CHAPTER 2

GENERAL METHODS OF ANALYSIS
2.1 INTRODUCTION

It is impractical to manipulate the variables, viz., sunspot number and atmospheric temperature, to investigate experimentally to know if one of them is measurable by varying the other. So a statistical study, using the correlation and regression analyses of the two variables, was undertaken, and an attempt to interpret the results based on probability and correlation measures was made in the present study.

Literature shows that various types of statistical methods have been used in the aerospace science by several workers. Inter-level correlations between various neutral atmospheric parameters have been studied by Quiroz and Miller (1968) who calculated height-lag correlations of density with pressure and temperature in the stratosphere using rocketsonde data. Justus and Woodrum (1973), using a statistical daily-variation method, analysed irregular variations of neutral atmospheric parameters in the 50–200 km region from rocketsonde and other data. Correlation methods have also been extensively used in the study of ionosphere drifts by Sastri and Rao (1971).

With specific reference to relationships between solar and geomagnetic activity and upper atmospheric parameters,
several statistical correlation analyses have also been carried out. Chandra and Krishnamurty (1968) analysed the thermospheric neutral density and temperature data obtained from the drag analysis of the Explorer - 39 satellite to determine the correlation with solar and geomagnetic activity characterised by the solar radio flux and the $K_p$ index. Detailed studies have been made on this subject by Jachia (1970). Investigations have also been carried out to find the response of the daily variations in neutral temperatures in the mesosphere to variations in solar and geomagnetic activities by using the correlation and regression analysis (Ramakrishna and Seshamani, 1973; Seshamani, 1976,1980). The sunspot cycle - related temperature variations in the stratosphere were also studied by using the correlation analysis (Angell and Korshover, 1978; Quiroz, 1979).

A method of analysis suitable for studies of middle atmospheric temperature data obtained by rocket soundings has been developed in this Chapter. A detailed description of the method applied to the study of relationship between the middle atmospheric temperature variations and variations in sunspot number by the present author is given as follows:
2.2 ANALYSING TECHNIQUES

2.2.1 Temperature data

The middle atmospheric temperature data, obtained from rocketsondes, which were processed and corrected for errors due to aerodynamic heating, lag, emissivity, self heating, conduction and long wave and short wave radiational heating, using computerised quality control procedures are collected from the appropriate publications, viz., (1) Bulletin of the Central Aerological Observatory (World Data Centre - B), Hydrometeoro logical Services, Moscow, USSR and (2) High Altitude Meteorological Rocket Data (World Data Centre - A), Asheville, North Carolina, USA. The data from the former publication are available in the form of individual (weekly) data, whereas the latter source gives statistical monthly mean values of temperature. Such temperature data are available at 1 km intervals in altitude in the middle atmosphere for a specific period, say N number of years.

In case there is non-availability of data for a period (due to failure of successful launching or otherwise) then the interpolated data have been used. This is done by means of graphical interpolation method in which temperature data averaged for a specific month for each level of the atmosphere over the entire period are plotted to enable to
draw a smoothed curve. From this curve, the data for the missing months and for each level are extracted. A feature of this interpolation technique is that short period oscillations do not influence the data and at the same time maintain the long period trend.

The weekly as well as monthly mean temperature data, as obtained above, are contaminated due to the effects of seasonal, semi-annual and other short period fluctuations such as sudden warming, planetary wave activity, auroral effect, etc. In order to filter out these short period oscillations and retain the long period trend in temperature variations, the data is averaged to annual mean values.

Let $T_{m,n,L}$ be the temperature observed from an individual sounding $m$, year $n$, and level $L$. The annual mean values can be calculated as

$$T_{n,L} = \frac{1}{M} \sum_{m=1}^{M} T_{m,n,L}$$

where $M$ is the total number of soundings made in an year or a season.

2.2.2 Sunspot number data

Monthly as well as annual mean Zürich sunspot number, $R_n$, data are taken from the 'Solar and Geophysical Data' published by the World Data Centre-A, Boulder, Colorado, USA for the required period.
2.3 COMPUTATION OF CORRELATION COEFFICIENT

The correlation coefficient, \( r(T_L, R, L, n) \) between the temperature, \( T_n, L \), at a given layer \( L \), and the sunspot number, \( R_n \), for a period of \( N \) number of years is calculated by using the Pearson's product moment expression,

\[
\begin{align*}
    r(T_L, R, L, n) &= \frac{\sum_{n=1}^{N} (T_{n, L} - \langle T_L \rangle)(R_n - \langle R \rangle)}{\sqrt{\sum_{n=1}^{N} (T_{n, L} - \langle T_L \rangle)^2 \sum_{n=1}^{N} (R_n - \langle R \rangle)^2}} \\
    &= \frac{\sum_{n=1}^{N} (T_{n, L} - \langle T_L \rangle)(R_n - \langle R \rangle)}{\sqrt{\sum_{n=1}^{N} (T_{n, L} - \langle T_L \rangle)^2 \sum_{n=1}^{N} (R_n - \langle R \rangle)^2}}^{1/2}
\end{align*}
\]

where the variable within the angle brackets denotes the average value over the entire period, consisting of \( N \) years, i.e.,

\[
\begin{align*}
    \langle T_L \rangle &= \frac{\sum_{n=1}^{N} T_{n, L}}{N} \\
    \langle R \rangle &= \frac{\sum_{n=1}^{N} R_n}{N}
\end{align*}
\]

2.3.1 Test of significance of the correlation coefficient

Assuming that the variations in temperature and those in sunspot number possess a joint normal distribution about the mean, according to the method for small samples, the resulting coefficients of correlation are treated with the 'Student's double tailed t-test' (Walker and Lev, 1958), which takes the form
\[ t(r) = \frac{r \sqrt{N-2}}{\sqrt{1-r^2}} \] ...

where \((N-2)\) is the effective number of degrees of freedom.

The computed \(t\)-value is then compared with the tabular value (available in standard statistical tables) for \((N-2)\) degrees of freedom, to find out the level of significance of the computed correlation coefficient. In the present analysis, a significance level of 95\% means the correlation coefficient is marginally significant, while a level of 99\% is considered as highly significant. The correlation coefficient whose level of significance is below 95\% limit is considered as practically insignificant.

2.4. REGRESSION ANALYSIS

The correlation index, \(C(\text{i.e., } C = r^2)\) is the direct measure of the degree (extent) to which the two variables are correlated. It tells only how much of the variation in temperature is correlated with the variation in sunspot number. So the correlation between the two variables show only how the two variables are associated, not that one causes changes in the other, and no causal relationships between them can be estimated unambiguously. Therefore one has to resort to linear regression analysis to obtain a quantitative estimate of the effect of variation in temperature on sunspot number.
Regression line equations of temperature on sunspot number at any atmospheric level L is represented by the expression

$$T'_{n,L} = b_{TR,L}(R_n - \langle R \rangle) + \langle T_L \rangle ... 2.5$$

where $T'_{n,L}$ is the estimated temperature at a level L on a specified year n; $b_{TR,L}$ is the linear regression coefficient (slope of the regression line) of temperature T on sunspot number R at level L; $(R_n - \langle R \rangle)$ is the change in sunspot number from its mean value.

The linear regression coefficient, $b_{TR,L}$ can be computed using the expression,

$$b(T_L, R, n, L) = \frac{\sum_{n=1}^{N} (T_{n,L} - \langle T_L \rangle) (R_n - \langle R \rangle)}{\sum_{n=1}^{N} (R_n - \langle R \rangle)^2} ... 2.6$$

2.4.1 Standard error of estimate (S.E.)

Since the regression line provides the values for estimating the atmospheric temperature ($T'$), the failure of the observed value (T) to fall on the regression line means that the estimates are in error by the degree to which each data point deviates from the regression line. A measure of the accuracy of estimated quantity when using the regression
line is given by the standard error of estimate, as

\[ S.E. = \frac{\sum_{n=1}^{N} (T - T')^2}{N - 1} \]

2.5 COMPUTATION OF LAPSE RATE (\(\Gamma\))

The negative vertical gradient of temperature, generally called the lapse rate of temperature, is an important factor for the estimation of the stability of the atmosphere. In general, the lapse rate is a negative quantity in the stratosphere, because the temperature in this region is increasing with altitude, and is positive in the mesosphere where the temperature is decreasing with height. In the mesosphere, the magnitude of the lapse rate is comparatively smaller than that of the troposphere. In the present investigation, the mean lapse rate at smaller height ranges in the stratosphere and mesosphere were calculated as follows:

Let \(T_{n,L1}\) and \(T_{n,L2}\) be the temperatures (in the Kelvin scale) at heights (in kilometers) \(L1\) and \(L2\), respectively, for the period \(n\). The mean lapse rate, \(\Gamma\), between \(L1\) and \(L2\) are computed by using the expression

\[ \Gamma = -\frac{dT}{dz} = \left[\frac{(T_{n,L1} - T_{n,L2})}{(L1 - L2)}\right] \]

The lapse rate at various layers of altitude in the stratosphere and mesosphere were computed during the solar
minimum and maximum period, and its vertical distribution in these periods is obtained.

2.6 COMPUTATION OF STATIC STABILITY \( (N^2) \)

Adiabatic oscillations of a fluid parcel about its equilibrium level in a stably stratified atmosphere are referred to as buoyancy oscillations. The squared buoyancy frequency, \( N^2 \), represents the measure of the static stability parameter. \( N^2 \) can be written as

\[
N^2 = \frac{g}{\Theta} \frac{d\Theta}{dz} \quad \cdots 2.8
\]

where \( \Theta \) is the potential temperature of the parcel and \( g \) is the acceleration due to gravity. The potential temperature \( \Theta \) can be defined as

\[
\Theta = \frac{T(P_o/P)R}{C_p} \quad \cdots 2.9
\]

By logarithmic differentiation with respect to \( z \), we get

\[
\frac{1}{\Theta} \frac{d\Theta}{dz} = \frac{1}{T} \frac{dT}{dz} - \frac{R}{C_p} \frac{1}{P} \frac{dP}{dz}
\]

\[
= \frac{1}{T} [dT/dz + g/C_p]
\]

But \( dT/dz = -\Gamma \), the environmental lapse rate, and \( g/C_p = \Gamma d \), the dry adiabatic lapse rate

Substituting in the above equation,
\[
\frac{1}{\theta} \frac{d\theta}{dz} = \frac{1}{T} [r_d - r]
\]
Then
\[
N^2 = \frac{g}{T} [r_d - r] \tag{2.9}
\]

In this expression, \(T\) represents the mean temperature (i.e., \((T_{n,L1} + T_{n,L2})/2\)) and \(r\) is the mean lapse rate of a small thickness of layer in the middle atmosphere. The constants \(r_d = 9.8 \text{ kmK}^{-1}\) and \(g = 9.8 \text{ mS}^{-1}\). From these values we can compute the static stability parameter at various altitude layers of the stratosphere and mesosphere.

A knowledge of the static stability variation at different heights in the middle atmosphere is very useful for further theoretical studies in this region, especially for the propagation of planetary scale waves. In earlier studies, the \(N^2\) value was assumed to be constant of about \(4 \times 10^{-4} \text{ S}^{-1}\) in the stratosphere (Holton, 1975; Bates, 1981). The main objective of the present investigation is to examine whether the \(N^2\) value can be taken as constant or not in the region of the stratosphere or mesosphere, in all latitude zones, and to study the difference in the stability of the middle atmosphere during a sunspot cycle maximum and a minimum period.

2.7 AUTO POWER SPECTRUM ANALYSIS

Power spectrum analysis is based on the premise that time series are not necessarily composed of a finite
number of oscillations, each with a discrete wavelength, but rather that they consist of virtually infinite number of small oscillations spanning a continuous distribution of wavelengths. The spectrum, therefore, yields a measure of the distribution of variance in a time series over a continuous domain of all possible wavelengths, ranging from an infinite wavelength to the shortest wavelength that can be resolved by any scheme of harmonic analysis.

Following the approach developed by Blackman and Tukey (1958), the procedures for computing power spectra are described as follows.

In a given series of N equally spaced values, one starts by computing all serial correlation coefficients for lags of 0 to m time units, where m<N. Then the cosine transforms of these m+1 lag correlation values are computed. This yields m+1 'raw estimates' of the power spectrum, the i\textsuperscript{th} value (0 ≤ i ≤ m) of which is a rough measure of the total variance in the original series. The i\textsuperscript{th} value is contributed by wavelength near the i\textsuperscript{th} harmonic of the fundamental wavelength of the analysis. The 'raw estimates' are then smoothed by a 3-term weighted moving average. This averaging operation is necessary in order to derive a consistent estimate of the final spectrum in terms of m+1 discrete estimates.

A mathematical description for the computation of power spectrum is given as follows.
For a series $x_i$ containing N terms, first compute the serial covariances $C_\tau$ for all lags $\tau = 0$ to $\tau = m$ (where $m < N$) according to some standard formula as given in WMO Tech. Note No. 79,

$$C_\tau = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} (x_i - \bar{x}) (x_{i+\tau} - \bar{x}) \quad \ldots \quad 2.10$$

where $\bar{x}$ is the mean of all $x_i$ in the series.

Raw spectral estimates, $S_k$ are then obtained directly from these $C_\tau$ values by the equations

$$\hat{S}_0 = \frac{1}{2m} (C_0 + C_m) + \frac{1}{m} \sum_{\tau=1}^{m-1} C_\tau$$

$$\hat{S}_k = \frac{C_0}{m} + \frac{2}{m} \sum_{\tau=1}^{m-1} C_\tau \cos \left( \pi k \frac{\tau}{m} \right) + \frac{1}{m} C_m (-1)^k \quad \ldots \quad 2.11$$

$$\hat{S}_m = \frac{1}{2m} [C_0 + (-1)^m C_m] + \frac{1}{m} \sum_{\tau=1}^{m-1} (-1)^\tau C_\tau$$

The first of these equations is used to compute the zeroth spectral estimate, which corresponds to virtually infinite wavelengths (trend); and the third equation is used to compute the last spectral estimate, which corresponds to the shortest wavelength resolvable in the spectrum. All the intervening $m-1$ spectral estimates are computed from the middle equation, by setting $k$ in the cosine argument to successive integral values $k = 1, 2, \ldots, m-1$. 
Final spectral estimate $S_k$ are then computed by smoothing the raw estimates with a 3-term weighted average, using the 'Hamming' method formulae, as

$$S_0 = 0.54 S_0 + 0.46 S_1$$

$$S_k = 0.23 S_{k-1} + 0.54 S_k + 0.23 S_{k+1}$$

$$S_m = 0.54 S_{m-1} + 0.46 S_m$$

where in the second equation $k$ is set successively equal to $1, 2, \ldots, m-1$.

2.7.1 Significance test applied to power spectra

The best procedure for evaluating the results of a power spectrum analysis is detailed as follows. Firstly, fitting a null hypothesis continuum to the spectrum is done. If the lag-one serial correlation coefficient $r_1$ of the series does not differ from zero by a statistically significant amount, the series should be regarded as free from persistence. In this case, the appropriate 'null' continuum is that of white noise, or, in other words, a horizontal straight line whose value is everywhere equal to the average of the values of all $m+1$ raw spectral estimates in the computed spectrum.

On the other hand, if the lag-one serial correlation coefficient $r_1$ of the series differs from zero by a statistically
significant amount, one should check if the coefficients for the first one or two lags greater than one approximate to the experimental relation, such as \( r_2 = r_1^2; r_3 = r_1^3; \) etc. If this is the case, the appropriate null continuum should be assumed to be that of Markov 'red noise', whose shape depends on the (unknown) value of the lag-one correlation coefficient for the population \( \theta \) (WMO Tech. Note No. 79). The continuum can be constructed by the following approximate procedure.

Assuming that the sample lag-one coefficient \( r_1 \) is an unbiased estimate of \( \theta \), the following equation for various choices of harmonic number \( k \) between \( k = 0 \) and \( k = m \) can be evaluated.

\[
S_k = \bar{S} \left[ \frac{1 - r_1^2}{1 + r_1^2 - 2 r_1 \cos \frac{\pi k}{m}} \right] \quad \ldots \quad 2.13
\]

In this equation \( \bar{S} \) is the average of all \( m+1 \) raw spectral estimations \( \hat{S}_k \) in the computed spectrum. The resulting values of \( S_k \) may then be plotted. It is superposed on the sample spectrum, and a smooth curve passed through the computed values of \( S_k \) to arrive at the required null continuum.

After choosing the null continuum and superposing on the spectrum, the spectrum is examined for its consistency with the continuum. In other words, the value of each spectral estimate \( S_k \) is compared with the local value of the null continuum. In case none of the \( m+1 \) spectral estimates is found to deviate by a statistically significant amount from the continuum, it
means the continuum does in fact approximate to the true spectrum of the population series. If, on the other hand, one or more of the spectral estimates deviate significantly from the continuum, then the continuum chosen for comparison is not a satisfactory approximation to the true spectrum of the population series.

The ratio of the magnitude of the spectral estimate to the local magnitude of the continuum is found to be distributed as \( \chi^2 \) (chi-square) divided by \( \gamma \) number of degrees of freedom (Tukey, 1950).

The number of degrees of freedom \( \gamma \) of each estimate of a spectrum that is based on a record length of \( N \) values and on a maximum lag of \( m \) time units is given closely by

\[
\gamma = \frac{2N - m/2}{m}
\]

Therefore, if it is assumed that local value of the continuum is the true magnitude of the population spectrum at that particular wavelength, it is implied that this value corresponds to the 50% of the \( \chi^2/\gamma \) distribution appropriate to size \( N \) of the sample and maximum lag \( m \).

The ratio of any sample spectral estimate \( S_k \) to its local value of the continuum is then compared with the critical percentage levels of a \( \chi^2/\gamma \) distribution for the proper value of \( \gamma \). This comparison establishes the level of statistical significance required. In practice, critical levels of the \( \chi^2/\gamma \)
distribution for arbitrary $\gamma$ can be obtained from tables, available in standard statistical books.

In a given sample spectrum, the 95 per cent point of the $\chi^2/\gamma$ distribution is the same for all spectral estimates $S_k$. In other words, the 95 per cent confidence limit for the null continuum is given by a second continuum whose value for any wavelength in the spectrum is equal to a certain fixed multiple of the value of the null continuum at that wavelength. Other confidence limits can be added above and below the null continuum if desired, which have values equal to other constant ratios of the null continuum at all wavelengths in the spectrum.

2.8 CROSS SPECTRUM ANALYSIS

When only a single series is observed, as in the case of autospectrum, then only the internal structure of the series can be found. Estimated spectra, in this case, give information only about the oscillations in individual series. On the other hand, when two series are observed, the internal structure of each as well as their joint structure, or the dependence of either series on the other may be of interest. In other words, the similarities between the spectra of the two series, such as peaks at similar frequencies, may raise the possibility that the series are related. To investigate such possibilities it is useful to compute estimates of the cross spectrum of the two series. This is an extension of the definition of the spectrum.
and is usually estimated by smoothing the cross periodogram.

2.8.1 Mathematical description for the computation of cross spectrum

Let \( x_i \) and \( y_i \) be the two series, each having \( N \) equally spaced values. The cross co-variances of \( x \) and \( y \) (i.e. \( C(x,y)_\tau \)) and also \( y \) and \( x \) (i.e., \( C(y,x)_\tau \)), for all lags \( \tau = 0 \) to \( \tau = m \) time units (where \( m < N \)) are computed using the standard formula given below:

\[
C(x,y) = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} (x_i - \bar{x})(y_{i+\tau} - \bar{y}) 
\]

\[
C(y,x) = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} (x_{i+\tau} - \bar{x})(y_i - \bar{y}) 
\]

These cross-covariances can be separated into two co-variant components.

\[
R(x,y)_\tau = \frac{[C(x,y)_\tau + C(y,x)_\tau]}{2} 
\]

\[
Q(x,y)_\tau = \frac{[C(x,y)_\tau - C(y,x)_\tau]}{2} 
\]

Now the raw cross spectral densities of \( R(x,y)_\tau \) and \( Q(x,y)_\tau \) can be obtained by calculating the corresponding \( S_k^\wedge \) values, as mentioned in equation 2.11. The results of \( R(x,y)_\tau \) and \( Q(x,y)_\tau \) represent the real and imaginary parts of the cross-spectrum, usually called the 'cospectrum' and 'quadrature spectrum', respectively. The cospectrum measures
the extent to which there are oscillations with the same phase in the two series (or with opposite sign, that is, with a phase shift of a half cycle), while the quadrature spectrum measures the extent to which there are oscillations with a phase difference of a quarter cycle (in either direction).

The raw cross-spectra corresponding to $R(x,y)_\gamma$ and $Q(x,y)_\gamma$ are then smoothed by using the Hamming method (see equation 2.12).

2.8.2 Estimation of coherence factor ($\gamma^2$)

If the spectral weights are non-negative, then the square of the cross-spectral density for a given frequency is less than or equal to the product of the individual spectrum of the series for the corresponding frequency. Thus, it is convenient to consider the ratio.

$\gamma^2 = \frac{R(x,y)^2 + Q(x,y)^2}{C_x C_y}$ ...

2.17

which is called the 'squared coherency' or 'coherence' of $x$ and $y$. Coherence being analogous to the correlation coefficient necessarily lies between 0 and 1. The value of 0 and 1 corresponds to these being no dependence or complete dependence, respectively, for given frequency.

The detailed computer programme in FORTRAN IV for the estimation of the auto and cross power spectral densities and the coherence factor is given in Appendix-B for running in the computer system CYBER 170.