gesis Leibniz-Institut für Sozialwissenschaften

GESIS *Papers* 2016|02

Zur Messung der Komplexität von Sequenzen

Georgios Papastefanou

GESIS Papers 2016|02

Zur Messung der Komplexität von Sequenzen

Georgios Papastefanou

GESIS – Leibniz-Institut für Sozialwissenschaften 2016

GESIS Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Abteilung „Dauerbeobachtung der Gesellschaft“
German Microdata Lab
Postfach 12 21 55
68072 Mannheim
Telefon: 0621 / 1246 - 553
Telefax: 0621 / 1246 - 577
E-Mail: georgios.papastefanou@gesis.org

ISSN: 2364-3781 (Online)
Herausgeber,
Druck und Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

Inhalt

1 Einleitung

Sequenzen werden für die sozialwissenschaftlichen Forschung relevant, wenn Längsschnitt-Daten verfügbar sind. Und zwar dann, wenn relativ dichte Reihen von Messwiederholungen erfasst sind, sei es durch retrospektive Lebensverlaufs- oder Tagesrekonstruktionen (z.B. durch die Day Reconstruction Method), prozessproduzierte Datensammlungen oder durch Aktivitätsaufzeichnungen mittels Tagebuchaufzeichnungen bzw. experience sampling-Verfahren.

Diese Verfahren haben gemeinsam, dass Zustandsinformationen für eine relativ große Anzahl von Zeitpunkten wiederholt erfasst werden. In der Zeitbudgetforschung werden z.B. durch Tagebuchaufzeichnung über 24 Stunden hinweg in jeweils 10-Minuten-Intervallen die Aktivitäten zu 144 Zeitpunkten erfasst. Indem die alltäglichen Aktivitäten mit einem sehr detaillierten Schema klassifiziert werden, kommt dieses Messverfahren an die phänomenologische Gestalt der realen Lebenswelt als Aktivitätsfolgen recht nah heran. Der Lebensalltag wird als Abfolge von einzelnen Aktivitätstypen mit einer je eigenen zeitlichen Lagerung erfasst, also als Sequenz von verschiedenen Aktivitätsperioden mit je spezifischem Anfang und Ende.

Es wird argumentiert, dass das Muster der Abfolge der verschiedenen Alltags- bzw. Lebensaktivitäten ein holistisches Merkmal darstellt. Indem es auf die Gesamtheit einer Abfolge in ihrer Charakteristik abhebt, und diese nicht in vielzähligen Übergänge bzw. Zustandsvariationen zersplittert, verspricht die Sequenzbetrachtung einen zusätzlichen Erkenntnisgewinn (Scherer/Brüderl 2010, Stegmann et al. 2013).

Dabei wird unterstrichen, dass durch die Sequenzbetrachtung der realweltliche Phänotyp von Verhaltensweisen adäquater repräsentiert wird, als wenn in einzelne Ereignisse oder Aktivitätswechsel „zerlegt" wird. Man könne eine Sequenz als spezielle Abfolge von Wechseln und Zuständen mit je eigener Beharrung, als supra- oder Amalgam-Determinante von sozial relevanten outcomes wie Lebenszufriedenheit und soziale Integration oder innerfamilialer Lebensqualität verstehen, die als genuiner Lebensstilfaktor einen zusätzlichen sozial differenzierende Wirkfaktor betrachten, der traditionelle vertikale Dimensionen wie Einkommen, Bildung und berufliche Position ergänzt.

Umgekehrt gerät in einer Kausalperspektive auch die Aktivitätssequenz als Explanandum in den analytischen Blick. Dabei stellt sich die Herausforderung, wie aus der kategorialen Musterbeschreibung eine quantifizierte Variable zu und in den weiteren statistischen Analysen als Messgröße modelliert werden kann.

Eine Betrachtung der Sequenz als kategoriale Variable hilft nicht weiter, da aufgrund der Permutation von Zustandskategorien und Zustandsmomenten Kategorien mit nahezu individuell spezifische Sequenzmuster entstehen, mit denen statistische Zusammenhänge mit anderen Merkmalen nicht mehr abgebildet werden können.

Ein Lösungsweg in diesem Dilemma besteht darin, die individuellen Sequenzen über deren Ähnlichkeit z.B. mit Optimal-Matching-Verfahren zu Klassen zu gruppieren, mit denen eine überschaubare Zahl an übergreifenden Sequenzmustern erzielt wird. Diese neue Variable, die verschiedene typische Sequenzmuster repräsentiert, lässt sich als abhängige Variable mittels multinomialer Regressionsverfahren bzw. als kategoriale unabhängige Variablen als Block von binären Vergleichen mit einer Referenzkategorie analysieren. Das Problem bei diesem Weg besteht darin, dass das holistische, eine gesamte Sequenz als solche beschreibende Merkmal, mittels eines Verfahrens hergestellt wurde, dessen Ergebnis eine konzeptuelle Interpretation der Daten darstellt. Damit wird jedoch vorab eine konzeptuell bedingte Unschärfe in die Analyse von Sequenz-Einflüssen eingebaut.

Ein anderer Weg zur datennahen Messung der spezifischen Sequenzmuster-Information bietet sich mit den Ansätzen an, die die Komplexität von Sequenzen zu erfassen suchen.

Hierbei kommen insbesondere Beiträge die

Komplexitätsmaße von Gabadinho et al. (2011) und Elzinga (2010) in Betracht. Im vorliegenden Paper soll auf die Konstruktion dieser Maße detailliert eingegangen werden. Darüber hinaus sollen die beiden diskutierten Komplexitätsindizes in einer vergleichenden in einer empirischen Analyse daraufhin untersucht werden, ob sie nur verschiedene numerische Wege, aber im Endeffekt gleich wirksame Abbildungen einer holistischen Sequenzeigenschaft darstellen oder ob sie einen jeweils eigenen Zusatzbeitrag hinsichtlich ihrer kausalen Bestimmtheit bereitstellen.

2 Strukturmerkmale von Sequenzen

An dieser Stelle erscheint es sinnvoll, zunächst die grundlegenden Daten-Konzepte zu definieren, die in Bestimmung von Sequenzen eingehen, nämlich Zustand (Status) und Alphabet.

Ein **Zustand** ist i.d.R. auf einen bestimmten Zeitpunkt oder eine bestimmtes Zeitintervall bezogen, je nachdem mit welcher zeitlichen Granularität Zustände aufgezeichnet worden sind. Jeder Zustand ist zeitlich oder hinsichtlich seiner Position in der Sequenzreihe eindeutig bestimmt. Man kann deshalb von der Ordnungsnummer eines Zustandes sprechen. Im Fall von Tagebuchaufzeichnungen z.B. der Zeitverwendungserhebung 2001/2002 des Statistischen Bundesamtes ist ein Zustand immer auf ein 10-Minuten-Intervall bezogen. Die Ordnungsnummer ergibt sich aus der zeitlichen Lagerung eines 10-Minuten-Intervals. In Tabelle 1 sind verschiedene exemplarische Sequenzen aufgeführt, deren Eigenschaften im Folgenden untersucht werden. So z.B. eine Sequenz mit wenigen Zustandsmomenten (Sequenz 1 oder 2) und Sequenzen mit deutlich mehr Zustandsmomenten (z.B. Sequenz 3-8).

Das **Alphabet** bezeichnet die beschreibt die kategoriale Qualität der Zustände. Es umfasst also alle Arten von Zuständen, die einer Stichprobe von Personen beobachtet werden können bzw. beobachtet worden sind.

Aus den Elementen (Buchstaben) eines Alphabets und ihrer geordneten Abfolge über eine Reihe von Zustandsmomenten (Positionen) ergibt sich also eine spezifische Sequenz.

Dabei wird das spezifische Muster einer Sequenz durch verschiedene Strukturmerkmale der Abfolge bestimmt, die im Folgenden anhand der exemplarischen Sequenzen erläutert werden sollen.

Wir sehen in Tabelle 1, dass sich Sequenzen bzgl. des Umfangs des Alphabets unterscheiden können. So besteht das Alphabet von Sequenz 1 aus zwei Buchstaben, während das Alphabet von Sequenz 4 aus sechs Buchstaben besteht. Der Umfang des Alphabets deutet auf die inhaltliche Vielfalt in einer Sequenz hin.

Tabelle 1:

Sequenz-Nr.	Sequenz mit distinktiver Zeit-Zustand-Abfolge	Sequenz mit distinkter Zustands-abfolge	Anzahl distinkter Zustände (Alphabet)	Gesamt-dauer von Zustand „A“	Anzahl der Episoden	Anzahl der Episoden mit Zustand „A“
1	AABBAABBAABB	ABABAB	2	6	6	3
2	AABBCCBBAABB	ABCBAB	3	4	6	2
3	AAAAABBBBBAAAAABBBBBAAAAABBBBB	ABABAB	2	15	6	3
4	AAAAABBBBBCCCCCDDDDDEEEEEFFFFF	ABCDEF	6	5	6	1
5	AAAAABBBBBCCCCCAAAAABBBBBCCCCC	ABCABC	3	10	6	2
6	AAAAAAAAAABBBCCAAAAAAAAAABBBCC	ABCABC	3	20	6	2
7	AAAAAAAAAABBBAAAAAAAAAABBBCCCC	ABABC	3	20	5	2
8	AAAAAAAAAABBBAAAAAAAAAABBBCCCB	ABABCB	3	20	5	2
9	AABBAABBAABBAABB	ABABABAB	2	8	8	4

Ein weiterer Aspekt, der das Abfolgemuster einer Sequenz kennzeichnet, ist unmittelbar sichtbar als Wechsel von einem Zustand zu einem differierenden Zustand. Diese Zustandswechsel definieren auch die Anzahl von Episoden, also Moment/Positions- Intervallen mit gleicher Zustandskategorie. In diesem Sinne können wir in Tabelle 1 sehen, dass die Sequenzen 1 bis 6, so unterschiedlich sie sich insgesamt darstellen, die gleiche Anzahl von Wechsel bzw. Episoden aufweisen. Sequenz 7 hingegen, die der Sequenz 6 auf den ersten Blick sehr ähnelt, weist nur 5 Episoden bzw. 4 Wechsel auf. Die Anzahl der Wechsel verweist auf die zeitliche Variabilität als genuines Merkmal einer Sequenz.

Weiterhin fällt auf – wenn man Sequenz 1, Sequenz 3, Sequenz 5 oder Sequenz 6 miteinander vergleicht – dass auch der Rhythmus von Wiederholungen von Zuständen bzw. Zustands-Teilsequenzen ein Charakteristikum einer Sequenz sein kann. Bei Sequenz 1 wiederholen sich in regelmäßiger Folge die Subsequenzen AA und BB, bei Sequenz3 sind es AAAA und BBBB bei gleichem Zyklus wie in Sequenz 1, allerdings mit längeren Episoden. Bei Sequenz 5 liegt ebenfalls ein Wiederholungszyklus vor, allerdings mit einer höheren Anzahl der verschiedenen Zustände.

Sequenz 6 hat das gleiche Wiederholungsmuster wie Sequenz 5, dennoch unterscheiden sie diese beiden Sequenzen, weil in der Sequenz 6 die Dauer der Episode A länger ist als bei Sequenz 5. Diese Differenz deutet daraufhin, dass auch die **Dauer der einzelnen Zustandsepisoden** innerhalb einer Sequenz ein weiteres differenzierendes Attribut einer Sequenz sein kann.

3 Zwei komposite Indikatoren der Komplexität von Sequenzen

Es gibt zwei wichtige Ansätze zur Quantifizierung des Strukturmusters von Sequenzen im Sinne von Sequenzkomplexität. Es handelt sich einerseits um den Komplexitätsindex von Gabadinho et al. (2011) und andererseits um den Komplexitätsindex von Elzinga (2010). In einer Arbeit von 2006 hat Elzinga seinen Komplexitätsindex noch als Turbulence-Index bezeichnet, später (2010) aber die Bezeichnung in Komplexitätsindex geändert (und den Autor dieses Papers auch in persönlicher Kommunikation auf die veränderte Bezeichnung hingewiesen). Um die beiden hier zu diskutierenden Indikatoren der Strukturkomplexität von Sequenzen auch sprachlich zu trennen, folge ich der Bezeichnung von Elzinga (2010) Komplexitätsindex, füge jedoch den Zusatz T hinzu, um ihn vom Komplexitätsindex von Gabadinho et al (2011) auch sprachlich zu unterscheiden, wobei letzterer als Komplexitätsindex C bezeichnet werden soll.

3.1 Der Komplexitätsindex C

Der Komplexitätsindex von Gabadinho et al. (2011) wird nach der Formel -1- bzw. -1a- gebildet:

$$\text{-1-} \qquad C(s) = \sqrt{\left(\frac{q(s) * h(s)}{q_{max} * h_{max}}\right)}$$

$$\text{-1a-} \qquad C(s) = \sqrt{\left(\frac{q(s)}{q_{max}}\right) * \left(\frac{h(s)}{h_{max}}\right)}$$

S bezeichnet die Sequenz, q die Anzahl der Episodenwechsel innerhalb einer individuellen Sequenz und h bezeichnet die Entropie (nach Shannon) einer individuellen Sequenz.

Wie aus der Formel 1a ersichtlich ist, gehen zwei dynamische Strukturmerkmale in den Komplexitätsindex ein, nämlich die zeitliche Variabilität einerseits, gemessen durch die Wechselhäufigkeit und andererseits die Varietät, gemessen durch die Shannon Entropie. Zudem berücksichtigt diese Indexbildung auch die Tatsache, dass individuelle Sequenzen unterschiedlich lang sein können. Zu diesem Zwecke wird die einfache Wechselhäufigkeit an der maximal möglichen Zahl von Wechseln innerhalb der spezifischen Sequenz normiert (was man auch als individuelle Wechselrate bezeichnen kann) und die Entropie an der maximal möglichen Entropie.

Die Verknüpfung der beiden normierten Komponenten Wechselintensität und Entropie geschieht per geometrischem Mittelwert, nämlich als Quadratwurzel des Produktes normierter Wechselintensität und normierter Entropie.

Insgesamt gehen in den Komplexitätsindex C die Variabilität und die Varietät einer Sequenz ein. Dabei wird vermieden, die Varietät durch einfache Zahl der distinkten Zustände abzubilden, indem mit der Entropie als Indikator der Varietät wird auch die zeitliche Verteilung eines Zustandes berücksichtigt wird. Die zeitliche Variabilität der Zustände bzw. Zustandswiederholung wird beim Komplexitätsindex von Gabadinho et al. (2011) durch die (normierte) Häufigkeit des Zustandswechsels abgebildet.

Bei der deutschen Zeitverwendungserhebung (GTUS) beispielsweise wurden Aktivitäten (Zustände) innerhalb zehn Minuten Abschnitten erhoben. Ein Übergang ist dabei als Wechsel zwischen verschiedenen Aktivitätstypen definiert. Ein möglicher Aktivitätswechsel kann sich dabei von jedem Zehn-Minuten-Intervall zum nächsten Intervall vollziehen. Die maximale Anzahl von

Übergängen ist dabei begrenzt durch die Zeit in der Aktivitätswechsel durchgeführt werden können. Da Aktivitätswechsel nur in dem Zeitabschnitt in dem Personen wach sind durchgeführt werden können, bestimmt die individuelle Wachzeit (die Anzahl der Zehn-Minuten-Intervalle minus 1) das Maximum der möglichen Übergänge. Durch die Division der faktischen Wechsel durch die maximal mögliche Anzahl an Übergängen wird damit der unterschiedlichen Gesamtzeit der Akteure, im Fall des GTUS beispielsweise der individuell schwankenden Wachzeit Rechnung getragen. Insgesamt gilt: je größer der Subindikator q/qmax ist, desto größer die Komplexität einer Sequenz.

Der zweite Subindikator h/h_{max} zielt darauf ab, die Verteilung der Aktivitäten über die Zeit zu erfassen. Als der Verteilung der Zustände über die Zeit wird die Entropie genutzt. Dabei gilt: wenn die unterschiedlichen Zustände gleichmäßig über die ganze Sequenz verteilt sind, besitzt die Entropie ihren maximalen Wert. Das bedeutet anders ausgedrückt, dass die Wahrscheinlichkeit, einen spezifischen Zustand zu beobachten vom Zeitpunkt der Sequenz unabhängig ist. So ist bei einem Alphabet von zwei Zuständen eine sehr hohe Entropie dann erreicht, wenn zu jeden Moment an jeder Stelle der Sequenz beide unterschiedliche Zustände gleich wahrscheinlich beobachtet werden können. Ein minimales Niveau der Entropie wäre hingegen dann vorhanden, wenn nur ein Zustand über die ganze Sequenz zu beobachten wäre. In diesem Fall wäre die Wahrscheinlichkeit den Zustand zu beobachten gleich 1, was folglich „keine Varietät" bedeutet. Die Messung der Varietät mittels des Entropiemaßes unterscheidet sich also von der bloßen Zählung der Anzahl unterschiedlicher Aktivitäten darin, dass neben der Häufigkeit von unterschiedlichen Aktivitäten auch die Ausdehnung einer Aktivität in der Zeit berücksichtigt wird. Wenn eine Person zum Beispiel zwei unterschiedliche Aktivitäten über zehn Zeiteinheiten durchführt, wird ihr bei der Häufigkeit ein Wert zwei zugeordnet, unabhängig davon ob sie beide Aktivitäten jeweils über fünf Zeiteinheiten verteilt, oder ob sie für eine Aktivität neun Zeiteinheiten und für die andere nur eine Einheit verwendet hat. Die Entropie hingegen wäre im zweiten Fall niedriger, weil die Zeit gleichmäßiger über die Aktivitäten verteilt wäre. Es erscheint plausibel, dass eine Person, die zwei Aktivitäten mit der gleichen Zeitdauer ausführt, variabler in ihrer Aktivitätensequenz definiert wird, als eine Person die zwar auch zwei Aktivitäten verfolgt hat, aber eine Aktivität zeitlich nur marginal. Weil auch hier (ähnlich wie beim ersten Subindikator) die individuelle Entropie einer Sequenz abhängig von der maximal möglichen Entropie ist, wird die Entropie normiert, d.h. die maximal mögliche Entropie in das Kalkül einbezogen. Der Komplexitätsindex ist demnach ein Maß, welches Variabilität und Varietät einer Sequenz berücksichtigt, indem es die Anzahl der Übergänge und die Anzahl der verschiedenen Zustände mit deren zeitlicher Verteilung als Entropie in der Sequenz berücksichtigt. Dabei wird die Sequenz-Entropie an der maximal möglichen Entropie eines gegebenen Alphabets (also aller in einer Stichprobe möglichen Zustände) normiert. Die Zahl der Wechsel wird ebenfalls normiert, und zwar an der maximal möglichen Anzahl einer gegebenen Sequenz, d. i. Sequenzlänge minus 1.

Tabelle 2: Komponenten bei der Berechnung des Komplexitätsindizes C

Sequenz-Nr.	Sequenzmuster	Zahl der Zustands-wechsel	Sequenz-länge	Sequenz-basierte maximale Zahl der Wechsel	Sequenz-normierte Über-gangsrate	Sample-basierte maximale Zahl der Wechsel	Sample-Normierte Über-gangsrate
1	AABBAABBAABB	5	12	11	0.45	29	0.17
2	AABBCCBBAABB	5	12	11	0.45	29	0.17
3	AAAAABBBBBAAAAABBBBBAAAAABBBBB	5	30	29	0.17	29	0.17
4	AAAAABBBBBCCCCCDDDDDEEEEEFFFFF	5	30	29	0.17	29	0.17
5	AAAAABBBBBCCCCCAAAAABBBBBCCCCC	5	30	29	0.17	29	0.17
6	AAAAAAAAAABBBCCAAAAAAAAAABBBCC	5	30	29	0.17	29	0.17
7	AAAAAAAAAABBBAAAAAAAAAABBBCCCC	4	30	29	0.13	29	0.13
8	AAAAAAAAAABBBAAAAAAAAAABBBCCCB	5	30	29	0.17	29	0.17
9	AABBAABBAABBAABB	7	16	15	0.47	16	0.24

Tabelle 2b: Komponenten bei der Berechnung des Komplexitätsindizes C

Sequenz-Nr.	Sequenzmuster	Anzahl distinkter Zustände (Alphabet)	Entropie	Sample-basierte Entropie-Maximum	Sample normierte Entropie	Sequenz-normierte Entropie	Sequenz-basiertes Entro-piemaxi-mum
1	AABBAABBAABB	2	1.00	2.58	0.39	1.00	1.00
2	AABBCCBBAABB	3	1.46	2.58	0.57	0.92	1.58
3	AAAAABBBBBAAAAABBBBBAAAAABBBBB	2	1.00	2.58	0.39	1.00	1.00
4	AAAAABBBBBCCCCCDDDDDEEEEEFFFFF	6	2.58	2.58	1.00	1.00	2.58
5	AAAAABBBBBCCCCCAAAAABBBBBCCCCC	3	1.58	2.58	0.61	1.00	1.58
6	AAAAAAAAAABBBCCAAAAAAAAAABBBCC	3	1.24	2.58	0.48	0.78	1.58
7	AAAAAAAAAABBBAAAAAAAAAABBBCCCC	3	1.24	2.58	0.48	0.78	1.58
8	AAAAAAAAAABBBAAAAAAAAAABBBCCCB	3	1.30	2.58	0.50	0.78	1.58
9	AABBAABBAABBAABB	2	1.00	2.58	0.39	1.00	1.00

Festzuhalten bleibt, dass die Anzahl der Wechsel in einer Sequenz an der jeweiligen Sequenzlänge normiert wird, hingegen die Entropie einer Sequenz am Alphabet, also an der maximalen Entropie der Stichprobe. Um die Folgen dieser unterschiedlichen Vorgehensweise konkret zu vergleichen, haben wir eine Variante des Komplexitätsindizes C2 berechnet, bei der die Wechselhäufigkeit durch die maximale Länge in der Stichprobe normiert wird. D.h. die individuelle Sequenzlänge wird nicht berücksichtigt.

Tabelle 3: Komplexitätswerte für exemplarische Sequenzen, zwei Varianten der Berechnung von Komplexität C

Sequenz-Nr.	Sequenzmuster	C1, bei sequenznormierter Wechselzahl und samplenormierter Entropie	C2, bei samplenormierter Zahl der Wechsel und samplenormierter Entropie
1	AABBAABBAABB	0.42	0.26
2	AABBCCBBAABB	0.51	0.31
3	AAAAABBBBBAAAAABBBBBAAAAABBBBB	0.26	0.26
4	AAAAABBBBBCCCCCDDDDDEEEEEFFFFF	0.41	0.41
5	AAAAABBBBBCCCCCAAAAABBBBBCCCCC	0.32	0.32
6	AAAAAAAAAABBBCCAAAAAAAAAABBBCC	0.29	0.29
7	AAAAAAAAAABBBAAAAAAAAAABBBCCCC	0.25	0.25
8	AAAAAAAAAABBBAAAAAAAAAABBBCCCB	0.25	0.25
9	AABBAABBAABBAABB	0.42	0.20

Wie wir in Tabelle 3 sehen können erhalten Sequenz 1 und 4 (zwei Sequenzen die man intuitiv als deutlich unterschiedlich bezeichnen würde) einen nahezu gleichen Komplexitätswert, hingegen wird die Sequenz 3, die vom Abfolgemuster identisch mit Sequenz 1 ist, als deutlich weniger komplexe Sequenz bestimmt wird. Dies deutet daraufhin, dass die Sequenzlänge sich methodisch im Komplexitätswert C1 sehr markant auswirken kann. Es stellt sich die Frage, ob die Länge einer Sequenz tatsächlich als dominierendes Strukturmuster in die Komplexitätsberechnung eingehen soll. Vor allem stellt sich die Frage, ob es eine akzeptable Repräsentation der Unterschiede zwischen Sequenz 2 und 4 ist, wenn Sequenz 2 deutlich komplexer als Sequenz quantifiziert wird.

Normiert man hingegen an der Länge der längsten vorkommenden Sequenz, dann erhalten die Sequenzen Komplexitätswerte C2, die sie in einer anderen Ordnung darstellen lassen. Sequenz 4 ist demnach die komplexeste Sequenz und Sequenzen 1, 3 und 7 wären die am wenigsten komplexe Sequenzen. Es zeigt sich, dass einerseits die relative Entropie eine Rolle spielt, also auf wieviel unterschiedliche Zustände wie gleichmäßig die Zeitmomente verteilt werden, andererseits auch die Zahl der Wechsel. Auch beim Vergleich von Sequenz 5 und 6 erscheint der C2-Wert plausibel, insofern Sequenz 5 einen höheren Komplexitätswert als Sequenz 6 erhält, weil in Sequenz 5 die gleiche Anzahl verschiedener Zustände gleichmäßiger verteilt ist.

4 Der Komplexitätsindex von Elzinga

Der Komplexitätsindex T von Elzinga (2010) wird nach folgender Formel -2- berechnet:

$$T(s) = log_2\left(\varphi \cdot \frac{s^2{}_{t,max}(s)+1}{s^2{}_t(xs)+1}\right) \qquad \text{-2-}$$

wobei gilt:

Log₂ == Logarithmus zur Basis 2

φ == Anzahl der Subsequenzen mit distinkten sukzessiven Zuständen

$s^2{}_{t,max}(s)$ == maximale Varianz von Episodendauern in einer Sequenz bei gegebener Anzahl von Episoden

$s^2{}_t(xs)$ == Varianz der Dauern der gegebenen Episoden innerhalb einer Sequenz

Für **$s^2{}_{t,max}(s)$** gilt: $s^2_{t,max}(x) = \left(\ell_d(x) - 1\right)\left(1 - \bar{t}(x)\right)^2$

Der Quotient aus $s^2{}_{t,max}(s)$ und $s^2{}_t(xs)$ stellt die inverse normalisierte Varianz der Dauern der gegebenen Episoden (also Subsequenzen mit sukzessiver Konstanz des Zustands) einer Sequenz dar.

Die maximale Varianz hängt direkt von der Sequenz-Länge ab.

Im Folgenden soll zunächst auf die Bestimmung von Phi und dessen impliziten Informationsgehalts bei der Beschreibung eines Sequenzmusters eingegangen werden. Ausgangspunkt der Berechnung von Phi ist eine Sequenz als Reihe distinkt-sukzessiver Zustände, d.h. ohne die eventuelle vorhandene sukzessive Wiederholung einzelner Zustände. Damit wird der Varietät in einer Sequenz Rechnung getragen werden. Als Maßzahl der Varietät dient die Anzahl der distinkten Subsequenzen. Diese ergibt sich aus der Potenzmenge, die in einer Abfolge der distinkten Zustände einer Sequenz beinhaltet ist, d.h. die Menge aller Kombinationen, die zwischen den Elementen einer gegebenen geordneten Reihe möglich ist. Zur Potenzmenge gehört auch die Leermenge. Je mehr unterschiedliche Elemente in der Sequenzreihe vorhanden sind, umso größer ist Potenzmenge, d.h. umso mehr Subsequenzen sind möglich, wobei zunächst hier auch Doppelungen von Subsequenzen gehören. Die Menge der distinkten Subsequenzen (DSS) distinkt-sukzessiver Zustände, ergibt sich somit als Potenzmenge einer Zustands-Reihe abzüglich der doppelten Subsequenzen, plus Leermenge.

Die Sequenz ABA als Beispiel zur Veranschaulichung der Bestimmung distinkter Subsequenzen (DSS).

Tabelle 4: Potenzmenge der nicht distinkten und distinkten Subsequenzen der Sequenz ABA

Potenzmenge der Subsequenzen	Potenzmenge der distinkten Subsequenzen
A	A
B	B
A (Doppel)	
A B	A B
A A	A A
B A	B A
A B A	A B A
Leersequenz	Leersequenz

Man kann in Tabelle 4 sehen, dass die Potenzmenge der Sequenz aus insgesamt acht Subsequenzen besteht, wobei die Leermenge dazu zählt. Eine Sequenz kommt jedoch doppelt vor, nämlich die Sequenz A, diese Doppelung wird nicht berücksichtigt, sodass die abschließende Zahl der distinkten Subsequenzen 7 beträgt.

In Tabelle 5 ist die Anzahl der DSS-Werte (Elzinga's Phi) für die einzelnen Beispiel-Sequenzen aufgeführt (in Tabelle 1 des Anhangs sind alle distinkten-sukzessiven Subsequenzen aufgeführt). Man sieht, dass sich in diesem Wert deutlich die Varietät der einzelnen Sequenzen ausdrückt, so hat Sequenz 4 mit sechs verschiedenen Element des Alphabets auch den höchsten Phi-Wert. Allerdings ist das nur teilweise eine Reflexion der Varietät, wie man im Vergleich der Sequenzen 2,5,6,7 und 8 sieht. Diese weisen drei verschiedene Buchstaben auf, die unterschiedlich zeitlich verteilt sind, wobei es eine Entsprechung zum Entropie-Wert zu geben scheint. Allerdings ist ein gravierender Unterschieden festzustellen. Während Sequenzen 6 und 7 den gleichen Entropiewert aufweisen, unterscheiden diese sich deutlich bezüglich des Phi-Wertes, wobei der relativ kleine Phi-Wert von 24 die relativ kürzere DSS-Länge widerspiegelt. Erweitert man die Sequenz 7 um einen Zustand B, d.h. ersetzt man das letzte C in der Sequenzreihe durch ein B, so ergibt dies einen höheren Phi-Wert, der aber immer noch niedriger liegt als bei Sequenz 6. Der Strukturmuster-Unterschied liegt darin, dass bei Sequenz 6 ein längerer Wiederholungszyklus ABC-ABC vorliegt als bei Sequenz 8, mit AB AB und CB. Bei Sequenzen mit striktem kurzphasigem Wiederholungsmuster (Sequenzen 1 und 3) fällt der Phi-Wert am kleinsten aus.

Wir können festhalten, dass sich Entropie und Phi hinsichtlich der Erfassung von Varietät ähneln, Phi jedoch auch das Ausmaß von Regularität (zyklische Wiederholung) in einer Sequenz zu erfassen scheint. Beim Phi-Wert ist jedoch zu berücksichtigen, dass es auch von der Länge der Sequenz distinkt-sukzessiver Zustände abhängt, wie ein Vergleich der Phi-Werte von Sequenz 1 und 7 verdeutlicht.

Tabelle 5: Bestimmung der Anzahl distinkter Subsequenzen

Sequenz-Nr.	Sequenz mit vollständigem Sequenzmuster	Sequenz mi distinktem Sequenz-muster	Anzahl distinkt-sukzessiver Zustände	Phi Anzahl der distinkten Subsequenzen einer Sequenz mit distinkt-sukzessiven Zuständen	Sample normierte Entropie
1	AABBAABBAABB	ABABAB	6	**33**	0.39
2	AABBCCBBAABB	ABCBAB	6	**46**	0.57
3	AAAAABBBBBAAAAABBBBBAAAAABBBBB	ABABAB	6	**33**	0.39
4	AAAAABBBBBCCCCCDDDDDEEEEEFFFFF	ABCDEF	6	**64**	1.00
5	AAAAABBBBBCCCCCAAAAABBBBBCCCCC	ABCABC	6	**52**	0.61
6	AAAAAAAAAABBBCCAAAAAAAAAABBBCC	ABCABC	6	**52**	0.48
7	AAAAAAAAAABBBAAAAAAAAAABBBCCCC	ABABC	5	**24**	0.48
8	AAAAAAAAAABBBAAAAAAAAAABBBCCCB	ABABCB	6	**41**	0.47
9	AABBAABBAABBAABB	ABABABAB	8	**88**	0.39

Im nächsten Schritt soll betrachtet werden, wie der zweite Term der Berechnung von Komplexität T, nämlich die Varianz-Gewichtung berechnet wird und in die Berechnung von T eingeht. Dabei ist hilfreich, sich Elzingas konzeptuelle Idee dieser Varianzgewichtung zu vergegenwärtigen. Wie Gabadinho et al. (2011) schreiben, geht Elzingas Konzept von einem "prediction point of view" aus, so dass gilt " *the higher the differences in state durations and hence the higher*

their variance, the less uncertain the sequence. In that sense, small duration variance indicates high complexity" (S. 23). D.h. mit der Episodendauer-Varianz wird der Aspekt der Variabilität einer Sequenz erfasst. Insofern ist hier auch eine Parallelität zum Entropie-Konzept festzustellen, das ebenfalls eine Aussage über die Gleichmäßigkeit der Zustandsverteilung zulässt: Je höher die Entropie, umso gleichmäßiger die Verteilung der verschiedenen Buchstabenzustände, umso geringer die Varianz der Buchstaben-Episoden. Um in diesem Sinne die Varianz der Episodendauern als verstärkender Indikator der Varietät (die mittels des Phi-Wertes erfasst wird) einzusetzen – und den Effekt der Gesamtlänge einer Sequenz auszugleichen – wird die Episodendauer-Varianz reziprok zur maximalen Varianz ins Verhältnis gesetzt. In diesem Sinne entsteht mit dem inversen Varianzquotienten ein Varianzfaktor, dessen Wert den Grad der normierten Gleichmäßigkeit darstellt: Je höher dieser Wert, umso gleichmäßiger die Verteilung der Episodendauern und/oder die maximale Varianz als Ausdruck der Anzahl der Episoden und der Durchschnittsdauer der Episoden, womit die Gesamtsequenzlänge repräsentiert ist (siehe Tabelle 6).

Tabelle 6: Komponenten der Berechnung des Komplexitätsindizes T

Sequenz-Nr.	Sequenz mit vollständigem Sequenzmuster	Sequenz mi distinktem Sequenz-muster	Anzahl distinkt-sukzessiver Episoden	Maximale Episoden-dauer-Varianz	Varianz der Episoden-dauern	Inverse Vari-anzquotient: (Vmax+1)/(V+1)
1	AABBAABBAABB	ABABAB	6	5	0	6
2	AABBCCBBAABB	ABCBAB	6	5	0	6
3	AAAAABBBBBAAAAABBBBBAAAAABBBBB	ABABAB	6	80	0	81
4	AAAAABBBBBCCCCCDDDDDEEEEEFFFFF	ABCDEF	6	80	0	81
5	AAAAABBBBBCCCCCAAAAABBBBBCCCCC	ABCABC	6	80	0	81
6	AAAAAAAAAABBBCCAAAAAAAAAABBBCC	ABCABC	6	80	12.7	5.9
7	AAAAAAAAAABBBAAAAAAAAAABBBCCCC	ABABC	5	100	10.8	8.6
8	AAAAAAAAAABBBAAAAAAAAAABBBCCCB	ABABCB	6	100	10.8	8.6
9	AABBAABBAABBAABB	ABABABAB	8	9	0	10

Im letzten Schritt kombiniert der Komplexitätsindex T den Subindikator der Varietät/Regularität und den inversen Varianzfaktor multiplikativ, was man als Varianzgewichtung der Varietät/Regularität betrachten kann. D.h. eine Sequenz mit hoher Varietät (in der sich auch die Länge der Sequenz distinkt-sukzessiver Zustände niederschlägt) wird verstärkt oder abgeschwächt, je nachdem ob eine geringe Varianz (bzw. der Längen sukzessiv gleichen Subsequenzen) oder hohe Varianz vorliegt. Ein hoher T-Wert einer Sequenz bedeutet also, dass es sich um eine Sequenz mit hoher Varietät und hoher Gleichförmigkeit handelt.

Die binär-basierte logarithmische Transformation dient dazu, die Verteilung der Multiplikationswerte mit großer Spannweite auszugleichen.

Tabelle 6 (ff): Berechnung des Komplexitätsindizes T für Sequenzbeispiele

Sequenz-Nr.	Sequenz mit vollständigem Sequenzmuster	Sequenz mi distinktem Sequenzmuster	Phi*InvVarianz-quotient	Log_2(phi*InvVarianz-quotient)
1	AABBAABBAABB	ABABAB	198	7.63
2	AABBCCBBAABB	ABCBAB	276	8.11
3	AAAAABBBBBAAAAABBBBBAAAAABBBBB	ABABAB	2673	11.38
4	AAAAABBBBBCCCCCDDDDDEEEEEFFFFF	ABCDEF	5184	12.34
5	AAAAABBBBBCCCCCAAAAABBBBBCCCCC	ABCABC	4212	12.04
6	AAAAAAAAAABBBCCAAAAAAAAAABBBCC	ABCABC	307	8.26
7	AAAAAAAAAABBBAAAAAAAAAABBBCCCC	ABABC	205	7.68
8	AAAAAAAAAABBBAAAAAAAAAABBBCCCB	ABABCB	351	8.46
9	AABBAABBAABBAABB	ABABABAB	880	9.78

5 Gegenüberstellung der Komplexitätsindizes C und T

Wenn man das Ranking der Beispiel-Sequenzen anhand des Komplexitätsindizes C dem Ranking anhand des Komplexitätsindizes T gegenüberstellt wird deutlich, dass T und C nicht in gleichsinniger Weise die Komplexität einer Sequenz erfassen. Der Rangkorrelationskoeffizient dieser Beispiel-Sequenzen beträgt 0.02, d.h. dass man je nach Index zu differenten Einschätzung des Komplexitätsgrades der einzelnen Sequenzen gelangt. Wie schon oben angedeutet wurde, sind beide kompositen Indizes anfällig für die Werte ihrer spezifischen Komponenten. Bei C ist es vor allem die Wechselrate, in die sich die Länge der Sequenz negativ niederschlägt, d.h. je länger die Sequenz umso geringer fällt die Wechselrate tendenziell aus. Dies kann aber zu irrigen Schlüssen führen, wie man im Vergleich von Sequenz 1 und 3 sehen kann: bei gleicher Wechselzahl erhalten wir unterschiedliche Wechselraten und deshalb – bei gleicher Entropie – deutlich unterschiedliche Komplexitätsranges. Bei T ist ein methodisches Artefakt dahingehend festzustellen, dass das Phi, also die Anzahl der DSS, deutlich von der Episodenzahl abhängt. Dies führt zum Beispiel dazu, dass bei Sequenzen 1 und 9 bei ähnlicher Wechselrate und gleicher Entropie, aber unterschiedlicher Episodenzahl, es zu sehr unterschiedlichem Phi-Wert und Varianzfaktor kommt, mit dem Ergebnis, dass Sequenz 9 eine höhere Komplexitätsrang erhält als Sequenz 1, welche die geringste T-Komplexität aufweist. Interessanterweise werden Sequenz 1 und Sequenz 9 beim C-Komplexitätswert beide eher in höherem Komplexitätsrang als beim T-Komplexitätswert eingeordnet.

Eine Situation, in der zwei komposite Indizes, die darauf abzielen die Komplexität einer Sequenz zu quantifizieren, eine nur geringe Kovarianz aufweisen, erscheint methodisch unbefriedigend. Es stellt sich die Frage nach der Validität dieser Indizes. Die Ergebnisse könnten womöglich als methodische Artefakte eingeordnet werden. Um diese Frage näher zu beleuchten, soll im nächsten Abschnitt eine empirische Analyse auf der Basis einer repräsentativen und umfangreichen Stichprobe von Aktivitätssequenzen durchgeführt werden.

Tabelle 7: Komplexitätswerte und ihre Komponenten bei den Komplexitätsindizes C und T

Sequenz-Nr.	Sequenz mit vollständigem Sequenzmuster	Sequenz mit distinktem Sequenz-muster	Wechsel-rate	Entropie (sample-normiert)	Komplexi-tätsindex C	Ordnungs-Nr. nach C	Phi (# DSS[1])	Varianz-faktor	Komplexi-tätsindex T	Ordnungs-Nr. nach T
1	AABBAABBAABB	ABABAB	0.45	0.39	0.42	7	33	6	7.63	1
7	AAAAAAAAAABB BAAAAAAAAAABBBCCCC	ABABC	0.13	0.48	0.25	1	24	8.6	7.68	2
2	AABBCCBBAABB	ABCBAB	0.45	0.57	0.51	8	46	6	8.11	3
6	AAAAAAAAAABBBCCAAAAA AAAAABBBCC	ABCABC	0.17	0.48	0.29	4	52	5.9	8.26	4
8	AAAAAAAAAABB-BAAAAAAAAAABBBCCCB	ABABCB	0.17	0.47	0.28	3	41	8.6	8.46	5
9	AABBAABBAABBAABB	ABABABAB	0.47	0.39	0.66	9	88	10	9.78	6
3	AAAAABBBBBAAAAABBBB BAAAAABBBBB	ABABAB	0.17	0.39	0.26	2	33	81	11.38	7
5	AAAAABBBBBCCCCCAAAAAB BBBBCCCCC	ABCABC	0.17	0.61	0.32	5	52	81	12.04	8
4	AAAAABBBBBCCCCCDDDDDE EEEEFFFFF	ABCDEF	0.17	1	0.41	6	64	81	12.34	9

Anmerkung: 1) Anzahl der distinkten sukzessiven Subsequenzen

6 Empirische Analyse zur Komplexitätsmessung bei Freizeitaktivitätssequenzen

Im Folgenden soll untersucht werden, wie die oben diskutierten Komplexitätsindizes von Elzinga (2009) und Gabadinho et al: (2011) sich bei einer empirischen Analyse zur sozialstrukturellen Determinierung der Komplexität von Sequenzen verhalten. Darüber hinaus soll untersucht werden, in welchem Maße die Subkomponenten der Komplexitätsindizes mit den sozialstrukturellen Faktoren kovariieren.

Als Ausgangspunkt greifen wir die Frage nach der Komplexität der persönlichen Freizeitgestaltung am Wochenende (Papastefanou, Gruhler 2014) auf. Wir nutzen hierzu die Daten der Zeitbudget-Erhebung des Statistischen Bundesamtes von 2001/2002. Als Alphabet der Freizeit-Aktivitäten am Sonntag werden folgende Aktivitäten definiert: Lesen, Musik hören, Fernsehen, Computer nutzen, Hobby nachgehen, Sport treiben sowie die Residualkategorie „andere Aktivitäten".

Auf dieser Basis berechnen wir die Komplexitätsindizes C und T. Als Indikatoren der sozialstrukturellen Lage werden folgende Variablen herangezogen: Geschlecht, Alter, Familienstand, Haushaltsnettoeinkommen (Intervallkategorien), Haushaltsgröße, allgemeinbildender Schulabschluss, beruflicher Bildungsabschluss und beruflicher Status. Zur einfacheren Interpretation beschränken wir die Zielgruppe auf Personen über 17 Jahre, die in Vollzeit beschäftigt sind.

Zunächst können wir eine hohe Kovariation der beiden Komplexitätsindizes C und T feststellen: für die Gruppe der Vollzeitbeschäftigten über 17 Jahre beträgt der Pearson Korrelationskoeffizient r=.94. Die beiden kompositen Komplexitätsindizes scheinen damit austauschbar zu sein.

Betrachten wir nun, wie sich Komplexitätsindizes in der Kovariation mit exogenen Variablen verhalten. Die Kovariation mit den Indikatoren der Sequenzkomplexität schätzen wir als multivariates OLS Regressionsmodell (siehe Tabelle 8).

Tabelle 8: Sozio-demographische Effekte auf komposite Komplexitätsindizes und deren Komponenten (b-Werte, OLS-Regression)

VARIABLES	(1) Komplexität C	(2) Komplexität T	(3) Wechsel-rate	(4) rel. Entropie	(5) Log_2(phi)	(6) Log_2(Varianz-faktor)
Geschlecht	**-0.0283*** **	**-1.359*** **	**-0.0133*** **	**-0.0654*** **	**-0.766*** **	**-0.622*** **
	(0.00472)	(0.245)	(0.00384)	(0.00870)	(0.141)	(0.181)
Alter	0.000281	**0.0226***	0.000106	**0.000995** **	**0.0216*** **	0.0152
	(0.000259)	(0.0136)	(0.000211)	(0.000477)	(0.00778)	(0.00991)
ledig	**0.0105***	-0.0727	0.00463	0.0118	0.00494	-0.208
	(0.00635)	(0.331)	(0.00517)	(0.0117)	(0.190)	(0.243)
geschieden	-0.00510	0.00606	-0.00782	-0.00822	-0.0378	-0.178
	(0.00746)	(0.384)	(0.00605)	(0.0137)	(0.220)	(0.284)
verwitwet	-0.0141	0.100	-0.00992	-0.00983	-0.274	0.658
	(0.0210)	(1.072)	(0.0174)	(0.0393)	(0.615)	(0.816)
Dauernd getrennt	-0.0174	-0.879	-0.00292	-0.0356	-0.412	-0.0629
	(0.0183)	(0.954)	(0.0151)	(0.0343)	(0.547)	(0.712)
Haushaltsgröße	-0.000589	-0.0905	-0.000336	-0.00119	-0.0300	-0.0180
	(0.00185)	(0.0960)	(0.00151)	(0.00343)	(0.0551)	(0.0712)
Mittlerer Abschluss	0.00167	0.356	-0.000253	0.00224	**0.351** **	0.0659
	(0.00532)	(0.278)	(0.00431)	(0.00977)	(0.160)	(0.203)
Abitur	**0.0110***	**0.939*** **	0.00698	0.0192	**0.722*** **	**0.420***
	(0.00649)	(0.339)	(0.00527)	(0.0119)	(0.195)	(0.248)
Fachschule/Meister	-0.00107	0.204	-0.00334	0.00116	0.162	0.0590
	(0.00669)	(0.351)	(0.00541)	(0.0123)	(0.201)	(0.255)
Universitätsabschluss	-0.00629	-0.00485	-0.00310	-0.0135	0.0646	-0.00153
	(0.00782)	(0.407)	(0.00630)	(0.0143)	(0.234)	(0.297)
Beamte	**0.0220*** **	**0.769** **	**0.0116** **	**0.0410*** **	**0.520** **	0.363
	(0.00708)	(0.368)	(0.00570)	(0.0129)	(0.211)	(0.268)
Angestellte	**0.0233*** **	**1.083*** **	**0.0162*** **	**0.0527*** **	**0.696*** **	**0.914*** **
	(0.00614)	(0.319)	(0.00497)	(0.0113)	(0.183)	(0.234)
Arbeiter	**0.0274*** **	**1.388*** **	**0.0190*** **	**0.0538*** **	**0.892*** **	**1.001*** **
	(0.00694)	(0.363)	(0.00561)	(0.0127)	(0.208)	(0.264)
Auszubildende	0.0167	0.916	0.00877	**0.0433** **	0.311	**1.007** **
	(0.0112)	(0.584)	(0.00902)	(0.0204)	(0.335)	(0.424)
Wehr-/Zivildienst	0.0167	1.064	0.00264	0.0338	0.716	-0.304
	(0.0206)	(1.074)	(0.0160)	(0.0363)	(0.616)	(0.754)
Constant	0.177***	1.472	0.0868***	0.352***	4.719***	-4.772***
	(0.0201)	(1.050)	(0.0163)	(0.0370)	(0.602)	(0.767)
Observations	2,088	1,998	2,170	2,170	1,998	2,170
R-squared	0.033	0.032	0.016	0.045	0.043	0.021

Standardfehler in Klammern. *** p<0.001, ** p<0.01, * p<0.1

In der Schätzung des Modells 1 mit dem Komplexitätsindex C als abhängige Variable, können wir feststellen, dass folgende Merkmale einen signifikanten Effekt haben:

- Frauen im Vergleich zu Männern (geringere Komplexität)
- Ledige im Vergleich zu Verheirateten (höhere Komplexität)
- Mit Abitur im Vergleich zu Hauptschulabschluss (höhere Komplexität)
- Beamte, Angestellte, Arbeiter im Vergleich zu Selbstständigen (höhere Komplexität)
- Auszubildenden und Wehr- bzw. Ersatzdienstleistende unterscheiden sich nicht von Selbständigen hinsichtlich der Komplexität ihrer Freizeitaktivitäten am Sonntag.

Betrachtet man die Ergebnissen des Modells 2, bei dem der Komplexitätsindex T als abhängige Variable in Abhängigkeit von den sozialstrukturellen Variablen modelliert wird, kann man folgende signifikante Ergebnisse feststellen:

- Frauen im Vergleich zu Männern (geringere Komplexität)
- Jüngere im Vergleich zu Älteren (geringere Komplexität)
- Mit Abitur im Vergleich zu Hauptschulabschluss (höhere Komplexität)
- Beamte, Angestellte, Arbeiter im Vergleich zu Selbstständigen (höhere Komplexität)
- Auszubildenden und Wehr- bzw. Ersatzdienstleistende unterscheiden sich nicht von Selbständigen hinsichtlich der Komplexität ihrer Freizeitaktivitäten am Sonntag.

Im Vergleich der beiden Modellergebnisse stellen wir teilweise eine Übereinstimmung der geschätzten Effekte fest, allerdings gibt es jedoch auch gravierende Unterschiede.

Im Modell der Komplexität T wird ein linearer Effekt des Alters als signifikant geschätzt, im Modell der Komplexität C jedoch nicht. Umgekehrt legt die Schätzung des Modells der Komplexität C nahe, dass Ledige signifikanter höhere Komplexität aufweisen als Verheiratete, bezüglich der Komplexität T jedoch nicht. Abschließend sei noch darauf hingewiesen, dass beim Modell mit T eine deutlichere Differenzierung in den Unterschieden zwischen den beruflichen Statusgruppen Beamte, Angestellte und Arbeiter festzustellen ist. Nach diesem Modell hätten die Arbeiter die höchste Komplexität, während beim Modell mit C sich diese drei Statusgruppen nicht wesentlich voneinander unterschieden.

Dieses Ergebnis zeigt, dass die beiden Komplexitätsindizes zwar hoch korreliert sind, jedoch mit exogenen Kovariaten teilweise sehr unterschiedlich korrelieren. Ein Befund, der hinsichtlich wichtiger sozialer Struktureffekte zu unterschiedlichen Schlüssen führt, je nachdem welchen Komplexitätsindex man verwendet.

Um diese Kovariationsunterschiede genauer zu verstehen, haben wir auch die einzelnen Komponenten der beiden Indizes als separate Modelle mit ihrer Kovariation modelliert (Modelle 3 bis 5 in Tabelle 8).

Betrachten wir zunächst die Komponenten des Indizes C, Wechselrate und relative Sequenzentropie. Im Vergleich der Modelle 3 und 4 sehen wir, dass der Geschlechts- und Statuseffekt sowohl bei der Wechselrate wie auch bei der relativen Entropie signifikant ist, der Ledigenstatus und die allgemeinschulische Bildung jedoch nicht. D.h. diese Effekte werden erst in der Kombination von Wechselrate und Entropie signifikant. Andererseits können wir feststellen, dass bei der relativen Entropie der Alterseffekt signifikant wird, ebenso wie der Unterschied der Auszubildenden gegenüber den Selbstständigen. Offenbar repräsentieren Wechselrate und die relative Entropie in den Abfolgen von in-home Freizeitaktivitäten, als Operationalisierungen von Abwechslungs- und Varietätsdisposition, teilweise unterschiedliche Prozesse.

Betrachten wir nun die Effekte auf die Anzahl der distinkten sukzessiven Subsequenz(DSS), und zwar als Logarithmus auf der Basis 2, wie er als Komponente Phi in den Komplexitätsindikator von Elzinga eingeht. Zusätzlich untersuchen wir den Varianzfaktor, der aus dem Quotienten von maximaler und der faktischer Varianz der Episodenlängen gebildet wird. Auch bilden wir den Logarithmus auf der Basis 2, da er als solcher in den Komplexitätsindex T eingeht.

Bei den Ergebnissen zum log2(phi) (Modell 4) sehen wir, dass auch hier die gleichen signifikanten Effekte gefunden werden wie beim kompositen Index T, mit dem Unterschied, dass der Bildungseffekt deutlicher linear ausfällt, indem die mittlere Bildungsgruppe ebenfalls signifikant abweicht von der unteren, aber auch kleineren Effekt hat im Vergleich zur höchsten Bildungsgruppe der Personen mit Abiturabschluss.

Hinsichtlich des Varianzfaktors, also der Streuung der Episodendauer, stellen wir den Geschlechtseffekt, den einfachen Bildungseffekt, sowie Unterschiede zwischen Angestellten bzw. Arbeitern und Selbstständigen bzw. Beamten fest. Außerdem ergibt sich diesbezüglich ein signifikanter Unterschied der Auszubildenden von den Selbständigen, der beim kompositen Komplexitätsindex T nicht festzustellen ist. Ein Alterseffekt ist bei der Episodendauer-Varianz nicht feststellbar.

Insgesamt kann man folgendes feststellen:

- Die beiden kompositen Indizes C und T sind zwar hoch korreliert, bzgl. der Kovariation mit exogener Variablen scheinen sie jedoch teilweise unterschiedliche Prozesse zu sein.
- Die Teil-Komponenten der Indizes T und C kovariieren z.T. unterschiedlich mit den untersuchten Sozialstrukturvariablen in ihren entsprechenden Kompositindizes.

7 Schlussfolgerungen

Ausgehend von der Frage, wie das Strukturmuster von individuellen Zustands-Sequenzen quantifiziert werden kann, haben wir zwei Vorschläge zur Konzeptualisierung und Operationalisierung von Komplexität untersucht. Ausgangspunkt war dabei die begriffliche Unterscheidung von Zustand, Alphabet und Position als Grundelemente, aus deren Spezifikation sich spezifische Sequenzen ergeben. Mit ausgewählten Beispielen wurde deutlich gemacht, dass zwei wesentliche Struktur-Facetten von Sequenzmustern die zeitliche Variabilität und die inhaltliche Varietät sind. Die untersuchten kompositen Indizes entwickeln unterschiedliche operationale Indikatoren für diese Facetten und kombinieren diese auf unterschiedlicher Weise zu einem Gesamtindex der Komplexität. Demnach wäre Komplexität als Kombination von Variabilität und Varietät zu verstehen. Bei der Untersuchung des Komplexitätsindizes T nach Elzinga wurde ein zusätzliches Strukturmerkmal von Sequenzmuster erkennbar, das man als Regularität bezeichnen könnte. In der Regularität drückt sich das Ausmaß zyklischer Wiederholung von Zuständen bzw. Zustandsfolgen aus.

Somit haben wir folgende Situation, bei der die betrachteten kompositen Komplexitätsindizes in unterschiedlicher Weise zur quantifizierten Operationalisierung der Sequenzmuster-Facetten führen.

Tabelle 9: Sequenzstrukturdimensionen und deren Operationalisierung mit den Komplexitätsindizes C und T

	Komplexitätsindex C	Komplexitätsindex T
Variabilität	Episodenwechselrate	Varianz der Episodendauer
Varietät	Entropie	Anzahl der distinkten sukzessiven Subsequenzen
Regularität	n/a	Anzahl der distinkten sukzessiven Subsequenzen

Bei der vergleichenden Betrachtung der Indizes C und T bezüglich der systematisch ausgewählten Sequenzmuster zeigte sich, dass die Rangordnung der Sequenz hinsichtlich ihrer Komplexität deutlich verschieden ausfällt.

In der empirischen Analyse jedoch, basierend auf einer Substichprobe der ZVE von 2001/2002 zeigte sich eine sehr hohe Korrelation der beiden Indizes, die augenscheinlich eine Substituierbarkeit von C und T nahelegte. In der multivariaten Betrachtung der beiden Indizes als abhängiger Variable konnten wir weiterhin feststellen, dass C und T nur teilweise mit den gleichen exogenen Variablen kovariieren. Bei wichtigen sozialstrukturellen Variablen wie Alter und allgemeinen Bildungsabschluss zeigen sich markante Unterschiede. Eine multivariate Modellierung der Teilkomponenten der Indizes zeigte weiterhin auf, dass einige Variablen wie Geschlecht und berufliche Stellung durchgehend mit allen Komponenten direkt kovariieren, allerdings Entropie oder DSS zusätzliche sozialstrukturelle Einflusseffekte offenbaren.

Insgesamt können wir aus diesen Befunden nicht den Schluss ziehen, dass C und T austauschbar sind. Die differierenden Ergebnisse der separaten Komponenten-Analyse legen darüber hinaus nahe, auf die kompositen Indizes zu verzichten, um falsche Schlussfolgerungen hinsichtlich der exogenen Effekte zu vermeiden. Dieser Aspekt wird auch konzeptuell untermauert, da sich die Frage stellt, ob zeitliche Variabilität, inhaltliche Varietät und Regularität nicht drei genuin verschiedene Sequenzstrukturphänomene sind, die dementsprechend hinsichtlich ihrer Kovariation

mit exogenen Variablen separat zu analysieren wären. Auf der anderen Seite stellt sich mit den vorgeschlagenen Operationalisierungen dabei ein Konfundierungsproblem, das die Interpretation erschwert. So ist z.B. der Entropie-Wert als Indikator der Varietät auch mit Variabilität der Episodendauer korreliert oder die Episodenwechselrate mit der Anzahl der DSS. Ein Ausweg aus diesem Dilemma könnte darin bestehen, gleichsinnige Komponenten in einer anderen Weise zu kombinieren bzw. die Konfundierung durch die Integration der endogenen Variable als Kovariate. Diese Perspektiven sollen in einer nächsten Analyse untersucht werden.

8 Literatur

Elzinga Cees H. (2006). *Turbulence In Categorical Time Series*, Department of Social Science Research Methods, Vrije Universiteit Amsterdam, The Netherlands.

Elzinga Cees H. (2010). Complexity Of Categorical Time Series, *Sociological Methods Research February 2010,* 38(3), 463-481

Gabadinho, A., Ritschard, G., Müller, N. S. & Studer M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software,* 40(4), 1-37.

Papastefanou, G., Gruhler, J. (2014). Social status differentiation of leisure activities variation over the weekend – Approaching the voraciousness thesis by a sequence complexity measure. *Electronic International Journal of Time Use Research*, 11, 1-12.

Scherer, S., Brüderl, J. (2011) Sequenzdatenanalyse. In Christof Wolf, Henning Best (Hrsg.). *Handbuch der sozialwissenschaftlichen Datenanalyse*. Hamburg, Springer Verlag.

Stegmann, M., Werner, J., Müller, H. (2013). *Sequenzmusteranalyse. Einführung in Theorie und Praxis.* München, Rainer Hampp Verlag.

Anhang

Tabelle1: Distinkte Sukzessive Subsequenzen (DSS) bei verschiedenen Sequenzen

Sequenz-Nr. ->		1	2	3	4	5	6	7	8	9
Sequenz mit distinkt sukzessiven Zuständen ->		ABABAB	ABCBAB	ABABAB	ABCDEF	ABCABC	ABCABC	ABABC	ABABCB	ABABABAB
Anzahl DSS ->		33	46	33	64	52	52	24	41	88
		a aa aaa aaab aab aaba aabab aabb ab aba abaa abaab abab ababa ababab abab abba abbab abbb b ba baa baab bab baba babab babb bb bba bbab bbb	a aa aab ab aba abab abb abba abbab abbb abc abca abcab abcb abcba abcbab abcbb ac aca acab acb acba acbab acbb b ba bab bb bba bbab bbb bc bca bcab bcb bcba bcbab bcbb c ca cab cb cba cbab cbb	a aa aaa aaab aab aaba aabab aabb ab aba abaa abaab abab ababa ababab abab abba abbab abbb b ba baa baab bab baba babab babb bb bba bbab bbb	a ab abc abcd abcde abcdef abcdf abce abcef abcf abd abde abdef abdf abe abef abf ac acd acde acdef acdf ace acef acf ad ade adef adf ae aef af b bc bcd bcde bcdef bcdf bce bcef bcf bd bde bdef bdf be bef bf c cd cde cdef cdf ce cef cf d de def df e ef f	a aa aab aabc aac ab aba abab ababc abac abb abbc abc abca abcab abcabc abcac abcb abcbc abcc ac aca acab acabc acac acb acbc acc b ba bab babc bac bb bbc bc bca bcab bcabc bcac bcb bcbc bcc c ca cab cabc cac cb cbc cc	a aa aab aabc aac ab aba abab ababc abac abb abbc abc abca abcab abcabc abcac abcb abcbc abcc ac aca acab acabc acac acb acbc acc b ba bab babc bac bb bbc bc bca bcab bcabc bcac bcb bcbc bcc c ca cab cabc cac cb cbc cc	a aa aab aabc aac ab aba abab ababc abac abb abbc abc ac b ba bab babc bac bb bbc bc c	a aa aab aabb aabc aabcb aac aacb ab aba abab ababb ababc ababcb abac abacb abb abbb abbc abbcb abc abcb ac acb b ba bab babb babc babcb bac bacb bb bbb bbc bbcb bc bcb c cb	a aa aaa aaaa aaaab aaab aaaba aaabab aaabb aab aaba aabaa aabaab aabab aababa aababab aababb aabb aabba aabbab aabbb ab aba abaa abaaa abaaab abaab abaaba abaabab abaabb abab ababa ababaa ababaab ababab abababa abababab ababbab ababb abbba abbbab abbbb b ba baa baaa baaab baab baaba baabab baabb bab baba babaa babaab babab bababa babaab bababb babb babba babbab babbb bb bba bbaa


b b a a b
b b a b
b b a b a
b b a b a b
b b a b b
b b b
b b b a
b b b a b
b b b b