Time-Series Analysis of Psychoanalytic Treatment Processes: Sampling Problems and First Findings in a Single Case*

Hans-Joachim Grünzig

1. Introduction

For several years our research group at Ulm University has been investigating long-term psychoanalytic treatment processes. The principal data base consists of tape-recorded treatment sessions with subsequent verbatim transcriptions (see Kächele et al. 1973; Kächele et al. 1975; Kächele et al., this volume). There are of course many different problems to be solved, only one of which is the focus of this contribution. The verbatim transcription of the tape-recorded psychotherapeutic dialogue is an extraordinarily time consuming endeavor; it takes approximately 15 to 20 hours to transcribe one 50-minute session but the transcribing permits subsequent computerized studies. For economic reasons long-term psychoanalytic treatments extending over several hundreds of sessions can therefore be investigated only by means of time sampling. Different methods of time-series sampling may be considered e.g. random samples in different variations or a typical selection of treatment sessions in which characteristics occur which are specific for a given research question. Unfortunately, there is no sampling theory for time-series from which methods for arriving at time-series samples might be derived; the existing attempts to formulate a psychoanalytic process theory (see Thomä und Kächele 1975, 1987; Fürstenau 1977) are not yet precise enough in order to yield sampling criteria. For practical research problems we formulated a few sampling criteria which are not strictly founded in a theoretical sense:

1. The sample should cover the entire treatment process in a representative way; sessions from all important treatment phases should

---

*Supported by the German Research Foundation, Sonderforschungsbereich 129, project B3.

be included. However, the notion of representativity is not defined since the parameters of the distribution of the variables are not known for a single case.

2. For clinical reasons the sample should consist of blocks of several consecutive therapy sessions. Indeed, it is of special clinical interest to investigate short-term changes of variables across a number of sessions, e.g. variables such as "intervention effects," "mutative interpretation," etc.

3. Since the sampling error also is undefined and we therefore had no statistical reason for choosing any particular number of sampled sessions, we arbitrarily chose to include about 20% of the total number of sessions.

Once such criteria have led to selecting a particular sample of therapy sessions then the central question to be answered is: to what extent is there a correspondence between the view of the treatment process resulting from the investigation of the sample sessions and that which would result from an examination of the entire case? Or, to put it another way, would a different sample of different sessions arrived at by using the same criteria lead to the same process description? We have tried to answer these questions empirically by using selected variables in a single psychotherapy case. And the limitations of a single-case study should be kept in mind: the findings from this one psychotherapy can neither be necessarily generalized to other therapies, nor can the results for these particular variables be necessarily generalized to another set of variables. This report represents just a beginning exploration of the problems of time-series sampling in psychotherapy process research.

For these studies we employed the methods of computer assisted content analysis (CACA) using both the variables in the Ulm Anxiety Topic Dictionary (ATD) (see Angstthemenwörterbuch (ATW), Speidel 1979; Grünzig and Mergenthaler 1986) and a count of the number of words spoken by the patient. The ATD consists of Gottschalk and Gleser's (1969) four anxiety scales, i.e., *guilt, shame, mutilation anxiety,* and *separation anxiety*, operationalized as lists of individual words meant to represent each category. A computer program, EVA (see Grünzig et al. 1976), similar to the better known General Inquirer (Stone et al. 1966), performs the CACA by (1) comparing the transcribed text of each therapy session word by word with the ATD entries, (2) keeping track of the occurrences of text words that match the words in each of the four ATD categories, (3) summing the counts for each category,

(4) computing percentages of total text words matched in each category, and (5) counting the number of patient words spoken.

2. The First Two Samples

Several years ago the practical considerations mentioned above led to the following time sample *1a:* out of every 25 therapy sessions the first block of 5 consecutive sessions was transcribed and included in the total. Thus, after 5 consecutive sessions 20 sessions were omitted in the sample, followed by the next 5 consecutive sessions and so on; this time sample *1a* includes N = 110 sessions out of the total of 517 therapy sessions. With the exception of the end of the therapy the number of sample sessions is exactly 20% of all of the sessions and the sample blocks are evenly distributed across the entire therapy.

Another time sample, *1b,* consisted of blocks of 8 consecutive sessions interrupted by variable, random length intervals of omitted sessions. The block length of 8 sessions was chosen to allow for a reliable estimation of autocorrelations (see below) and the random intervals were chosen to compare with the 20-session fixed length omissions in sample *1a*. This time sample *1b* includes N = 112 sessions out of the total 517 therapy sessions. Although the length of the intervals between blocks was determined by random numbers (within a range), in order to minimize transcribing effort, we made minor corrections so that several sessions from the 5-session blocks that had been previously transcribed could be used for this alternative sample *1b*. Therefore these two samples were not entirely independent of each other. Table 1 shows these two samples; the correspondences in the selected sessions selected can be seen.

Our goal was to determine whether certain statistical analyses of these two time-series would lead to similar results. The theoretical basis for our analyses was the ARIMA methodology of Box and Jenkins (1976) (see also Glass et al. 1975; McCleary and Hay 1980).

For those readers who are not familiar with the ARIMA methodology, a brief summary together with a rough explanation of some technical terms is inserted. The aim of this methodology is to investigate regularities, i.e., serial dependencies within a given series of time ordered data values. A consistent increase or decrease of the time-series values over time is called a *trend* and is denoted by the letter *I* in the ARIMA abbreviation; *I* stands for *integrated* process and can adopt the values 0 (no trend; stationary time-series), 1 (for a linear trend), 2 (for a quadratic trend), etc. The value of *I* denotes the order of the integrated process.

Table 1 Schematic Presentation of the 5-Session-Block, Sample 1a and the 8-Session-Block, Sample 1b

Sample 1a 5-session-block N = 110	Sample 1b 8-session-block N = 112	Sample 1a 5-session-block	Sample 1b 8-session-block
I 1 - 5	1 - 8 **I**	**XII** 276 - 280	
II 26 - 30	25 - 32 **II**	**XIII** 300 - 304	297 - 304 **X**
III 51 - 55		**XIV** 326 - 330	
IV 76 - 80	73 - 80 **III**	**XV** 351 - 355	348 - 355 **XI**
V 101 - 105	98 - 105 **IV**	**XVI** 376 - 380	376 - 383 **XII**
	109 - 116 **V**		
VI 126 - 130		**XVII** 401 - 405	
VII 151 - 155	150 - 157 **VI**	**XVIII** 421 - 425	
VIII 176 - 180	172 - 179 **VII**	**XIX** 445 - 449	442 - 449 **XIII**
IX 202 - 206	202 - 209 **VIII**	**XX** 476 - 480	
X 221 - 225		**XXI** 502 - 506	
XI 251 - 255	249 - 256 **IX**	**XXII** 513 - 517	510 - 517 **XIV**

For a further description of serial dependencies, two different mathematical models are suggested: the *autoregressive* (AR) model and the *moving average* (MA) model. The *AR* model implies that the value of a variable at a given time point is dependent on the value(s) of the variable at the preceding time point(s); the number of the preceding time points determining the value of the given data point denotes the order of the AR model. For example, a 1st order AR model (AR(1) model) means that only the immediately preceding data point is determining the value at the given time point. An AR process may be thought of as a time-series having a kind of memory of the preceding data value(s).

The *MA* model implies that the value at a given time point is composed of a certain data level (which is assumed to be constant for the entire time-series) with a deviation caused by so called random shocks that occurred at the preceding time point(s); the number of the preceding time points at which those random shocks with an influence on the given data point occurred denotes the order of the MA model. For example, a 1st order MA model (MA(1) model) means that such a random shock occurred only at the immediately preceding time point. An MA process may be thought of as a homeostatic process over time, the time-series continuously trying to restore its values at the assumed constant level. Of course, all three models may be present in a given time-series.

In order to identify the presence of a serial dependency model, the autocorrelation function and the partial autocorrelation function have to be computed. Computing an *autocorrelation* means to correlate the values of the time-series with the same values being shifted for 1, 2, 3 or more time points (*lag*); at lag = 0 the autocorrelation yields a value of $r = 1$. The set of autocorrelations for several lags is called an autocorrelation function (*ACF*). For the computation of *partial autocorrelations*, similar to the partial correlations in classical statistics, the determining influence of the 1, 2, 3 or more preceding data points (for lag = 2, 3, 4 or more) on the given data point is eliminated and a, so to say, "uncontaminated" autocorrelation for a given lag is computed. At lag = 1, the autocorrelation is the same as the partial autocorrelation. The set of partial autocorrelations for several lags is called a partial autocorrelation function (*PACF*). The complex problem of estimating the model parameters is omitted at this place; the model parameters are further specifications of the time-series model identified.

After eliminating the identified serial dependency model from the time-series, the resulting *residual* time-series is said to be serially independent (white noise) if the ACF and the PACF of the residuals are non-significant; at the same time, this is a test of correctness for the dependency model identified.

3. Comparing the Samples *1a* and *1b*

For all 5 variables in both samples the time series are stationary without trend and with no clear periodicities. In the computation of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) the "lags" between the individual session blocks are taken into consideration. Table 2 shows the ACF and the PACF up to lag 4 for the 5 variables in samples *1a* and *1b*.

For the *guilt* category there are no serial dependencies; evidently this variable consists mainly of white noise. The high values for lag 3 of *mutilation* are hints for a periodicity which, however, is not supported by values of the higher order lags; the entire ACF up to lag 50 in both samples is not significant however. The ACF and PACF of *shame* are not typical and, with their low values, do not indicate a clear dependency model; a preliminarily computed moving-average model of first order, MA(1), in both samples leads to a significant parameter and to serially independent residuals. The same is true for *separation*. A clear autoregressive dependency model of first order, AR(1), can be seen, however, in the *patient word count*. Now, we have a first answer to our central question: "Is there a correspondence between the two time samples?" In all 5 variables the values of the ACF and the PACF in the two samples correspond to a very high degree; if at all, there is a slight difference for *patient word count*.

In Table 3 the model parameters as well as the ACF and the PACF of the residuals in the two samples are shown, after the computation of a MA(1) model for *shame* and for *separation* and an AR(1) model for *patient word count*. A constant and very high correspondence between the two samples can be seen in Table 3.

Table 2 Comparison of the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) in the Two Samples 1a and 1b for the 5 Variables

	A C F		P A C F	
1a	1b	1a	1b	
5-Sess. Block N = 110	8-Sess. Block N = 112	5-Sess. Block N = 110	8-Sess. Block N = 112	
Shame	.208	.208	.208	.208
-.003	.040	-.048	-.087	
.102	.036	-.076	.022	
.098	.119	-.031	.101	
Mutilation	.181	.151	.181	.151
.048	.116	.015	.096	
.243	.197	.212	.152	
.080	-.026	-.051	-.152	
Guilt .056	.048	.056	.048	
.078	.156	.075	.154	
.011	.027	-.001	.012	
-.015	-.055	-.033	-.102	
Separation	.257	.241	.257	.241
.138	.065	.077	.007	
-.009	.027	-.148	-.116	
.020	.064	-.095	.025	
Patient Word	.368	.461	.362	.461
Count	.245	.372	.127	.203
.202	.215	-.083	-.417	
.133	.198	-.353	-.963	

Table 3 Comparison of the Model Parameters and the Residuals in the Two Samples 1a and 1b

	Model Parameters	Residuals A C F	P A C F	
1a	1b 1a	1b 1a	1b	
Shame	X 8.68 9.12	-.022 .004	-.022 .004	
MA (1)	s 2.68 2.88	-.036 -.033	-.094 -.033	a_1
-.24*	-.25* .116 .169	.113 -.168		
.050	.073 .047 .074			
Separation	X 10.46 10.63	.026 .004	.026 .004	
MA (1)	s 3.52 3.99	.101 -.03	.100 -.03	
a_1	-.24* -.25* -.025 -.17	-.030	-.17	
	-.115 -.07	-.125	.07	
Patient Word Count	X 2937 2930	-.105 -.114	-.105 -.114	
Count	s 910 962	.079 .142	.069 .131	
AR (1)	a_1 .483* .492*	.219 .097	.238 .130	
	.051 .088	.101 .098		

4. Comparing Split-Half (odd-even) Samples, *2a* and *2b*

It is of course clear that a good part of this high degree of correspondence is to be accounted for by the overlapping sessions in the two samples. In order to establish two entirely independent time samples, in the sense of not having any sessions in common, sample 1b – 8-session blocks with variable intervals between – was split into two samples. Sample *2a* consisted of the therapy sessions with the even numbered blocks, and sample *2b* consisted of the therapy sessions with the odd numbered blocks. Since the initial and the final phase of a psychotherapy are typically conducted under special psychodynamic and technical conditions, the first and the last 8-session blocks were eliminated from this comparison of the two samples. Moreover these two blocks would have fallen in different samples and this could have led to unjustified differences between them.[1] Therefore these two samples *2a* and *2b* each included 6 blocks, i.e., 48 sessions. For the purpose of clarity the presentation of the findings will be restricted to the variable *patient word count* since for the *shame*, *guilt* and *separation* no clear dependency model emerged in these two samples; remarkably, *mutilation* in sample *2a* shows a clear AR(1) model although sample *1b* as well as sample *2b* are not significant in this respect.

As can be seen in Table 4 the ACF and the PACF of *patient word spoken* in sample *2a* – as was expected – indicates an AR(1) model; the parameter is significant and leads to white noise residuals. The values of the comparison sample *2b* again correspond very highly; the slightly larger differences are presumably accounted for by the smaller number of sample elements. Therefore comparing these two independent time samples again led to very similar results.

As the most conservative test of correspondence between these two samples a kind of cross-validation was performed: the autoregressive model of sample *2a* was used to predict sample *2b*. According to the AR(1) model this prediction was made as follows: the last value of the second block was taken to predict the first value of the third block; the first value of the third block was taken to predict the second value of this block, etc. Restricting the prediction to the immediately following value was necessary because a prediction over more than one time point leads to a steady increase in the error of prediction from AR(1) model. The confidence interval of the prediction therefore also steadily increases and the prediction cannot then be falsified (see methodical problems of time-series prediction: McCleary and Hay 1980).

If there really is a correspondence between the two samples the serial independence of the residuals in the "predicted" sample *2b* is to be expected. And in fact, the values of the ACF and the PACF demonstrate the serial independence of the residuals in the "predicted" sample *2b*. The mean of the predicted *patient word count* is 2787 words and the standard error of the prediction is 637; this means that the 95% confidence interval of the prediction comprises about 2500 spoken words; 15 of the 48 predicted values, that is, less than one third, are located outside of this confidence interval. In general it is true that the univariate prediction of

[1]In the previous sample comparison these two blocks could reasonably be considered because they were contained in both samples, and any irregularities would have affected both samples in a similar way.

Table 4 Comparison of the Split-Half (odd-even) Samples 2a and 2b

		Patient Word Count	
2a even numbered blocks N = 48	odd numbered blocks N = 48	2b	
ACF .348 .102	.377 .115	.231 .171 .191	.369
PACF .264	.384 -.135 -.726	.377 .098 -.233	-.376
PARAMETER	a_1 .347*	.464*	
Residuals ACF -.066 .083 .074	-.115 .217 .012		
PACF .079 .085	-.066 .206 .059	-.115	

stationary time-series without any periodicity includes substantial error variance and this error was of course present in this case. In fact when the mean of all of the values in sample *2a* was used as a constant predictor for the values in sample *2b*, only 16 values were located outside of the 95% confidence interval, just *one more* than we found by using the AR (1) model. Nonetheless these findings demonstrate the similarity between the two independent time samples and permit us to answer our central question positively. In all essential characteristics samples *2a* and *2b* have very similar corresponding values and it is therefore plausible that both represent the same actual therapy process.

5. Comparing the First with the Second Half of the Therapy

Next, consecutive samples were investigated. The 8-session block sample *1b*, covering the entire therapy process, was divided into two at the middle of the sample, omitting the first and last 8-session blocks. Again the two resulting samples were independent in the sense of not having any sessions in common. In contrast to the split-half (odd-even block) samples we could now compare the time-series analyses of the first half with those of the second half of the therapy.

On the grounds of simple clinical reasoning it is to be expected that the similarities between the first and second half samples would be smaller than those between the split-half samples, each of which covered most of the time span of the therapy. This is particularly true in the case of long-term psychoanalytic treatment (in this case 517 hours over about 4 years) in which one expects that the patient will change and that the measured variables will reflect that change. Thus it is reasonable to suppose that the serial dependencies that are detected by the autoregressive model will not remain constant over the course of the therapy. Rather it is to be expected that different treatment phases would yield different autoregressive models. We therefore examined the two halves separately.

Table 5 shows the ACF and the PACF for the two halves of the sample for the 5 variables. *Guilt* again is evidently mostly white noise (random variation), but the other 4 variables show more or less clear differences between the two samples. *Shame* follows a clear AR (1) model in the second part of the therapy whereas in the first half it does not. *Mutilation* in the first half as well as *separation* in the second half are characterized by a weak AR (1) process. *Patient word count* however appears to be quite a homogeneous process; both in the first and in the second half samples as well as in sample *1b* (covering the entire therapy time span) there is a clear first order autoregressive process which is considerably stronger in the second half (a_1 = .560) than in the first half (a_1 = .314). In both halves the residuals are serially independent.

Table 5 Comparison of Two 'Consecutive' Samples Out of the 1st and the 2nd Half of the Therapy

	A C F		P A C F		
1st half N = 48	2nd half N = 48	1st half N = 48	2nd half N = 48		
Shame -.305 .036	.145 .075 .257	.314 -.036 -.247	.145 .130	.314 -.277 .045 -.052	.167 .204
Mutilation .101 -.182	.230 .171 -.033	.092 .205 -.300	.230 .143	.092 .149 .111 .110	.178 .211
Guilt .152 -.001 -.310	-.101 .121 -.158	.152 -.042	-.101 .012	.355 .009 -.052 .026	.340 .134
Separation -.034 -.195	.099 .000 .172	.229 -.044 -.051	.099 -.055 .137	.229 .192 -.134 -.019	 .194
Patient Word Count -.011	.288 .270 -.240	.380 .395 .215	.288 .204 .054	.380 .293 .194 -.126 -.021	 .208

6. Discussion

Thus the results of our time-series analyses using 5 computer generated content measures have provided a reasonably clear (if nonetheless limited) answer to the important question: will different methods of sampling the sessions of an entire long-term psychoanalytic case yield similar or different results? And the answer is, *both*, depending on the kind of sample. In both samples *1a* and *1b*, each of which covered the entire span of the treatment, but which differed in the number of sequential sessions in each block and in whether a fixed or random number of sessions were skipped between each block, the 4 content variables were all stationary with no trends or periodicities while the number of *patient words spoken* followed an autoregressive dependency model. And in samples *2a* and *2b*, each derived from the odd-even blocks of *1b* (and therefore also covered the entire treatment span), the results were essentially similar to those of *1a* and *1b*. Thus, all four samples that spanned the whole treatment, though differing in other respects, were, by these measures, essentially similar.

On the other hand the last two samples each covered only half of the time span of the treatment. In these samples the results for 4 of the variables were different in the two halves, supporting the clear commonsense expectation that if the patient changes, i.e., gets better or worse, one ought to find differences between the first and second halves of the therapy if the variables used are at all appropriate, as these appear to have been.

Although the application of time-series analyses here has been limited to answering the one question about different methods of sampling, there are many other potential applications of this technology in the field of psychotherapy research, especially that of long-term psychoanalytic treatment where there will always be a great many potential data points. There is now a range of questions that are of great interest for which time-series analyses are both powerful and appropriate. In this study we used only the patient's text, but many of the most pressing questions in psychotherapy research involve the interaction between the therapist and the patient. Does the therapist suggest content to a patient who then follows with his own elaboration? Does the therapist's focus on the transference (see Hoffman and Gill, this volume) have consequences different from his not doing so? Are the referential activity levels (see Bucci, this volume) of the therapist and patient related to the therapeutic outcome? Are the different consequences of the application of techniques derived from different theories of change in psychotherapy and psychoanalysis (see Silberschatz, et al., this volume)? Indeed, Hohage and Kübler's (this volume) study of the relationship between the emotional and cognitive aspects of insight yielded the kind of data that would be suitable for rigorous time-series analyses.

Whenever one has reason to believe that two or more extended series of events occurring over time are related, as is the case with the interactions of therapist and patient, one is always interested in the *causal* nature of the interaction. Who mainly influences or follows whom? Does the patient adapt to the therapist or vice versa or neither? And so on. It is in principle within the power of appropriate time-series analyses to answer such questions, provided one wisely chooses the variables to measure. Of course such wisdom is not evenly distributed and that is an obstacle. But the larger practical obstacle at this time is the widespread ignorance among psychotherapy researchers of the techniques of time-series analyses. The reader is referred to Gottman's (1981) introduction to the subject for social scientists for nearly painless access to these important research tools.