Die Statistische Methodik von Civey

Eine Einordnung im Kontext gegenwärtiger Debatten über das Für und Wider internetbasierter nicht-probabilistischer Stichprobenziehung

1 Einleitung

Die moderne Umfrageforschung befindet sich in einer Phase des Umbruchs: Lange Zeit als verlässlich geltende Methoden geraten zunehmend in die Kritik, insbesondere da sich die Ausschöpfungsquoten bei klassischen Erhebungen mittlerweile auf einem historisch niedrigem Niveau bewegen. Gleichzeitig drängen neue Anbieter auf den Markt, die diese Probleme mit neuen Erhebungsmethoden lösen wollen, dabei aber zugleich vor neuen statistischen Herausforderungen stehen. Die Diskussion wird sowohl in der Wissenschaft als auch in den Fachorganen der Markt- und Meinungsforschungsbranche kontrovers geführt. Im Folgenden möchten wir einen kurzen Überblick über die zentralen Streitpunkte der Debatte geben und die von Civey entwickelte Methodologie innerhalb des Diskurses verorten.

2 Die Debatte

2.1 Probability Sampling

Der klassische Ansatz der Stichprobenziehung geht auf den Statistiker Neyman (1934) zurück. Die von ihm entwickelte Theorie des Probability-Samplings und nahezu jedes seither publizierte Standardwerk folgen dabei dem gleichen Grundgedanken: Über das Design der Stichprobe kann jedem Individuum eine klare Wahrscheinlichkeit zugeordnet werden, zufällig in diese ausgewählt zu werden. Auf dieser Basis können aus der statistischen Theorie verlässliche Aussagen über relevante Größen in der Grundgesamtheit, zum Beispiel der Anteil von Zustimmung zu einer Aussage, hergeleitet werden.
Die Formulierung des Probability-Samplings repräsentiert einen Meilenstein der Survey-Statistik: Durch die Befragung einer kleinen Zufallsauswahl von Personen war es erstmals auf seriöser Grundlage möglich, Aussagen über eine Grundgesamtheit zu treffen. Über mehrere Jahrzehnte erfreute sich die in Neymans Tradition stehende Methode deshalb großer Beliebtheit in der Markt- und Meinungsforschung und galt insbesondere in der Privatwirtschaft häufig kombiniert mit Callcenter-basierten Telefonbefragungen als Branchenstandard.

Aus dem Grundgedanken des design-basierten Ansatzes folgen zwei zentrale Annahmen: Sämtliche Personen innerhalb der interessierenden Grundgesamtheit müssen einerseits über eine Auflistung (Frame) bekannt sein und andererseits über eine positive und berechenbare Wahrscheinlichkeit verfügen, in die Stichprobe zu gelangen. In Zeiten flächendeckender Festnetzverbreitung, verpflichtender Eintragungen in Telefonbücher und hoher Bereitschaft zur Teilnahme an Umfragen konnte diese zentrale Basis des Probability-Samplings als gegeben angenommen werden, sodass valide und verlässliche Rückschlüsse auf die Bevölkerung möglich waren.
Heutzutage ist telefonbasiertes Probability-Sampling jedoch mit schwerwiegenden Herausforderungen konfrontiert: Zuallererst ist hier die schwindende Bereitschaft der Bevölkerung zu nennen, an Telefonumfragen teilzunehmen. Aus den USA existieren hierzu detaillierte Zahlen: So sank der Anteil der tatsächlich erzielten Interviews an der ursprünglich ausgewählten Stichprobe (Ausschöpfungsquote) für die renommierte Gallup Poll Social Series von 28 % (1997) auf 7 % (2017), der Anteil der Erreichten, die sich bereit erklärten, am Interview teilzunehmen, halbierte sich in der gleichen Zeit (Marken, 2018). Auch die Ausschöpfungsquoten von Pew Research fielen von 1997-2018 von 36% auf gerade einmal 6% (Kennedy and Hartig, 2019). Für die Marktforschung insgesamt spricht der Branchenverband AAPOR alleine im Zeitraum 2008-2015 von einem Einbruch von 15.7% auf 9.3% (Festnetz) bzw. 11.7% auf 7.0% (Mobilfunk).
Dies ist bedenklich, da ein Zusammenhang zwischen der Ausschöpfungsquote und der durchschnittlichen Verzerrung innerhalb einer Umfrage existieren kann (Brick and Tourangeau, 2017). Je geringer die Bereitschaft zur Teilnahme, desto größer das Risiko, dass diejenigen, die tatsächlich in die Stichprobe gelangen, systematische Unterschiede zur Grundgesamtheit aufweisen. Im Kampf gegen Umfragemüdigkeit müssen deshalb teure Anreize bereitgestellt oder aufwendige flexible Survey-Designs genutzt werden, die den Sampling-Prozess adaptiv an das Antwortverhalten anpassen (Kinney and Cooney, 2019). Telefonumfragen leiden weiterhin unter einer sinkenden Festnetz-Penetration, welche komplexe Dual-Frame-Ansätze notwendig macht, durch die zusätzlich Mobilfunknutzer befragt werden. Smartphone-Betriebssysteme, Apps und Provider blockieren jedoch zunehmend potentielle Spam-Anrufe unbekannter Nummern, sodass Befragte in Zukunft noch schwieriger zu erreichen sein werden (Dutwin et al., 2018).
Etablierte wissenschaftliche Großbefragungen können die durch diese veränderten Umstände entstehenden, teils erheblichen Kostensteigerungen (Blohm and Koch, 2015) zu Gunsten verlässlicher Ergebnisqualität in Kauf nehmen. Dies garantiert weiterhin eine qualitativ hochwertige Grundlage für sozial- und verhaltenswissenschaftliche Forschung. Private Unternehmen sind hierzu oftmals nicht in der Lage, ohne ihre Wirtschaftlichkeit zu gefährden. Die einzige kostenneutrale Alternative zum Ausgleich der niedrigen Ausschöpfungsquoten sind signifikante Nachgewichtungen, welche jedoch Neymans Grundidee zuwiderlaufen, da sie von Annahmen abhängen, die auch im Kontext des Non-Probability Samplings notwendig sind. Es ist deshalb fraglich, inwieweit die tatsächliche Praxis der Markt- und Meinungsforschung heutzutage noch den Ansprüchen des Prob-

ability Samplings genügen kann oder ob die Voraussetzungen dieses Paradigmas so vollumfänglich verletzt sind, dass dieser Begriff nicht mehr angebracht ist.

2.2 Online Non-Probability Sampling

Sowohl vor als auch nach Neymans Durchbruch wurden auch andere Methoden der Stichprobenziehung genutzt, die nicht allen Elementen der Grundgesamtheit eine eindeutige, positive Auswahlwahrscheinlichkeit zuordnen (Stephan and McCarthy, 1958). Diese werden auch als Non-Probability Sampling-Methoden bezeichnet. Während sie nach der Etablierung des Probability-Samplings zunächst nicht mehr als eine Randerscheinung waren, änderte sich dies mit der wachsenden Verbreitung des Internets und dem fortschreitenden Fall der Ausschöpfungsquoten.
Mittlerweile mehren sich Stimmen in der Forschung, die sich für die Nutzung von Online Non-Probability-Sampling aussprechen (Wang et al., 2015; Ansolabehere and Rivers, 2013; Ansolabehere and Schaffner, 2014; Goel et al., 2015). Neben beeindruckenden Erfolgen in der Prädiktion von Wahlergebnissen spricht hierfür die gerade im Kontext der Marktforschung große Zweckeignung (fit for purpose) im Hinblick auf praktisch benötigte Präzision, Ergebnisgranularität, Durchführbarkeit, Geschwindigkeit und Kosten (AAPOR, 2015). Online-Non-Probability Surveys sind häufig günstiger, ermöglichen die Erhebung großer Stichproben und die Befragung kleiner Zielgruppen und können schnell und unkompliziert durchgeführt werden. Zudem verringert die Anonymität der Online-Befragung das Risiko unehrlichen Antwortverhaltens als Ergebnis sozialer Erwünschtheit sowohl im Vergleich zu Telefon- (Keeter, 2015) als auch papierbasierten Umfragen (Gnambs and Kaspar, 2015).
Andere Autoren widersprechen dieser Initiative auf Basis von empirischen (MacInnis et al., 2018) und theoretischen (Quatember, 2019) Argumenten. Probability-Sampling kann als einheitlicher theoretischer Rahmen betrachtet werden, da sämtliche verwendeten Methoden von der Existenz einer bekannten Ziehungswahrscheinlichkeit größer Null für jedes Individuum in der Grundgesamtheit ausgehen. Die Ansätze des Non-Probability-Samplings vereint hingegen primär die Abwesenheit jener Wahrscheinlichkeit. Das bedeutet, dass sie auf eine sorgfältige Berücksichtigung des Selektionsmechanismus angewiesen sind, der bestimmt, ob ein Individuum Teil des Samples wird oder nicht. Inwieweit dieser Nachteil in der Praxis der Marktforschung von Relevanz ist, kann allerdings bezweifelt werden. So müssen (vermeintliche) Probability-Samples wie bereits angedeutet auf vergleichbare Korrekturen zurückgreifen, um die bereits angesprochenen Verzerrungen im Antwortverhalten bedingt durch extrem niedrige Ausschöpfungsquoten auszugleichen (Rivers, 2013).
Zudem wurden in den vergangenen Jahren maßgebliche Fortschritte in der theoretischen Fundierung des Non-Probability-Samplings gemacht, sodass heute gemeinhin zwischen zwei etablierten Methoden für valide Schlüsse aus diesen unterschieden wird: der auf Sample-Matching basierenden Quasi-Randomisierung

und der Superpopulationsmodellierung[1]: Während Quasi-Randomisierungsansätze mit Hilfe einer probabilistischen Stichprobe von hoher Qualität versuchen, künstliche Ziehungswahrscheinlichkeiten für Personen in einem Non-Probability-Sample zu bestimmen, schätzen Superpopulationsansätze unter Verwendung von in der Bevölkerung entweder auf Individual- oder auf Aggregatebene bekannten Informationen und einem im Non-Probability-Sample gebildeten Modell die Antworten aller Personen in der Grundgesamtheit.

Hierauf aufbauend existieren klare Voraussetzungen, unter denen Non-Probability-Sampling valide Ergebnisse liefern kann. Zentral ist insbesondere die Annahme der Ignorable Sample Selection bzw. Selection at Random. Diese ist eng mit dem im Kontext der Kategorisierung fehlender Daten genutzten Missing at Random-Konzept verbunden und wird zum Beispiel von Rivers (2013) ausführlich beschrieben. Praktisch bedeutet dies, dass gegeben der im Rahmen des Modells oder des Matchings verwendeten Variablen der Selektionsmechanismus unabhängig vom Antwortverhalten (bzgl. der interessierenden Analysevariable) der Nutzer sein muss, um valide Ergebnisse zu garantieren. Diese Annahme wird im Kontext des bei Civey verwendeten Modells im technischen Anhang genauer diskutiert.

3 Methodisches Vorgehen bei Civey

Die bei Civey verwendete Methodik basiert im Kern auf einer Kalibration via Raking zur Berechnung der regulären Ergebnisse (welche man als implizite Superpopulationsmodellierung betrachten kann, wie im Anhang genauer erläutert wird) und einer expliziten Superpopulationsmodellierung durch Mehrebenen-Regression und Poststratifizierung für kleinräumige Schätzungen auf Bundesland-, Bezirks- und Kreisebene. Sie lässt sich somit in die beschriebene Systematik der Non-Probability-Sampling-Theorie einfügen und orientiert sich an den Anforderungen der aktuellen Fachliteratur.
Die zentralen Elemente des methodischen Vorgehens bei Civey werden im Folgenden kurz erläutert, eine umfassende mathematische Darstellung findet sich im Anhang.

3.1 Rekrutierung des Panels

3.1.1 River-Sampling

Der erste zentrale Schritt ist die Rekrutierung von Umfrageteilnehmern. Civey-Umfragen werden tagtäglich auf mehr als 25 000 einzigartigen URLs über ein Netzwerk von mehreren Dutzend reichweitenstarken Seiten von Medienpartnern, Email-Providern und zahlreichen Privatnutzern und Blogs eingebunden, die unterschiedliche thematische Schwerpunkte aufweisen. Dieses auch als "River-Sampling" bezeichnete Verfahren erlaubt den Zugriff auf einen deutlich größeren und

[1] Eine detaillierte Beschreibung beider Ansätze kann bei Valliant et al. (2018) gefunden werden.

diverseren Pool von Befragten; gerade kleine demographische Gruppen können so effizient erreicht werden (Callegaro, 2014). Dabei wird sichergestellt, dass Umfragen über die URLs gleichmäßig an die zu befragende Zielgruppe und über die erhobene Zeit hinweg ausgespielt werden.

3.1.2 Incentivierung

Nahezu sämtliche Online Non-Probability-Panels sind gezwungen, potentiellen Nutzern eine Form von finanziellem Anreiz (Geld, Gutscheine, Spenden, etc.) anzubieten, um sie zur Beantwortung von Fragen zu motivieren. Stammt ein signifikanter Teil der Antworten im Panel von Personen, welche ausschließlich aus monetärem Eigeninteresse an der Umfrage teilnehmen, sog. "Professionelle Befragte" (Hillygus et al., 2014), könnte dies negative Auswirkungen auf die Datenqualität haben, da professionelle Befragte unter Umständen primär an einer Maximierung ihrer finanziellen Belohnung und nicht an sorgfältigen und korrekten Antworten interessiert sind. Dies kann sich unter anderem in bewussten Falschantworten zur Vergrößerung der Menge von zu beantwortenden Umfragen äußern (Guin et al., 2006).
Ein wesentliches Alleinstellungsmerkmal des bei Civey genutzten Vorgehens ist der komplette Verzicht auf derartige Maßnahmen, sodass eine Verzerrung durch professionelle Befragte ausgeschlossen werden kann. Der Anreiz zur Umfrageteilnahme bei Civey entsteht hingegen durch die exklusive Sicht auf die gewichteten Umfrageergebnisse.

3.1.3 Verhinderung von Manipulation

Zur Verhinderung von Manipulation greifen wir auf ein breites Spektrum von dem Stand der Forschung entsprechenden Methoden zurück, wie sie zum Beispiel von Teitcher et al. (2015) beschrieben werden.
In den Berechnungen werden ausschließlich verifizierte Teilnehmende berücksichtigt. Diese Verifizierung umfasst eine niedrigschwellige Registrierung, bei der Befragte grundlegende Soziodemographika und ein Einverständnis zur Datenverarbeitung abgeben. Weitere Abstimmungen ordnet Civey über E-Mail-Adressen, Authentifizierungs-Token und Cookies den entsprechenden Befragten zu. Darüber hinaus prüfen wir im Rahmen einer fortlaufenden Verifizierung, ob der einzelne Nutzer eine echte Person ist, ausreichend Daten für eine spätere Gewichtung vorliegen und mit welcher Wahrscheinlichkeit seine Angaben der Wahrheit entsprechen. Hierfür werden technische, statistische und inhaltliche Plausibilitätschecks genutzt, darunter Kriterien wie Mausbewegungen des Nutzers, das Klickverhalten und Geschwindigkeit der Teilnahme sowie die inhaltliche Plausibilität beziehungsweise Widersprüche in gegebenen Antworten. Mehrfache Antworten eines Nutzers auf die gleiche Frage sind innerhalb eines fragespezifischen Zeitraums nicht möglich. Eine gezielte Manipulation der Ergebnisse ist somit ausgeschlossen.

3.2 Berechnung von Anteilen

3.2.1 Quota-Sampling

Aus allen innerhalb eines frageabhängigen Zeitfensters abgegebenen Antworten wird eine quotierte Stichprobe von einer vorher festgelegten Größe (üblicherweise 5000 Befragte) gezogen. Dabei muss die Verteilung der Stichprobe im Hinblick auf bestimmte demographische Variablen wie Alter, Geschlecht und Wahlverhalten der (z.B. aus administrativen Daten) bekannten Verteilung innerhalb der Grundgesamtheit entsprechen. Da durch den inhaltlichen Kontext der Einbindung unserer Umfragen ins journalistische Angebot unserer Medienpartner Framing-Effekte (Stalans, 2012) nicht ausgeschlossen werden können, wird die Antwort eines Nutzers für die weitere Analyse nur dann verwendet, wenn ihm diese durch den Civey-Algorithmus randomisiert ausgespielt wurde. Die Information über die erste beantwortete Frage eines Nutzers innerhalb einer zusammenhängenden Sitzung wird somit üblicherweise nicht verwendet.[2]
Dieser Prozess hat nur wenig Gemeinsamkeiten mit der klassischen quotierten Stichprobe der analogen Welt, in welcher ein Interviewer nach eigenem Ermessen entscheidet, ob eine Person befragt werden soll oder nicht. Einige gut gewählte Quoten können so online bereits eine signifikante Verzerrung des Ergebnisses verhindern und vermeiden exzessive Gewichtung im weiteren Verlauf der Berechnungen. (Rivers, 2007)

3.2.2 Raking

Etwaige nach dem Quota-Sampling noch verbleibende Abweichungen zwischen Stichprobe und Grundgesamtheit im Hinblick auf bekannte Variablen (z.B. Alter, Geschlecht,Wahlverhalten, Parteineigung und geographische Verteilung) werden durch eine Raking-basierte Gewichtung beseitigt. Diese ursprünglich von Deming and Stephan (1940) eingeführte Methode wird klassischerweise im Kontext des Probability Samplings zur Korrektur von Verzerrungen durch niedrige Ausschöpfungsquoten genutzt. Sie kalibriert die Gewichtung der Beobachtungen in einer Stichprobe so, dass sie mit den Randverteilungen einer oder mehrerer Variablen übereinstimmt, welche zum Beispiel aus der amtlichen Statistik bekannt sind. In der Stichprobe unterrepräsentierte Gruppen erhalten somit ein höheres Gewicht, überrepräsentierte ein niedrigeres.
Der Non-Probability-Terminologie von Valliant et al. (2018) folgend kann das Raking als eine Art implizites Superpopulationsmodell betrachtet werden. Superpopulationsmodelle zeichnen sich generell dadurch aus, dass der Zusammenhang zwischen dem Antwortverhalten und einer Reihe von über die Befragten

[2] Die einzige Ausnahme hierzu ist die kurze Phase in den ersten Stunden direkt nach der Veröffentlichung einer Umfrage, welche in einen populären Artikel eingebunden ist. In dieser Situation kann es theoretisch passieren, dass eine große Mehrheit aller Antworten nicht zufällig ausgespielt worden ist. Um dem Nutzer in dieser Situation trotzdem eine Auswertung der Ergebnisse zu präsentieren, werden unter Umständen ebenfalls Daten genutzt, die aus ersten Fragen stammen. In diesem Fall wird die Unsicherheit in den Ergebnissen durch eine konservative Korrektur des statistischen Fehlers widergespiegelt.

vorliegenden Informationen explizit (zum Beispiel mit Hilfe einer linearen Regression, wie beim sog. GREG-Schätzer, Särndal et al. (1992)) modelliert wird, um Vorhersagen für jedes Individuum (oder Totalwerte) in der Grundgesamtheit vorherzusagen (Elliott and Valliant, 2017). Obwohl vergleichsweise simpel und robust, liefert Raking Ergebnisse, die vergleichbar mit jenen deutlich komplexerer Superpopulationsmodelle sind (Valliant, 2019). Ein detaillierter Überblick über den Zusammenhang von Superpopulationsmodellen, ihren Annahmen und Raking ist im technischen Anhang zu finden.
Die so erzeugten Gewichte können allerdings eine erhebliche Variation aufweisen: Manche Beobachtungen in stark unterrepräsentierten Gruppen verfügen unter diesen Umständen über einen deutlich größeren Einfluss auf das Ergebnis als der durchschnittliche Befragte, was die Unsicherheit der Ergebnisse erhöht. Zu diesem Zweck wird auf das Verfahren des Weight-Trimmings zurückgegriffen, welches einen Maximal- und einen Minimalwert für die kalibrierten Gewichte festsetzt[3].

3.2.3 Unsicherheitsmaß des Ergebnisses

Es existiert gegenwärtig kein allgemein akzeptierter Standard zur Bestimmung von Unsicherheitsmaßen für die Schätzung interessierender Größen in nicht-probabilistischen Stichproben. Deshalb orientieren wir uns an den von der AAPOR aufgestellten Richtlinien zur Angabe von Unsicherheitsmaßen für Non-Probability-Surveys (AAPOR, 2015) und berechnen die Unsicherheit unserer Ergebnisse mit Hilfe eines bayesianischen 95%-Kredibilitätsintervall basierend auf einem Beta-Binomial-Modell unter Verwendung einer nicht-informativen Bayes-Laplace Prior-Verteilung(Tuyl et al., 2008) mit maximal konservativen Annahmen und einer Pseudo-Design-Effekt-Korrektur nach Kish (1992), wodurch der durch die Gewichtung gestiegenen Varianz Rechnung getragen wird. Die mathematischen Details dieses Ansatzes sind im technischen Anhang dargelegt. Obwohl in der Markt- und Meinungsforschung in verschiedenen Ausprägungen gängige Praxis[4], existiert keine rigorose theoretische Begründung dieser Korrektur. Interne Tests zeigen jedoch, dass die so generierte Maßzahl extrem nahe an mit Hilfe komplexerer Verfahren generierter Schätzungen des statistischen Fehlers liegt, insbesondere dem Generalized Raking-Varianzschätzer.

3.2.4 Kleinräumige Schätzung

Neben der Bestimmung von Bevölkerungsanteilen und ihrer Aufschlüsselung nach demographischen Subgruppen berechnet Civey auf Anfrage kleinräumige geographische Schätzungen für Bundesländer, Regierungsbezirke und Landkreise.

[3] Verschiedene Methoden des Weight-Trimmings existieren, ohne dass sich ein dominanter Ansatz etabliert hätte, einige werden z.B. von Potter (1990) beschrieben. Wir orientieren uns an der unter anderem vom Sozio-Ökonomischen Panel verwendeten Obergrenze vom 10-fachen Wert des Median-Gewichts und ergänzen diese um eine Untergrenze in Höhe des 0.1-fachen. Die getrimmten Werte werden auf alle anderen Gewichte iterativ umverteilt, sodass die Summe konstant bleibt.

[4] Vgl. zum Beispiel die Nutzung bei Pew Research (Smith, 2010) und YouGov (2015)

Häufig liegen jedoch nicht genug Beobachtungen vor, um verlässliche Ergebnisse mit Hilfe der oben beschriebenen Kombination aus Quota-Sampling und Raking für jeden Regierungsbezirk oder gar jeden der 401 deutschen Landkreise und kreisfreie Städte zu erhalten. Derartige Probleme werden in der Survey-Statistik unter dem Schlagwort Small Area Estimation[5] erforscht, eine Disziplin, welche sich seit Jahrzehnten konstant weiterentwickelt (Pfeffermann, 2013). Die Lösungen der Small Area Estimation bauen dabei meist auf der effizienten Nutzung aller verfügbaren Daten durch partial Pooling[6] und der Verwendung von Hilfsinformationen (Arbeitslosigkeit, Kaufkraft, etc.) auf, die für jedes der Gebiete, für die eine Vorhersage erfolgen soll, vorliegen.
Insbesondere im Kontext von Non-Probability-Stichproben hat sich der von Park et al. (2004) entwickelte Mehrebenen-Regression und Poststratifizierungs-Ansatz (MRP) als beliebte Methode der Small Area Estimation erwiesen (Wang et al., 2015; Shirley et al., 2014; Hoover and Dehghani, 2018), welche den spezifischen Anforderungen von Civey entsprechend in einer modifizierten Version adaptiert wurde. Zentrale Veränderungen betreffen dabei die automatische Auswahl der im Modell verwendeten Kovariaten mit Hilfe von L1-Regularisierung (LASSO) nach Vincent and Hansen (2014) und die Verwendung einer Approximation des rechenintensiven vollständig bayesianischen multinomialen gemischten logistischen Regressionsmodells durch mehrere binäre gemischte logistische Regressionen. Weiterhin wird eine synthetische Poststratifizierung (Leemann and Wasserfallen, 2017) basierend auf mehreren partiellen gemeinsamen Verteilungen, für welche amtliche Daten existieren, durchgeführt, sofern keine vollständige Poststratifizierung möglich ist. Die Details dieser Methode sind im technischen Anhang nachzuvollziehen.

4 Fazit

Die Markt- und Meinungsforschung ist geprägt von einer weitreichenden Methodendebatte: Der klassische, erprobte und statistisch seit Jahrzehnten umfassend erforschte Ansatz des Probability-Samplings ist in Zeiten dramatisch fallender Ausschöpfungsquoten unter ökonomischen Rechtfertigungsdruck geraten. Online Non-Probability-Sampling verspricht vor diesem Hintergrund eine den praktischen Erfordernissen der Meinungsforschung häufig besser entsprechende Alternative. In den vergangenen Jahren wurden wesentliche Schritte zur statistischen Fundierung des Ansatzes unternommen. Die von Civey genutzte Methodik baut auf diesem theoretischen Fundament auf und liefert mit Hilfe einer Kombination aus Quota-Sampling und Raking-basierter Superpopulationsmodellierung unter klar definierbaren Annahmen (insbesondere der Unabhängigkeit von Antwortverhalten und Selektion ins Panel unter Berücksichtigung der Gewichtungsvariablen) valide Ergebnisse. Diese werden durch kleinräumige Schätzungen auf Basis einer adaptierten Form von Multilevel-Regression und Poststratifizierung ergänzt.

[5] Siehe das Standardwerk von Rao and Molina (2015) für einen umfassenden Überblick
[6] Siehe die Einführung von Gelman and Hill (2006) für eine genauere Definition.

5 Technischer Anhang

5.1 Superpopulationsmodellierung

Im Folgenden geben wir einen Überblick über die grundsätzlichen Annahmen der Superpopulationsmodellierung. In der Praxis greifen wir wie bereits angedeutet mit dem sog. Raking auf eine Methode zurück, die klassischerweise im Kontext der Kalibration eingeordnet wird. In einem Non-Probability-Kontext lassen sich die Annahmen dieses Ansatzes jedoch schlüssiger über Superpopulationsmodelle motivieren.

Im Rahmen des Superpopulationsansatzes[7] unterstellen wir einen Zusammenhang zwischen der interessierenden Variable $y_k, k = 1, ..., K$ und einer Reihe von L Kovariaten $x_k = x_{k1}, ..., x_{kL}$ auf Basis von Parametern Θ:

$$E(Y) = f(X; \Theta) \tag{1}$$

Hierbei ist Y der $1 \times K$-Vektor der abhängigen Variable und X die $K \times L$-Matrix der Kovariaten.
Im Kontext einer begrenzten Grundgesamtheit U interessieren wir uns für alle $k \in U$. U könnte in diesem Sinne z.B. die Wahlbevölkerung Deutschlands und K die Anzahl der Personen sein, die dieser Grundgesamtheit zugehörig sind. Ein Teil dieser Grundgesamtheit bildet die Stichprobe $s \subset U$ (wobei $s \cup \bar{s} = U$), für welche y_k, x_k als bekannt gelten können. x_k kann für alle $k \in \bar{s}$ vorliegen, meist ist jedoch nur eine Reihe von Totalwerten t_{x_U} (z.B. die Anzahl von Männern und Frauen) als aggregierte Information aus externen Quellen bekannt.

5.1.1 Annahmen

Sei $\gamma_s = \gamma_1, ...\gamma_K$ der Vektor binärer Indikatorvariablen, die die Zugehörigkeit zu s angeben. Die gemeinsame Dichte von γ_s gegeben X, Y auf Basis von Parametern Φ ist $f(\gamma_s|X, Y; \Phi)$. Das gemeinsame Modell für Y, γ_s ist somit:

$$f(Y, \gamma_s|X; \Theta, \Phi) = f(Y|X; \Theta) f(\gamma_s|Y, X; \Phi) \tag{2}$$

Unser Interesse gilt Rückschlüssen über Y in der Grundgesamtheit U, welche sich aus den Beobachtungen in der Stichprobe Y_s und allen weiteren Beobachtungen $Y_{\bar{s}}$ zusammensetzt. Hierfür sind Prädiktionen für die $Y_{\bar{s}}$ notwendig. Dies setzt voraus, dass unser Modell $E(Y_s) = f(X_s; \Theta)$ als Modell eine unverzerrte Vorhersage $\hat{Y}_{\bar{s}}$ zulässt. Hierfür muss Y bedingt auf X unabhängig von γ_s sein:

$$f(\gamma_s|Y, X; \Phi) = f(\gamma_s|X; \Phi) \tag{3}$$

Dies erlaubt es, den Selektionsmechanismus $f(\gamma_s|Y, X; \Phi)$ im weiteren Verlauf im Sinne der Selection at Random-Annahme zu ignorieren (Rubin, 1976).

[7] Sofern nicht anders angegeben, orientieren sich die folgenden Ausführungen an Valliant et al. (2018).

5.1.2 Generalisierte Regression

Ist eine Verzerrung des Ergebnisses durch den Selektionsprozess ausgeschlossen, kann z.B. mit Hilfe eines Regressionsmodells Y durch X vorhergesagt werden. Ausgehend von einem linearen und identischen Zusammenhang zwischen x und y für alle $k \in s, k \in \bar{s}$ ist $\Theta = \beta$, sodass

$$E(y_k) = x_k'\beta. \tag{4}$$

Setzen wir X_s als Matrix aller $x_k, k \in s$, so lässt sich β mit dem klassischen OLS-Schätzer bestimmen:

$$\hat{\beta} = (X_s'X_s)^{-1}X_s'y_s. \tag{5}$$

Das geschätzte Total von y, $\hat{t}_y$ ergibt sich somit als

$$\hat{t}_y = \sum_{k\in s} y_k + \sum_{k\in\bar{s}} \hat{y}_k = \sum_{k\in s} y_k + (t_{x_U} - t_{x_s})'\hat{\beta} \tag{6}$$

Aus diesem Totalwert können ohne weiteres Anteile oder Durchschnitte ermittelt werden.
$\hat{t}_y$ lässt sich zudem als gewichtete Summe der beobachteten y_k ausdrücken:

$$\hat{t}_y = \sum_{k\in s} w_k y_k \tag{7}$$

Aus Gleichung (5) und (6) folgt, dass die w_k sich als unabhängig von y darstellen lassen:

$$w_k = 1 + (t_{x_U} - t_{x_s})'(X_s X_s')^{-1}x_k \tag{8}$$

Diese Ergebnisse entsprechen dem Generalisierten Regressionsschätzer (GREG) im Kontext des Probability-Samplings unter Annahme einer einfachen Zufallsstichprobe, also in Abwesenheit von Design-Effekten.

5.1.3 Raking

Die Gewichtung anhand univariater Häufigkeiten kategorialer Variablen (Raking) kann als Spezialfall dieses GREG-Modells interpretiert werden, wie Särndal et al. (1992) erläutert. Betrachten wir hierzu o.B.d.A. die Situation, in der die Randverteilungen $N_{g,*}, N_{*,j}$ zweier Hilfsvariablen, x_1, x_2 mit den Kategorien $g = 1, ..., G$ bzw. $j = 1, ...J$, jedoch nicht die gemeinsame Verteilung $N_{g,j}$, bekannt sind. Der Kovariatenvektor x für eine Beobachtung k entspricht dann einer Reihe von $G + J$ Indikatorvariablen δ, welche die Zugehörigkeit von k zu einer Kategorie j bzw. g mit dem Wert 1 markieren und ansonsten für die restlichen $G + J - 2$ δ gleich 0 sind:

$$x_k = (\delta_{1,*}, ..., \delta_{G,*}, \delta_{*,1}, \delta_{*,J}) \tag{9}$$

Das Total von x_k entspricht somit der Häufigkeitsverteilung von x_1 und x_2:

$$\sum_{k \in U} x_k = (N_{1,*}, ..., N_{G,*}, N_{*,1}, ..., N_{*,J}) \tag{10}$$

Die Gewichte können dann entsprechend Gleichung (8) berechnet werden und sind sehr eng verbunden mit dem von Deming and Stephan (1940) vorgeschlagenen iterativen Raking-Verfahren[8], auf welches Civey zurückgreift. Insofern lassen sich die zuvor genannten Grundannahmen des Superpopulationsansatzes auf das River-Sample von Civey und die genutzte Methodik übertragen, sodass von einem impliziten Superpopulationsmodell gesprochen werden kann.

5.2 Korrigiertes Kredibilitätsintervall

Wir sind primär an der Schätzung von Anteilen interessiert. Deren Varianz wird über ein bayesianisches Beta-Binomial-Modell bestimmt und dient zur Quantifizierung der Unsicherheit unserer Ergebnisse. Als Statistischen Fehler (MoE) für eine Stichprobe der Größe n nutzen wir deshalb das 95%-Kredibilitätsintervall, welches sich aus der Differenz $Q_{beta(\alpha,\beta)_{0.975}} - Q_{beta(\alpha,\beta)_{0.025}}$, dem 0.975- und 0.025-Quantil der Beta-Verteilung mit folgender Dichte ergibt:

$$f(x, \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{11}$$

Hierbei entsprechen $\alpha = \beta = 1 + \frac{n}{2}$ (die 1 ergibt sich aus der Laplace-Bayes-Prior-Verteilung (Tuyl et al., 2008) des Beta-Binomial-Modells) und bewusst nicht der Anzahl der Befragten, die eine entsprechende Antwortoption gewählt bzw. nicht gewählt haben: Denn mit $\alpha = \beta = 1 + \frac{n}{2}$ erreicht das Kredibilitätsintervall seine maximale Breite und ist damit möglichst konservativ.
Die Verwendung der Gewichte w_k führt in vielen Fällen zu einer erhöhten Variabilität der Ergebnisse. Kish (1992) schlägt zur Quantifizierung dieses Phänomens den sogenannten Design-Effekt vor. Dieser misst die durch Gewichtung bedingte Erhöhung der Variabilität des Schätzers in der Abweichung von einer perfekten, ungewichteten Zufallsstichprobe. Wir berechnen ihn approximativ als

$$deff_w = \frac{n^{-1} \sum_{k \in s} w_k^2}{(n^{-1} \sum_{k \in s} w_k)^2} \tag{12}$$

und vergrößern das Kredibilitätsintervall entsprechend, sodass der statistische Fehler für eine Stichprobe mit Gewichten w_k sich wie folgt ergibt:

$$MoE = (Q_{beta(\alpha,\beta)_{0.975}} - Q_{beta(\alpha,\beta)_{0.025}}) * \sqrt{deff_w} \tag{13}$$

5.3 Multilevel-Regression und Poststratifizierung

Multilevel-Regression und Poststratifizierung ist ein modellbasiertes Verfahren zur Generierung von verlässlichen Ergebnissen für Gebiete mit einer geringen

[8] Hierfür muss die Invertibilität von $X_s' X_s$ sichergestellt werden, vgl. Särndal et al. (1992, S. 282)

Zahl an Befragten, welches sich in den vergangenen Jahren zunehmender Beliebtheit erfreut. Es kann im weiteren Sinne ebenfalls als Ausprägung des Superpopulations-Ansatzes verstanden werden.
Wir sind hierbei interessiert an kleinräumigen Schätzungen der Anteile p_{ci} für eine Variable y (z.B. der Zustimmung zu einer bestimmten politischen Maßnahme) mit den Ausprägungen $i = 1, ..., I$ für eine Reihe von geographischen Regionen $c = 1, ..., C$ (z.B. Landkreisen) und verfügen über eine Auswahl von Hilfsinformationen z_c (Wirtschaftsleistung, Kriminalität, etc.) über diese. Ebenfalls sind individuelle Ausprägungen der Analysevariable y_k (Zustimmung einzelner Befragter zur Politik), sowie für M kategoriale Variablen im $1 \times M$-Vektor x_k (Altersgruppe, Geschlecht, etc. , sowie die Zugehörigkeit zu einem c) über eine Stichprobe s für alle $k \in s$ bekannt, die sich über die c verteilen, sodass jedem k ein Vektor z_c zugeordnet werden kann.

5.3.1 Multilevel-Regression

Wir beginnen mit der Schätzung der Wahrscheinlichkeiten eine gewisse Antwort $P(y_k = i)$ in Abhängigkeit der individuellen Kovariaten x_k zu wählen. Über $I-1$ generalisierte gemischte Regressionsmodelle[9] (Multilevel-Regression) lässt sich die Wahrscheinlichkeit der Wahl aller Kategorien i bedingt auf die Kovariaten x_k effizient bestimmen.

$$p_{ki} = P(y_k = i) = h(z_c b + x_k u) \tag{14}$$

Hierbei ist $h(*)$ die Umkehrfunktion des Logit-Links $g(*)$:

$$g(*) = log(\frac{*}{1-*}) \tag{15}$$

$$h(*) = \frac{e^{(*)}}{1+e^{(*)}} \tag{16}$$

Während die Effekte b der Hilfsvariablen z als separate Parameter modelliert werden, wird für u eine multivariate Normalverteilung mit Kovarianzmatrix G angenommen, sodass die Effekte der individuellen Kovariaten x_k als Zufallsvariablen, sogenannte Random Effects, modelliert werden können.

$$u \sim \mathcal{N}(0, G) \tag{17}$$

Dieses Modell kann mit bestehender Software geschätzt werden, um $\hat{p}_{ki}$ für alle i gegeben x_k, z_c zu erhalten.

5.3.2 Poststratifizierung

Wir können nun für jede beliebige Kombination von x-Werten in jeder Region c eine Wahrscheinlichkeit $\hat{p}_i$ bestimmen, eine Antwort i zu wählen. Sei nun aus externen Quellen wie dem (Mikro-)Zensus die Häufigkeit dieser Kombinationen

[9] Eine ausgezeichnete Einführung findet sich bei Searle et al. (2009).

aller x für alle C Regionen bekannt[10]. Dies ergibt $\prod_{m=1}^{M} l_m = J$ Zellen als Produkt der l möglichen Ausprägungen jedes x, für welche die Häufigkeiten $N_j, j = 1, ..., J$ gegeben sind.
Bestimmen wir nun für jedes j $\hat{p}_{ij}$ und multiplizieren dies mit N_j, so erhalten wir die absolute Zahl von Personen k mit der entsprechenden Kombination von x in c, für die $y_k = i$ gilt. Dieser Wert kann auf Ebene der Regionen aggregiert werden, um die interessierenden Anteile $\hat{p}_{ic}$ für eine bestimmte Region c zu erhalten:

$$\hat{p}_{ic} = \frac{\sum_{j \in c} \hat{p}_{ij} N_j}{\sum_{j \in c} N_j} \tag{18}$$

Die Anzahl der Poststratifizierungszellen $\prod_{m=1}^{M} l_m = J$ wächst in Abhängigkeit von M und l_m multiplikativ und kann bereits bei einer niedrigen zweistelligen Zahl von Prädiktorvariablen die Grenzen der verfügbaren Rechenleistung überschreiten. Zu diesem Zweck kann eine vorgeschaltete Variablenselektion verwendet werden, um M und l_m zu verkleinern und nur jene x in der weiteren Berechnung zu verwenden, die einen signifikanten Zusammenhang zu y besitzen. Hierfür wird im Rahmen der bei Civey verwendeten Methode auf ein multinomiales logistisches Regressionsmodell mit Sparse Group-LASSO-Penalty (Vincent and Hansen, 2014) zurückgegriffen, woduch sowohl nicht mit y assoziierte Ausprägungen l_m einzelner x als auch gesamte x entfernt werden können. Somit verkleinert sich M zu $M*$ mit $l*_{m*}$ Ausprägungen, sodass $\prod_{m*=1}^{M*} l*_{m*} = J*$ deutlich kleiner als J ohne signifikante Qualitätseinbußen ist.

[10] Häufig ist diese Datenbasis nicht gegeben. Stattdessen sind ausschließlich partielle gemeinsame Verteilungen der x bekannt. Diese können mit geringen Verlusten von Genauigkeit unter Annahme partieller Unabhängigkeit zu einer synthetischen Poststratifzierungsverteilung (Leemann and Wasserfallen, 2017) zusammengefügt werden.

Literatur

AAPOR (2015). Guidance on Reporting Precision for Nonprobability Samples. Technical report, American Association for Public Opinion Research.

Ansolabehere, S. and Rivers, D. (2013). Cooperative Survey Research. *Annual Review of Political Science*, 16(1):307–329.

Ansolabehere, S. and Schaffner, B. F. (2014). Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison. *Political Analysis*, 22(03):285–303.

Blohm, M. and Koch, A. (2015). Führt eine höhere Ausschöpfung zu anderen Umfrageergebnissen? In *Nonresponse Bias*, pages 85–129. Springer.

Brick, J. M. and Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, 33(3):735–752.

Callegaro, M. (2014). *Online Panel Research: A Data Quality Perspective.*

Deming, W. E. and Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Ann. Math. Statist.*, 11(4):427–444.

Dutwin, D., Blum, M., Copeland, K., Fienberg, H., Jackson, C., Jodts, E., Koly, O., Malarek, D., Holzbaur, G., Marken, S., Matuzak, J., Pierannunzi, C., Ridenhour, J., Sheppard, D., Staehli, E. M., Stalone Lynn, Thompson, J., and Vrudhula, S. (2018). Spam Flagging and Call Blocking and Its Impact on Survey Research. Technical report, American Association for Public Opinion Research.

Elliott, M. R. and Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science*, 32(2):249–264.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.*

Gnambs, T. and Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: a meta-analysis. *Behavior research methods*, 47(4):1237–1259.

Goel, S., Obeng, A., and Rothschild, D. (2015). Non-Representative Surveys: Fast, Cheap, and Mostly Accurate. Technical report.

Guin, T. D.-L., Mechling, J., and Baker, R. (2006). Great results from ambiguous sources - cleaning internet panel data. In *ESOMAR: Panel Research 2006.*

Hillygus, D. S., Jackson, N., and Young, M. (2014). Professional respondents in non-probability online panels.

Hoover, J. and Dehghani, M. (2018). The Big, The Bad, and The Ugly: Geographic estimation with flawed psychological data.

Keeter, S. (2015). From Telephone to the Web: The Challenge of Mode of Interview Effects in Public Opinion Polls. Technical report, Pew Research.

Kennedy, C. and Hartig, H. (2019). Response rates in telephone surveys have resumed their decline. Technical report, Pew Research.

Kinney, S. K. and Cooney, D. A. (2019). Nonresponse Bias in Sample Surveys. *New Directions for Institutional Research*, 2019(181):35–46.

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, 8(2):183.

Leemann, L. and Wasserfallen, F. (2017). Extending the Use and Prediction Precision of Subnational Public Opinion Estimation. *American Journal of Political Science*, 61(4):1003–1022.

MacInnis, B., Krosnick, J. A., Ho, A. S., and Cho, M.-J. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, 82(4):707–744.

Marken, S. (2018). Still Listening: The State of Telephone Surveys. Technical report, Gallup.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4):558.

Park, D. K., Gelman, A., and Bafumi, J. (2004). Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. *Political Analysis*, 12:375–385.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*.

Potter, F. J. (1990). A Study of Procedures to identify and trim extreme sampling weights.

Quatember, A. (2019). Inferences based on Probability Sampling or Nonprobability Sampling–Are They Nothing but a Question of Models? *Survey Methods: Insights from the Field (SMIF)*.

Rao, J. N. and Molina, I. (2015). *Small Area Estimation: Second Edition*.

Rivers, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings*.

Rivers, D. (2013). Comment. *Journal of Survey Statistics and Methodology*, 1(2):111–117.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3):581–592.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model assisted survey sampling.

Searle, S. R., Casella, G., and McCulloch, C. E. (2009). *Variance components*, volume 391. John Wiley & Sons.

Shirley, K. E., York, N., and Gelman, A. (2014). Hierarchical models for estimating state and demographic trends in US death penalty public opinion. Technical report.

Smith, A. (2010). Government Online. Technical report, Pew Research.

Stalans, L. J. (2012). Frames, framing effects, and survey responses. In *Handbook of survey methodology for the social sciences*, pages 75–90. Springer.

Stephan, F. F. and McCarthy, P. J. (1958). *Sampling opinions: An analysis of survey procedure.* John Wiley, Oxford, England.

Teitcher, J. E. F., Bockting, W. O., Bauermeister, J. A., Hoefer, C. J., Miner, M. H., and Klitzman, R. L. (2015). Detecting, preventing, and responding to fraudsters in internet research: ethics and tradeoffs. *The Journal of law, medicine & ethics : a journal of the American Society of Law, Medicine & Ethics*, 43(1):116–133.

Tuyl, F., Gerlach, R., and Mengersen, K. (2008). A Comparison of Bayes–Laplace, Jeffreys, and Other Priors. *The American Statistician*, 62(1):40–44.

Valliant, R. (2019). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*.

Valliant, R., Dever, J. A., and Kreuter, F. (2018). Nonprobability Sampling. pages 565–603. Springer, Cham.

Vincent, M. and Hansen, N. R. (2014). Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis*, 71:771–786.

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

YouGov (2015). The Methodology of the 2016 YouGov/CBS News Battleground Tracker. Technical report, YouGov.