Politiker nicht aktiv. In anderen Ländern ist dies durchaus der Fall. Für den künftigen US-Präsidenten Donald Trump, die Vorsitzende des französischen Front National Marine Le Pen und auch den österreichischen Außenminister Sebastian Kurz existiert ein Instagram-Account. Soziale Netzwerke und andere Onlinequellen können analysiert werden, um das aktuelle Stimmungsbild und das Ausmaß an Reaktionen über bestimmte Aktionen auszuwerten. Häufig genutzte Netzwerke in Deutschland sind Facebook und Twitter. Die Anzahl der Nutzer in Deutschland für die Netzwerke ist nicht genau bekannt. Außerdem muss zwischen täglichen, gelegentlichen und denjenigen Nutzern, die zwar einen Account erstellt haben, das Netzwerk aber nicht wirklich nutzen, unterschieden werden. Die reine Angabe der Nutzerzahl ohne die Berücksichtigung der Nutzungsintensität ist daher nicht besonders aufschlussreich. Das Statistikportal Statista geht von 27 Millionen Facebook-Nutzern im Jahr 2016 aus (Statista 2016). Für Twitter sind die Zahlen unbekannt. Mit dem Thema Twitter und der deutschen Bundestagswahl setzten sich (Tumasjan, Sprenger, Sandner et al. 2010) auseinander. Sie nutzten knapp über 100.000 Tweets, die in den Wochen vor der Bundestagswahl 2009 veröffentlicht wurden. Sie kommen zu dem Schluss, dass Twitter in Deutschland als Medium für politische Meinungen genutzt wird. Die Meinungsäußerungen sind aber bei wenigen Nutzern konzentriert. So verfassten nur 4% der Nutzer 40% der politischen Nachrichten. Dies muss bei der Betrachtung der Tweets berücksichtigt werden. Ein zunehmend von Unternehmen genutztes Verfahren, das auch Parteien nutzen können, ist die Sentiment-Analyse. Dabei handelt es sich um die Aufgabe, die Meinungen von Personen über ein Thema herauszufinden. Die Erfassung auf verschiedenen sozialen Medien öffentlich geäußerten Meinungen ermöglichen einer Organisation, in Echtzeit Feedback auf ihr Handeln zu erhalten. Dadurch kann schnell auf Entwicklungen reagiert werden. Parteien können verfolgen, wie Wähler zu bestimmten politischen Themen eingestellt sind und wie in den sozialen Medien auf die öffentlichen Äußerungen und Handlungen eines Politikers oder einer Partei reagiert wird. (Feldman 2013)

Ebenso wie die Plattform mybarackobama.com bieten einige deutsche Parteien für ihre Plattform einen Login via Facebook an. Dann werden die Facebook-Daten, die öffentlichen Daten eines Facebook-Nutzers an die Partei übertragen. Wenn eine Partei ihre Parteimitglieder und Unterstützer mit ihren Accounts in den sozialen Medien verknüpfen kann, erhält sie weitere Einblicke über diese Personen. Ebenso wie bei der Obama-Kampagne können deutsche Parteien Facebook-Anwendungen erstellen, die auf die Daten der Anwendungsnutzer zugreifen können, sofern diese ihre Zustimmung geben.

4.5.6. Privatwirtschaftliche Unternehmen

In Deutschland existiert ein Markt für den Adresshandel. Ebenso wie bei Spenden und Parteimitgliedsbeiträgen gilt hier §28 des Bundesdatenschutzgesetzes. Eine Zustimmung des Betroffenen, dessen Daten gesammelt werden, ist nicht erforderlich; er kann aber der Verwendung der Daten widersprechen. So wird über eine Vielzahl an Bürgern Daten gesammelt, ohne dass diesen die Sammlung bewusst ist. Im Folgenden werden nun einige wichtige Adresshändler vorgestellt.

Die Deutsche Post AG hat mit der Deutschen Post Direkt GmbH ein Tochterunternehmen, das Adressen verkauft und vermietet, die mit verschiedenen weiteren Merkmalen angereichert sind. Das Unternehmen sammelt Daten in einer Datenbank, das es micro-Dialog-Datenbank nennt. Dort sind im Schnitt 6,6 Haushalte zu einer sogenannten Mikrozelle zusammengefasst. Die Zusammenfassung basiert auf der Annahme, dass ähnliche Personen nahe beieinander wohnen. Für eine Mikrozelle sind verschiedene Informationen verfügbar, die in der folgenden Aufzählung genannt sind.

- Soziodemographische Daten wie Alter, Kaufkraft, Familienstruktur
- Konsumdaten über 22 verschiedene Sortimentsaffinitäten wie Kleidung, Haustierartikel, Technik und Haushaltsgeräte, Bestellwert
- Strukturdaten wie Gebäudetyp und Höhe der Kaltmiete
- Regionale Daten wie Postleitzahlengebiete und Bundesländer
- Informationen zu Pkw-Besitz, Fahrzeugdichte, Alter und Leistung der Pkw, Anzahl der Halter eines Pkw, Pkw-Klasse, Segmentdichte, Kaufverhalten wie Neuwagenkäufer, Jahreswagenkäufer, Zweitwagenbesitzer, Firmenwagenbesitzer
- Vertretene Werte wie Gesundheit und Nachhaltigkeit
- Verhalten in Banken-, Versicherungs- und Finanzmärkten
- Lebensauffassung und Lebensweise anhand des Sinus-Modells
- Sieben Zielgruppensegemente basierend auf Motiven, Wünschen und Emotionen im Zusammenhang mit Kaufverhalten

Die Mikrozellen mit im Schnitt 6,6 Haushalten sind sehr feingranular und sehr nahe an der individuellen Ebene. Angeboten werden auch Umzugsdaten, da die Deutsche Post Umzüge in der Regel sehr gut nachfolgen kann, selbst wenn sich der Betroffene nicht beim Einwohnermeldeamt ummeldet. Die Daten werden häufig verwendet, um zielgerichtete Haushaltswerbung durchzuführen. (Deutsche Post Direkt 2015)

Die AZ Direct GmbH gehört zu dem Unternehmen Arvato, das wiederum eine Tochtergesellschaft des deutschen Medienkonzerns Bertelsmann ist. Die AZ Direct GmbH bietet 34 Millionen Adressen an, die sie „Konsumer-Daten" nennt. Sie betreibt eine Datenbank, das Audience Targeting System AZ DIAS, die zu ca. 40 Millionen Haushalten, 70 Millionen Personen und 20 Millionen Gebäuden Profildaten enthält. Für diese Daten werden ca. 600 Merkmale erfasst. Die Abdeckung von 70 Millionen Personen bedeutet, dass für einen Großteil der deutschen Bevölkerung Daten vorhanden sind. Es wird jedoch keine genauere Information gegeben, welche Merkmale erfasst werden. (Arvato Bertelsmann 2016)

Ein weltweit aktiver Datenbroker ist Acxiom, in Deutschland vertreten durch die Acxiom Deutschland GmbH. Sie gibt an, über „qualitativ hochwertige Adressdaten mit zahlreichen Zusatzinformationen" zu verfügen, welche das Konsumverhalten der Haushalte anzeigen. Für weitere Informationen ist aber eine Kontaktaufnahme nötig. Da detaillierte Auskünfte wahrscheinlich nur an zahlende Kunden gegeben werden, wurde eine Anfrage unterlassen. Basierend auf der Datenbank ist eine Marktsegmentierung möglich. Die Datenbank nutzt Informationen von Quellen wie Adressdatenbanken, Markt-Media-Studien, anonymisierten und aggregierten Transaktionsdaten, aggregierte Statistiken aus den amtlichen Melderegistern und Veröffentlichungen des Statischen Bundesamts, der Statischen Landesämter, der Bundesagentur für Arbeit und des Kraftfahrt-

Bundesamts. Aus diesen Informationen erfolgt die Zuweisung eines Haushalts in einen von 14 Segmenten. Diese sind beispielsweise „Jung & Berufseinstieg“, „Kinderlos & Aktiv“ oder „Junge Senioren & Kleines Budget“. Das Unternehmen nutzt also verschiedene öffentliche und nicht-öffentliche Quellen, um ein integriertes Bild über Personen zu ermöglichen. (acxiom 2016)

Eine in Deutschland häufig genutzte Dienstleistung, die den Datensammlungen in den USA ähnlich ist, ist die Sammlung über Daten des Konsumverhaltens mittels Payback. Mit dieser Karte werden Daten des Kunden über sein Kaufverhalten gesammelt. An der Erfassung sind viele Partnerunternehmen aus dem stationären Einzelhandel, aber auch Tankstellen und Online-Shops beteiligt. Die Daten werden jedoch nur an Partnerunternehmen weitergegeben und können demnach nicht von politischen Parteien genutzt werden. (Selk 2016)

Die bestehenden Datenangebote ermöglichen eine zielgerechte Kundenansprache für Unternehmen. Es ist aber unklar, ob auch politische Interessen der Bewohner abgebildet werden. Im Gegensatz zu den USA konnten keine Unternehmen identifiziert werden, welche gezielt Daten zur Unterstützung des Wahlkampfes für deutsche Parteien bereitstellen oder auf genau diesen Zweck zugeschnittene Software anbieten. Wie in Unterkapitel zu Datenschutz beschrieben ist die Nutzung von Auskünften aus Melderegistern zum Zwecke des Adresshandels nicht ohne Einwilligung möglich. Dies gilt jedoch erst seit November 2015. Zuvor war bei der Weitergabe von Daten an Adresshändler ein expliziter Widerspruch erforderlich. Es ist davon auszugehen, dass dieser Widerspruch nur einem Teil der Bürger bewusst war und von einem noch geringeren Teil wirklich eingelegt wurde. Da die Neuregelung noch nicht so lange zurückliegt und nur wenige Bürger einen Widerspruch eingelegt haben, sollte der Qualitätsverlust der Daten bei den Adresshändlern momentan noch nicht allzu gravierend ausfallen.

4.5.7. Forschungsinstitute

In Deutschland gibt es eine Reihe an Forschungsinstituten, die sich mit Markt- Sozial- und Politikforschung beschäftigen. Diese haben alle ein relativ ähnliches Profil und werden im Folgenden knapp vorgestellt. Dabei wird nur auf die mit Politik verbundenen Dienstleistungen eingegangen. Die Forschungsgruppe Wahlen ist hauptsächlich für die wissenschaftliche Betreuung und Beratung der Wahlsendungen des Fernsehsenders ZDF zuständig. Es werden aber auch Umfragen für andere Auftraggeber durchgeführt. Die Themenschwerpunkte liegen bei Prognosen und Hochrechnungen zu Wahlen, dem Politbarometer zur aktuellen politischen Stimmung in Deutschland, die Beobachtung gesellschaftlicher Trends und die Erforschung von Wählerverhalten. Das Pendant zur Forschungsgruppe Wahlen ist das politische Umfrageinstitut Infratest dimap, das die Wahlberichterstattung der ARD mit der Erstellung von Prognosen, Hochrechnungen und Wahlanalysen unterstützt. Auch die aproxima Gesellschaft für Markt- und Sozialforschung bietet Politikforschung an. Sie erforscht auf Langzeitbeobachtungen gestützte Informationen zu Wahl- und Parteipräferenzen in der wahlberechtigten Bevölkerung. Daneben werden Informationen zu politischen Einstellungen und Werteorientierungen geboten. Außerdem werden Einschätzungen zu persönlichem und gesellschaftlichem Reformbedarf und Empfehlungen für die kurz-, mittel- und langfristige Politikgestaltung gegeben. Das Institut für Demoskopie Allensbach ist auf repräsentative Befragungen spezialisiert. Politische Meinungsforschung spielt eine untergeordnete Rolle. Für Bundes- und Landesministerien werden regelmäßig Umfragen durchgeführt. Das Institut für Markt und Sozialforschung Insa Consulere bietet neben Umfragen und Analysen für Unternehmen auch Politikberatung an. Sie decken die Bereiche Wählerbefragung, Potentialanalyse und Datenerhebungen ab. TNS Emnid ist im Bereich Politik- und Sozialforschung tätig. Im politischen Bereich bietet es „politische Meinungsforschung zu lokalen, nationalen und internationalen Themen, Parteien- und Kandidatenprofile, Wählergruppen, Politikern, Regierung“. IM Leipzig ist ein Markt-forschungsinstitut, welches auch im Bereich Politik- und Sozialforschung tätig ist. Eine regelmäßige Studie des Instituts ist das Meinungsbarometer, das die Einstellungen und Stimmungen in der

volljährigen Bevölkerung zu politischen, wirtschaftlichen und gesellschaftlichen Themen widerspiegelt. Als Dienstleistung im politischen Bereich wird angegeben, dass Ministerien mittels fundierter Analysen darüber beraten werden können, wie die Gesellschaft auf ihre Aktivitäten reagieren wird. Konkret angeboten werden unter anderem Einstellungs- und Werteerhebungen, Reputationsstudien und Zielgruppenbefragungen. Es existieren also zahlreiche Forschungsinstitute, bei denen Umfragen in Auftrag gegeben werden können, die das Stimmungsbild in der Bevölkerung zu bestimmten politischen oder sozialen Themen wiedergeben können. Außerdem bieten die Institute Politikberatung an. Bei Betrachtung der Leistungsangebote fällt auf, dass die Datenanalyse in den Händen der Forschungsinstitute bleibt. Die Datenanalyse und die Ableitung von Empfehlungen ist von den Auftraggebern getrennt. Im Sinne des US-amerikanischen Modells wäre eine Bereitstellung der Daten an die Parteien und eine Integration dieser Daten in den parteilichen Datenbestand. (aproxima Gesellschaft für Markt- und Sozialforschung 2016, Forschungsgruppe Wahlen 2016, IM Leipzig 2016, infratest dimap 2016, Insa Consulere 2016, Institut für Demoskopie Allensbach 2016, TNS Emnid 2016)

Nützliche Informationen zum Thema Wahlkampf können auch von Fachzeitschriften der Politikwissenschaft stammen. Wichtige Zeitschriften sind in diesem Zusammenhang die Politische Vierteljahresschrift, die Zeitschrift für Politik, die Zeitschrift für Politikwissenschaft und die Zeitschrift für Parlamentsfragen, die alle vom Nomos-Verlag verlegt werden. In der US-amerikanischen Politikwissenschaft gab es Experimenten mit den Ergebnissen, dass sozialer Druck, das Streben nach Normkonformität und eine persönliche Ansprache in den USA genutzt werden kann, um die Wahlbeteiligung zu erhöhen. Ob die Ergebnisse auf Deutschland übertragbar sind, ist zu prüfen. Die Prüfung ist allerdings dadurch erschwert, dass im Gegensatz zu den USA nicht feststellbar ist, welche Personen gewählt haben und welche nicht.

4.5.8. Zusammenfassung der Datenquellen

Die vorherigen Ausführungen zeigen, dass deutsche Parteien auf mehrere Datenquellen zugreifen können, die ihnen Informationen über Wähler liefern. Diese unterscheiden sich nach ihrer Verfügbarkeit, den Beschaffungskosten und ihrem Volumen. Die Datenquellen sind in der folgenden Tabelle noch einmal zusammengefasst.

Name	Verfügbarkeit	Beschaffungskosten	Volumen
Parteimitglieder	Parteiintern	Kostenlos	17.500 – 460.000 aktuelle Parteimitglieder, zusätzlich ausgetretene oder verstorbene Mitglieder
Parteispenden	Parteiintern, Gesamtsumme öffentlich extern, Spenden über 10.000€ öffentlich extern	Kostenlos	Für das Jahr 2014 je nach Partei 8-231 Spender, zusätzlich Spender aus vorherigen Jahren und bei einem Wert über 10.000€ Spenden an andere Parteien
Bundeswahlleiter	Öffentlich extern	Kostenlos	Bundeswahlstatistiken der letzten Wahlen
Statistikämter	Öffentlich extern	Je nach Statistik kostenlos oder kostenpflichtig	Zahlreiche Publikationen auf verschiedenen Gebietsebenen
Soziale Medien	Öffentlich extern	Kostenlos	Hoch
Privatwirtschaftliche Unternehmen	Nicht-öffentlich extern	Kostenpflichtig	Mehrere Millionen Adressen, erweitert mit Konsum-Informationen
Forschungsinstitute	Nicht-öffentlich extern	Kostenpflichtig	Je nach Auftragsgröße

Tabelle 6 Zusammenfassung der für deutsche Parteien verfügbare Datenquellen

Hinsichtlich der Verfügbarkeit kann zwischen parteiinternen, parteiöffentlichen, öffentlichen externen und nicht-öffentlichen externen Daten unterschieden werden. Parteiinterne Daten werden im Rahmen von Parteieintritten, den Daten in den sozialen Netzwerken der Parteien und bei Spenden an Parteien und anderen Einnahmeformen gesammelt. Die Anzahl der Parteimitglieder ist in den letzten Jahren stark gesunken und beläuft sich auf 17.500 bis 460.000 Mitglieder. Bei Spenden müssen die Gesamtspendensumme pro Jahr und Spenden über 10.000 Euro veröffentlicht werden. Parteien können neben dem internen Datenbestand auch auf externe Datenangebote zurückgreifen. Externe öffentliche Daten stammen dabei vom Bundeswahlleiter, statistischen Ämtern und sozialen Medien. Das Angebot des Bundeswahlleiters beschränkt sich im Wesentlichen auf die repräsentative Wahlstatistik. Die Angebote der deutschen Statistikämter sind vielfältig. Daten aus sozialen Medien können mittels einer Sentiment Analyse genutzt werden oder es wird durch entsprechende Anwendungen auf Nutzerdaten zugegriffen. Schließlich sind externe Daten zu nennen, die nicht-öffentlich sind. Privatwirtschaftliche Unternehmen, in Deutschland insbesondere Adresshändler, verkaufen Daten über Haushalte. Für die Haushalte sind verschiedene Eigenschaften gespeichert, die ihre Konsumpräferenzen widerspiegeln. Forschungsinstitute führen Umfragen durch, die die Einstellung der Bevölkerung zu bestimmten Themen erheben. Je nach Größe der Befragung werden unterschiedlich viele Informationen gesammelt. Der Erwerb der Daten ist bei parteiinternen Daten, abgesehen von den Kosten für Personal und Technik, kostenlos. Öffentliche externe Daten sind mit Ausnahme von bestimmten kostenpflichtigen Statistiken ebenfalls kostenlos zu erwerben. Nicht-öffentliche externe Daten sind kostenpflichtig.

4.6. Kommunikation mit dem Wähler

Deutschen Parteien stehen für die Kommunikation mit den Wählern verschiedene Kanäle zur Verfügung. Einige wichtige Kanäle werden im folgenden Abschnitt vorgestellt, wobei zwischen Offline- und Online-Kanälen unterschieden wird.

4.6.1. Offline-Kanäle

Eine wichtige Rolle im US-amerikanischen Wahlkampf nahm der Tür-zu-Tür-Wahlkampf ein, bei dem Freiwillige in ihrer Umgebung an die Haustüren der Wähler klopften und mit ihnen auf vorgegebene Weise kommunizierten. Dieses Vorgehen ist in veränderter Form von der SPD übernommen worden. Bei der Oberbürgermeisterwahl in Wiesbaden wurden im Jahr 2013 mehr als 8000 Hausbesuche getätigt. Ziel dieses Tür-zu-Tür-Wahlkampfs war die Mobilisierung von Nichtwählern und nicht die inhaltliche Diskussion von Sachthemen. Zur Führung des Gesprächs stand ein Gesprächsleitfaden zur Verfügung, die konkrete Formulierung des Gesprächs war aber dem SPD-Wahlkämpfer vor Ort überlassen. Im Gegensatz zu den USA sind also keine komplett vorformulierten Gesprächsskripte vorgegeben. Das Bestreben zur Wählermobilisierung entstand daraus, dass in der vorherigen Bundestagswahl im Jahr 2009 zwei Millionen Bürger, die im Jahr 2005 noch die SPD gewählt hatten, nicht mehr zur Wahl gingen. Zur Koordination des Tür-zu-Tür-Wahlkampfs wurde die Plattform mitmachen.spd.de erstellt, welche eine leicht erkennbare Anlehnung an mybarackobama.com ist. Die Orte, an denen die Wählermobilisierung stattfand, wurden auf Stimmbezirk-Ebene ermittelt. Für die Stimmbezirke, die Mobilisierungsbezirke genannt wurden, wurde ein Mobilisierungsindex ermittelt, der sich aus der Wahlbeteiligung und dem Stimmanteil für die SPD zusammensetzte. Es wurden Stimmbezirke mit einer geringen Wahlbeteiligung und einem hohen Stimmanteil für die SPD aufgesucht. Um zu vermeiden, dass Hausbewohner vom Besuch vollkommen überrascht werden, wurden Ankündigungsflyer über die Aktion verteilt. Zusätzlich wurde empfohlen, eine Pressemitteilung zu machen, eine Anzeige in der Lokalzeitung zu schalten und die Ankündigung der Aktion in den sozialen Medien. (Janssen, Schlote und Stolzenberg 2013)

Eine Übernahme des Tür-zu-Tür-Wahlkampfes durch alle Parteien könnte problematisch sein. In Deutschland gibt es derzeit sechs Parteien, die wahrscheinlich in den Bundestag einziehen werden. Der Tür-zu-Tür-Wahlkampf dient der Mobilisierung von Wahlberechtigten und der Überzeugung von noch unentschlossenen Bürgern. Für den Fall, dass ein Bürger bis kurz vor der Wahl unentschlossen ist und alle sechs Parteien dies erkennen, würde bei dem Bürger in der Zeit vor dem Wahlkampf sechs Gruppen von jeweils einer anderen Partei stehen. Dies könnte den Effekt haben, dass auf die Hausbesuche negativer reagiert wird als wenn nur zwei Parteien einen Besuch durchführen würden. Generell kann aber gesagt werden, dass Parteien wie in den USA die Interaktionen mit den Wählern erfassen können. Beim Tür-zu-Tür-Wahlkampf können die Reaktionen der Bürger in einem bestimmten Gebiet erfasst werden.

Das Nutzen von automatisierten Werbeanrufen von Parteien, wie sie Unternehmen aus den USA anbieten, würde in Deutschland eher befremdlich wirken. Etwas, das in der Literatur zum Wahlkampf in den USA nicht vorkam, aber in Deutschland weit verbreitet ist, sind Informationsstände an öffentlichen Plätzen, vor allem auch in Innenstädten. Dies wird auch dadurch belegt, dass viele Parteien in ihren Webshops Material für diese Infotheken verkaufen. In diesen Infoständen könnten wie auch in den USA mobile Apps zur Datenerfassung genutzt werden. Gesammelt werden können die Reaktionen von Bürgern auf bestimmte Themen und soziodemographische Merkmale der Personen zu den Reaktionen. Diese Merkmale sind das geschätzte Alter und das Geschlecht und eventuelle auffällige äußerliche Merkmale. Die Erfassung der Daten muss anonym erfolgen.

4.6.2. Online-Kanäle

Deutsche Parteien betreiben Webseiten, auf denen aktuelle Informationen bereitgestellt werden und Mitgliedschaftsanträge ausgefüllt werden können. Außerdem gibt es ein Online-Formular für Parteispenden. Überdies betreiben deutsche Parteien wie in den USA Online-Communities, über die sich ihre Mitglieder und Freiwillige vernetzen können. In Deutschland ist ein Großteil der Funktionen nur für Parteimitglieder zugänglich, weshalb der Umfang und die Funktionen der Communities nicht abgeschätzt werden konnten. Bei der SPD gibt es die Community mitmachen.spd.de. Dort ist es möglich, sich als Freiwilliger zu registrieren, um sich mit den Freiwilligen in seiner Umgebung zu vernetzen. Der Unterbezirk Darmstadt-Stadt hatte am 31.09.2016 insgesamt 18 Mitglieder. Die CDU betreibt CDUPlus, „die Online- und Serviceplattform der CDU Deutschlands". Dort können sich CDU-Mitglieder mit ihrer Mitgliedsnummer und Unterstützer ohne Mitgliedschaft registrieren, wobei die zweite Gruppe nur einen beschränkten Zugang zum Angebot hat. Zum Account der Plattform können die Auftritte eines Nutzers in den sozialen Netzwerken hinzugefügt werden. Auf der Webseite der CSU ist es möglich, sich im Bereich „Meine CSU" einzuloggen. Jedoch ist dies nur für Parteimitglieder möglich. Bei der AfD und den Grünen wurde kein Mitgliederportal auf der Webseite gefunden. Die FDP betreibt das soziale Netzwerk „meine freiheit". Ein Login ist mit den sozialen Netzwerken Google+, Twitter und Facebook möglich. Auch hier gibt es für Nicht-Parteimitglieder nur einen eingeschränkten Zugriff. Bei der Linkspartei gibt es keine Online-Community, aber es ist möglich, an der Kampagne teilzunehmen. Dafür kann sich eine interessierte Person auf der Webseite anmelden und wird dann über aktuelle Aktionen der Partei per Mail informiert.

Für interessierte Bürger besteht die Möglichkeit, sich auf den Webseiten der Parteien für einen Newsletter anzumelden. Die Anmeldung ist dabei häufig sehr gut auf der Homepage sichtbar und kann dort gleich vorgenommen werden. Die Newsletter werden in der Regel wöchentlich versandt. Sie werden oft zum Negative Campaigning genutzt. Das heißt, es wird versucht, gegnerische Parteien durch negative Botschaften schlechter dastehen zu lassen. Als Nachteil dieser Methode wird gesehen, dass die Glaubwürdigkeit und Seriosität der Negative Campaigning betreibenden Partei sinken könnte. Die FDP gibt an, die Zugriffe auf ihren Newsletter und Zählpixel zu nutzen und wird die „Daten in anonymisierter Form zu Optimierungs- und Studienzwecken sammeln und speichern." Die Datenschutzerklärung der CDU macht folgende Aussage: „Wenn Sie den Newsletter öffnen oder einen Link darin anklicken, wird dies über unseren Webserver protokolliert (Datum, Uhrzeit, Mail-Adresse). Das dient internen statistischen Zwecken. Diese Daten werden nicht zu persönlichen Nutzungsprofilen zusammengeführt.". Die SPD gibt Folgendes an: „In unserem Newsletter werden Zählpixel zur Messung der Zugriffe auf den Newsletter eingesetzt. Zählpixel werden von uns allein zu statistischen Zwecken genutzt, um Zugriffszahlen zu erheben". Die AfD macht zu Newslettern keine Angaben. Bei der Linken wird auf Newsletter eingegangen. Die Zugriffe auf den Newsletter werden aber offenbar nicht analysiert. Die CSU und Grünen geben ebenfalls nicht an, eine Analyse der Newsletter-Nutzung vorzunehmen. In den Newslettern ist ersichtlich, dass die regierenden Parteien von ihrer erfolgreichen Regierungsarbeit sprechen, während diese von den Oppositionsparteien kritisiert wird. Die Newsletter informieren außerdem über aktuelle Parteiereignisse. Einige Parteien werten die Nutzung der Newsletter aus und sehen so, wie sehr bestimmte Themen auf Interesse stoßen. Es wurde jedoch nicht in irgendeiner Form um Spenden gebeten. (Christlich Demokratische Union 2016, Freie Demokratische Partei 2016, Sozialdemokratische Partei Deutschlands 2016)

Deutsche Bürger können externe Plattformen nutzen, um sich über Partei- und Politikerpositionen zu Themen zu informieren und sich mit Politikern auszutauschen. Die Informationsportale arbeiten dabei mit den Politikern zusammen. Die wichtigsten bekannten Möglichkeiten sind dabei die Webseiten abgeordnetenwatch.de sowie der Wahl-O-Mat. Beide erhöhen die Transparenz politischer Entscheidungen und Programme für den Bürger. Für Parteien ermöglichen sie eine weitere Form der Vermittlung ihres Wahlprogramms und der Wählerkommunikation. Der Wahl-O-Mat wird von der

Bundeszentrale für politische Bildung seit dem Jahr 2002 betrieben. Er dient als Informationsdienst, um die Parteipositionen zu wichtigen Themen für die aktuelle Wahl zu erfahren und die Übereinstimmung mit den eigenen Positionen abzugleichen. Das Angebot ist sehr beliebt, insgesamt wurde es bisher mehr als 47 Millionen Mal genutzt. Einem Nutzer des Dienstes werden politische Thesen gezeigt. Der Nutzer kann auswählen, ob er der Aussage zustimmt, ihr neutral gegenübersteht, ihr nicht zustimmt oder sie überspringen möchte. Eine These ist beispielsweise „Die Anzahl der erforderlichen Stimmen bei Volksentscheiden soll gesenkt werden". In der Regel nehmen alle größeren Parteien, aber auch Randparteien, am Wahl-O-Mat teil. (Bundeszentrale für politische Bildung 2016)

Bei abgeordnetenwatch.de steht der einzelne Bundestagsabgeordnete im Vordergrund. Bürger können das Abstimmungsverhalten des Abgeordneten bei den Bundestagsabstimmungen nachvollziehen und dem Abgeordneten Fragen stellen. Die mit Fördergeldern finanzierte Plattform ist in elf Bundesländern aktiv. Das Portal veröffentlicht, wie die einzelnen Abgeordneten bei Entscheidungen im Bundestag abgestimmt haben. Es wird gezeigt, ob ein Politiker mit Ja oder Nein gestimmt hat, sich nicht an der Abstimmung beteiligt oder sich enthalten hat. Das Abstimmungsverhalten der Bundestagsabgeordneten Brigitte Zypries, die über die Erststimme im Wahlkreis Darmstadt in den Bundestag eingezogen ist, ist in der folgenden Abbildung dargestellt.

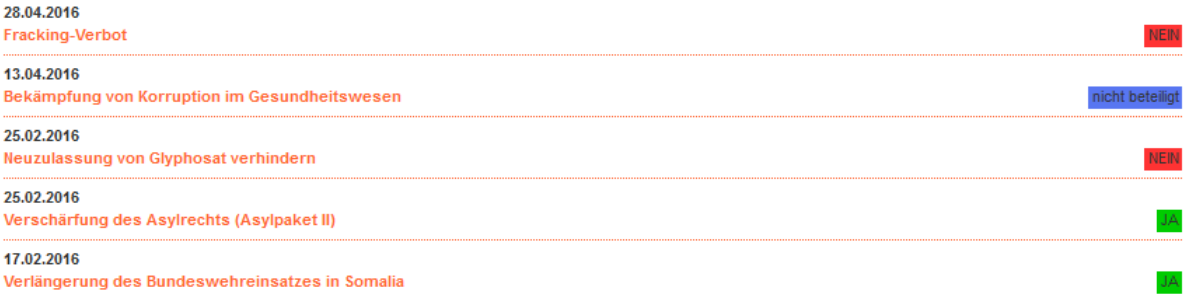


Abbildung 9 Abstimmungsverhalten der Bundestagsabgeordneten am Beispiel von Brigitte Zypries (abgeordnetenwatch.de 2016)

Im Fall von Brigitte Zypries gab es bei den aufgeführten Abstimmungen in den letzten zwei Jahren keine Enthaltungen. Die zweite Funktion des Angebots besteht darin, dass Einwohner eines Wahlkreises ihrem Wahlkreisabgeordneten oder einem anderen Politiker Fragen stellen können. Eine Beispielfrage ist in Abbildung 10 aufgeführt. Ein Politiker kann dadurch herausfinden, welche Themen die Bürger interessieren. Bisher haben über 500 Politiker Antworten verfasst. Die zehn aktivsten Politiker haben jeweils über 100 Fragen beantwortet. Dabei handelt es sich auch um richtige Antworten, da Standardantworten, bei denen lediglich auf andere Kommunikationswege verwiesen wird, gesondert gezählt werden. Durch die direkte Interaktion mit einzelnen Politikern scheint die Plattform besonders in Bezug auf Erststimmen interessant.

Frage zum Thema Frauen
Von: Monika Frank
29.08.2016

Guten Tag!

Sehr geehrte Frau Zypries, ich erbitte die Beantwortung meiner Fragen.

Als SPD MdB gehören Sie ja fast zu den Frauen um unsere Mutter des Grundgesetzes Selbert, die die Gleichberechtigung von Frau und Mann ins Grundgesetz durchsetzte. Daher frage ich Sie, wie kann es sein, dass wir vor das Grundgesetz zurückgehen und die Verschleierung von Frauen, die hier bei uns leben,zulassen? Das Grundgesetz steht über allem anderen! Also doch auch über der religösen Ausrichtung einzelner Bevölkerungsgruppen, zumal dann, wenn man bedenkt, dass das Grundgesetz die Trennung von Staat und Kirche ausdrücklich vorsieht und das gilt für alle, die in diesem Land leben wollen und sehr willkommen sind. Mich macht das so unendlich traurig, wenn ich den Kontakt zu voll- oder auch teilverschleierten Frauen suche und feststelle, dass sie durch ihr früheres Leben so geprägt wurden, so dass sie jetzt erst langsam beginnen, darüber nachzudenken, warum sie hier bei uns immer noch ihren Körper so bedecken müssen und ihre Männer helfen ihnen nicht, die meisten von ihnen wollen keine Veränderungen! Frau Zypries, wir, wer sonst müssen diesen Frauen helfen, wir haben doch alle Rechte dazu, nämlich das Grundgesetz. Wie können wir diese Errungenschaft an die Frauen weitergeben? Vermutlich nicht nur durch Gespräche!
Was würden Sie als nächste schnelle Veränderungen vorschlagen?

Vielen Dank schon einmal für Ihre Antwort und mit freundlichen Grüßen
Monika Frank

Antwort von Brigitte Zypries
08.09.2016
1 Empfehlung

Sehr geehrte Frau Frank ,

es ist nicht ganz so einfach. Wie Sie wissen darf man sich in Deutschland grundsätzlich so kleiden, wie man möchte – auch mit Ganzkörperschleier.

Juristisch betrachtet können sich Trägerinnen von Burka oder Niqab auf das im Grundgesetz verbürgte Recht ungestörter Religionsausübung berufen – dieses Grundrecht ist gar nicht durch ein einfaches Bundesgesetz einschränkbar. Ein generelles Verbot wäre verfassungswidrig und deshalb nicht nur keine schnelle Lösung, sondern überhaupt keine Lösung des Problems.

Ich finde aber, dass Sie grundsätzlich die richtige Frage stellen: Was hilft uns weiter beim Einsatz gegen die Unterdrückung von Frauen? Das ist es nämlich, was unser eigentliches Anliegen sein sollte! Zu oft wird das Thema "Burkaverbot" leider ganz anders benutzt: Als willkommenes Wahlkampfthema, mit dem man verunsicherten oder verärgerten potentiellen Wählern Entschlossenheit vorspielt. Ich finde es angebrachter, darüber nachzudenken, wie man Menschen erreicht, die sich von der Gesellschaft abgekapselt haben und Traditionen pflegen, die mit unserem Verständnis von freiheitlichem Leben nicht zusammengehen. Denn ja, ich bin Ihrer Meinung: Die Vollverschleierung ist ein Symbol der Unterdrückung von Frauen. Ich wünsche mir, dass dieses Phänomen der Vergangenheit angehört, bei uns und überall sonst auf der Welt auch. Aber: Nicht alles, was wir ablehnen, können wir verbieten.

Mit freundlichen Grüßen
Brigitte Zypries

Abbildung 10 Interaktion mit Politikern am Beispiel von Brigitte Zypries (abgeordnetenwatch.de 2016)

4.7. Zusammenfassung des Vergleichs

Zusammenfassend bestehen die wesentlichen Unterschiede zwischen den USA und Deutschland darin, dass die Parteien in Deutschland mehr staatliche Unterstützung erhalten, der deutsche Datenschutz weitreichender ist und den Parteien auch in Folge des stärkeren Datenschutzes weniger Daten, insbesondere auf individueller Ebene, zur Verfügung stehen. Die staatliche Parteifinanzierung macht in Deutschland einen großen Anteil der Parteieinnahmen aus. Um staatliche Zuwendungen zu sammeln, muss eine Partei jedoch selbst Einnahmen generieren. Dafür werden Einnahmeformen wie Mitgliedsbeiträge und Spenden genutzt. Die Finanzierung des US-Wahlkampfs basiert hingegen im Wesentlichen auf privaten Spenden, wodurch das Generieren von Spenden für einen Präsidentschaftskandidaten von zentraler Wichtigkeit ist. Aus finanzieller Sicht besteht für das massive Sammeln und Auswerten von Personendaten, wie es in den USA geschieht, also keine Notwendigkeit. Einige Aspekte wie das experimentelle Variieren von Themen und Textbausteinen bei Spendenaufrufen können aber durchaus übernommen werden. Der Erfolg von verschiedenen Maßnahmen ist in diesem Fall sehr gut nachvollziehbar. Ein fundamentaler Unterschied besteht im Datenschutz der beiden Länder. Der Schutz personenbezogener Daten in den USA ist schwach und gilt nicht für nicht-kommerzielle Organisationen. Politische Organisationen können ohne Einschränkungen riesige Mengen an Daten über Individuen sammeln. Bereits vor längerer Zeit sind kommerzielle Datenbroker entstanden, die basierend auf Wählerverzeichnissen, Daten über in der Regel 200 Millionen individuelle Bürger mit mehreren Tausend Attributen anbieten. Zusätzlich sind in den letzten Jahren Unternehmen entstanden, die Daten und Software für den informationsgetriebenen Wahlkampf anbieten. Von dieser Entwicklung ist Deutschland weit entfernt. Das Recht auf informationelle Selbstbestimmung, die in der Regel notwendige Zustimmung und Zweckbezogenheit bei der Verarbeitung personenbezogener Daten und das Prinzip der Datensparsamkeit stehen einem „Big Data für Wählerdaten“ entgegen. Ein bedeutender Unterschied besteht in der Verfügbarkeit von Wählerregistern. Diese bilden in den USA die Ausgangsbasis für die Datensammlung. Dort ist ersichtlich, ob eine Person an den letzten vier Wahlen teilgenommen hat und je nach Staat muss sogar eine Parteipräferenz angegeben werden. In Deutschland werden auf Gemeindeebene Wählerverzeichnisse geführt, die aber von Parteien nicht genutzt werden dürfen. Durch das Fehlen der Informationen von Wählerverzeichnissen ist es auch schwerer, den Erfolg einer Wählermaßnahme zu messen. In den USA kann basierend auf der Wahlteilnahme festgestellt werden, wie gut Personen auf Mobilisierungsmaßnahmen ansprechen. Außerdem wird die Information genutzt, um eine Wahrscheinlichkeit zu berechnen, mit der eine Person an der nächsten Wahl teilnehmen wird. In Deutschland kann der Erfolg von Mobilisierungsmaßnahmen höchstens indirekt aus Wahlstatistiken abgeleitet werden. Die Daten, die deutschen Parteien zur Verfügung stehen, sind selten auf individueller Ebene. Nur zu den eigenen Parteimitgliedern gibt es weitergehende Daten. Zwar gibt es mehrere Datenquellen, die Parteien einen Rahmen für ihr Handeln setzen können. Zum Beispiel gibt es in Deutschland verschiedene Adresshändler, die Konsumentendaten sammeln und auch Nutzeraktivitäten in sozialen Medien können ausgewertet werden. Jedoch ist dies nicht mit dem sehr großen Bestand an Daten für individuelle Wähler in den USA zu vergleichen, der es ermöglicht, für Individuen verschiedene Scores zu berechnen und mit ihnen individuell interagieren zu können. In Deutschland zeigt sich wie in den USA eine sinkende Wahlbeteiligung, wobei diese mit circa 70% noch vergleichsweise hoch ist. Zusätzlich ist in den letzten Jahrzehnten ein drastisches Sinken der Parteimitgliedschaften zu beobachten. Einige Entwicklungen deuten darauf hin, dass sich Parteien beim Wahlkampf hin zum US-Wahlkampf entwickeln. Zum einen betreiben deutsche Parteien Online-Communities, in denen freiwillige Wahlkampfhelfer ihre Aktionen koordinieren können. Andererseits hat beispielsweise die SPD mit Verweis auf die USA den Tür-zu-Tür-Wahlkampf übernommen. Der Vergleich der beiden Länder in dieser Arbeit unterliegt aber einer generellen Limitation. Es wurden nur öffentlich verfügbare Informationen verwendet. Parteien halten ihren Umgang mit Daten geheim. Dadurch kann nicht festgestellt werden, mit welchen Mitteln und Ergebnissen die Daten ausgewertet werden. Die genaue Datenverwendung ist nur den Parteien selbst bekannt. Die Voraussetzungen, die für die Nutzung von Daten bestehen, können jedoch gut von außen eingeschätzt werden.

5. Praktische Umsetzung eines Vorhersagemodells

Im ersten Teil der Arbeit wurde theoriebasiert erörtert, inwiefern die umfassende Datennutzung im Wahlkampf in den Vereinigten Staaten von Amerika auf das deutsche Rennen um das Bundeskanzleramt übertragbar ist. Im Folgenden wird ein Teil der bestehenden öffentlichen Daten genutzt, um selbst eine Datenauswertung vorzunehmen. Diese ist an den Unterstützung-Score aus dem US-Wahlkampf angelehnt, der für ein Individuum die Unterstützung für einen Kandidaten berechnet hat. In Deutschland gibt es jedoch nicht die gleiche Datenbasis wie in den USA und für ein Individuum kann, gerade bei ausschließlicher Verwendung öffentlicher Daten, kein solcher Score berechnet werden. Daher musste eine Ebene gewählt werden, zu der öffentliche Daten verfügbar sind. Dies ist bei Gemeinden der Fall. Demgemäß wurde ein Unterstützungs-Score auf Gemeindeebene errechnet, der aus verschiedenen Merkmalen einer Gemeinde ihre Unterstützung für eine Partei voraussagt. Das Vorgehen bei der Erstellung des prädiktiven Modells folgt dabei grob dem Vorgehensmodell CRISP-M, das ein Data-Mining-Projekt in verschiedene Phasen einteilt. Zunächst wird also die fachliche Sicht des Data-Mining-Projekts beschrieben. Danach werden Daten gesammelt und bearbeitet. Daraufhin wird das Klassifikationsproblem modelliert und ein Klassifizierer ausgewählt. Mit diesem wird auf den Daten ein Klassifikationsmodell gelernt. Das Modell wird schließlich evaluiert.

5.1. Fachliche Ziele des Data-Mining-Projekts

Zunächst werden die Ziele eines Data-Mining-Projekts festgelegt. Das Data-Mining-Projekt hat zum Ziel, mittels eines durch eine Verfahren des maschinellen Lernens erstelltes Modell zu zeigen, dass der informationsbasierte Wahlkampf in den USA in Teilen auch auf Deutschland übertragbar ist. Für die vorliegende Arbeit sind im Wesentlichen zwei verschiedene Herangehensweisen denkbar. Dabei handelt es sich um deskriptive und prädiktive Verfahren. Beide Ansätze sind für Parteien potentiell nützlich. Bei deskriptiven Modellen können Parteien herausfinden, welche Eigenschaften ihre Wählergruppen ausmachen. In früheren Zeiten war für Parteien leichter ersichtlich, von welchen Personengruppen sie gewählt werden. Vereinfacht gesagt, wählten Kirchgänger in der Regel CDU und die Arbeiterschaft die SPD. In der Gegenwart trifft eine so simple Einteilung der Wählerschaft nicht mehr zu. Daher ist es für Parteien interessant zu sehen, ob es Eigenschaften gibt, die ihre Wähler charakterisieren. Mit einem prädiktiven Modell hingegen kann aus einer Menge an Attributen ein Wert oder eine Klasse vorhergesagt werden. Für die vorliegende Arbeit wurde ein prädiktives Verfahren ausgewählt. Es wird versucht, aus den soziodemographischen Merkmalen einer Gemeinde die Stimmverteilung auf die sechs größten Parteien vorherzusagen. Zur Modellierung werden die Daten zur Bevölkerung aus dem Zensus 2011 verwendet, aus denen die Ergebnisse bei der Bundestagswahl 2013 für alle Gemeinden vorhergesagt werden sollen. Daraus kann eine Partei die Unterstützung der Gemeinde für sich und für die anderen Parteien ableiten. Theoretisch kann gegen diesen Ansatz gehalten werden, dass die Unterstützung einer Gemeinde einfach aus ihrem Wahlergebnis entnommen werden kann. Ein Vorhersagemodell hat aber zwei zusätzliche Nutzen. Zum einen ist es möglich, auch bei Veränderungen in Gemeinden, wie dem Zusammenlegen von zwei Gemeinden, die Unterstützung vorherzusagen. Zum anderen ist das Modell nicht von neuen Wahlergebnissen abhängig, um die Unterstützung einer Gemeinde zu bestimmen. Bei Änderungen in den soziodemographischen Daten kann anhand des Modells die neue Stimmenverteilung vorhergesagt werden. Die vorhergesagte Stimmverteilung hat für eine Partei also den fachlichen Zweck, ihr unabhängig von aktuellen Wahlterminen Hinweise über die Stimmverteilungen und damit den Grad an Unterstützung in allen etwa 11.000 Gemeinden in Deutschland zu geben. Ein Partei kann außerdem die Stimmverteilung für die anderen Parteien nachvollziehen und mit einem früheren Ergebnis vergleichen. Falls sich herausstellt, dass bei einer gegnerischen Partei ein hoher Anstieg der Stimmen durch veränderte soziodemographische Merkmale zu erwarten ist, kann eine Partei ihr Handeln gezielt an diese Entwicklung anpassen. Zum Beispiel können ortsbasiert negative Aspekte dieser gegnerischen Partei hervorgehoben werden oder auch die eigenen Stärken in Bereichen betont werden,

in denen die gegnerische Partei weniger Kompetenzen hat. Bezüglich der Vorhersage ist klar, dass das Wahlergebnis nicht alleine von den für die Gemeinde verwendeten soziodemographischen Merkmalen abhängig ist, sondern von zahlreichen Faktoren beeinflusst wird. Jedoch kann das eigene Vorgehen als erste Anwendung von prädiktiver Modellierung verstanden werden, die künftig ausgebaut oder mit weiteren Vorhersagemodellen ergänzt werden kann.

5.2. Erstellung der Datenbasis

Um das fachliche Ziel des Projekts zu erreichen, müssen passende Daten gesammelt werden. In einem weiteren Schritt müssen die Daten vorverarbeitet werden, damit sie in der richtigen Form für die Erstellung des Modells vorliegen. Die Phasen der Datensammlung und Datenvorverarbeitung sind meist sehr zeitaufwendig, was auch bei dieser Arbeit der Fall war. Die Datensammlung besteht aus dem Suchen nach Daten über soziodemographische Merkmale von Gemeinden und nach den Wahlergebnissen der Gemeinden. Der Schritt der Datenvorverarbeitung betrifft das Zusammenfügen der Daten, das Löschen einiger Daten, die Konstruktion passender Attribute und schließlich die Formatierung der Daten in ein Format, das zur Weiterverarbeitung genutzt werden kann.

5.2.1. Zensusdaten

Die soziodemographischen Daten für die Gemeinden liefert der Zensus 2011. Der Zensus 2011 war eine Volkszählung, die im Mai 2011 in Deutschland durchgeführt wurde. Verantwortlich für den Zensus waren die statistischen Ämter des Bundes und der Länder. Werden die im vierten Kapitel erörterten Datenquellen für deutsche Parteien berücksichtigt, so sind die Quelle der Zensusdaten die Statistikämter. Hauptziele des Zensus waren die Ermittlung der Einwohnerzahlen für Bund, Länder und Gemeinden und das Festhalten von demographischen Informationen zu der in Deutschland lebenden Bevölkerung. Mit dem Zensus werden die amtlichen Einwohnerzahlen festgestellt, die beispielsweise bei der Einteilung der Wahlkreise bedeutend sind. Dafür wurden unter anderem Haushaltsbefragungen durchgeführt, bei denen über alle Bewohner eines Haushalts Informationen ermittelt wurden. Der Zensus fand nicht in Form einer Totalerhebung statt, sondern lief registergestützt ab. Insgesamt wurden durch Haushaltsbefragungen Daten bei circa zehn Prozent der Bevölkerung in Deutschland erhoben. Die Bevölkerungsdaten für alle Gemeinden können auf der Webseite des Zensus 2011 heruntergeladen werden (Zensus 2011 2016). Der Datensatz enthält verschiedene soziodemographische Daten. Diese werden für Gesamtdeutschland, die einzelnen Bundesländer, alle Landkreise, kreisfreie Städte und Stadtkreise und für alle Gemeinden angegeben. Insgesamt gibt es 12544 Instanzen. Der Datensatz umfasst 223 Attribute, die in regionale Merkmale und Personenmerkmale unterteilt werden können. Die regionalen Daten beinhalten die Attribute 12-stelliger amtlicher Gemeindeschlüssel, Bundesland, Regierungsbezirk, Kreisfreie Stadt/Stadtkreis/ Landkreis, Gemeindeverband, Gemeinde, Gebiet und Regionalebene. Bei diesen Daten handelt es sich um nominale Daten. Die Personenmerkmale sind wiederum in verschiedene Unterbereiche unterteilt und umfassen die nachfolgenden in Spiegelstrichen aufgeführten Informationen.

- Einwohnerzahl zum 09.Mai 2011
- Bevölkerung nach Geschlecht
- Bevölkerung nach Familienstand (ausführlich) und Geschlecht
- Bevölkerung nach Alter (10er-Jahresgruppen) und Geschlecht
- Bevölkerung nach 11 Altersklassen und Geschlecht
- Bevölkerung nach Staatsangehörigkeitsgruppen
- Bevölkerung nach Geburtsland (Gruppen)
- Bevölkerung nach Religion
- Bevölkerung nach Migrationshintergrund und -erfahrung
- Personen mit Migrationserfahrung nach Zuzugsjahrzehnt
- Bevölkerung mit Migrationshintergrund nach Regionen
- Bevölkerung nach Erwerbsstatus und Geschlecht

- Erwerbstätige nach Stellung im Beruf
- Erwerbstätige Bevölkerung nach Beruf
- Erwerbstätige nach Wirtschaftszweig
- Personen in schulischer Ausbildung nach Klassenstufen
- Personen in schulischer Ausbildung nach Schulform
- Personen ab 15 Jahren nach dem höchsten schulischen Abschluss
- Personen ab 15 Jahre nach dem höchsten beruflichen Abschluss.

Die Daten erfassen demgemäß die Unterbereiche Geschlecht, Familienstand, Alter, Staatsangehörigkeit, Religion, Migrationshintergrund, Beruf und schulische Ausbildung. Es handelt sich um quantitative Daten. Jedes der Attribute bei den Personenmerkmalen gibt eine Anzahl an Personen an. Der Unterbereich enthält mehrere Attribute. Der Unterbereich Familienstand enthält verschiedene Attribute für Familienstände. Dabei wird jeder Familienstand jeweils für die Gesamtanzahl an Personen und getrennt nach Frauen und Männern angegeben. Zu jedem Familienstand gibt es dementsprechend drei Attribute. Die einzelnen Familienstände sind ledig, verheiratet, verwitwet, geschieden, in einer eingetragenen Lebenspartnerschaft lebend, eingetragener Lebenspartner verstorben, eigetragene Lebenspartnerschaft aufgehoben und ohne Angabe. Der Familienstand wird also sehr genau erfasst. Für die einzelnen Unterbereiche ist festzustellen, dass unterschiedlich genaue Informationen zur Verfügung stehen. Bei der Religionszugehörigkeit werden drei Gruppen unterschieden. Die erste Gruppe beinhaltet Menschen römisch-katholischen Glaubens, die zweite Gruppe Menschen mit evangelischem Glauben. Die dritte Gruppe fasst Angehörige sonstiger Religionen, keiner Religion und diejenigen Personen zusammen, die keine Angabe zu ihrem Glauben gemacht haben. Die dritte Gruppe ist dementsprechend sehr heterogen. Sie vereint streng religiöse Menschen und überzeugte Atheisten. Beim Erwerbsstatus wird zwischen zwei Personengruppen, den Erwerbspersonen und den Nichterwerbspersonen, unterschieden. Nichterwerbspersonen werden nicht zu den Erwerbspersonen gezählt, da sie noch schulpflichtig, berufsunfähig oder arbeitsunfähig sind. Erwerbspersonen werden in Erwerbstätige und Erwerbslose untereilt. Erwerbslose stehen dem Arbeitsmarkt zur Verfügung. Sie haben aktuell kein Arbeitsverhältnis, aber suchen danach. Erwerbstätige Personen sind diejenigen, die einer auf wirtschaftlichen Erwerb ausgerichteten Tätigkeit nachgehen (Gabler Wirtschaftslexikon 2016). Für den Bereich Migration werden als Gruppen Deutschland, EU-27-Länder ohne Deutschland, das sonstige Europa und die sonstige Welt unterschieden. Die Unterscheidung hinsichtlich nicht-europäischer Einwanderer ist damit in keiner Weise differenziert, obwohl viele unterschiedliche Kulturen davon umfasst sind. Hinsichtlich der Schulform ist eine Vergleichbarkeit der Daten zwischen den einzelnen Bundesländern nicht unbedingt vollständig gegeben. Die Regelungen zu existierenden Schulformen und Lehrplänen sind Aufgabe der Länder. Der Anspruch von gleichnamigen Schulabschlüssen variiert von Land zu Land. Neben den Unterschieden zwischen den Ländern kann es auch innerhalb eines Landes Unterschiede im Niveau eines Schulabschlusses geben, je nachdem wann dieser erreicht wurde. Diese Unterschiede werden in der Arbeit aber vernachlässigt. Attribute über Migrationshintergrund, Erwerbstätigkeit und schulische und berufliche Abschlüsse sind nur für circa 10% der Datenobjekte verfügbar. Die soziodemographischen Daten des Zensus liefern eine gute Charakterisierung der in einem Gebiet lebenden Bevölkerung.

5.2.2. Wahlergebnisse

Neben den Gemeindedaten des Zensus wurden anschließend die Wahlergebnisse der einzelnen Gemeinden für die Bundestagswahl 2013 gesammelt. Die Bundestagswahl 2013 wurde ausgewählt, da sie die letzte zurückliegende Wahl dieser Art ist. Bei Bundestagswahlen hat jeder wahlberechtigte Bürger eine Erst- und eine Zweitstimme. Die Erststimme gilt dabei der direkten Wahl eines Wahlkreiskandidaten. Die Zweitstimme dient zur Wahl einer Partei. Für die Masterarbeit wurden nur die Zweitstimmen ausgewählt. Die Erststimme kann mit der Sympathie oder der Kompetenz eines einzelnen Wahlkreiskandidaten zusammenhängen, die unabhängig von der Wahl der Partei ist. Die Zweitstimme hingegen ist nicht von einer einzelnen Person abhängig und damit bestehen über alle Gemeinden hinweg die gleichen Wahlvoraussetzungen. Die Bereitstellung der Wahlergebnisse erfolgt durch die Statistikämter der einzelnen Länder und die Landeswahlleiter. Benötigt wurde die Anzahl der Zweitstimmen aller Gemeinden eines Bundeslandes für die sechs Parteien CDU beziehungsweise CSU, SPD, FDP, LINKE, Grüne und AfD. Bei zwei Bundesländern waren die Daten in der benötigten Form jedoch nicht auffindbar. Der Grund dafür war, dass die Zweitstimmen für die AfD, die im Jahr 2013 zum ersten Mal bei einer Bundestagswahl antrat, nur unter den sonstigen Parteien aufgeführt war. Bei diesen sonstigen Parteien waren die Stimmen mehrerer Parteien mit geringem Stimmanteil zusammengefasst. Für die Arbeit war es jedoch nötig, die genaue Zweitstimmenzahl für die AfD zu verwenden. Im Falle von fehlenden öffentlichen Daten wurden die Landesämter direkt kontaktiert. Dies betraf die Landesämter für Statistik in Nordrhein-Westfalen und Sachsen. Diese stellten daraufhin die Daten in der benötigten Form bereit. Die einzelnen Datenquellen sind in der nachfolgenden Tabelle aufgeführt.

Bundesland	Datenquelle
Baden-Württemberg	(Statistisches Landesamt Baden-Württemberg 2016)
Bayern	(Bayerisches Landesamt für Statistik 2016)
Berlin	(Die Landeswahlleiterin für Berlin 2016)
Brandenburg	(Der Landeswahlleiter für Brandenburg 2016)
Bremen	(Statistisches Landesamt Bremen 2016)
Hamburg	(Statistisches Amt für Hamburg und Schleswig-Holstein 2013)
Hessen	(Hessisches Statistisches Landesamt 2016)
Mecklenburg-Vorpommern	(Landesamt für innere Verwaltung Mecklenburg-Vorpommern 2016)
Niedersachsen	(Landesamt für Statistik Niedersachsen 2016)
Nordrhein-Westfalen	Schriftliche Anfrage beim nordrhein-westfälischen Landesamt für Statistik wegen Fehlen der AfD
Rheinland-Pfalz	(Landeswahlleiter Rheinland-Pfalz 2016)
Saarland	(Die Landeswahlleiterin 2013)
Sachsen	Schriftliche Anfrage beim sächsischen Landesamt für Statistik wegen Fehlen der AfD
Sachsen-Anhalt	(Landeswahlleiterin Statistisches Landesamt Sachsen-Anhalt 2016)
Schleswig-Holstein	(Statistisches Amt für Hamburg und Schleswig-Holstein 2016)
Thüringen	(Thüringer Landesamt für Statistik 2016)

Tabelle 7 Datenquellen für die Wahlergebnisse der Bundesländer

Dass die Wahlergebnisse der Ebene der Gemeinden gewählt wurden, hat zwei Gründe. Zum einen handelt es sich um die kleinste Ebene, bei der sowohl Daten zur Soziodemographie der Bevölkerung als auch zum Wahlergebnis vorhanden sind. Es wird angestrebt, die Voraussage für ein möglichst

kleines Gebiet zu machen. Auf Seiten der Wählerstimmen gibt es mit den Wahlbezirken ein noch kleineres Gebiet als die Gemeinden. Den Wahlbezirken stehen aber keine äquivalenten Zensusdaten gegenüber. Zum anderen ist die Wahl der Gemeinden am einfachsten. Würden anstelle von Gemeinden die 299 Wahlkreise verwendet werden, so wäre ein umständliches Matching mit den Zensusdaten notwendig. Da sich die Zensusdaten nicht nach Wahlkreisen richten, hätte für alle Gemeinden geprüft werden müssen, in welchem Wahlkreis sie liegen. Überdies läge eine weitere Schwierigkeit darin, dass sich die Bildung von Wahlkreisen nicht zwingend an Stadtgrenzen orientiert und die Wahlberechtigten aus einer Stadt unterschiedlichen Wahlkreisen angehören können. Dies wird gut durch die Wahlkreise 29, 30 und 35 deutlich gemacht. Diese Wahlkreise sind „Cuxhaven – Stade II", „Stade I – Rotenburg II" und „Rotenburg I – Heidekreis". Die Städte Stade und Rotenburg sind also auf jeweils zwei Wahlkreise aufgeteilt. Eine derartige Aufteilung der Städte auf mehrere Wahlkreise, wobei ein Wahlkreis dann aus einem Teil der Stadt und einem weiteren, eigenständigen Gebiet besteht, ist in den Wahlkreisen häufig zu finden. Durch diese Aufteilung könnten die Zensusdaten den Wahlkreisen nur unzureichend zugeordnet werden.

5.2.3. Datenvorverarbeitung

Nachdem die Daten gesammelt wurden, werden diese nun vorverarbeitet. Die Vorverarbeitung beinhaltet die Schritte der Datenauswahl, Datenreinigung, Datenkonstruktion und Datenintegration. Bei den Zensusdaten wurden diejenigen Instanzen gelöscht, die keine Gemeinde waren. Zusätzlich war eine Zuordnung bestimmter Datenobjekte zueinander erforderlich. Zwischen dem Zensus 2011 und der Bundestagswahl 2013 liegt ein Zeitraum von 23 Monaten. In dieser Zeit fanden mehrere Änderungen in der Gemeindegliederung statt. Die häufigsten Gründe für eine Änderung war dabei die Eingemeindung einer Gemeinde in eine größere, bereits existierende Gemeinde oder das Zusammenschließen mehrerer Gemeinden in eine neue Gemeinde. Es kam auch vor, dass eine Gemeinde in der Zwischenzeit umbenannt wurde, ohne dass sich eine sonstige Änderung stattgefunden hat. Diese Änderungen mussten bei der Zusammenführung der Daten berücksichtigt werden. Falls drei Gemeinden zu einer neuen Gemeinde zusammengeschlossen wurden, wurden diese Gemeinden aus dem Datensatz entfernt. Für den Fall, dass eine kleine Gemeinde in eine viel größere Gemeinde eingegliedert wurde, wurde die größere Gemeinde unter der Annahme einer geringen Auswirkung im Datensatz behalten. Die Änderungen fanden überwiegend in den Bundesländern im Norden und Osten der Republik statt.

Die Attribute sind für alle Gemeinden als ganze Zahlen angegeben. Sie geben die Anzahl an Personen an, die eine bestimmte Eigenschaft, zum Beispiel „verwitweter Mann" erfüllen. Um die Gemeinden untereinander vergleichbar zu machen, wurde anstelle der Personenzahl der Anteil eines Attributs an der Einwohnerzahl einer Gemeinde angegeben. Eine Auflistung aller verwendeten Attribute ist in Anhang A zu finden. Ein Attribut konnten aufgrund der Redundanz weggelassen werden. Die Geschlechtsverteilung teilt sich in einen Anteil an Männern und einen Anteil an Frauen auf. Da dadurch der Frauenanteil aus dem Anteil an Männern direkt hervorgeht, wurde dieses Attribut weggelassen. Zwischen den Daten von „Bevölkerung nach Alter (10er-Jahresgruppen) und Geschlecht" und „Bevölkerung nach 11 Altersklassen und Geschlecht" besteht eine starke Abhängigkeit. Der später verwendete Klassifizierer hat die Eigenschaft, dass er bei möglichst unabhängigen Attributen bessere Ergebnisse als bei stark abhängigen Attributen erzielt. Daher wurden die 10er-Jahresgruppen entfernt, da diese weniger genau sind und die elf Altersklassen eine Anordnung anhand von homogeneren Personengruppen darstellen, die als Kleinkind, Kind, Teenager, junger Erwachsener usw. interpretiert werden können. Bei der Datenintegration wurden die Informationen aus den verschiedenen Tabellen zusammengefügt. Es gab eine Tabelle für die Zensus-Daten sowie jeweils eine Datei für die Wahlergebnisse in den einzelnen Bundesländern. Eine Ausnahme bildete das Land Rheinland-Pfalz, bei dem die Daten aus drei einzelnen Dateien zusammengefügt werden mussten. Insgesamt wurden die Daten also aus 19 verschiedenen Tabellen zusammengeführt. Auch hinsichtlich der Wahldaten waren einige Anpassungen nötig. Die

Zweitstimmen für die einzelnen Parteien waren als ganze Zahl angegeben. Hier wurde daher der prozentuale Anteil der Zweitstimmen einer Partei an den abgegebenen gültigen Zweitstimmen errechnet. Eine Instanz, die Gemeinde Gröde, wurde aus dem Datensatz entfernt, da bei der Bundestagswahl 2013 niemand von den neun Wahlberechtigten dort einen gültigen Stimmzettel abgegeben hat. Die Wahldaten aus den einzelnen Bundesländern unterschieden sich auch in der Angabe der Stimmen aus Briefwahlen. Das Statistikamt für Thüringen teilt die Briefwahlstimmen ihren Gemeinden zu. In der Veröffentlichung des Statistikamtes von Baden-Württemberg werden die Zweitstimmen aus der Briefwahl hingegen für mehrere Gemeinden zusammengefasst. Diese Stimmen wurden daher entfernt, da eine Zuordnung der Stimmen zu den Gemeinden nicht möglich war. Beim Datensatz von Schleswig-Holstein waren die Briefwahlstimmen von größeren Gemeinden diesen Gemeinden zugeordnet. Die Briefwahlstimmen kleinerer Gemeinden waren jedoch zusammengefasst. Die zusammengefassten Briefwahlstimmen wurden nicht berücksichtigt. Die Stimmen, die zugeordnet werden konnten, flossen in die Berechnung mit ein. Im Datensatz für das Land Brandenburg waren Urnen- und Briefwahl untereinander angeordnet, was eine Bearbeitung erleichterte. Bei den anderen Ländern waren Briefwahlstimmen in den Datensätzen nicht explizit aufgeführt. Im Laufe der Arbeit waren weitere Anpassungen der Daten nötig, die durch die gewählte Software und das Klassifikationsverfahren nötig waren. Auf diese Anpassungen wird an entsprechender Stelle eingegangen. Der Aufbau der Daten ist in der folgenden Tabelle dargestellt.

Attribut	Datentyp
Bundesland	Nominal
Einwohnerzahl	Numerisch
Männliche Bevölkerung	Numerisch, im Intervall [0,1]
Ledige Personen	Numerisch, im Intervall [0,1]
…	…
Hochschulabschluss als höchster beruflicher Abschluss	Numerisch, im Intervall [0,1]
Promotion als höchster beruflicher Abschluss	Numerisch, im Intervall [0,1]
Zweitstimmen für CDU	Numerisch, im Intervall [0,1]
Zweitstimmen für SPD	Numerisch, im Intervall [0,1]
Zweitstimmen für FDP	Numerisch, im Intervall [0,1]
Zweitstimmen für Grüne	Numerisch, im Intervall [0,1]
Zweitstimmen für Linke	Numerisch, im Intervall [0,1]
Zweitstimmen für AfD	Numerisch, im Intervall [0,1]

Tabelle 8 Aufbau des Datensatzes

Der Datensatz beinhalt insgesamt 11148 Gemeinden. Für jede Gemeinde sind 161 Attributen gespeichert. Bei allen Attributen handelt es sich, mit Ausnahme eines einzigen nominalen Attributs, um numerische Attribute. Die Attributwerte der numerischen Attribute befinden sich mit Ausnahme der Einwohnerzahl im Intervall [0,1]. Die Attribute wurden aus den Zensusdaten konstruiert und geben einen Prozentwert in Abhängigkeit einer Gesamtheit an. Das nominale Attribut erfasst das Bundesland und die einzelnen Attributwerte sind die 16 Bundesländer. Die letzten sechs Attribute stehen für die einzelnen Parteien. Für diese ist jeweils der Zweitstimmenanteil angegeben. Zu beachten ist hierbei, dass die Parteien CDU und CSU zur Vereinfachung gemeinsam als CDU aufgeführt werden.

5.3. Modellierung des Klassifikationsproblems und Modellerstellung

In diesem Abschnitt wird die Modellierung des Klassifikationsproblems als probabilistische Klassifikation vorgenommen und auf die Erstellung des Klassifikationsmodells eingegangen. Die Erstellung beinhaltet die Schritte der Auswahl der Modellierungstechnik, den Bau des Modells und dessen Bewertung. Außerdem ist eine Aufteilung und Anpassung der Daten notwendig. Zum Bau des Modells wurde das Softwaretool Weka, die Waikato Umgebung zur Wissensanalyse, in der aktuellen Version 3.9.0 genutzt. Dieses wird häufig in der akademischen Forschung verwendet. Es ist eine Sammlung von Algorithmen des maschinellen Lernens für Aufgaben des Data Mining. Als Modell wird ein Random Forest genutzt, der für die Gemeinden den Wahlausgang in Form der Stimmverteilung der Zweitstimmen voraussagt. Bei der nachfolgenden Behandlung der Funktionsweise eines Random Forest wird auch darauf eingegangen, wie dieser mittels Weka umgesetzt wird. (Hall, Frank, Holmes et al. 2009)

Modellierung des Klassifikationsproblems

Das Ziel der Klassifikation ist die Vorhersage der Stimmanteile über die sechs Parteien CDU, SPD, FDP, Grüne, Linke und AfD für die Gemeinden in Deutschland. Demgemäß sollen, basierend auf den Informationen, die für die Gemeinden zur Verfügung stehen, sechs verschiedene numerische Werte ausgegeben werden, die als Stimmanteile für die Parteien interpretiert werden können. Dies wird durch eine Modellierung des Klassifikationsproblems als probabilisitische Klassifikation erreicht. Bei einer probabilistischen Klassifikation wird für ein Datenobjekt nicht nur ein diskreter Wert als Klasse ausgegeben. Stattdessen wird eine Wahrscheinlichkeitsverteilung über alle vorhandenen Klassen ermittelt. Jeder Klasse wird so ein Wahrscheinlichkeitswert zugeordnet, der die Wahrscheinlichkeit angibt, mit der ein Datenobjekt zu dieser Klasse gehört. Die Wahrscheinlichkeit einer Klasse x liegt im Intervall $[0, 1]$. Die Summe der Wahrscheinlichkeiten über alle Klassen ergibt 1.

$$\sum_{x \epsilon X} P = (x = X) = 1$$

Dem Datenobjekt wird diejenige Klasse zugeordnet, die den höchsten Wahrscheinlichkeitswert hat. Die probabilistische Klassifikation führt also zu der Zuweisung eines diskreten Wertes, wobei zusätzlich die Information angegeben wird, mit welcher Wahrscheinlichkeit der Klassifizierer die zugeordnete Klasse für zutreffend hält. Einem Datenobjekt würde also zum Beispiel die Klasse CDU zugeordnet werden und die Information, dass diese Klasse eine Wahrscheinlichkeit von 70% hat.

Für die Masterarbeit ergibt sich ein Unterschied im Umgang mit den ausgegebenen Werten. Anders als bei der üblichen Vorgehensweise wird nicht nur der Wert mit der höchsten Wahrscheinlichkeit betrachtet und dem Datenobjekt als Klasse zugeordnet. Vielmehr sind alle Wahrscheinlichkeiten, die den einzelnen Klassen zugeordnet werden, von Bedeutung. Die Wahrscheinlichkeitsverteilung über die Klassen wird als Verteilung der Zweitstimmen über die einzelnen Parteien interpretiert. Die Wahrscheinlichkeit für eine Klasse gibt dementsprechend an, mit wie viel Prozent der Zweitstimmen eine Partei in einer Gemeinde gewählt wird. Es wird also nicht eine einzelne Zielvariable betrachtet, die als Label die einzelnen Parteien hat und einer Gemeinde genau eine Partei zuordnet. Stattdessen existieren sechs verschiedene Zielvariablen, wobei jede der sechs Parteien CDU, SPD, FDP, Grüne, Linke und AfD jeweils eine numerische Zielvariable darstellt, der ein Wert zugeordnet wird. Somit wird ein Klassifikationsverfahren dazu genutzt, um numerische Werte für insgesamt sechs Zielvariablen vorherzusagen. Um die Güte des Modells bewerten zu können, werden die ausgegebenen Werte für jede Partei mit den von ihr erhaltenen Stimmanteilen verglichen. Im Vergleich zum üblichen Vorgehen ergibt sich lediglich eine weitergehende Interpretation des Ergebnisses, die über eine reine

Betrachtung des höchsten Wahrscheinlichkeitswertes hinausgeht. Eine Klassifikation, bei der ein Klassifizierer Inputdaten mit Angabe einer diskreten Klasse benötigt, findet nach wie vor statt.
Eine Klassifikation setzt eine genau zugewiesene Klasse für jedes Datenobjekt, das zum Lernen des Modells genutzt wird, voraus. Dies ist bei den vorliegenden Daten nicht der Fall. Anstelle des Vorhandenseins genau eines Klassenlabels liegen für jedes Datenobjekt sechs verschiedene numerische Werte vor. Um die beschriebene Klassifikation durchführen zu können, müssen die Inputdaten daher in ein passendes Format gebracht werden, ohne aber die Informationen der Stimmverteilung zu verlieren. Dies wird durch eine Versechsfachung jeder Instanz, die zum Lernen des Modells genutzt wird, und dem Zuweisen von Gewichten gelöst. Jede Unterinstanz e_i hat dann ein Gewicht w_i. Das Gewicht liegt im Wertebereich $[0,1]$. Die Gesamtsumme der Gewichte w für eine Instanz e hat den Wert, der der Summe der Stimmanteile für die sechs Klassen entspricht. Wurden die sechs betrachteten Parteien in einer Gemeinde beispielsweise mit insgesamt 91 Prozent der Stimmen gewählt, so ergibt die Summe der Gewichte 0,91. Das Prinzip der Aufteilung ist in der nachfolgenden Abbildung dargestellt. Die oberste Instanz stellt die Ursprungsinstanz e dar. Sie besteht aus den Attributen und hat jeweils den Stimmanteil für alle Parteien gespeichert. Bei der Transformation wird eine Instanz dann in sechs Unterinstanzen zerlegt. Anstelle der Stimmverteilung wird jeder Unterinstanz e_i jeweils eine andere Partei als Klasse zugeordnet. Für jede Unterinstanz e_i wird zudem ein Gewicht w_i hinzugefügt. Dieses Gewicht entspricht dem Zweitstimmenanteil der Klasse.

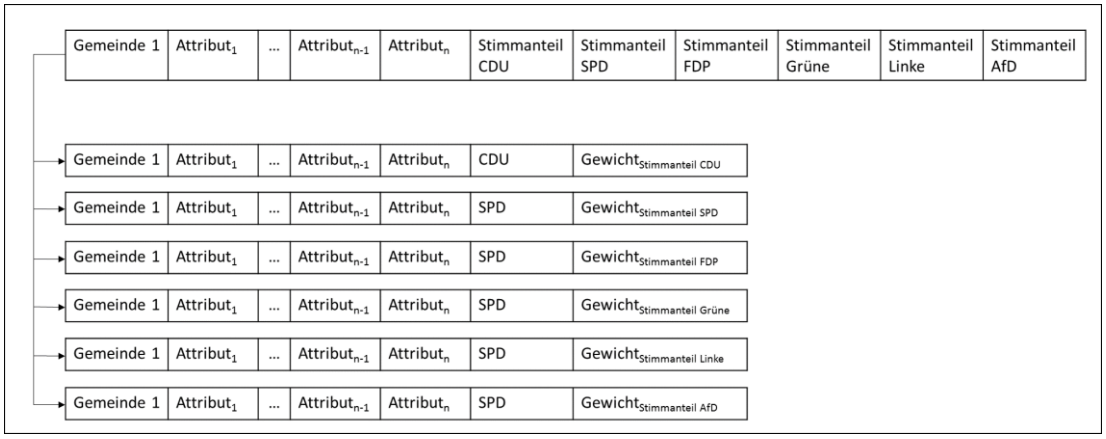


Abbildung 11 Transformation der Testdaten

Unterinstanzen, die für eine Gemeinde eine Partei mit einem hohen Stimmanteil als Klasse haben, gehen durch ihr Gewicht stärker in die Bildung des Modells ein. Instanzen, denen eine Partei mit einem geringen Stimmanteil als Klasse zugeordnet ist, gehen dementsprechend mit einem geringeren Gewicht ein. So wird gewährleistet, dass die Stimmverteilungen der Gemeinden bei der Erstellung des Modells berücksichtigt wird. In den Fällen, in denen eine Partei für eine Gemeinde null Prozent der Stimmen erhalten hat und dementsprechend mit dem Wert Null gewichtet würde, kann die Unterinstanz mit dieser Partei als Klasse weggelassen werden. Die Vervielfachung der Instanzen und Zuordnung einer jeweils anderen Partei als Klasse gewährleistet, dass jede Partei für eine Gemeinde berücksichtigt wird. Die Gewichtung stellt sicher, dass jede Partei gemäß ihres Stimmanteils in die Bildung des Modells eingeht.

Random Forests

Random Forests wurden in ihrer bestehenden Form von Leo Breiman entwickelt. Kurz gesagt, besteht ein Random Forest aus mehreren zufallsbedingt erzeugten Entscheidungsbäumen, die zusammen einen Wald bilden. Bei der Klassifikation eines Datenobjekts gibt jeder Baum eine Stimme darüber ab, zu welcher Klasse ein neues Datenobjekt seinem Modell nach zugeordnet werden soll. Das Datenobjekt wird der Klasse mit den meisten Stimmen zugeordnet. (Breiman 2001)

Die Erstellung und Funktion eines Random Forest wird durch den folgenden Algorithmus aufgezeigt.

1. Für b = 1 bis B:
   (a) Ziehe ein Bootstrap-Sample Z* der Größe N aus den Trainingsdaten
   (b) Erstelle einen Baum T_b aus den mittels Bootstrapping ausgewählten Daten durch rekursive Wiederholung der folgenden Schritte für jeden Blattknoten des Baums, bis die minimale Knotengröße n_{min} erreicht ist.
      i. Wähle zufällig m Variablen aus den p Variablen.
      ii. Wähle die beste Variable/Trennwert aus den m.
      iii. Trenne den Knoten in zwei Kindknoten auf.
2. Gib das Komitee aus Bäumen $\{T_b\}_1^B$ aus.

Um eine Klassifikation für einen neuen Punkt x vorzunehmen:
Sei $\hat{C}_b(x)$ die Klassenvorhersage des b-ten Baums.
Dann $\hat{C}_{rf}^B(x)$ = Mehrheitswahl $\{\hat{C}_b(x)\}_1^B$.

Abbildung 12 Algorithmus zur Erstellung eines Random Forest nach (Hastie, Tibshirani und Friedman 2008, S. 588)

Im Folgenden werden die einzelnen Schritte des Algorithmus näher erläutert und darauf eingegangen, wie sie in Weka implementiert sind. Die Anzahl der Bäume ist durch die Variable B festgelegt. Ein Bootstrap-Sample Z* ist eine Menge an Datenobjekten, die aus der Trainingsmenge ausgewählt wird. Dabei werden die Datenobjekte mit Zurücklegen gezogen. Das heißt, dass ein Datenobjekt, nachdem es aus der Trainingsmenge gezogen wurde und in das Bootstrap-Sample aufgenommen wurde, wieder zur Trainingsmenge zurückgelegt wird und erneut gezogen werden kann. Daher ist es möglich, dass in einem Bootstrap-Sample einige Datenobjekte mehrfach vorhanden sein können. Im Schritt 1. (b) i wird für jeden neuen Knoten eine zufällige Teilmenge der Attribute gewählt. Die übrigen Attribute werden nicht mit einbezogen. Standardmäßig werden $log_2(Anzahl\ Gesamtattribute) + 1$ Attribute ausgewählt. Das heißt, dass durch die Verwendung von Bootstrap-Samples und der zufälligen Auswahl der Attribute jeder Knoten auf Basis einer Teilmenge der vorhandenen Datenobjekte und einer Teilmenge der vorhandenen Attribute gelernt wird. Die Datenobjekte sind für jeden Knoten desselben Baumes gleich, die Attribute, abhängig vom Ergebnis der zufälligen Auswahl, in der Regel unterschiedlich. Für den Schritt 1. (b) ii. existieren verschiedene Maße, mit denen die beste Variable bzw. der beste Trennwert gewählt werden kann. In Weka wird dafür das Maß Information Gain verwendet. Der Information Gain nimmt die Trennung der Daten auf Basis der Entropie vor. Die Entropie gibt die Ordnung für eine Klassenverteilung an. Bei einer Gleichverteilung der Klassen hat die Entropie den Wert 1. Wenn die Instanzen alle dieselbe Klasse haben, hat die Entropie den Wert 0. Eine Datenmenge wird also an der Stelle geteilt, für die die Entropie am geringsten ist. Für die gegebenen Daten kann jede Instanz sechs verschiedenen Klassen, den sechs verschiedenen Parteien, zugeordnet werden.

Die Entropie für eine Klassenverteilung S wird durch die folgende Formel berechnet. Der Anteil einer Klasse i an den Gesamtdaten wird durch p_i angegeben.

$$E(S) = -\sum_{i=1}^{6} p_i \log_2 p_i$$

Nach der Berechnung der Entropie der Klassenverteilung S, wird für jedes Attribut A die durchschnittliche Entropie berechnet.

$$I(S,A) = \sum_i \frac{|S_i|}{|S|} E(S_i)$$

Aus diesen Formeln kann der Information Gain errechnet werden, der sich durch die Subtraktion der durchschnittlichen Entropie der Klassenverteilung eines Attributs von der Entropie der Klassenverteilung ergibt.

$$Information\ Gain\ (S,A) = E\ (S) - I\ (S,A)$$

Die Datenmenge wird dann an demjenigen Attributwert, der den größten Information Gain hat, in zwei Teile getrennt. Die Wahl des höchsten Information Gain sorgt für eine größtmögliche Minimierung der Entropie. (Quinlan 1986)

Bei nominalen Attributen wird die Datenmenge direkt anhand eines ihrer Attributwerte in Untermengen getrennt. Beim Attribut Bundesländer werden die Instanzen einer Datenmenge danach aufgeteilt, ob sie dem Bundesland mit dem höchsten Information Gain angehören oder nicht. Bei Random Forests ist festgeschrieben, dass ein Knoten nur zwei Kindkonten haben darf. Bei Erstellung von Bäumen außerhalb von Random Forests kann ein Knoten, der auf Basis eines nominalen Attributs unterteilt wird, mehr als zwei Kindknoten haben. Bei numerischen Daten erfolgt die Teilung anhand eines numerischen Trennwerts. Die Aufteilung in Untermengen erfolgt danach, ob ein Attributwert kleiner oder gleich oder größer oder gleich als der Trennwert ist.

Die in Schritt 1 erstellten Entscheidungsbäume bilden ein sogenanntes Komitee, das die Klassifizierung von Datenobjekten vornimmt. Für ein unbekanntes Datenobjekt, das klassifiziert werden soll, gibt jeder Baum eine Stimme darüber ab, zu welcher Klasse das Datenobjekt gehört. Die Stimmen der einzelnen Bäume werden dann gemittelt. Daraus ergibt sich für jede Klasse ein Wahrscheinlichkeitswert, mit dem das Datenobjekt zu der Klasse gehört. Das Datenobjekt wird der Klasse mit dem höchsten Wahrscheinlichkeitswert zugeordnet.

Die Generalisierbarkeit des Verfahrens, also die Übertragbarkeit auf unbekannte Daten, hängt von zwei Faktoren ab. Der erste Faktor ist die Stärke der individuellen Bäume. Der andere Faktor ist die Korrelation zwischen den einzelnen Bäumen. Diese sollte so niedrig wie möglich sein. Beim Erlernen eines Random Forest ist darüber hinaus das sogenannte Bias-Varianz-Dilemma bedeutend. Der Bias ist der Fehler, der durch ein schlechtes Modell verursacht wird. Der Bias wird erhöht, wenn bei der Erstellung der Entscheidungsbäume zu wenige Attribut verwendet werden. Das Modell stellt dann eine nicht zutreffende Beziehung zwischen den Trainingsdaten und ihren Klassen her. Die Varianz ist der Fehler, der durch die zur Verfügung stehenden Daten verursacht wird. Fehler entstehen, wenn der Algorithmus auf die Daten überangepasst wird und damit das Rauschen in den Daten zu stark abbildet. Sowohl ein hoher Bias als auch eine hohe Varianz sollten vermieden werden, da sie die Qualität der Voraussage verschlechtern. Sie sind aber voneinander abhängig. Modelle mit einer hohen Varianz haben häufig einen niedrigen Bias und umgekehrt. Diese Beziehung ist auch für Random Forests bedeutend. Die Anzahl der ausgewählten Attribute beeinflusst den Bias. Wenige Attribute pro Baum bedeuten einen höheren Bias im Vergleich zu mehr Attributen. Dadurch, dass mehrere Klassifizierer und nicht nur ein einzelner Entscheidungsbaum gelernt wird, wird der Bias gesenkt. Die

Konstruktion mehrerer Datensätze senkt die Varianz, da das potentielle Rauschen des Gesamtdatensatzes durch die Verwendung von verschiedenen zufälligen Bootstrap-Samples ausgeglichen wird.

Unterteilung in Trainings- und Testmenge

Der zuvor erstellte Datensatz muss zunächst in eine Trainings- und eine Testmenge unterteilt werden. Zwei Drittel der Beispiele werden als Trainingsmenge genutzt und das übrige Drittel bildet die Testmenge. Die Aufteilung in die Trainings- und Testmenge sollte dabei stratifiziert sein. Das bedeutet, dass die Wahrscheinlichkeitsverteilungen in den beiden Datensätzen gleich sein sollten. Eine nicht-stratifizierte Aufteilung erhöht die Varianz in den Daten. Dies kann zu folgender beispielhafter Situation führen: Wenn die Testmenge nur Beispiele enthält, in denen die CDU sehr schlecht abschneidet, wird das gelernte Modell Schwierigkeiten bei der Klassifizierung von Beispielen haben, die einen hohen CDU-Stimmanteil haben. Bei einem einzelnen Klassenattribut ist die Stratifizierung einfach. Die Datenmenge wird dann gemäß der Attributwerte für das Klassenattribut aufgeteilt, wobei das Verhältnis der einzelnen Attributwerte beibehalten wird. Im Falle einer probabilistischen Verteilung über mehrere Klassenattribute gestaltet sich die Aufteilung schwierig. Als Lösung wurde die verhältnismäßige Aufteilung nach dem Klassenranking gewählt. Dafür werden die Klassenattribute in eine Reihenfolge beginnend vom höchsten bis zum niedrigsten Wert gebracht. Die Partei mit den meisten Stimmen ist also auf Platz eins, die Partei mit den zweitmeisten Stimmen auf Platz zwei und so weiter. Durch das Bilden der Reihenfolge können die Datenobjekte in Gruppen eingeteilt werden. Jedes mögliche Ranking stellt eine Gruppe dar. Die Anzahl der Gruppen beträgt 103. Die verschiedenen Ranking-Gruppen sollen in der Trainings- und Testmenge jeweils das Verhältnis von 2:1 aufweisen. Beim Bilden der Rankings stellt sich heraus, dass circa 10% der Datenobjekte für zwei oder mehr Klassenattribute den gleichen Wert haben. Dies passiert, wenn Parteien eine identische Anzahl an Stimmen erhalten haben. Damit nehmen mehrere Klassen den gleichen Platz ein und es ist keine klare Reihenfolge gegeben. In diesem Fall wurde zur Vereinfachung die Reihenfolge der Klassen eingehalten, welche auch dem Ergebnis der Bundestagswahl entspricht. Die Reihenfolge der Bundestagswahl ist von oben absteigend CDU, SPD, Linke, Grüne, FDP, AfD. Diese Vereinfachung dient dazu, die ohnehin schon hohe Anzahl von 103 verschiedenen Rangfolgen nicht weiter zu erhöhen. Außerdem wird angenommen, dass der mögliche negative Effekt vernachlässigbar ist. Eine Ausnahme bilden drei Beispiele, bei denen von den sechs Parteien nur die CDU gewählt wurde. Bei diesen werden zwei Beispiele der Trainings- und eines der Testmenge zugeordnet. Die Häufigkeit der unterschiedlichen Rangfolgen kann Anhang B entnommen werden.

Dateiformat

Nachdem die Unterteilung der Daten erfolgt ist, müssen die Trainings- und Testdaten in das von Weka verwendete Arff-Format gebracht werden. Arff steht dabei für attribute-relation file format. Die Umwandlung ist schnell erledigt, da sich eine Arff-Datei bis auf einige zusätzlich erforderliche Angaben nicht von einer csv-Datei unterscheidet. Neben einer Bezeichnung für den Datensatz mit @relation zu Dokumentbeginn werden alle Attribute mit @attribute und dem passenden Datentyp annotiert. Vor die Auflistung der Instanzen wird die Annotation @data gesetzt. Wie zuvor beschrieben, findet bei den Trainingsdaten eine Anpassung der Daten statt, die aus einer Versechsfachung jeder Instanz, dem Zuordnen einer jeweils anderen Partei zu jeder Instanz und dem Hinzufügen eines Gewichts besteht. In der arff-Datei kann den einzelnen Instanzen ein Gewicht zugeordnet werden, indem es in geschwungenen Klammern an das Ende der Zeile geschrieben wird. Abbildung 13 zeigt eine stark reduzierte Version der verwendeten Arff-Datei für die Trainingsmenge. Es ist erkennbar, dass eine Vervielfachung der Instanzen stattgefunden hat, die sich nur in der Klasse und dem Gewicht unterscheiden. Die tatsächliche Arff-Datei umfasst durch die Vervielfachungen mehr als 44500 Instanzen.

```
@relation Trainingsmenge

@attribute Einwohnerzahl numeric
@attribute Alter11_15_17Insg numeric
@attribute StaatEU27 numeric
@attribute Klasse {CDU, SPD, FDP, GRUENE, LINKE, AFD}

@data
2345, 0.04234, 0.09322, CDU, {0.543071}
2345, 0.04234, 0.09322, SPD, {0.11985}
2345, 0.04234, 0.09322, FDP, {0.018727}
2345, 0.04234, 0.09322, GRUENE, {0.007491}
2345, 0.04234, 0.09322, LINKE, {0.164794}
2345, 0.04234, 0.09322, AFD, {0.06367}
879, 0.03929, 0.12429, CDU, {0.380089}
879, 0.03929, 0.12429, SPD, {0.340126}
```

Abbildung 13 Verkürzte Arff-Datei der Trainingsmenge

Bezüglich der Testdaten besteht im Vergleich zu den Trainingsdaten ein geringer Unterschied. Bei der Verwendung von Weka ist es notwendig, dass die Testdaten exakt den gleichen Aufbau wie die Trainingsdaten haben. Die Testdaten benötigen dieselbe Anzahl an Attributen. Die Attribute müssen auch vom selben Namen und Typ sein wie bei den Trainingsdaten. Daher kann in der arff-Datei der Testdaten nicht die Wahrscheinlichkeitsverteilung über die Parteien abgespeichert werden. Deshalb wird der Aufbau der Trainingsdaten übernommen und an die Stelle der Partei wird ein Fragezeichen eingefügt. Das Fragezeichen bedeutet, dass die Klasse unbekannt ist. Eine Zuordnung von Gewichten ist bei den Testdaten nicht notwendig. Der Aufbau der Testdaten ist in verkürzter Form in Abbildung 14 dargestellt.

```
@relation Testmenge

@attribute Einwohnerzahl numeric
@attribute Alter11_15_17Insg numeric
@attribute StaatEU27 numeric
@attribute Klasse {CDU, SPD, FDP, GRUENE, LINKE, AFD}

@data
72045, 0.06855, 0.01232, ?
98365, 0.08992, 0.12429, ?
```

Abbildung 14 Verkürzte Arff-Datei der Testmenge

Parameter

Nachdem die Vorbereitung der Daten abgeschlossen ist, wird die Trainingsmenge verwendet, um ein Vorhersagemodell zu lernen. In Weka können bei der Erstellung eines Random Forest verschiedene Parameter eingestellt werden, wobei auf die wichtigsten Parameter kurz eingegangen werden soll. Diese sind die Anzahl und Tiefe der Bäume und die Anzahl der verwendeten Attribute pro Knoten. Bezüglich der Anzahl an verwendeten Bäumen kann gesagt werden, dass eine höhere Anzahl an Bäumen zu einem geringeren Klassifikationsfehler führt. Dieser Effekt hält aber nur bis zu einer bestimmten Zahl an Bäumen an, danach verbleibt der Fehler im Wesentlichen auf einem konstanten Niveau und sinkt nicht weiter. Die Tiefe des Baums gibt an, wie viele Tests maximal durchlaufen werden müssen, bis ein Wurzelknoten erreicht wird. Bäume ohne eine Beschränkung der Tiefe können sehr groß werden und zu einem Overfitting der Daten und damit einer schlechten Generalisierbarkeit neigen. Die Anzahl der für jeden Knoten verwendeten Features ist standardmäßig

$log_2(Anzahl\ Gesamtattribute) + 1$. Im Allgemeinen führt ein Random Forest zu besseren Ergebnissen, wenn die von einem Baum verwendeten Attribute eine möglichste geringe Korrelation haben. Neben theoretischen Überlegungen und einem praktischen Ausprobieren verschiedener Parameterkombinationen ist die Wahl der Parameter durch die vorhandenen Rechenkapazitäten limitiert. Je mehr Bäume berechnet werden müssen und je tiefer diese Bäume sind, desto mehr von der begrenzten Rechenkapazität muss zur Erstellung des Random Forest verwendet werden.

Modelle

Es wurden 19 verschiedene Modelle erstellt, die sich in der Anzahl und Tiefe der Bäume unterscheiden. Die Anzahl der verwendeten Attribute wurde nicht verändert, da sich diese in der Anwendung bewährt hat. Die nachfolgende Tabelle zeigt die erstellten Modelle mit ihren Parametern.

Name des Modells	Baumanzahl	Baumtiefe
Modell_01001	100	1
Modell_01002	100	2
Modell_01003	100	3
Modell_01004	100	4
Modell_01005	100	5
Modell_01006	100	6
Modell_01007	100	7
Modell_01008	100	8
Modell_01009	100	9
Modell_10010	100	10
Modell_10015	100	15
Modell_10020	100	20
Modell_10025	100	25
Modell_10050	100	50
Modell_10088	100	Unbegrenzt
Modell_05015	50	15
Modell_20015	200	15
Modell_25015	250	15
Modell_25088	250	Unbegrenzt

Tabelle 9 Erstellte Modelle und ihre Parameter

Alle in der Tabelle aufgeführten Modelle stellen jeweils einen Random Forest dar, der mit einer bestimmten Parameterkombination gelernt wurde. Die Baumanzahl reicht dabei von 50 bis 250 Bäumen. Die Tiefe der Bäume beginnt bei 1 und reicht bis zu einer theoretisch unbegrenzten Tiefe. Der Name eines Modells gibt jeweils dessen Konfiguration an. Zuerst wird dabei die Anzahl der Bäume genannt und danach ihre Tiefe. Bei einer unendlichen Tiefe wurde die Zahl 88 gewählt, da die 8 dem Unendlichkeitssymbol ähnelt, aber eine acht alleine schon zur Angabe der Tiefe von acht verwendet wird. Damit alle Modellnamen alle gleich lang sind, wurde bei kürzeren Namen eine führende null eingefügt. Das Modell „Modell_05015“ steht dementsprechend für die Anzahl von 50 Bäumen mit einer maximalen Tiefe von 15.

5.4. Evaluierung der Modelle

Die Evaluierung der Modelle erfolgt zunächst auf der Trainingsmenge und anschließend auf der Testmenge. Hinsichtlich der Trainingsmenge kann allerdings keine aussagekräftige Evaluierung stattfinden. Daher wird nur kurz auf eine Besonderheit eingegangen, die sich bei der Bewertung des Modells auf der Trainingsmenge ergibt. Zur Evaluierung der Modelle auf der Testmenge werden verschiedene Fehlermaße berechnet. Nachdem exemplarisch einige Ergebnisse für einzelne Gemeinden aufgezeigt werden, werden die durchschnittlichen Klassifikationsfehler der Modelle über alle Gemeinden hinweg berechnet. Danach wird ein näherer Blick auf die Klassifikationsgenauigkeit der Modelle bezüglich der einzelnen Parteien geworfen.

Evaluierung der Ergebnisse auf der Trainingsmenge

Bezüglich der Bewertung auf der Trainingsmenge ist es dem Modell praktisch nicht möglich, eine Vorhersagegenauigkeit von 100% zu erreichen, selbst wenn das Modell einfach alle Beispiele auswendig lernen würde. Bei den Trainingsdaten ist jeweils sechs identischen Unterinstanzen e_i eine andere Klasse zugeordnet. Das Modell müsste, um jede Unterinstanz richtig vorherzusagen, für sechs identische Datenobjekte sechs Mal eine andere Klasse vorhersagen. Dieses Vorhersageverhalten wäre unlogisch. Aufgrund dieses Umstands der sich widersprechenden Trainingsmenge kann eine Evaluierung des Modells auf der Trainingsmenge nur eingeschränkt stattfinden.

Wie zuvor beschrieben, wird jedem Datenobjekt ein Gewicht zugeordnet. Die Datenobjekte gehen jeweils mit einem unterschiedlich hohen Gewicht in das Modell ein. Die Summe der Gewichte beträgt bei der CDU 3410,42; für die SPD 1786,38; bei den Linken 615,11 und für die Grünen 481,35. Bei der AfD beläuft sich der Wert auf 363,46 und bei der FDP auf 341,01. Die Gewichtung ähnelt der Verteilung des Bundesergebnisses, ist aber nicht ganz identisch. Aus der Summe der Gewichte und ihrem Verhältnis zueinander lässt sich ableiten, dass die Stimmverteilung durch die Stratifizierung angenähert, aber nicht perfekt wiedergegeben wurde.

Da die Instanzen mit der Klasse FDP nur 4,9% des Gesamtgewichts ausmachen, wie es auch nahezu ihrem Wahlergebnis entspricht, wäre es ein Zeichen eines schlechten Modells, wenn 16,7% der Beispiele als FDP klassifiziert würden, was der Modellierung nach geschehen müsste. Genauso verhält es sich mit den anderen Parteien, denen nach der Modellierung jeweils ein Sechstel der Instanzen zugeordnet werden müssten, was aber nicht ihrer Gewichtung entspricht. Es besteht also nicht nur hinsichtlich der Widersprüchlichkeit der Klassenzuordnung, sondern auch durch die sehr unterschiedliche Gewichtung der einzelnen Klassen eine Einschränkung für die Bewertung des Modells auf der Trainingsmenge. Eine aussagekräftige Ergebnisevaluierung kann dementsprechend bei der Trainingsmenge wenn überhaupt nur grob durchgeführt werden. Die tatsächliche Evaluierung, bei der die verschiedenen Modelle untereinander mit aussagekräftigen Ergebnissen verglichen werden, kann nur bei der Testmenge stattfinden.

Die Modelle wurden mittels einer zehnfachen Kreuzvalidierung evaluiert. Dafür wird die Trainingsmenge in zehn Untermengen aufgeteilt. Von diesen zehn Untermengen werden dann neun Untermengen genutzt, um eine Modell zu lernen. Dieses Modell wird auf der zehnten, nicht zum Lernen verwendeten, Untermenge getestet. Dieser Vorgang wird weitere neun Male wiederholt, bis jede der Untermengen einmal zum Testen des Modells verwendet wurde. Jede Instanz wird auf diese Weise neun Mal zum Lernen und einmal zum Testen des Modells verwendet. Aus jeder Kreuzvalidierung ergeben sich bestimmte Fehlerwerte. Schließlich werden die Fehlerwerte gemittelt und es ergibt sich die Bewertung des Modells.

Eine Bewertung kann anhand einer Konfusionsmatrix stattfinden. Diese zeigt auf, wie die einzelnen Datenobjekte klassifiziert wurden. Das Prinzip von Konfusionsmatrizen soll kurz beispielhaft dargestellt werden. Angenommen wird ein Beispiel, in dem es 6000 Datenobjekte gibt, wobei jeweils

1000 Datenobjekten eine andere Partei als Klasse zugeordnet ist. Auf diesen Daten wird das Ergebnis eines Modells evaluiert. Zwei beispielhafte Konfusionsmatrizen sind in Abbildung 15 dargestellt. Die Werte sind erfunden und dienen allein zur Veranschaulichung.

Klasse \ Klassifiziert als	CDU	SPD	Grüne	LINKE	FDP	AfD
CDU	1000	0	0	0	0	0
SPD	0	1000	0	0	0	0
Grüne	0	0	1000	0	0	0
LINKE	0	0	0	1000	0	0
FDP	0	0	0	0	1000	0
AfD	0	0	0	0	0	1000

Klasse \ Klassifiziert als	CDU	SPD	Grüne	LINKE	FDP	AfD
CDU	800	0	100	0	100	0
SPD	0	1000	0	0	0	0
Grüne	0	0	1000	0	0	0
LINKE	0	0	0	1000	0	0
FDP	100	0	0	0	700	200
AfD	0	100	0	0	200	700

Abbildung 15 Bewertung einer Klassifikation mit Konfusionsmatrizen

Jede Zeile gibt die Anzahl der Datenobjekte für eine Klasse an. Jede Spalte gibt für eine Klasse an, wie viele Datenobjekte dieser Klasse zugeordnet wurden. Daraus lässt sich ablesen, wie die einzelnen Datenobjekte klassifiziert wurden. Bei der linken Konfusionsmatrix wurden alle Beispiele richtig klassifiziert. Zum Beispiel gibt es 1000 Datenobjekte mit der Klasse FDP, die alle der richtigen Klasse zugeordnet wurden. In der rechten Konfusionsmatrix ist zu sehen, dass einige Beispiele falsch klassifiziert wurden. Bei den Datenobjekten mit der Klasse CDU wurden 800 Datenobjekte richtig klassifiziert. 100 Datenobjekten wurde jedoch fälschlicherweise die Klasse Grüne zugeordnet und weiteren 100 Datenobjekten wurde die Klasse FDP zugewiesen. Bei den Datenobjekten mit den Klassen FDP und AfD kam es ebenfalls zu falschen Klassifikationen, bei den Parteien Grüne und Linke verlief die Klassifikation korrekt. Diese Matrix kann bei der Bewertung des Modells verwendet werden, wenn es auf die Trainingsdaten angewendet wird. Zur Evaluierung der Testdaten kann eine derart aufgebaute Konfusionsmatrix nicht verwendet werden, da es bei den Testdaten nicht um die Zuordnung zu genau einer Klasse geht, sondern eine Klassenverteilung zu bewerten ist.

Die folgende Konfusionsmatrix zeigt das Ergebnis für das Modell_10025. Das Modell hat eine Baumtiefe von 25 und eine Baumanzahl von 100.

Klasse \ Klassifiziert als	CDU	SPD	Grüne	LINKE	FDP	AfD
CDU	34,59	2530,13	161,85	589,44	62,72	31,67
SPD	1617,11	0,34	78,40	48,92	24,65	16,96
Grüne	400,52	72,15	0,00	5,51	2,41	0,77
LINKE	522,26	84,56	1,98	0,00	0,81	5,5
FDP	286,92	48,10	2,30	3,30	0,00	0,40
AfD	303,24	48,11	2,00	9,51	0,49	0,00

Abbildung 16 Konfusionsmatrix der Trainingsdaten bei Modell_10025

Als erstes fällt auf, dass keine ganzen Zahlen, sondern Werte mit Nachkommastellen angegeben sind. Dies ist durch die zuvor durchgeführte Gewichtung der Instanzen zu erklären. Eine Instanz geht nicht mit dem Wert eins in die Konfusionsmatrix ein, sondern mit dem Gewicht, das ihr zuvor zugeteilt wurde. Die Instanzen einer Klasse sind jeweils über eine Zeile verteilt. Aus der Summe der Werte einer Zeile ergibt sich ihr Gesamtgewicht. Dies soll kurz anhand der AfD dargelegt werden. Das Gesamtgewicht der Instanzen mit der Klasse AfD ist 363,46. Davon wurden Instanzen mit dem Gesamtgewicht von 303,24 der Klasse CDU zugeordnet. Instanzen mit dem summierten Gewicht von 48,11 wurden der SPD zugeordnet. Bei der FDP beträgt der Wert 0,49; bei den Grünen 2,00; bei den Linken 9,51 und bei der AfD selbst 0,00. Durch diese Darstellung kann die Verteilung der Gewichte

sehr gut beurteilt werden. Demgegenüber kann keine konkrete Aussage über die Anzahl der richtig und falsch klassifizierten Instanzen gemacht werden.

Die Analyse der zu den Modellen gehörigen Konfusionsmatrizen führt zu mehreren Beobachtungen, wovon eine erwartet und die anderen unerwartet sind. Bei Modellen mit einer niedrigen Baumtiefe wird ausschließlich die CDU vorhergesagt. Bei einer Erhöhung der Baumtiefe werden zunehmend auch die anderen Klassen vorhergesagt wie die obige Konfusionsmatrix zeigt. Diese Beobachtung ist nicht verwunderlich, da Bäume mit zunehmender Tiefe immer mehr Attribute berücksichtigen und differenziertere Entscheidungen treffen. Diese weichen dann davon ab, immer die am höchsten gewichtete Klasse zu wählen. Unerwartet hingegen ist die Tatsache, dass eine Instanz fast nie richtig ihrer eigenen Klasse zugeordnet wird. Dies ist für alle Klassen zu beobachten. Die Instanzen, denen die vier kleineren Parteien zugeordnet sind, werden in keinen Fall richtig zugeordnet. Die Werte in den Feldern, in denen die Gewichte der Instanzen stehen, die die Klasse Grüne, Linke, FDP oder AfD haben und auch als solche vorhergesagt werden, sind jeweils null. Bei den Parteien CDU und SPD sind die Werte an den entsprechenden Stellen höher als null. Dennoch haben sie von allen möglichen Zuordnungen das jeweils geringste Gewicht. Zudem ist zu erkennen, dass für alle Klassen, mit Ausnahme der CDU selbst, die CDU mit dem mit Abstand höchsten Gewicht vorhergesagt wird. Die Instanzen mit der Klasse CDU werden bevorzugt der SPD zugeordnet, gefolgt von den Linken. Obwohl die Klassifikation unter widersprüchlichen Bedingungen stattfindet, ist das Ergebnis dennoch überraschend. Dies betrifft zum einen die Tatsache, dass Datenobjekte mit der Klasse CDU bevorzugt der SPD zugeordnet werden und dass nur in den seltensten Fällen überhaupt eine Klasse richtig zugeordnet wird.

Normalisierung

Bevor die Ergebnisse der Modelle auf der Testmenge evaluiert werden können, muss eine Normalisierung der Daten in der Testmenge stattfinden. Der Random Forest gibt die Wahrscheinlichkeitsverteilung über die Klassen normiert auf 100% an. Ein direkter Vergleich mit den Wahlergebnissen würde zu einer fehlerhaften Einschätzung führen, da die Summe der Stimmanteile bei den Testdaten nicht unbedingt 100% beträgt. Die betrachteten sechs Parteien haben bei der Bundestagswahl 93.7% der Stimmen erhalten. Daher müssen die Testdaten auf 100% normalisiert werden. Um einen normalisierten Stimmanteil a_x* zu erhalten, muss für jede Instanz jeder Stimmanteil a für eine Partei x mit einen Wert multipliziert werden, der von der Summe der Stimmanteile der sechs Parteien abhängt.

$$normalisierter\ Stimmanteil\ a_x *= \ a_x \times (\,2 - \ (\sum_{i=1}^{6} a_i))$$

Dieser Wert ist dann direkt mit dem Wert vergleichbar, der vom Random Forest ausgegeben wird. Diese nachträgliche Normalisierung ist einer Normalisierung der Trainingsmenge vorzuziehen. Eine Normalisierung der Trainingsmenge auf 100% würde zu Verzerrungen führen, deren Stärke vom Stimmanteil für die großen Parteien abhängt. Dann würden Parteien den höchsten Stimmzuwachs bekommen, wenn sehr viele Stimmen an die kleineren Parteien gegangen sind. Im Extremfall würden die großen Parteien mit 1% der Stimmen gewählt und die nicht betrachteten kleinen Parteien mit 99%. Bei einer Normalisierung der Daten würde dieses 1% zu 100% werden. Dies führt zu einer Verzerrung und ist aus fachlicher Sicht nicht sinnvoll. Durch die nachträgliche Normalisierung kann diese Verzerrung umgangen werden und es besteht für alle Instanzen die gleiche Vergleichsgrundlage. Anstelle der Testdaten könnten auch die ausgegebenen Werte des Random Forest angepasst werden. Dies würde allerdings einen höheren Aufwand bedeuten, da bei den Testdaten nur einmal normalisiert werden muss und diese Werte mit den Ausgaben aller Modelle verglichen werden können. Wären die Ausgaben des Random Forest angepasst worden, so wäre dies für jedes Modell erneut notwendig gewesen.

Anwendung des Modells auf die Testmenge

Die mit den Trainingsdaten gelernten Modelle werden nun zur Klassifikation der Testmenge verwendet. Als Ergebnis wird eine Wahrscheinlichkeitsverteilung ausgegeben. Diese gibt an, für wie wahrscheinlich es der Random Forest hält, dass eine Instanz zu einer bestimmten Klasse gehört. Der Instanz wird diejenige Klasse zugeordnet, die den höchsten Wahrscheinlichkeitswert hat. Zur Bewertung der Klassifikation können verschiedenen Fehlermaße verwendet werden. Für die Masterarbeit werden die Fehlerwerte berechnet, die üblicherweise auch von Weka verwendet werden. (Witten, Frank und Hall 2011, S. 180 ff.)

Dabei handelt es sich um den mittleren quadratischen Fehler, den relativen quadratischen Fehler, den durchschnittlichen absoluten Fehler, den relativen absoluten Fehler, die Wurzel des mittleren quadratischen Fehlers und die Wurzel des relativen quadratischen Fehlers. Diese Maße beziehen sich auf die für die Testinstanzen vorhergesagten Klassenwerte p_1, p_2, …, p_n und ihre tatsächlichen Klassenwerte a_1, a_2, …, a_n. Die Variable $\bar{a}$ gibt das arithmetische Mittel der tatsächlichen Klassenwerte an.

$$mittlerer\ quadratischer\ Fehler = \frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}$$

Der mittlere quadratische Fehler misst das Mittel der quadrierten Differenzen zwischen den vorhergesagten und tatsächlichen Werten.

$$relativer\ quadratischer\ Fehler = \frac{(p1 - a1)^2 + \cdots + (pn - an)^2}{(a1 - \bar{a})^2 + \cdots + (an - \bar{a})^2}$$

Der relative quadratische Fehler gibt den Fehler in Abhängigkeit zum Fehler eines Klassifizierers an, der jedem Attribut einfach den Mittelwert der tatsächlichen Werte zugeordnet hätte.

$$durchschnittlicher\ absoluter\ Fehler = \frac{|p1 - a1| + \cdots + |pn - an|}{n}$$

Der durchschnittliche absolute Fehler mittelt die Höhe der individuellen Fehler.

$$relativer\ absoluter\ Fehler = \frac{|p1 - a1| + \cdots + |pn - an|}{|a1 - \bar{a}| + \cdots + |an - \bar{a}|}$$

Der relative absolute Fehler gibt den Fehler in Abhängigkeit zum Fehler eines Klassifizierers an, der jedem Attribut einfach den Mittelwert der tatsächlichen Werte zugeordnet hätte. Er verhält sich zum durchschnittlichen absoluten Fehler wie der relative quadrierte Fehler zum mittleren quadratischen Fehler.

$$Wurzel\ des\ mitteleren\ quadratischen\ Fehlers = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (pn - an)^2}{n}}$$

Die Wurzel des mittleren quadratischen Fehlers gewichtet große Unterschiede zwischen den tatsächlichen und vorhergesagten Wert in höherem Maß als kleine Unterschiede. Dies ist bei den vorherigen Fehlermaßen nicht der Fall.

$$Wurzel\ des\ relativen\ quadratischen\ Fehlers = \sqrt{\frac{(p1-a1)^2+\cdots+(pn-an)^2}{(a1-\bar{a})^2+\cdots+(an-\bar{a})^2}}$$

Die Wurzel des relativen quadratischen Fehlers setzt den Fehler wieder in Abhängigkeit zum Fehler, bei dem immer der Mittelwert vorausgesagt wird.

Die Berechnung der Werte ist in Weka für eine Klassenverteilung nicht möglich und wird daher selbst vorgenommen. Dies ist mit einer kleinen Erweiterung der Fehlermaße möglich.

$$Erweiterter\ Fehler\ = \frac{\sum_{m=1}^{m} Fehlermaß}{m}$$

Die vorhandenen Maße geben den Vorhersagefehler über alle Klassen wieder. Bei der Erweiterung wird für jede Instanz m_1, m_2, …, m_n aus den vorhergesagten Wahrscheinlichkeitswerten p_1, p_2, …, p_n und ihren tatsächlichen Wahrscheinlichkeitswerten a_1, a_2, …, a_n ein Fehler berechnet. Die Fehler aller Instanzen werden dann summiert und durch die Anzahl der Instanzen geteilt.

Evaluierung eines Modells anhand ausgewählter Beispiele

Bevor eine Modellbewertung auf der gesamten Testemenge erfolgt, wird zunächst exemplarisch auf die Vorhersagen für einzelne Gemeinden eingegangen. Damit soll ein Eindruck darüber vermittelt werden, wie gut das Vorhersagemodell unter verschiedenen Bedingungen funktioniert. Die Auswahl versucht dabei, eine gewisse Vielfalt der Gemeindemerkmale und der Wahlergebnisse abzubilden. Es werden die Vorhersagen von Modell_25015 wiedergegeben. In den folgenden Ausführungen gibt die mittlere Zeile jeweils das tatsächliche normalisierte Wahlergebnis an und die untere Zeile das Ergebnis, das durch das Modell vorhergesagt wurde.

Für die Bundeshauptstadt Berlin zeigt sich die größte Abweichung für die Linke. Auch bei der CSU zeigt sich ein großer Unterschied zwischen vorhergesagtem und tatsächlichem Wert. Das Ergebnis für die AfD wurde hingegen ziemlich genau vorhergesagt. Der durchschnittliche absolute Fehler liegt bei 0,04347. Das bedeutet, dass das vorhergesagte Ergebnis durchschnittlich um einen absoluten Wert von 4,3% vom tatsächlichen Ergebnis abweicht.

Berlin	Partei	CDU	SPD	FDP	Grüne	Linke	AfD
	Wahlergebnis (normalisiert)	0,3063	0,2646	0,0383	0,1329	0,1990	0,0530
	Voraussage	0,3874	0,2964	0,0561	0,0955	0,1090	0,0557

Tabelle 10 Tatsächliches und vorhergesagtes Wahlergebnis für Berlin

Die Besonderheit an der Gemeinde Birtlingen, die 76 Einwohner hat, ist das vergleichsweise sehr hohe Wahlergebnis der AfD. Dieses wird vom Modell überhaupt nicht erkannt. Für die Partei die Linke, die keine einzige Stimme erhalten hat, wird ein zu hohes Wahlergebnis ausgegeben. Die Stimmanteile für die FDP und die Grünen werden hingegen sehr genau vorhergesagt. Der durchschnittliche absolute Fehler für die Gemeinde beträgt 0,07924.

Birtlingen	Partei	CDU	SPD	FDP	Grüne	Linke	AfD
	Wahlergebnis (normalisiert)	0,4691	0,1173	0,0586	0,0586	0	0,2932
	Voraussage	0,5364	0,2037	0,0626	0,0573	0,0816	0,0583

Tabelle 11 Tatsächliches und vorhergesagtes Wahlergebnis für Birtlingen

In der rheinland-pfälzischen Gemeinde Nusbaum wählte nur ein Fünftel der Wahlberechtigten CDU. Mit diesem Ergebnis hat der Klassifizierer große Probleme. Obwohl die SPD mehr als doppelt so viele Stimmen wie die CDU erhalten hat, sagt der Klassifizierer einen deutlich höheren Stimmanteil für die CDU voraus. Auch bei den Grünen wird das tatsächliche Ergebnis stark unterschätzt. Der durchschnittliche absolute Fehler hat eine Höhe von 0,08578.

Nusbaum	Partei	CDU	SPD	FDP	Grüne	Linke	AfD
	Wahlergebnis (normalisiert)	0,1992	0,4665	0,0472	0,1625	0,0734	0,0472
	Voraussage	0,4393	0,2992	0,0546	0,0776	0,0703	0,0591

Tabelle 12 Tatsächliches und vorhergesagtes Wahlergebnis für Nusbaum

Besser funktioniert die Erkennung eines relativ niedrigen Ergebnisses für die CDU bei Saarbrücken. Hier ist der vorausgesagte Wert für die CDU nicht viel größer als der tatsächliche normalisierte Wert. Der nahezu gleich große Stimmanteil für CDU und SPD wird gut wiedergegeben. Die größte Differenz für den vorhergesagten und den tatsächlichen Wert ergibt sich bei der Linken. Der absolute Fehler beträgt 0,02562.

Saarbrücken	Partei	CDU	SPD	FDP	Grüne	Linke	AfD
	Wahlergebnis (normalisiert)	0,3362	0,3254	0,0534	0,1016	0,1305	0,0481
	Voraussage	0,3691	0,3569	0,0423	0,0835	0,0851	0,0631

Tabelle 13 Tatsächliches und vorhergesagtes Wahlergebnis für Saarbrücken

Die Gemeinde Zweifelscheid liegt in Rheinland-Pfalz und hatte im Jahr 2011 eine Bevölkerungszahl von 47. In dem Ort wurden drei der sechs Parteien nicht gewählt. Das Modell teilt diese Parteien dennoch insgesamt circa 15,5 Prozent der Stimmen zu. Das Ergebnis der SPD wird zu hoch eingeschätzt und das Ergebnis für die CDU um fast 20 Prozentpunkte zu niedrig. Die vorausgesagten Werte sind schlecht. Es ergibt sich ein durchschnittlicher absoluter Fehler von 0,07370.

Zweifelscheid	Partei	CDU	SPD	FDP	Grüne	Linke	AfD
	Wahlergebnis (normalisiert)	0,7589	0,1598	0,0799	0	0	0
	Voraussage	0,5633	0,2270	0,0552	0,0632	0,0637	0,0277

Tabelle 14 Tatsächliches und vorhergesagtes Wahlergebnis für Zweifelscheid

Die Stadt Sonneberg mit ihren im Jahr 2011 knapp 22.000 Einwohnern liegt in Thüringen. Die größte Abweichung zwischen vorhergesagten und tatsächlichem Wert besteht bei der CDU. Insgesamt ist das vorhergesagte Ergebnis sehr nahe am tatsächlichen Ergebnis. Der hohe Stimmanteil der Linken wird sehr gut abgebildet. Der durchschnittliche absolute Fehler hat einen Wert von 0,01248.

Sonneberg	Partei	CDU	SPD	FDP	Grüne	Linke	AfD
	Wahlergebnis (normalisiert)	0,3960	0,1881	0,0271	0,0367	0,2826	0,0634
	Voraussage	0,4327	0,1748	0,0291	0,0384	0,2662	0,0588

Tabelle 15 Tatsächliches und vorhergesagtes Wahlergebnis für Sonnerberg

Für die bayerische Landeshauptstadt München zeigen sich gute Ergebnisse. Die höchsten Abweichungen gibt es bei der CDU, den Grünen und den Linken. Der vorhergesagte Wert für die SPD zeigt nur eine geringe Differenz zum tatsächlichen Wert. Der durchschnittliche absolute Fehler beträgt 0,02514.

München	Partei	CDU	SPD	FDP	Grüne	Linke	AfD
	Wahlergebnis (normalisiert)	0,4060	0,2569	0,0828	0,1510	0,0494	0,0483
	Voraussage	0,4522	0,2393	0,0635	0,1153	0,0740	0,0557

Tabelle 16 Tatsächliches und vorhergesagtes Wahlergebnis für München

Die nordhessische Gemeinde Oberweser wurde ausgewählt, da hier das Wahlergebnis der SPD sehr hoch ist. Die Höhe des Ergebnisses für die SPD wird vom Modell jedoch um zehn Prozentpunkte zu niedrig eingeschätzt. Dafür wird das Ergebnis für die CDU um etwa zehn Prozentpunkte zu hoch eingeschätzt. Der durchschnittliche absolute Fehler für die Gemeinde beträgt 0,04055.

Oberweser	Partei	CDU	SPD	FDP	Grüne	Linke	AfD
	Wahlergebnis (normalisiert)	0,2800	0,4912	0,0455	0,0694	0,0652	0,0466
	Voraussage	0,3781	0,3807	0,0438	0,0771	0,0569	0,0634

Tabelle 17 Tatsächliches und vorhergesagtes Wahlergebnis für Oberweser

Für die bayerische Gemeinde Konnersreuth, die 1874 Einwohner hat, ist erkennbar, dass das hohe Ergebnis der CDU sehr genau abgebildet wird und auch das eher niedrige Ergebnis für die SPD vorhergesagt wird. Auch für die anderen Parteien sind die Ergebnisse gut. Der durchschnittliche absolute Fehler beläuft sich auf 0,01812.

Konnersreuth	Partei	CDU	SPD	FDP	Grüne	Linke	AfD
	Wahlergebnis (normalisiert)	0,7073	0,1693	0,0250	0,0291	0,0311	0,0210
	Voraussage	0,7010	0,1374	0,0452	0,0568	0,0290	0,0306

Tabelle 18 Tatsächliches und vorhergesagtes Wahlergebnis für Konnersreuth

Zusammenfassend ist festzustellen, dass das Vorhersagemodell die Wahlergebnisse mit Einschränkungen häufig gut abbildet. Probleme hat das Modell bei der Vorhersage ungewöhnlicher Ergebnisse. Dies ist zum Beispiel der Fall, wenn eine oder mehrere Parteien gar keine Stimme erhalten. In diesem Fall wird immer ein höherer Wert vorausgesagt. Außerdem wurden ein sehr hohes Ergebnis für die AfD und ein relativ niedriges Ergebnis für die CDU nicht einmal im Ansatz abgebildet. In der Mehrheit der Fälle wird das Ergebnis für die CDU überschätzt. Das Ergebnis für die SPD ist teilweise recht genau, teils zu hoch und teils zu niedrig. Für die FDP und die AfD gibt es in der Regel eine geringe absolute Differenz zwischen vorhergesagtem und tatsächlichem Wert. Das Ergebnis für die Grünen wird relativ gut abgebildet, es ergeben sich für manche Gemeinden aber auch größere Unterschiede. Ähnliches gilt für die Linke. Für diese Partei ist zu bemerken, dass auch hohe Stimmanteile gut abgebildet werden. Der absolute Fehler für die betrachteten Gemeinden reicht von 0,01248 bis 0,08578. Je nach Gemeinde ergeben sich demnach deutliche Unterschiede in der Vorhersagegenauigkeit.

Evaluierung der Ergebnisse über alle Gemeinden

Nachdem ein Überblick über die Qualität der Ergebnisse anhand einzelner Beispiele gegeben wurde, findet nun eine systematische Bewertung statt, die die gesamten Testdaten einbezieht. Um die Güte eines Modells bewerten zu können, ist es nützlich, seine Fehlerwerte mit den Fehlerwerten von sogenannten Baseline-Klassifizierern zu vergleichen. Diese Fehlerwerte werden auch Baseline-Fehler genannt. Ein Baseline-Klassifizierer ist ein sehr einfaches Modell. Es dient als Vergleich zur Bewertung der erstellten Modelle. Modelle, die schlechter oder genauso gut wie der Baseline-Klassifizierer sind, sind nicht sinnvoll. Sie erreichen gemessen an ihrer Komplexität nur ein unzureichendes Ergebnis. Als Vergleichswerte werde die Fehlerwerte von vier verschiedenen Klassifizierern verwendet. Zunächst wird ein Klassifizierer angenommen, der jeder neuen Gemeinde die Stimmverteilung zuordnet, die der Stimmverteilung auf Bundesebene entspricht. Zusätzlich wird der Fehler berechnet, der bei einer Gleichverteilung aller Stimmen über die Parteien auftritt. Schließlich werden die Fehler bestimmt, die entstehen, wenn nur der stärksten oder nur der schwächsten Partei alle Stimmen zugeordnet werden. In Tabelle 19, die sich auf der nächsten Seite befindet, bilden die ersten vier Einträge die Baseline-Fehler. Danach folgen die Fehlerwerte für die einzelnen Modelle. Für jedes Fehlermaß wurde das Modell mit dem kleinsten Fehler grün eingefärbt.

	Mittlerer quadr. Fehler	Relativer quadr. Fehler	Durchschn. Absoluter Fehler	Relativer absoluter Fehler	Wurzel des mitteleren quadr. Fehlers	Wurzel des relativen quadr. Fehlers
Baseline-Fehler						
Bundes-ergebnis	0,00769	0,11850	0,05861	0,90213	0,08194	0,31509
Gleich-verteilung	0,03473	0,42298	0,14689	1,63927	0,18285	209,70504
Nur CDU	0,05214	0,42059	0,15537	1,56096	0,22127	0,66169
Nur AfD	0,21214	11,72442	0,31537	4,44201	0,46016	5,85145
Fehler der Modelle						
Variation der Baumtiefe						
Modell_01001	0,00463	0,09915	0,04621	0,82704	0,06176	0,28110
Modell_01002	0,00367	0,09324	0,04057	0,79249	0,05415	0,27124
Modell_01003	0,00315	0,09002	0,03718	0,77168	0,04960	0,26686
Modell_01004	0,00280	0,08796	0,03475	0,75703	0,04634	0,26482
Modell_01005	0,00246	0,08599	0,03256	0,74367	0,04320	0,26171
Modell_01006	0,00230	0,08483	0,03130	0,73591	0,04147	0,26131
Modell_01007	0,00210	0,08395	0,02995	0,72792	0,03949	0,25957
Modell_01008	0,00197	0,08253	0,02885	0,72113	0,03799	0,25899
Modell_01009	0,00196	0,08309	0,02886	0,72166	0,03792	0,26007
Modell_10010	0,00189	0,08304	0,02840	0,71940	0,03721	0,26012
Modell_10015	0,00179	0,08272	0,02788	0,71865	0,03617	0,26063
Modell_10020	0,00182	0,08322	0,02841	0,72261	0,03665	0,26071
Modell_10025	0,00183	0,08325	0,02855	0,72427	0,03681	0,26066
Modell_10050	0,00182	0,08325	0,02833	0,72292	0,03663	0,26140
Modell_10088	0,00182	0,08325	0,02833	0,72292	0,03663	0,26140
Variation der Baumanzahl						
Modell_05015	0,00192	0,08421	0,02930	0,72989	0,03783	0,26536
Modell_20015	0,00175	0,08217	0,02731	0,71393	0,03553	0,25865
Modell_25015	0,00174	0,08190	0,02720	0,71252	0,03542	0,25833
Maximaleinstellung						
Modell_25088	0,00173	0,08188	0,02724	0,71380	0,03535	0,25832

Tabelle 19 Fehlerwerte über alle Gemeinden

Fehler der Baseline-Klassifizierer

Für die Baseline-Klassifizierer zeigen sich unterschiedlich hohe Fehlerwerte. Wenn die AfD als unwahrscheinlichste Klasse mit einem Ergebnis von 100% vorhergesagt wird, ist der Fehler mit Abstand am höchsten. Bei einer reinen Vorhersage der CDU, die die wahrscheinlichste Klasse darstellt, ist der Fehler weitaus geringer. Noch ein wenig näher am richtigen Ergebnis ist eine Gleichverteilung der Stimmen auf alle Klassen. Die Zuordnung des Bundesergebnisses den geringsten Fehler. Der durchschnittliche absolute Fehler ist von den Fehlermaßen derjenige, der am einfachsten bewertet werden kann. Beim Bundesergebnis hat er einen Wert von 0,05861. Das bedeutet, dass die prozentuale Zuordnung der Stimmen im Schnitt um 5,9% vom tatsächlichen Ergebnis abweicht. Die Fehlerwerte der erstellten Modelle sollten geringer als die Fehlerwerte des besten Baseline-Klassifizierers sein.

Fehler der Modelle

Alle erstellten Modelle haben, verglichen mit den Fehlerwerten des Bundesergebnisses, für alle Fehlermaße einen niedrigeren Fehler. Jedes der erstellten Modelle stellt also eine Verbesserung zum besten Baseline-Klassifizierer dar. Bei den Modellen mit 100 Bäumen zeigt sich, dass die Fehlerwerte mit steigender Baumtiefe abnehmen. Der niedrigste Fehler bei den Modellen mit 100 Bäumen ergibt sich bei einer Baumtiefe mit 15. Hier beträgt der durchschnittliche absolute Fehler 0,02788. Für zwei der sechs Fehlerwerte zeigt das Modell_01008 ein besseres Ergebnis. Bei einer höheren Baumtiefe als 15 steigen die Fehler wieder. Beim Vergleich der Modelle Modell_10050 und Modell_10088 ist festzustellen, dass diese identische Fehlerwerte haben. Daraus lässt sich schließen, dass auch ohne Limitierung der Baumtiefe kein Baum eine höhere Tiefe als 50 hat.

Neben der Veränderung der Baumtiefen wurde auch die Baumanzahl variiert. Die Erstellung von Modell_05015, Modell_20015 und Modell_25015 geschah mit der Überlegung, die ermittelte optimale Baumtiefe mit einer anderen Zahl an Bäumen zu kombinieren. Dabei ist klar, dass Baumtiefe und Baumanzahl nicht völlig voneinander unabhängig sind, sodass bei einer Anzahl an 200 Bäumen die Tiefe von 15 nicht zwingend die optimale Tiefe ist. Jedoch konnten nicht sämtliche Kombinationen getestet werden und die weitere Verwendung der Baumtiefe 15 stellte einen praktikablen Ansatz dar. Hinsichtlich der unterschiedlichen Baumzahlen zeigt sich, dass das Ergebnis bei einer Senkung der Baumanzahl auf 50 schlechter wird. Wird die Baumanzahl auf 200 verdoppelt, verbessert sich das Ergebnis. Dasselbe trifft bei einer weiteren Erhöhung auf 250 Bäume zu. Modell_25015 hat über alle Fehlermaße hinweg ein besseres Ergebnis als Modell_10015.

Für das zuletzt erstellte Modell_25088 wurde die maximal mögliche Baumanzahl und eine unbegrenzte Tiefe der Bäume gewählt. Die maximal mögliche Baumanzahl beträgt 250 und wird durch die Rechenleistung des verwendeten Rechners festgelegt. Das so erstellte Modell hat bei vier der sechs verwendeten Fehlermaße einen besseren Wert als das zuvor beste Modell_25015. Dieses hat dagegen beim durchschnittlichen absoluten Fehler und relativen absoluten Fehler den besseren Wert. Die Ergebnisse der beiden Modelle sind extrem nahe beieinander. Aus dem absoluten Fehler ergibt sich, dass das Modell_25015 das Wahlergebnis einer Gemeinde im Schnitt um 2,720 Prozentpunkte und das Modell_25088 im Schnitt um 2,724 Prozentpunkte falsch vorhersagt. Im Schnitt weicht das vorhergesagt Ergebnis vom tatsächlichen Ergebnis also um einen absoluten Wert von 2,720% bzw. 2,724% ab. Dieser Unterschied ist vernachlässigbar klein. Zur Vorhersage der Parteiergebnisse würde also das Modell_25015 genutzt werden. Bei zwei Modellen, die gleiche Ergebnisse erzielen, wird in der Regel das einfachere von beiden gewählt.

Die Fehlerwerte zeigen, dass das Klassifikationsmodell auf unbekannten Daten gute Ergebnisse liefert. Wird jedoch noch ein weiterer Bewertungsmaßstab zur Rate gezogen, zeigen sich auch die Schwächen des Modells. Neben der reinen Fehlerberechnung sollte auch gegeben sein, dass das Verhältnis der Parteien auf dem ersten Rang in etwa gewahrt wird. Dies ist beim verwendeten Random Forest nicht der Fall. Unabhängig von der Anzahl der Bäume und der Baumtiefe erhält mit sehr wenigen Ausnahmen die CDU den größten Prozentsatz. Bei den tatsächlichen Wahlergebnissen hat die SPD bei 7% der Gemeinden den höchsten Wert. Bei der Klassifikation erreicht die SPD weniger als 1%. Wenn auch ein exakt abgebildetes Verhältnis der Parteien auf dem ersten Rang nicht unbedingt geleistet werden muss, so sollte es dennoch grob erkennbar sein. An dieser Stelle besteht also noch Verbesserungsbedarf. Bei den übrigen Parteien ist der Anteil am ersten Rang vernachlässigbar klein und muss daher vom Modell nicht zwingend erkannt werden.

Evaluierung der Ergebnisse nach Parteien

Zusätzlich zu den bisherigen Betrachtungen ist die Bewertung der Modelle im Hinblick auf die einzelnen Parteien interessant. Um die Vorhersagegenauigkeit für die einzelnen Parteien zu bestimmen, wurde der durchschnittliche absolute Fehler für jede Partei über alle Gemeinden berechnet. Die Ergebnisse sind in der nachfolgenden Tabelle dargestellt. Die niedrigsten Fehler sind dabei wieder grün markiert.

	Absoluter Fehler CDU	Absoluter Fehler SPD	Absoluter Fehler FDP	Absoluter Fehler Grüne	Absoluter Fehler Linke	Absoluter Fehler AfD
Baseline-Fehler						
Bundes-Ergebnis	0,13301	0,08304	0,01594	0,03214	0,07211	0,01544
Gleich-verteilung	0,36303	0,06650	0,12272	0,10335	0,10859	0,11716
Nur CDU	0,47030	0,20965	0,04411	0,06369	0,09477	0,04970
Nur AfD	0,52970	0,20965	0,04411	0,06369	0,09477	0,95030
Fehler der Modelle						
Variation der Baumtiefe						
Modell_01001	0,09237	0,07475	0,01521	0,02362	0,05549	0,01583
Modell_01002	0,08272	0,06651	0,01444	0,02195	0,04239	0,01540
Modell_01003	0,07633	0,06168	0,01388	0,02095	0,03511	0,01514
Modell_01004	0,07228	0,05798	0,01358	0,02020	0,02938	0,01509
Modell_01005	0,06635	0,05378	0,01328	0,01956	0,02754	0,01487
Modell_01006	0,06341	0,05207	0,01335	0,01894	0,02522	0,01481
Modell_01007	0,05984	0,04911	0,01318	0,01877	0,02399	0,01481
Modell_01008	0,05724	0,04699	0,01301	0,01867	0,02245	0,01477
Modell_01009	0,05697	0,04694	0,01310	0,01865	0,02263	0,01486
Modell_10010	0,05550	0,04589	0,01325	0,01854	0,02218	0,01505
Modell_10015	0,05281	0,04359	0,01411	0,01861	0,02185	0,01631
Modell_10020	0,05314	0,04326	0,01489	0,01938	0,02304	0,01675
Modell_10025	0,05361	0,04350	0,01486	0,01988	0,02259	0,01687
Modell_10050	0,05392	0,04284	0,01476	0,01928	0,02213	0,01703
Modell_10088	0,05392	0,04284	0,01476	0,01928	0,02213	0,01703
Variation der Baumanzahl						
Modell_05015	0,05464	0,04524	0,01559	0,01989	0,02290	0,01755
Modell_20015	0,05230	0,04307	0,01324	0,01814	0,02142	0,01570
Modell_25015	0,05213	0,04309	0,01308	0,01805	0,02141	0,01542
Maximaleinstellung						
Modell_25088	0,05255	0,04188	0,01351	0,01842	0,02119	0,01589

Tabelle 20 Absoluter Fehler nach Parteien

Das Ergebnis zeigt, dass sich die Fehler zwischen den einzelnen Parteien deutlich unterscheiden. Bei der FDP und der AFD ist der durchschnittliche absolute Fehler mit 0,01301 beziehungsweise 0,01477 am niedrigsten. Dies sind die Parteien, die den geringsten Stimmanteil haben. Es ist zu beobachten, dass mit einem steigenden Stimmanteil auch ein Anstieg des absoluten Fehlers zu beobachten ist. Die CDU hat dementsprechend den höchsten absoluten Fehler. Diese unterschiedlich großen Fehler für die Parteien sind im Sinne des fachlichen Ziels. Das Ergebnis wäre schlechter gewesen, wenn der absolute Fehler für alle Parteien gleich groß wäre. Bei der AfD als kleinsten Partei fiele ein absoluter

Fehler von 0,03000 viel stärker ins Gewicht als bei einer großen Partei. Bei einem Stimmanteil von 4,8% macht eine absolute Abweichung von 3% einen deutlichen Unterschied. Daher ist es als positiv zu bewerten, dass sich die Höhe des Fehlers nach der Höhe der Stimmverteilung richtet. Dass kleinere Parteien niedrigere Fehler als größere Parteien haben, ist auch dem Stimmanteil der Parteien geschuldet. Bei einem höheren Stimmergebnis kann die Voraussage stärker abweichen. Daher müssen die Unterschiede auch in Relation zu den Stimmanteilen gesehen werden.

Bei der CDU liegt der Fehler bei der Zuteilung des Bundesergebnisses bei 0,13301 und kann mit dem besten Modell auf 0,05213 gesenkt werden. Bei der SPD verringert sich der Fehler von 0,08304 auf 0,04188. Für die Linke sinkt der Fehler von 0,07211 auf 0,02119. Bei den Grünen sinkt der absolute Fehler von 0,03214 auf 0,01805. Für die Grünen ist die erreichte Verbesserung also vergleichsweise niedrig. Die Linke und die Grünen haben mit 8,6% bzw. 8,4% ungefähr gleich viele Stimmen bei der Bundestagswahl erhalten. Trotzdem ist der absolute Fehler bei den Linken beim Baseline-Fehler des Bundesergebnisses mehr als doppelt so hoch. Das liegt daran, dass sich das Ergebnis für die Linkspartei zwischen den Gemeinden stark unterscheiden kann, je nachdem in welchem Bundesland sich die Gemeinde befindet. Durch diese Ungleichverteilung weicht das beste Ergebnis für die Linken stärker vom durchschnittlichen Bundesergebnis ab als bei den Grünen. Der Baseline-Fehler ist für die Linke fast so hoch wie für die SPD, obwohl diese mehr als doppelt so viele Stimmen erhalten hat. Bei der FDP wird der absolute Fehler von 0,01594 auf 0,01301 gesenkt. Bei der AfD sinkt der absolute Fehler von 0,01544 auf 0,01477. Der Fehler bei diesen beiden Parteien war also bereits schon bei der Zuteilung des Bundesergebnisses sehr gering. Durch das Modell konnte eine Verbesserung erreicht werden, die allerdings nicht sehr groß ist.

Aus dem Ergebnis ist außerdem ersichtlich, dass die optimale Tiefe des Baumes für die beiden kleinsten Parteien anders als für die anderen Parteien ist. Sowohl bei der AfD als auch FDP wird das beste Ergebnis bei einer Baumtiefe von acht, also mit Modell_01008 erreicht. Besonders bei den zwei größten Parteien CDU und SPD zeigt sich bei einer weiteren Erhöhung der Baumtiefe ein erkennbar besseres Ergebnis. Bei den beiden mittleren Parteien Linke und Grüne verbessert sich das Ergebnis mit steigender Baumanzahl ebenso, aber nicht so stark wie bei CDU und SPD. Das Ergebnis zeigt also einen Zusammenhang zwischen der Vorhersagegenauigkeit eines Random Forest abhängig von seiner Baumtiefe für unterschiedlich wahrscheinliche Klassen.

Wie bei der vorherigen parteiübergreifenden Analyse sind die Modelle Modell_25015 und Modell_25088 hier die beste Wahl. Eine Ausnahme ergibt sich jedoch hinsichtlich der beiden kleinsten Parteien, bei denen eine geringere Baumtiefe in der Höhe von acht optimal ist, da sich das Ergebnis danach wieder verschlechtert. Für diese Parteien erreicht das Modell_01008 die besten Ergebnisse. Je nachdem, auf welchen Aspekt der Vorhersage am meisten Wert gelegt wird, ist also die Wahl unterschiedlicher Modelle sinnvoll.

6. Diskussion der Ergebnisse

In diesem Abschnitt werden die Ergebnisse der Masterarbeit diskutiert und die Limitationen der Arbeit aufgezeigt. Für den US-amerikanischen Wahlkampf wurde festgestellt, dass er durch das Vorhandensein von öffentlichen Wählerregistern, die als Grundlage der Datensammlung dienen, und einen schwachen Datenschutz geprägt ist. Durch weitere Informationen, die aus der Interaktion mit Wählern, dem Verhalten von Personen im Web und das Zukaufen von Datenbrokern gewonnen werden, wird die Erstellung von prädiktiven Scores ermöglicht. Basierend darauf werden passende Aktionen wie beispielsweise die gezielte Ansprache einer bestimmten Personengruppe ausgeführt. Das Volumen der verwendeten Daten ist dabei extrem groß. Alleine die Datenbroker haben für jeweils etwa 200 Millionen individuelle US-Amerikaner mehrere Tausend Datenpunkte gespeichert. Als wichtiger Faktor stellt sich außerdem heraus, dass die einzelnen Daten nicht in voneinander getrennten Datensilos gespeichert werden, sondern dass mit einer integrierten aktuellen Datenbasis gearbeitet wird. Eine wichtige Rolle nahm auch die Freiwilligenarbeit ein, durch die eine Vielzahl von Personen entweder im Gespräch oder über soziale Medien kontaktiert wurde. Die richtige Ansprache von Bürgern wurde darüber hinaus mittels Experimenten verbessert. Bezüglich der Betrachtung des US-amerikanischen Wahlkampfs in den USA ist jedoch zu bemerken, dass neben der Behandlung der grundsätzlichen Bedingungen im Wahlkampf im Speziellen nur auf die Wahlkampagne des Präsidentschaftskandidaten Barack Obama eingegangen wurde. Die anderen Kandidaten bei der Präsidentschaftswahl in den Jahren 2008 und 2012 wurden nicht berücksichtigt. Ebenso wurde nicht auf die aktuelle Präsidentschaftswahl im Jahr 2016 eingegangen, obwohl sich seit dem Jahr 2012 neue technische Möglichkeiten ergeben haben. Dies betrifft insbesondere auch die Entwicklung, dass im Vergleich zu 2012 Smartphones sehr weit verbreitet sind. Mittels Smartphones und der für sie entwickelten Apps, die teilweise sehr viele Daten sammeln und unbegrenzt an Dritte weitergeben, lassen sich noch deutlich mehr Daten über Personen sammeln. Dazu zählen das Erstellen von genauen Bewegungsprofilen und der Zugriff auf die gespeicherten Kontaktlisten.

In Deutschland darf eine Partei hingegen nicht auf die Wählerverzeichnisse zugreifen. Der Datenschutz beinhaltet das Recht auf informationelle Selbstbestimmung, die in den meisten Fällen nötige Zustimmung bei der zweckbezogenen Verarbeitung personenbezogener Daten und das Prinzip der Datensparsamkeit. Die Auswertung von Daten ist in Deutschland hauptsächlich zur Wählermobilisierung und Wählergewinnung notwendig. Hinsichtlich der Parteifinanzierung spielen die staatliche Teilfinanzierung und die Mitgliedsbeiträge von Parteimitgliedern die größte Rolle. Deutschen Parteien stehen diverse Informationsquellen zur Verfügung, um Informationen über die Bevölkerung und Wähler in Deutschland zu erhalten. Zum einen ergeben sich Informationen aus den Daten der Parteimitglieder. Bezüglich der Mitgliederzahlen deutscher Parteien ist aber zu beobachten, dass diese sinkend sind. Weitere Quellen sind Parteispenden und die repräsentative Wahlstatistik des Bundeswahlleiters. Ein großes Datenangebot stellen die statistischen Ämter des Bundes und der Länder bereit. Überdies können soziale Medien zur Datengewinnung genutzt werden. Informationen können zudem von Adresshändlern und Forschungsinstitutionen bezogen werden. Adresshändler sammeln Daten mit dem Ziel, Personen in verschiedene Konsumentengruppen einzuteilen. Diese Informationen werden dann an Organisationen verkauft, die sich neue Kundengruppen erschließen oder zielgerichtet werben wollen. Forschungsinstitutionen bieten vor allem Umfragedienste an und leisten Politikberatung. Für die Interaktion mit Wählern stehen Parteien Online-Kanäle und Offline-Kanäle zur Verfügung, die sich zu Teilen mit den in den USA gängigen Kommunikationsmethoden überschneiden. Trotz der Vielfältigkeit der Datenquellen steht in Deutschland ein viel geringeres Datenvolumen zur Verfügung, was vor allem durch einen Mangel an Informationen über Individuen begründet ist. Insgesamt konnten zu einigen Datenquellen nur oberflächliche Aussagen gemacht werden. Bezüglich der Adresshändler wurden nur Informationen aus öffentlich verfügbaren Dokumenten entnommen. Es konnte daher nicht bewertet werden, wie vollständig, aktuell und detailliert die von diesen Unternehmen gesammelten Daten sind. Auch bei den Forschungsinstituten

im Politikbereich wurden nur die auf den Webseiten der Institute angegeben Informationen verwendet. Ausgelassen wurde zudem eine Betrachtung der Unternehmen in Deutschland, die sich unabhängig von politischen Themen der Marktforschung widmen.

Um die Möglichkeit von Vorhersagen für den deutschen Wahlkampf praktisch zu evaluieren, wurde ein Vorhersagemodell entwickelt, das an den Unterstützungs-Score aus dem US-amerikanischen Wahlkampf angelehnt ist. Das Klassifikationsproblem wurde dabei als probabilistische Klassifikation modelliert. Als Klassifizierer wurde ein Random Forest verwendet, der für eine Gemeinde aus ihren soziodemographischen Daten das Wahlergebnis vorhersagt. Der Random Forest gibt bei einem Datenobjekt für jede Klasse eine Wahrscheinlichkeit an, mit der das Datenobjekt zu dieser Klasse gehört. Die Wahrscheinlichkeit, dass ein Datenobjekt einer Klasse angehört wurde als der Anteil an Zweitstimmen interpretiert, den eine Partei in einer Gemeinde erhält. Um die Klassifikation durchführen zu können, war eine Transformation der Trainingsdaten notwendig. Dabei wurde eine Instanz in sechs Unterinstanzen zerlegt und es wurde jeder Unterinstanz eine andere Partei sowie der Stimmanteil dieser Partei zugeordnet. Mit dieser Methode wurden gute Ergebnisse erzielt. Würde jeder Gemeinde das Bundesergebnis zugeordnet werden, ergäbe sich ein durchschnittlicher absoluter Fehler von 0,059. Durch das Vorhersagemodell konnte dieser Fehler auf knapp 0,027 gesenkt werden. Jedoch sind bei genauerem Hinsehen Schwächen des Modells zu erkennen. Die Modelle sagen die CDU zu oft als stärkste Partei voraus. Der Random Forest teilt in über 99% der Fälle der CDU das höchste Wahlergebnis zu, obgleich dies bei den realen Daten nur in 93% der Fälle richtig ist. Insgesamt wird das Ergebnis für die CDU in vielen Fällen überschätzt. Hier besteht also noch Verbesserungsbedarf. Probleme bereiten auch ungewöhnliche Wahlergebnisse. Dies ist beispielsweise der Fall, wenn die AfD einen sehr hohen Stimmanteil erhält oder drei der sechs Parteien gar nicht gewählt werden. Außerdem ist festzustellen, dass die vorhergesagten Werte je nach Partei um 1,3 bis 5,2 Prozentpunkte vom tatsächlichen Ergebnis abweichen.

Bezüglich des Ergebnisses der Konfusionsmatrix, die sich aus einer zehnfachen Kreuzvalidierung ergibt, konnte keine schlüssige Erklärung gefunden werden. Diese ist für das Endergebnis aber von untergeordneter Wichtigkeit. Der Grund für das sehr eigenartige Ergebnis könnte in der Widersprüchlichkeit der Trainingsdaten liegen. Die Stratifizierung der Daten anhand des Rankings geschah ohne eine Anlehnung an bestehende Forschung oder Praxis. Die Trennung auf Basis eines Rankings berücksichtigt, abgesehen von ihrer Reihenfolge, nicht die tatsächlichen Distanzen zwischen den einzelnen Parteien. Die Unterteilung aufgrund des Rankings wurde gewählt, da sie die Instanzen grob nach ihrer Klassenverteilung trennt und einfach durchzuführen ist. Es bleibt aber unklar, inwieweit eine Unterteilung der Daten, die weitere Aspekte als nur die Reihenfolge der Klassen berücksichtigt, zu besseren Klassifikationsergebnissen geführt hätte. Hinsichtlich der verwendeten Daten wäre denkbar gewesen, die Instanzen zu den Stadtstaaten und anderen sehr großen Städten ab einer Einwohnerzahl von 500.000 weiter aufzugliedern. Die Tatsache, dass die Daten auf Gemeinde-Ebene vorliegen führt dazu, dass die Einwohnerzahlen der Instanzen teilweise bedeutend voneinander abweichen. Beispielsweise hat die Gemeinde Juliusburg 178 Einwohner und Berlin über 3 Millionen Einwohner. Die Instanzen werden bei der Analyse als gleich wichtig angesehen, was aber nicht ihrer politischen Wichtigkeit entspricht. Für einige größere Städte sind Zensus- und Wahldaten auch auf der Ebene der Stadtteile verfügbar. Diese Daten hätten für ein genaueres Ergebnis anstelle von Gemeindedaten verwendet werden können. Neben dem Argument der Einwohnerzahl ist auch davon auszugehen, dass es sich bei der Bevölkerung in großen Städten um keine homogenen Gruppen handelt und demnach große Unterschiede in den einzelnen Stadtteilen bestehen können. Außerdem wäre es denkbar gewesen, die Einwohnerzahl oder die Wahlbeteiligung eines Ortes in das Modell miteinfließen zu lassen, was allerdings die Komplexität erhöht hätte. Für das Modell wurde nur ein kleiner Teil der theoretisch vorhandenen statistischen Daten verwendet. Zudem handelt es sich bei den Daten um eine Momentaufnahme. Die politische Meinungsentwicklung ist jedoch höchst dynamisch und von mehr als soziodemographischen Faktoren geprägt. Die Wahlergebnisse und die Zensusdaten wurden außerdem zu unterschiedlichen Zeitpunkten erhoben, die mehr als zwei Jahre

auseinander liegen. Das Ziel der Arbeit war nicht die Beachtung der aktuellen politischen Entwicklungen, sondern die potentielle Nutzung von Daten anhand eines Beispiels aufzuzeigen. Dies ist gelungen. Die Masterarbeit kann damit als ein Ansatz zur Möglichkeit der Verwendung von Daten im deutschen Wahlkampf angesehen werden. Zu beachten ist dabei, dass die Arbeit ohne Kenntnis darüber entstanden ist, mit welcher Intensität die Parteien in Deutschland die vorgestellten Datenquellen bereits nutzen und in welchem Maß Personen beschäftigt werden, die in Informationstechnologie oder Statistik ausgebildet sind.

7. Zusammenfassung und Ausblick

In der vorliegenden Arbeit wurde die Möglichkeit eines datengetriebenen Wahlkampfs in Deutschland nach dem Vorbild der USA erörtert. Dabei wurde festgestellt, dass die Rahmenbedingungen der US-amerikanischen Präsidentschaftswahl und der deutschen Bundestagswahl unterschiedlich sind. In den USA werden auf professionelle Weise sehr viele Daten über Individuen zur Optimierung des Wahlkampfes ausgewertet. Auf der deutschen Seite gibt es mehr staatliche Unterstützung bei der Parteienfinanzierung, einen weiterreichenden Datenschutz und viel weniger Informationen, die über individuelle Bürger vorhanden sind. Zugleich ist zu beobachten, dass deutsche Parteien durch das Anbieten von Online-Communities zur Organisation und Vernetzung von Freiwilligen und Parteimitgliedern, das Auswerten von Newsletteraufrufen und die Vernetzung mit sozialen Medien in der digitalen Welt angekommen sind. Mit dem Tür-zu-Tür-Wahlkampf wurde auch eine Methode nach US-amerikanischem Vorbild übernommen.

Darüber hinaus wurde im Rahmen dieser Masterarbeit ein Random Forest erstellt, der die Zweitstimmenverteilung für die Gemeinden in Deutschland basierend auf Zensus-Daten vorhersagt. Das Vorhersagemodell erzielt gute Ergebnisse. Die Planung und Umsetzung des Modells nahm wenige Monate in Anspruch, verwendete nur einen kleinen Teil der in Deutschland verfügbaren statistischen Daten und wurde von einer einzelnen Person umgesetzt. Gemessen daran wird ersichtlich, dass Data Mining für deutsche Parteien nutzbringend eingesetzt werden kann. Wichtig dabei ist herauszufinden, welche Aspekte des Wahlkampfes mit Data Mining sinnvoll unterstützt werden können und welche Daten dafür verwendet werden können. Aus dieser Fragestellung ergeben sich Anknüpfungspunkte für die weitere Forschung. Wissenschaftliche Veröffentlichungen über den erfolgreichen Einsatz von Data Mining zum Zwecke des Wahlkampfes werden aber immer limitiert sein. Wahlkampf ist ein Wettkampf zwischen mehreren Parteien und eine Partei wird nutzbringende Erkenntnisse geheim halten und zu ihrem Vorteil nutzen. Bezüglich der Datennutzung im US-Wahlkampf ist nur bekannt, dass extrem viele Daten verarbeitet wurden und bestimmte Handlungen auf Basis der Ergebnisse durchgeführt wurden. Wie genau die einzelnen Verfahren des maschinellen Lernens operierten und auf welche Daten sie konkret zurückgriffen, ist unbekannt. Aus technischer Sicht ergeben sich weitere Forschungsansätze für die Verwendung von probabilisitischer Klassifikation zur Vorhersage mehrerer numerischer Werte. Diese betreffen auch die Unterteilung von Daten in stratifizierte Untermengen und die Interpretation der Konfusionsmatrix.

Abschließend ist zu sagen, dass aus Parteiensicht ein immer besseres Verständnis einzelner Wähler erstrebenswert ist. Sollten die Entwicklungen der Datenanalyse in den USA jedoch weiter voranschreiten und Deutschland diesem Vorgehen durch eine Anpassung der Rahmenbedingungen nacheifern, so ergeben sich durch den zunehmend gläsernen Wähler nicht nur moralische Bedenken, sondern auch eine Gefährdung der Demokratie mit ihren Wahlprinzipen. Am Ende des Wahlkampfs stehen aber nicht nur Datenanalysen, sondern immer auch Themen und die Auseinandersetzung des Bürgers mit diesen zur politischen Meinungsbildung. Ebenso wichtig wie die kritische Beobachtung der Ausschöpfung der technischen Möglichkeiten zur Wähleranalyse ist die Aufrechterhaltung einer Gesellschaft, die sich mit politischen Themen auseinandersetzt.

Literaturverzeichnis

abgeordnetenwatch.de (2016). "Brigitte Zypries (SPD) Abgeordnete Bundestag." Abrufdatum: 15.09.2016, von http://www.abgeordnetenwatch.de/brigitte_zypries-778-78592.html.

acxiom (2016). "Personicx." Abrufdatum: 22.09.2016, von http://d1fe5u1jnbojul.cloudfront.net/wp-content/uploads/2015/05/Acxiom-Personicx-2015.pdf.

aproxima Gesellschaft für Markt- und Sozialforschung (2016). "aproxima Gesellschaft für Markt- und Sozialforschung Weimar. Full-Service Institut.". Abrufdatum: 27.05.2016, von http://www.aproxima.de/index.php.

Aritstotle (2016). "We power democracy. Providing technology, data, and strategy for your campaign and public affair needs.". Abrufdatum: 31.10.2016, von http://aristotle.com/.

Arvato Bertelsmann (2016). "Adressqualifizierung - erfahren, was Kunden bewegt."

Bayerisches Landesamt für Statistik (2016). "GENESIS-Online Datenbank, Allgemeine Bundestagswahlstatistik." Abrufdatum: 24.09.2016, von https://www.statistikdaten.bayern.de/genesis/online/data?operation=statistikAbruftabellen&levelindex=0&levelid=1474710456626&index=1.

BDSG "Bundesdatenschutzgesetz (BDSG)." 2015.

Bennett, Colin J (2015). "Trends in Voter Surveillance in Western Societies: Privacy Intrusions and Democratic Implications." Surveillance & Society **13**(3/4): 370.

BMG "Bundesmeldegesetz (BMG)."

Boyd, Danah und Crawford, Kate (2012). "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." Information, communication & society **15**(5): 662-679.

Breiman, Leo (2001). "Random forests." Machine learning **45**(1): 5-32.

Bundeskanzlerin, Die (2016, 12.03.2016). "Merkel: Wir müssen uns sputen." von https://www.bundeskanzlerin.de/SiteGlobals/Forms/Webs/BKin/Suche/DE/Solr_Mediathek_formular.html?id=1923720&cat=podcasts&doctype=AudioVideo.

Bundesverfassungsgericht (1983). BVerfGE 65.1. Entscheidung des Bundesverfassungsgerichts. Bundesverfassungsgericht.

Bundeszentrale für politische Bildung (2016). "Interaktives Wahltool Wahl-O-Mat." Abrufdatum: 01.10.2016, von http://www.bpb.de/politik/wahlen/wahl-o-mat/.

BWahlG Bundeswahlgesetz.

BWO "Bundeswahlordnung (BWO)."

Cambridge Analytica (2016). "About us." Abrufdatum: 02.07.2016, von https://cambridgeanalytica.org/about.

Catalist (2016). "Who we are.". Abrufdatum: 31.10.2016, von http://www.catalist.us/.

Christlich Demokratische Union (2016). "Datenschutz." Abrufdatum: 10.10.2016, von https://www.cdu.de/datenschutz.

CRISP-DM (2016). "CRISP-DM Methodology." Abrufdatum: 21.09.2016, von http://crisp-dm.eu/home/crisp-dm-methodology/.

Der Bundeswahlleiter (2016). "Parteiunterlagen zum Download." Abrufdatum: 22.09.2016, von https://www.bundeswahlleiter.de/de/parteien/parteien_downloads.html.

Der Landeswahlleiter für Brandenburg (2016). "Bundestagswahl in Brandenburg am 22. September 2013. Ergebnisse der Bundestagswahl zum Download.". Abrufdatum: 24.09.2016, von https://www.wahlergebnisse.brandenburg.de/wahlen/BU2013/ErgebnisBerichte.asp?sel1=2155&sel2=0700.

Deutsche Post Direkt (2015). "Adressvermietung Consumer-Adressen Anfrage." Abrufdatum: 22.09.2016, von https://www.deutschepost.de/content/dam/dpag/images/D_d/DDP/Downloads/consumer/formular_anfrage_consumer-adressen-2016.pdf.

Deutscher Bundestag (2016). "Fundstellenverzeichnis der Rechenschaftsberichte." Abrufdatum: 22.09.2016, von http://www.bundestag.de/bundestag/parteienfinanzierung/rechenschaftsberichte/.

Die Landeswahlleiterin für Berlin (2016). "Wahlbezirksergebnisse ab 1990.". Abrufdatum: 22.09.2016, von https://www.wahlen-berlin.de/historie/hist_wahlendownload.asp?sel1=9500&sel2=1610.

Die Landeswahlleiterin, Statistisches Amt Saarland, , (2013). "Bundestagswahl Downloads." 24.09.2016, von http://www.statistikextern.saarland.de/wahl/internet_saar/BTW_BUND/download.html.

EStG "Einkommensteuergesetz (EStG)."

EU-DSGVO Verordnung zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung). Rat, Europäisches Parlament und Europäischer.

Fayyad, Usama, Piatetsky-Shapiro, Gregory und Smyth, Padhraic (1996). "From data mining to knowledge discovery in databases." AI magazine **17**(3): 37.

Feldman, Ronen (2013). "Techniques and applications for sentiment analysis." Communications of the ACM **56**(4): 82-89.

Forschungsgruppe Wahlen (2016). "Startseite." Abrufdatum: 27.05.2016, von http://www.forschungsgruppe.de/Startseite/.

Freie Demokratische Partei (2016). "frei-brief abonnieren." Abrufdatum: 10.10.2016, von https://www.fdp.de/page/newsletter.

Fürnkranz, Johannes, Gamberger, Dragan und Lavrač, Nada (2012). Foundations of rule learning, Springer Science & Business Media.

Gerber, Alan S und Green, Donald P (2000). "The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment." American Political Science Review **94**(03): 653-663.

Gerber, Alan S, Green, Donald P und Larimer, Christopher W (2008). "Social pressure and voter turnout: Evidence from a large-scale field experiment." American Political Science Review **102**(01): 33-48.

GG Grundgesetz für die Bundesrepublik Deutschland.

GRCh "Charta der Grundrechte der Europäischen Union."

GRECO Staatengruppe gegen Korruption (2016). Dritte Evaluierungsrunde. Zweiter Umsetzungsbericht zu Detuschland. "Kriminalisierung (SEV 173 und 91, Leitlinie 2)". "Transparenz der Parteienfinanzierung". Europarat.

Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter und Witten, Ian H (2009). "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter **11**(1): 10-18.

Hastie, Travor, Tibshirani, Robert und Friedman, Jerome (2008). The Elements of Statistical Learning, Springer.

Herrera, Franciso, Carmona, Cristóbal José, González, Pedro und Del Jesus, María José (2011). "An overview on subgroup discovery: foundations and applications." Knowledge and information systems **29**(3): 495-525.

Hessisches Statistisches Landesamt (2016). "Bundestagswahl." Abrufdatum: 22.09.2016, von https://statistik.hessen.de/zahlen-fakten/bundestagswahl.

IM Leipzig (2016). "Full Service. Alles aus einer Hand.". Abrufdatum: 26.06.2016, von http://www.imleipzig.de/de/leistungen/full-service/full-service.

infratest dimap (2016). "Leistungen." Abrufdatum: 27.05.2016, von http://www.infratest-dimap.de/leistungen/.

Insa Consulere (2016). "Leistungen." Abrufdatum: 27.05.2016, von http://www.insa-consulere.de/leistungen.html.

Institut für Demoskopie Allensbach (2016). "Das Institut für Demoskopie Allensbach - Porträt." Abrufdatum: 26.06.2016.

Janssen, Job, Schlote, Sara und Stolzenberg, Melanie (2013). "Die SPD klopft an. Von Tür zu Tür im neuen Stil." Abrufdatum: 22.09.2016, von https://mitmachen.spd.de/uploads/tx_news/Wegweiser_Tuer-zu-Tuer.pdf.

Keim, Nina und Rosenthal, Adrian (2016). Memes, Big Data und Storytelling. Rückblick auf den digitalen US-Wahlkampf 2012. Die US-Präsidentschaftswahl 2012, Springer.

Kosinski, Michal, Stillwell, David und Graepel, Thore (2013). "Private traits and attributes are predictable from digital records of human behavior." Proceedings of the National Academy of Sciences **110**(15): 5802-5805.

Lammert, Nobert (2016). Drucksache 18/8295. Unterrichtung durch den Präsidenten des Deutschen Bundestages., Deutscher Bundestag.

Landesamt für innere Verwaltung Mecklenburg-Vorpommern, Die Landeswahlleiterin, , (2016). "Endgültige Ergebnisse. Bundestagswahl am 22. September 2013.". Abrufdatum: 24.09.2016, von http://www.mv-laiv.de/Wahlen/Bundestagswahlen/2013/Ergebnisseite/.

Landesamt für Statistik Niedersachsen (2016). "LNS-Online Regionaldatenbank." Abrufdatum: 24.09.2016, von http://www1.nls.niedersachsen.de/statistik/default.asp.

Landeswahlleiter Rheinland-Pfalz (2016). "Bundestagswahl 2013. Wahlergebnisse im CSV-Format.". Abrufdatum: 24.09.2016, von http://www.wahlen.rlp.de/btw/wahlen/2013/downloads/index.html.

Landeswahlleiterin Statistisches Landesamt Sachsen-Anhalt (2016). "Wahl des 18. Deutschen Bundestages am 22. September 2013. Ergebnisse in Sachsen-Anhalt.". Abrufdatum: 22.09.2016, von http://www.statistik.sachsen-anhalt.de/wahlen/bt13/index.html.

Mitchell, Thomas M (1997). "Machine learning." New York.

Mitchell, Tom Michael (2006). The discipline of machine learning, Carnegie Mellon University, School of Computer Science, Machine Learning Department.

NationBuilder (2016). "Everything you need to win your election.". Abrufdatum: 20.09.2016, von http://nationbuilder.com/software_for_political_campaigns.

NationBuilder (2016). "Voter data use terms and conditions." Abrufdatum: 21.09.2016, von http://nationbuilder.com/voterdata.

NGP VAN (2016). "NGP VAN." Abrufdatum: 18.09.2016, von https://www.ngpvan.com/about.

Nickerson, David W und Rogers, Todd (2010). "Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making." Psychological Science **21**(2): 194-199.

Nickerson, David W und Rogers, Todd (2014). "Political campaigns and big data." The Journal of Economic Perspectives **28**(2): 51-73.

Niedermayer, Oskar (2015). "Parteimitglieder in Deutschland."

Nikiforakis, Nick, Kapravelos, Alexandros, Joosen, Wouter, Kruegel, Christopher, Piessens, Frank und Vigna, Giovanni (2013). Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. Security and privacy (SP), 2013 IEEE symposium on, IEEE.

PartG Gesetz über die politischen Parteien (Parteiengesetz).

Quinlan, J. Ross (1986). "Induction of decision trees." Machine learning **1**(1): 81-106.

reddit (2012). "I am Barack Obama, President of the United States -- AMA." Abrufdatum: 01.10.2016, von https://www.reddit.com/r/IAmA/comments/z1c9z/i_am_barack_obama_president_of_the_united_states/.

Rogers, Todd, Fow, Craig R und Gerber, Alan S (2013). "Rethinking why people vote." The behavioral foundations of public policy: 27.

Rubinstein, Ira S (2014). "Voter privacy in the age of big data." Wis. L. Rev.: 861.

Sagiroglu, Seref und Sinanc, Duygu (2013). Big data: A review. Collaboration Technologies and Systems (CTS), 2013 International Conference on, IEEE.

Selk, Robert (2016). "Datenschutz bei Payback." Abrufdatum: 21.10.2016, von https://www.payback.de/pb/id/252514/.

Sozialdemokratische Partei Deutschlands (2016). "Datenschutz." Abrufdatum: 10.10.2016, von https://www.spd.de/site/datenschutz/.

Statista, Das Statistik-Portal, , (2016). "Anzahl der Nutzer von Facebook und Instagram in Deutschland im Jahr 2016 (in Millionen)." Abrufdatum: 21.10.2016, von https://de.statista.com/statistik/daten/studie/503046/umfrage/anzahl-der-nutzer-von-facebook-und-instagram-in-deutschland/.

Statista, Das Statistik-Portal, , (2016). " Wahlbeteiligung bei US-Präsidentschaftswahlen von 1908 bis 2012 ". Abrufdatum: 21.07.2016, von https://de.statista.com/statistik/daten/studie/2184/umfrage/wahlbeteiligung-bei-us-praesidentschaftswahlen/.

Statistische Ämter des Bundes und der Länder (2015). "Zensus 2011. Methoden und Verfahren.".

Statistisches Amt für Hamburg und Schleswig-Holstein (2013). "Ergebnisse der Bundestagswahl 2013 in Hamburg." Abrufdatum: 29.09.2016, von http://www.statistik-nord.de/wahlen/wahlen-in-hamburg/bundestagswahlen/2013/.

Statistisches Amt für Hamburg und Schleswig-Holstein (2016). "Bundestagswahl 2013. Endgültiges Ergebnis für Schleswig-Holstein.". Abrufdatum: 24.09.2016, von http://www.statistik-nord.de/wahlen/wahlen-in-schleswig-holstein/bundestagswahlen/2013/.

Statistisches Bundesamt (2016). "Der Mikrozensus stellt sich vor." Abrufdatum: 23.10.2016, von https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Bevoelkerung/Mikrozensus.html.

Statistisches Bundesamt (2016). "Genesis-Online Datenbank. Themen.". Abrufdatum: 29.10.2016, von https://www-genesis.destatis.de/genesis/online/data;jsessionid=4BDA3D3012D8D49FDD88D85DA5C474DA.tomcat_GO_1_1?operation=statistikenVerzeichnis.

Statistisches Landesamt Baden-Württemberg (2016). "Ergebnisse der Bundestagswahlen 2013 und 2009 als Datei." Abrufdatum: 24.09.2016, von https://www.statistik-bw.de/Wahlen/Bundestag/Download.jsp.

Statistisches Landesamt Bremen (2016). "Bundestagswahl (Zweitstimmen)." Abrufdatum: 24.09.2016, von http://www.statistik-bremen.de/soev/abfrage_csv.cfm?tabelle=25200.

Thüringer Landesamt für Statistik (2016). "Wahlen im Freistaat Thüringen." Abrufdatum: 24.09.2016, von http://www.wahlen.thueringen.de/bundestagswahlen/bw_wahlergebnisse.asp.

TNS Emnid (2016). "Über uns." Abrufdatum: 27.05.2016, von https://www.tns-emnid.com/ueber-uns/.

Tufekci, Zeynep (2014). "Engineering the public: Big data, surveillance and computational politics." First Monday **19**(7).

Tumasjan, Andranik, Sprenger, Timm Oliver, Sandner, Philipp G und Welpe, Isabell M (2010). "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." ICWSM **10**: 178-185.

United States Government (2006). "Register to vote in your state by using this postcard form and guide. For U.S. citizens.". Abrufdatum: 21.09.2016, von https://vote.usa.gov/assets/downloads/fvr-6-25-14-eng.pdf.

United States Government (2016). "Vote for you, your family, your community.". Abrufdatum: 27.09.2016, von https://vote.usa.gov/.

wahlrecht.de (2016). "Sonntagsfrage Bundestagswahl." Abrufdatum: 21.09.2016, von http://www.wahlrecht.de/umfragen/.

Weinmann, Philipp (2016). Wahlen und Direkte Demokratie: Demokratische Teilhabe im Spannungsfeld politischer Machtinteressen. Handbuch Politik USA, Christian Lammert, Markus B. Siewert, Boris Vormann: 243-263.

Witten, Ian H, Frank, Eibe und Hall, Mark A (2011). Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann.

Woyke, Wichard (1998). "Stichwort: Wahlen. 10." Auflage, Opladen: Leske und Budrich.

WStatG Gesetz über die allgemeine un die repräsentative Wahlstatistik bei der Wahl zum Deutschen Bundestag und bei der Wahl der Abgeordneten des Europäischen Parlaments aus der Bundesrepublik Deutschland (Wahlstatistikgesetz - WStatG).

Zensus 2011 (2016). "Ergebnisse des Zensus 2011 zum Download." von https://www.zensus2011.de/SharedDocs/Aktuelles/Ergebnisse/DemografischeGrunddaten.html.

Anhang A: Attribute im Datensatz

Themenbereich	Attribute
Bundesland	Bundesländer der Gemeinde, codiert mit den Zahlen 1-16
Bevölkerung	Einwohnerzahl zum 09.05.2011 Anteil an Männern an der Bevölkerung
Familienstand und Geschlecht	Ledige Personen/Männer/Frauen Verheiratete Personen/Männer/Frauen Verwitwete Personen/Männer/Frauen Geschiedene Personen/Männer/Frauen Personen/Männer/Frauen in eingetragener Lebenspartnerschaft Personen/Männer/Frauen mit verstorbenem eingetragenem Lebenspartner Personen/Männer/Frauen mit aufgehobener eingetragener Lebenspartnerschaft
Bevölkerung in 11 Altersklassen und Geschlecht	Personen/Männer/Frauen unter 3 Jahren Personen/Männer/Frauen zwischen 3 und 5 Jahren Personen/Männer/Frauen zwischen 6 und 14 Jahren Personen/Männer/Frauen zwischen 15 und 17 Jahren Personen/Männer/Frauen zwischen 18 und 24 Jahren Personen/Männer/Frauen zwischen 25 und 29 Jahren Personen/Männer/Frauen zwischen 30 und 39 Jahren Personen/Männer/Frauen zwischen 40 und 49 Jahren Personen/Männer/Frauen zwischen 50 und 64 Jahren Personen/Männer/Frauen zwischen 65 und 74 Jahren Personen/Männer/Frauen ab 75 Jahren
Bevölkerung nach Staatsangehörigkeitsgruppen	Deutsche Staatsangehörigkeit Staatsangehörigkeit aus einem EU27-Land Staatsangehörigkeit aus dem sonstigen Europa Staatsangehörigkeit aus der sonstigen Welt Sonstige Staatsangehörigkeit
Bevölkerung nach Geburtsland	Deutschland, EU27-Land, Sonstiges Europa, Sonstige Welt, Sonstige
Bevölkerung nach Religion	Anhänger der römischen-katholischen Kirche Anhänger der evangelischen Kirche Anhänger mit sonstigen Religionen, ohne Religion und Personen ohne Angabe
Bevölkerung nach Migrationshintergrund und -erfahrung	Personen ohne Migrationshintergrund Personen mit Migrationshintergrund Ausländer Ausländer mit eigener Migrationserfahrung Ausländer ohne eigene Migrationserfahrung Deutsche mit Migrationshintergrund Deutsche mit Migrationshintergrund mit eigener Migrationserfahrung Deutsche mit Migrationshintergrund ohne eigene Migrationserfahrung Deutsche mit Migrationshintergrund ohne eigene Migrationserfahrung mit beidseitigem Migrationshintergrund Deutsche mit Migrationshintergrund ohne eigene Migrationserfahrung mit einseitigem Migrationshintergrund
Personen mit Migrationserfahrung nach Zuzugsjahrzehnt	1956 bis 1959, 1960 bis 1969, 1970 bis 1979,1980 bis 1989, 1990 bis 1999, 2000 bis 2011, Unbekannter Zuzugszeitraum

Bevölkerung mit Migrationshintergrund nach Regionen	EU27-Land, Sonstiges Europa, Sonstige Welt, Unbekanntes Ausland
Bevölkerung nach Erwerbsstatus und Geschlecht	Ewerbsstatus Mann Erwerbsstatus Frau Erwerbspersonen Insgesamt/Männer/Frauen Erwerbstätige Personen/Männer/Frauen Erwerbslose Personen/Männer/Frauen Nichterwerbspersonen Insgesamt/Männer/Frauen
Erwerbstätige nach Stellung im Beruf	Stellung als Angestellte und Arbeiter Stellung als Beamte Stellung als Selbstständige mit Beschäftigten Stellung als Selbstständige ohne Beschäftigte Stellung als mithelfende Familienangehörige
Erwerbstätige Bevölkerung nach Beruf	Führungskräfte Akademische Berufe Techniker und gleichrangige nichttechnische Berufe Bürokräfte und verwandte Berufe Dienstleistungsberufe und Verkäufer Fachkräfte in Land-/Forstwirtschaft und Fischerei Handwerks- und verwandte Berufe Bediener von Anlagen/Maschinen und Montageberufe Hilfsarbeitskräfte Angehörige der regulären Streitkräfte
Erwerbstätige nach Wirtschaftszweig	Zweig Land-/Forstwirtschaft, Fischerei Produzierendes Gewerbe Bergbau und Verarbeitendes Gewerbe Energie-/Wasserversorgung, Abfallentsorgung Baugewerbe Zweig Handel, Gastgewerbe und Verkehr; IuK Handel, Reparatur von KFZ, Gastgewerbe Verkehr und Lagerei, Kommunikation Sonstige Dienstleistungen Finanz- und Versicherungsdienstleistungen Grundstücks-/Wohnungswesen, wirtschaftliche Dienstleistungen Öffentliche Verwaltung u.ä. Öffentliche und private Dienstleistungen (ohne öffentliche Verwaltung)
Personen in schulischer Ausbildung nach Klassenstufen	Klasse 1 bis 4 Klasse 5 bis 9 bzw. 10 (Sekundarstufe II) Klasse 11 bis 13 (Gymnasiale Oberstufe)
Personen in schulischer Ausbildung nach Schulform	Grundschule, Hauptschule, Realschule, Gymnasium, Gesamtschule, Sonstige Schule
Personen ab 15 Jahren nach höchstem schulischen Abschluss	Ohne Abschluss Noch in schulischer Ausbildung Haupt-/Volksschulabschluss Realschul- oder gleichwertiger Abschluss Schüler/-innen der gymnasialen Oberstufe Fachhochschulreife Allgemeine/fachgebundene Hochschulreife
Personen ab 15 Jahren nach höchstem beruflichen Abschluss	Ohne beruflichen Abschluss Lehre, Berufsausbildung im dualen System Fachschulabschluss Abschluss einer Fachakademie oder Berufsakademie Fachhochschulabschluss, Hochschulabschluss, Promotion

Tabelle 21 Verwendete Attribute

Anhang B: Rangfolgen des Wahlergebnisses

Die nachfolgende Abbildung zeigt die Rangfolgen der Parteiwahl und die Anzahl der Gemeinden, die die Parteien in dieser Rangfolge gewählt haben. Rangfolgen mit weniger als 10 Gemeinden wurden dabei aus Gründen der Übersichtlichkeit entfernt. Die am weitesten links stehende Partei erhielt prozentual die meisten Zweistimmen. Nach rechts hin nimmt der Stimmanteil ab.

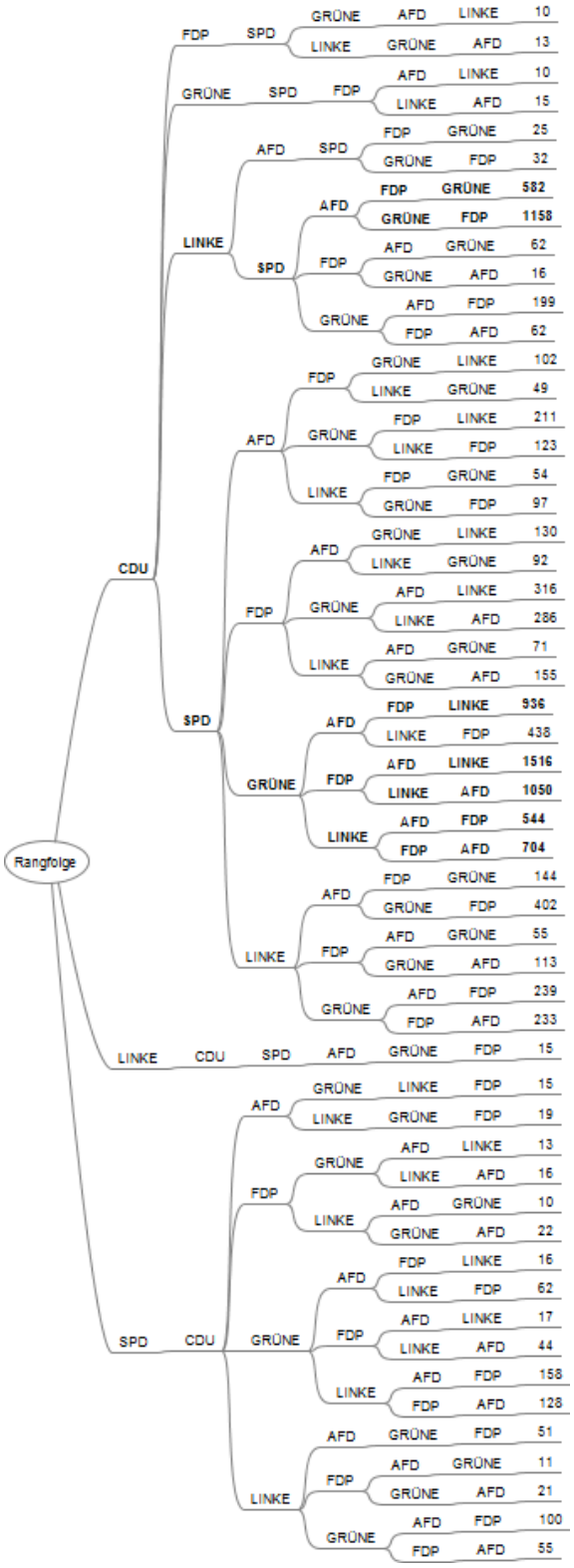


Abbildung 17 Rangfolge der Parteiwahl inklusive der Gemeindeanzahl