# STATISTICAL CONSIDERATIONS IN THE ANALYSIS OF SOLAR OSCILLATION DATA BY THE SUPERPOSED EPOCH METHOD*

S. E. FORBUSH, M. A. POMERANTZ, S. P. DUGGAL†, and C. H. TSAO

*Bartol Research Foundation of The Franklin Institute, University of Delaware, Newark, Delaware 19711, U.S.A.*

**Abstract.** Although the method of superposed epochs (Chree analysis) has been utilized for seven decades, a procedure to determine the statistical significance of the results has not been available heretofore. Consequently, various subjective methods have been utilized in the interpretation of Chree analysis results in several fields. The major problem in the statistical treatment of Chree analysis results arises from the fact that in most studies of natural phenomena, data are neither random nor sequentially independent. In this paper, a statistical procedure which takes this factor into account is developed.

## 1. Introduction

For investigating the possible relationship between two sets of geophysical observations, Chree (1912, 1913) introduced a procedure for analyzing one set of measurements during epochs which were selected on the basis of a specific type of feature in the second set of measurements. The method of superposed epochs can also be used for investigating basic periodicities in time series of data (see e.g., Chapman and Bartels, 1940). This version of the superposed epoch technique is germane in the present context, since it has been utilized by several groups for investigating certain aspects of global solar oscillations (Severny *et al.*, 1976; Scherrer *et al.*, 1979; Grec *et al.*, 1980).

However, unfortunately, despite its long history, a proper statistical test for evaluating the significance level of the results obtained by superposed epoch analysis has not been available heretofore. This lack of a quantitative 'figure of merit' of the results of applications of the Chree procedure has led to controversial situations arising from different interpretations of the reality of an apparent signal. The fact that proper statistical methods have generally not been available for assessing Chree analysis results arises from a fundamental problem: Data representing natural phenomena are neither random nor sequentially independent. Consequently, the basic criterion for the application of standard statistical procedures is, in fact, violated.

The pitfalls of ignoring the non-randomness of data representing observations of natural phenomena were first demonstrated by Bartels (1935; see also Chapman and Bartels, 1940). He introduced the concept of quasi-persistency and developed a procedure for calculating the standard error by evaluating the extent of its effect. In this

* Proceedings of the 66th IAU Colloquium: *Problems in Solar and Stellar Oscillations*, held at the Crimean Astrophysical Observatory, U.S.S.R., 1–5 September, 1981.

† Shakti P. Duggal died, July 11, 1982.

*Solar Physics* **82** (1983) 113–122. 0038–0938/83/0821–0113$01.50.

paper we will describe a statistical procedure based on analysis of variance that takes into account the quasi-persistency and is suitable for testing the significance of Chree analysis results. (For a complete review, see Forbush *et al.*, 1982.)

It should be emphasized that the purpose here is not to discuss previous analyses of solar oscillation data by superposed epoch analysis. Rather, we wish to issue a caveat to the solar-physics community that conclusions based upon superposed epoch analysis must be viewed with extreme skepticism unless it is unambiguously demonstrated that both the nature of the statistical tests that are applied, and the assignment of error bars or other indices of the probable reality of a signal are strictly legal.

## 2. Chree Analysis

Let us assume that on the basis of some observational criterion (e.g., an unusually high value of a particular geomagnetic index) $N$ key-days are associated with some variation in the data under investigation (e.g., the cosmic ray intensity). In the method of superposed epoch analysis, each key-day is designated as the center of an epoch (day zero), the length of which is selected on the basis of a physically plausible period (e.g., the 27-day solar rotation period). We then list the data in the form of a matrix in which the rows $r_j$ represent the epochs, and the columns $c_i$ represent days before and after the individual key-days $c_{13}$ as in Table I.

The column averages of this matrix, which will invariably show some variations, represent the Chree analysis result. We will refer to this result as the signal. The objective of the procedures described in this paper is to determine its significance level (i.e. the probability that it did not occur by chance).

Classical statistical procedures may be (and ordinarily are) followed for evaluating the apparent significance level of the variance attributable to the signal. However, as will become clear later, this is grossly erroneous because the data for any epoch are not sequentially independent. In general, there are real effects, in addition to the one under study, which can cause the measured phenomenon to vary over different time scales. This is the nub of the problem.

## 3. Statistical Test

### A. Analysis of Variance

Table I represents the data matrix in a typical Chree analysis. In this example, we assume that there are 150 epochs ($r = \sum r_j = 150$) each comprising 27 days ($c = \sum c_i = 27$). The statistical test of the resulting signal can be performed as follows:

(1) Remove the linear trend, if any, from each row $r_j$.

(2) Calculate the variance of the population $S_c^2$ from the column means $\bar{c}_i$:

$$S_c^2 = r \sum_{i=1}^{c} \frac{(\Delta \bar{c}_i)^2}{c-1} , \tag{1}$$

https://doi.org/10.1017/S0252921100095452 Published online by Cambridge University Press

where

$r$ = total number of rows,

$c$ = total number of columns.

(3) Calculate the variance $S_r^2$ of the population from the row means $\bar{r}_j$:

$$S_r^2 = c \sum_{j=1}^{r} \frac{(\Delta \bar{r}_j)^2}{r-1} . \tag{2}$$

(4) Calculate the total variance $S_T^2$ from individual data points $(x_{ij})$:

$$S_T^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(\Delta x_{ij})^2}{rc-1} . \tag{3}$$

(5) Calculate the residual variance of the population $S_R^2$:

$$S_R^2 = \frac{[(rc-1)S_T^2] - [(c-1)S_c^2] - [(r-1)S_r^2]}{(c-1)(r-1)} . \tag{4}$$

(6) Test whether the variances of single rows and columns are homogeneous. One possible test is described in Appendix I.

(7) At this stage, let us first assume for simplicity that there is no quasi-persistency in the data. Under this assumption (which in most cases is not valid) the signal variance which is represented by $S_c^2$ can be compared with the residual variance $S_R^2$ of the data by using the $F$ test with $(c-1)$, $(c-1)(r-1)$ degrees of freedom, df. If this test reveals

TABLE I

Data matrix that is generally used in Chree analysis. $x_{ij}$ represents a single data point for column $i$ and row $j$. $\bar{c}_i$ and $\bar{r}_j$ represent the averages for columns and rows respectively. Day 0 is termed key day.

CHREE MATRIX – $\alpha_{ij}(1)$

| DAY → | | | | | |
|---|---|---|---|---|---|
| EPOCH ↓ | – 13 | – 12 | ---------- 0 ---------- | + 12 | + 13 |
| | 1 | 2 | ---------- 13 ---------- | 26 | 27 |
| 1 | | | | | $\bar{r}_1$ |
| 2 | | | | | $\bar{r}_2$ |
| , , , , , , | | | $x_{ij}$ | | $\bar{r}_j$ |
| 150 | $\bar{c}_1$ | $\bar{c}_2$ | $\bar{c}_i$ | $\bar{c}_{26}$ | $\bar{c}_{27}$ |

https://doi.org/10.1017/S0252921100095452 Published online by Cambridge University Press

that the signal is not significant, there is no need to proceed further, since the determination of quasi-persistency leads only to an increase in the residual variance.

## B. QUASI-PERSISTENCY

In order to provide a physical picture, let us assume that the signal in each row can be represented by a sine wave. In this case, following Bartels (1935), we define quasi-persistence as a periodicity which repeats for a certain number of epochs with approximately the same phase and amplitude, each such sequence ending more or less abruptly without any relation to other sequences. An example of quasi-persistent vectors derived from simulated data (see Appendix 2) is shown in Figure 1. To determine the

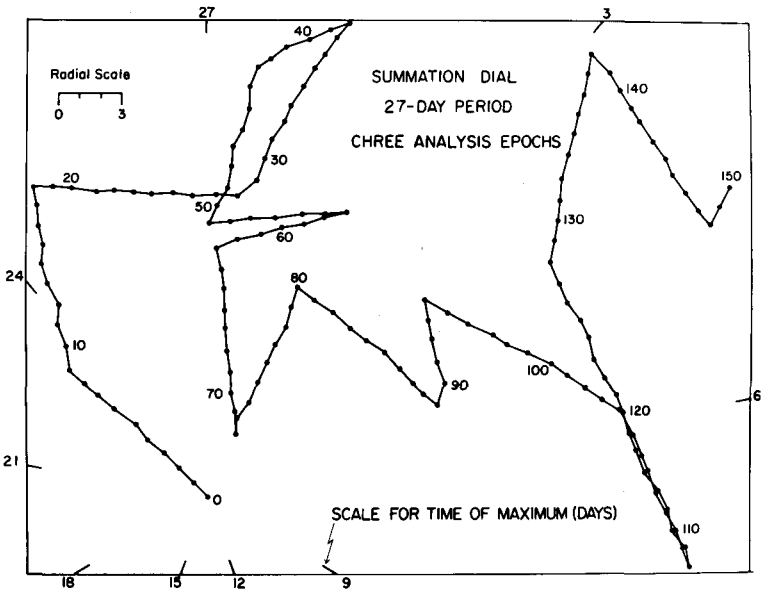


Fig. 1. Summation dial for the 27-day period. Vectors represent the 150 epochs of the simulated data. Every 10th epoch is marked at the end of the corresponding vector.

realistic residual variance, $S_R^2$ for quasi-persistent data, we transform the original Chree analysis matrix $\alpha_{ij}$ (1) into a new matrix in which the first row represents the mean of the first two rows of the original matrix, i.e.:

$$X'_{i1}\,(\text{new matrix}) = \frac{(X_{i1} + X_{i2})}{2},$$

where the columns $i$ extend from 1 to 27.

Similarly, the 2nd row of the new matrix $\alpha_{ij}$ (2), represents the average of the 3rd and 4th rows of $\alpha_{ij}$ (1). Following the same procedure, we construct $h$ matrices:

$\alpha_{ij}$ (1) – Original data matrix with $r$ rows.

$\alpha_{ij}$ (2) – Each row represents average of 2 consecutive rows of matrix $\alpha_{ij}$ (1). Total number of rows = $r/2$.

https://doi.org/10.1017/S0252921100095452 Published online by Cambridge University Press

$\alpha_{ij}(3)$ – Each row represents average of 3 consecutive rows. Total number of rows = $r/3$.

$\alpha_{ij}(h)$ – Each row represents average of $h$ rows. Total number of rows = $r/h$.

For each matrix, we repeat the first five steps described in the last section to obtain the residual variances $S_R^2(1), S_R^2(2), S_R^2(3) \ldots S_R^2(h)$ corresponding to the aforementioned $h$ matrices. It is evident from Appendix III that if the data contain quasi-persistency, and if it is assumed that there is no persistent signal in the epochs, the ratio

$$\frac{S_r(h)h^{1/2}}{S_r(1)} = \frac{\zeta(h)}{\zeta(1)} \tag{5}$$

will exhibit some relationship with $h^{1/2}$. Note that in the analysis of variance, the signal is eliminated from the residual variance (4). However, this relationship will break down at some limiting value: $\zeta(h)/\zeta(1) = \zeta(\infty)/\zeta(1)$. In fact the quasi-persistency is negligible beyond $\sigma = [\zeta(\infty)/\zeta(1)]^2$ rows.

In the above discussion, it has been assumed that the chronological order of the epochs has been maintained in all the matrices $\alpha_{ij}(1), \alpha_{ij}(2) \ldots \alpha_{ij}(h)$.

## C. Data Analysis

To clarify the procedure described thus far, and to demonstrate the pitfalls of using ordinary statistical tests for evaluating the significance level of Chree analysis results, let us consider the vectors shown in Figure 1. The data corresponding to these vectors (Appendix II) consist of a matrix with $c = 27$ and $r = 150$. The Chree analysis results derived from these data, i.e., the column means of this matrix, are plotted as a function of the day (column number $c_j$) in Figure 2. At first sight, Figure 2 reveals an impressive trend. In general, there appears to be a significant difference between the column means before and after the key-day (0-day). In fact, the plot in Figure 2 has the distinct appearance of a sine wave. Now let us examine, by using an ordinary statistical test, whether the variation evident in Figure 2 is actually significant. In other words, let us ignore the quasi-persistency in the data and perform the calculations enumerated in the seven tests outlined in Section A above. The final results are shown in Table II (random data). Application of the $F$-test reveals that the probability that the signal in Figure 2 has appeared by chance is very low (3 in $10^{15}$). On the basis of this result, the 'signal' would be accepted as real.

An examination of Figure 1 suggests that this result cannot be valid because the vectors show abrupt change in direction after each sequence. Let us now apply the new statistical test described in the preceding section. To evaluate the quasi-persistency, we plot $\zeta(h)/\zeta(1)$ vs $h^{1/2}$ in Figure 3. It is clear from this figure that $\zeta(h)/\zeta(1)$ shows a definitive relationship with $h^{1/2}$ up to a point ($h \approx 4$) where the relationship breaks down. The break occurs at $\zeta(h)/\zeta(1) = \zeta(\infty)/\zeta(1) \approx 2.4 = \sqrt{\sigma}$. Thus the equivalent length of sequences is $[\zeta(\infty)/\zeta(1)]^2 = 5.76$. In other words, the quasi-persistency lasts for about six rows, which is consistent with the appearance of the summation dial in Figure 1. On the basis of this derived equivalent length of sequences, the results listed in Table II are

https://doi.org/10.1017/S0252921100095452 Published online by Cambridge University Press

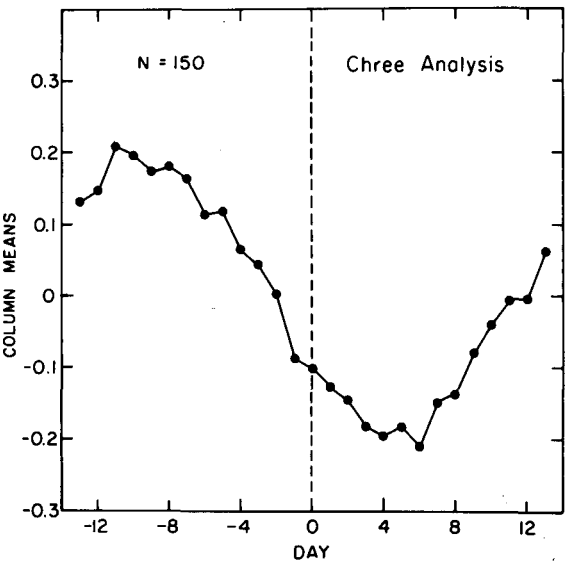


Fig. 2. Column means ($\bar{c}_i$) representing Chree analysis results are plotted as a function of days (Table I). Total number of epochs for this analysis is 150.

TABLE II

Comparison of the signal variance in the Chree analysis result with the residual variance in order to evaluate the probability that the signal has appeared by chance.

| Analysis of variance | Random data: quasi-persistency ignored | Non-random data: quasi-persistency included |
|---|---|---|
| Signal variance, df* = 26 | 2.94 | 2.94 |
| Residual Variance, df = 3874 | 0.59 | 3.40 (0.59 × 5.76) |
| F(26, 3874) | 4.96 | 1.16** |
| Probability that the signal has appeared by chance (1 − $P$) | $3 \times 10^{-15}$ | $3 \times 10^{-1}$ |

* df = degrees of freedom.

** Note that for this case, residual variance is larger than the signal variance, hence F(3874, 26) = (residual variance/signal variance) > 1, in accordance with standard practice (Hald, 1952).

obtained. The probability that the 'signal' has appeared by chance has increased to 0.33; hence, the apparent effect displayed in Figure 2 is *not* significant.

## 4. Discussion and Conclusion

For obtaining information concerning periodicities, or for understanding the relationship between two phenomena, superposed epoch analysis is unquestionably a useful procedure. However, this method of analysis yields meaningful results only if the

https://doi.org/10.1017/S0252921100095452 Published online by Cambridge University Press

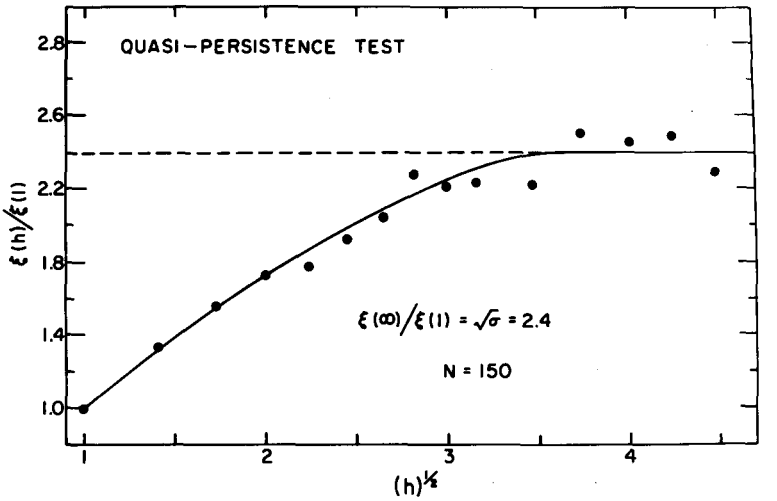


Fig. 3. Plot of the ratio $\zeta(h)/\zeta(1)$ vs $h^{1/2}$ derived from 150 epochs of Table I. The increase in the ratio for low values of $h^{1/2}$ indicates quasi-persistency in the data.

inherent quasi-persistency of natural phenomena is properly taken into account in the evaluation of the result. In fact, as was originally emphasized by Bartels (1935), the proper evaluation of quasi-persistency is of utmost importance in all types of analysis of problems in geophysics (and astrophysics). We have demonstrated here that the standard error can be grossly underestimated by ignoring the almost inevitable quasi-persistency in the data. In a statistical analysis of simulated data, we showed that ordinary (textbook) statistical tests led to the incorrect conclusion that the signal in Figure 2 is highly significant, whereas in reality it is not.

A new method, based on two-way classification analysis of variance, has been developed to determine the quasi-persistency (equivalent length of sequences) in the data. This analysis revealed that the effective (independent) number of epochs (rows) in the Chree matrix is much less than that determined under the invalid assumption that the data are strictly random. Thus, the standard error is modified, and the new effective standard error is found to be *larger* than the signal. It should be pointed out that an alternative method based on vectorial representation can be used to test the Chree analysis result. A study is in progress to determine the relative merits of the two procedures in various cases. It is hoped that application of the procedure developed here for including the effects of quasi-persistency in evaluating the statistical uncertainty of superposed epoch results will lead to more objective conclusions in future studies utilizing this powerful analytical tool.

## Acknowledgements

This research is supported by the National Science Foundation's Division of Polar Programs under grants DPP–7923218–01 and DPP–7822467 and Atmospheric Research Section under grant ATM–8005866.

## Appendix I: Test for Homogeneity

The hypothesis to be tested is that the variances of $k$ normally distributed populations are equal. If there is no quasi-persistency in the data, Bartlett's test (see e.g., Hald, 1952; Dixon and Massey, 1957) can be utilized to test this hypothesis.

Let the variance of the $i$th sample of size $n_i$ be given by $S_i^2$. Note that the sample size will take care of the fact that data in some rows or columns are missing.

Let

$$\eta = (N-k)\ln S_p^2 - \sum (n_i - 1)\ln S_i^2 ,$$

$$S_p^2 = \sum (n_i - 1) S_i^2/(N-k) ,$$

$$A = \frac{1}{3(k-1)}\left[\sum \frac{1}{n_i - 1} - \frac{1}{N-K}\right] ,$$

$$v_1 = k - 1 ,$$

$$v_2 = \frac{k+1}{A^2} ,$$

$$b = \frac{v_2}{1 - A + (2/v_2)} ,$$

$$N = \sum n_i .$$

Then the sampling distribution of $F = v_2\,\eta / v_1(b - \eta)$ is approximately $F(v_1, v_2)$. It should be emphasized that this test is not valid for non-independent data. In case there is quasi-persistency, the equivalent length of sequences ($\sigma$) can be evaluated. Bartlett's test can then be applied to sets of $k/\sigma$ independent samples.

## Appendix II: Data Simulation

The simulated data $D(t)$ for each epoch, for the tests described in this paper, are generated from:

$$D(t) = R_q \sin\left[\omega t + \phi_q(t)\right] + \zeta(t) + \beta t ,$$

where $R_q$, $\phi_q(t)$ represent amplitude and phase of a quasi-persistent signal, $\omega = 2\pi/27$ and $\zeta(t)$ and $\beta t$ represent random and linear effects in each epoch.

Harmonic analysis of these simulated data, after linear term corrections, yields the 27-day period vectors in the summation dial in Figure 1. Note that each vector represents a single epoch row in Table I.

https://doi.org/10.1017/S0252921100095452 Published online by Cambridge University Press

## Appendix III: Evaluation of Quasi-Persistency

Let

$M_i(h)$ = $i$th among $N/h$ means of $h$ consecutive means,
$N$ = total number of $M_i(1)$,
$N(h)$ = total number of $M_i(h)$,
$r_i(1)$ = contribution of random effects to $M_i(1)$,
$r_i(h)$ = contribution of random effects to $M_i(h)$,
$q_i(1)$ = the quasi-persistent contribution to $M_i(1)$,
$q_i(h)$ = the quasi-persistent contribution to $M_i(h)$,
$m$ = the contribution of the persistent wave to $M_i(1)$, constant for all $i$ from 1 to $N$,
$c^2(1)$ = variance of $M_i(1)$,
$c^2(h)$ = variance of the means of $h$ successive sequential means,

$$M_i(1) = [m + q_i(1) + r_i(1)] ,$$

$$c^2(1) = \frac{1}{N} \sum_1^N [m + q_i(1) + r_i(1)]^2$$

$$= \frac{1}{N} \sum \{[(m + q_1(1)]^2 + r_i^2(1)\} .$$

Since for large $N$, $\sum m r_i(1) = 0$ and $\sum r_i(1) q_i(1) = 0$,

$$c^2(1) = \frac{1}{N} \sum [m^2 + 2mq_i(1) + q_i^2(1) + r_i^2(1)]$$

$$= m^2 + 2m\bar{q}(1) + S_q^2(1) + S_r^2(1) . \tag{1A}$$

Since

$$\frac{2m \sum q_i(1)}{N} = 2m\bar{q}(1) ,$$

$\bar{q}(1)$ = mean of all quasi-persistent steps

and

$$\sum r_i^2(1)/N = S_r^2(1) ,$$
$$\sum q_i^2(1)/N = S_q^2(1).$$

Similarly

$$c^2(h) = \frac{1}{N/h} \sum_1^{N/h} [m^2 + 2mq_i(h) + q_i^2(h) + r_i^2(h)]$$

$$= m^2 + 2m\bar{q} + S_q^2(h) + S_r^2(h) . \tag{2A}$$

https://doi.org/10.1017/S0252921100095452 Published online by Cambridge University Press

Since

$$2m \sum_{1}^{N/h} q_i(h) = 2m\bar{q} .$$

Then

$$\begin{aligned} c^2(h)h &= m^2h + 2mh\bar{q} + S_q^2(h)h + S_r^2(h)h , \\ c^2(h)h &= h(m^2 + 2m\bar{q}) + S_q^2(h)h + S_r^2(1) . \end{aligned} \tag{3A}$$

As a special case, assume that the data contain no persistent wave, i.e. $m = 0$, then

$$c^2(h)h = hS_q^2(h) + S_r^2(1) . \tag{4A}$$

For large values of $h$, the right-hand side becomes constant, i.e.

$$c^2(h)h = \text{const.} = c^2(1)\sigma , \tag{5A}$$

where $\sigma$ is defined as 'equivalent length of sequences' (Bartels, 1935).

Equation (5A) can be written as

$$c(h)h^{1/2}/c(1) = \zeta(h)/\zeta(1) \approx \zeta(\infty)/\zeta(1) = \sigma^{1/2} .$$

## References

Bartels, J.: 1935, *Terrs. Magnetism Atmospheric Electricity* **40**, 1.
Chapman, S. and Bartels, J.: 1940, *Geomagnetism*, Vol. II, Oxford University Press.
Chree, C.: 1912, *Phil. Trans. London* **A212**, 75.
Chree, C.: 1913, *Phil. Trans. London* **A213**, 245.
Dixon, W. J. and Massey, F. J.: 1957, *Introduction to Statistical Analysis*, McGraw-Hill Book Co.
Forbush, S. E., Duggal, S. P., Pomerantz, M. A., and Tsao, C. H.: 1982, *Rev. Geophys. Space Phys.*, in press.
Grec, G., Fossat, E., and Pomerantz, M.: 1980, *Nature* **288**, 541.
Hald, A.: 1952, *Statistical Theory with Engineering Applications*, John Wiley and Sons.
Scherrer, P. M., Wilcox, J. J., Kotov, V. A., Severny, A. B., and Tsap, T. T.: 1979, *Nature* **277**, 635.
Severny, A. B., Kotov, V. A., and Tsap, T. T.: 1976, *Nature* **259**, 8.