

# Methoden zur Rekonstruktion von Wählerströmen aus Aggregatdaten

17

Politik  
Politique  
Politica



OFS BFS UST

Office fédéral de la statistique  
Bundesamt für Statistik  
Ufficio federale di statistica  
Uffizi federal da statistica  
Swiss Federal Statistical Office

Neuchâtel, 2003

Die vom Bundesamt für Statistik (BFS) herausgegebene Reihe «Statistik der Schweiz» gliedert sich in folgende Fachbereiche:

---

- |    |   |    |  |
|----|---|----|--|
| 0  | Statistische Grundlagen und Übersichten | 11 | Verkehr und Nachrichtenwesen                       |
| 1  | Bevölkerung                             | 12 | Geld, Banken, Versicherungen                       |
| 2  | Raum und Umwelt                         | 13 | Soziale Sicherheit                                 |
| 3  | Arbeit und Erwerb                       | 14 | Gesundheit   |
| 4  | Volkswirtschaft                         | 15 | Bildung und Wissenschaft                           |
| 5  | Preise                                  | 16 | Kultur, Medien, Zeitverwendung                     |
| 6  | Industrie und Dienstleistungen          | 17 | Politik  |
| 7  | Land- und Forstwirtschaft               | 18 | Öffentliche Verwaltung und Finanzen                |
| 8  | Energie                                 | 19 | Rechtspflege                                       |
| 9  | Bau- und Wohnungswesen                  | 20 | Einkommen und Lebensqualität der Bevölkerung       |
| 10 | Tourismus                               | 21 | Nachhaltige Entwicklung und regionale Disparitäten |
-

# Methoden zur Rekonstruktion von Wählerströmen aus Aggregatdaten

**Dr. Mathias Ambühl**

Consult AG Bern

Mathematische Statistik

---

**Herausgeber:** Bundesamt für Statistik (BFS)  
**Auskunft:** Werner Seitz, Monique Graf, BFS, Tel. 032 713 65 85  
E-mail: [werner.seitz@bfs.admin.ch](mailto:werner.seitz@bfs.admin.ch), [monique.graf@bfs.admin.ch](mailto:monique.graf@bfs.admin.ch)  
**Autor:** Dr. Mathias Ambühl  
**Vertrieb:** Bundesamt für Statistik, CH-2010 Neuchâtel  
Tel. 032 713 60 60 / Fax 032 713 60 61 / E-mail: [order@bfs.admin.ch](mailto:order@bfs.admin.ch)  
**Bestellnummer:** 598-0300  
**Preis:** Fr. 12.–  
**Reihe:** Statistik der Schweiz  
**Fachbereich:** 17 Politik  
**Originaltext:** Deutsch  
**Layout:** Claude Maier  
**Copyright:** BFS, Neuchâtel 2003  
Abdruck – ausser für kommerzielle Nutzung – unter Angabe der Quelle gestattet  
**ISBN:** 3-303-17028-2

---

# INHALT

<b>Vorwort .....</b>	<b>5</b>
<b>Problemeinführung .....</b>	<b>7</b>
<b>1 Ausgangslage und Zusammenfassung .....</b>	<b>15</b>
1.1 Ausgangslage .....	15
1.2 Zusammenfassung .....	16
<b>2 Rekonstruktionsmethoden für (2x2)-Tabellen .....</b>	<b>23</b>
2.1 Deterministische Betrachtungen in einer (2x2)-Tabelle .....	23
2.2 Mehrere Wahlkreise .....	26
2.3 Diskussion .....	48
<b>3 Verallgemeinerungen für mehr als zwei Wahlalternativen .....</b>	<b>50</b>
3.1 Ein Zahlenbeispiel: Nationalratswahlen 1995 und 1999 Kanton Zürich .....	50
3.2 Deterministische Betrachtungen in einer (p x q)-Tabelle .....	51
3.3 Verallgemeinerung des ökologischen Regressionsansatzes .....	52
3.4 Allgemeiner Ansatz von Thomsen .....	55
3.5 Weitere Modellvorschläge .....	59
3.6 Modifizierte Probit- und Logit-Modelle .....	59
3.7 Diskussion .....	71
<b>4 Weitere Probleme in praktischen Anwendungen .....</b>	<b>73</b>
4.1 Nichtwähler .....	73
4.2 Zeitliche Veränderung der Population der Wahlberechtigten .....	73
4.3 Separate Analyse in homogenen Teilgebieten .....	75
4.4 Kleinere Parteien .....	77
4.5 Betrachtung von Wanderungssalden anstelle entgegengesetzter Wanderungen .....	77
<b>5 Wie kann die Panaschierstatistik genutzt werden? .....</b>	<b>79</b>
5.1 Konstruktion von Kovariablen aus den Panaschierdaten .....	79
5.2 Berücksichtigung von Kovariablen im Regressionsmodell mit mehreren Parteien .....	80
5.3 Ein Modellvorschlag .....	80
<b>6 Publierte Wählerstromanalysen im deutschsprachigen Raum .....</b>	<b>82</b>
6.1 Anwendung des Regressionsmodells .....	82
6.2 Anwendungen von Thomsens Methode .....	82
6.3 Kohlsche: Eigene Methode mit Elementen aus Thomsen und Regression .....	82
<b>7 Empfehlungen .....</b>	<b>87</b>
<b>8 Literaturverzeichnis .....</b>	<b>88</b>



# Vorwort

Seit gut zwei Jahrzehnten sind auch in der Politik Spuren des gesellschaftlichen Individualisierungsprozesses feststellbar: Die politische Beteiligung der Bürgerinnen und Bürger ist gesunken, und die Bindungen an die Parteien nehmen ab. Wenn eine Partei bei Wahlen erfolgreich sein will, so muss sie heute nicht nur ihre Stammwählerschaft zum Urnengang bewegen können, sie muss auch für Wählende anderer Parteien attraktiv sein und einige von diesen für sich gewinnen können («Wechselwähler»), und schliesslich muss sie möglichst viele von jenen, die sich politisch nicht mehr beteiligen, zum Urnengang und zur Stimmabgabe mobilisieren können.

Bei der Analyse der Wahlergebnisse wird so neben der Frage nach den Per-Saldo-Gewinnen und -Verlusten der einzelnen Parteien auch die Frage nach der so genannten Wählerwanderung immer wichtiger: Von welchen Parteien hat die siegreiche Partei A Wählende übernehmen können, welche Partei mobilisierte am besten bei den Nicht-Wählenden und an wen haben die Verliererparteien Wählende verloren? Solche Fragen können am besten mit der Befragung von repräsentativ ausgewählten Stimmbürgerinnen und Stimmbürgern beantwortet werden. Wissenschaftliche Meinungsumfragen sind jedoch relativ aufwändig und teuer und bleiben oft auf grössere räumliche Einheiten beschränkt; zudem sind sie für zeitlich zurückliegende Wahlen nicht mehr durchführbar.

Seit einiger Zeit erhält nun die ökologische Aggregatdatenanalyse wieder etwas Aufwind, welche versucht, die Wahlergebnisse von geographischen Einheiten (z.B. Gemeinden, Wahlkreisen) durch charakteristische soziale, kulturelle oder ökonomische Eigenschaften dieser Einheiten zu beschreiben und zu erklären. Neuerdings beleben auch Modelle die politologische Diskussion, welche mittels ausgeklügelter statistischer Verfahren versuchen, aus Aggregatdaten Informationen über Wählerströme zu gewinnen. Mit solchen Modellen wurde das Bundesamt für Statistik in jüngster Zeit verschiedentlich konfrontiert.

Das BFS hat deshalb den Statistiker Dr. Mathias Ambühl von Consult AG Bern beauftragt, (1) einen Literaturüberblick zu den bekannten Methoden betreffend die Rekonstruktion von Wählerströmen zu erstellen, (2) Kriterien zu erarbeiten, welche es erlauben, die Solidität der Modelle einzuschätzen, (3) abzuklären, inwieweit die Panaschierdaten des BFS allenfalls für ein solches Modell verwendet werden könnten und schliesslich (4), falls dies möglich ist, ein Modell mit den Wahldaten des BFS zu entwickeln.

Die vorliegende Studie, welche in Zusammenarbeit mit den beiden BFS Mitarbeitenden Dr. Monique Graf von der Sektion «Statistische Methoden» und Dr. Werner Seitz, Leiter des Bereichs «Wahlen und Abstimmungen», entstanden ist, kommt zu einem skeptischen Schluss: Die bekannten und hier diskutierten Modelle beruhen alle auf (begründeten) Annahmen und Schätzungen – und stehen und fallen mit diesen letztlich auch. Die getroffenen Annahmen sind in jedem Fall mit Aggregatdaten nur begrenzt überprüfbar; deshalb können die Resultate einer ökologischen Inferenz nicht als gleichwertige Information wie die erhobenen Wahldaten betrachtet werden. Das heisst zwar nicht, dass sämtliche Modelle zur Rekonstruktion der Wählerströme aus Aggregatdaten verworfen werden müssen. Es wird aber unmissverständlich gefordert, dass wer solche Modelle anwendet, die Methode und die getroffenen Annahmen offen legt und die Grenzen des Mo-

dells aufzeigt. Wer dieser wissenschaftlichen Selbstverständlichkeit nicht nachkommt, setzt sich dagegen dem Verdacht der Unseriosität aus.

Diese Studie soll einen Beitrag auf der Ebene der international geführten wissenschaftlichen Diskussion über die ökologische Inferenz leisten und dazu beitragen, dass auch in der Hektik der Wahlberichterstattungen der kritische Blick der Medien auf die Modelle, welche den verschiedenen Analysen zu Grunde liegen, geschärft werde bzw. erhalten bleibe. In diesem Sinne möchte ich dem Autor, Dr. Mathias Ambühl, für seine interessante und seriöse Arbeit danken. Ein herzliches Dankeschön richtet sich auch an Dr. Peter Selb von der Universität Zürich, der als methodisch versierter Politikwissenschaftler und Leiter der Schweizer Wahlstudie (Swiss Electoral Studies, Selects) eine verständliche Einführung in dieses sehr komplexe Gebiet der Rekonstruktion von Wählerströmen geschrieben hat.

Bundesamt für Statistik  
Dr. Adelheid Bürgi-Schmelz  
Direktorin



# Problemeinführung

## Die Analyse von Wählerströmen

Insbesondere dann, wenn aus demokratischen Wahlen grössere Verschiebungen im politischen Kräfteverhältnis zwischen den Parteien hervorgehen, beherrscht die Frage nach Ausmass und Richtung von Wählerwanderungen sowie deren Ursachen nicht nur die strategischen Gremien der betroffenen Parteien, sondern auch die politische Berichterstattung der Medien. Solche Verschiebungen sind in der Schweiz seit den 90er Jahren bei Wahlen auf allen föderativen Ebenen verstärkt zu beobachten – häufig zugunsten der SVP. Mit steigendem öffentlichen Interesse gerät die Meinungs- und Wahlforschung unter zunehmenden Druck, denn die Befragungsdaten, die üblicherweise im Umfeld von Wahlen erhoben werden, sind ausgerechnet zur Beantwortung der Frage nach den Wanderungsbewegungen der Wählerschaft nur sehr begrenzt geeignet.

Denn *erstens* handelt es sich bei diesen Daten in der Regel um *Querschnittsdaten*, die zu einem Zeitpunkt – etwa kurz nach den betreffenden Wahlen – gesammelt werden. Zwar geben Querschnittsdaten Aufschluss über den aktuellen Parteientscheid von Wählenden; Informationen zu deren Verhalten bei den vorangegangenen Wahlen können dabei aber allenfalls retrospektiv ermittelt werden. Die so gewonnenen Daten bergen ein hohes Mass an Unsicherheit, denn Wahlen finden üblicherweise nur alle vier Jahre statt und entsprechend getrübt ist häufig das Erinnerungsvermögen der Befragten. In der Tat haben viele Befragungsteilnehmer die Neigung, ihr berichtetes Verhalten bei den vorangegangenen Wahlen mit ihrem aktuellen Parteientscheid in Einklang zu bringen. Das Ausmass tatsächlicher Wanderungsbewegungen wird daher mittels solcher Querschnittsdaten aller Wahrscheinlichkeit nach unterschätzt. Dieses Problem lässt sich nur in längsschnittlichen Panelstudien beheben, in denen dieselben Respondenten im Abstand von vier Jahren jeweils zu den aktuellen Wahlen befragt werden. Eine solche Untersuchung wird innerhalb der Schweizer Wahlstudie *Selects* 2003 erstmals durchgeführt, bisher existieren derartige Daten jedoch nicht.

*Zweitens* stossen repräsentative Befragungen mit einem üblichen Stichprobenumfang von 1'000 bis 2'000 Respondenten gerade in relativ stabilen Mehrparteiensystemen wie der Schweiz hinsichtlich der Frage der Wählerströme schnell an die Grenzen ihrer Aussagekraft. *Tabelle 1* stellt dieses Problem beispielhaft anhand von Daten dar, die im Rahmen von *Selects* anlässlich der National- und Ständeratswahlen 1999 erhoben wurden. Obwohl die Anzahl der verwertbaren Fälle hier mit 1'950 Befragten relativ hoch ist, ist die Besetzung der Zellen ausserhalb der Diagonalen der Tabelle, auf denen die Beobachtung von Wählerwanderungen massgeblich basiert, relativ schwach: Unter den 1'704 Respondenten, die 1995 bereits wahlberechtigt waren (und ihre Wahlteilnahme bzw. ihren Parteientscheid erinnern können), berichten allein zwei Drittel ein stabiles Wahlverhalten, d.h. sie befinden sich auf einem der sechs Felder der Diagonalen. Nur ein Drittel der Befragten verteilt sich auf die übrigen 30 Zellen, die hier durchschnittlich mit nur knapp 20 (manche sogar mit deutlich weniger!) Respondenten besetzt sind. Zuverlässige Rückschlüsse auf das Verhalten der Wählerschaft als der interessierenden Grundgesamtheit dieser Befragung sind hier kaum mehr möglich, da statistische Rückschlüsse wesentlich auf Bernoullis *Gesetz der grossen Zahl* beruhen.

Tabelle 1: Wählerströme zwischen den Nationalratswahlen 1995 und 1999; Wanderungstabelle aus querschnittlichen Befragungsdaten 1999 (absolute Zahlen).

	1999: FDP	CVP	SP	SVP	andere Parteien	hat nicht gewählt	total
<b>1995:</b>							
FDP	<b>162</b>	5	4	26	6	33	236
CVP	6	<b>85</b>	7	10		21	129
SP	8	7	<b>231</b>	10	29	27	312
SVP	4	2	2	<b>112</b>	6	16	142
andere Parteien	24	18	33	41	<b>91</b>	136	343
hat nicht gewählt	20	13	27	28	14	<b>440</b>	542
Neuwähler	1	3	9	10	6	39	68
kann sich nicht erinnern	23	16	23	37	19	60	178
total	248	149	336	274	171	772	1'950

Quelle: Schweizer Wahlstudie *Selects 1999*, Stichprobenumfang N=2'048, 98 Fälle mit fehlenden Angaben

Alternativ zu Befragungs- oder *Individualdaten* werden in den vergangenen Jahren immer häufiger auch *Aggregatdaten* genutzt, um Wählerströme zu analysieren. Anlässlich der National- und Ständeratswahlen 1999 etwa publizierten einige Schweizer Zeitungen erstmals die Wählerstromanalysen von Andreas Kohlsche (*Institut für Wahl-, Sozial- und Methodenforschung*), die den Schluss nahelegten, die SVP habe vor allem von der Mobilisierung ehemaliger Nichtwähler und von den Wählerschaften kleinerer rechter Parteien profitiert. Auf diese Untersuchung folgten weitere Analysen von kantonalen Wahlen – zuletzt im April 2003 für Zürich, Luzern und das Tessin. Im März 2002 erregte auch das *GfS-Forschungsinstitut* mit einer Aggregatstudie zur UNO-Abstimmung im Kanton Aargau grosses Aufsehen, welche zu dem Ergebnis kam, dass die CVP-Wählerschaft mehrheitlich nicht der Abstimmungspareole ihrer Partei folgte (auch eine Art von Wählerwanderung!) und damit die Ablehnung der Initiative durch den Kanton bewirkte.

Aggregatdaten beziehen sich nicht wie Befragungsdaten auf individuelle Wahlberechtigte, sondern auf räumliche Einheiten wie Wahlbezirke. Im Gegensatz zu Befragungsdaten, die eigens zu Forschungszwecken erhoben werden müssen, sind solche Aggregatdaten *prozessgeneriert*, d.h. sie fallen während der Auszählung von Wahl- und Abstimmungsergebnissen ohnehin an und sind als amtliche Statistiken in der Regel frei verfügbar. Hier wird der erste Vorzug der Aggregatanalyse von Wählerströmen offenkundig: Sie sind wesentlich kostengünstiger (und schneller verfügbar) als befragungsbasierte Analysen. Daneben fussen Aggregatanalysen auf vollständigen Datenquellen, während Befragungen anhand von Stichproben durchgeführt werden. Wird eine solche Stichprobe – wie üblich – auf nationaler Ebene gezogen, dann erlaubt sie vor allem Rückschlüsse auf eben diese nationale Ebene. Folgerungen für Subgruppen der Stichprobe, wie z.B. die Stimmberechtigten eines einzelnen Kantons, werden dabei mit abnehmender Gruppengrösse immer unsicherer. So würden beispielsweise bei einer nationalen Zufallsstichprobe mit einem Umfang von 2'000 Respondenten auf den Kanton Schaffhausen, in welchem um ein Prozent der Schweizer Wahlberechtigten ihr Stimmrecht haben, nur etwa 20 Personen entfallen. Selbst simple Analysen wie diejenige, die in *Tabelle 1* dargestellt ist, sind hier schlicht nicht mehr möglich. In Aggregatanalysen von Wählerströmen stellt die Fokussierung auf Kantone hingegen kein Problem dar, da innerhalb der Kantone genügend Untersuchungseinheiten zur Verfügung stehen – seien es Wahlbezirke oder sogar einzelne Stimmlokale, die in diesem Kontext die kleinstmöglichen Beobachtungseinheiten darstellen. Die Möglichkeit der Regionalisierung des Untersuchungsschwerpunkts erhält gerade vor dem Hintergrund des *Ständemehrs* bei nationalen Abstimmungen in der Schweiz besondere Relevanz. Neben diesen Vorteilen bringen Aggregatanalysen jedoch

auch gravierende Nachteile mit sich, allem voran die Gefahr des sogenannten *ökologischen Fehlschlusses*. Diese Problematik soll in den folgenden Abschnitten kurz umrissen werden.

## Der ökologische Fehlschluss

Zwischen den Wahlen wandern Wahlberechtigte, und nicht Wahlbezirke oder Stimmlokale. Daher stellt jede Aussage über Wählerbewegungen, die auf der Beobachtung von Aggregat-einheiten wie Wahlbezirken beruht, einen *ökologischen Rückschluss* (*ökologische Inferenz*) dar, d.h. einen Rückschluss über verschiedene Untersuchungsebenen hinweg. Zusammenhänge, die auf der Ebene von Wahlbezirken beobachtet werden können, müssen nicht zwingendermassen den Zusammenhängen entsprechen, die innerhalb dieser Wahlbezirke auf der Ebene individueller Wahlberechtigter zu finden sind. Wenn nun Aussagen über Wählende auf Basis von einfachen Aggregatbeziehungen in Situationen getroffen werden, in denen sich diese Zusammenhänge unterscheiden, dann liegt ein *ökologischer Fehlschluss* vor. Dieses Phänomen lässt sich leicht anhand eines fiktiven Zweiparteiensystems bestehend aus nur zwei gleichgrossen Wahlbezirken illustrieren, das in *Tabellen 2a* und *b* dargestellt ist. Die abgebildeten Randverteilungen für die beiden Wahljahre seien aus offiziellen Wahlstatistiken bekannt. In Aggregatanalysen von Wählerströmen geht es nun im Wesentlichen darum, von diesen Randverteilungen auf die Besetzung der inneren Zellen der Wanderungstabelle für das gesamte Wahlsystem respektive die *Übergangswahrscheinlichkeiten* zwischen diesen Zellen zu schliessen. In dem dargestellten Beispiel sind die Wahlen 1999, was die offiziellen Ergebnisse betrifft, offenbar ein unspektakulärer Neuaufguss der Wahlen 1995: Die beiden Parteien A und B haben in beiden Wahlen und in beiden Wahlbezirken jeweils identische Stimmentotale und –anteile erhalten.

*Tabellen 2a und b: Fiktive Randverteilungen der abgegebenen Stimmen für zwei Parteien über zwei Wahlen 1995 und 1999 in zwei Wahlbezirken (absolute Zahlen).*

### a. Wahlbezirk 1

	1999:		total
	Partei A	Partei B	
1995:			
Partei A	?	?	5'000
Partei B	?	?	5'000
total	5'000	5'000	10'000

### b. Wahlbezirk 2

	1999:		total
	Partei A	Partei B	
1995:			
Partei A	?	?	500
Partei B	?	?	9'500
total	500	9'500	10'000

*Tabellen 3 a und b: Zwei mögliche Zellenbesetzungen für die Randverteilung im fiktiven Wahlsystem (absolute Zahlen).*

### a. Option 1

	1999:		total
	Partei A	Partei B	
1995:			
Partei A	5'500	0	5'500
Partei B	0	14'500	14'500
total	5'500	14'500	20'000

### b. Option 2

	1999:		total
	Partei A	Partei B	
1995:			
Partei A	0	5'500	5'500
Partei B	5'500	9'000	14'500
total	5'500	14'500	20'000

Wie gefährlich es nun sein kann, aus dieser Beobachtung zu schliessen, es hätten von 1995 bis 1999 keine Wählerwanderungen zwischen den beiden Parteien stattgefunden, zeigen *Tabellen 3a* und *b*, welche zwei (sehr drastische) der zahlreichen möglichen Besetzungen der inneren Zellen bei gegebener Randverteilung für das gesamte Wahlsystem wiedergeben. Der erste Fall reflektiert die Situation, die die Randverteilungen in den beiden Wahlbezirken auf den ersten Blick sugge-

rieren, nämlich keine Wanderungen. Die Übergangswahrscheinlichkeit von Partei A 1995 nach Partei A 1999 ( $p_{AA}$ ) ist hier ebenso wie  $p_{BB}$  gleich 1;  $p_{AB}$  und  $p_{BA}$  sind 0 (siehe Ambühl, Kap. 2.1). Im zweiten Fall sind hingegen beträchtliche Wanderungsbewegungen zu beobachten. Tatsächlich verliert hier Partei A 1999 die gesamte Wählerschaft von 1995 an Partei B, d.h.  $p_{AA} = 0$  und  $p_{AB} = 1$ . Partei B verliert aber gleichzeitig 5'500 von ihren 14'500 ehemaligen Wählenden an Partei A, also  $p_{BA} = 5'500/14'500 = .38$  und  $p_{BB} = 1 - 5'500/14'500 = .62$ . In diesem zweiten Fall läge ein verheerender ökologischer Fehlschluss vor, würde man von der beobachteten Aggregatstabilität auf die interessierende Individualbeziehung schliessen.

Diesem Problem ist man sich in den Sozialwissenschaften spätestens seit der Publikation von William S. Robinsons *Ecological Correlation and the Behavior of Individuals* 1950 durchaus bewusst. Tatsächlich sah man aber in der Folgezeit die Lösung für dieses Problem in erster Linie in der Erschliessung neuer Datenquellen, nämlich in stichprobenbasierten Befragungen, die seit-her mehr oder weniger zum „Königsweg der Sozialforschung“ (Erwin K. Scheuch) avancierten. Leider sind Befragungsdaten aber teuer und – wie weiter oben dargelegt – nicht für sämtliche Fragestellungen in geeigneter Qualität verfügbar. Nicht zuletzt aus diesem Grund erlebt die Aggregatanalyse von an sich individuellen Phänomenen seit den 90er Jahre eine Renaissance. Vor allem innerhalb der Politikwissenschaft wurden in den vergangenen Jahren einige innovative Methoden zur Rekonstruktion individueller Verhaltensweisen aus Aggregatdaten hervorgebracht, welche Mathias Ambühl in der vorliegenden Expertise in einem Überblick vorstellt, kritisch evaluiert und stellenweise weiterentwickelt. Der abschliessende Teil dieser Problemeinführung soll es nun der Leserschaft ausserhalb der statistischen Fachgemeinschaft erleichtern, diesen hervorragenden – aber notwendigerweise sehr stark formalisierten – Bericht hinsichtlich seiner praktischen Relevanz für solide Schlüsse von Aggregatdaten auf Wählerströme einzuordnen.

## Methoden zur Rekonstruktion von Wählerströmen aus Aggregatdaten

Durch Methoden der ökologischen Inferenz wird versucht, von der auf Ebene der Wahlbezirke beobachteten Beziehung zwischen den Wahlergebnissen bei zwei typischerweise aufeinander folgenden Wahlen auf die nicht-beobachtete Beziehung zwischen dem individuellen Wahlverhalten bei den beiden Wahlen zu schliessen. Solche Rückschlüsse stützen sich notwendigerweise auf *Annahmen* zur Beschaffenheit des Zusammenhangs zwischen den Beziehungen auf Aggregat- und Individualebene. In der obigen Illustration des fiktiven Zweiparteiensystems wurde – wie in regressionsbasierten Verfahren der ökologischen Inferenz üblich – angenommen, dass die individuellen Übergangswahrscheinlichkeiten in sämtlichen Wahlbezirken jeweils identisch sind. Diese Annahme wird unter *Option 2 (Tabelle 3b)* verletzt. *Tabellen 4a* und *b* geben die Besetzungen der inneren Zellen der Wanderungstabellen der beiden Wahlbezirke A und B wieder, die sich unter dieser Option zwangsläufig ergäben. Für den Wahlbezirk 1 wäre  $p_{AA} = 0$ ,  $p_{AB} = 1$ ,  $p_{BA} = 1$  und  $p_{BB} = 0$ . Für den 2. Bezirk hingegen wäre  $p_{AA} = 0$ ,  $p_{AB} = 1$ ,  $p_{BA} = .05$  und  $p_{BB} = .95$ . Unter solchen Bedingungen führt die Schätzung von Übergangswahrscheinlichkeiten mittels ökologischer Regressionen häufig zu allein unter logischen Gesichtspunkten unsinnigen Ergebnissen, d.h. zu Wahrscheinlichkeiten kleiner 0 oder grösser 1. Während sophistiziertere Methoden der ökologischen Inferenz eine Lockerung dieser Annahme erlauben (siehe Ambühl, Kap. 2.2), ist die Annahme der Unabhängigkeit der Stimmenanteile für die Parteien und den Übergangswahrscheinlichkeiten in den einzelnen Wahlbezirken in sämtlichen regressions-basierten Verfahren zur ökologischen Inferenz zentral. Auch diese würde im letztgenannten Beispiel verletzt, denn hier zeichnet sich – zumindest in der Tendenz – ein systematischer Zusammenhang zwischen diesen beiden Grössen ab. Die Übergangswahrscheinlichkeiten zu der jeweils anderen Partei sind gross für diejenigen Parteien, die ihren Wahlbezirk 1995 nicht dominieren (Parteien A

und B im Bezirk 1; Partei A im Bezirk 2); während sie für die dominierenden Parteien gering sind (Partei B in Bezirk 2). Eine Verletzung dieser Annahme hat Verzerrungen (*bias*) bei der Schätzung der Übergangswahrscheinlichkeiten durch Aggregatmodelle zur Folge.

*Tabellen 4a und b: Besetzung der inneren Zellen der Wanderungstabellen in den zwei Wahlbezirken A und B, unter Option 2 (Tabelle 3b) .*

*a. Wahlbezirk 1*

	1999:		
	Partei A	Partei B	total
1995: Partei A	0	5'000	5'000
Partei B	5'000	0	5'000
total	5'000	5'000	10'000

*b. Wahlbezirk 2*

	1999:		
	Partei A	Partei B	total
1995: Partei A	0	500	500
Partei B	500	9'000	9'500
total	500	9'500	10'000

Die Zuverlässigkeit von ökologischen Inferenzen hängt also davon ab, inwieweit die Modellannahmen zutreffen. Deren Gültigkeit kann aber wiederum mittels der aggregierten Daten selbst nicht zweifelsfrei überprüft werden. Tatsächlich ist eine verlässliche Überprüfung der Annahmen nur in den Fällen möglich, in denen die Besetzungen der inneren Zellen der Wanderungstabellen bekannt sind. Zwar erübrigt sich in diesen Situationen an sich die ökologische Inferenz, zur Methodenevaluation sind solche Situationen aber äusserst wertvoll. So wäre es beispielsweise grundsätzlich möglich, die unterschiedlichen Modelle der ökologischen Inferenz, die Ambühl (*Kap. 3*) anhand von offiziellen Ergebnissen der Nationalratswahlen 1995 und 1999 auf Ebene der 171 Gemeinden des Kantons Zürich schätzt, in einem zweiten Schritt einer kantonalen Wanderungstabelle gegenüberzustellen, die aus der Befragung der Zürcher Stichprobe im Rahmen der Schweizer Wahlstudie *Selects* 1999 gewonnen wurde. Der praktische Nutzen eines solchen Vergleichs wäre aber nur sehr begrenzt, da diese querschnittlichen Befragungsdaten mit einem Umfang der kantonalen Stichprobe von nur etwa 350 verwertbaren Fällen die weiter oben geschilderten Probleme bei der Verwendung von Befragungsdaten für Wählerstromanalysen verstärkt in sich tragen würden. Abweichungen zwischen Modellprognosen und beobachteter Besetzung der inneren Zellen der Wanderungstabelle liessen sich daher nicht mehr eindeutig auf die Schwächen der Modelle zurückführen. Denn ebenso gut könnten diese auf die Unzuverlässigkeit der Befragungsdaten in dieser Situation zurück gehen. Einen Ausweg aus diesem Dilemma können nur Befragungsdaten bieten, die den hohen Datenansprüchen der Fragestellung gerecht werden. Dem Anspruch der längsschnittlichen Erhebung wird dabei in absehbarer Zeit durch die Schweizer Wahlstudie *Selects* 2003 nachgekommen. Dem Anspruch des hohen Stichprobenumfangs hingegen wird jedoch auch die Panelstudie im Rahmen von *Selects* 2003 nicht gerecht werden, da hier selbstverständlich nur diejenigen 2'048 Respondenten erneut befragt werden können, die bereits 1999 befragt wurden. Allerdings ist bei weitem nicht davon auszugehen, dass sämtliche ehemaligen Befragten nach einem Zeitraum von immerhin 4 Jahren erneut erreicht werden können und zudem auskunftswillig bzw. -fähig sind.

Wesentlich grundlegender bleiben aber berechtigte Zweifel, ob Methoden der ökologischen Inferenz prinzipiell eine befriedigende Problemlösung darstellen können, wenn sich deren Zuverlässigkeit nicht alleine anhand modellinhärenter Kriterien – wie der *Erwartungstreue* und *Konsistenz* des Schätzverfahrens – prüfen lässt, da die Gültigkeit der Modellannahmen immer an den jeweiligen Einzelfall gebunden bleibt. Daher ist die Schlussfolgerung Ambühls, dass „auch in Zukunft nicht mit einer vollauf befriedigenden Lösung des Problems zu rechnen ist“ durchaus nachvollziehbar.

Dieses pessimistische Fazit soll jedoch keinesfalls über den wertvollen Beitrag hinweg täuschen, den Ambühl für die Bewertung und vor allem die innovative Weiterentwicklung von Modellen zur Rekonstruktion von Wählerströmen aus Aggregatdaten leistet. Dieser Beitrag ist insbesondere in der Berücksichtigung bisher häufig vernachlässigter Probleme bei der Analyse von Wählerströmen, wie etwa der Veränderung der Wählerschaft innerhalb der Wahlbezirke zwischen den Wahlen (siehe *Kap. 4*), und in der Integration der Panaschierstimmen als zusätzliche Informationen in Wählerstrommodelle zu sehen (siehe *Kap. 5*).

Dr. Peter Selb  
Projektverantwortlicher  
Schweizer Wahlstudie *Selects* 2003

# **Methoden zur Rekonstruktion von Wählerströmen aus Aggregatdaten**





# 1 Ausgangslage und Zusammenfassung

## 1.1 Ausgangslage

Seit einigen Jahren werden in den Sozialwissenschaften verschiedene Verfahren angeboten – u.a. vom Amerikaner Gary King –, welche es erlauben sollen, von Aggregatdaten zu Schlüssen auf der individuellen Ebene zu gelangen. Im deutschen Sprachraum publiziert u.a. der Politologe Dr. Andreas Kohlsche regelmässig Wählerstromanalysen, welche auf den aggregierten Daten vergangener Wahlen basieren; für die Analysen in der Schweiz verwendet Kohlsche hauptsächlich die Wahlergebnisse des BFS (Wahlzettel und Panaschierdaten).

### 1.1.1 Auftrag

Der vorliegende Bericht beruht auf einem Auftrag des Bundesamtes für Statistik (BFS) an die Consult AG Bern, welcher die folgenden Punkte umfasst:

- Es soll ein Überblick gegeben werden über die bekannten Methoden, welche die Rekonstruktion von Wählerströmen erlauben sollen (Literaturüberblick).
- Es sollen Kriterien erarbeitet werden, welche es erlauben, die Solidität der Modelle einzuschätzen.
- Es soll abgeklärt werden, wie die Panaschierdaten in einem solchen Modell berücksichtigt werden können, um ihre Genauigkeit zu verbessern.
- Erarbeitung und Diskussion eines Vorschlags für ein Modell mit den Wahldaten BFS.

### 1.1.2 Kriterien zur Beurteilung der Rekonstruktionsmethoden

Die Beurteilung der in der Literatur vorgeschlagenen Methoden zur Rekonstruktion von Wählerströmen erfolgt in diesem Bericht in erster Linie durch die Analyse ihrer mathematisch-statistischen Eigenschaften. Vereinzelt werden Berichte von Erfahrungen mit den vorgestellten Methoden aus der Literatur zitiert.

Eher exemplarischen Charakter haben die präsentierten Ergebnisse aus der Anwendung verschiedener Methoden, mit welchen versucht wird, die Wählerwanderungen im Kanton Zürich zwischen den Nationalratswahlen 1995 und 1999 zu rekonstruieren. Der Aufwand eines breit abgestützten empirischen Vergleichs der verschiedenen Methoden mit zahlreichen Datensätzen würde den Rahmen dieser Arbeit sprengen.

Die verschiedenen Methoden werden bezüglich folgender Kriterien untersucht:

1. **Modell:** Die Methodik soll eine sinnvolle Basis in Form eines mathematisch-statistischen Modells besitzen. Insbesondere sollte klar werden, auf welchen formellen Annahmen die Resultate basieren.
2. **Überprüfbarkeit:** Die Gültigkeit der getroffenen Annahmen in einer Anwendung sollte überprüfbar sein. Es sollte also aus den vorliegenden Daten beurteilt werden können, ob die Voraussetzungen zur Verwendung der Methode gegeben sind.

**3. Schätzung:** Beurteilt wird zudem die bei der Ermittlung der vorgeschlagenen Lösung verwendete Methodik. Die Lösung sollte transparent sein und auf einem objektiven mathematisch-statistischen Kriterium basieren. Ausserdem interessieren die statistischen Eigenschaften der Schätzungen, insbesondere Erwartungstreue und Konsistenz:

- Von einer *erwartungstreuen* Schätzung spricht man in der Statistik dann, wenn der Erwartungswert einer Schätzung mit der zu schätzenden Zahl identisch ist. Erwartungstreue sagt nichts aus über die Genauigkeit eines Schätzers, schliesst aber aus, dass der zu schätzende (wahre) Wert systematisch über- oder unterschätzt wird.
- Eine Schätzung heisst (*statistisch*) *konsistent*<sup>1</sup>, falls ihre Varianz (d.h. Ungenauigkeit) bei wachsendem Stichprobenumfang gegen 0 und ihr Erwartungswert gegen den wahren Wert strebt. Unter „Stichprobenumfang“ ist im vorliegenden Fall von Wahlanalysen die Anzahl der Wahlkreise zu verstehen.

## 1.2 Zusammenfassung

### 1.2.1 Allgemeines

Grundlage aller betrachteten Rekonstruktionsmethoden bilden die Wähleranteile in zwei Wahlen, welche separat für eine grössere Anzahl von Teilgebieten (Wahlkreise) des Untersuchungsgebiets vorliegen. Das Ziel besteht darin, die vollständige Wanderungstabelle zu schätzen, d.h. den Anteil der Wahlberechtigten, die in der ersten Wahl eine Partei  $p$  und in der zweiten eine Partei  $q$  wählen.

Es liegt also für jeden Wahlkreis eine Tabelle der folgenden Form vor, wobei  $P$  die Anzahl der betrachteten Wahlalternativen in der ersten Wahl und  $Q$  diejenige in der zweiten Wahl ist<sup>2</sup>:

Wahl 2	Wahl 1 Partei 1	Partei 2	...	Partei $P$	Total
Partei 1	?	?	...	?	$n_{21}$
Partei 2	?	?	...	?	$n_{22}$
...	...	...	...	...	...
Partei $Q$	?	?	...	?	$n_{2Q}$
Total	$n_{11}$	$n_{12}$	...	$n_{1P}$	$n$

Die Randtotale  $n_{11}, \dots, n_{1P}$  aus Wahl 1 sowie  $n_{21}, \dots, n_{2Q}$  aus Wahl 2 sind bekannt, die Häufigkeiten in den Innenfeldern sollen geschätzt werden. Das primäre Interesse liegt dabei nicht bei den Innenfeldern jedes einzelnen Wahlkreises, sondern bei den Summen über alle Wahlkreise hinweg, d.h. die interessierenden Grössen sind die Wählerströme im gesamten Untersuchungsgebiet.

<sup>1</sup> Um Missverständnisse zu vermeiden, wird in diesem Bericht jeweils der Begriff „statistisch konsistent“ verwendet, wenn von der hier beschriebenen Eigenschaft die Rede ist. Der Begriff der Konsistenz wird gelegentlich als Eigenschaft einer Wanderungstabelle verwendet, welche dadurch definiert ist, dass die Spalten- und Zeilensummen der rekonstruierten Zelleneinträge den beobachteten Randhäufigkeiten entsprechen.

<sup>2</sup> In diesem Bericht werden Wanderungstabellen jeweils so dargestellt, dass die Wahlalternativen in Wahl 1 den Spalten und diejenigen in Wahl 2 den Zeilen entsprechen. In der Politologie ist die Darstellungsweise mit umgekehrten Rollen von Zeilen und Spalten üblich. Die hier gewählte Ausrichtung entspricht derjenigen von Übergangsmatrizen von Markov-Ketten in der Wahrscheinlichkeitstheorie.

Betrachtet man anstelle absoluter Stimmenzahlen relative Anteile bezüglich der Population der Wahlberechtigten und bezeichnet den Stimmenanteil einer Partei  $p$  in der ersten Wahl mit  $x_p$  sowie denjenigen einer Partei  $q$  in der zweiten mit  $y_q$ , so ergibt sich für jeden Wahlkreis eine Darstellung wie in der untenstehenden Tabelle. Dabei handelt es sich jeweils bei einer der betrachteten „Parteien“ um die Nichtwähler. Offensichtlich muss dann gelten, dass

$$\sum_{p=1}^P x_p = \sum_{q=1}^Q y_q = 1.$$

Wahl 2	Wahl 1 Partei 1	Partei 2	...	Partei $P$	Total
Partei 1	?	?	...	?	$y_1$
Partei 2	?	?	...	?	$y_2$
...	...	...	...	...	...
Partei $Q$	?	?	...	?	$y_Q$
Total	$x_1$	$x_2$	...	$x_P$	1

Die Rekonstruktion von Wählerströmen aus Wähleranteilen stellt einen Spezialfall der sogenannten „**Ökologischen Inferenz**“ (*Ecological Inference*) dar, bei der versucht wird, aus den Randhäufigkeiten von zwei- oder höherdimensionalen Häufigkeitstabellen Erkenntnisse über die Belegung der einzelnen Zellen zu gewinnen. Es ist zu beachten, dass nicht jede statistische Analyse von Wähleranteilen zu dieser Methodenklasse zu zählen ist, wie beispielsweise Katz und King (1999) betonen.

### 1.2.2 Die zwei Haupttypen von Modellen

Die betrachteten Methoden lassen sich in zwei Kategorien einteilen, welche auf verschiedenen methodischen Zugängen basieren.

- Mit dem Modell der „ökologischen Regression“ und seinen zahlreichen Modifikationen wird versucht, die Anteile in der zweiten Wahl als lineare Funktion der Anteile in der ersten Wahl zu erklären:

$$y_q \approx p_{(1,q)}x_1 + p_{(2,q)}x_2 + \dots + p_{(P,q)}x_P.$$

Die Koeffizienten  $p_{(p,q)}$  können dabei als Übergangswahrscheinlichkeiten interpretiert werden, d.h. als Wahrscheinlichkeit, dass eine Person, die ihre Stimme in der ersten Wahl der Partei  $p$  gab, sich in der zweiten Wahl für Partei  $q$  entscheidet.

- Der zweite Zugang geht von der Annahme aus, dass die politische Einstellung der Wahlberechtigten sich in Form einer (unbekannten) latenten Variablen  $z$  beschreiben lässt. Diese wird üblicherweise als mehrdimensionale numerische Grösse angenommen. Die Entscheidungen der Wahlberechtigten in den beiden Wahlen werden dann als eine Funktion dieser Einstellungsvariable modelliert.

Wichtig ist die Feststellung, dass das Problem der ökologischen Inferenz nur durch die Festlegung gewisser Annahmen gelöst werden kann. Bei allen Methoden wird versucht, mittels Modellannahmen aus den Abhängigkeitsstrukturen *über die Wahlkreise hinweg* auf solche *innerhalb der Wahlkreise* zu schliessen. Die Zuverlässigkeit der Resultate ist folglich davon abhängig, ob diese Annahmen dem tatsächlichen Wahlverhalten entsprechen. Aus diesem Grund ist die Möglichkeit

einer Überprüfung der den jeweiligen Berechnungen zugrundeliegenden Annahmen wünschenswert.

### 1.2.3 Das Regressionsmodell und seine Modifikationen

Der grösste Teil der in der Literatur vorgestellten statistischen Methoden für ökologische Inferenz basiert auf einem linearen Modellansatz, bei dem die Wähleranteile in der zweiten Wahl als lineare Funktion derjenigen in der ersten Wahl modelliert werden. Diese Modelle weisen den Vorteil einfacher Interpretierbarkeit auf, und ausserdem ist die Schätzung mit der weit verbreiteten statistischen Methode der linearen Regression möglich.

Das einfache Regressionsmodell besticht auf den ersten Blick durch seine Einfachheit und Interpretierbarkeit, doch es weist etliche ungünstige Eigenschaften auf. Insbesondere erweist sich die zentrale Voraussetzung gleicher Übergangswahrscheinlichkeiten in allen Wahlkreisen als problematisch. Ist diese Annahme verletzt, so führt das Regressionsmodell in vielen Fällen zu geschätzten Übergangswahrscheinlichkeiten kleiner als 0 oder grösser als 1. Dies ist typischerweise dann der Fall, wenn ein Zusammenhang zwischen den tatsächlichen Übergangswahrscheinlichkeiten und den Wähleranteilen besteht. In solchen Fällen spricht man von „Aggregationsbias“. Aggregationsbias liegt beispielsweise dann vor, wenn die Bereitschaft der A-Wähler aus der ersten Wahl, sich in der zweiten Wahl für B zu entscheiden, in Wahlkreisen mit hohem A-Anteil in Wahl 1 grösser ist als in andern. Ob in einer Anwendung Aggregationsbias vorliegt, lässt sich anhand der Wähleranteile beider Wahlen nicht beantworten.

Bezüglich der Methodik bei der Parameterschätzung ist nichts gegen die verwendeten Kleinstquadrat-Schätzungen einzuwenden. Diese sind erwartungstreu und konsistent, solange die Modellvoraussetzungen erfüllt sind.

Verschiedene Massnahmen zur Behebung der genannten Schwierigkeiten des einfachen Regressionsmodells sind vorgeschlagen worden. Ein Ansatz besteht darin, die Übergangswahrscheinlichkeiten als lineare Funktion von Kovariablen (numerische Merkmale der Wahlkreise) zu betrachten. Als mögliche Kovariablen sind dabei etwa Merkmale denkbar, welche die soziale Struktur der Wahlkreise widerspiegeln. Diese Modellerweiterung kann in gewissen Fällen zu besseren Schätzungen der Wählerströme führen, doch die Schwächen des Regressionsmodells bleiben bestehen, insbesondere das Auftreten von unmöglichen Schätzungen der Übergangswahrscheinlichkeiten, die Verfälschungseffekte wegen Aggregationsbias und die Unmöglichkeit der Beurteilung, ob Aggregationsbias vorliegt.

Grosse Beachtung hat die Methode von Gary King gefunden, welche ebenfalls als Weiterentwicklung des Regressionsmodells zu betrachten ist (King 1997). Bemerkenswert daran ist, dass die geschätzten Übergangswahrscheinlichkeiten immer im Bereich zwischen 0 und 1 liegen, womit eine Schwierigkeit des Regressionsmodells behoben ist. Kings Lösungsansatz führt allerdings wie das Regressionsmodell bei Aggregationsbias zu falschen Resultaten, und die vorgeschlagenen diagnostischen Graphiken zur Beurteilung von Aggregationsbias sind nicht in allen Fällen zuverlässig. Dies wird übereinstimmend von verschiedenen Autoren berichtet, die Kings Verfahren in praktischen Anwendungen testeten. Deren Erfahrungen weisen darauf hin, dass Kings Methode in Fällen, in denen das einfache Regressionsmodell zu falschen Resultaten führt, häufig auch versagt. King selbst berichtet dagegen ausschliesslich von positiven Erfahrungen bei der praktischen Anwendung seines Modells. Eine wesentliche Einschränkung für den Einsatz der Methode in praktischen Anwendungen bedeutet die Tatsache, dass sich die von King angebotene Software nur für Probleme der Dimensionen  $2 \times 2$  und  $2 \times 3$  eignet.

### 1.2.4 Modelle mit latenten Variablen

Zum Zugang zur ökologischen Inferenz mit latenten Variablen sind in der Literatur wesentlich weniger Vorschläge zu finden. Der bedeutendste Beitrag kommt vom Dänen Søren Thomsen (1987).

In Thomsens multinomialen Logit-Modell wird davon ausgegangen, dass das Wählerverhalten durch eine mehrdimensionale, die politische Einstellung widerspiegelnde latente Variable erklärt werden kann. Für diese latente Einstellungsvariable wird eine Verteilungsannahme formuliert, und die Wahlentscheidung wird als eine Funktion dieser Einstellung modelliert. Die Abwesenheit von Aggregationsbias wird nicht vorausgesetzt, wodurch das Modell zumindest theoretisch über das Potenzial verfügt, zu guten Resultaten zu führen, wo Regression versagt.

Thomsen formuliert auf theoretischer Ebene ein transparentes Modell, seine Argumentationen zur mathematisch-statistischen Begründung seines Lösungsansatzes enthalten aber einen Widerspruch. Was mit seinem Vorgehen letztlich geschätzt wird, bleibt deshalb unklar. Ausserdem ist die Angabe von Streuungsmassen und Vertrauensbereichen nicht möglich, d.h. es handelt sich um eine rein deterministische Schätzung.

Es handelt sich um ein abstraktes Modell, in dem nicht beobachtbare Grössen involviert sind. Die Voraussetzungen können deshalb nicht überprüft werden, selbst bei Vorliegen der vollen Wanderungstabelle nicht (bei Regressionsmodellen ist eine Überprüfung der Voraussetzungen in diesem Fall möglich).

Thomsens multinomialer Logit-Modellansatz weist auf theoretischer Ebene interessante Eigenschaften auf und setzt insbesondere die Abwesenheit von Aggregationsbias nicht voraus, der vorgeschlagene Lösungsweg ist aber unbefriedigend. Deshalb wurde im Rahmen dieser Arbeit nach Alternativen für die Formulierung und Schätzung des Modells gesucht. Es konnte kein vollauf zufriedenstellender Lösungsweg gefunden werden, und gewisse Fragen bezüglich der theoretischen Eigenschaften des Modells mussten unbeantwortet bleiben. Die Ergebnisse im Anwendungsbeispiel deuten darauf hin, dass der verwendete Modellansatz sich nicht zur Modellierung von Wahlzeiten eignet, da er nicht imstande ist, gleichzeitig einen genügend starken Zusammenhang zwischen den Resultaten der beiden Wahlen einerseits und die tatsächlich beobachtete Streuung der Wahlergebnisse der verschiedenen Wahlkreise andererseits zu erfassen. Diese Beurteilung ist aber mit einer gewissen Unsicherheit behaftet, da sie teilweise auf vermuteten und unbewiesenen Modelleigenschaften beruht.

### 1.2.5 Weitere Aspekte in praktischen Anwendungen

Statistische Modelle gehen von einer fiktiven Idealsituation aus, die in der Realität nie exakt erfüllt ist. Falls die Abweichungen zwischen Modell und Wirklichkeit nur geringfügig sind, wird die statistische Analyse trotzdem zu vernünftigen Ergebnissen führen; bei schwerwiegenden Verletzungen der Modellannahmen können sich hingegen völlig falsche Resultate ergeben. Deshalb soll an dieser Stelle der Frage nachgegangen werden, welche Eigenschaften von realen Wahl- oder Abstimmungsdaten zu Verfälschungseffekten führen können, und wie die resultierenden Probleme zu handhaben sind.

- **Nichtwähler:** Der Tatsache, dass die Wähler als Alternative zur Wahl einer Partei die Möglichkeit der Stimmenthaltung haben, kann Rechnung getragen werden, indem die Nichtwähler als eine weitere „Partei“ betrachtet werden. Die Anzahl der Nichtwähler lässt sich leicht ermitteln, nämlich als Differenz der Summe der Stimmen aller Parteien zur Anzahl der Stimmberechtigten.

- Zeitliche Veränderung der Population der Wahlberechtigten: In den Modellen der ökologischen Inferenz muss üblicherweise davon ausgegangen werden, dass die Gesamtheit der Wahlberechtigten zu den Zeitpunkten beider Wahlen identisch war. In Wirklichkeit ist diese Voraussetzung natürlich nicht erfüllt. Mutationen treten auf durch den Umzug von Personen zwischen Wahlkreisen, in das betrachtete Wahlgebiet oder aus dem Wahlgebiet hinaus sowie durch das Erlangen der Wahlberechtigung oder den Tod von Wahlberechtigten zwischen den beiden Wahlterminen.

Liegen die entsprechenden Daten vor, so kann diesem Umstand Rechnung getragen werden, indem für jede Wahl eine zusätzliche Kategorien („Partei“) eingeführt wird:

- Für die erste Wahl diejenige der Personen, die bei der zweiten Wahl im betreffenden Wahlkreis die Wahlberechtigung hatten, in der ersten jedoch nicht,
- für die zweite Wahl diejenige der Personen, die bei der ersten Wahl im betreffenden Wahlkreis die Wahlberechtigung hatten, in der zweiten jedoch nicht.

Sind diese Angaben hingegen nicht verfügbar, so ist es unumgänglich, die Veränderung der Population im Modell zu ignorieren, was zu einer systematischen Verfälschung der Resultate führen kann. Diese ist um so bedeutender, je grösser die Zahl der Mutationen ist, und je stärker das Wahlverhalten der aus der Population Ausgeschiedenen sowie der Neuwähler sich von demjenigen der gesamten Population unterscheidet.

- Betrachtung homogener Teilgebiete: Unterschiedet sich das Wahlverhalten in den verschiedenen Regionen des Untersuchungsgebiets stark, so kann dies einen Widerspruch zu den Modellannahmen bedeuten, in welchen üblicherweise die Homogenität gewisser Aspekte des Wahlverhaltens unterstellt wird. In solchen Fällen kann es sinnvoll sein, eine ökologische Inferenz separat in Teilgebieten vorzunehmen, innerhalb welcher das Wahlverhalten relativ homogen ist. Beispielsweise können die Teilgebiete städtischen und ländlichen Gebieten entsprechen. Die Einteilung der Wahlkreise in Gruppen kann entweder in Übereinstimmung mit einer bestehenden geographischen oder administrativen Einteilung gewählt werden oder z.B. mit dem statistischen Verfahren der Clusteranalyse ausgehend von den Stimmenanteilen der Parteien in den beiden Wahlen vorgenommen werden. Voraussetzung ist dabei eine gewisse Mindestzahl von Wahlkreisen pro Teilgebiet.
- Kleinere Parteien: Sind bei den untersuchten Wahlen kleinere Parteien beteiligt, deren gemeinsamer Stimmenanteil nur wenige Prozent der Wählerstimmen ausmacht, so stellt sich die Frage, ob und wie diese Parteien bei der Analyse berücksichtigt werden sollen. Die übliche Praxis in Wahlanalysen besteht darin, kleinere Parteien zusammenzufassen, und zwar entweder in eine gemeinsame Gruppe „Übrige“, oder in Gruppen von Parteien ähnlicher politischer Ausrichtung. Da ein statistisches Modell nicht mehr zu schätzende Parameter enthalten sollte als die unmittelbar interessierenden, empfiehlt sich dieses Vorgehen insbesondere dann, wenn zahlreiche kleine und kleinste Parteien vorhanden sind, deren Unterscheidung von geringem Interesse ist.

## 1.2.6 Berücksichtigung der Panaschierdaten

In 1.1.1 wurde als ein Teil des Auftrags die Frage gestellt, wie die Information der Panaschierdaten berücksichtigt werden kann, um die Qualität der Rekonstruktion von Wählerströmen zu verbessern. Dazu müssen in einem ersten Schritt aus der Information der Panaschierdaten sinnvolle Masszahlen für Parteiaffinitäten definiert werden, d.h. für die Tendenz der Wähler einer Partei A, Kandidaten von Partei B zu panaschieren. Das einzige uns bekannte Modell der ökologischen Inferenz, welches die Berücksichtigung solcher numerischer Information als Kovariablen bei mehr als zwei Parteien zulässt, ist das Regressionsmodell. Ausgehend von der Arbeit von Burger

(2001) wurde ein erweitertes Regressionsmodell erarbeitet, welches aus den Panaschierdaten gebildete Kovariablen berücksichtigt. Dieser Modellvorschlag erwies sich aber in der exemplarischen Anwendung mit den Daten der Nationalratswahlen 1995/99 im Kanton Zürich als wenig erfolgversprechend. Daraus folgt nicht zwingend, dass kein Zusammenhang zwischen der Panaschierstatistik und den Wählerwanderungen besteht. Falls aber ein solcher besteht, so ist er nicht von der Gestalt, welche in unserem Modell angenommen wird.

### 1.2.7 Publierte Wählerstromanalysen im deutschsprachigen Raum

- Anwendung des Regressionsmodells: Das SORA (Institute for Social Research and Analysis) in Wien stellt auf seiner Internetseite die Resultate verschiedener Wählerstromanalysen in Österreich vor. Den Ausführungen zur Methodik ist zu entnehmen, dass die Resultate mittels Regression ermittelt wurden.
- Anwendungen von Thomsens Methode: Es liegen uns zwei Wahlanalysen vor, welche von Thomsens Modell Gebrauch machen: eine vom Statistischen Informationsdienst der Stadt Freiburg im Breisgau und eine der agis Arbeitsgruppe interdisziplinäre Sozialstrukturforschung, Universität Hannover. In beiden Publikationen wird abgesehen von der Angabe Thomsens (1987) als Quelle nicht auf die verwendete Methodik eingegangen.
- Andreas J. Kohlsche vom Institut für Wahl-, Sozial- und Methodenforschung in Kaufbeuren (Deutschland) hat im deutschsprachigen Raum u.a. in verschiedensten Tageszeitungen Resultate von Wählerwanderungsrekonstruktionen publiziert. Eine vollständige Beschreibung seiner Methode existiert nicht. Kohlsche publiziert keine vollständigen Wanderungstabellen, sondern nur Wanderungssalden, d.h. die Differenzen jeweils zweier entgegengesetzter Wählerwanderungen. Er verwendet Elemente sowohl aus der Regression als auch aus Thomsens Methode. Diese werden in ein Verfahren integriert, welches von Kohlsche als „Die endgültige Lösung des Problems der ökologischen Inferenz“ präsentiert wird. Allerdings enthält Kohlsches Vorgehen verschiedene aus theoretischer Sicht fragwürdige Punkte. Die wesentlichsten Einwände sind dabei:
  - Es wird kein konkretes Modell angegeben, d.h. es wird nicht gesagt, auf welchen Annahmen die Resultate basieren.
  - Die verwendeten Optimalitätskriterien sind nicht ohne weiteres als objektive Beurteilung der Anpassung des Modells an die beobachteten Daten erkennbar. Besonders problematisch scheint uns dabei, dass die Resultate durch die Anwendung von Kohlsches Kriterien inhaltlich in eine bestimmte Richtung gelenkt werden, so beispielsweise bei der Maximierung von Stammwähleranteilen.

Den in 1.1.2 formulierten Anforderungen wird Kohlsches Methode damit in keiner Weise gerecht. Es besteht keine transparente Grundlage in Form eines mathematisch-statistischen Modells, und nichts ist über die Bedingungen bekannt, unter welchen das Verfahren erfolgreich eingesetzt werden kann.

### 1.2.8 Schlussfolgerungen

Bei allen Methoden der ökologischen Inferenz wird versucht, aus beobachtbaren Zusammenhängen in den Wahldaten *über die Wahlkreise hinweg* Erkenntnisse über die unbekannte Abhängigkeitsstruktur *innerhalb der Wahlkreise* zu gewinnen. Solche Schlüsse können nur ausgehend von bestimmten Annahmen gezogen werden, weshalb dem gewählten theoretischen Modellansatz in der ökologischen Inferenz eine zentrale Rolle zukommt. Die Möglichkeit der Überprüfung der getroffenen Annahmen ist bei sämtlichen betrachteten Methoden begrenzt. Eine verlässliche Veri-

fizierung ist bestenfalls bei Vorliegen der vollständigen Wanderungstabelle möglich, in diesem Fall erübrigt sich jedoch eine ökologische Inferenz. Somit beruht jede ökologische Inferenz bis zu einem gewissen Grad auf Vermutungen. Diese Schwierigkeiten liegen weitgehend in der Natur der Aufgabenstellung, so dass unserer Meinung nach auch in Zukunft nicht mit einer vollauf befriedigenden Lösung des Problems zu rechnen ist.

Für eine mögliche Berücksichtigung der Panaschierdaten wurde ein Modellvorschlag erarbeitet, der sich aber in der exemplarischen Anwendung mit konkreten Wahldaten als wenig erfolgversprechend herausstellte.

Aufgrund unserer Untersuchungen können wir keines der untersuchten Verfahren für die routinemässige Anwendung mit den Wahldaten des BFS empfehlen.



## 2 Rekonstruktionsmethoden für (2x2)-Tabellen

Um die verschiedenen Rekonstruktionsmethoden einzuführen und ihre Stärken und Schwächen zu diskutieren, beschränken wir uns vorerst auf den einfachsten Spezialfall, in dem davon ausgegangen wird, dass die Individuen einer unveränderten Population bei zwei Wahlen sich zwischen zwei Parteien A und B zu entscheiden haben. Weiter gehen wir davon aus, dass die wahlberechtigte Population in beiden Wahlen identisch ist, und dass die Stimmbeteiligung 100% beträgt. In Abschnitt 3 erfolgt dann die Verallgemeinerung auf Fälle mit mehr als zwei Wahlalternativen.

### 2.1 Deterministische Betrachtungen in einer (2x2)-Tabelle

Wir betrachten die folgende Situation: eine Population von  $n$  Individuen musste sich in zwei Wahlen jeweils für eine der Parteien A oder B entscheiden. Bekannt sind die Randhäufigkeiten der untenstehenden (2x2)-Tabelle, unbekannt sind  $n_{AA}, n_{BA}, n_{AB}, n_{BB}$ .

Wahl 2	Wahl 1		Total
	Partei A	Partei B	
Partei A	$n_{AA} = ?$	$n_{BA} = ?$	$n_{A2}$
Partei B	$n_{AB} = ?$	$n_{BB} = ?$	$n_{B2}$
Total	$n_{A1}$	$n_{B1}$	$n$

Bezeichnet man die Anteile von Partei A bei den beiden Wahlen als  $x = n_{A1}/n$  bzw.  $y = n_{A2}/n$ , so ergibt sich die folgende Notation, wobei die Einträge nun den relativen Anteilen bezüglich der gesamten Population der Grösse  $n$  entsprechen:

Wahl 2	Wahl 1		Total
	Partei A	Partei B	
Partei A	$p_{AA}x$	$p_{BA}(1-x)$	$y$
Partei B	$(1-p_{AA})x$	$(1-p_{BA})(1-x)$	$1-y$
Total	$x$	$1-x$	$1$

$p_{AA}$  ist dabei der Anteil der A-Wähler der ersten Wahl, die in der zweiten Wahl wieder A wählen.  $p_{BA}$  der Anteil der B-Wähler der ersten Wahl, die in der zweiten Wahl zu A wechseln. Offensichtlich existiert keine eindeutige Lösung. Aus einer der zwei Unbekannten  $p_{AA}$  und  $p_{BA}$  ergibt sich jeweils die andere. Die Zeilenweise Summenbildung in der Tabelle liefert die folgenden linearen Beziehungen:

$$\begin{aligned}
 y &= x \cdot p_{AA} + (1-x) \cdot p_{BA} \\
 1-y &= x \cdot (1-p_{AA}) + (1-x) \cdot (1-p_{BA})
 \end{aligned}$$

Aufgelöst nach  $p_{AA}$  entspricht dies (die Grenzfälle  $x = 0$  und  $x = 1$  werden im Folgenden jeweils ausgeschlossen):

$$p_{AA} = \frac{y}{x} - \frac{1-x}{x} \cdot p_{BA}.$$

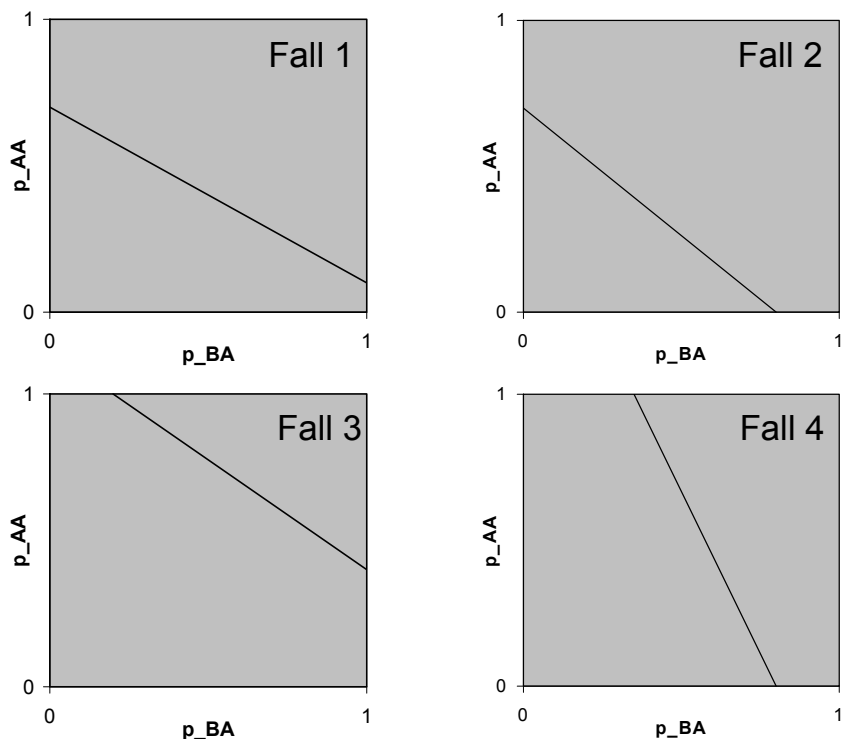
Die möglichen Lösungen für die bivariate Unbekannte  $(p_{BA}, p_{AA})$  liegen auf dem Abschnitt der entsprechenden Geraden innerhalb des Einheitsquadrates  $[0,1] \times [0,1]$ . In Abhängigkeit der bekannten Randhäufigkeiten ergeben sich deterministische Beschränkungen für  $p_{AA}$  und  $p_{BA}$ , wobei vier Fälle unterschieden werden können (vgl. Abbildung 1):

Fall 1:  $x \geq y \geq 1-x \Rightarrow \frac{x+y-1}{x} \leq p_{AA} \leq \frac{y}{x}, \quad 0 \leq p_{BA} \leq 1.$

Fall 2:  $x \geq y, y \leq 1-x \Rightarrow 0 \leq p_{AA} \leq \frac{y}{x}, \quad 0 \leq p_{BA} \leq \frac{y}{1-x}.$

Fall 3:  $x \leq y, y \geq 1-x \Rightarrow \frac{x+y-1}{x} \leq p_{AA} \leq 1, \quad \frac{y-x}{1-x} \leq p_{BA} \leq 1.$

Fall 4:  $x \leq y \leq 1-x \Rightarrow 0 \leq p_{AA} \leq 1, \quad \frac{y-x}{1-x} \leq p_{BA} \leq \frac{y}{1-x}.$



**Abbildung 1:** Illustration der vier Fälle von deterministischen Grenzen für  $p_{AA}$  und  $p_{BA}$ .

In kürzerer Notation lauten diese deterministischen Grenzen:

$$\max\left(0, \frac{x+y-1}{x}\right) \leq p_{AA} \leq \min\left(\frac{y}{x}, 1\right)$$

$$\max\left(0, \frac{y-x}{1-x}\right) \leq p_{BA} \leq \min\left(\frac{y}{1-x}, 1\right)$$

Diese Grenzen schränken die Menge der möglichen Lösungen zwar ein, präzise Informationen liefern sie jedoch nur bei gewissen Konstellationen der Randhäufigkeiten, und auch dann nur für einen der beiden Anteile  $p_{AA}$  und  $p_{BA}$ .

Beispiel:

Wahl 2 Partei A	Wahl 1		Total
	Partei A	Partei B	
Partei B			$y = 80\%$
			$1 - y = 20\%$
Total	$x = 90\%$	$1 - x = 10\%$	100%

Es liegt hier der Fall 1 vor, somit kommen für  $p_{AA}$  Werte zwischen  $\frac{x+y-1}{x} = 77.8\%$  und  $\frac{y}{x} = 88.9\%$  in Frage, für  $p_{BA}$  ist jede Lösung zwischen 0 und 1 möglich.

## 2.2 Mehrere Wahlkreise

Wir betrachten nun ein Untersuchungsgebiet, welches in  $m$  Teilgebiete unterteilt ist, aus denen die Resultate zweier Wahlen bekannt sind. Die Teilgebiete können Gemeinden, Bezirke, usw. sein. Wir verwenden in diesem Bericht die Bezeichnung „Wahlkreise“.

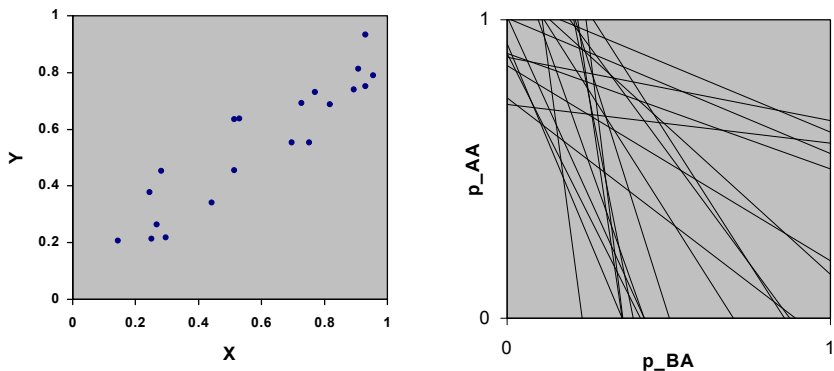
Für jeden der Wahlkreise  $j = 1, \dots, m$  liegt also eine  $(2 \times 2)$ -Tabelle vor, von der nur die Randhäufigkeiten bekannt sind. Die Notation aus dem vorangegangenen Abschnitt wird nun um einen Index  $j$  zur Bezeichnung der Wahlkreise erweitert. Sei also  $x_j$  der Anteil der A-Wähler in Wahlkreis  $j$  bei der ersten,  $y_j$  bei der zweiten Wahl. Die Tabelle im Wahlkreis  $j$  sieht folgendermassen aus:

Wahl 2: Partei A	Wahl 1: Partei A                      Partei B		Total
	$n_{AAj} = p_{AAj} x_j n_j$	$n_{BAj} = p_{BAj} (1 - x_j) n_j$	
Partei B	$n_{ABj} = (1 - p_{AAj}) x_j n_j$	$n_{BBj} = (1 - p_{BAj}) (1 - x_j) n_j$	$n_{B2j} = (1 - y_j) n_j$
Total	$n_{A1j} = x_j n_j$	$n_{B1j} = (1 - x_j) n_j$	$n_j$

Bekannt sind  $x_j, y_j$  und  $n_j$ , gesucht sind  $p_{AAj}$  und  $p_{BAj}$ . Die obige Tabelle mit relativen Anteilen an der Gesamtpopulation anstelle absoluter Häufigkeiten lautet:

Wahl 2: Partei A	Wahl 1: Partei A                      Partei B		Total
	$p_{AAj} x_j$	$p_{BAj} (1 - x_j)$	
Partei B	$(1 - p_{AAj}) x_j$	$(1 - p_{BAj}) (1 - x_j)$	$1 - y_j$
Total	$x_j$	$1 - x_j$	1

Diese Daten können mit einem Punkteschwarm der Wähleranteile  $y_j$  in Abhängigkeit der  $x_j$  dargestellt werden (Abbildung 2, links). Eine Illustration der möglichen Lösungen ist in Abbildung 2 rechts gegeben. Eine Schar von  $m$  Geradenabschnitten im Einheitsquadrat zeigt die zulässigen Wertepaare des Parameterpaares  $(p_{AAj}, p_{BAj})$  für jeweils einen Wahlkreis.



**Abbildung 2:** Beispiel eines Punkteschwarms der Wähleranteile von Partei A in der ersten (X) und zweiten (Y) Wahl, und der daraus resultierenden Schar von Geraden im Einheitsquadrat (fiktive Daten).

Der Anteil der Personen, die in Wahl 2 wieder A wählen, an den A-Wählern der ersten Wahl lautet in der gesamten Population

$$\bar{p}_{AA} = \frac{1}{S} \sum_j p_{AAj} x_j n_j \quad \text{mit } S = \sum_{j=1}^m n_j x_j.$$

Derjenige der B-Wähler aus Wahl 1, die zu A wechseln ergibt sich analog als

$$\bar{p}_{BA} = \frac{1}{T} \sum_j p_{BAj} (1-x_j) n_j. \quad \text{mit } T = \sum_{j=1}^m n_j (1-x_j) = \sum_{j=1}^m n_j - S.$$

$\bar{p}_{AA}$  und  $\bar{p}_{BA}$  sind in Anwendungen diejenigen Grössen, die in erster Linie von Interesse sind. In den folgenden Abschnitten wird jeweils hauptsächlich die Schätzung der Grössen  $p_{AAj}$  und  $p_{BAj}$  in den Wahlkreisen diskutiert; die entsprechenden Schätzungen für die Anteile über alle Wahlreise hinweg ergeben sich aus den geschätzten  $p_{AAj}$  und  $p_{BAj}$  mit den obigen Formeln.

### 2.2.1 Method of bounds

Obere und untere Grenzen für  $p_{AAj}$  und  $p_{BAj}$  innerhalb jedes Wahlkreises können wie oben beschrieben bestimmt werden:

$$\begin{aligned} \max\left(0, \frac{x_j + y_j - 1}{x_j}\right) &\leq p_{AAj} \leq \min\left(\frac{y_j}{x_j}, 1\right) \\ \max\left(0, \frac{y_j - x_j}{1-x_j}\right) &\leq p_{BAj} \leq \min\left(\frac{y_j}{1-x_j}, 1\right) \end{aligned}$$

Die mit den Gewichten  $n_j x_j$  gewichtete Summe der unteren Grenzen von  $p_{AAj}$  ergibt eine deterministische untere Grenze für  $\bar{p}_{AA}$ , analog wird die obere Grenze ermittelt (Duncan/Davis 1953). In der entsprechenden Rechnung für  $\bar{p}_{BA}$  sind die Gewichte  $n_j (1-x_j)$  zu verwenden. Die

Menge der möglichen Lösungen kann auf diese Weise etwas eingeschränkt werden, präzise Informationen über die Populationsparameter  $\bar{p}_{AA}$  und  $\bar{p}_{BA}$  sind auf die Art aber im Allgemeinen nicht zu erwarten.

Sollen präzisere Erkenntnisse gewonnen werden, als sie diese deterministischen Grenzen liefern, so ist es unumgänglich, gewisse Annahmen über die Struktur der unbekannten Inhalte der Innenfelder der Tabelle zu treffen. Im Folgenden wird eine Auswahl von Methoden diskutiert, welche auf verschiedenen Annahmen beruhen und folglich zu unterschiedlichen Resultaten führen.

## 2.2.2 „Ökologischer“ Regressionsansatz nach Goodman

W.S. Robinson stellte 1950 in einer viel beachteten Arbeit fest, dass es in soziologischen Forschungsarbeiten weit verbreitet war, den Korrelationskoeffizienten mit aggregierten (oder „ökologischen“) Daten zu verwenden und die Modellinterpretation auf der Stufe der Individuen vorzunehmen. Anhand konkreter Zahlenbeispiele zeigte er auf, wie dies zu Fehlschlüssen führen kann („ökologischer Fehlschluss“). Beispielsweise sei der Anteil der Analphabeten in den USA unter den im Ausland geborenen Einwohnern 1930 höher gewesen als in der Gesamtpopulation. Der Korrelationskoeffizient zwischen Analphabetenrate und dem im Ausland geborenen Bevölkerungsanteil nach Bundesstaat betrug dagegen  $-0.526$ , da die Einwanderer sich häufiger in den reicheren Staaten niederliessen, in denen der Anteil der Analphabeten unter dem Landesdurchschnitt lag. Goodman (1953, 1959) untersuchte in der Folge, unter welchen Annahmen die Verwendung von linearer Regression mit aggregierten Daten zulässig ist.

Anmerkung zur Notation: Die Wähleranteile von Partei A in den beiden Wahlen werden jeweils als Grossbuchstaben  $X_j$  und  $Y_j$  geschrieben, wenn sie in einem Modell als Zufallsvariablen auftreten. Mit  $x_j$  und  $y_j$  werden sie bezeichnet, wenn sie innerhalb des jeweiligen Modells als feste beobachtete Werte betrachtet werden.

Modell:  $p_{AA}$  und  $p_{BA}$  sind Populationsparameter, welche in allen Wahlkreisen identisch seien.  $x_j$  wird als fest vorgegeben betrachtet (also nicht stochastisch), und  $Y_j$  folgendermassen modelliert:

$$Y_j = p_{AA}x_j + p_{BA}(1 - x_j) + U_j,$$

wobei die  $U_j$  unabhängig und identisch verteilte Zufallsvariablen mit Erwartungswert  $E(U_j) = 0$  sind. Dies entspricht der Regression

$$Y_j = \alpha + \beta x_j + U_j \quad \text{mit} \quad \alpha = p_{BA} \quad \text{und} \quad \beta = p_{AA} - p_{BA}.$$

Für die Koeffizienten  $\alpha$  und  $\beta$  können die üblichen Kleinstquadrate-Schätzungen  $\hat{\alpha}$  und  $\hat{\beta}$  verwendet werden. Bei Gültigkeit der Annahmen werden dann  $p_{AA}$  und  $p_{BA}$  erwartungstreu und statistisch konsistent geschätzt durch  $\hat{p}_{AA} = \hat{\alpha} + \hat{\beta}$  und  $\hat{p}_{BA} = \hat{\alpha}$ .

Offensichtlich gilt in diesem Modell  $\bar{p}_{AA} = p_{AA}$  und  $\bar{p}_{BA} = p_{BA}$ , so dass sich ein weiterer Rechenschritt zur Ermittlung der Parameter  $\bar{p}_{AA}$  und  $\bar{p}_{BA}$  erübrigt.

Graphisch bedeutet dies, dass sich die Geraden aller Wahlkreise abgesehen von den durch den Fehlerterm  $U_j$  verursachten Abweichungen in einem gemeinsamen Punkt  $(p_{BA}, p_{AA})$  kreuzen.

Wie sind die Parameter  $p_{AA}$  und  $p_{BA}$  in diesem Modell zu interpretieren? Im Fall einer einzigen Tabelle haben wir  $p_{AA}$  und  $p_{BA}$  als die *exakten Anteile* der A-Wähler (bzw. der B-Wähler) aus Wahl A betrachtet, welche in der zweiten Wahl A wählen. Bei dieser Betrachtungsweise müsste im Fall mehrerer Wahlkreise die Gleichung

$$y_j = p_{AA}x_j + p_{BA}(1 - x_j),$$

für jedes  $j$  exakt erfüllt sein, d.h. der Punkt  $(p_{BA}, p_{AA})$  müsste auf der entsprechenden Geraden liegen. In der Darstellung von Abbildung 2 rechts würden sich somit sämtliche Geraden in diesem Punkt kreuzen. Dies ist in der Praxis kaum je erfüllt. Interpretiert man hingegen  $p_{AA}$  und  $p_{BA}$  als Wahrscheinlichkeiten, dass sich ein A- bzw. B-Wähler aus der ersten Wahl in der zweiten Wahl für A entscheidet, so sind Abweichungen von der obigen Gleichung zulässig und können als zufällige Abweichungen vom Erwartungswert aufgefasst werden. Wir werden demnach  $p_{AA}$  und  $p_{BA}$  als **Übergangswahrscheinlichkeiten** bezeichnen.

#### Einige Bemerkungen und Überlegungen zu diesem Modell:

1. Lösungen für  $p_{AA}$  und  $p_{BA}$  ausserhalb von  $[0,1]$  sind möglich. Bei annähernd gleichen Übergangswahrscheinlichkeiten in allen Wahlkreisen sollten sie (bei genügend grosser Anzahl Wahlkreise) aufgrund der statistischen Konsistenz der Schätzungen nur in Ausnahmefällen auftreten. Als eine mögliche Massnahme gegen dieses „**Out of bounds**“-Problem bietet sich die Verwendung nichtlinearer Transformationen  $f: (0,1) \rightarrow \mathbf{R}$  an:

$$f(Y_j) = \alpha + \beta f(x_j) + U_j$$

(z.B. Thomsen 1987, p. 19). Problematisch an diesem Ansatz ist jedoch, dass die Parameter  $\alpha$  und  $\beta$  keine eindeutige Interpretation wie im linearen Modell mehr besitzen. Es ist nicht klar, wie die interessierenden Wählerströme aus den Parameterschätzungen von  $\alpha$  und  $\beta$  zu ermitteln sind.

2. Anwendungen zeigen, dass die getroffenen Annahmen in der Praxis häufig zu stark sind. Oft sind die  $p_{AAj}$  bzw.  $p_{BAj}$  in verschiedenen Wahlkreisen nicht identisch, sondern weisen einen Zusammenhang mit den  $x_j$  auf. In der englischsprachigen Literatur spricht man dann von „aggregation bias“, also von **Aggregationsbias**. In solchen Fällen sind die üblichen Kleinstquadrat-Schätzungen aus dem Regressionsmodell für  $p_{AA}$  und  $p_{BA}$  weder erwartungstreu noch statistisch konsistent. Auch im Fall eines Zusammenhangs zwischen den Übergangswahrscheinlichkeiten und der Populationsgrösse  $n_j$  in den Wahlkreisen kann es zu einer systematischen Verfälschung der Schätzungen der  $\bar{p}_{AA}$  und  $\bar{p}_{BA}$  kommen.
3. Im Modell liegt eine Asymmetrie in der Betrachtung von  $X$  und  $Y$  vor. Es ist offensichtlich, dass die Wähleranteile in der ersten und zweiten Wahl gleichermassen zufälligen Schwankungen unterworfen sind, und diesem Umstand wird nicht bei beiden Variablen in gleichem Mass Rechnung getragen. Die  $x_j$  werden als feste Werte betrachtet, die  $Y_j$  hingegen als Zufallsvariablen. Diese Modellformulierung kann aber damit begründet werden, dass die Betrachtungen bedingt über die Wahlresultate der ersten Wahl erfolgen.

4. Die Symmetrie ist in einer weiteren Hinsicht nicht gegeben: Vertauscht man die Rollen von  $X$  und  $Y$  als Ziel- und Einflussgrösse, so ändert sich auch die Grundannahme des Regressionsmodells. Bezeichnet man mit  $a_j$  den Populationsanteil, der zweimal A wählt, so lautet die Annahme gleicher Übergangswahrscheinlichkeiten im Regressionsmodell mit Zielgrösse  $Y$   $a_j / x_j \approx \text{const.}$ , im Modell mit Zielgrösse  $X$  hingegen  $a_j / y_j \approx \text{const.}$ . Diese zwei Annahmen sind nur dann (annähernd) identisch, wenn sich  $x_j$  und  $y_j$  annähernd proportional zueinander verhalten.
5. Werden  $p_{AA}$  und  $p_{BA}$  als *Übergangswahrscheinlichkeiten* aufgefasst, so könnte man auf Stufe Individuum das folgende Modell annehmen: Sei  $i$  ein Index zur Bezeichnung der Individuen (Wähler), und  $W_{1ji}, W_{2ji}$  seien Zufallsvariablen, die den Wert 1 annehmen, falls sich Wähler  $i$  in Wahlkreis  $j$  in der ersten bzw. zweiten Wahl für Partei A entscheidet, 0 sonst. Weiter geht man davon aus, dass die Wähler ihre Entscheidungen unabhängig voneinander treffen. Dann gilt

$$P(W_{2ji} = 1 | W_{1ji} = 1) = p_{AA} \text{ und } P(W_{2ji} = 1 | W_{1ji} = 0) = p_{BA}.$$

Die folgenden Betrachtungen erfolgen bedingt über den Ausgang der ersten Wahl.  $n_{AAj} = \sum_{\{i|W_{1ji}=1\}} W_{2ji}$  folgt einer Binomialverteilung  $B(n_{A1j}, p_{AA})$  und  $n_{BAj} = \sum_{\{i|W_{1ji}=0\}} W_{2ji}$  stammt aus einer  $B(n_{B1j}, p_{BA})$ . Die Anzahl der A-Wähler in der zweiten Wahl,  $n_j Y_j = n_{AAj} + n_{BAj}$ , ist somit die Summe zweier binomialverteilter Zufallsvariablen, und es gilt:

$$\text{Var}(n_j Y_j) = n_j x_j p_{AA} (1 - p_{AA}) + n_j (1 - x_j) p_{BA} (1 - p_{BA})$$

Wegen  $\text{Var}(U_j) = \text{Var}(Y_j)$  wiedergibt die Gleichung

$$\text{Var}(U_j) = \frac{x_j p_{AA} (1 - p_{AA}) + (1 - x_j) p_{BA} (1 - p_{BA})}{n_j}$$

die Heteroskedastizität, die nach diesem Modellansatz in der Regressionsbeziehung zu erwarten wäre. Gemäss Achen/Shively (1995, p.49) macht es jedoch wenig Sinn, solche Betrachtungen in die Berechnungen einfließen zu lassen, da in Anwendungen die Variabilität sehr viel grösser ist als nach der obigen Formel. Zugunsten des Regressionsmodells könnte argumentiert werden, die Annahme völliger Unabhängigkeit der Wahlentscheidungen der Wähler innerhalb eines Wahlkreises sei unrealistisch, und die deutlich grössere Streuung in dieser Regression entstehe durch Abhängigkeiten in diesen Wahlentscheidungen.

Diese Betrachtungen werfen zudem die Frage auf, ob ein theoretisches Modell auf Stufe Individuum als Rechtfertigung des verwendeten Ansatzes auf der Stufe aggregierter Daten erforderlich bzw. wünschenswert ist. Die Meinungen verschiedener Autoren gehen in dieser Frage auseinander. Achen/Shively (1995) berufen sich beispielsweise wiederholt auf Modelle auf Stufe Individuum, um die Qualitäten eines Modells mit aggregierten Daten zu beurteilen. King (1997) hingegen findet, das Postulieren eines Modells für Individuen sei nicht wesentlich für die Beurteilung eines Modells mit aggregierten Daten (p. 119-122).

6. Man kann sich fragen, ob es sinnvoll wäre, die Regression mit der Anzahl Wähler in den Wahlkreisen ( $n_j$ ) zu gewichten. Dies ist dann sinnvoll, wenn die Annahme gleicher Vertei-



lung aller  $U_j$  im Regressionsmodell durch die Bedingung  $\text{Var}(U_j) = \sigma^2 / n_j$  ersetzt wird. Diese hat eine gewisse Verwandtschaft mit der Formel in Punkt 5. Die Verwendung einer Gewichtung ist von geringer Bedeutung, wenn die verschiedenen Wahlkreise von ähnlicher Grösse sind. Sind jedoch wesentliche Unterschiede in deren Grösse vorhanden, so sind aus der gewichteten Regression in der Regel realistischere Schätzungen der Populationsparameter  $\bar{p}_{AA}$  und  $\bar{p}_{BA}$  zu erwarten. Eine ausführliche Diskussion der Problematik der Gewichtung in der ökologischen Regression findet man bei Achen und Shively (1995), p.57ff.

7. Ein einfaches Illustrationsbeispiel für das Versagen des Regressionsmodells nach Goodman bei nicht erfüllten Voraussetzungen geben Achen/Shively (1995, p.79/80):

Es wird angenommen, dass jeder Wähler eine (nicht zwingende) Präferenz für eine der beiden Parteien A und B hat (latente Variable „Parteiidentifikation“). Die Anteile der A-Sympathisanten  $z_j$  in den Wahlkreisen seien unterschiedlich ( $0 \leq z_j \leq 1$ ), im gesamten Untersuchungsgebiet betrage dieser Anteil 50%, d.h. es gelte  $\sum_j n_j z_j / \sum_j n_j = \frac{1}{2}$ . Weiter wird angenommen, dass jeder Wähler mit einer Wahrscheinlichkeit von 80% für „seine“ Partei stimmt, mit einer Wahrscheinlichkeit von 20% für die andere, und ausserdem fallen alle Teilnehmer ihre Wahl unabhängig voneinander. Ist nun die Anzahl Wähler innerhalb jedes Wahlkreises so gross, dass der zufällige Fehlerterm vernachlässigbar wird, so gilt  $y_j \approx x_j$ , und die Punkte der ökologischen Regression liegen fast genau auf einer Geraden durch den Koordinatenursprung mit Steigung 1. Somit wird (abgesehen von vernachlässigbaren Abweichungen)  $\hat{\alpha} = 0$  und  $\hat{\beta} = 1$ , und es folgt  $\hat{p}_{AA} = \hat{p}_{BB} = 1$  und  $\hat{p}_{BA} = \hat{p}_{AB} = 0$ . Es lässt sich jedoch leicht berechnen, dass tatsächlich gilt  $\bar{p}_{AA} = \bar{p}_{BB} \approx 0.8 \cdot 0.8 + 0.2 \cdot 0.2 = 0.68$  und  $\bar{p}_{BA} = \bar{p}_{AB} \approx 0.8 \cdot 0.2 + 0.2 \cdot 0.8 = 0.32$ . Der Ansatz nach Goodman führt somit hier zu völlig falschen Resultaten. Dies liegt daran, dass die Grundannahme gleicher  $p_{AAj}$  und  $p_{BAj}$  in allen Wahlkreisen verletzt ist. Es besteht ein positiver Zusammenhang zwischen  $p_{AAj}$  und  $x_j$ , d.h. es liegt Aggregationsbias vor:

$$p_{AAj} \approx \frac{0.8^2 \cdot z_j + 0.2^2 \cdot (1 - z_j)}{0.8 \cdot z_j + 0.2 \cdot (1 - z_j)} = \frac{0.04 + 0.6 \cdot z_j}{0.2 + 0.6 \cdot z_j} = 1 - \frac{0.16}{0.2 + 0.6 \cdot z_j} \approx 1 - \frac{0.16}{x_j}.$$

Dieses Beispiel zeigt ausserdem, dass eine hohe Korrelation zwischen den  $x_j$  und den  $y_j$  keinesfalls als eine Bestätigung des Modellansatzes gewertet werden darf.

### **Fazit zu 2.2.2:**

1. **Modell:** Das einfache Regressionsmodell ist einfach, gut interpretierbar und profitiert sicher auch vom allgemein hohen Bekanntheitsgrad der Methode der linearen Regression. Für die Anwendung zur Rekonstruktion von Wählerströmen scheint jedoch die Annahme gleicher Übergangswahrscheinlichkeiten in allen Wahlkreisen häufig zu stark. Insbesondere im Fall von Korrelationen der Übergangswahrscheinlichkeiten mit Wähleranteilen bei der ersten Wahl (Aggregationsbias) oder mit der Populationsgrösse der Wahlkreise kann die Methode irreführend sein.
2. **Überprüfbarkeit:** Die Überprüfung der Annahme gleicher Übergangswahrscheinlichkeiten ist aus den Randtotalen nicht möglich. Liegen die geschätzten Übergangswahrscheinlichkeiten (= Regressionskoeffizienten) mehrheitlich und deutlich ausserhalb des Intervalls  $[0,1]$ , so ist da-

von auszugehen, dass die Annahme verletzt ist. Der umgekehrte Schluss ist nicht zulässig: auch wenn alle Koeffizienten im Einheitsintervall liegen, besteht keine Garantie für die Gültigkeit der Annahmen.

3. Schätzung: Bei erfüllten Modellvoraussetzungen werden die Wählerströme erwartungstreu und statistisch konsistent geschätzt. Bei Verletzungen der Modellbedingungen ist dies nicht der Fall, so etwa im Fall von Aggregationsbias.

### 2.2.3 Modellerweiterung mit Kovariablen

Ist die Annahme identischer Übergangswahrscheinlichkeiten  $p_{AA}$  und  $p_{BA}$  in allen Wahlkreisen zu stark, so kann diesem Problem begegnet werden, indem man  $p_{AA}$  und  $p_{BA}$  als lineare Funktion einer Kovariablen  $z_j$  modelliert:

$$p_{AAj} = \gamma_1 + \delta_1 z_j \quad \text{und}$$

$$p_{BAj} = \gamma_2 + \delta_2 z_j.$$

Die Kovariable  $z_j$  ist ein numerisches Merkmal der Wahlkreise, bei der ein Zusammenhang mit den Übergangswahrscheinlichkeiten  $p_{AAj}$  und  $p_{BAj}$  vermutet wird. Diese Annahmen in die Hauptgleichung der ökologischen Regression eingesetzt führen zum multiplen Regressionsmodell

$$Y_j = (\gamma_1 + \delta_1 z_j)x_j + (\gamma_2 + \delta_2 z_j)(1 - x_j) + U_j$$

$$\alpha + \beta_1 x_j + \beta_2 z_j + \beta_3 x_j z_j + U_j$$

$$\text{mit } \alpha = \gamma_2, \beta_1 = \gamma_1 - \gamma_2, \beta_2 = \delta_2 \text{ und } \beta_3 = \delta_1 - \delta_2.$$

Die Berücksichtigung von Kovariablen kann bei Vorliegen von Aggregationsbias im einfachen Regressionsansatz Abhilfe schaffen. Aggregationsbias kann aber auch im erweiterten Modell vorkommen: Jeder Zusammenhang des Fehlerterms  $U_j$  mit einer Kovariablen ( $x_j, z_j$  oder  $x_j z_j$ ) bedeutet Aggregationsbias und kann zu verzerrten Schätzungen führen.

Ist  $z_j = x_j$ , so liegt ein Identifikationsproblem vor. Der Steigungsparameter der Kovariablen  $x_j$  lautet  $\gamma_1 - \gamma_2 + \delta_2$ , und die  $p_{AAj}$  und  $p_{BAj}$  können nicht eindeutig geschätzt werden. Somit führt dieser Ansatz nicht zum Ziel, es sei denn, eine weitere a priori-Annahme über die Parameter wird eingeführt, etwa  $\gamma_1 = \gamma_2$  oder  $\delta_2 = 0$ .

Das Out of bounds-Problem, d.h. die Möglichkeit von Schätzungen der  $p_{AAj}$  und  $p_{BAj}$  ausserhalb des Einheitsintervalls, bleibt erhalten. Als mögliche Massnahme bietet sich die Verwendung einer bijektiven Funktion  $f: (0,1) \rightarrow \mathbf{R}$  (z.B. Logit- oder Probit) an. Ersetzt man im Goodman-Ansatz  $p_{AA}$  durch  $f^{-1}(\gamma_1 + \delta_1 z_j)$  sowie  $p_{BA}$  durch  $f^{-1}(\gamma_2 + \delta_2 z_j)$ , so erhält man

$$Y_j = f^{-1}(\gamma_1 + \delta_1 z_j)x_j + f^{-1}(\gamma_2 + \delta_2 z_j)(1 - x_j) + U_j \quad \text{resp.}$$

$$Y_j = \alpha(z_j) + \beta(z_j)x_j + U_j$$

$$\text{mit } \alpha(z_j) = f^{-1}(\gamma_2 + \delta_2 z_j) \text{ und } \beta(z_j) = f^{-1}(\gamma_1 + \delta_1 z_j) - f^{-1}(\gamma_2 + \delta_2 z_j).$$

Dies führt zu einem nichtlinearen Modell in den Parametern  $\gamma_1, \delta_1, \gamma_2, \delta_2$ . Die Verwendung von  $x_j$  als Kovariable führt hier in der Regel nicht mehr zu einem singulären Modell, doch eine

starke Kollinearität in der Einflussgrößen bleibt erhalten. Achen/Shively (1995, p.45) schreiben hierzu: „estimation will be delicate and inference doubtful“. Modelle dieser Art scheinen in der Literatur nicht näher untersucht worden zu sein.

Allgemeinere Modelle mit mehreren Kovariablen oder unterschiedlichen Kovariablen für  $p_{AAj}$  und  $p_{BAj}$  sind denkbar. Ihre Eigenschaften sind grundsätzlich die gleichen, welche hier für den Spezialfall einer identischen Kovariable für  $p_{AAj}$  und  $p_{BAj}$  diskutiert wurden.

### **Fazit zu 2.2.3:**

1. **Modell:** Das erweiterte einfache Regressionsmodell ist immer noch relativ einfach, verständlich und gut interpretierbar. Die Berücksichtigung von Kovariablen kann in manchen Fällen bei Vorliegen von Aggregationsbias im einfachen Regressionsmodell Abhilfe schaffen, Aggregationsbias kann aber auch hier vorkommen.
2. **Überprüfbarkeit:** Die Überprüfung der Modellannahmen aus den Randtotalen ist nicht möglich. Liegen die geschätzten Übergangswahrscheinlichkeiten mehrheitlich und deutlich ausserhalb des Intervalls  $[0,1]$ , so ist davon auszugehen, dass die Annahmen verletzt sind. Der umgekehrte Schluss ist nicht zulässig.
3. **Schätzung:** Bei erfüllten Modellvoraussetzungen werden die Wählerströme erwartungstreu und statistisch konsistent geschätzt.

### **2.2.4 Modelle mit zufälligen Übergangswahrscheinlichkeiten**

Eine weitere Möglichkeit, die starke Annahme gleicher Übergangswahrscheinlichkeiten  $p_{AA}$  und  $p_{BA}$  in allen Wahlkreisen zu lockern, besteht darin, die Grössen  $p_{AA}$  und  $p_{BA}$  nicht wie bisher als feste Koeffizienten zu betrachten, sondern als Zufallsvariablen, welche in verschiedenen Wahlkreisen unterschiedliche Werte annehmen können.

**Modell:**  $(p_{AAj}, p_{BAj})$ ,  $j = 1, \dots, m$ , seien bivariate Zufallsvektoren, unabhängig aus einer Verteilung  $F$  auf dem Einheitsquadrat. Setze  $E(p_{AAj}) = p_{AA}$ ,  $\text{Var}(p_{AAj}) = \sigma_{AA}^2$ ,  $E(p_{BAj}) = p_{BA}$ ,  $\text{Var}(p_{BAj}) = \sigma_{BA}^2$  und  $\text{Cov}(p_{AAj}, p_{BAj}) = \sigma$ .

$U_j$  seien zufällige, von den  $(p_{AAj}, p_{BAj})$  unabhängige Fehlerterme mit  $E(U_j) = 0$  und  $\text{Var}(U_j) = \sigma_U^2$ . Eingesetzt in die übliche Gleichung der ökologischen Regression ergibt sich

$$\begin{aligned} Y_j &= p_{AAj}x_j + p_{BAj}(1-x_j) + U_j \\ &= p_{AA}x_j + p_{BA}(1-x_j) + (p_{AAj} - p_{AA})x_j + (p_{BAj} - p_{BA})(1-x_j) + U_j \\ &= p_{AA}x_j + p_{BA}(1-x_j) + V_j \end{aligned}$$

Es wird ersichtlich, dass dieser Ansatz einer gewichteten Variante der Goodman-Regression entspricht, in welcher der Fehlerterm  $V_j$  Erwartungswert 0 und die folgende Varianz hat:

$$\text{Var}(V_j) = x_j^2 \sigma_{AA}^2 + (1-x_j)^2 \sigma_{BA}^2 + 2x_j(1-x_j)\sigma + \sigma_U^2.$$

In einer alternativen Formulierung, bei dem die Realisationen von  $(p_{AAj}, p_{BAj})$  nicht als Wahrscheinlichkeiten, sondern als exakte Anteile aufgefasst werden, entfällt der Fehlerterm  $U_j$  in der Modellgleichung, und dementsprechend fehlt der Summand  $\sigma_U^2 = \text{Var}(U_j)$  in der Varianz von  $V_j$ . In diesem Fall beschränkt sich die bedingte Verteilung von  $(p_{AAj}, p_{BAj})$  gegeben  $Y_j = y_j$  auf eine Gerade, welche durch  $y_j = p_{AAj}x_j + p_{BAj}(1-x_j)$  definiert ist.

Achen/Shively (1995, p. 51ff.) schlagen ein iteratives Verfahren zur Schätzung des obigen Modells vor. Dabei wird zuerst eine ungewichtete Regression gerechnet, dann werden aus den Residuen die Gewichte geschätzt. Diese zwei Schritte werden iteriert, bis die Resultate konvergieren. Weiter schreiben sie, dass sich die gewichtete Rechnung insbesondere im Fall von zwei Parteien in der Praxis kaum vom ungewichteten Ansatz nach Goodman unterscheidet. Insbesondere werden die folgenden Probleme des einfachen Goodman-Ansatzes nicht behoben:

- Out of bounds-Lösungen für  $p_{AA}$  und  $p_{BA}$  ausserhalb von  $[0,1]$  bleiben möglich.
- Besteht ein Zusammenhang von  $p_{AAj}$  und  $p_{BAj}$  mit den  $x_j$  (Aggregationsbias), so sind die  $(p_{AAj}, p_{BAj})$  nicht mehr unabhängig und identisch verteilt, d.h. die Modellvoraussetzungen sind verletzt.

#### **Fazit zu 2.2.4:**

Modelle mit zufälligen Übergangswahrscheinlichkeiten können als alternative Formulierung von Regressionsmodellen betrachtet werden und führen nicht zu wesentlich anderen Lösungen als der einfache Regressionsansatz. Ihre Beurteilung anhand der Kriterien Modell, Überprüfbarkeit und Schätzung ist damit weitgehend identisch mit dem, was in 2.2.2 gesagt wurde.

### **2.2.5 Gary Kings Ecological Inference (EI)**

Gary King (1997) stellt einen etwas weiterentwickelten Ansatz eines Modells mit zufälligen Übergangswahrscheinlichkeiten vor. Er geht von der Gleichung

$$Y_j = p_{AAj}x_j + p_{BAj}(1-x_j)$$

aus, und wählt als Verteilung der  $(p_{AAj}, p_{BAj})$  eine auf das Einheitsquadrat beschränkte Normalverteilung, d.h. eine Verteilung mit bivariater Dichte

$$f(p_{AAj}, p_{BAj}) = \varphi_2(p_{AAj}, p_{BAj}; \mu, \Sigma) \cdot \frac{1_{[0,1]}(p_{AAj}) \cdot 1_{[0,1]}(p_{BAj})}{\int_0^1 \int_0^1 \varphi_2(\pi_1, \pi_2; \mu, \Sigma) d\pi_1 d\pi_2},$$

wo  $\varphi_2(\cdot, \cdot; \mu, \Sigma)$  die Dichte einer bivariaten Normalverteilung mit Mittelwertvektor  $\mu$  und Kovarianzmatrix  $\Sigma$  bezeichnet. Sein Lösungsansatz besteht aus einem Maximum Likelihood- und einen Simulationsschritt. Im ersten Schritt werden die Maximum-Likelihood-Schätzungen der fünf

Parameter in  $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$  und  $\Sigma = \begin{pmatrix} \sigma_{AA}^2 & \sigma \\ \sigma & \sigma_{BA}^2 \end{pmatrix}$  ermittelt. Aus der resultierenden bedingten Verteilung von  $(p_{AAj}, p_{BAj})$  gegeben  $Y_j = y_j$  generiert er Simulationen, aus denen er dann durch Mit-

telwertbildung die geschätzten Werte  $\hat{p}_{AAj}$  und  $\hat{p}_{BAj}$  bestimmt. Die Standardabweichung der simulierten Werte verwendet er als standard errors. Dieses Vorgehen hat die schöne Eigenschaft, dass „out of bounds“-Lösungen nicht möglich sind, d.h. alle Schätzungen von  $(\hat{p}_{AAj}, \hat{p}_{BAj})$  liegen im Einheitsquadrat. King stellt weiter graphische Darstellungsweisen und Diagnostiken vor, welche die Überprüfung der Modellvoraussetzungen ermöglichen sollen. Um das Problem des Aggregationsbias anzugehen, erweitert King zudem sein Modell mit der Möglichkeit, die Übergangswahrscheinlichkeiten analog zur Regression (vgl. 2.2.3) als lineare Funktion von Kovariablen zu modellieren.

Die Veröffentlichung des Buches, in dem seine Methode vorgestellt wird (King 1997), fand grosses Echo, und da King selbst im Internet entsprechende Software anbietet, kann davon ausgegangen werden, dass das Verfahren in der Praxis zum Einsatz kommt. Trotz einigen offensichtlich positiven Eigenschaften, namentlich die Einbindung der deterministischen Grenzen in ein regressionsähnliches Modell, ist Kings Modell unter Fachleuten jedoch umstritten.

Die Erfahrungen, die einige Autoren bei der Anwendung von Kings Methode machten, sollen hier zitiert werden:

- Cho (1998) setzt sich sehr differenziert mit Kings Methode auseinander und kommt zu den folgenden Erkenntnissen. „El represents a genuine advancement to ecological inference in that it incorporates two elements that have never previously existed together in aggregate data models. The combination of the method of bounds and allowance for varying parameters brings a new degree of efficiency to aggregate data analysis.“ Weiter sei die Abwesenheit von Aggregationsbias die Hauptvoraussetzung dafür, dass das Verfahren gute Resultate liefert. Die vorgeschlagenen Diagnostiken erlauben jedoch nicht immer eine zuverlässige Beurteilung: „However, the diagnostics are problematic in that they do not always signal deviations from the model even when they do exist. Alternatively, the diagnostics sometimes point toward a poor model fit when the estimates are actually quite reasonable“. Ein weiteres Problem ortet Cho für den Anwender, der den Mangel eines vermuteten Aggregationsbias beheben will, indem er das Modell mit Kovariablen erweitert. Verschiedene Modelle mit verschiedenen Kovariablen führen zu unterschiedlichen Resultaten, und hierbei stellt sich das Problem fehlender Anhaltspunkte für die Wahl einer bestimmten Lösung. „An obvious problem with adding covariates is that King provides no method of choosing covariates (outside of utilizing qualitative information and substantive beliefs). However, selecting the proper covariates, or determining the proper specification, is the heart of the problem. [...] Formal and objective tests are necessary. Visual examinations of the tomography plots is simply not enough“.
- Palmquist (1999) stellt fest, dass Kings Lösungsansatz oft in denselben Situationen zu falschen Resultaten führt wie die einfache ökologische Regression, namentlich bei Vorliegen von Aggregationsbias.
- Freedman et al. (1998) testen Kings Methode an zwei realen Datensätzen, in denen die genauen Daten bekannt sind (dabei geht es zwar nicht um die Frage von Wählerwanderungen, sondern um Wähleranteile in verschiedenen Bevölkerungsgruppen, doch das Problem ist bezüglich der Aufgabenstellung identisch). Sie kommen zum Schluss, dass „his [King's] method produces results that are far from truth, and diagnostics are unreliable“. Kings Methode liefere Resultate, die sich nicht wesentlich von denjenigen aus dem einfachen Goodman-Ansatz unterscheiden.

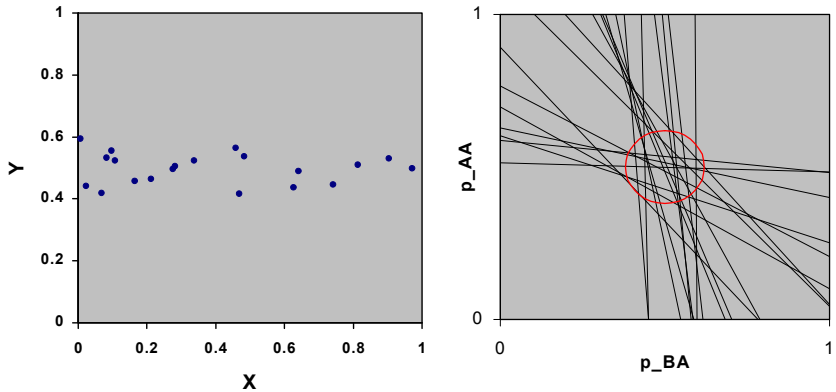
- McCue (2001) untersucht die in Kings Buch dargelegte mathematisch-statistische Methodik im Detail und findet etliche Ungenauigkeiten und Fehler. Er kommt zum folgenden Schluss: Kings Methode „is in fact an application of the standard statistical theory of prediction, though with many statistical errors (...)“.

Die Erfahrungen der zitierten Autoren mit Kings Methode stehen teilweise in Widerspruch zu dem, was King selbst über die Anwendung an realen Beispielen mit vollständig bekannten Daten berichtet. Zwar stellt King bereits auf der ersten Seite der Einleitung zu seinem Buch fest, dass „Because the ecological inference problem is caused by the lack of individual-level information, no method of ecological inference, including that introduced in this book, will produce precisely accurate results in every instance“ (p. xv). Das Fazit, welches er weiter unten aus der Anwendung an fünf Datenbeispielen zieht, lautet dann aber : „The method works in practice: it gives accurate answers and correct assessments of uncertainty even when existing methods lead to incorrect inferences or impossible results“ (p. xvii).

Aus der Lektüre von Kings Buch entsteht der Eindruck, dass die Beurteilung seiner eigenen Methode nicht ganz „unverfälscht“ ist. Aussagen wie „all components of the proposed model are in large part verifiable in aggregate data (...) using diagnostic tests to evaluate the appropriateness of the model to each application“ (p. 20) sind in dieser Allgemeingültigkeit ungenügend fundiert, wie die Erfahrungen anderer Autoren zeigen. Die Wirksamkeit der Diagnostiken ist mit künstlichen Extrembeispielen illustriert (z.B. p.176). In der Tat lassen sich ohne weiteres Datenbeispiele konstruieren, in denen Kings Methode zu falschen Resultaten führt, ohne dass dies aus den entsprechenden diagnostischen Abbildungen ersichtlich würde.

Im Kapitel 9 „What can go wrong?“ werden mit künstlichen Daten verschiedene „Worst case scenarios“ diskutiert, wobei entweder eine Verletzung der Modellannahmen aus einem diagnostischen Plot ersichtlich wird (9.1.1), oder die Schätzung der Parameter  $\bar{p}_{AA}$  und  $\bar{p}_{BA}$  trotz unerfüllter Voraussetzungen akkurat ausfällt (9.1.2). Eine tatsächliche „Worst case“-Situation ergibt sich aus einer Modifikation des in 9.1.2 diskutierten Falls. Dort beschreibt King die in Abbildung 3 dargestellte Situation, in der sämtliche Linien, welche die möglichen Punkte für die  $(p_{AAj}, p_{BAj})$  im Einheitsquadrat enthalten<sup>3</sup>, mit sehr unterschiedlichen Steigungen durch den zentralen Bereich  $\{(p_{AAj}, p_{BAj}) : 0.4 \leq p_{AAj}, p_{BAj} \leq 0.6\}$  des Einheitsquadrats verlaufen. Die mit seiner Methode geschätzte Verteilung der  $(p_{AAj}, p_{BAj})$  konzentriert sich auf diesen (in der Graphik rechts in Abbildung 3 eingezeichneten) Bereich, die entsprechende bivariate Normalverteilung wird nur unwesentlich gestutzt.

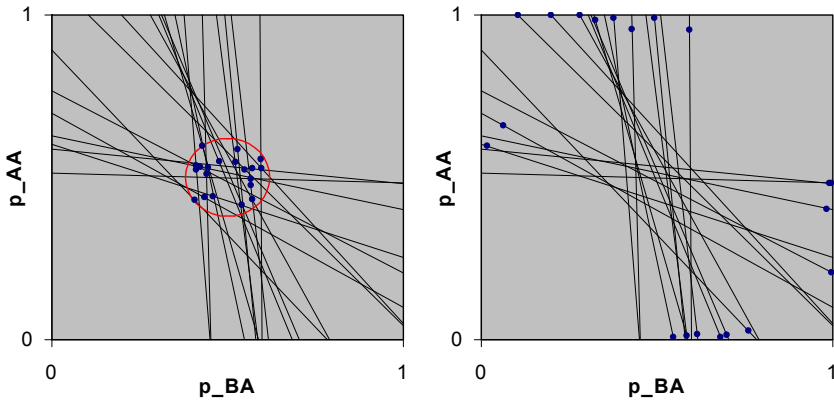
<sup>3</sup> King bezeichnet diese Linien als *tomography lines* und die entsprechende Graphik als *tomography plot*.



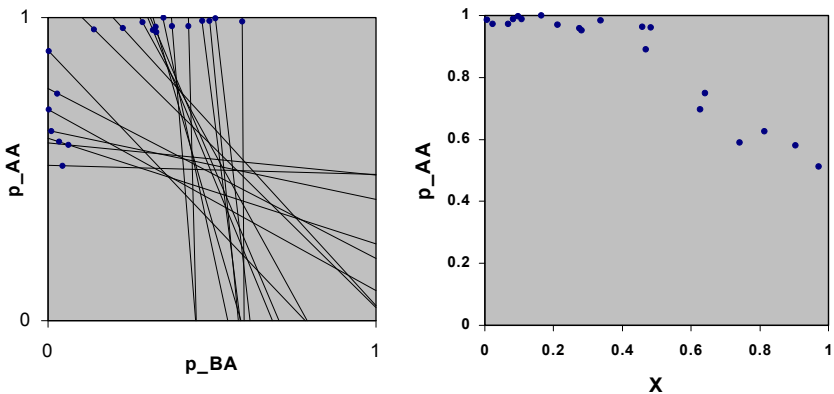
**Abbildung 3:** Qualitative Wiedergabe der in King (1997), Bsp 9.1.2, diskutierten Situation.

Gemäss Kings Modell müssten die Punkte  $(p_{AAj}, p_{BAj})$  für alle Wahlkreise in diesen zentralen Bereich fallen, wie dies in Abbildung 4 links illustriert ist. Als „Worst of Distributional Violations“ betrachtet King den Fall von Abbildung 4 rechts, in dem sich die tatsächlichen Punkte  $(p_{AAj}, p_{BAj})$  an den Rändern des Einheitsquadrats befinden, und zwar so, dass die  $p_{AAj}$  je etwa zur Hälfte in der Nähe ihrer oberen und unteren deterministischen Grenze liegen. Er stellt dabei fest, dass die Schätzung der Parameter  $\bar{p}_{AA}$  und  $\bar{p}_{BA}$  gut ausfällt, obwohl die geschätzte Verteilung der  $(p_{AAj}, p_{BAj})$  offensichtlich falsch ist.

Werden nun stattdessen alle Punkte auf dieselbe Seite des Zentrums gesetzt, so dass z.B. alle  $p_{AAj}$  wie in Abbildung 5 links in der Nähe ihrer oberen deterministischen Grenze liegen, so wird  $\bar{p}_{AA}$  systematisch unter- und  $\bar{p}_{BA}$  überschätzt, ohne dass diese Verletzung einer Modellvoraussetzung (Aggregationsbias, d.h. es besteht ein Zusammenhang zwischen  $p_{AAj}$  und  $x_j$ , wie das Bild rechts in Abbildung 5 zeigt) aus den entsprechenden Diagnostiken sichtbar würde. Diese können einen vorhandenen Aggregationsbias nur dann aufdecken, wenn die Normalverteilung von  $(p_{AAj}, p_{BAj})$  an den Rändern des Einheitsquadrats genügend stark gestützt ist.



**Abbildung 4:** Lage der Punkte  $(p_{AAj}, p_{BAj})$  in Übereinstimmung mit Kings Modell (links), und der von King diskutierte „Worst of Distributional Violations“-Fall (rechts).



**Abbildung 5:** Modifikation der Situation aus Abbildung 4 (links), und Illustration des Zusammenhangs der  $x_j$  und  $p_{AAj}$  (rechts).

Etwas voreilig mutet auch Kings Feststellung an, sein Modell sei robust gegenüber Aggregationsbias. Diese gründet offenbar einzig auf der Untersuchung *eines* realen Zahlenbeispiels (chapter 11), in dem seine Methode gute Schätzungen der (in diesem Fall bekannten)  $p_{AAj}$  und  $p_{BAj}$  liefert, obwohl Aggregationsbias vorliegt.



### Fazit zu 2.2.5:

1. Modell: Kings Modell stellt insofern einen deutlichen Fortschritt dar, dass der Ansatz des Regressionsmodells und die deterministischen Grenzen der Übergangswahrscheinlichkeiten in einem sauber definierten Modell vereint werden.
2. Überprüfbarkeit: Die vorgestellten Diagnostiken können Modellverletzungen aufdecken, falls diese sehr ausgeprägt sind. Wie verschiedene Autoren übereinstimmend berichten, können diese Diagnostiken aber auch irreführend sein. Das Problem des Aggregationsbias bleibt wie bei der Regression die grösste Schwierigkeit in praktischen Anwendungen.
3. Schätzung: Diese beruht auf dem in der Statistik weit verbreiteten Maximum Likelihood-Kriterium, welches gemäss Theorie zu nicht zwingend erwartungstreuen, aber statistisch konsistenten Schätzungen führt. Die Anwendung des Kriteriums ist allerdings nicht ganz korrekt, falls man McCue (2001) Glauben schenkt.

### 2.2.6 Modelle mit latenten Variablen

Als Alternative zu Regressionsmodellen kann der Zugang zur ökologischen Inferenz mit latenten Variablen gewählt werden. Bei diesem Ansatz wird nicht versucht, die Anteile der zweiten Wahl direkt mit denjenigen der ersten Wahl zu erklären, sondern es wird von der Existenz einer quantifizierbaren Grösse „politische Einstellung“ der Wahlberechtigten (latente Variable  $z$ ) ausgegangen, welche für das Wahlverhalten in beiden Wahlen verantwortlich ist.

Ein einfaches Beispiel eines solchen Modells wurde bereits in Punkt 7 im Abschnitt 2.2.2 (Seite 31) präsentiert. Dort wurde angenommen, dass alle Wählenden eine (nicht zwingende) Präferenz für eine der beiden Parteien A und B haben (latente Variable „Parteiidentifikation“). Die dort geschilderte Situation kann etwas allgemeiner formuliert werden, indem in Wahl 1 nun jeder A-Sympathisant mit einer Wahrscheinlichkeit  $p_1$  für „seine“ Partei A und jede B-Sympathisantin mit einer Wahrscheinlichkeit  $q_1$  für B stimmt. Die entsprechenden Wahrscheinlichkeiten in der zweiten Wahl lauten  $p_2$  und  $q_2$ . Unter diesen Voraussetzungen können die Wähleranteile von Partei A in beiden Wahlen in Abhängigkeit der latenten Variable  $Z_j$  „Anteil der Population, der Partei A nahesteht“ folgendermassen dargestellt werden:

$$X_j = p_1 Z_j + q_1 (1 - Z_j) + U_j$$

$$Y_j = p_2 Z_j + q_2 (1 - Z_j) + V_j$$

Gegenüber den auf Regression beruhenden Modellen werden in diesem Ansatz die Variablen  $X$  und  $Y$  gleichwertig behandelt, d.h. es besteht Symmetrie zwischen diesen beiden Variablen. In der Parametrisierung

$$X_j = q_1 + (p_1 - q_1) Z_j + U_j$$

$$Y_j = q_2 + (p_2 - q_2) Z_j + V_j$$

wird die faktorenanalytische Form des Ansatzes deutlich. Achen/Shively (1995, chapter 7) geben eine auf einem Modell auf Stufe der Individuen basierende Formel für die Varianzen von  $U_j$  und  $V_j$  an und diskutieren das Problem der eindeutigen Identifikation der Parameter  $p_1, p_2, q_1, q_2$  (Faktorladungen). Die in der Faktorenanalyse übliche Skalierung von  $Z_j$  ( $E(Z_j) = 0$  und  $\text{Var}(Z_j) = 1$ ) macht hier wenig Sinn, falls die  $z_j$  als Anteile der Wählenden, die A nahestehen, und die Faktorladungen als Wahrscheinlichkeiten interpretiert werden sollen. Achen und Shively bieten kein konkretes Schätzverfahren für diesen Modellansatz.

Bemerkenswert am beschriebenen Modell ist die Tatsache, dass die in der ökologischen Regression und ähnlichen Modellen problematische Annahme der Abwesenheit von Aggregationsbias nicht getroffen werden muss. Achen und Shively meinen deshalb: „Such models have a stronger substantive interpretation than ecological regression and a better record of producing plausible estimates“ (1995, p.188), vorausgesetzt jedoch, dass für die zahlreichen offenen methodischen Probleme eine Lösung gefunden werden kann. Bis heute scheinen jedoch keine vielversprechenden Arbeiten zu diesem Modellansatz vorzuliegen.

## 2.2.7 Søren Thomsens Logit / Probit-Modell

Einen anderen Zugang mit latenten Variablen als in 2.2.6 präsentiert Søren Thomsen (1987, 2000). Sein Vorschlag soll hier vorerst für den Spezialfall einer 2x2-Tabelle dargestellt werden.

Die Werte von  $k$  latenten Variablen von Person  $i$  im Wahlkreis  $j$  werden im Zufallsvektor  $Z_{ji} = (Z_{ji}^{(1)}, \dots, Z_{ji}^{(k)})$  zusammengefasst. Diese sind als persönliche Positionen auf einer  $k$ -dimensionalen numerischen Skala der politischen Einstellung zu verstehen ( $k \geq 2$ ).

Verteilungsannahmen: Die Verteilung der latenten Variablen der Individuen in Wahlkreis  $j$  folgen einer  $k$ -dimensionalen Normalverteilung

$$Z_{ji} | \mu_j \sim N_k(\mu_j, \Sigma), \quad \Sigma = \text{Diag}_k(\sigma^2) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \sigma^2 \end{pmatrix},$$

wobei die (bedingten) Wahlkreis-Erwartungswerte  $\mu_j$  ihrerseits aus einer  $k$ -dimensionalen Normalverteilung

$$\mu_j = (\mu_j^{(1)}, \dots, \mu_j^{(k)}) \sim N_k(0, I_k), \quad I_k = \text{Diag}_k(1)$$

stammen. Die unbedingte Verteilung der  $Z_{ji} = (Z_{ji}^{(1)}, \dots, Z_{ji}^{(k)})$  lautet folglich

$$Z_{ji} \sim N_k(0, \text{Diag}(1 + \sigma^2)).$$

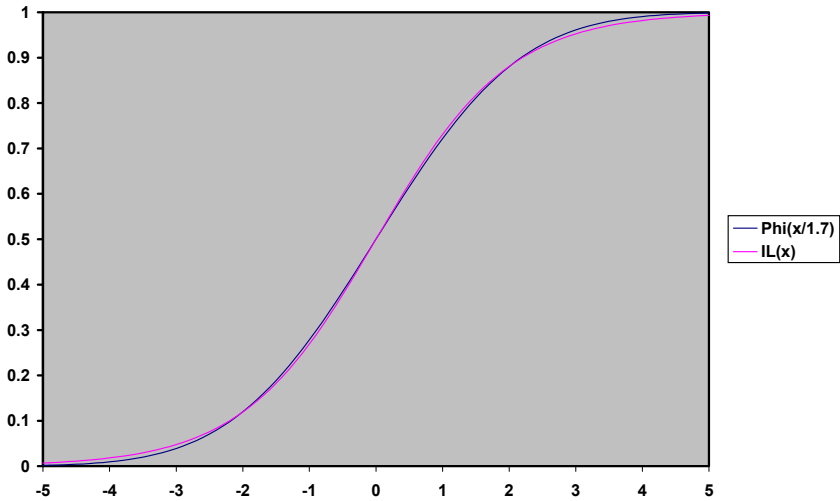
Wahl von Person  $i$  in Wahlkreis  $j$  in Wahl  $t = 1, 2$ : Seien  $W_{1ji}, W_{2ji}$  Zufallsvariablen, welche die Wahl von Person  $i$  in Wahlkreis  $j$  wiedergeben:  $W_{1ji} = 1$ , falls sich diese in der ersten Wahl für A entscheidet, 0 sonst.  $W_{2ji}$  entspricht analog der Entscheidung in der zweiten Wahl. Der Modellansatz lautet

$$P(W_{tji} = 1) = \Phi(\alpha_t + \beta_t' Z_{ji}) = \Phi\left(\alpha_t + \sum_{d=1}^k \beta_t^{(d)} Z_{ji}^{(d)}\right), \quad \beta_t = (\beta_t^{(1)}, \dots, \beta_t^{(k)}),$$

wo  $\alpha_t, \beta_t^{(1)}, \dots, \beta_t^{(k)}$  feste reelle Koeffizienten und  $\Phi$  die Verteilungsfunktion der standardisierten Normalverteilung sind. Thomsen verwendet die inverse Logit-Funktion<sup>4</sup>  $IL(x) = \frac{e^x}{1 + e^x}$  anstelle von  $\Phi(x)$ , was hinsichtlich einer Verallgemeinerung des Modells für mehr als zwei Parteien sinnvoll ist, macht aber bei der Modellschätzung von der Approximation  $\Phi(x/1.7) \approx IL(x)$

<sup>4</sup>  $y = IL(x)$  ist die Umkehrfunktion der Logit-Funktion  $L(y) = \log\left(\frac{x}{1-x}\right)$ .

Gebrauch, die in Abbildung 6 illustriert ist. Wir formulieren hier das Modell direkt mit der bei der Schätzung letztlich verwendeten Funktion  $\Phi(x)$ , wobei der Skalierungsfaktor 1.7 ignoriert wird, da er durch die entsprechende Anpassung der Koeffizienten  $\beta_t^{(d)}$  berücksichtigt werden kann.



**Abbildung 6:** Illustration der Approximation der inversen Logit-Funktion  $IL(x) = e^x / (1 + e^x)$  durch die (skalierte) Verteilungsfunktion der Normalverteilung,  $\Phi(x/1.7)$ .

Unter den dargelegten Voraussetzungen kann gezeigt werden, dass die folgenden Resultate gelten:

$$P(W_{ji} = 1 | \mu_j) = \Phi \left( \frac{\alpha_t + \beta'_t \mu_j}{\sqrt{1 + \beta'_t \Sigma \beta'_t}} \right) = \Phi \left( \frac{\alpha_t + \sum_{d=1}^k \beta_t^{(d)} \mu_j^{(d)}}{\sqrt{1 + \sigma^2 \cdot \sum_{d=1}^k (\beta_t^{(d)})^2}} \right),$$

$$P(W_{ji} = W_{2ji} = 1 | \mu_j) = \Phi_2 \left( \frac{\alpha_1 + \beta'_1 \mu_j}{\sqrt{1 + \beta'_1 \Sigma \beta'_1}}, \frac{\alpha_2 + \beta'_2 \mu_j}{\sqrt{1 + \beta'_2 \Sigma \beta'_2}}; \rho_w \right),$$

mit

$$\begin{aligned} \rho_w &= \frac{\beta'_1 \Sigma \beta'_2}{\sqrt{1 + \beta'_1 \Sigma \beta'_1} \sqrt{1 + \beta'_2 \Sigma \beta'_2}} = \frac{\sigma^2 \cdot \beta'_1 \beta'_2}{\sqrt{1 + \sigma^2 \cdot \beta'_1 \beta'_1} \sqrt{1 + \sigma^2 \cdot \beta'_2 \beta'_2}} \\ &= \frac{\beta'_1 \beta'_2}{\sqrt{1/\sigma^2 + \beta'_1 \beta'_1} \sqrt{1/\sigma^2 + \beta'_2 \beta'_2}} \end{aligned}$$

( $\Phi_2(x, y; \rho)$  ist die kumulative Verteilungsfunktion  $P(X \leq x, Y \leq y)$  für einen bivariat normalverteilten Zufallsvektor  $(X, Y)$  mit  $\mu_X = \mu_Y = 0$ ,  $\sigma_X^2 = \sigma_Y^2 = 1$  und  $\rho_{XY} = \rho$ ). Der Anteil der

Wähler in Wahlkreis  $j$ , die zweimal Partei A wählen, entspricht etwa der Wahrscheinlichkeit  $P(W_{1ji} = W_{2ji} = 1 | \mu_j) = \Phi_2 \left( \frac{\alpha_1 + \beta'_1 \mu_j}{\sqrt{1 + \beta_1 \Sigma \beta'_1}}, \frac{\alpha_2 + \beta'_2 \mu_j}{\sqrt{1 + \beta_2 \Sigma \beta'_2}}; \rho_w \right)$ . Da die  $\mu_j$  aus einer Normalverteilung stammen, kann  $\rho_w$  als tetrachorischer Korrelationskoeffizient der (2x2)-Tabelle

Wahl 2 Partei A	Wahl 1 Partei A	Partei B	
	$a_j$	$y_j - a_j$	
Partei B	$x_j - a_j$	$1 - y_j - x_j + a_j$	$1 - y_j$
	$x_j$	$1 - x_j$	1

für Wahlkreis  $j$  aufgefasst werden<sup>5</sup>, wo  $a_j = p_{AAj} \cdot x_j$  dem Populationsanteil in Wahlkreis  $j$  entspricht, der zweimal A wählt. Eine Möglichkeit der Schätzung des tetrachorischen Korrelationskoeffizienten aus den Zellenwerten bietet das Mass „Yule's  $Q$ “, welches im vorliegenden Fall folgendermassen lautet:

$$\rho_w \approx \frac{a_j(1 + a_j - x_j - y_j) - (x_j - a_j)(y_j - a_j)}{a_j(1 + a_j - x_j - y_j) + (x_j - a_j)(y_j - a_j)} \quad \forall j.$$

Um nun eine Schätzung von  $\rho_w$  aus den Wähleranteilen  $x_j$  und  $y_j$  zu ermöglichen, stellt Thomsen die folgenden Überlegungen an: Mit dem Gesetz der grossen Zahlen folgt aus

$$P(W_{ijt} = 1 | \mu_j) = \Phi \left( \frac{\alpha_t + \beta'_t \mu_j}{\sqrt{1 + \beta_t \Sigma \beta'_t}} \right) \text{ für } \bar{W}_{ijt}, \text{ dem A-Wähleranteil in Wahlkreis } j \text{ in der Wahl } t:$$

$$\bar{W}_{ijt} \approx \Phi(\tilde{\alpha}_t + \tilde{\beta}'_t \mu_j)$$

$$\text{mit } \tilde{\alpha}_t = \frac{\alpha_t}{\sqrt{1 + \beta_t \Sigma \beta'_t}} \text{ sowie } \tilde{\beta}_t = \frac{\beta_t}{\sqrt{1 + \beta_t \Sigma \beta'_t}}.$$

Dies lässt sich umformulieren zu  $\Phi^{-1}(\bar{W}_{ijt}) \approx \tilde{\alpha}_t + \tilde{\beta}'_t \mu_j$ , d.h. mit der Notation  $\bar{W}_{1j} = X_j$  und  $\bar{W}_{2j} = Y_j$  gilt:

$$\begin{aligned} \Phi^{-1}(X_j) &\approx \tilde{\alpha}_1 + \tilde{\beta}'_1 \mu_j \\ \Phi^{-1}(Y_j) &\approx \tilde{\alpha}_2 + \tilde{\beta}'_2 \mu_j \end{aligned}$$

Die Funktion  $\Phi^{-1}()$  ist die Quantilfunktion der standardisierten Normalverteilung und wird als Probit-Funktion bezeichnet. Die Verteilung von  $(\Phi^{-1}(X_j), \Phi^{-1}(Y_j))$  kann also durch eine bivariate Normalverteilung approximiert werden, deren Korrelationskoeffizient

<sup>5</sup> Der Begriff des tetrachorischen Korrelationskoeffizienten einer (2x2)-Tabelle basiert auf der Annahme, die Tabelle sei durch eine Dichotomisierung beider Variablen einer bivariaten Normalverteilung entstanden. Der tetrachorische Korrelationskoeffizient entspricht dann dem (Pearson-)Korrelationskoeffizienten dieser ursprünglichen Normalverteilung. Für Details s. z.B. Kotz/Johnson/Read 1988, p.223ff.

$$\rho_e = \text{Corr}(\Phi^{-1}(X_j), \Phi^{-1}(Y_j)) = \text{Corr}(\tilde{\beta}'_1 \mu_j, \tilde{\beta}'_2 \mu_j) = \frac{\beta'_1 \beta_2}{\sqrt{\beta'_1 \beta_1} \sqrt{\beta'_2 \beta_2}}$$

lautet. Nun vollzieht Thomsen den Grenzübergang  $\sigma \rightarrow \infty$  mit der Begründung, dies entspreche der Annahme, dass der Quotient  $\frac{\text{Var}(\mu_j)}{\text{Var}(Z_{ij} | \mu_j)} = \frac{1}{\sigma^2}$  gegen 0 strebt, d.h. der Grenzfall wiedergebe die Situation gleicher Verteilung der  $Z_{ji}$  in allen Wahlkreisen. Wegen

$$\lim_{\sigma^2 \rightarrow \infty} \rho_w = \lim_{\sigma^2 \rightarrow \infty} \frac{\beta'_1 \beta_2}{\sqrt{1/\sigma^2 + \beta'_1 \beta_1} \sqrt{1/\sigma^2 + \beta'_2 \beta_2}} = \frac{\beta'_1 \beta_2}{\sqrt{\beta'_1 \beta_1} \sqrt{\beta'_2 \beta_2}} = \rho_e$$

könne  $\rho_w$  unter der genannten Annahme gleichgesetzt werden mit  $\rho_e$ , welches als Pearson-Korrelation zwischen den  $\Phi^{-1}(X_j)$  und den  $\Phi^{-1}(Y_j)$  geschätzt werden könne. Da  $\rho_w$  andererseits als tetrachorischer Korrelationskoeffizient aufgefasst werden kann, der sich durch Yule's  $Q$  approximieren lässt, ergibt sich für alle Wahlkreise  $j = 1, \dots, m$ :

$$\rho_e = \frac{a_j(1 + a_j - x_j - y_j) - (x_j - a_j)(y_j - a_j)}{a_j(1 + a_j - x_j - y_j) + (x_j - a_j)(y_j - a_j)}.$$

Dies kann nach  $a_j$  aufgelöst werden, und es ergibt sich als eindeutige Lösung ( $\rho_e(a_j)$  ist monoton für  $0 \leq a_j \leq \min(x_j, y_j)$ ):

$$a_j = \frac{1}{4\rho_e} \left( 1 - \rho_e + 2\rho_e(x_j + y_j) - \sqrt{(1 - \rho_e + 2\rho_e(x_j + y_j))^2 - 8\rho_e(1 + \rho_e)x_j y_j} \right)$$

(resp.  $a_j = x_j y_j$  falls  $\rho_e = 0$ ).

Thomsens Methode scheint zwar auf den ersten Blick elegant, doch der Grenzübergang, welcher zur asymptotischen Identität von  $\rho_w$  und  $\rho_e$  führt, hält einer genauen Überprüfung nicht stand. Die Verteilung der Variablen  $W_{ji}$  degeneriert im Grenzwert  $\sigma \rightarrow \infty$  und bei konstant gehaltenen  $\alpha_i$  und  $\beta_i$  zu einer Bernoulliverteilung mit Parameter  $p = \frac{1}{2}$  bei völliger Unabhängigkeit aller  $W_{ji}$ . Bei Gültigkeit dieses Modells ist zwar die asymptotische Identität von  $\rho_w$  und  $\rho_e$  korrekt, doch die Korrelation  $\rho_e$  wird zwischen den im Grenzfall unkorrelierten Variablen  $\Phi^{-1}(X_j)$  und  $\Phi^{-1}(Y_j)$  ermittelt und kann folglich keine substantielle Information über die Stärke des Zusammenhangs in der Tabelle liefern.

Soll nun der Grenzübergang  $\sigma = \sqrt{\text{Var}(Z_{ji})} \rightarrow \infty$  so vollzogen werden, dass die beobachtbare Korrelation  $\rho_e$  nicht gegen 0 strebt, so dürfen die Zufallsvariablen  $P(W_{1ji} = 1) = \Phi(\alpha_1 + \beta'_1 Z_{ji})$  und  $P(W_{2ji} = 1) = \Phi(\alpha_2 + \beta'_2 Z_{ji})$  im Grenzwert nicht unabhängig werden. Dies ist dann gewährleistet, wenn die Koeffizienten  $\beta_i$  proportional zu  $\sigma$  mitwachsen, so dass  $\dot{\beta}_i = \beta_i / \sigma$  konstant bleibt. In diesem Fall degeneriert das Modell nicht, d.h.  $\rho_e$  strebt nicht gegen 0. Allerdings strebt dann

$$\rho_w = \frac{\sigma^2 \cdot \beta'_1 \beta_2}{\sqrt{1 + \sigma^2 \cdot \beta'_1 \beta_1} \sqrt{1 + \sigma^2 \cdot \beta'_2 \beta_2}} = \frac{\dot{\beta}'_1 \dot{\beta}_2}{\sqrt{1 + \dot{\beta}'_1 \dot{\beta}_1} \sqrt{1 + \dot{\beta}'_2 \dot{\beta}_2}}$$

nicht gegen  $\rho_e = \frac{\beta'_1 \beta_2}{\sqrt{\beta'_1 \beta_1} \sqrt{\beta'_2 \beta_2}} = \frac{\dot{\beta}'_1 \dot{\beta}_2}{\sqrt{\dot{\beta}'_1 \dot{\beta}_1} \sqrt{\dot{\beta}'_2 \dot{\beta}_2}}$ , d.h. die asymptotische Identität von  $\rho_w$  und  $\rho_e$  ist nicht gewährleistet.

In Thomsens Lösungsvorschlag wird nun einerseits durch die Interpretation von  $\rho_e$  als Mass für die Stärke der Abhängigkeit der Entscheidung in der ersten und zweiten Wahl unterstellt, dass  $\rho_e$  nicht gegen 0 strebt, was aufgrund obiger Argumentation impliziert, dass der Quotient  $\beta_i / \sigma$  konstant bleibt. Andererseits wird bei der Gleichsetzung von  $\rho_e$  und  $\rho_w$  der Grenzübergang so interpretiert, dass der Übergang  $\sigma \rightarrow \infty$  bei unveränderten  $\beta_i$  erfolgt. Thomsens Argumentation enthält damit einen eindeutigen Widerspruch. Dieses Problem wird auch von Achen und Shively (1995, p.187) diskutiert.

Der zweite Fall in der obigen Schilderung kann auch anhand einer etwas allgemeineren Formulierung der Verteilungsannahme illustriert werden:

$$\mu_j = (\mu_j^{(1)}, \dots, \mu_j^{(k)}) \sim N_k(0, T) \text{ mit } T = \text{Diag}_k(\tau^2),$$

Mit dem Grenzübergang  $\tau^2 \rightarrow 0$  findet man

$$\rho_w = \frac{(\sigma^2 + \tau^2) \cdot \beta'_1 \beta_2}{\sqrt{1 + (\sigma^2 + \tau^2) \beta'_1 \beta_1} \cdot \sqrt{1 + (\sigma^2 + \tau^2) \beta'_2 \beta_2}} = \frac{\beta'_1 \beta_2}{\sqrt{1/(\sigma^2 + \tau^2) + \beta'_1 \beta_1} \cdot \sqrt{1/(\sigma^2 + \tau^2) + \beta'_2 \beta_2}}$$

und der Grenzwert lautet

$$\lim_{\tau^2 \rightarrow 0} \rho_w = \frac{\beta'_1 \beta_2}{\sqrt{1/\sigma^2 + \beta'_1 \beta_1} \sqrt{1/\sigma^2 + \beta'_2 \beta_2}} \neq \frac{\beta'_1 \beta_2}{\sqrt{\beta'_1 \beta_1} \sqrt{\beta'_2 \beta_2}} = \rho_e.$$

### **Fazit zu 2.2.7:**

1. Modell: Auf theoretischer Ebene wird ein klares Modell formuliert, welches vom Vorhandensein einer mehrdimensionalen, die politische Einstellung widerspiegelnden latenten Variablen ausgeht. Für diese latente Einstellungsvariable wird eine Verteilungsannahme formuliert, und die Wahlentscheidung wird als eine Funktion dieser Einstellung modelliert. Die Modellannahmen implizieren nicht die Abwesenheit von Aggregationsbias und können somit zumindest theoretisch in Situationen zu guten Resultaten führen, in denen Regression versagt.
2. Überprüfbarkeit: Das Modell ist abstrakt, die Voraussetzungen nicht überprüfbar, selbst bei Vorliegen der vollen Wanderungstabelle nicht. Überprüft werden kann hingegen die aus den Modellannahmen folgende Eigenschaft der Normalverteilung der transformierten Wähleranteile  $\Phi^{-1}(X_j)$  und  $\Phi^{-1}(Y_j)$ .
3. Schätzung: Es wurde oben gezeigt, dass die Begründung von Thomsen Schätzverfahren einen Widerspruch enthält. Was mit dem beschriebenen Vorgehen letztlich genau geschätzt wird, bleibt unklar. Abgesehen davon handelt es sich um eine rein deterministische Schätzung, d.h. die Angabe von Streuungsmassen (standard errors) und Vertrauensbereichen ist nicht möglich.

### 2.2.8 Exemplarischer Vergleich der einfachen Regression mit Thomsens Modell

An dieser Stelle soll als Illustration ein kleiner Vergleich der Resultate vorgenommen werden, welche aus diesen zwei Modellen hervorgehen. Wir interessieren uns für den Wähleranteil  $a_j$  in einem Wahlkreis  $j$ , der in beiden Wahlen für Partei A stimmt. Dieser wird geschätzt

- im Regressionsmodell als

$$\begin{aligned}\hat{a}_j &= \hat{p}_{AAj} \cdot x_j = (\hat{\alpha} + \hat{\beta}) \cdot x_j \\ &= \left( \bar{y} + (1 - \bar{x}) \cdot r_{xy} \cdot \frac{s_x}{s_y} \right) \cdot x_j\end{aligned}$$

- gemäss Thomsens Methode als

$$\hat{a}_j = \frac{1}{4\rho_e} \left( 1 - \rho_e + 2\rho_e(x_j + y_j) - \sqrt{(1 - \rho_e + 2\rho_e(x_j + y_j))^2 - 8\rho_e(1 + \rho_e)x_j y_j} \right)$$

resp.  $\hat{a}_j = x_j y_j$  falls  $\rho_e = 0$ .

Die folgenden Betrachtungen gehen von der Annahme aus, dass die empirischen Korrelationskoeffizienten  $r_{xy} = \text{Corr}(X, Y)$  und  $\rho_e = \text{Corr}(\Phi^{-1}(X), \Phi^{-1}(Y))$  ungefähr gleich sind. Dies ist dann der Fall, wenn die Ausgänge beider Wahlen ausgeglichen sind, d.h. die Anteile nicht nahe bei 0 oder 1 liegen.

In den Spezialfällen  $\rho_e \approx r_{xy} \approx -1, 0$  und  $1$  führt dies zu den folgenden Ergebnissen für  $\hat{a}_j$ :

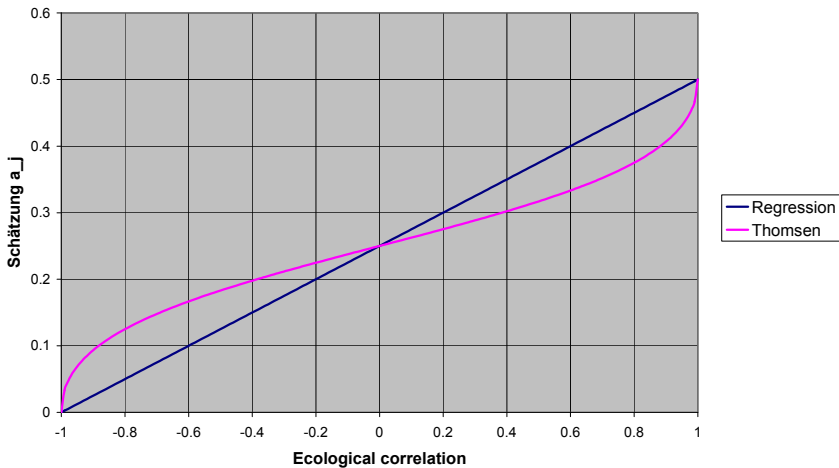
	$\rho_e \approx r_{xy} \approx -1$	$\rho_e \approx r_{xy} \approx 0$	$\rho_e \approx r_{xy} \approx 1$
Regression	$\bar{y} - (1 - \bar{x}) \cdot \frac{s_x}{s_y}$	$x_j \cdot \bar{y}$	$\bar{y} + (1 - \bar{x}) \cdot \frac{s_x}{s_y}$
Thomsen	0	$x_j \cdot y_j$	$\min(x_j, y_j)$

Die Resultate aus Thomsens Modell in diesen Spezialfällen sind leicht zu interpretieren: im Fall  $\rho_e = 0$  wird aus dem fehlenden Zusammenhang *zwischen* den Resultaten verschiedener Wahlkreise geschlossen, dass auch kein Zusammenhang *innerhalb* der einzelnen Tabellen besteht. Demnach ist die Setzung  $\hat{a}_j = x_j \cdot y_j$  (entspricht dem in 2.2.9 erwähnten „Neighbourhood Model“) konsequent. Die Resultate in den Fällen  $\rho_e = -1$  und  $\rho_e = 1$  entsprechen den Grenzen des Bereichs der möglichen Werte von  $a_j$ .

Das Resultat des Regressionsmodells im Fall  $r_{xy} = 0$  lässt sich damit begründen, dass die Modellannahme gleicher Übergangswahrscheinlichkeit in allen Wahlkreisen eingehalten werden muss, insofern ist die Lösung,  $\hat{a}_j = p_{AA} x_j = x_j \cdot \bar{y}$  erklärbar. Für die Lösungen in den Fällen  $\rho_e = r_{xy} = -1$  und  $\rho_e = r_{xy} = 1$  lässt sich allerdings keine einfache Erklärung finden. Diese liegen zudem häufig ausserhalb des zulässigen Bereichs.

Die Abbildung 7 zeigt den Verlauf der Lösungen beider Modelle in Abhängigkeit von  $\rho_e \approx r_{xy}$  bei konstant gehaltenen Werten  $x_j = y_j = \bar{x} = \bar{y} = \frac{1}{2}, s_x = s_y = 0.1$ . Die konstanten Werte sind so

gewählt, dass die Regression keine unzulässigen Ergebnisse liefern kann. Die Resultate in den drei Punkten  $-1$ ,  $0$  und  $1$  sind identisch. Trotzdem sind dazwischen deutliche Abweichungen festzustellen. Bei  $\rho_e \approx r_{xy} \approx 0.8$  etwa findet man mit dem Regressionsmodell  $\hat{a}_j = 0.45$ , mit Thomsons Vorschlag  $\hat{a}_j = 0.375$ .



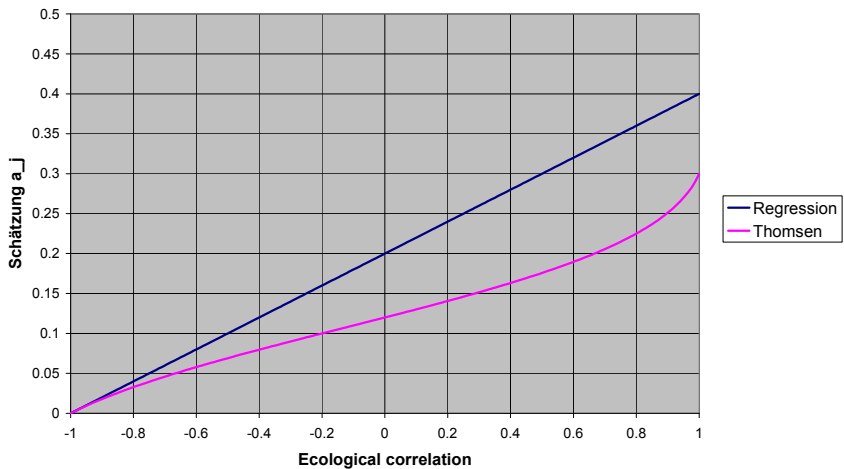
**Abbildung 7:** Numerischer Vergleich der Resultate aus einer Regression und Thomsons Methode mit den Werten  $x_j = y_j = \bar{x} = \bar{y} = \frac{1}{2}, s_x = s_y = 0.1$  (vgl. Haupttext).

Abbildung 8 zeigt dieselbe Betrachtung mit den konstanten Werten  $x_j = 0.4, y_j = 0.3, \bar{x} = \bar{y} = \frac{1}{2}, s_x = s_y = 0.1$ . Die Regression ergibt im Bereich  $r_{xy} > 0.5$  unmögliche Resultate von  $\hat{a}_j > 0.3$ , während Thomsons Ansatz den grössten zulässigen Wert  $\hat{a}_j = 0.3$  im Extremfall perfekter (positiver) Korrelation liefert.

Aus diesen Ausführungen wird klar, dass von Thomsons Modell zumindest in gewissen Fällen sinnvollere Resultate zu erwarten sind als von einer einfachen Regression. Von welcher Bedeutung diese Probleme in der Praxis tatsächlich sind, lässt sich jedoch aufgrund unserer begrenzten Betrachtungen nicht beurteilen. Abbildung 7 zeigt ausserdem, dass die Abweichungen selbst in Fällen, wo das Regressionsmodell sinnvolle Resultate ergibt, beträchtlich sein können.

Trotz des offensichtlichen Fehlers in der Herleitung von Thomsons Lösung führt die Anwendung seiner Methode zumindest in den betrachteten Fällen nicht zu grundsätzlich abwegigen Resultaten. Möglicherweise liesse sich eine angepasste Version des Modells finden, aus welcher die von Thomsen vorgeschlagene Lösung als eine sinnvolle und sauber begründbare Schätzung hervorgehe.





**Abbildung 8:** Numerischer Vergleich der Resultate aus einer Regression und Thomsens Methode mit den Werten  $x_j = 0.4$ ,  $y_j = 0.3$ ,  $\bar{x} = \bar{y} = \frac{1}{2}$ ,  $s_x = s_y = 0.1$  (vgl. Haupttext).

## 2.2.9 Weitere Modellansätze

- „Aggregated compound multinomial model“, Brown and Payne (1986):  
Dieses Modell wird in der Literatur zur ökologischen Inferenz häufig zitiert. Achen/Shively (1995, p. 62ff.) haben gezeigt, dass dieses Modell als Spezialfall eines Regressionsmodells mit zufälligen Übergangswahrscheinlichkeiten (Abschnitt 2.2.6) betrachtet werden kann.
- „Neighbourhood model“, Freedman 1999:

Modell:  $p_{AAj} = p_{BAj}$

Es wird angenommen, dass die Wähleranteile der ersten Wahl irrelevant für die Schätzung der Anteile der zweiten Wahl sind.

Lösung:  $p_{AAj} = p_{BAj} = y_j$

Die Annahmen des Neighbourhood Model sind in Wählerwanderungsproblemen kaum je erfüllt. Interessant an diesem Modell ist aber, dass es die Rolle der Annahmen illustriert, auf denen die verschiedenen Methoden beruhen. Gemäss der Regression lassen sich die Wahlwahrscheinlichkeiten eines Individuums in der zweiten Wahl am ehesten aus seiner Entscheidung in der ersten Wahl ableiten, der dominierende Effekt ist also derjenige des Individuums. Im "Neighbourhood Model" hingegen sind die aktuellen lokalen Rahmenbedingungen im betreffenden Wahlkreis ausschlaggebend, d.h. der räumliche Faktor ist entscheidend. Beide Argumentationen können je nach untersuchter Situation plausibel sein. In beiden Fällen handelt es sich nicht um eine allgemeingültige korrekte Rekonstruktion der Realität aus der partiellen Information der Randhäufigkeiten aller Wahlkreise, sondern um die Rekonstruktion unter gewissen Annahmen.

- Binomial-Beta Hierarchical Models, King et al. 1999:  
Hierbei handelt es sich um ein Modell mit zufälligen Übergangswahrscheinlichkeiten, wobei als Verteilung der Übergangswahrscheinlichkeiten Beta-Verteilungen angesetzt werden, deren Parameter ihrerseits wieder aus Exponentialverteilungen stammen. Die numerische Handhabung dieser Methodik ist schwierig und erfolgt mit MCMC-Methoden (Markov chain Monte Carlo). Dieser Ansatz ist relativ neu und dürfte in praktischen Anwendungen noch wenig zum Einsatz gekommen sein.

## 2.3 Diskussion

In Abschnitt 2 wurde der Spezialfall der ökologischen Inferenz behandelt, in welchem versucht wird, aus den bekannten Randhäufigkeiten von  $m$  (2x2)-Tabellen (Aggregatdaten) Schlüsse über die Zellhäufigkeiten innerhalb dieser Tabellen (Individualdaten) zu ziehen. Da zwischen Abhängigkeitsstrukturen *über die Wahlkreise hinweg* und solchen *innerhalb der Wahlkreise* nicht zwingend ein Zusammenhang bestehen muss, ist ökologische Inferenz nur ausgehend von bestimmten Annahmen möglich. In den verschiedenen Methoden werden die Wahrscheinlichkeiten der Entscheidungen der Individuen in der zweiten Wahl als Produkt gewisser Einflüsse aufgefasst:

- In der einfachen ökologischen Regression (Abschnitt 2.2.2) wird die Entscheidung in der ersten Wahl als einziger Einfluss betrachtet.
- In der erweiterten ökologischen Regression in Abschnitt 2.2.3 werden als weitere Einflüsse bekannte numerische Merkmale (Kovariablen) auf der Ebene der Wahlkreise berücksichtigt.
- In der ökologischen Regression mit zufälligen Übergangswahrscheinlichkeiten (Abschnitt 2.2.4) fließt keine zusätzliche Information ein, nur die Handhabung der zufälligen Variabilität ändert.
- Auch in Kings Modell (Abschnitt 2.2.5) fließt keine zusätzliche Information ein, doch die Schätzung erfolgt so, dass keine a priori unmöglichen Lösungen resultieren können.
- Thomsen (Abschnitt 2.2.7) geht von einem Modell aus, in dem die Wahl der Individuen als Funktion einer Reihe latenter Variablen erklärt wird, welche die politische Einstellung der Wählenden widerspiegelt. Dann zeigt er, wie unter den gegebenen Annahmen aus der Korrelation über Wahlkreise hinweg auf die Abhängigkeit innerhalb der Tabellen geschlossen werden kann, da die Abhängigkeitsstrukturen beider Ebenen das Produkt derselben latenten Variablen sind. Auf diese Weise ist die konkrete Schätzung der latenten Variablen nicht nötig.
- Beim „Neighbourhood Model“ schliesslich wird a priori angenommen, dass innerhalb der Tabellen keine Abhängigkeiten zwischen der ersten und zweiten Wahl bestehen.

Keines dieser Modelle ist grundsätzlich richtig oder falsch. In manchen Anwendungen wird ein Modell Ergebnisse näher an der Realität liefern, in anderen Fällen wird eine andere Methode besser abschneiden, je nachdem, welche Annahmen der tatsächlichen Struktur des Wahlverhaltens nahekommen.

Aus den Ausführungen in diesem Abschnitt geht hervor, dass alle besprochenen Methoden gewisse Schwächen aufweisen, und dass keine den in 1.1.2 gestellten Anforderungen gerecht wird.

Das Regressionsmodell und seine Weiterentwicklungen weisen zwei grosse Schwächen auf:

1. die Möglichkeit unmöglicher Resultate in der Form von Übergangswahrscheinlichkeiten kleiner als 0 oder grösser als 1,
2. ihre Empfindlichkeit auf Verletzungen der Voraussetzung fehlenden Aggregationsbias.

Die Überprüfbarkeit der Modellvoraussetzungen, insbesondere hinsichtlich eines möglichen Aggregationsbias, sind sehr limitiert. Liegen die geschätzten Übergangswahrscheinlichkeiten mehrheitlich und deutlich ausserhalb des Intervalls  $[0,1]$ , so kann dies als Indiz für nicht erfüllte Voraussetzungen gewertet werden. Der umgekehrte Schluss ist jedoch nicht zulässig: auch wenn alle Koeffizienten im Einheitsintervall liegen, besteht keine Garantie für die Gültigkeit der Annahmen.

Mit Kings EI-Methode wird nur eines der beiden angesprochenen Probleme der Regression behoben, nämlich dasjenige unzulässiger Resultate („Out of bounds-Problem“). Bezüglich der Anfälligkeit auf Aggregationsbias sowie der Überprüfbarkeit dieser Voraussetzung schneidet es jedoch kaum besser ab.

Thomsens Modellansatz schliesslich ist zwar in der Lage, Aggregationsbias zu modellieren, die Möglichkeit einer Überprüfung der Modellvoraussetzungen ist aber selbst bei Vorliegen der vollständigen Wanderungstabelle nicht gegeben. Der wesentlichste Schwachpunkt ist allerdings methodischer Natur. Die Herleitung von Thomsens Lösung basiert auf widersprüchlichen Argumentationen. Trotz dieses offensichtlichen Fehlers führt die Anwendung von Thomsens Methode nicht zu grundsätzlich abwegigen Resultaten. Möglicherweise liesse sich eine angepasste Version des Modells finden, aus welcher die von Thomsen vorgeschlagene Lösung als eine sinnvolle und sauber begründbare Schätzung hervorginge.

Der einfache Spezialfall  $2 \times 2$  genügt in der Praxis kaum je. Selbst in Zweiparteiensystemen existiert jeweils die dritte Möglichkeit der Nichtwahl. Dass dieser Situation hier trotzdem so viel Aufmerksamkeit geschenkt wird, liegt einerseits daran, dass sich ein grosser Teil der Fachliteratur damit befasst. Andererseits ist es wichtig, Erkenntnisse in diesem leicht erfassbaren Spezialfall zu gewinnen, da sie sich häufig auf allgemeinere Situationen übertragen lassen.

### 3 Verallgemeinerungen für mehr als zwei Wahlalternativen

Die Situation aus dem letzten Abschnitt soll nun auf den allgemeinen Fall von  $P$  Wahlalternativen in der ersten und  $Q$  in der zweiten Wahl verallgemeinert werden. Im Folgenden wird gelegentlich von „Parteien“ gesprochen, obwohl es sich bei gewissen dieser Wahlalternativen etwa um „Nichtwahl“ oder „keine Wahlberechtigung“ handeln kann.

#### 3.1 Ein Zahlenbeispiel: Nationalratswahlen 1995 und 1999 Kanton Zürich

Als Zahlenbeispiel soll in diesem Bericht wiederholt die Schätzung der Wählerwanderungen zwischen den Nationalratswahlen von 1995 und 1999 im Kanton Zürich betrachtet werden. Die Daten liegen auf Gemeindeebene für die 171 Gemeinden des Kantons vor (Quelle: BFS). Die Stimmenanteile der Parteien lauten (Parteien absteigend nach Parteistärke geordnet):

NRW ZH 1995			NRW ZH 1999		
Partei	Anteil		Partei	Anteil	
	Stimmenanteil	Wahlberechtigte		Stimmenanteil	Wahlberechtigte
SVP	25.46%	10.95%	SVP	32.48%	14.66%
SPS	23.07%	9.93%	SPS	25.63%	11.56%
FDP	18.10%	7.79%	FDP	17.82%	8.04%
GPS	6.52%	2.81%	CVP	5.12%	2.31%
LdU	5.31%	2.29%	GPS	4.14%	1.87%
CVP	4.87%	2.10%	EVP	3.44%	1.55%
EVP	3.74%	1.61%	Übrige	3.03%	1.37%
FPS	3.54%	1.52%	LdU	2.05%	0.93%
SD	3.29%	1.42%	FGA	1.79%	0.81%
FGA	2.72%	1.17%	EDU	1.76%	0.79%
EDU	1.86%	0.80%	SD	1.55%	0.70%
Übrige	0.87%	0.37%	FPS	0.82%	0.37%
LPS	0.47%	0.20%	CSP	0.22%	0.10%
CSP	0.16%	0.07%	LPS	0.15%	0.07%
Total	100.00%	43.03%	Total	100.00%	45.13%

Die Tatsache, dass die Wählenden mehr als eine Stimme abgeben, bedeutet keine grundsätzliche Änderung der Aufgabenstellung. Betrachtet man die Anteile der „fiktiven Wählenden“ einer Partei gemäss

$$\text{Fiktive Wählende} = \text{erhaltene Stimmen} \times \frac{\text{Total gültige Wahlzettel}}{\text{Total abgegebene Stimmen}},$$

so wird die Situation rechnerisch auf diejenige einer Wahl zurückgeführt, bei welcher die Wählenden nur eine Stimme abgeben<sup>6</sup>. Die Anteile bezüglich der Population der Wahlberechtigten werden mit der Formel

$$\text{Anteil bzgl. Wahlberechtigten} = \text{Anteil bzgl. Wählenden} \times \frac{\text{Total Wählende}}{\text{Total Wahlberechtigte}}$$

ermittelt.

Für unsere Berechnungen werden wir jeweils nur die fünf wählerstärksten Parteien SVP, SPS, FDP, CVP und GPS gesondert betrachten und die restlichen Parteien in der Kategorie „Übrige“ zusammenfassen. Die Resultate verlieren dadurch zwar an Informationsgehalt, doch dieser wird in Kauf genommen, da der primäre Zweck der Berechnungen die exemplarische Vorführung der vorgestellten Methoden ist, und diese kann mit einer kleineren Anzahl Parteien überblickbarer gestaltet werden. Auf eine Diskussion der Resultate wird verzichtet, der Vergleich der mit verschiedenen Methoden erhaltenen Ergebnisse wird dem Leser überlassen.

### 3.2 Deterministische Betrachtungen in einer (P x Q)-Tabelle

Wir verwenden die folgende Notation. Sei  $n_{1p}$  die Zahl der Wähler von Partei  $p$  in der ersten Wahl,  $n_{2q}$  sei die entsprechende Zahl in der zweiten Wahl. Für die (in der Regel unbekannte) Zahl der Wähler, die in der ersten Wahl  $p$  und in der zweiten  $q$  wählen, verwenden wir das Symbol  $n_{(p,q)}$ . Die vollständige Wanderungstabelle kann somit folgendermassen dargestellt werden:

Wahl 2	Wahl 1 Partei 1	Partei 2	...	Partei P	Total
Partei 1	$n_{(1,1)} = ?$	$n_{(2,1)} = ?$	...	$n_{(P,1)} = ?$	$n_{21}$
Partei 2	$n_{(1,2)} = ?$	$n_{(2,2)} = ?$	...	$n_{(P,2)} = ?$	$n_{22}$
...	...	...	...	...	...
Partei Q	$n_{(1,Q)} = ?$	$n_{(2,Q)} = ?$	...	$n_{(P,Q)} = ?$	$n_{2Q}$
Total	$n_{11}$	$n_{12}$	...	$n_{1P}$	$n$

Bezeichnet man mit  $p_{(p,q)}$  den Anteil der  $p$ -Wähler der ersten Wahl, die in der zweiten Wahl  $q$  wählen, so ergeben sich die linearen Bedingungen

$$n_{2q} = n_{11}p_{(1,q)} + n_{12}p_{(2,q)} + \dots + n_{1P}p_{(P,q)}, \quad q = 1, \dots, Q.$$

Offensichtlich muss ausserdem gelten:

$$p_{(p,1)} + p_{(p,2)} + \dots + p_{(p,Q)} = 1, \quad p = 1, \dots, P$$

Der Raum der möglichen Lösungen  $p_{(1,1)}, \dots, p_{(P,Q)}$  ist von der Dimension  $(P-1)(Q-1)$ . Es wird eingeschränkt durch die Randbedingungen  $0 \leq p_{(p,q)} \leq 1$ . Deterministische Grenzen können wie im Spezialfall  $P=Q=2$  berechnet werden, liefern aber kaum präzise Informationen, es sei denn eine Partei ist sehr dominant.

<sup>6</sup> Zu den fiktiven Wählenden vgl. Seitz (2002), S. 43.

### 3.3 Verallgemeinerung des ökologischen Regressionsansatzes

**Notation:** Für die Situation, in der für jeden von  $m$  Wahlkreise eine  $(P \times Q)$ -Tabelle vorliegt, wird die folgende Notation verwendet. Der Index  $j$  steht jeweils für die Nummer des betreffenden Wahlkreises. Die relativen Wähleranteile bezüglich der Gesamtpopulation des betreffenden Wahlkreises sind  $x_{pj}$  für Partei  $p$  in der ersten bzw.  $y_{qj}$  für Partei  $q$  in der zweiten Wahl. Die Übergangswahrscheinlichkeiten von Partei  $p$  zu Partei  $q$  werden als  $p_{(p,q)j}$  bezeichnet. Eine Übersicht über die Bezeichnungen gibt die untenstehende Tabelle.

Wahl 2	Wahl 1 Partei 1	Partei 2	...	Partei $P$	Total
Partei 1	$p_{(1,1)j}x_{1j}$	$p_{(2,1)j}x_{2j}$	...	$p_{(P,1)j}x_{Pj}$	$y_{1j}$
Partei 2	$p_{(1,2)j}x_{1j}$	$p_{(2,2)j}x_{2j}$	...	$p_{(P,2)j}x_{Pj}$	$y_{2j}$
...	...	...	...	...	...
Partei $Q$	$p_{(1,Q)j}x_{1j}$	$p_{(2,Q)j}x_{2j}$	...	$p_{(P,Q)j}x_{Pj}$	$y_{Qj}$
Total	$x_{1j}$	$x_{2j}$	...	$x_{Pj}$	1

Der Anteil der Personen, die in Wahl 2 wieder Partei  $q$  wählen an den Wählern von Partei  $p$  aus der ersten Wahl in der gesamten Population lautet

$$\bar{p}_{(p,q)} = \frac{1}{S_p} \sum_j p_{(p,q)j} x_{pj} n_j \quad \text{mit } S_p = \sum_{j=1}^m n_j x_{pj}.$$

Im ökologischen Regressionsmodell geht man davon aus, dass die  $p_{(p,q)j}$  für alle Wahlkreise identisch sind. Es ergeben sich aus der obigen Darstellung die  $Q$  Regressionsgleichungen ( $q = 1, \dots, Q$ )

$$Y_{qj} = p_{(1,q)}x_{1j} + p_{(2,q)}x_{2j} + \dots + p_{(P,q)}x_{Pj} + U_{qj}, \quad j = 1, \dots, m,$$

wobei die  $U_{qj}$  als unabhängig und identisch verteilte Zufallsvariablen mit Erwartungswert 0 betrachtet werden. Die geschätzten Regressionsparameter erfüllen die Bedingungen

$$\hat{p}_{(p,1)} + \hat{p}_{(p,2)} + \dots + \hat{p}_{(p,Q)} = 1, \quad p = 1, \dots, P.$$

Da die letzte Regressionsgleichung von den ersten  $Q-1$  abhängig ist, genügt es, die ersten  $Q-1$  Regressionen zu rechnen. Die geschätzten Koeffizienten der letzten Gleichung können dann als  $\hat{p}_{(p,Q)} = 1 - (\hat{p}_{(p,1)} + \dots + \hat{p}_{(p,Q-1)})$  ermittelt werden.

Sämtliche in 2.2.2 diskutierten Punkte lassen sich auf diesen allgemeineren Fall übertragen, und die dort besprochenen Probleme bleiben bestehen. Dies gilt insbesondere für die Probleme von Aggregationsbias und von Schätzungen der  $p_{(p,q)}$  kleiner als 0 oder grösser als 1. Zu Schwierigkeiten, die bei der Anwendung der Methode auftreten können, siehe das Beispiel in 3.3.1.

Verallgemeinerungen dieses Modells mit von Kovariablen anhängigen (vgl. 5.2) oder stochastischen  $p_{(p,q)}$  sind möglich. Die Diskussion entspricht weitgehend dem, was für den Spezialfall  $P = Q = 2$  gesagt worden ist.

### 3.3.1 Anwendung auf die Nationalratswahlen im Kanton Zürich 1995/99

Die untenstehende Tabelle enthält die als Regressionskoeffizienten einer ökologischen Regression mit den 171 Gemeinden des Kantons Zürich errechneten Übergangswahrscheinlichkeiten. Die Regression wurde dabei gewichtet mit dem Mittelwert der Anzahl Wahlberechtigter in den beiden Wahlen<sup>7</sup>.

Wahl 1999	Wahl 1995						
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler
CVP	88.42%	2.71%	-7.52%	-1.93%	0.63%	-1.76%	1.33%
FDP	-14.36%	98.92%	-16.46%	2.54%	2.81%	0.02%	0.88%
GPS	-5.80%	-1.29%	52.62%	4.59%	-2.07%	3.69%	0.06%
SPS	-23.10%	0.44%	10.96%	122.96%	-8.64%	10.71%	-0.86%
SVP	8.00%	13.85%	20.22%	-44.20%	93.04%	18.67%	9.01%
Übrige	8.77%	-2.27%	34.39%	8.48%	-2.60%	62.19%	-1.25%
Nichtwähler	38.07%	-12.36%	5.80%	7.56%	16.83%	6.48%	90.84%
	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Eine ganze Reihe der Koeffizienten liegt ausserhalb des Intervalls [0,1]. Dies deutet darauf hin, dass die Voraussetzung gleicher Übergangswahrscheinlichkeiten kaum erfüllt ist. Um trotzdem eine „mit Regression geschätzte“ sinnvolle Wanderungstabelle zu finden, können die folgenden pragmatischen Anpassungsschritte vorgenommen werden:

Werte über 100% und unter 0% sind unmöglich und werden deshalb durch die entsprechenden Grenzen ersetzt:

Wahl 1999	Wahl 1995						
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler
CVP	88.42%	2.71%	0.00%	0.00%	0.63%	0.00%	1.33%
FDP	0.00%	98.92%	0.00%	2.54%	2.81%	0.02%	0.88%
GPS	0.00%	0.00%	52.62%	4.59%	0.00%	3.69%	0.06%
SPS	0.00%	0.44%	10.96%	100.00%	0.00%	10.71%	0.00%
SVP	8.00%	13.85%	20.22%	0.00%	93.04%	18.67%	9.01%
Übrige	8.77%	0.00%	34.39%	8.48%	0.00%	62.19%	0.00%
Nichtwähler	38.07%	0.00%	5.80%	7.56%	16.83%	6.48%	90.84%
	143.26%	115.92%	123.99%	123.17%	113.31%	101.76%	102.12%

Die Spaltensummen entsprechen den Summen der Übergangswahrscheinlichkeiten der Wähler der jeweiligen Partei 1995 und sollten deshalb 100% ergeben. Dies ist hier nicht mehr der Fall, weshalb die Daten spaltenweise normiert werden, so dass die Spaltensummen wieder 100% ergeben. Aufgrund der resultierenden Übergangsmatrix und der Wähleranteile von 1995 wird dann eine provisorische Wanderungstabelle bestimmt:

<sup>7</sup> Die Anwendung des Regressionsmodells ohne Gewichtung führt zu etwas anderen Resultaten, an der grossen Anzahl Regressionskoeffizienten ausserhalb der Einheitsintervalls ändert sich nichts.

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	1.29%	0.18%	0.00%	0.00%	0.06%	0.00%	0.74%	2.28%
FDP	0.00%	6.65%	0.00%	0.20%	0.27%	0.00%	0.49%	7.62%
GPS	0.00%	0.00%	1.19%	0.37%	0.00%	0.34%	0.03%	1.94%
SPS	0.00%	0.03%	0.25%	8.06%	0.00%	0.99%	0.00%	9.33%
SVP	0.12%	0.93%	0.46%	0.00%	9.00%	1.73%	5.03%	17.27%
Übrige	0.13%	0.00%	0.78%	0.68%	0.00%	5.77%	0.00%	7.37%
Nichtwähler	0.56%	0.00%	0.13%	0.61%	1.63%	0.60%	50.68%	54.21%
<b>Total 1995</b>	<b>2.10%</b>	<b>7.79%</b>	<b>2.81%</b>	<b>9.93%</b>	<b>10.96%</b>	<b>9.45%</b>	<b>56.97%</b>	<b>100.00%</b>

In dieser Wanderungstabelle stimmen nun die Randtotal der Wahl 1995 mit den exakten Zahlen überein, nicht aber diejenigen von 1999. Letztere lauten in Wahrheit:

#### Exakte Anteile 1999:

CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler
2.31%	8.04%	1.87%	11.56%	14.66%	6.68%	54.88%

Für die Anpassung der Tabellenwerte an die vorgegebenen Randtotale werden zwei Möglichkeiten betrachtet:

Um die Randsummen der obigen Wanderungstabelle an die wahren Anteile anzupassen, wird ein iteratives Verfahren angewandt. Dieses besteht darin, die Zelleneinträge abwechselnd zeilen- und spaltenweise durch Multiplikation mit einem geeigneten Faktor an das jeweilige Randtotal anzupassen, bis die Abweichung der Zeilen- und Spaltensummen nahe genug an den vorgegebenen Werten liegen. Wird beispielsweise gefordert, dass die relative Abweichung der Summen von ihrem Sollwert höchstens 0.1% beträgt, so benötigt der Algorithmus 38 Iterationsschritte. Es resultiert die folgende Tabelle, welche nur noch geringe Abweichungen von den tatsächlichen Zahlen in den Zeilensummen besitzt:

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	1.31%	0.16%	0.00%	0.00%	0.07%	0.00%	0.77%	2.31%
FDP	0.00%	6.94%	0.00%	0.11%	0.40%	0.00%	0.60%	8.04%
GPS	0.00%	0.00%	1.23%	0.22%	0.00%	0.37%	0.04%	1.87%
SPS	0.00%	0.06%	0.49%	8.98%	0.00%	2.02%	0.00%	11.55%
SVP	0.09%	0.63%	0.28%	0.00%	8.53%	1.10%	4.03%	14.66%
Übrige	0.15%	0.00%	0.71%	0.35%	0.00%	5.47%	0.00%	6.68%
Nichtwähler	0.55%	0.00%	0.10%	0.27%	1.96%	0.48%	51.53%	54.89%
<b>Total 1995</b>	<b>2.10%</b>	<b>7.79%</b>	<b>2.81%</b>	<b>9.93%</b>	<b>10.96%</b>	<b>9.45%</b>	<b>56.97%</b>	<b>100.00%</b>

Der beschriebene Algorithmus wird manchmal als *iterative proportional fitting* beschrieben und wurde ursprünglich von Deming (1943) vorgeschlagen als (zu jener Zeit) leichter zu berechnende Variante des unten beschriebenen Kleinstquadrat-Ansatzes. Die bei Konvergenz gefundene Lösung mit den angepassten Zellhäufigkeiten  $\pi_{pq}$  minimiert die Summe  $\sum_p \sum_q \pi_{pq} \cdot (\log(\pi_{pq} / p_{pq}) - 1)$  unter den gegebenen Nebenbedingungen, wo  $p_{pq}$  die Zelleneinträge in der Wanderungstabelle vor der Anpassung bezeichnet (Deville et al. 1993).



Andere Ansätze für die Anpassung werden ebenfalls von Deville et al. (1993) diskutiert. Ein anschaulicherer Ansatz besteht darin, die Summe der quadrierten Abweichungen

$$\sum_p \sum_q \left( \frac{\pi_{pq} - p_{pq}}{\sqrt{p_{pq}}} \right)^2$$

zu minimieren (Kleinstquadrate-Kriterium). Bei dessen Anwendung ergeben sich im vorliegenden Fall zwei Probleme. Erstens sind gewisse  $p_{pq}$  im Nenner gleich 0. Diesem Problem kann begegnet werden, indem die angepassten Werte in den entsprechenden Zellen unverändert belassen werden. Zweitens werden gewisse  $\pi_{pq}$  negativ, weitere Anpassungsschritte sind somit notwendig.

Wir werden im Folgenden bei Anwendungen der Regressionsmethode von der zuerst vorgestellten iterativen Methode Gebrauch machen.

### Fazit zu 3.3:

Bezüglich der Kriterien Modell, Überprüfbarkeit und Schätzung gilt, was am Ende von Abschnitt 2.2.2 für den Zweiparteienfall gesagt wurde. Die Berechnungen mit konkreten Wahldaten haben gezeigt, dass die Anwendung der ökologischen Regression in der Praxis nicht immer so elegant und reibungslos erfolgt, wie es bei der Modellformulierung den Anschein macht. Falls nicht alle aus der Regression resultierenden Übergangswahrscheinlichkeiten im Einheitsintervall liegen, so sind weitere Anpassungsschritte erforderlich, um sinnvoll interpretierbare Resultate zu erhalten, die gleichzeitig in Übereinstimmung mit den bekannten Randhäufigkeiten stehen.

## 3.4 Allgemeiner Ansatz von Thomsen

Die formelle Herleitung der Verallgemeinerung von Thomsens Methode (vgl. 2.2.7) basiert auf einem multinomialen Logit-Modell, einer Verallgemeinerung des Logit-Modells im Zweiparteienfall. Um die Notation nicht unnötig zu belasten, wird im folgenden davon ausgegangen, dass in beiden Wahlen die gleiche Anzahl Wahlalternativen betrachtet wird, d.h. es ist  $P = Q$ .

Die politische Ausrichtung der Person  $i$  in Wahlkreis  $j$  kann als latente Variable  $Z_{ji}$  in  $k$  Dimensionen ausgedrückt werden, wobei  $Z_{ji}$  ein normalverteilter Zufallsvektor mit Verteilung

$$Z_{ji} \sim N(\mu_j, \text{Diag}(\sigma^2)).$$

ist. Die  $\mu_j$  sind ihrerseits normalverteilt gemäss  $\mu_j \sim N(0, \text{Diag}(1))$ . Die Wahrscheinlichkeit, dass Person  $i$  in Wahlkreis  $j$  in Wahl  $t$  für Partei  $p$  stimmt, laute

$$P(W_{tpji} = 1) = \frac{\exp(\alpha_{tp} + \beta'_{tp} Z_{ji})}{\sum_{q=1}^P \exp(\alpha_{tq} + \beta'_{tq} Z_{ji})},$$

wo  $\alpha_{tp}$  ( $t=1,2; p=1,\dots,P$ ) reelle Zahlen und  $\beta_{tp}$  ( $t=1,2; p=1,\dots,P$ ) reelle Vektoren der Dimensionen  $k$  sind. Ohne Einschränkung der Allgemeinheit kann  $\alpha_{tp} = \beta_{tp} = 0$  gesetzt werden. Es gilt dann:

$$P(W_{tpji} = 1) = \frac{\exp(\alpha_{tp} + \beta'_{tp} Z_{ji})}{1 + \sum_{q=1}^{P-1} \exp(\alpha_{tq} + \beta'_{tq} Z_{ji})} \quad p = 1, \dots, P-1,$$

und für die Referenzpartei (Partei  $P$ )

$$P(W_{tpji} = 1) = \frac{1}{1 + \sum_{q=1}^{P-1} \exp(\alpha_{tq} + \beta'_{tq} Z_{ji})}.$$

Welche Partei als Referenzpartei  $P$  gewählt wird, ist für die Wahrscheinlichkeiten  $P(W_{tpji} = 1)$  irrelevant<sup>8</sup>.

Thomsen führt die Lösung des Problems näherungsweise auf den einfachen Fall zweier Wahlalternativen zurück. Die Betrachtungen aus Abschnitt 2.2.7 können für jede (2x2)-Untertabelle der Form

Wahl 2	Wahl 1		
	Partei $p_1$	Partei $p_2$	
Partei $p_3$	$a$	$y_{p_3j} - a$	$y_{p_3j}$
Partei $p_4$	$x_{p_1j} - a$	$1 - x_{p_2j} - y_{p_4j} + a$	$1 - y_{p_4j}$
	$x_{p_1j}$	$1 - x_{p_2j}$	1

vorgenommen werden, wobei die zusätzliche Schwierigkeit auftritt, dass die Randhäufigkeiten dieser (2x2)-Untertabellen unbekannt sind. Thomsen (1987) stellt ein iteratives Verfahren zur Lösung dieses Problems vor. Die resultierenden Lösungen weisen die Eigenschaft auf, dass die ökologische Korrelation  $\rho_e$  (gerechnet mit dem Logit der gefitteten Randhäufigkeiten) in einer Auswahl von (2x2)-Untertabellen der Approximation des tetrachorischen Korrelationskoeffizienten durch Yule's  $Q$  entsprechen. Dabei wird wieder der problematische Grenzübergang  $\sigma \rightarrow \infty$  vollzogen; der in 2.2.7 aufgezeigte Widerspruch in Thomsens Lösungsvorschlag bleibt erhalten. Für beide Wahlen muss eine Referenzpartei gewählt werden, und das Resultat hängt nicht unwesentlich von der Wahl dieser Referenzpartei ab, obwohl dies auf der Ebene des theoretischen Modells nicht der Fall ist. Thomsen empfiehlt die „neutrale“ Alternative der Nichtwähler als Referenz. Dieser Mangel scheint auch für den Autoren selbst nicht zufriedenstellend: „Currently, work is in progress to design a method for ecological inference that does not require the choice of a reference party“ (Thomsen 2000).

<sup>8</sup> Die Verteilung, welche der Vektor  $\pi_t = (\pi_{t1}, \dots, \pi_{tp})$  mit  $\pi_{tp} = P(W_{tpji} = 1)$  hier aufweist, ist in der Analyse von Anteilsdaten (compositional data, z.B. in der Geologie) gebräuchlich und wird als Logistic-normal distribution oder Aitchison-Verteilung bezeichnet, vgl. z.B. Aitchison 1986.

### 3.4.1 Anwendung auf die Nationalratswahlen im Kanton Zürich 1995/99

Mit dem (frei verfügbaren) Programm ECOL, Version 3.0, heruntergeladen von Thomsens Internetseite unter <http://www.ps.au.dk/srt/Ecology.htm>, findet man die folgende Wanderungstabelle. Als „Referenzpartei“ wurden dabei gemäss Thomsens Empfehlung die Nichtwähler gewählt.

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	0.69%	0.05%	0.02%	0.08%	0.01%	0.11%	1.35%	2.30%
FDP	0.04%	6.69%	0.09%	0.05%	0.05%	0.14%	0.95%	8.00%
GPS	0.02%	0.02%	0.54%	0.11%	0.02%	0.25%	0.89%	1.86%
SPS	0.14%	0.07%	0.22%	7.65%	0.01%	0.85%	2.59%	11.52%
SVP	0.03%	0.19%	0.32%	0.02%	10.07%	0.40%	3.57%	14.60%
Übrige	0.07%	0.05%	0.26%	0.32%	0.03%	4.03%	1.90%	6.66%
Nichtwähler	1.10%	0.68%	1.34%	1.63%	0.71%	3.61%	45.98%	55.06%
<b>Total 1995</b>	<b>2.08%</b>	<b>7.74%</b>	<b>2.79%</b>	<b>9.87%</b>	<b>10.89%</b>	<b>9.40%</b>	<b>57.22%</b>	<b>100.00%</b>

Die entsprechenden Wanderungstabellen mit anderen Referenzpartei sind unten angeführt:

- Referenzpartei CVP:

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	0.57%	0.15%	0.02%	0.14%	0.08%	0.10%	1.24%	2.30%
FDP	0.12%	4.47%	0.19%	0.30%	0.98%	0.43%	1.52%	8.01%
GPS	0.03%	0.07%	0.69%	0.23%	0.17%	0.23%	0.44%	1.86%
SPS	0.31%	0.45%	0.31%	3.78%	0.37%	1.22%	5.08%	11.52%
SVP	0.11%	1.30%	0.62%	0.24%	3.32%	1.48%	7.53%	14.60%
Übrige	0.11%	0.28%	0.42%	0.98%	0.75%	1.64%	2.46%	6.64%
Nichtwähler	0.84%	1.02%	0.54%	4.20%	5.23%	4.29%	38.95%	55.07%
<b>Total 1995</b>	<b>2.09%</b>	<b>7.74%</b>	<b>2.79%</b>	<b>9.87%</b>	<b>10.90%</b>	<b>9.39%</b>	<b>57.22%</b>	<b>100.00%</b>

- Referenzpartei FDP:

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	1.09%	0.19%	0.01%	0.06%	0.00%	0.05%	0.90%	2.30%
FDP	0.22%	5.62%	0.43%	0.18%	0.22%	0.34%	1.01%	8.02%
GPS	0.01%	0.17%	0.94%	0.13%	0.02%	0.28%	0.31%	1.86%
SPS	0.12%	0.22%	0.24%	7.23%	0.01%	0.98%	2.72%	11.52%
SVP	0.02%	0.80%	0.42%	0.02%	10.21%	0.37%	2.77%	14.61%
Übrige	0.04%	0.18%	0.35%	0.39%	0.03%	4.48%	1.19%	6.66%
Nichtwähler	0.58%	0.56%	0.40%	1.87%	0.41%	2.90%	48.33%	55.05%
<b>Total 1995</b>	<b>2.08%</b>	<b>7.74%</b>	<b>2.79%</b>	<b>9.88%</b>	<b>10.90%</b>	<b>9.40%</b>	<b>57.23%</b>	<b>100.02%</b>

- Referenzpartei GPS:

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	0.86%	0.12%	0.02%	0.15%	0.02%	0.12%	1.01%	2.30%
FDP	0.07%	4.57%	0.25%	0.25%	0.79%	0.37%	1.70%	8.00%
GPS	0.02%	0.10%	0.40%	0.22%	0.15%	0.20%	0.78%	1.87%
SPS	0.19%	0.33%	0.24%	4.88%	0.08%	1.16%	4.63%	11.51%
SVP	0.04%	0.96%	0.65%	0.08%	6.70%	0.68%	5.49%	14.60%
Übrige	0.08%	0.29%	0.26%	0.68%	0.30%	1.95%	3.09%	6.65%
Nichtwähler	0.81%	1.37%	0.98%	3.60%	2.87%	4.91%	40.51%	55.05%
Total 1995	2.07%	7.74%	2.80%	9.86%	10.91%	9.39%	57.21%	99.98%

- Referenzpartei SPS:

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	1.09%	0.04%	0.01%	0.33%	0.00%	0.08%	0.73%	2.28%
FDP	0.04%	6.93%	0.12%	0.22%	0.12%	0.23%	0.35%	8.01%
GPS	0.01%	0.03%	0.99%	0.36%	0.03%	0.26%	0.17%	1.85%
SPS	0.44%	0.20%	0.49%	6.58%	0.07%	1.18%	2.55%	11.51%
SVP	0.02%	0.34%	0.69%	0.26%	10.17%	0.84%	2.28%	14.60%
Übrige	0.05%	0.06%	0.31%	0.56%	0.06%	4.78%	0.83%	6.65%
Nichtwähler	0.42%	0.14%	0.18%	1.56%	0.44%	2.02%	50.30%	55.06%
Total 1995	2.07%	7.74%	2.79%	9.87%	10.89%	9.39%	57.21%	99.96%

- Referenzpartei SVP:

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	1.23%	0.06%	0.02%	0.09%	0.17%	0.12%	0.60%	2.29%
FDP	0.05%	6.89%	0.13%	0.06%	0.36%	0.18%	0.34%	8.01%
GPS	0.01%	0.04%	0.88%	0.13%	0.29%	0.24%	0.26%	1.85%
SPS	0.10%	0.07%	0.27%	8.30%	0.15%	1.08%	1.56%	11.53%
SVP	0.32%	0.50%	0.96%	0.14%	8.82%	0.94%	2.92%	14.60%
Übrige	0.05%	0.06%	0.34%	0.34%	0.26%	4.72%	0.89%	6.66%
Nichtwähler	0.31%	0.12%	0.19%	0.81%	0.84%	2.11%	50.65%	55.03%
Total 1995	2.07%	7.74%	2.79%	9.87%	10.89%	9.39%	57.22%	99.97%

- Referenzpartei Übrige:

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	1.00%	0.05%	0.01%	0.09%	0.01%	0.20%	0.94%	2.30%
FDP	0.05%	6.61%	0.16%	0.09%	0.22%	0.32%	0.55%	8.00%
GPS	0.02%	0.04%	0.90%	0.18%	0.06%	0.39%	0.27%	1.86%
SPS	0.18%	0.11%	0.19%	7.45%	0.02%	0.91%	2.67%	11.53%
SVP	0.03%	0.45%	0.70%	0.03%	9.45%	0.97%	2.97%	14.60%
Übrige	0.19%	0.17%	0.51%	0.57%	0.19%	3.28%	1.75%	6.66%
Nichtwähler	0.62%	0.31%	0.31%	1.46%	0.95%	3.33%	48.07%	55.05%
Total 1995	2.09%	7.74%	2.78%	9.87%	10.90%	9.40%	57.22%	100.00%

Ein Vergleich der Tabellen zeigt, wie die Wahl der Referenzpartei sich auf die Schätzung auswirkt.

### Fazit zu 3.4:

Was hinsichtlich der Kriterien Modell, Überprüfbarkeit und Schätzung am Ende von Abschnitt 2.2.7 für den Zweiparteienfall gesagt wurde, gilt weiterhin. Die Verallgemeinerung auf der Ebene des theoretischen Modells ist unproblematisch. Die Tatsache, dass die Anwendung der Methode bei unterschiedlichen Referenzparteien zu unterschiedlichen Resultaten führt, kontrastiert mit einer theoretischen Eigenschaft des Modells, gemäss welcher die Wahl der Referenzpartei keine Auswirkungen auf die Wählerströme hat.

## 3.5 Weitere Modellvorschläge

### 3.5.1 Verallgemeinerung von Kings Modell

King (1997) stellt nur einige allgemeine Überlegungen zum Fall mit mehreren Wahlalternativen in beiden Wahlen an (chapter 15: „I do not yet have an extensive experience with these larger tables. As a result, parts of this chapter are more conjectural“). Laut seinen Ausführungen ist eine Verallgemeinerung seines Ansatzes problemlos möglich. Im Gegensatz zu den Problemen der Dimensionen  $2 \times 2$  und  $2 \times 3$  bietet King allerdings keine Software an.

### 3.5.2 Andere Modelle

Alle drei in Abschnitt 2.2.9 erwähnten Modellansätze lassen sich auf den Mehrparteienfall erweitern. Die Verallgemeinerung der Binomial-Beta Hierarchical Models von King et al. (1999) für beliebige  $(P \times Q)$ -Tabellen wird von Rosen et al. (2000) eingeführt.

## 3.6 Modifizierte Probit- und Logit-Modelle

Da Thomsens multinomialer Logit-Modellansatz auf theoretischer Ebene interessante Eigenschaften aufweist (Unabhängigkeit von Referenz, Berücksichtigung von Aggregationsbias, Symmetrie in den Variablen  $X$  und  $Y$ ), aber der vorgeschlagene Lösungsweg unbefriedigend ist, sollen in diesem Abschnitt Alternativen für die Formulierung und Schätzung des Modells gesucht werden.

Auf theoretischer Ebene besteht die zentrale Anpassung darin, dass der von Thomsen vollzogene problematische Grenzübergang  $\sigma \rightarrow \infty$  nicht vollzogen wird. Der vorgeschlagene Lösungsweg ist grundsätzlich anders als bei Thomsen. Es erfolgt eine Schätzung des gesamten Modells.

### 3.6.1 Modifiziertes Probit-Modell für 2 Parteien

Wir gehen von einer bivariaten latenten Variablen  $Z_{ji}$  aus, welche die politische Ausrichtung von Person  $i$  in Wahlkreis  $j$  widerspiegelt. Deren Verteilung laute

$$Z_{ji} \sim N\left(\begin{pmatrix} \mu_{j1} \\ \mu_{j2} \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right),$$

wobei die  $\mu_{j1}$  und  $\mu_{j2}$  jetzt als fixe Parameter betrachtet werden.

$$\tau^{(d)} = \sqrt{\frac{1}{m} \sum_{j=1}^m (\mu_j^{(d)} - \bar{\mu}^{(d)})^2} \quad (d=1,2)$$

ist die Standardabweichung, welche die Heterogenität der Wahlkreise bezüglich ihrer mittleren politischen Ausrichtung in Dimension  $d$  wiedergibt. Die latenten Variablen werden so skaliert, dass  $\bar{\mu}^{(d)} = 0$  und  $\tau^{(d)} = 1$  ( $d = 1, 2$ ).  $\sigma$  entspricht dann der Heterogenität (bezüglich politischer Ausrichtung) der Wähler innerhalb der Wahlkreise im Verhältnis zur Heterogenität der Wahlkreise. Der Parameter  $\sigma$  muss im hier vorgestellten Ansatz a priori vorgegeben werden. Hier liegt der wesentlichste Unterschied zur Modellformulierung bei Thomsen.

Die Wahrscheinlichkeit, dass Person  $i$  in Wahlkreis  $j$  in Wahl  $t$  ( $t = 1, 2$ ) für Partei 1 stimmt, lautet

$$\Phi(\alpha_i + \beta'_i Z_{ji}).$$

Wegen der einfacheren mathematischen Handhabung verwenden wir im Zweiparteienfall wieder die Probit- anstelle der inversen Logit-Funktion. Die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person  $i$  in Wahlkreis  $j$  in Wahl  $t$  für Partei 1 stimmt, ist

$$\Phi\left(\frac{\alpha_i + \beta'_i \mu_j}{\sqrt{1 + \sigma^2 \beta'_i \beta_i}}\right).$$

Geht man nun davon aus, dass die Zahl der Wähler in Wahlkreis  $j$  gross genug ist, damit die Varianz des Stimmenanteils  $\bar{W}_{ij}$  von Partei 1 vernachlässigbar wird, so gilt

$$\bar{W}_{ij} \approx \Phi\left(\frac{\alpha_i + \beta'_i \mu_j}{\sqrt{1 + \sigma^2 \beta'_i \beta_i}}\right).$$

Aus diesen Überlegungen ergibt sich ein Modellansatz faktorenanalytischer Gestalt:

$$\Phi^{-1}(\bar{W}_{ij}) \approx \tilde{\alpha}_i + \tilde{\beta}'_i \mu_j \quad t = 1, 2.$$

$$\text{mit } \tilde{\alpha}_i = \frac{\alpha_i}{\sqrt{1 + \sigma^2 \beta'_i \beta_i}} \text{ und } \tilde{\beta}_i = \frac{\beta_i}{\sqrt{1 + \sigma^2 \beta'_i \beta_i}}.$$

$\tilde{\alpha}_i$  kann als Mittelwert der  $\Phi^{-1}(\bar{W}_{ij})$  ermittelt werden, die Schätzung von  $\tilde{\beta}_i$  und  $\mu_j$  erfolgt mit einer Hauptkomponentenanalyse:  $\tilde{\beta}_i^{(1)}$  ist der erste Eigenvektor der Kovarianzmatrix von  $(\Phi^{-1}(X), \Phi^{-1}(Y))$  multipliziert mit der Wurzel des zugehörigen Eigenwerts, die Bestimmung von  $\tilde{\beta}_i^{(2)}$  erfolgt analog. Der Anteil der Stammwähler von Partei 1 in Wahlkreis  $j$  lautet

$$a_j = \Phi_2(\tilde{\alpha}_1 + \tilde{\beta}'_1 \mu_j, \tilde{\alpha}_2 + \tilde{\beta}'_2 \mu_j, \rho_w) \text{ mit } \rho_w = \sigma^2 \tilde{\beta}'_1 \tilde{\beta}_2.$$

Zwei wichtige Eigenschaften dieses Modells sollen nun festgehalten werden:

**E1: Eindeutigkeit der Lösung.** Die geschätzten Wanderungstabellen sind durch

$$a_j = \Phi_2(\tilde{\alpha}_1 + \tilde{\beta}'_1 \mu_j, \tilde{\alpha}_2 + \tilde{\beta}'_2 \mu_j, \rho_w) \text{ und } \rho_w = \sigma^2 \tilde{\beta}'_1 \tilde{\beta}_2$$

eindeutig festgelegt. Zwar sind bei der Bestimmung der Koeffizienten  $\tilde{\beta}_i$  der Hauptkomponentenanalyse verschiedene Parametrisierungen möglich (Richtung der Eigenvektoren), welche aber zu identischen Lösungen für die Anteile  $a_j$  der Stammwähler führen.

**E2: Obere Schranke für  $\sigma$ .** Aus der Tatsache, dass  $\beta'_i \beta_i = \frac{\tilde{\beta}'_i \tilde{\beta}_i}{1 - \sigma^2 \tilde{\beta}'_i \tilde{\beta}_i}$  positiv sein muss, der

Darstellung  $\Phi^{-1}(\bar{W}_{ij}) \approx \tilde{\alpha}_i + \tilde{\beta}'_i \mu_j$  sowie der Gleichung  $\tau^{(d)} = 1$  ergibt sich eine obere Schranke für  $\sigma$ :

$$\sigma^2 < (\tilde{\beta}'_i \tilde{\beta}_i)^{-1} = \frac{1}{v_i^2} \quad \text{mit } v_i^2 = \frac{1}{m} \sum_{j=1}^m (\Phi^{-1}(\bar{W}_{ij}) - \bar{\Phi}_i)^2 \text{ und } \bar{\Phi}_i = \frac{1}{m} \sum_{j=1}^m \Phi^{-1}(\bar{W}_{ij}).$$

Der maximale Wert, den  $\sigma$  annehmen kann, lässt sich in diesem Modell somit direkt aus den Wähleranteilen  $\bar{W}_{ij}$  berechnen, nämlich als Kehrwert der empirischen Standardabweichung ihrer Probit-Transformation  $\Phi^{-1}(\bar{W}_{ij})$ .

Mit dem skizzierten Verfahren lassen sich die Modellparameter wie gesagt eindeutig schätzen, es können aber keine Standard Errors ermittelt werden. Es handelt sich somit um eine deterministische Lösung des Modells (welches aber stochastische Elemente enthält).

Der Spezialfall  $\sigma = 0$  entspricht dem in 2.2.9 erwähnten „Neighbourhood model“. Verschiedene Werte für  $\sigma$  entsprechen unterschiedlichen Annahmen bezüglich der Homogenität der Wahlkreise (bezüglich der politischen Einstellung). Mit wachsendem  $\sigma$  wird  $\rho_w$  grösser, was zur Folge hat, dass die Anteile der Stammwähler zu- und diejenigen der Wechselwähler abnehmen. Die Tatsache, dass sich die obere Schranke für  $\sigma$  als negative Funktion der empirischen Varianz  $v_i^2$  der Probit-transformierten Wähleranteile ermitteln lässt, widerspiegelt den folgenden Zusammenhang: Ist  $\sigma$  gross, so liegt eine grosse Heterogenität der Wähler innerhalb der Wahlkreise relativ zu denjenigen zwischen den Wahlkreisen vor. Dies entspricht geringer Heterogenität zwischen den Wahlkreisen relativ zu denjenigen innerhalb der Wahlkreise, und damit geringen Unterschieden im Wahlverhalten. Bei grosser Streuung zwischen den Wahlergebnisse der verschiedenen Wahlkreise kann folglich die Heterogenität der Wähler innerhalb der Wahlkreise nicht allzu ausgeprägt sein, und genau dies wird in der Ungleichung  $\sigma^2 < (\tilde{\beta}'_i \tilde{\beta}_i)^{-1} = \frac{1}{v_i^2}$  ausgedrückt.

Eine Gewichtung nach Anzahl Wahlberechtigter bei der Bestimmung der Modellparameter ist denkbar, aber nicht zwingend notwendig (vgl. dazu die entsprechende Diskussion der Gewichtungproblematik bei der Regression in Abschnitt 2.2.2, Punkt 6).

### 3.6.2 Modifiziertes multinomiales Logit-Modell für $P \geq 2$ Parteien

In diesem Abschnitt wird eine Verallgemeinerung des Modells aus 3.6.1 für den Mehrparteienfall präsentiert. Wie bei Thomsen wird die multinomiale Logit-Transformation verwendet. Beim vorgestellten Verfahren handelt es sich wie in 3.6.1 um eine deterministische Lösung eines stochastischen Modells mit a priori zu wählendem Parameter  $\sigma$ . Wir setzen  $P = Q$ , d.h. es wird angenommen, dass die Anzahl der Parteien in beiden Wahlen identisch ist.

Wir gehen weiterhin davon aus, dass die politische Ausrichtung der Person  $i$  in Wahlkreis  $j$  als latente Variable  $Z_{ji}$  ausgedrückt werden kann.  $Z_{ji}$  sei dabei ein normalverteilter Zufallsvektor in  $k = 2(P-1)$  Dimensionen mit Verteilung

$$Z_{ji} \sim N_k(\mu_j, \text{Diag}(\sigma^2)).$$

Die  $\mu_j^{(d)}$  sind dabei fixe Parameter, welche als mittlere Ausprägung der Einstellung der Wahlberechtigten in Wahlkreis  $j$  in Dimension  $d$  zu verstehen sind.  $\tau^{(d)} = \sqrt{\frac{1}{m} \sum_{j=1}^m (\mu_j^{(d)} - \bar{\mu}^{(d)})^2}$  ist dann wieder die Standardabweichung zwischen Wahlkreisen, welche wir  $=1$  setzen, und ebenso wird  $\bar{\mu}^{(d)} = 0$  angenommen.  $\sigma$  entspricht also der Heterogenität (bezüglich politischer Ausrichtung) der Wähler innerhalb der Wahlkreise im Verhältnis zur Heterogenität der Wahlkreise.

Die Wahrscheinlichkeit, dass Person  $i$  in Wahlkreis  $j$  in Wahl  $t$  für Partei  $p$  stimmt, laute

$$P(W_{tpji} = 1) = \frac{\exp(\alpha_p + \beta'_p Z_{ji})}{\sum_{q=1}^P \exp(\alpha_q + \beta'_q Z_{ji})},$$

wo  $\alpha_p$  ( $t=1,2; p=1,\dots,P$ ) reelle Zahlen und  $\beta_p$  ( $t=1,2; p=1,\dots,P$ ) reelle Vektoren der Dimensionen  $k$  sind. Ohne Einschränkung der Allgemeinheit kann  $\alpha_{ip} = \beta_{ip} = 0$  gesetzt werden, es gilt dann

$$P(W_{tpji} = 1) = \frac{\exp(\alpha_p + \beta'_p Z_{ji})}{1 + \sum_{q=1}^{P-1} \exp(\alpha_q + \beta'_q Z_{ji})}.$$

Geht man wieder davon aus, dass die Zahl der Wählenden in Wahlkreis  $j$  gross genug ist, damit die Varianz des Stimmenanteils  $\bar{W}_{tpj}$  von Partei  $p$  vernachlässigbar wird, so lautet der Wähleranteil in Wahlkreis  $j$  gemäss den Modellvoraussetzungen

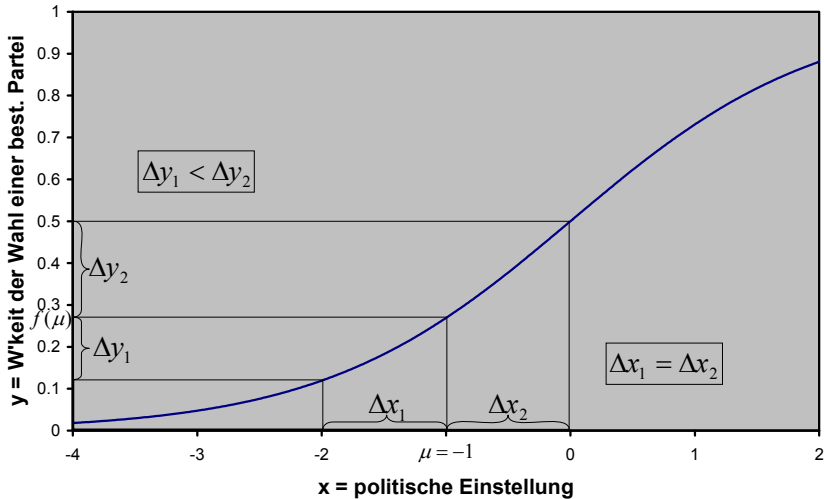
$$\bar{W}_{tpj} \approx E \left( \frac{\exp(\alpha_p + \beta'_p Z_{ji})}{\sum_{q=1}^P \exp(\alpha_q + \beta'_q Z_{ji})} \right) =: P_{tpj}.$$

Dieser Erwartungswert kann nicht als geschlossene Formel in den auftretenden Parametern dargestellt werden (Aitchison 1986, p. 116). Bei bekannten Parameterwerten kann er aber durch Simulation näherungsweise bestimmt werden. Falls  $\sigma > 0$ , gilt zudem die Ungleichung

$$P_{tpj} \neq \frac{\exp(\alpha_p + \beta'_p \mu_{ji})}{\sum_{q=1}^P \exp(\alpha_q + \beta'_q \mu_{ji})},$$



d.h. die Wähleranteile in einem Wahlkreis  $j$  sind nicht gleich den Wahlwahrscheinlichkeiten eines Wählers, dessen Einstellung dem Durchschnitt in diesem Wahlkreis entspricht. Wegen der Nicht-linearität der inversen Logit-Funktion werden die Wähleranteile nach Wahlkreis in die Richtung des Zentrums der Skala tendieren, und diese Tendenz ist um so ausgeprägter, je grösser  $\sigma$  ist. Eine Illustration dieser Tatsache gibt Abbildung 9.



**Abbildung 9:** Der Wähleranteil in einem Wahlkreis ist nicht gleich der entsprechenden Wahlwahrscheinlichkeit eines Wählers, dessen Einstellung dem Durchschnitt in diesem Wahlkreis entspricht. Die gleiche Abweichung nach unten und oben auf der Einstellungsskala ( $x$ ) führt zu unterschiedlichen Abweichungen auf der Skala der Wahlwahrscheinlichkeit ( $y$ ). Bei der Mittelwertbildung auf der  $y$ -Skala wird der Mittelwert in Richtung der Mitte ( $y = \frac{1}{2}$ ) tendieren.

Mit den Bezeichnungen  $\pi_{tpj} = \frac{\exp(\alpha_{tp} + \beta'_{tp}\mu_j)}{\sum_{q=1}^P \exp(\alpha_{tq} + \beta'_{tq}\mu_j)}$  und  $\lambda_{tpj} = \log\left(\frac{\pi_{tpj}}{\pi_{tPj}}\right)$  ergibt sich die

faktorenanalytische Modellgleichung im Mehrparteienfall:

$$\lambda_{tpj} = \alpha_{tp} + \beta'_{tp}\mu_j, \quad t = 1, 2; \quad p = 1, \dots, P; \quad j = 1, \dots, m.$$

Weitere Eigenschaften dieses Modells:

- Mit  $\sigma^2 = 0$  gilt  $P_{tpj} = \pi_{tpj}$  und  $\lambda_{tpj} = \log(P_{tpj} / P_{tPj}) =: L_{tpj}$ . In diesem Fall reduziert sich die Bestimmung der Modellparameter auf eine Hauptkomponentenanalyse und das Modell selbst ist äquivalent zum „Neighbourhood model“ aus Abschnitt 2.2.9.
- Für fest vorgegebene Parameter  $\alpha_{tp}$ ,  $\beta_{tp}$  und  $\mu_j$  gilt  $P_{tpj} \xrightarrow{\sigma \rightarrow \infty} \bar{P}_{tp}$ ,  $L_{tpj} \xrightarrow{\sigma \rightarrow \infty} \bar{L}_{tp}$  mit  $\bar{P}_{tp}, \bar{L}_{tp}$  unabhängig von  $\mu_j$ .

Die Schätzung des Modells besteht nun darin, Werte für die Parameter  $\alpha_{ip}$ ,  $\beta_{ip}$  und  $\mu_j$  zu finden, für welche die Gleichungen  $\bar{W}_{ipj} = E\left(\frac{\exp(\alpha_{ip} + \beta'_{ip} Z_{ji})}{\sum_{q=1}^P \exp(\alpha_{iq} + \beta'_{iq} Z_{ji})}\right)$  mit dem a priori gewählten  $\sigma$  erfüllt sind.

Beim Probit-Modell aus 3.6.1 handelt es sich um eine Approximation des multinomialen Logit-Modells im Fall  $P = 2$ . In Analogie zu den Eigenschaften E1 und E2 des Probit-Modells können die folgenden beiden Eigenschaften des multinomialen Logit-Modells vermutet werden:

**V1: Eindeutigkeit der Schätzung.** Zu gegebenen Wähleranteilen  $\bar{W}_{ipj}$  existiert eine eindeutige Lösung des Modells in Form von Werten der Parameter  $\alpha_{ip}, \beta_{ip}, \mu_j$ , mit welchen (zusätzlich zu den an die  $\mu_j$  gestellten Anforderungen) die Gleichungen

$$\bar{W}_{ipj} = E\left(\frac{\exp(\alpha_{ip} + \beta'_{ip} Z_{ji})}{\sum_{q=1}^P \exp(\alpha_{iq} + \beta'_{iq} Z_{ji})}\right)$$

erfüllt sind (die Eindeutigkeit ist dabei zu verstehen als Eindeutigkeit bis auf Vorzeichenwechsel der Koeffizienten  $\beta_{ip}$  bei gleichzeitigem „Umpolen“ der entsprechenden Komponenten der Vektoren  $\mu_j$ ). Die Eindeutigkeit ist im Fall  $\sigma = 0$  gegeben, für  $\sigma > 0$  liegt kein Beweis vor.

**V2: Obere Schranke für  $\sigma$ .** Ist  $\sigma$  grösser als ein bestimmter (unbekannter) Wert, so kann das Modell nicht mehr an die beobachteten Daten angepasst werden, da die Streuung der Anteile über Wahlkreise hinweg zu gering wird.

Für diese beiden vermuteten Eigenschaften des multinomialen Logit-Modells konnten leider keine Beweise gefunden werden.

### 3.6.3 Modellschätzung<sup>9</sup>

Die Schätzung des Modells erfolgt iterativ durch wiederholte lineare Approximation der Grössen  $L_{ipj} = \log\left(\frac{P_{ipj}}{P_{ipj}}\right)$  in Abhängigkeit der  $\lambda_{ipj}$ . Um die Notation nicht zu überlasten werden im folgenden gelegentlich die Indizes weggelassen. Ausserdem werden die Einstellungsvektoren  $Z_{ji}$  additiv zerlegt in den Wahlkreis-Mittelwert  $\mu_j$  und die individuelle Komponente  $\zeta_{ji}$ :  $Z_{ji} = \mu_j + \zeta_{ji}$ .

Aus den  $\lambda_{ipj}$  können die  $\alpha_{ip}$  als Mittelwerte, die  $\beta_{ip}$  aus den Koeffizienten der Hauptkomponentenanalyse (durch Multiplikation mit der Wurzel des zugehörigen Eigenwerts) und die  $\mu_j$  aus den factor scores (durch Division durch die Wurzel des zugehörigen Eigenwerts) ermittelt wer-

<sup>9</sup> Dieser Abschnitt erfordert mehr Vorkenntnisse aus der Mathematik als der Rest des Berichts und kann bei der Lektüre übersprungen werden.

den, d.h. die  $\lambda_{tpj}$  enthalten die volle Information der Parametrisierung des Modells. Wir suchen eine Lösung  $\{\lambda_{tpj}\}$ , für welche die Gleichungen

$$P_{tpj} = E \left( \frac{\exp(\alpha_{tp} + \beta'_{tp} Z_{ji})}{\sum_{q=1}^P \exp(\alpha_{tq} + \beta'_{tq} Z_{ji})} \right) = E \left( \frac{\exp(\lambda_{tpj} + \beta'_{tp} \zeta_{ji})}{\sum_{q=1}^P \exp(\lambda_{tqj} + \beta'_{tq} \zeta_{ji})} \right)$$

erfüllt sind. Es liegt somit die Situation vor, dass zu einer Funktion  $L(\lambda)$  das Argument  $\lambda^*$  gesucht ist, für welches gilt  $L(\lambda^*) = L$ .  $L$  ist bekannt, da es sich aus den Wähleranteilen berechnen lässt, und Werte der Funktion  $L_{tpj}$  können bei bekanntem  $\lambda_{tpj}$  mittels Simulation von

$$P_{tpj} = E \left( \frac{\exp(\lambda_{tpj} + \beta'_{tp} \zeta_{ji})}{\sum_{q=1}^P \exp(\lambda_{tqj} + \beta'_{tq} \zeta_{ji})} \right)$$

und anschließender Transformation  $L_{tpj} = \log(P_{tpj} / P_{tPj})$  näherungsweise ermittelt werden. Die lineare Taylor-Approximation von  $L(\lambda)$  an der Stelle  $\lambda_0$  lautet

$$L(\lambda) \approx L(\lambda_0) + \left. \frac{dL}{d\lambda} \right|_{\lambda=\lambda_0} (\lambda - \lambda_0).$$

Falls die Matrix  $\frac{dL}{d\lambda}$  vollen Rang hat, kann diese Gleichung nach  $\lambda$  aufgelöst werden und lautet dann

$$\lambda \approx \lambda_0 + \left. \frac{d\lambda}{dL} \right|_{\lambda=\lambda_0} (L(\lambda) - L(\lambda_0)).$$

Die Bestimmung der Ableitungen  $\frac{dP}{d\lambda}$  muss ebenfalls mittels Simulation erfolgen, wobei wir von der folgenden Approximation Gebrauch machen:

$$\frac{dP_{tpj}}{d\lambda_{sqj}} \approx E \left\{ \frac{\exp(\lambda_{tpj} + \beta'_{tp} \zeta_{ji})}{\sum_{r=1}^P \exp(\lambda_{trj} + \beta'_{tr} \zeta_{ji})} \cdot \left( 1_{\{p=q\}} - \frac{\exp(\lambda_{sqj} + \beta'_{sq} \zeta_{ji})}{\sum_{r=1}^P \exp(\lambda_{srj} + \beta'_{sr} \zeta_{ji})} \right) \cdot 1_{\{s=t\}} \right\},$$

(die Differentiation kann innerhalb des Integrals erfolgen, da  $P_{tpj}$  beschränkt und stetig in  $\lambda_{sqj}$  ist). Die Approximation in der obigen Formel für  $\frac{dP}{d\lambda}$  besteht darin, dass der indirekte Einfluss ignoriert wird, den  $\lambda$  über den Hauptkomponentenanalyse-Koeffizienten  $\beta$  auf  $P$  ausübt. Der Effekt von Einzelbeobachtungen auf die Koeffizienten der Hauptkomponentenanalyse sollte bei genügender Anzahl Beobachtungen so klein sein, dass die durch die Approximation verursachte Ungenauigkeit die Konvergenz des Verfahrens nicht beeinträchtigt.  $\frac{dL}{d\lambda}$  lässt sich aus  $\frac{dP}{d\lambda}$  unter Anwendung der Kettenregel schätzen.

Die Schätzung der  $\lambda$  erfolgt dadurch, dass iterativ  $L(\lambda)$  und  $\frac{dL}{d\lambda}$  simuliert werden und aufgrund der nach  $\lambda$  aufgelösten Taylor-Approximation ein neues  $\lambda$  bestimmt wird, bis die simulierten Werte  $P(\lambda)$  nahe genug an den beobachteten Stimmenanteilen liegen. Dieses Vorgehen entspricht einer Anwendung des Newton-Raphson-Verfahrens zum Auffinden von Nullstellen einer Funktion.

Die einzelnen Schritte bei der Lösungsfindung lauten im Detail:

- Wähle als Ausgangswerte  $\pi_{tpj}^{(0)} = P_{tpj} = \overline{W}_{tpj}$ ,  $\lambda_{tpj}^{(0)} = L_{tpj} = \log(P_{tpj} / P_{tpj})$  sowie  $\alpha_{ip}^{(0)}$ ,  $\beta_{ip}^{(0)}$  und  $\mu_j^{(0)}$  aus der Hauptkomponentenanalyse

$$\lambda_{tpj}^{(0)} = \alpha_{ip}^{(0)} + \beta_{ip}^{(0)} \mu_j^{(0)} \quad (t = 1, 2, p = 1, \dots, P-1).$$

- Wiederhole die folgenden Schritte zur Schätzung des Gesamtmodells solange, bis die Schätzungen  $\hat{P}_{tpj}^{(i)}$  befriedigend nahe bei den  $P_{tpj} = \overline{W}_{tpj}$  sind. Der hochgestellte Index ( $i$ ),  $i = 1, 2, \dots$ , gibt jeweils die Nummer der Iteration an.

1. Schätze  $P_{tpj}^{(i)}$  durch Simulation des Modells

$$\hat{P}_{tpj}^{(i)} = E \left( \frac{\exp(\lambda_{tpj}^{(i-1)} + \beta_{ip}^{(i-1)} \zeta_{ji})}{\sum_{r=1}^P \exp(\lambda_{trj}^{(i-1)} + \beta_{tr}^{(i-1)} \zeta_{ji})} \right) \text{ und bestimme}$$

$$\hat{L}_{tpj}^{(i)} = \log \left( \frac{P_{tpj}^{(i)}}{P_{tpj}^{(i)}} \right).$$

Die  $\zeta_{ji}$  sind dabei unabhängige Zufallsvariablen aus der Normalverteilung  $N(0, \sigma^2)$ .

2. Schätze die  $\left( \frac{dL_{tpj}}{d\lambda_{sqj}} \right)_j^{(i)}$  durch Simulation von

$$\left( \frac{dP_{tpj}}{d\lambda_{sqj}} \right)_j^{(i)} = E \left\{ \frac{\exp(\lambda_{tpj}^{(i-1)} + \beta_{ip}^{(i-1)} \zeta_{ji})}{\sum_{r=1}^P \exp(\lambda_{trj}^{(i-1)} + \beta_{tr}^{(i-1)} \zeta_{ji})} \cdot \left( 1_{\{p=q\}} - \frac{\exp(\lambda_{sqj}^{(i)} + \beta_{sq}^{(i)} \zeta_{ji})}{\sum_{r=1}^P \exp(\lambda_{srj}^{(i)} + \beta_{sr}^{(i)} \zeta_{ji})} \right) \cdot 1_{\{s=t\}} \right\}$$

und anschließender Anwendung der Kettenregel:

$$\left( \frac{dL_{tpj}}{d\lambda_{sqj}} \right)_j^{(i)} = \frac{dL_{tpj}^{(i)}}{dP_{tpj}^{(i)}} \left( \frac{dP_{tpj}}{d\lambda_{sqj}} \right)_j^{(i)} = \left( \frac{dP_{tpj}}{d\lambda_{sqj}} \right)_j^{(i)} \Bigg/ P_{tpj}^{(i)} - \left( \frac{dP_{tpj}}{d\lambda_{sqj}} \right)_j^{(i)} \Bigg/ P_{tpj}^{(i)}.$$

3. Bilde die Matrizen  $\left(\frac{d\mathbf{L}}{d\boldsymbol{\lambda}}\right)_j^{(i)}$  der Dimension:  $(k \times k) = (2(P-1) \times 2(P-1))$  mit den Elementen  $\left(\frac{dL_{tpj}}{d\lambda_{sqj}}\right)_j^{(i)}$ . Ermittle die inversen Matrizen  $\left(\frac{d\boldsymbol{\lambda}}{d\mathbf{L}}\right)_j^{(i)} = \left(\left(\frac{d\mathbf{L}}{d\boldsymbol{\lambda}}\right)_j^{(i)}\right)^{-1}$ .

4. Bestimme die Vektoren  $\boldsymbol{\lambda}_j^{(i)}$  gemäss:

$$\boldsymbol{\lambda}_j^{(i)} = \boldsymbol{\lambda}_j^{(i-1)} + \left(\frac{d\boldsymbol{\lambda}}{d\mathbf{L}}\right)_j^{(i)} (\mathbf{L}_j - \hat{\mathbf{L}}_j^{(i)}).$$

$\boldsymbol{\lambda}_j^{(i)}$  ist dabei ein Vektor der Dimension  $k = 2(P-1)$  mit den Komponenten  $\lambda_{tpj}^{(i)}$ ,  $t = 1, 2$ ;  $p = 1, \dots, P-1$ .  $\boldsymbol{\lambda}_j^{(i-1)}$ ,  $\mathbf{L}_j$ ,  $\hat{\mathbf{L}}_j^{(i)}$  setzen sich analog aus den Komponenten  $\lambda_{tpj}^{(i-1)}$ ,  $L_{tpj}$ ,  $\hat{L}_{tpj}^{(i)}$  zusammen.

5. Bestimme  $\alpha_p^{(i)}$ ,  $\beta_p^{(i)}$  und  $\mu_j^{(i)}$  mit der Hauptkomponentenanalyse

$$\lambda_{tpj}^{(i)} = \alpha_p^{(i)} + \beta_p^{(i)} \mu_j^{(i)} \quad (t = 1, 2, \quad p = 1, \dots, P-1).$$

- Ermittle den Anteil der Wahlberechtigten in Wahlkreis  $j$ , die in der ersten Wahl Partei  $p$  und in der zweiten  $q$  wählen, durch Simulation von

$$\hat{W}_{(p,q)j} = E \left[ \frac{\exp\left(\hat{\lambda}_{1pj} + \hat{\lambda}_{2qj} + (\hat{\beta}_{1p} + \hat{\beta}_{2q})' \zeta_{ji}\right)}{\sum_{r=1}^P \exp(\hat{\lambda}_{1rj} + \hat{\beta}'_{1r} \zeta_{ji}) \cdot \sum_{r=1}^P \exp(\hat{\lambda}_{2rj} + \hat{\beta}'_{2r} \zeta_{ji})} \right].$$

### 3.6.4 Anwendung auf die Nationalratswahlen im Kanton Zürich 1995/99

5 der 171 Gemeinden müssen aus technischen Gründen weggelassen werden, da eine der berücksichtigten Parteien in einer Wahl 0 Stimmen erhielt, und die Berechnung der entsprechenden Grösse  $L_{tpj} = \log(\overline{W}_{tpj} / \overline{W}_{tpj})$  in diesen Fällen nicht möglich ist. Da diese 5 Gemeinden nur knapp 0.2% der Wahlberechtigten umfassen, fällt dies kaum ins Gewicht.

Das in 3.6.3 skizzierte Verfahren wurde mit der Statistiksoftware SAS programmiert. Die präsentierten Resultate ergaben sich aus maximal 10 Iterationsschritten, die Simulationsschritte umfassten jeweils 100'000 Replikationen. Die vorgenommenen Rechnungen sind ausserordentlich rechenintensiv und nahmen viel Zeit in Anspruch.

$\sigma=0$  (Neighbourhood model):

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	0.05%	0.18%	0.06%	0.23%	0.25%	0.22%	1.32%	2.31%
FDP	0.17%	0.63%	0.23%	0.80%	0.88%	0.76%	4.58%	8.04%
GPS	0.04%	0.15%	0.05%	0.19%	0.20%	0.18%	1.06%	1.87%
SPS	0.24%	0.90%	0.32%	1.15%	1.27%	1.09%	6.58%	11.56%
SVP	0.31%	1.14%	0.41%	1.46%	1.61%	1.39%	8.35%	14.66%
Übrige	0.14%	0.52%	0.19%	0.66%	0.73%	0.63%	3.81%	6.68%
Nichtwähler	1.15%	4.28%	1.54%	5.45%	6.01%	5.19%	31.26%	54.88%
<b>Total 1995</b>	<b>2.10%</b>	<b>7.79%</b>	<b>2.81%</b>	<b>9.93%</b>	<b>10.96%</b>	<b>9.45%</b>	<b>56.97%</b>	<b>100.00%</b>

$\sigma=1$ :

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	0.07%	0.20%	0.06%	0.24%	0.21%	0.21%	1.33%	2.32%
FDP	0.18%	0.84%	0.23%	0.79%	0.86%	0.75%	4.38%	8.03%
GPS	0.04%	0.14%	0.06%	0.20%	0.22%	0.18%	1.03%	1.87%
SPS	0.25%	0.91%	0.34%	1.27%	1.19%	1.11%	6.52%	11.61%
SVP	0.26%	1.11%	0.44%	1.36%	1.99%	1.39%	8.01%	14.57%
Übrige	0.14%	0.52%	0.20%	0.68%	0.75%	0.67%	3.75%	6.69%
Nichtwähler	1.16%	4.09%	1.48%	5.38%	5.78%	5.14%	31.90%	54.92%
<b>Total 1995</b>	<b>2.09%</b>	<b>7.82%</b>	<b>2.81%</b>	<b>9.90%</b>	<b>10.99%</b>	<b>9.45%</b>	<b>56.92%</b>	<b>100.00%</b>

$\sigma=2$ :

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	0.14%	0.24%	0.04%	0.25%	0.12%	0.19%	1.34%	2.33%
FDP	0.21%	1.26%	0.23%	0.81%	0.72%	0.73%	4.07%	8.03%
GPS	0.03%	0.14%	0.09%	0.21%	0.25%	0.19%	0.95%	1.87%
SPS	0.27%	0.96%	0.37%	1.48%	1.12%	1.12%	6.29%	11.61%
SVP	0.19%	0.95%	0.51%	1.26%	2.79%	1.44%	7.43%	14.56%
Übrige	0.13%	0.50%	0.22%	0.69%	0.79%	0.74%	3.62%	6.68%
Nichtwähler	1.14%	3.77%	1.36%	5.20%	5.20%	5.02%	33.23%	54.92%
<b>Total 1995</b>	<b>2.10%</b>	<b>7.82%</b>	<b>2.81%</b>	<b>9.90%</b>	<b>11.00%</b>	<b>9.44%</b>	<b>56.92%</b>	<b>100.00%</b>

$\sigma=3$ :

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	0.33%	0.30%	0.02%	0.24%	0.03%	0.14%	1.24%	2.31%
FDP	0.25%	2.19%	0.21%	0.80%	0.45%	0.67%	3.44%	8.01%
GPS	0.02%	0.12%	0.14%	0.25%	0.31%	0.21%	0.82%	1.87%
SPS	0.27%	0.99%	0.42%	1.90%	0.93%	1.16%	5.96%	11.62%
SVP	0.09%	0.65%	0.61%	1.06%	4.53%	1.47%	6.18%	14.58%
Übrige	0.10%	0.44%	0.25%	0.72%	0.83%	0.90%	3.43%	6.67%
Nichtwähler	1.04%	3.12%	1.14%	4.95%	3.97%	4.88%	35.82%	54.93%
<b>Total 1995</b>	<b>2.09%</b>	<b>7.80%</b>	<b>2.81%</b>	<b>9.92%</b>	<b>11.06%</b>	<b>9.42%</b>	<b>56.89%</b>	<b>100.00%</b>

$\sigma=4$  (keine Konvergenz!):

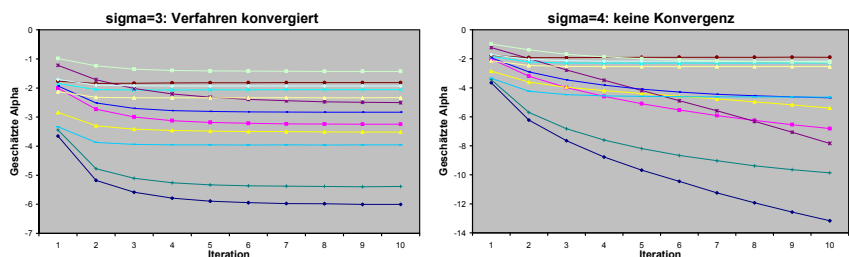
Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	0.86%	0.39%	0.00%	0.18%	0.00%	0.06%	1.01%	2.51%
FDP	0.28%	4.24%	0.15%	0.64%	0.25%	0.44%	2.21%	8.21%
GPS	0.01%	0.07%	0.26%	0.26%	0.42%	0.23%	0.60%	1.85%
SPS	0.27%	0.93%	0.47%	2.56%	0.74%	1.15%	5.36%	11.48%
SVP	0.02%	0.27%	0.67%	0.61%	8.15%	1.26%	3.77%	14.74%
Übrige	0.07%	0.33%	0.33%	0.73%	0.90%	1.18%	3.07%	6.61%
Nichtwähler	0.93%	2.12%	0.84%	4.38%	2.66%	4.42%	39.25%	54.60%
<b>Total 1995</b>	<b>2.44%</b>	<b>8.35%</b>	<b>2.73%</b>	<b>9.35%</b>	<b>13.11%</b>	<b>8.75%</b>	<b>55.27%</b>	<b>100.00%</b>

$\sigma=5$  (keine Konvergenz!):

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	2.34%	0.74%	0.00%	0.14%	0.01%	0.02%	1.15%	4.40%
FDP	0.52%	7.16%	0.13%	0.35%	0.35%	0.16%	1.11%	9.78%
GPS	0.01%	0.05%	0.43%	0.19%	0.62%	0.22%	0.28%	1.80%
SPS	0.49%	0.92%	0.54%	2.93%	0.69%	0.79%	3.93%	10.28%
SVP	0.02%	0.14%	0.52%	0.15%	13.05%	0.84%	1.50%	16.20%
Übrige	0.10%	0.27%	0.48%	0.60%	1.00%	1.49%	2.22%	6.15%
Nichtwähler	1.64%	1.84%	0.71%	2.78%	2.17%	2.78%	39.49%	51.40%
<b>Total 1995</b>	<b>5.11%</b>	<b>11.11%</b>	<b>2.80%</b>	<b>7.13%</b>	<b>17.87%</b>	<b>6.30%</b>	<b>49.66%</b>	<b>100.00%</b>

Das multinomiale Logit-Modell mit  $\sigma=0$  entspricht dem Neighbourhood model, welches auf der unrealistischen Annahme basiert, dass zwischen Wähleranteilen der ersten und zweiten Wahl kein Zusammenhang besteht (vgl. Abschnitt 2.2.9). Durch sukzessive Erhöhung des Parameters  $\sigma$  auf die Werte 1, 2 und 3 wird der Zusammenhang zwischen den Anteilen beider Wahlen verstärkt. Als Folge davon nehmen die Werte auf der Diagonalen etwas zu, sind aber bei  $\sigma=3$  noch recht tief; die Stammwähleranteile steigen auf maximal 40% des jeweiligen Wähleranteils (bei der SVP 1995), bleiben aber ansonsten deutlich tiefer.

Für  $\sigma \leq 3$  konvergiert das Verfahren und die erhaltene Lösung stimmt in den Randhäufigkeiten gut mit den wahren Werten überein. Ab  $\sigma=4$  hingegen treten Konvergenzprobleme auf. Dies ist in Abbildung 10 am Beispiel der Entwicklung der geschätzten Koeffizienten  $\alpha_p^{(i)}$  in Abhängigkeit der Nummer der Iteration  $i$  illustriert.



**Abbildung 10:** Bei  $\sigma=3$  konvergieren die geschätzten Koeffizienten  $\alpha_p^{(i)}$  gegen eine Lösung, welche gut mit den Randhäufigkeiten übereinstimmt. Bei  $\sigma=4$  zeichnet sich nach 10 Iterationsschritten keine Konvergenz ab.

Ausgehend von der vermuteten Oberen Schranke für den Parameter  $\sigma$  (vermutete Eigenschaft V2) lautet eine mögliche Erklärung für das Ausbleiben von Konvergenz bei  $\sigma \geq 4$ , dass die Streuung der Wähleranteile zwischen den Gemeinden zu gross ist, und dass deshalb keine Lösung mit  $\sigma \geq 4$  existiert. Indizien hierfür sind:

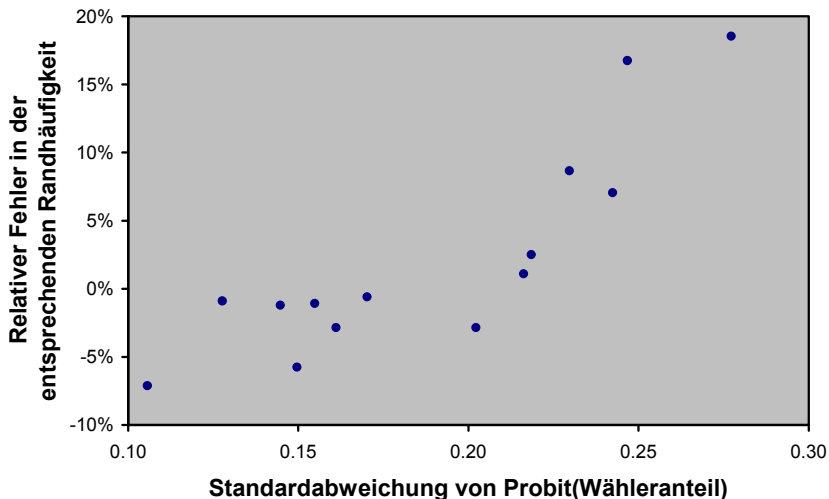
- Die Grössenordnung der oberen Schranke für  $\sigma$  kann durch die entsprechende obere Schranke im Probit-Modell, welches die Wähleranteile jeweils einer Partei untersucht, abgeschätzt werden, also mit der Formel

$$\sigma^2 < (\tilde{\beta}_{ip}^* \tilde{\beta}_{ip})^{-1} = \frac{1}{v_{ip}^2} \quad \text{mit} \quad v_{ip}^2 = \frac{1}{m} \sum_{j=1}^m (\Phi^{-1}(\bar{W}_{ipj}) - \bar{\Phi}_{ip})^2 \quad \text{und} \quad \bar{\Phi}_{ip} = \frac{1}{m} \sum_{j=1}^m \Phi^{-1}(\bar{W}_{ipj}).$$

Derjenige Wähleranteil mit der maximalen Varianz nach Probit-Transformation ist der Anteil der SVP 1995 mit  $v_{ip}^2 = 0.0768$ . Daraus ergibt sich eine obere Schranke für  $\sigma$  von  $1/\sqrt{0.0768} = 3.61$ .

- Bei näherer Untersuchung der geschätzten Wanderungstabelle zu  $\sigma=4$  stellt sich heraus, dass die Schätzung der Wähleranteile durch das Modell in denjenigen Fällen besonders schlecht ausfällt, in denen der Wert  $1/v_{ip}$  unter oder nahe bei 4 liegt. Dies ist in Abbildung 11 illustriert. Diese Tatsache steht in Übereinstimmung mit der Vermutung, dass zu grosse Streuung der  $\Phi^{-1}(\bar{W}_{ipj})$  bei der fehlenden Konvergenz eine Rolle spielt.

Bei diesen Überlegungen handelt es sich natürlich keinesfalls um einen Beweis für die vermutete Ursache des Ausbleibens von Konvergenz. Diese Indizien lassen aber die Annahme plausibel erscheinen, die fehlende Konvergenz in den Fällen  $\sigma=4$  und  $\sigma=5$  sei auf zu grosse Streuung gewisser Wähleranteile zurückzuführen.



**Abbildung 11:** Zusammenhang zwischen der Überschätzung der Randtotale und der empirischen Standardabweichung  $v_{ip}$  der zugehörigen  $\Phi^{-1}(\bar{W}_{ipj})$ . Jeder Punkt entspricht einer Kombination einer Partei und einer Wahl, z.B. entspricht der Punkt in der oberen rechten Ecke der SVP 1995.



Geht man davon aus, dass in den Fällen  $\sigma \geq 4$  tatsächlich keine Lösung existiert, so lautet die Erkenntnis aus den Berechnungen folgendermassen: Der verwendete Modellansatz eignet sich nicht zur Modellierung dieser Daten. Dabei liegt das Problem darin, dass der Parameter  $\sigma$  zwei völlig unterschiedliche Rollen im Modell übernimmt. Eine Erhöhung von  $\sigma$  bedeutet einerseits eine stärkere Abhängigkeit zwischen den Wahlresultaten in der ersten und zweiten Wahl (dies ist die primäre, beabsichtigte Rolle des Parameters im Modell). Andererseits ist bei grossem  $\sigma$  die mögliche Streuung der Wähleranteile zwischen Wahlkreisen begrenzt. Mit den vorliegenden Daten ist der multinomiale Logit-Ansatz nicht imstande, einerseits einen genügend starken Zusammenhang zwischen den Resultaten der beiden Wahlen und andererseits die tatsächlich beobachtete Streuung der Wahlergebnisse der verschiedenen Wahlkreise gleichzeitig zu modellieren.

### **Fazit zu 3.6:**

1. Modell: Das verwendete theoretische Modell ist demjenigen von Thomsen sehr ähnlich (vgl. hierzu das Fazit zu 2.2.7 und 3.4). Als zusätzliches Element kann der Parameter  $\sigma$  variiert werden, welcher die Stärke der Abhängigkeit der Wahlentscheidungen in den beiden Wahlen festlegt.
2. Überprüfbarkeit: Ist nicht gegeben.
3. Schätzung: Es handelt sich um eine rein deterministische Schätzung mit der Eigenschaft, dass Parameterwerte gesucht werden, mit denen das Modell in Übereinstimmung mit den bekannten Randhäufigkeiten steht. Ob die Lösungen eindeutig sind, ist jedoch unklar.

Unsere Ergebnisse im vorliegenden Anwendungsbeispiel deuten darauf hin, dass der verwendete Modellansatz sich nicht zur Modellierung dieser Daten eignet. Diese Erkenntnis ist aber mit einer gewissen Unsicherheit behaftet, da die beiden wichtigen Eigenschaften der Eindeutigkeit einer Modelllösung und der Existenz einer oberen Schranke für den Parameter  $\sigma$  nicht bewiesen werden konnten.

## **3.7 Diskussion**

Zum allgemeinen Fall der ökologischen Inferenz mit  $P$  Wahlalternativen in der ersten und  $Q$  in der zweiten Wahl liegt deutlich weniger Literatur vor als zum Spezialfall  $P = Q = 2$ , obwohl letzterer die meisten Anwendungen nicht erfasst.

Im Zusammenhang mit der Regressionsmethode und ihren Verallgemeinerungen treten keine substantiell neuen Diskussionspunkte auf. Die Anwendung der Regressionsmethode im Beispiel zeigt, wie für die Schätzung des scheinbar einfachen Modells in der Praxis zusätzliche Anpassungsschritte erforderlich werden, wenn ein Teil der aus der Regression resultierenden Übergangswahrscheinlichkeiten nicht im Einheitsintervall liegt. Die Tatsache, dass im Anwendungsbeispiel mehrere Übergangswahrscheinlichkeiten teilweise deutlich ausserhalb des Einheitsintervalls liegen, kann als Indiz dafür gewertet werden, dass die Modellvoraussetzungen in unseren Daten der Nationalratswahlen im Kanton Zürich nicht erfüllt sind.

Die Verallgemeinerung von Kings Methode liegt nur in groben Zügen vor; entsprechende Software existiert im Gegensatz zum Zweiparteienfall nicht.

Thomsens multinomiales Logit-Modell für den Mehrparteienfall ist auf der Ebene des theoretischen Modells unproblematisch. Sein Lösungsvorschlag weist aber den Mangel auf, dass die Wahl unterschiedlicher Referenzparteien zu unterschiedlichen Resultaten führt, was im Gegensatz

zu einer theoretischen Eigenschaft des Modells steht, gemäss welcher die Wahl der Referenzpartei keine Auswirkungen auf die Wählerströme hat. Ausserdem bleibt das im Zweiparteienfall erwähnte Problem in Thomsens Lösungsweg bestehen.

Unsere Berechnungen mit einer leicht modifizierten Variante des multinomialen Logit-Modells im vorliegenden Anwendungsbeispiel deuten zudem auf ein weiteres Problem dieses Modells hin. Dieses besteht darin, dass der Modellansatz nicht gleichzeitig starke Abhängigkeiten der Resultate beider Wahlen und grössere Streuungen in den Wahlergebnissen verschiedener Wahlkreise zu erfassen vermag. Diese Erkenntnis ist aber mit einer gewissen Unsicherheit behaftet, da zwei vermutete Eigenschaften des theoretischen Modells nicht formell bewiesen werden konnten.

## 4 Weitere Probleme in praktischen Anwendungen

### 4.1 Nichtwähler

Die Tatsache, dass die Wähler als zusätzliche Alternative zur Wahl einer Partei die Möglichkeit der Stimmenthaltung haben, muss in Wählerstromanalysen berücksichtigt werden, um eine vollständige Übersicht über die Wählerwanderung zu erhalten. Dies gilt um so mehr, je grösser der Anteil der Nichtwähler ist. Methodisch stellt dies kein Problem dar, da die Nichtwähler als eine weitere Partei betrachtet werden können.

Liegen Angaben über die Anzahl der Stimmberechtigten vor, so lässt sich die Anzahl der Nichtwähler als deren Differenz zur Summe der Stimmen aller Parteien ermitteln. Alternativ lässt sich die Zahl der Nichtwähler aus der Wahlbeteiligung ermitteln. Da Angaben über die Zahl der Stimmberechtigten oder die Wahlbeteiligung in aller Regel verfügbar sind, dürfte die Zahl der Nichtwähler allgemein kein bedeutendes Hindernis bei der Rekonstruktion von Wählerströmen darstellen.

### 4.2 Zeitliche Veränderung der Population der Wahlberechtigten

Bisher sind wir beim Vergleich von Wähleranteilen davon ausgegangen, dass die Gesamtheit der Wahlberechtigten zum Zeitpunkt der zwei betrachteten Wahlen in allen Wahlkreisen vollkommen identisch war. In Wirklichkeit ist diese Voraussetzung natürlich kaum je erfüllt. Mutationen treten auf

- durch den Umzug von Personen zwischen Wahlkreisen, in das betrachtete Wahlgebiet oder aus dem Wahlgebiet hinaus,
- durch das Erlangen der Wahlberechtigung durch Personen in der Zeit zwischen den beiden Wahlterminen,
- durch den Tod von Wahlberechtigten.

#### 4.2.1 „Noch nicht Wahlberechtigte“ und „nicht mehr Wahlberechtigte“

Rein formell kann diesen Veränderungen Rechnung getragen werden, indem für jede Wahl eine zusätzliche Kategorie („Partei“) definiert wird, nämlich

- für die erste Wahl diejenige der noch nicht Wahlberechtigten, also derjenigen Personen, die bei der zweiten Wahl im betreffenden Wahlkreis wahlberechtigt waren, in der ersten jedoch nicht,
- für die zweite Wahl diejenige der nicht mehr Wahlberechtigten, also derjenigen Personen, die bei der ersten Wahl im betreffenden Wahlkreis die Wahlberechtigung hatten, in der zweiten jedoch nicht.

Auf diese Art wird erreicht, dass (zumindest formell) zwei Einteilungen derselben eindeutig definierten Population zu zwei Zeitpunkten vorliegen:

- Wähler der verschiedenen Parteien, Nichtwähler und noch nicht Wahlberechtigte in der ersten Wahl,
- Wähler der verschiedenen Parteien, Nichtwähler und nicht mehr Wahlberechtigte in der zweiten Wahl.

Die betreffende Population umfasst alle Personen, die im jeweiligen Wahlkreis in einer oder beiden Wahlen wahlberechtigt waren. Die Zelle (noch nicht Wähler / nicht mehr Wähler) in der Wanderungstabelle hat per definitionem die Häufigkeit Null.

Im Gegensatz zur Zahl der Nichtwähler sind Daten über die Anzahl der noch nicht bzw. nicht mehr Wahlberechtigten in der Regel nicht ohne weiteres zu beschaffen. Wird die zeitliche Veränderung der betrachteten Population der Wahlberechtigten aber ignoriert, so heisst das, dass in jeder Dimension der betreffenden Tabelle eine Kategorie nicht berücksichtigt wird. Die Folge sind systematische Verfälschungen der Resultate der ökologischen Analyse, wie im nächsten Unterabschnitt diskutiert werden soll.

#### 4.2.2 Verfälschungen aufgrund unberücksichtigter Änderungen der Population der Wahlberechtigten

Da in praktischen Anwendungen meistens keine Daten über die Anzahl der noch nicht und nicht mehr Wahlberechtigten vorliegen, muss die Analyse mit Anteilen aus der ersten und zweiten Wahl durchgeführt werden, die sich auf unterschiedliche Populationen beziehen. Sind die Populationsänderungen zwischen den beiden Wahlen gering, so fallen die Verfälschungseffekte nicht ins Gewicht. Je grösser der zeitliche Abstand zwischen zwei Wahlen ist, und je grösser die Mobilität der Bevölkerung ist, um so schwerwiegender ist das Problem.

Wie diese Problematik zu Verfälschungen führen kann, soll anhand eines kleinen Beispiels illustriert werden. Wir gehen von einem Zweiparteiensystem ohne Wahlabstinenz aus. Die Voraussetzungen der ökologischen Regression seien perfekt erfüllt, d.h. in sämtlichen Wahlkreisen werden sich genau  $p_{AA}=80\%$  der A-Wähler aus Wahl 1 sowie  $p_{BA}=10\%$  der B-Wähler aus Wahl 1 in der zweiten Wahl für Partei A entscheiden. Alle in der ersten Wahl wahlberechtigten Personen nehmen auch an der zweiten Wahl teil. In der zweiten Wahl werden sich zusätzlich ausnahmslos für Partei A stimmende Neuwähler beteiligen, deren Zahl in jedem Wahlkreis genau  $w \cdot 100\%$  der Population der bisher Wahlberechtigten entspricht. In der geschilderten Situation wird in Wahl 2 anstelle von  $y$ , dem Anteil der Wähler aus der ersten Wahl, die für A stimmt, eine Grösse

$\tilde{y} = \frac{y + w}{1 + w}$  beobachtet. Die folgende Tabelle zeigt die Auswirkungen auf die Schätzung der Übergangswahrscheinlichkeiten  $p_{AA}$  und  $p_{BA}$  mittels ökologischer Regression für verschiedene Neuwähleranteile  $w$ :

$w$	$\hat{p}_{AA}$	$\hat{p}_{BA}$
0%	80.0%	10.0%
5%	81.0%	14.3%
10%	81.8%	18.2%
15%	82.6%	21.7%
20%	83.3%	25.0%
25%	84.0%	28.0%

Natürlich handelt es sich beim skizzierten Szenario um eine nicht sehr realistische Extremsituation, doch es zeigt sich, dass schon ein geringer Anteil Neuwähler beträchtlichen Einfluss auf die Schätzung der Übergangswahrscheinlichkeiten haben kann. In unserem Extrembeispiel können schon 5% Neuwähler die Schätzung von  $p_{BA}$  derart verfälschen, dass es um über 40% überschätzt wird.

Der unproblematischste Fall liegt dann vor, wenn sich das Wahlverhalten der Neuwähler mit demjenigen der übrigen Wählerschaft deckt. Die Berücksichtigung der Neuwähler hat dann überhaupt keine Auswirkungen auf die Schätzung von  $p_{AA}$  und  $p_{BA}$ .

Das Ausmass dieses Problems in praktischen Anwendungen kann nicht ohne weiteres abgeschätzt werden. Da sich das Wahlverhalten der (mehrheitlich jüngeren) Neuwähler nicht wesentlich von demjenigen älterer Generationen unterscheidet, ist zweifellos ein gewisses Verfälschungspotential vorhanden.

### 4.3 Separate Analyse in homogenen Teilgebieten

Die verschiedenen Wahlkreise eines Untersuchungsgebiets können häufig in Gruppen von Wahlkreisen (Cluster) unterteilt werden, innerhalb derer das Wahlverhalten, d.h. die Stimmenanteile der verschiedenen Parteien, relativ homogen sind. Diese Einteilung kann zum Beispiel Unterschiede zwischen Stadt und Land oder in der sozialen Struktur der Wahlkreise widerspiegeln. Die Annahme ist naheliegend, dass auch das Muster der Wanderungen zwischen Parteien innerhalb solcher Cluster einheitlicher sind als im gesamten Untersuchungsgebiet.

Steht nun Datenmaterial aus einer grossen Anzahl von Wahlkreisen zur Verfügung, so bietet sich die Möglichkeit einer separaten Wählerstromanalyse für jedes Cluster. Für die Bildung der Cluster sind zwei grundsätzlich verschiedene Vorgehensweisen denkbar:

- Einteilung in vorgegebene geographische oder administrative Teilgebiete, von denen bekannt ist, dass sie sich bezüglich dem Wahlverhaltens der Bevölkerung unterscheiden. Dies ist besonders dann sinnvoll, wenn es sich dabei um Teilgebiete mit verschiedenen Kandidaten oder Wahllisten handelt, da in diesem Fall unterschiedliches Stimmverhalten schon aufgrund der Attraktivität der Kandidierenden zu erwarten ist.
- Einteilung mittels Clusteranalyse basierend auf den Stimmenanteilen der Parteien in den beiden Wahlen. Die Einteilung erfolgt in diesem Fall im Gegensatz zum vorangehenden Punkt nicht aufgrund externer Vorinformation, sondern direkt mit den zu analysierenden Daten. Thomsen (2000) schlägt die Bildung homogener Cluster basierend auf den Differenzen der Wähleranteile der verschiedenen Parteien vor.

Es stellt sich in diesem Zusammenhang die Frage nach einer sinnvollen Grösse der Cluster. Einerseits sollen die Cluster möglichst homogen sein, was für eine grössere Clusteranzahl spricht. Auf der anderen Seite muss die Anzahl der Wahlkreise in den Clustern ausreichend für eine statistische Analyse mit einem der in den Abschnitten 2 und 3 beschriebenen Verfahren sein.

#### 4.3.1 Nationalratswahlen im Kanton Zürich 1995/99: Regression separat in Bezirken

Der Kanton Zürich ist in 12 Bezirke eingeteilt, wobei die Stadt Zürich allein einen Bezirk bildet. Wird die Regression der Wähleranteile 1999 in Abhängigkeit der Wähleranteile 1995 separat in den Bezirken vorgenommen, so lautet die getroffene Annahme, dass die Übergangswahrscheinlichkeiten innerhalb der Bezirke konstant sind. Die Gemeinde Zürich wird bei dieser Rechnung weggelassen, da mit einer einzigen Beobachtung keine Regressionsgerade berechnet werden kann<sup>10</sup>. Auch in mehreren anderen Bezirken bewegt sich die Zahl von 10-12 Gemeinden an der unteren Grenze dessen, was als sinnvolle Clustergrösse bezeichnet werden kann.

Wird in jedem Bezirk separat ein Regressionsmodell berechnet, so bedeutet dies, dass wir die ursprüngliche Annahme gleicher Übergangswahrscheinlichkeiten in sämtlichen Gemeinden des Kantons Zürich lockern und nur noch von der schwächeren Annahme gleicher Übergangswahrscheinlichkeiten innerhalb der Bezirke ausgehen.

Von den  $11 \times 49 = 539$  Regressionskoeffizienten werden 248 negativ. Die Gültigkeit der Voraussetzungen muss also auch in diese allgemeineren Modell in Frage gestellt werden.

Berechnet man wie in Abschnitt 3.3.1 beschrieben für jeden Bezirk eine an die Randhäufigkeiten angepasste Wanderungstabelle und summiert diese gewichtet mit der Anzahl Stimmberechtigter auf, so erhält man für den ganzen Kanton ohne Stadt Zürich die folgende Wanderungstabelle:

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	0.91%	0.38%	0.04%	0.35%	0.20%	0.15%	0.32%	2.35%
FDP	0.14%	5.54%	0.18%	0.41%	0.60%	0.83%	0.76%	8.46%
GPS	0.03%	0.04%	0.64%	0.43%	0.08%	0.46%	0.16%	1.84%
SPS	0.27%	0.56%	1.05%	5.59%	0.26%	0.54%	1.82%	10.09%
SVP	0.14%	1.07%	0.38%	0.08%	8.75%	1.45%	4.46%	16.33%
Übrige	0.21%	0.27%	0.21%	0.57%	0.42%	4.13%	0.67%	6.47%
Nichtwähler	0.38%	0.41%	0.40%	1.55%	1.95%	1.73%	48.03%	54.47%
<b>Total 1995</b>	2.10%	7.79%	2.81%	9.93%	10.96%	9.45%	56.97%	100.00%

Im Vergleich dazu lautet die Wanderungstabelle, welche man ohne Unterscheidung nach Bezirken für den Kanton ohne Stadt Zürich erhält:

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	1.35%	0.15%	0.00%	0.05%	0.03%	0.00%	0.77%	2.35%
FDP	0.00%	7.17%	0.00%	0.00%	0.55%	0.00%	0.73%	8.46%
GPS	0.00%	0.00%	1.00%	0.32%	0.00%	0.50%	0.02%	1.84%
SPS	0.31%	0.35%	1.34%	7.39%	0.00%	0.56%	0.13%	10.08%
SVP	0.00%	0.60%	0.01%	0.00%	9.90%	1.70%	4.14%	16.34%
Übrige	0.20%	0.00%	0.55%	0.18%	0.00%	5.55%	0.00%	6.47%
Nichtwähler	0.23%	0.00%	0.00%	1.06%	1.77%	0.98%	50.42%	54.47%
<b>Total 1995</b>	2.10%	7.79%	2.81%	9.93%	10.96%	9.45%	56.97%	100.00%

<sup>10</sup> Das primäre Ziel unserer Berechnungen ist nicht die Ermittlung der Wählerströme im Kanton Zürich, sondern es geht darum, die Anwendung der präsentierten Methoden der ökologischen Inferenz exemplarisch zu demonstrieren.

## 4.4 Kleinere Parteien

Nehmen kleinere Parteien an den Wahlen teil, deren gemeinsamer Stimmenanteil nur wenige Prozente der Wählerstimmen ausmacht, so stellt sich die Frage, ob und wie diese Parteien bei der Analyse berücksichtigt werden sollen.

Die übliche Praxis in Wahlanalysen besteht darin, kleinere Parteien zusammenzufassen, und zwar entweder in eine gemeinsame Gruppe „Übrige“, oder in Gruppen mit Parteien ähnlicher politischer Ausrichtung. Da ein statistisches Modell im Allgemeinen nicht mehr zu schätzende Parameter enthalten sollte als die unmittelbar interessierenden, empfiehlt sich dieses Vorgehen insbesondere dann, wenn zahlreiche kleine und kleinste Parteien vorhanden sind, deren Unterscheidung bei der Analyse nicht von Interesse ist.

## 4.5 Betrachtung von Wanderungssalden anstelle entgegengesetzter Wanderungen

Ist man nicht an der vollen Information einer Wanderungstabelle interessiert, sondern möchte den Effekt der direkten Wanderungen zwischen jeweils zwei Parteien „unter dem Strich“ kennen, so ist es sinnvoll, Wanderungssalden zu betrachten. Nicht die Schätzung der vollen Tabelle wird in diesem Fall angestrebt, sondern nur diejenige der Differenzen von jeweils zwei „entgegengesetzten“ Zellen.

Voraussetzung für eine sinnvolle Saldierung ist, dass in beiden Wahlen die gleichen Parteien betrachtet werden. Insbesondere muss also  $P = Q$  gelten.

Betrachten wir beispielsweise die folgende Wanderungstabelle (es handelt sich um die mit Regression ermittelte Tabelle von Seite 54):

Wahl 1999	Wahl 1995							Total 1999
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler	
CVP	1.31%	0.16%	0.00%	0.00%	0.07%	0.00%	0.77%	2.31%
FDP	0.00%	6.94%	0.00%	0.11%	0.40%	0.00%	0.60%	8.04%
GPS	0.00%	0.00%	1.23%	0.22%	0.00%	0.37%	0.04%	1.87%
SPS	0.00%	0.06%	0.49%	8.98%	0.00%	2.02%	0.00%	11.55%
SVP	0.09%	0.63%	0.28%	0.00%	8.53%	1.10%	4.03%	14.66%
Übrige	0.15%	0.00%	0.71%	0.35%	0.00%	5.47%	0.00%	6.68%
Nichtwähler	0.55%	0.00%	0.10%	0.27%	1.96%	0.48%	51.53%	54.89%
<b>Total 1995</b>	<b>2.10%</b>	<b>7.79%</b>	<b>2.81%</b>	<b>9.93%</b>	<b>10.96%</b>	<b>9.45%</b>	<b>56.97%</b>	<b>100.00%</b>

Im Folgenden wird der Eintrag in der  $q$ -ten Zeile und  $p$ -ten Spalte dieser Tabelle als  $w_{pq}$  bezeichnet. Eine Saldierung dieser Tabelle besteht darin, dass man jeweils zwei entgegengesetzte Werte  $w_{pq}$  und  $w_{qp}$  betrachtet, und von beiden das Minimum  $\min(w_{pq}, w_{qp})$  subtrahiert. Der resultierende (absolute) Saldo lautet demnach  $s_{pq} = w_{pq} - \min(w_{pq}, w_{qp})$ . Für den grösseren der beiden Werte gilt somit  $s_{pq} = w_{pq} - w_{qp}$  und für den kleineren  $s_{pq} = 0$ . In den Diagonalen kann keine Saldierung erfolgen, weshalb wir in der folgenden Tabelle die Einträge unverändert belassen. In Zellen  $(p, q)$  mit  $w_{pq} < w_{qp}$  wurde der Wert 0 durch einen Punkt ersetzt.

Wahl 1999	Wahl 1995						
	CVP	FDP	GPS	SPS	SVP	Übrige	Nichtwähler
CVP	1.31%	0.16%	0.00%	0.00%	.	.	0.22%
FDP	.	6.94%	0.00%	0.04%	.	0.00%	0.60%
GPS	0.00%	0.00%	1.23%	.	.	.	.
SPS	0.00%	.	0.27%	8.98%	0.00%	1.67%	.
SVP	0.02%	0.24%	0.28%	0.00%	8.53%	1.10%	2.07%
Übrige	0.15%	.	0.34%	.	.	5.47%	.
Nichtwähler	.	.	0.06%	0.27%	.	0.48%	51.53%

Da die Einträge neben der Hauptdiagonalen um den Wert  $\min(w_{pq}, w_{qp})$  reduziert wurden, lautet die Summe aller Innenfelder nicht mehr 1, sondern

$$1 - \sum_{p=1}^P \sum_{\substack{q=1 \\ q \neq p}}^P \min(w_{pq}, w_{qp}).$$

Die Summe

$$\sum_{p=1}^P \sum_{\substack{q=1 \\ q \neq p}}^P \min(w_{pq}, w_{qp})$$

– nennen wir sie die Summe der „neutralisierten Wechselwähler“ – beträgt in unserem Beispiel 8.06%. Die Zeilen- bzw. Spaltentotale entsprechen nicht mehr den Wähleranteilen, sondern dem betreffenden Wähleranteil minus den Anteil der „neutralisierten Wechselwähler“ aus allen Kombinationen der betreffenden Partei mit einer anderen.

Natürlich bedeutet die Saldierung einer Tabelle immer einen Informationsverlust. Bezüglich ihres Informationsgehalts kann eine saldierte Wanderungstabelle als eine Zwischenstufe zwischen den üblicherweise bekannten Randhäufigkeiten und der vollständigen Wanderungstabelle betrachtet werden.



## 5 Wie kann die Panaschierstatistik genutzt werden?

In diesem Abschnitt soll der in 1.1.1 gestellten Frage nachgegangen werden, wie die Information der Panaschierdaten sinnvoll in den vorgestellten Verfahren berücksichtigt werden kann, um die Qualität der Rekonstruktion von Wählerströmen zu verbessern. Kann die Information, welche Kandidaten wie oft auf welchen Listen panaschiert wurden, näheren Aufschluss geben über die politischen Präferenzen der Wähler bzw. ihrem Verhältnis zu den Parteien?

Die Information der Panaschierstatistik ist auf geeignete Weise mittels Kovariablen in der ökologischen Inferenz zu berücksichtigen. Die Auswahl an entsprechenden Methoden, bei welchen der Einfluss von Kovariablen einbezogen werden kann, ist beschränkt. Als einzige der besprochenen Verfahren bietet sich im Mehrparteienfall der Regressionsansatz an.

### 5.1 Konstruktion von Kovariablen aus den Panaschierdaten

Die Panaschierdaten liegen in der Form „Anzahl Panaschierstimmen für Partei ... auf Liste ... in Wahlkreis (=Gemeinde) ...“ vor. Bevor sie für unsere Zwecke genutzt werden können, müssen sie aufgearbeitet werden, um eine sinnvolle numerische Angabe auf Ebene der Wahlkreise zu gewinnen über die Tendenz der Wähler von Partei A, Kandidaten von Partei B zu panaschieren. Diese sollen dann als Kovariablen im Modell der Wählerstromanalyse berücksichtigt werden.

Einige Vorschläge zur Quantifizierung von „Parteiaffinitäten“ und „Parteidisziplin“ sind Burger (2001) zu entnehmen. Als geeignetes Mass für die „Parteiaffinitäten“ der A-Wähler gegenüber einer Partei B kann dasjenige angesehen werden, welches seine Tabelle 7 (S.25) exemplarisch für die Resultate der Nationalratswahlen 1999 im Kanton Waadt zeigt: deren Einträge lauten

$$\frac{\text{Anzahl Panaschierstimmen, die eine Partei A an eine Partei B abgegeben hat}}{\text{Zahl der Wahlzettel für die Partei A} \cdot \text{Zahl der Kandidaten der Partei B}} \cdot 1000.$$

Dies ist ein Mass für die Tendenz der A-Wähler, Kandidaten der Partei B zu panaschieren. Je höher dieser Wert, um so mehr wurde panaschiert.

Als Mass für die „Parteidisziplin“ schlägt Burger den folgenden Quotienten vor („Parteidisziplin II“ in Burger 2001, S. 21):<sup>11</sup>

$$\frac{\text{Anzahl der auf Wahlzetteln einer Partei A abgegebenen Panaschierstimmen}}{\text{Anzahl Wahlzettel für Partei A} \cdot \text{Anzahl parteifremder Kandidaten}} \cdot 1000.$$

Ein tiefer Wert dokumentiert hier grosse Parteitreuere der Wähler von Partei A.

Liegen Panaschierdaten aus beiden Wahlen vor, so muss zwischen der Verwendung von Panaschierdaten aus der ersten oder zweiten Wahl (oder aus beiden) entschieden werden. Eine Möglichkeit besteht darin, die Parteiaffinitäten nach obigem Muster für beide Wahlen zu bestimmen

<sup>11</sup> Die Parteidisziplin kann auch anders definiert werden, was aber keinen Einfluss auf die folgenden Überlegungen hat:

$$\frac{\text{Anzahl der auf Wahlzetteln einer Partei A abgegebenen Panaschierstimmen}}{\text{Anzahl Wahlzettel für Partei A} \cdot \text{Anzahl Sitze pro Wahlkreis}} \cdot 1000$$

und deren Mittelwert als Kovariablen zu wählen. Im Sinne einer anschaulichen Modellinterpretation könnte zugunsten der Panaschierdaten der ersten Wahl argumentiert werden, diese eigneten sich am ehesten als Bestimmungsgrößen für die Übergangswahrscheinlichkeiten von A zu B, da sie die Bereitschaft der A-Wähler der ersten Wahl, ihre Stimme in einer späteren Wahl der Partei B zu geben, dokumentieren.

## 5.2 Berücksichtigung von Kovariablen im Regressionsmodell mit mehreren Parteien

Das Modell der ökologischen Regression im Mehrparteienfall ohne Kovariablen wurde in Abschnitt 3.3 besprochen. Im erweiterten Modell werden die Übergangswahrscheinlichkeiten  $p_{(p,q)j}$  als lineare Funktionen von  $k$  Kovariablen betrachtet, welche nicht für alle Übergangswahrscheinlichkeiten identisch sein müssen:

$$p_{(p,q)j} = \gamma_{(p,q)} + \delta_{(p,q)1} z_{(p,q)1j} + \dots + \delta_{(p,q)k} z_{(p,q)kj}.$$

Durch Einsetzen in die (Haupt-)Regressionsgleichungen mit Zielgrösse  $Y_{qj}$ ,

$$Y_{qj} = p_{(1,q)j} x_{1j} + p_{(2,q)j} x_{2j} + \dots + p_{(P,q)j} x_{Pj} + U_{qj} \quad (q = 1, \dots, Q-1),$$

ergeben sich Regressionen mit den folgenden Einflussgrößen:

- die Stimmenanteile der ersten Wahl  $x_1, \dots, x_P$  (mit Koeffizienten  $\gamma_{(1,q)}, \dots, \gamma_{(P,q)}$ ),
- sämtliche  $P \cdot k$  Produkte von der Form  $x_p \cdot z_{(p,q)l}$  (mit Koeffizienten  $\delta_{(p,q)l}$ ,  $p = 1, \dots, P$ ;  $l = 1, \dots, k$ ).

Bei der letzten Regressionsgleichung mit  $q = Q$  handelt es sich im Gegensatz zum Fall ohne Kovariablen nicht um eine Linearkombination der ersten  $Q-1$  Regressionen, weshalb die Summe der geschätzten Übergangswahrscheinlichkeiten  $\hat{p}_{(p,q)j}$ ,  $q = 1, \dots, Q$ , nicht mehr 1 ergeben muss.

Insgesamt sind in jeder der  $Q$  Regressionsgleichungen  $(k+1)P$  Parameter zu schätzen. Eine Verallgemeinerung dieses Modells mit unterschiedlicher Anzahl Kovariablen für die verschiedenen Übergangswahrscheinlichkeiten kann analog beschrieben werden.

Es ist klar, dass die Anzahl  $k$  der Kovariablen nicht allzu hoch angesetzt werden darf, um die Zahl der Regressoren im Modell im Rahmen zu halten. Ausserdem wird ein Modell angestrebt, welches einigermaßen interpretierbar sein soll.

## 5.3 Ein Modellvorschlag

Wir nehmen an, es gelte  $P = Q$  und der Index  $P$  entspreche in beiden Wahlen der Wahlabstimmungsart. Ist  $p, q < P$ , so wird die Übergangswahrscheinlichkeiten  $p_{(p,q)}$  in Abhängigkeit einer Kovariablen  $z_{(p,q)}$  betrachtet.  $z_{(p,q)}$  ist im Fall  $p \neq q$  die entsprechende Parteiaffinität aus der ersten Wahl, also die Tendenz der  $p$ -Wähler, Kandidaten der Partei  $q$  zu panaschieren. Im Fall  $p = q$  wird als Kovariable  $z_{(p,p)}$  die Parteidisziplin der Wähler von Partei  $p$  gewählt. Entspricht entwe-

der  $p$  oder  $q$  der Wahlalternative  $P$  (Nichtwahl), so kann keine solche Kovariable bestimmt werden.

Mit den Modellannahmen

$$\begin{aligned} p_{(p,q)j} &= \gamma_{(p,q)} + \delta_{(p,q)} z_{(p,q)j} && \text{falls } 1 \leq p, q < P, \\ p_{(p,q)j} &= \gamma_{(p,q)} && \text{falls } p = P \text{ oder } q = P \end{aligned}$$

sowie den üblichen Voraussetzungen über die Verteilung der  $U_{qj}$  ergeben sich damit die Regressionsgleichungen

$$\begin{aligned} Y_{qj} &= (\gamma_{(1,q)} + \delta_{(1,q)} z_{(1,q)j}) x_{1j} + \dots + (\gamma_{(P-1,q)} + \delta_{(P-1,q)} z_{(P-1,q)j}) x_{P-1j} + \gamma_{P,q} x_{Pj} + U_{qj} \\ &\text{für } q = 1, \dots, P-1 \text{ sowie} \\ Y_{Pj} &= \gamma_{(1,P)} x_{1j} + \dots + \gamma_{P,P} x_{Pj} + U_{Pj} \\ &\text{für } q = P. \end{aligned}$$

### **Bemerkungen:**

- Das Modell gründet auf den folgenden Annahmen: Eine hohe Parteiaffinität der A-Wähler in Wahl 1 gegenüber Partei B dokumentiert eine hohe Bereitschaft, in einer späteren Wahl B zu wählen. Analog wird angenommen, dass die Wahrscheinlichkeit, dass A-Wähler aus Wahl 1 in der zweiten Wahl sich wieder für A entscheiden, um so grösser ist, je weniger parteifremde Kandidaten sie (in Wahl 1) auf ihre Listen setzen.
- Dieses Modell besitzt sämtliche früher festgehaltenen Nachteile des ökologischen Regressionsansatzes und sollte folglich nur mit grosser Skepsis angewandt werden. Bei einer sinnvollen Modellschätzung sollten die geschätzten Übergangswahrscheinlichkeiten grösstenteils zwischen 0 und 1 liegen. Falls die Kovariablen den erhofften Effekt erzielen, werden sich die geschätzten Übergangswahrscheinlichkeiten von Gemeinde zu Gemeinde sichtbar unterscheiden.
- Ist die Zahl der Kandidierenden jeder Partei in allen Wahlkreis identisch, so hat die Division durch die Anzahl Kandidierender in der Formel für Parteiaffinität bzw. die Division durch die Anzahl parteifremder Kandidierender in der Formel für Parteidisziplin keine Auswirkungen auf die geschätzten Wählerwanderungen.

### **5.3.1 Anwendung auf die Nationalratswahlen im Kanton Zürich 1995/99**

Das oben beschriebene Regressionsmodell wurde mit den Daten der Nationalratswahlen 1995 und 1999 im Kanton Zürich angewendet. Der Anteil der geschätzten Übergangswahrscheinlichkeiten ausserhalb des Einheitsintervalls beträgt 45%. Die geschätzten Übergangswahrscheinlichkeiten sind ausserdem in allen Gemeinden nahezu identisch: die entsprechende Standardabweichung (berechnet aus den geschätzten Übergangswahrscheinlichkeiten jeweils zweier Parteien) beträgt maximal 1.66%, in den meisten Fällen liegt sie unter 0.1%. Damit erfüllt das Modell seinen Zweck nicht. Der Grund hierfür muss nicht unbedingt lauten, dass kein Zusammenhang zwischen der Panaschierstatistik und den Wählerwanderungen besteht. Falls aber ein solcher besteht, so ist er nicht von der Gestalt, welche in unserem Modell angenommen wird.

Auf die nötigen Anpassungsschritte zur Bestimmung einer sinnvollen Wanderungstabelle wie in 3.3.1 wird an dieser Stelle verzichtet.

## 6 Publiizierte Wählerstromanalysen im deutschsprachigen Raum

### 6.1 Anwendung des Regressionsmodells

Das SORA (Institute for Social Research and Analysis) in Wien stellt auf seiner Internetseite die Resultate verschiedener Wählerstromanalysen in Österreich vor. Der kurzen Methodenbeschreibung unter <http://www.sora.at/wahlen/wsa/> ist zu entnehmen, dass als Methode die multiple Regression verwendet wird:

„Die Gleichung für eine Wählerstromanalyse von der Nationalratswahl 1995 zur Nationalratswahl 1999 sähe für die ÖVP 1999 so aus:

$$\text{ÖVP99} = b_1 \times \text{SPÖ95} + b_2 \times \text{ÖVP95} + b_3 \times \text{FPÖ95} + b_4 \times \text{LIF95} + b_5 \times \text{Grüne95} + b_6 \times \text{Sonstige95} + b_7 \times \text{Nichtwähler95}“$$

### 6.2 Anwendungen von Thomsens Methode

Die folgenden Wahlanalysen machen von Thomsens Modell Gebrauch:

- Statistischer Infodienst Freiburg im Breisgau, 2001: „Schätzung der Freiburger Wählerwanderung zwischen den Landtagswahlen 1996/2001“, herausgegeben vom Amt für Statistik und Einwohnerwesen.

In diesem Dokument wird auf die Unsicherheit der präsentierten Resultate hingewiesen: „Allerdings ist deutlich darauf hinzuweisen, dass die im folgenden berichteten Zahlen lediglich Schätzwerte sind.“

- agis (Arbeitsgruppe interdisziplinäre Sozialstrukturforschung, Universität Hannover) und Landeshauptstadt Hannover, 2002: „Hannover hatte die Wahl. Ergebnisse und Analysen zur Bundestagswahl vom 22. September 2002“. (Internet: [http://www.agis.uni-hannover.de/wahlforschung/btw02/Wahlbericht\\_btw02\\_start.htm](http://www.agis.uni-hannover.de/wahlforschung/btw02/Wahlbericht_btw02_start.htm)). Der für uns relevante Teil befindet sich in Kapitel 7).

Hier wird Thomsen (1987) ohne weiteren Kommentar zur Methodik als Quelle zitiert.

### 6.3 Kohlsche: Eigene Methode mit Elementen aus Thomsen und Regression

Andreas J. Kohlsche vom Institut für Wahl-, Sozial- und Methodenforschung in Kaufbeuren (Deutschland) hat hauptsächlich im deutschsprachigen Raum in verschiedensten Tageszeitungen Resultate von Wählerwanderungsrekonstruktionen publiziert.

Eine vollständige Beschreibung von Kohlsches Methode existiert nicht. Im Folgenden wird versucht, die wesentlichen Elemente seiner Methode aufgrund zweier Dokumente (Kohlsche 1998 und 2002) nachzuvollziehen; auf Details wird verzichtet.

### 6.3.1 Kohlsches Verfahren

Kohlsches Verfahren umfasst zwei Hauptschritte. Im ersten Schritt wird mit einer Modifikation von Thomsons Methode der Anteil der Stammwähler geschätzt, d.h. der Anteil derjenigen Wähler, die in beiden Wahlen für dieselbe Partei stimmten. Im zweiten Schritt werden die Wanderungssalden zwischen den Parteien geschätzt, d.h. die Differenzen der Wählerwanderungen beider Richtungen zwischen zwei Parteien. In beiden Schritten erfolgt die Rechnung separat innerhalb von Clustern von Wahlkreisen, welche durch die Optimierung eines numerischen Kriteriums gebildet werden.

1. Im ersten Schritt wird die Anzahl der Stammwähler (vorläufig) geschätzt, indem für jede Partei Thomsons Idee folgend der Pearson-Korrelationskoeffizient der Wähleranteile beider Wahlen berechnet und mit Yule's Q aus der entsprechenden (2x2)-Tabelle der Form

Wahl 2 Partei $p$	Wahl 1 Partei $p$ Rest (inkl. Nichtwähler)		$y_{pj}$
	$a_{pj}$	$y_{pj} - a_{pj}$	
Rest (inkl. Nichtwähler)	$x_{pj} - a_{pj}$	$1 - x_{pj} - y_{pj} - a_{pj}$	$1 - y_{pj}$
	$x_{pj}$	$1 - x_{pj}$	1

gleichgesetzt wird (Kohlsche verwendet die untransformierten Wähleranteile  $X_{pj}$  und  $Y_{qj}$  anstatt wie Thomsen die transformierten  $\logit(X_{pj})$  und  $\logit(X_{qj})$ ; die Transformation sei „absolut überflüssig“, da sich die Resultate kaum unterscheiden; 1998, S. 2). Daraus ergibt sich durch Auflösen nach  $a_j$  für jeden Wahlkreis der Anteil der Stammwähler der betreffenden Partei. Diese Berechnungen erfolgen separat für eine Auswahl von Clustern, welche durch Iteration so gebildet werden, dass die gesamte Anzahl der Stammwähler maximal wird (Kohlsches Begründung: „Wirft man alle Teilgebiete in einen Topf, dann erhält man zu wenig Stammwähler“; 2002, Abschnitt a). Beim verwendeten iterativen Clusterverfahren werden in jedem Schritt die Cluster angepasst und die Stammwählerschätzungen aktualisiert.

Für den zweiten Schritt müssen die beteiligten Parteien in sogenannte „Lager“ eingeteilt werden, innerhalb welcher Parteien „mit ähnlicher Programmatik“ liegen.

2. Die in Schritt 1 ermittelten Stammwähler werden von den entsprechenden Randtotalen subtrahiert. Mittels multipler Regression der Wähleranteile jeder Partei in der zweiten Wahl (Zielgrösse) in Abhängigkeit der Wähleranteile aller anderen Parteien bei der ersten Wahl werden dann die übrigen Wählerwanderungen geschätzt. Die Regressionen erfolgen in beide Richtungen, d.h. einmal mit  $X$  und einmal mit  $Y$  als Zielgrösse. Kohlsche: „Die geeignete Regressionsrichtung ergibt sich in den allermeisten Fällen von selbst: Bei der falschen Richtung bewegt sich der Koeffizient ausserhalb des zulässigen Bereichs von 0 bis 1“ (2002, Abschnitt b). Es wird also separat für jeden einzelnen Parameter diejenige Richtung gewählt, welche eine zulässige Lösung ergibt. Sind beide Lösungen zulässig, „dann wird der Mittelwert [der beiden Lösungen] gebildet“ (1998, S. 4). Dabei wird wieder eine Clustereinteilung vorgenommen mit der Eigenschaft, dass „der Anteil der Wanderungssalden innerhalb der Lager an den Wanderungssalden innerhalb der Lager und zwischen den Lagern im Wahlgebiet maximal ausfällt“, denn: „Wirft man alle Teilgebiete in einen Topf, dann werden die Lager zu wenig getrennt.“ (2002, Abschnitt b). Zu diesem Zweck wird ein

iteratives Verfahren verwendet, bei welchem in jedem Schritt die Cluster aktualisiert und die Regressionen neu gerechnet werden.

Mit einem iterativen Verfahren werden schliesslich die Einträge der geschätzten Wanderungstabelle so angepasst, dass sie in Übereinstimmung mit den bekannten Randhäufigkeiten stehen. Zwischen je zwei Parteien werden schliesslich Wanderungssalden berechnet als Differenz der ermittelten Wanderungen zwischen den beiden Parteien. Obwohl Kohlsche vollständige Wanderungstabellen berechnet, publiziert er nur Angaben über Wanderungssalden. Dies wird damit begründet, dass es „definitiv nicht möglich ist, das komplette Wanderungstableau zu berechnen, sondern ausschliesslich Wanderungssalden“ (2002, Abschnitt b).

Für die Anzahl Cluster in den Schritten 1 und 2 hat sich laut Kohlsche ein Wert von 5 Clustern in der Praxis bewährt. Die Clusterbildung in den beiden Schritten erfolgt unabhängig, d.h. die beiden Einteilungen sind verschieden.

### 6.3.2 Diskussion

Offenbar geht Kohlsche von der Beobachtung aus, dass bei der Verwendung bekannter Modelle (Thomsen, Regression)

- die Stammwähleranteile sowie
- Wanderungen zwischen sich politisch nahestehenden Parteien

unterschätzt werden. Diesen Feststellungen könnte nun Rechnung getragen werden, indem ein konkretes Modell formuliert wird, welches diese Tendenzen berücksichtigt. Kohlsches Lösungsansatz besteht hingegen darin, die Analyse mit bestehenden, auf unterschiedlichen Modellannahmen basierenden Methoden (Thomsens Modell in Schritt 1, Regression in Schritt 2) separat für Cluster von Wahlgebieten vorzunehmen, welche so gebildet werden, dass (in Schritt 1) die Zahl der Stammwähler bzw. (in Schritt 2) die Wanderungen zwischen sich politisch nahestehenden Parteien maximal werden. Zweifellos ergeben sich auf diese Weise höhere geschätzte Stammwähleranteile und stärkere Wanderungen zwischen Parteien ähnlicher Ausrichtung. Es ist aber nicht einzusehen, wieso *das Ausmass* dieser „Korrektur“ mit einem solchen Vorgehen sinnvoll geschätzt werden sollte. Kohlsche argumentiert, dass seine Resultate sich besser mit Umfragedaten decken als diejenigen aller anderen Methoden (Mündliche Mitteilung März 2003). Der Tatsache, dass die Resultate nicht auf einem konkret spezifizierten Modell und einer auf objektiven Kriterien basierenden Schätzung beruhen, misst er keine grosse Bedeutung bei. Kohlsche: „Zudem ergibt sich höchst überraschend, dass nicht das in der mathematischen Statistik übliche Verfahren, den Fit eines Modells zu optimieren, zum Erfolg führt, sondern ein inhaltlich orientiertes Kriterium“ (1998, S. 3). Mit Kohlsches „inhaltlich orientierten Kriterien“ werden die Resultate gezielt in die Richtung gelenkt, in welcher die Lösung vermutet wird.

Aus theoretischer Sicht fragwürdig ist zudem die Art und Weise, wie Kohlsche in Schritt 2 zwei auf unterschiedlichen Annahmen basierende (entgegengesetzte) Regressionsmodelle berechnet und in Abhängigkeit der Resultate die Ergebnisse des einen oder anderen Modells weiterverarbeitet. Auf diesen Aspekt soll hier nicht weiter eingegangen werden. Stattdessen soll die Frage nach Sinn und Zweck konkreter Modelle im vorliegenden Zusammenhang diskutiert werden.

Die Aufgabe der ökologischen Inferenz besteht darin, aus dem Input „Wahldaten in der Form Parteistärke nach Wahlkreis“ mittels eines geeigneten Verfahrens einen Output in Form einer Wanderungstabelle zu generieren. Wie in Abschnitt 2.2.1 gesagt wurde, liegt es in der Natur der ökologische Inferenz, dass die Berechnungen *unter gewissen Annahmen* erfolgen. Die Anwen-

dung eines Verfahrens wird dann erfolgreich sein, wenn das Verfahren auf die Eigenschaften des realen, aber unbekannten Wählerverhaltens abgestimmt ist.

Gehen wir nun davon aus, dass Kohlsches Vorgehen bei Wählerwanderungen zu „guten“ Resultaten im Sinn von „nahe an der Realität“ führt. Dann hat er ein Verfahren gefunden, das der üblichen Struktur des Wählerverhaltens gerecht wird. Dann muss es aber auch möglich sein, die entsprechenden Bedingungen anzugeben, welche erfüllt sein müssen, damit das Verfahren erfolgreich ist. Sind diese Bedingungen einmal bekannt, so eröffnet dies für Forschungsarbeiten in diesem Bereich neue Perspektiven:

- Eine systematische Überprüfung und Beurteilung des Verfahrens wird möglich.
- Allfällige Gefahren bei der Anwendung des Verfahrens werden ersichtlich, d.h. Konstellationen der Wählerströme, in denen das Verfahren zu verfälschten Ergebnissen führt.
- Eine Überarbeitung des Verfahrens, welche sich an objektiven Optimalitätskriterien orientiert, wird möglich.

Solange jedoch keine transparente Basis in Form von Bedingungen, eines klar spezifizierten Modells und der Anwendung objektiver Kriterien bei der Schätzung ersichtlich ist, ist die Methode mit einer bestimmten Willkür behaftet. Von einem ausgereiften wissenschaftlichen Instrument kann aus mathematisch-statistischer Sicht keine Rede sein.

Zusammenfassend kann festgehalten werden, dass Kohlsches Wählerstromanalyse den in 1.1.2 formulierten Anforderungen in keiner Weise gerecht wird. Es besteht keine theoretische Grundlage in Form eines mathematisch-statistischen Modells, und nichts ist über die Bedingungen bekannt, unter welchen das Verfahren erfolgreich eingesetzt werden kann.

### **6.3.3 Nationalratswahlen im Kanton Zürich 1995/99: Kohlsches Resultate**

Die Resultate von Kohlsches Wählerwanderungsanalyse sind auf Seite 86 zu finden (Quelle: Internetseite des Statistischen Amtes des Kantons Zürich: <http://www.statistik.zh.ch/themen/b17/nrw99/wwzh.html>).

Es fällt auf, dass sich die Tabelleneinträge zu 100% summieren, was in einer saldierten Tabelle üblicherweise nicht zu erwarten ist (vgl. Abschnitt 4.5). Da es sich bei den Tabelleneinträgen neben der Diagonalen um Wechselwähleranteile in Prozent der Wahlberechtigten handelt (dies ist aus der unteren Tabelle ersichtlich), können die Zahlen auf der Diagonalen nicht Stammwähleranteile sein. Die Spalten- und Zeilentotale entsprechen (ungefähr) den tatsächlichen Wähleranteilen in den beiden Wahlen, so dass die Zahlen in der Diagonalen als „Stammwähleranteile zuzüglich der neutralisierten Wechselwähler“<sup>12</sup> aus allen Kombinationen der betreffenden Partei mit einer anderen aufzufassen sind. (Einzig im Spezialfall, in dem es keine „neutralisierten Wechselwähler“ gibt, handelt es sich bei den Diagonalelementen in Kohlsches Tabelle um Stammwähleranteile. Genau dann ergibt ausserdem die Summe aller saldierten Tabelleneinträge 100%.) Eine Erklärung der Zahlen auf der Diagonalen ist nicht zu finden, so dass diese leicht als Stammwähleranteile fehlinterpretiert werden könnten. Die geschätzten Stammwähleranteile hingegen sind aus der Tabelle nicht herauszulesen.

Sollen nun Kohlsches Resultate mit unseren Resultaten aus Berechnungen mit anderen Methoden verglichen werden, so müssen wir einerseits die mit anderen Methoden ermittelten Wandertabellen saldieren. Aus der Tabelle Seite 86 werden die Diagonalelemente entfernt, da es sich nicht um die Stammwähleranteile handelt. Ausserdem bleiben die saldierten Wechselwähleranteile von und zu den „Übrigen“ unbekannt, da unsere Gruppierung der Parteien sich von derje-

<sup>12</sup> Zu den „neutralisierten Wechselwählern“ vgl. Abschnitt 4.5.

nigen Kohlsches unterscheidet. Die damit verbleibenden saldierten Wählerwanderungen, die bei einem Vergleich verwendet werden können, sind der folgenden Tabelle zu entnehmen.

Wahl 1999	Wahl 1995					
	CVP	FDP	SPS	SVP	Übrige	Nichtwähler
CVP	?	0.08%	0.01%	.	?	0.30%
FDP	.	?	0.02%	0.06%	?	.
SPS	.	.	?	.	?	0.79%
SVP	0.29%	.	0.05%	?	?	1.52%
Übrige	?	?	?	?	?	?
Nichtwähler	.	0.13%	.	.	?	?

NRW99-Wählerwanderungen: Kanton Zürich

Wanderungssalden N 95 -> N 99 in % der Wahlberechtigten  
Wanderungsrichtung verläuft von der Partei oben zur Partei links

	Nicht	Ungült	Rest	SPS	Linke	FDP	Ö/P	Mitte	SVP	Rechte	N 99
Nicht	54,01	0,07	.	.	0,04	0,13	.	0,40	.	0,09	54,73
Ungült	.	0,42	.	0,01	.	.	0,01	.	.	.	0,44
Rest	0,36	0,01	0,22	0,07	0,43	.	.	0,14	0,13	.	1,36
SPS	0,79	.	.	9,59	0,38	.	.	0,46	.	0,26	11,49
Linke	.	0,07	.	.	2,58	.	.	0,02	.	.	2,66
FDP	.	0,14	0,08	0,02	0,10	7,41	.	.	0,06	0,18	7,99
CVP	0,30	.	0,04	0,01	0,10	0,08	1,69	.	.	0,08	2,30
Mitte	.	0,02	.	.	.	0,02	0,06	2,52	.	.	2,63
SVP	1,52	.	.	0,05	0,29	.	0,29	0,50	10,57	1,34	14,56
Rechte	.	0,04	0,03	.	.	.	.	0,06	.	1,72	1,85
N 95	56,97	0,77	0,37	9,75	3,91	7,65	2,06	4,09	10,76	3,68	100,00

Die wichtigsten Wanderungssalden für die Parteien

Partei	Partei			Schweiz	Abweichung
N 95	-> N 99	Stimmen	% der Wahlberechtigten	% der Wahlberechtigten	% der Wahlberechtigten
Nicht	-> SVP	11'800	1,52	1,25	0,26
Rechte	-> SVP	10'400	1,34	1,28	0,06
Nicht	-> SPS	6'200	0,79	0,29	0,50
Mitte	-> SVP	3'900	0,50	0,25	0,26
Mitte	-> SPS	3'600	0,46	0,13	0,33
Linke	-> Rest	3'400	0,43	0,18	0,25
Mitte	-> Nicht	3'100	0,40	0,31	0,09
Linke	-> SPS	2'900	0,38	0,08	0,30
Nicht	-> Rest	2'800	0,36	0,15	0,20
Nicht	-> CVP	2'300	0,30	0,27	0,03



## 7 Empfehlungen

In 1.1.1 wurde als Teil des Auftrags die Erarbeitung und Diskussion eines Vorschlags für ein Modell mit den Wahldaten des BFS verlangt.

Damit ein Verfahren zur routinemässigen Anwendung mit den Wahldaten des BFS empfohlen werden könnte, müssten aus unserer Sicht die folgenden Minimalanforderungen erfüllt sein:

- das Verfahren beruht auf einem klar definierten Modell und die Schätzung erfolgt aufgrund objektiver Kriterien,
- die getroffenen Annahmen sind plausibel,
- die Gültigkeit der Annahmen oder der Resultate ist in einigen (möglichst unterschiedlichen) Anwendungen überprüft worden (z.B. mit Umfragedaten).

Ob die Überprüfung der Resultate einer Wählerwanderungsanalyse zwischen zwei Wahlen im Abstand von vier Jahren überhaupt möglich ist, sei dahingestellt. Selbst wenn die genannten Anforderungen erfüllt sind, werden die Ergebnisse einer Wählerwanderungsanalyse eher spekulativen Charakter haben.

Aufgrund unserer Untersuchungen können wir keines der untersuchten Verfahren für die routinemässige Anwendung mit den Wahldaten des BFS empfehlen.

# Literaturverzeichnis

- Achen, C.H, Shively, W.P. (1995). Cross-Level Inference. University of Chicago Press, Chicago.
- Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Chapman and Hall, London.
- Brown, P.J., Payne, C.D. (1986). Aggregate Data, Ecological Regression and Voting Transitions. Journal of the American Statistical Association, 81, 452-460.
- Burger, R. (2001). Konzepte zur Analyse der Panaschierstatistik. Eine Studie mit Daten der Nationalratswahlen 1999. Bundesamt für Statistik, Neuchâtel.
- Cho, W.K.T. (1998). If the Assumptions Fit: A Comment in the King Ecological Inference Solution. Political Analysis, 7, 143-163.
- Deming, W.E. (1943). Statistical Adjustment of Data. John Wiley, New York.
- Dewille, J.-C., Särndal, C.-E., Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association 88, 1013-1020.
- Duncan, O.D., Davis, B. (1953). An Alternative to Ecological Correlation. American Sociological Review 18, 665-666.
- Freedman, D.A., Klein, S.P., Ostland, M., Roberts, M. (1998). On „Solutions“ to the Ecological Inference Problem. Journal of the American Statistical Association 93, 1518-1522.
- Goodman, L. (1953). Ecological Regressions and the Behavior of Individuals. American Sociological Review 18, 663-664.
- Goodman, L. (1959). Some Alternatives to Ecological Correlation. American Journal of Sociology 64, 610-624.
- Katz, J.N., King, G. (1999). A Statistical Model for Multiparty Electoral Data. American Political Science Review, 93, 15-32.
- King, G. (1997). A Solution to the Ecological Inference Problem. Reconstructing Individual Behavior from Aggregate Data. Princeton University Press, Princeton, New Jersey.
- King, G., Rosen, O., Tanner, M.A. (1999). Binomial-Beta Hierarchical Models for Ecological Inference. Sociological Methods & Research, 28, 61-90.
- Kohlsche, A.J. (1998). Die endgültige Lösung des Problems der ökologischen Inferenz. Vortragsmanuskript für die Tagung des Arbeitskreises „Wahlen und politische Einstellungen“ der DVPW, Mai 1998, Mainz.
- Kohlsche, A.J. (2002). Ökologische Inferenz I: Wählerwanderungen und Stimmensplitting (Internet: <http://www.wahlforschung.de/methodik.html>).
- Kotz/Johnson/Read (1988). Encyclopedia of Statistical Sciences, Volume 9. Wiley, New York.
- McCue, K.F. (2001). The Statistical Foundation of the EI Method. The American Statistician, 55, 106-110.
- Palmquist, B. (1999). Ecological Inference in Practice, April 15-17, 1999. Annual Meeting of the Midwest Political Sciences Association. Chicago, Ill.
- Robinson, W.S. (1950). Ecological Correlation and the Behavior of Individuals. American Sociological Review 15, 351-357.
- Rosen, O., Jiang, W., King, G., Tanner, M.A. (2001). Bayesian and frequentist inference for ecological inference: The  $R \times C$  Case. Statistica Neerlandica, 55, 134-156.
- Seitz, W. (2002). Nationalratswahlen 1999, Übersicht und Analyse. BFS, Neuchâtel.
- Thomsen, S.R. (1987). Danish Elections, 1920-79: A Logit Approach to Ecological Analysis and Inference. Politica, Aarhus.
- Thomsen, S.R. (2000). Issue Voting and Ecological Inference. Working Paper (Internet: <http://www.ps.au.dk/srt/Multi/thoms00.pdf>).

## Publikationsprogramm BFS

Das Bundesamt für Statistik (BFS) hat – als zentrale Statistikstelle des Bundes – die Aufgabe, statistische Informationen breiten Benutzerkreisen zur Verfügung zu stellen.

Die Verbreitung der statistischen Information geschieht gegliedert nach Fachbereichen (vgl. Umschlagseite 2) und mit verschiedenen Mitteln:

<i>Diffusionsmittel</i>	<i>Kontakt</i>	<i>Diffusionsmittel</i>	<i>Kontakt</i>
Individuelle Auskünfte	032 713 60 11 info@bfs.admin.ch	Publikationen zur vertieften Information (zum Teil auch als Diskette)	032 713 60 60 order@bfs.admin.ch
Das BFS im Internet	www.statistik.admin.ch	Online-Datenbank	032 713 60 86 www.statweb.admin.ch
Medienmitteilungen zur raschen Information der Öffentlichkeit über die neusten Ergebnisse	www.news-stat.admin.ch		

Nähere Angaben zu den verschiedenen Diffusionsmitteln liefert das laufend nachgeführte Publikationsverzeichnis im Internet unter der Adresse [www.statistik.admin.ch](http://www.statistik.admin.ch) >>News >>Neuerscheinungen.

## Politik

Ladner Andreas: *Kantonale Parteiensysteme im Wandel. Eine Studie mit Daten der Wahlen in den Nationalrat und in die kantonalen Parlamente 1971–2003*. Hrsg. BFS, Neuchâtel 2003, Bestell-Nr. 589-0300.

Armington Klaus: *Das Parteiensystem der Schweiz im internationalen Vergleich. Eine Studie mit Daten der Nationalratswahlen 1971–1999*. Hrsg. BFS, Neuchâtel 2003, Bestell-Nr. 586-9900.

*Nationalratswahlen 1999. Übersicht und Analyse*. Neuchâtel 2002, Bestell-Nr. 016-9904.

*Die Frauen bei den Nationalratswahlen 1999. Entwicklung seit 1971*. Neuchâtel 2000, Bestell-Nr. 016-9902.

*Nationalratswahlen 1999. Der Wandel der Parteienlandschaft seit 1971*. Neuchâtel 1999, Bestell-Nr. 016-9901.

*Nationalratswahlen 1999: Die «Voll- und Restmandate» der Parteien bei den Nationalratswahlen 1995 und die Entwicklung der Parteienlandschaft bei den kantonalen Parlamentswahlen (1996–1999)*. BFS-aktuell, Neuchâtel 1999.

Burger Rudolf: *Konzepte zur Analyse der Panaschierstatistik. Eine Studie mit Daten der Nationalratswahlen 1999*. Hrsg. BFS, Neuchâtel 2001, Bestell-Nr. 016-9903.

*Der lange Weg ins Parlament. Die Frauen bei den Nationalratswahlen von 1971 bis 1991. Im Anhang: 1) Frauen in den kantonalen Parlamenten (1961–1994), 2) Studie von Thanh-Huyen Ballmer-Cao/John Bendix über Determinanten der Frauenvertretung in den schweizerischen Legislativen*. Bern 1994, Bestell-Nr. 016-9102.

*Die Frauen in den Exekutiven der Schweizer Gemeinden 2001*. Neuchâtel 2001, Bestell-Nr. 221-0100.

*Die eidgenössischen Volksabstimmungen 1999*. Neuchâtel 2001, Bestell-Nr. 200-9900.

*Die eidgenössischen Volksabstimmungen 1998*. Neuchâtel 2000, Bestell-Nr. 200-9800.

---

Seit gut zwei Jahrzehnten ist die politische Beteiligung der Bürgerinnen und Bürger am Sinken, und die Bindungen an die Parteien nehmen ab. Bei der Analyse der Wahlergebnisse wird so neben der Frage nach den Per-Saldo-Gewinnen und -Verlusten der einzelnen Parteien auch die Frage nach der so genannten Wählerwanderung immer wichtiger. Diese Frage kann am besten mit der Befragung von repräsentativ ausgewählten Stimmbürgerinnen und Stimmbürgern beantwortet werden. Wissenschaftliche Meinungsumfragen sind jedoch relativ aufwändig und teuer und bleiben oft auf grössere räumliche Einheiten beschränkt; zudem sind sie für zeitlich zurückliegende Wahlen nicht mehr durchführbar.

Neuerdings beleben daher auch Modelle die politologische Diskussion, welche mittels ausgeklügelter statistischer Verfahren versuchen, aus Aggregatdaten Informationen über Wählerströme zu gewinnen. Die vorliegende Studie gibt einen kritischen Überblick über die Literatur zu den bekannten Methoden zur Rekonstruktion von Wählerströmen und versucht anhand ausgewählter Kriterien, die Solidität der Modelle einzuschätzen.