Papastefanou, G. (2016)

Measuring sequence complexity A conceptual and empirical comparison of two composite complexity indice

in G. Ritschard & M. Studer (eds), Proceedings of the International Conference on Sequence Analysis and Related Methods, Lausanne, June 8-10, 2016, pp 805-808.

Measuring sequence complexity

A conceptual and empirical comparison of two composite complexity indices

Georgios Papastefanou

GESIS Leibniz Institute for the Social Sciences, Mannheim/Germany

For a causality oriented analysis of sequences, namely by estimating their covariation with exogenous social variables like social-economic status, gender, age or attributes of family of origin, structural pattern, which characterizes a sequence, has to be represented by a quantifying indicator. One way is to capture a sequence's complexity, by constructing an indicator with a specific quantitative range. Two prominent approaches, Gabadinho et al. (2011) and Elzinga (2010) propose each a different approach to incorporate crucial features of sequence patterning like variety (qualitative differentiation of states), variability (temporal differentiation of states as episodes) and regularity (repetition of subsequences) .

Gabadinho et al. (2011) propose this formula:

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

As formula 1 indicates, two dynamic structural features are involved in the complexity index, namely the temporal variability on the one hand, as measured by the change of frequency and variety, as measured by the Shannon entropy. Moreover, this index takes into account the fact that individual sequences may vary in their lengths. Linking of the two normalized components change intensity and entropy is done by geometric mean, i.e. the square root of the product of normalized change intensity and normalized entropy.

The complexity index of T Elzinga (2010) is calculated using the following formula -2:

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

where:

Log$_2$== logarithm to the base 2
φ == number of sub-sequences with distinct successive states
$\boldsymbol{s^2{}_{t,max}(s)}$== maximum variance of episode durations in a sequence for a given number of episodes
$\boldsymbol{s^2{}_t(xs)}$== variance of the durations of episodes within a given sequence

Für $\boldsymbol{s^2{}_{t,max}(s)}$applies: $s^2_{t,max}(x) = \big(\ell_d(x) - 1\big)\big(1 - \bar{t}(x)\big)^2$

As both indicators want to be named as complexity indices, we will label the Gabadinho et al. approach as complexity C and Elzinga's definition as complexity T.

In a detailed conceptual analysis we discuss the foundation and restrictions of their components like transition rate, normalized entropy, number of distinct successive states and normalized episode duration variability. Further we examine interchangeability of C and T as is stated by Gabadinho et al (2011). We find – based on comparing C and T for nine systematically varied sequences – that there is nearly no co-variation between C and T.

Tabelle 1: Complexity indices C and T and their components for nine examplary sequences

sequence-Nr.	**Full sequence pattern**	**Distinct states sequenc pattern**	Transition rate	Normalized Entropy (sample-based)	**Complexity C**	**Ranking by C**	Phi (# DSS[1])	**Variance factor**	**Complexity T**	**Ranking by** nach T
1	AABBAABBAABB	ABABAB	0.45	0.39	**0.42**	**7**	33	6	**7.63**	**1**
7	AAAAAAAAAABBBAAAAAAAAAABBBCCCC	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	AAAAAAAAAABBBCCAAAAAAAAAABBBCC	ABCABC	0.17	0.48	**0.29**	**4**	52	5.9	**8.26**	**4**
8	AAAAAAAAAABBBAAAAAAAAAABBBCCCB	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	AAAAABBBBBAAAAABBBBBAAAAABBBBB	ABABAB	0.17	0.39	**0.26**	**2**	33	81	**11.38**	**7**
5	AAAAABBBBBCCCCCAAAAABBBBBCCCCC	ABCABC	0.17	0.61	**0.32**	**5**	52	81	**12.04**	**8**
4	AAAAABBBBBCCCCCDDDDDEEEEEFFFFF	ABCDEF	0.17	1	**0.41**	**6**	64	81	**12.34**	**9**

note: 1) number of distinct successive sub-sequences

The calculated rank correlation coefficient of these sample sequences is 0.02. This very low correlaton means, that one gets quite a different complexity ranking of these sequences, if one uses T or C.

For a more extensive test of the substitutability of T and C we did an empirical analysis of 2000 sequences of leisure activities on Sunday, based on the German Time Use Survey of 2001/2002. As a starting point we take the issue of complexity of the personal leisure time on weekends (Papastefanou, Gruhler 2014). For this purpose, we make use of the data collected in the time use survey of the Federal Statistical Office from 2001-2002. As alphabet of leisure activities on Sunday following activities are defined: reading, listening to music, watch television, computers, pursue hobbies, sports, and the residual category "other activities". As indicators of the socio-structural situation, the following variables are taken in account: gender, age, marital status, household income (interval categories), household size, general secondary education, vocational education and occupational status. For ease of interpretation, we restrict the target group to persons aged over 17 years who are employed full-time.

First, we find a high co-variation of both complexity indices C and T: for the group of full-time employees over 17 years the Pearson correlation coefficient is r = .94. The two composites complexity indices so appear to be interchangeable.

But a multivariate modeling of T and C of leisure time sequences on Sunday as outcome of social determinants like gender, age, family status, net household income, household size, education status and occupational status reveal significant differences between the models (see table 2).

Tabelle 2: Socio-demographic covariates of complexity scores and of their components (separate OLS-Regressions, b, standard error)

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Komplexität C	Komplexität T	Wechselrate	rel. Entropie	Log_2(phi)	Log_2(Varianzfaktor)
gender	**-0.0283*****	**-1.359*****	**-0.0133*****	**-0.0654*****	**-0.766*****	**-0.622*****
	(0.00472)	(0.245)	(0.00384)	(0.00870)	(0.141)	(0.181)
age	0.000281	**0.0226***	0.000106	**0.000995****	**0.0216*****	0.0152
	(0.000259)	(0.0136)	(0.000211)	(0.000477)	(0.00778)	(0.00991)
single	**0.0105***	-0.0727	0.00463	0.0118	0.00494	-0.208
	(0.00635)	(0.331)	(0.00517)	(0.0117)	(0.190)	(0.243)
divorced	-0.00510	0.00606	-0.00782	-0.00822	-0.0378	-0.178
	(0.00746)	(0.384)	(0.00605)	(0.0137)	(0.220)	(0.284)
widowed	-0.0141	0.100	-0.00992	-0.00983	-0.274	0.658
	(0.0210)	(1.072)	(0.0174)	(0.0393)	(0.615)	(0.816)
separated	-0.0174	-0.879	-0.00292	-0.0356	-0.412	-0.0629
	(0.0183)	(0.954)	(0.0151)	(0.0343)	(0.547)	(0.712)
Size of household	-0.000589	-0.0905	-0.000336	-0.00119	-0.0300	-0.0180
	(0.00185)	(0.0960)	(0.00151)	(0.00343)	(0.0551)	(0.0712)
Middle educational status	0.00167	0.356	-0.000253	0.00224	**0.351****	0.0659
	(0.00532)	(0.278)	(0.00431)	(0.00977)	(0.160)	(0.203)
High educational status	**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)
University degree	-0.00629	-0.00485	-0.00310	-0.0135	0.0646	-0.00153
	(0.00782)	(0.407)	(0.00630)	(0.0143)	(0.234)	(0.297)
Civil servant	**0.0220*****	**0.769****	**0.0116****	**0.0410*****	**0.520****	0.363
	(0.00708)	(0.368)	(0.00570)	(0.0129)	(0.211)	(0.268)
White collara	**0.0233*****	**1.083*****	**0.0162*****	**0.0527*****	**0.696*****	**0.914*****
	(0.00614)	(0.319)	(0.00497)	(0.0113)	(0.183)	(0.234)
Blue collar	**0.0274*****	**1.388*****	**0.0190*****	**0.0538*****	**0.892*****	**1.001*****
	(0.00694)	(0.363)	(0.00561)	(0.0127)	(0.208)	(0.264)
apprenticeship	0.0167	0.916	0.00877	**0.0433****	0.311	**1.007****
	(0.0112)	(0.584)	(0.00902)	(0.0204)	(0.335)	(0.424)
Military/ civil service	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

In brackets: Standard error. *** p<0.001, ** p<0.01, * p<0.1

We find, that in the model of the complexity of T a linear effect of age is estimated to be significant, which is not estimated significantly in the model of complexity of C. For example, the estimation of the model of the complexity C suggests that unmarried persons show significant higher complexity of leisure activities on Sunday as married ones. Also, it should be noted that the Model T is to determine a clearer differentiation in the differences between the professional status groups officials, employees and workers. According to this model, workers had the highest complexity, while the three status groups are not essentially different from each other for the model with C.

In sum it seems, that T and C might represent (at least partially) substantially different processes of sequence patterning. This assumption finds support in modeling the social effects on the components of C and T separately. In sum, we conclude that it might be more adequate to analyse the components of C and T separately instead of their joint incorporation into C and T, because the components especially like normalized entropy and number distinct successive subsequences seem to represent different processes of sequence differentiation like variety and regularity.

References

Elzinga Cees H., (2010), Complexity Of Categorical Time Series, *Sociological Methods Research February 2010 vol. 38 no. 3 463-481*

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