A Brief Overview of Cluster Analysis

Alain Sirois

Atmospheric Environment Service
Environment Canada
2121, Voie de Service Nord, Suite 500
Route Trans-canadienne,
Dorval, Quebec, Canada, H9P 1J3
Notice

This report is the background document for the presentation of the same name given at the WMO-EMEP Workshop on Advanced Statistical Methods and their Application to Air Quality Data Sets, Finnish Meteorological Institute, Helsinki, Finland, 14-18 September, 1998. The author can be reached at the following address: Dr. Alain Sirois (ARQI), Service de l'Environnement Atmosphérique, 2121, Voie de Service Nord, Suite 500, Route Trans-canadienne, Dorval, Quebec Canada, H9P 1J3. Tel: (514) 421-4674; Fax: (514) 421-2106; E-mail: Alain.Sirois@ec.gc.ca
Preface

Along, with principal component analysis, cluster analysis has been one of the most frequently used statistical analysis technique in the field of atmospheric chemistry. It has been employed mainly as a method to group atmospheric chemistry measurements by air mass origin using associated air mass back-trajectories. References to some of these analyses are given in the overview.

Although the subject of cluster analysis is an important and vast one, the author has had to limit himself to only touch lightly on the topic because of lack of space and time. However, he has tried to give a comprehensive, although brief, overview of the basic concepts of cluster analysis.

After defining the aim of cluster analysis in the introduction, the concept of distance or similarity measurements are discussed. These is followed by a brief description of available basic clustering techniques. A brief mention is then made of graphical techniques that can be used if the number of available measurements to classify is not too large. Finally, an example of the use of cluster analysis in the context of atmospheric chemistry is presented at the end of the overview. Many references to more comprehensive books on cluster analysis are given throughout the overview to compensate for its brevity.

As mentioned earlier, time for the preparation of this overview was limited. Although the author tried as much as possible to make sure that the equations, examples, and descriptions are correct, errors could easily have crept in, especially in unifying the notation used by the different authors consulted. The author would therefore recommend, in particular, that before using the equations presented herein, the reader verify them in the references cited.
## Contents

Preface ........................................ iii

Contents ....................................... v

1. Introduction ................................ 1

2. Definition of Distance or Similarity Measurements........... 2

3. Clustering Techniques ....................... 3

3.1 Hierarchical Techniques ................. 3

3.1.1 Agglomerative Methods ............... 3

3.1.1.1 Single Linkage or the Nearest-Neighbor Method .... 4

3.1.1.2 Complete Linkage or the Furthest-Neighbor Method 4

3.1.1.3 Centroid Method .......................... 5

3.1.1.4 Average Linkage ......................... 5

3.1.1.5 Ward’s Error Sum of Squares Method ......... 5

3.1.1.6 Lance and Williams Flexible Method ......... 6

3.1.2 Divisive Methods .......................... 6

3.2 Partitioning Methods ...................... 7

3.2.1 K-Means Clustering ..................... 7

3.2.2 Methods Based on Trace .................. 8

4. Graphical Methods ........................... 8

4.1 Glyphs and Metroglyphs .................... 9

4.2 Fourier Series ................................ 10

4.3 Chernoff Faces ................................ 11

5. Application to Atmospheric Chemistry: An example ....... 11

6. Final Remarks ................................ 16

7. References ................................... 17
1. **Introduction**

The starting point of cluster analysis is the same as in principal component analysis (PCA) or other reduction techniques; i.e., one has \( n \) observations of \( p \) variables that can be put into matrix form as:

\[
X_{(n \times p)} = \begin{pmatrix}
  x_{11} & x_{12} & \cdots & x_{1p} \\
  x_{21} & x_{22} & \cdots & x_{2p} \\
  \vdots   & \vdots   & \ddots & \vdots   \\
  x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix}
\]  

Cluster analysis differs from principal component analysis and other similar techniques, however, in that its aim is to group the observations (referred to as objects) into clusters, rather than to find links between the \( p \) variables. Cluster analysis can be summarized by the following four steps:

1. Collect \( n \) observations of \( p \) variables.
2. Create a similarity or distance matrix.
3. Form the clusters, which can either a) be mutually exclusive or b) form a hierarchy.
4. Look at the cluster profile.

The main goal of cluster analysis is to arrive at clusters in which the variation within clusters is smaller than that between clusters. This is illustrated in Figure 1.1 for a hypothetical two-dimensional case. In this example, it is clear that the variation between the objects in each cluster is smaller than the variation between objects in different clusters.

When doing a cluster analysis, one has to make two key decisions that will dominate the analysis. They are:

1. decide on a measure of inter-object distance or similarity.
2. define a procedure to define the clusters.

Many definitions of distance or similarity have been used in the past and many cluster algorithms have been developed. In the present overview, we will only mention the most important ones.

There is no generally accepted definition of a cluster. The meanings of such terms as cluster or distance (similarity) depends largely on individual judgments.

If we imagine each object or observation as a point in a \( p \)-dimensional space (\( p \) being the number of variables measured), we can describe clusters as continuous regions with a high density of points separated by regions having a low density of points (Figure 1.1). This defines what is known as *natural clusters* and does not impose any *a priori* restrictions on the structure of the data and clusters. As such, the data guide the analysis.

---

*Figure 1.1 Examples of three clusters*
Another definition is based on the *criterion of closeness*. This type of definition invokes the premise that objects or observations in a cluster should be closer to each other than to objects in other clusters. Note, however, that it is difficult to identify clusters that are not spherical when using this criterion.

In the following sections, we will first discuss the definition of distances and then present different techniques that have been proposed for estimating clusters.

2. **Definition of Distance or Similarity Measurements**

Since the variables measured in atmospheric chemistry are metric, we will limit our discussion of distance to these types of data. For more information on the definition of similarity for other types of data, the reader is referred to Dillon and Goldstein (1984).

The most common definition of distance between two objects or observations is the *Euclidean distance* between two objects i and j, defined as:

\[
d_{ij} = \left\{ \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right\}^{1/2}
\]  

(2.1)

Another interesting distance is the *absolute or city-block distance*, defined as:

\[
d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|
\]  

(2.2)

These two types of distance are special cases of the Minkowski metric, defined as:

\[
d_{ij} = \left\{ \sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \right\}^{1/r}
\]  

(2.3)

The reader should notice that the distances defined in 2.1 to 2.3 are not scale invariant. Therefore, if the unit of measurement of a variable is changed, the distance between the observations is also changed.

Using relations 2.1 to 2.3 has the disadvantage that the distance is dominated by the variable(s) with the largest range of variation (i.e., the greatest spread). For example, if most of the variables vary between 0 and 10 but one varies between 0 and 10000, the distance would be dominated by the latter. In these cases, it is recommended that the variables be standardized (by dividing each variable by its standard deviation) before the distances are calculated.

Another way of scaling the variables so that no one variable dominates the distance calculation is to use the *Mahalanbois distance*, given by:

\[
(X_i - X_j) S^{-1} (X_i - X_j)
\]  

(2.4)

where \( X_i \) and \( X_j \) are the respective vectors of observations i and j and \( S \) is the pooled within-group variance-covariance matrix. This distance measure has the advantage of explicitly account for any correlations that might exist between the variables. Note that if \( S = I \) (i.e., the identity matrix), the Mahalanbois distance reduces to the Euclidean distance. Figure 2.1 presents the geometry of different definitions of distances. In the case of the Euclidean distance, all the observations lying on a circle around observation C are at the same distance.
from it. If the data have been standardized, the circles become ellipses with the axes proportional in length to the variance of the variable. The Mahalanbois distance rotates the ellipses by an angle proportional to the correlation between the two variables.

Other definitions of distances can be created in special cases. An example of a special definition of distance is the mean angle distance defined by Siros and Bottenheim (1995) in the case of air mass back-trajectories. The reader is referred to their paper for more details.

3. Clustering Techniques

The different types of clustering techniques can be grouped into two categories, namely, hierarchical techniques and partitioning techniques. We will discuss each of these in the rest of this section.

3.1 Hierarchical Techniques

In hierarchical techniques, one performs a succession of fusions or divisions of the clusters. An important characteristic of hierarchical techniques is that the allocation of an observation to a cluster is irrevocable; that is, once an object is allocated, it cannot be reallocated.

Two types of hierarchical clustering exist, namely, agglomerative methods and divisive methods. In the first, one starts with each object constituting a cluster and then groups them until every object is in one cluster. In divisive methods, one starts with one cluster and divides clusters until all the resulting clusters contain only one object. These techniques are illustrated in Figure 3.1.

3.1.1 Agglomerative Methods

In agglomerative methods, one starts with n clusters; in other words, each observation is a cluster. The first step is to merge the two closest clusters to form (n-1) clusters, then one merges the next two closest clusters to form (n-2) clusters, and so on. This process is continued until all the objects or observations are in the same cluster. This is illustrated in Figure 3.1a. The fusion of two clusters has to be made n-1 times.
Various methods differ, however, in the way they measure the distance between two clusters. (Note that if the two clusters contain one object each, all the distances are the same.)

Examples of the use of agglomerative methods in atmospheric chemistry can be found in Slanina et al. (1983), Moody and Galloway (1988), and Moody and Samson (1989).

3.1.1.1 Single Linkage or the Nearest-Neighbor Method

In this method, the distance between clusters is defined as the distance between their nearest members. This is illustrated in Figure 3.2 and can be written mathematically by:

\[ D_{ij} = \min_{n \in \{1, \ldots, n_i\}, m \in \{1, \ldots, n_j\}} \left( d_{nm} \right) \]

(3.1)

where \( D_{ij} \) is the distance between clusters \( i \) and \( j \); \( n_i \) and \( n_j \) are the number of observations in clusters \( i \) and \( j \), respectively; and \( d_{nm} \) is the distance between the objects \( n \) and \( m \). Note that \( n \) and \( m \) index the observations in clusters \( i \) and \( j \) respectively.

This method is subject to chaining (chaining forms linear groups of objects). An example of chaining is shown in Figure 3.3.

3.1.1.2 Complete Linkage or the Furthest-Neighbor Method

The definition of distance in this method is the opposite of the one given in the preceding section. Here, the distance between two clusters is defined as the distance between the most distant pair of observations. This is illustrated in Figure 3.4 and can be written mathematically as:

\[ D_{ij} = \max_{n \in \{1, \ldots, n_i\}, m \in \{1, \ldots, n_j\}} \left( d_{nm} \right) \]

(3.2)

This method tends to overcome the problem of chaining.
3.1.1.3 Centroid Method

In this method, the distance between two clusters is defined as the distance between each group's centroid. This is illustrated in Figure 3.5. The distance between clusters \( i \) and \( j \) can be written as:

\[
D_{ij} = \sum_{k=1}^{p} (\bar{x}_k^i - \bar{x}_k^j)^2
\]

(3.3)

where \( \bar{x}_k^i \) is the \( k \)th coordinate of the centroid for cluster \( i \). It is equal to:

\[
\bar{x}_k^i = \frac{1}{n_i} \sum_{n=1}^{n} x_{nk}
\]

(3.4)

The summation is performed on the \( n_i \) objects in cluster \( i \).

3.1.1.4 Average Linkage

In this method, the distance between two clusters is defined as the average distance between all pairs of objects, one from one cluster and one from the other. This can be written as:

\[
D_{ij} = \frac{1}{n_in_j} \sum_{n=1}^{n} \sum_{m=1}^{n} d_{ij}
\]

(3.5)

This method tries to minimize within-group variance and maximize between-group variance.

3.1.1.5 Ward's Error Sum of Squares Method

This method was developed by Ward (1963). The technique estimates the loss of information resulting from grouping objects into clusters by summing the squared deviations of every observation from the mean of the cluster to which they belong. The assignment rule rests on the increase in the error sum of squares induced from combining every possible pair of clusters. At each step in the analysis, the union of every possible pair of clusters is tested and the two clusters whose fusion results in the minimum increase in the error sum of squares are combined.

If we denote by \( x_{ijk} \) the observation of the \( i \)th of the \( p \) variables for the \( j \)th of the \( n_k \) objects in the \( k \)th of the \( h \) clusters, the mean for variable \( i \) and cluster \( k \) is:

\[
\bar{x}_{ik} = \frac{1}{n_k} \sum_{j=1}^{n_k} x_{ijk}
\]

(3.6)

The error sum of squares for cluster \( k \) is given by:

\[
E_k = \sum_{i=1}^{p} \sum_{j=1}^{n_k} (x_{ijk} - \bar{x}_{ik})^2 = \sum_{i=1}^{p} \sum_{j=1}^{n_k} x_{ijk}^2 - n_k \sum_{i=1}^{p} \bar{x}_{ik}^2
\]

(3.7)

and the error sum of squares is:
Table 3.1 Coefficients for the Lance and Williams recurrence formula

<table>
<thead>
<tr>
<th>Method</th>
<th>$\alpha_i$</th>
<th>$\alpha_j$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Linkage</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>-1/2</td>
</tr>
<tr>
<td>Complete Linkage</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>1/2</td>
</tr>
<tr>
<td>Centroid</td>
<td>$\frac{n_i}{n_i + n_j}$</td>
<td>$\frac{n_j}{n_i + n_j}$</td>
<td>$-\alpha_i \alpha_j$</td>
<td>0</td>
</tr>
<tr>
<td>Average Linkage</td>
<td>$\frac{n_i}{n_i + n_j}$</td>
<td>$\frac{n_j}{n_i + n_j}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ward's method</td>
<td>$\frac{n_k + n_i}{n_i + n_j}$</td>
<td>$\frac{n_k + n_j}{n_i + n_j}$</td>
<td>$-n_k$</td>
<td>0</td>
</tr>
</tbody>
</table>

$$ESS = \sum_{k=1}^{h} E_k$$ (3.8)

Ward's method tends to join clusters with a small number of observations and is strongly biased toward producing clusters with roughly the same number of observations. The method is very sensitive to outliers.

3.1.1.6 **Lance and Williams Flexible Method**

Lance and Williams (1967) developed a flexible algorithm that can be used for all the techniques presented in the preceding sections. The distance measures between groups used by these methods satisfy a recurrence formula for the distance between a cluster k and a cluster (ij) formed by the fusion of clusters i and j. The recurrence formula of Lance and Williams is:

$$D_{k(ij)} = \alpha_i D_{ki} + \alpha_j D_{kj} + \beta D_{ij} + \gamma |D_{ki} - D_{kj}|$$ (3.9)

where the coefficients $\alpha$, $\beta$ and $\gamma$ are given in Table 3.1 for the different methods.

The Lance and Williams algorithm is very useful because the same program can be used for all the different agglomerative techniques presented in this section. Only the coefficients in relation 3.9 have to be changed.

3.1.2 **Divisive Methods**

In divisive methods, one starts with all the observations in one cluster. The first step in the method is to divide the cluster into two groups or clusters. If there are n observations, there are $2^{n-1} - 1$ different ways to form these two subsets. Even if n is not particularly large, the most powerful computer would have trouble in testing all these combinations. Thus, any divisive method must give an efficient algorithm for dividing the first cluster into two groups.

Once the initial split is made, objects are moved from one cluster to another, or finer subdivisions of the already formed clusters are made.
The various divisive methods differ according to how the initial split is effected and how already formed clusters are subdivided.

One of the simpler methods is the Splinter-Average Distance Method of MacNaughton-Smith et al. (1962). With this technique, one starts by splintering out that object having the largest distance from the other objects. That operation leaves two groups containing the splintered object and the other containing the remaining objects. Next, one computes two distances: (1) the average distance of each object in the main group to the object(s) in the splinter group and (2) the average distance of each object in the main group to the other observations in the group. If an object's average distance to the object(s) in the splinter group is less than its average distance to the other objects in the main cluster, it is moved to the splinter group. This process is continued until the composition of the two clusters has stabilized, which is the case when each object's average distance to objects in its own cluster is smaller than its average distance to objects in the other cluster. The process is then repeated, with each of the two clusters being splintered again. The process is continued until all clusters contained only one object.

Another example of the divisive method is Automatic Interaction Detection (AID). Sonquist and Morgan (1964) give details of this technique.

To the knowledge of the author, no divisive methods have been used in atmospheric chemistry.

3.2 Partitioning Methods

In partitioning methods, the allocation of an observation to a cluster is not irrevocable. Objects may be reallocated if the initial assignments were inaccurate.

Partitioning techniques partition the data by optimizing some formal and predefined criterion. The use of partitioning techniques assumes that the number of final clusters is known and specified in advance, although some methods allow the number to vary during the course of the analysis.

Partitioning techniques differ with respect to (1) how clusters are initialized, (2) how objects are allocated to clusters, and (3) how some or all of the objects already clustered are reallocated to other clusters.

An example of the use of partitioning cluster analysis can be found in Dorling et al. (1992a, 1992b).

3.2.1 K-Means Clustering

Suppose that one has n observations of p variables, denoted as \(X_{ij}\) (\(i = 1, 2, \ldots, n\); \(j = 1, 2, \ldots, p\)). We will denote by \(\bar{X}^k_j\) the mean value for variable \(j\) in cluster \(k\) (\(k = 1, 2, \ldots, h\)). The process starts by generating an initial mean vector, called a seed, for each of the h clusters. Different techniques have been proposed to generate them. The easiest uses the first \(h\) observations as the seeds. A second technique is to select \(h\) observations randomly and use them as seeds. A third technique is to use the \(h\) observations that are mutually furthest apart. A final technique is to employ prior knowledge to generate the \(h\) seeds.

Once the \(h\) seeds have been selected, each observation is allocated to the closest cluster. New mean values for each cluster can be calculated using the observations in it. The
observations are then re-allocated to the closest cluster, and new mean vectors for each cluster are obtained. This process is continued until no observation can be transferred from one cluster to another.

This technique is very simple but it can take a long time to achieve convergence if the number of observations is large. In some cases there may be problems in achieving convergence. These usually occur when a few observations are alternatively shifted between two clusters because they are at the periphery of the two clusters.

3.2.2 Methods Based on Trace

Some partitioning algorithms are based upon minimizing (or maximizing) within-group (between-group) dispersion. If \( T \) denotes the total dispersion matrix, \( W \) the within-clusters dispersion matrix, and \( B \) the between-clusters dispersion matrix, then:

\[
T = W + B
\]  

(3.10)

where, assuming \( h \) clusters, \( W = \sum_{k=1}^{h} W_k \). For any given set of data, \( T \) is fixed, so the clustering criteria can be defined in terms of \( W \) or \( B \). Several possibilities have been suggested, including:

1. **Trace of \( W \).** This criterion attempts to minimize the trace of the pooled within-cluster matrix of sums-of-squares and cross-products. Note that minimizing \( \text{tr}(W) \) is equivalent to maximizing \( \text{tr}(B) \), since \( \text{tr}(T) = \text{tr}(W) + \text{tr}(B) \).

2. **Determinant of \( W \).** Minimizing \( |W| \) is equivalent to maximizing \( |T|/|W| \). Friedman and Rubin (1967) suggest that this latter criterion be modified to \( \log(\max(|T|/|W|)) \).

3. **Trace of \( BW^{-1} \).** This criterion may also be expressed in terms of the eigenvalues, \( \lambda_1, \lambda_2, \ldots, \lambda_p \), of \( BW^{-1} \), since \( \text{tr}(BW^{-1}) = \sum \lambda_i \).

All of these techniques attempt to minimize or maximize some criterion. However, not every partition of the data is evaluated, and consequently less-than-optimal solutions may result. Most algorithms are designed to search for local optima of the criterion by employing hill-climbing or hill-descending methods, depending on whether the criterion is maximized or minimized. In either case, the existing partitions are rearranged, and only those new arrangements that yield an improvement in the criterion are kept. The problem of local optima can be avoided with the use of a dynamic programming algorithm developed by Jensen (1969), which guarantees a global optimum.

Note that the use of certain of the proposed criteria restricts the shape of the clusters formed. For example, use of the \( \text{tr}(W) \) criterion will mean that the uncovered clusters will be only spherical, even if natural clusters of other shapes are more appropriate. The \( |W| \) does not restrict the clusters to being spherical, but it does assume that all clusters have the same shape. For this reason, Scott and Symons (1971) suggest that \( \prod_{i=1}^{h} |W_i|^{n_i} \), where \( n_i \) is the number of objects in cluster \( i \), be used.

4. Graphical Methods
Table 4.1 Percentage of people agreeing with eleven statements about eight brands of cereals (from Chakrapani and Ehrenberg, 1981)

<table>
<thead>
<tr>
<th>Statement</th>
<th>Corn Flakes</th>
<th>Weet Abix</th>
<th>Rice Krispies</th>
<th>Shredded Wheat</th>
<th>Sugar Puffs</th>
<th>Special K</th>
<th>Frosties</th>
<th>All Bran</th>
</tr>
</thead>
<tbody>
<tr>
<td>come back to taste nice</td>
<td>65</td>
<td>64</td>
<td>59</td>
<td>60</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>popular with all the family</td>
<td>64</td>
<td>40</td>
<td>32</td>
<td>24</td>
<td>20</td>
<td>19</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>very easy to digest nourishing</td>
<td>40</td>
<td>50</td>
<td>17</td>
<td>33</td>
<td>18</td>
<td>19</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>natural flavor reasonably priced</td>
<td>47</td>
<td>39</td>
<td>11</td>
<td>12</td>
<td>6</td>
<td>15</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>a lot of food value stays crispy in milk helps to keep you fit fun for children to eat</td>
<td>27</td>
<td>38</td>
<td>9</td>
<td>26</td>
<td>9</td>
<td>17</td>
<td>7</td>
<td>17</td>
</tr>
</tbody>
</table>

Another set of techniques to classify objects or observations uses the graphical methods. Many graphical techniques have been proposed in the past. The most frequently used are glyphs and metroglyphs, Fourier series, and Chernoff faces.

To illustrate these techniques, we will use the data given in Table 4.1. These data are the percentage of people agreeing with eleven statements about eight brands of cereals in the United States.

Note that the usefulness of these techniques is greatly reduced when the number of observations becomes large.

4.1 Glyphs and Metroglyphs

This graphical method was developed by E. Anderson (1954, 1957, 1960). A glyph is a point (or circle of fixed radius) with rays emanating from it. The rays correspond to the characteristics or variables measured: the position of a ray labels the characteristic, while its length reflects its values. The angles between adjacent rays are equal, and the number of rays corresponds to the number of variables. This is illustrated in Figure 4.1. Usually, the star is completed by connecting the extremities of the rays to form a polygon.

The glyphs for the data in Table 4.1 are shown in Figure 4.2. One can see the connection between the Corn Flakes and the Weet Abix. The results for Frosties and Sugar Puffs are also quite similar. Both of these cereals are very popular with children.

![Figure 4.1 Illustration of glyphs.](image-url)
4.2 Fourier Series

This graphical technique was suggested by Andrews (1972). He proposed to transform a $p$-dimensional response vector (i.e., $p$ observed variables) by Fourier series as:

$$f_X(t) = \frac{1}{\sqrt{2}} + X_1 \sin t + X_2 \cos t + X_3 \sin 2t + X_4 \cos 2t + \cdots$$

(4.1)

over the range $-\pi \leq t \leq \pi$.

Figure 4.3 presents the Fourier graph for the cereal attitude data of Table 4.1. On this plots, one can detect three groups or clusters. The first is formed by the Corn Flakes and the Weet Abix (continuous lines). A second cluster (dashed lines) contains Rice Krispies, Sugar Puffs and Frosties, the cereals that are popular with children. The last cluster (short-long dashed lines) contains the Shredded Wheat, the Special K, and the All Bran.

Andrews' representation has a number of useful properties that make it well suited for exploratory analysis. In particular:

1. The function $f_X(.)$ preserves means in the sense that if $\overline{X}$ is the mean of a set of $n$ multivariate observations $X_i$, then $f_X(t) = \frac{1}{n} \sum f_{X_i}(t)$.
2. The function $f_\chi(.)$ preserves distances. The implication of this property is that close points in a Euclidean sense will appear as close functions, and distant points as distant functions. This may be very useful for detecting outliers.

3. For a given $t_0$, $f_\chi(t_0)$ is proportional to the length of the projection of the vector $(X_1, X_2, \ldots, X_p)$ on the vector $f_\chi(t_0) = (1/\sqrt{2}, \sin t_0, \cos t_0, \sin 2t_0, \cos 2t_0, \ldots)$.

4. The function $f_\chi(.)$ preserves variances.

4.3 *Chernoff Faces*

This graphical method was introduced by Chernoff (1973). In its original form, up to 18 dimensions in a response vector were allowed. Each dimension becomes associated with one of 18 facial features.

The Chernoff faces for the cereal attitude data of Table 4.1 are shown in Figure 4.4. The association between the statements in Table 4.1 and the face characteristics are given in Table 4.2.

The connection between the *Corn Flakes* and the *Weet Abix* is easy to see. One can also see how similar the results are for the *Rice Krispies*, the *Sugar Puffs* and the *Frosties*. Additional similarity is apparent in the results for *Shredded Wheat*, the *Special K*, and the *All Bran*.

5. **Application to Atmospheric Chemistry: An Example**

To illustrate the use of cluster analysis in atmospheric chemistry, we will present the results of different clustering techniques applied to one year (1980) of 925 mb three-day back-trajectories from Chalk River, Ontario. Note that there is four back-trajectories for each day. Figure 5.1 presents all the three-day back trajectories for the year. It is

<table>
<thead>
<tr>
<th>Statement</th>
<th>Face Characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>come back to taste nice</td>
<td>Area of face</td>
</tr>
<tr>
<td>popular with all the family</td>
<td>shape of face</td>
</tr>
<tr>
<td>very easy to digest</td>
<td>length of nose</td>
</tr>
<tr>
<td>nourishing</td>
<td>location of mouth</td>
</tr>
<tr>
<td>natural flavor</td>
<td>curve of smile</td>
</tr>
<tr>
<td>reasonably priced</td>
<td>width of mouth</td>
</tr>
<tr>
<td>a lot of food value</td>
<td>location of eyes</td>
</tr>
<tr>
<td>stays crispy in milk</td>
<td>separation of eyes</td>
</tr>
<tr>
<td>helps to keep you fit</td>
<td>angle of eyes</td>
</tr>
<tr>
<td>fun for children to eat</td>
<td>shape of eyes</td>
</tr>
</tbody>
</table>

|                |                |
|                |                |
| Corn Flakes    | Shredded Wheat |
| Frosties       |                |

|                               |                |
|                               |                |
| Weet Abix       | Sugar Puffs    |
|                 | All Bran       |

|                               |                |
|                               |                |
| Rice Krispies | Special K      |

*Figure 4.4* Four Chernoff faces for the cereal attitude data in Table 4.1.
easily seen that the back-trajectories cover much of the area west of the site and only a smaller area east of the site.

The first cluster analysis was made using an agglomerative hierarchical method (Ward's error sum of squares technique). The merging of clusters was stopped when the back-trajectories were grouped into six clusters. The Euclidean metric was used to measure distances between the back-trajectories. A stereographic projection true at 60° North was used to project the back-trajectories before the analysis was made.

Figure 5.2 presents the positions of the mean back-trajectories for each cluster. The mean positions were drawn for every 6 hours for 3 days. The ellipse around each mean value for time \( t \) would contain 68% of the corresponding (in time) positions of the back-trajectories within the cluster, if these positions followed a bi-normal distribution.

**Figure 5.1** 925mb three-day back-trajectories for Chalk River, Ontario, in 1980.

**Figure 5.2** Mean back-trajectories for the six clusters obtained using the Ward's technique and Euclidean distances.
Figure 5.3 Three-day back-trajectories from Chalk River, Ontario, for the six clusters obtained using the Ward's technique and Euclidean distances.

From these results, one sees that the clusters differ by either direction or speed of displacement of the air. For example, the air in clusters 2 and 4 came from about the same direction, but the air came from a greater distance in cluster 2 than in cluster 4. The same is

Figure 5.4 Mean back-trajectories for the six clusters obtained using the K-Means technique and Euclidean distances.
true for clusters 1 and 6. The three-day back-trajectories for the six clusters are shown in Figure 5.3.

The next step in the analysis was to redo it using the K-mean method with

Figure 5.6 Mean back-trajectories for the six clusters obtained using the K-Means technique and angle distances.
Euclidean distances. As in the preceding case, the back-trajectories were projected before the analysis.

The results of the clustering are presented in Figures 5.4 and 5.5. Some of the clusters are similar to the ones obtained using the hierarchical technique. For example, Clusters 2 and 6 of Figure 5.5 are similar to clusters 1 and 4 of Figure 5.3 respectively. Other clusters are combinations of portions of clusters in Figures 5.2 and 5.4. For example, cluster 3 in Figures 5.4 and 5.5 is a mixture of part of cluster 2 and all of cluster 5 in Figures 5.2 and 5.3.

The preceding example illustrates the fact that two different methods produce different results even if the same distance metric is used. In the same way, the use of a different distance metric produces different results even if the same technique is used. This is illustrated in Figures 5.6 and 5.7. These figures present the result of clustering using the K-means method with the angle distance metric defined in Sirois and Bottenheim (1995). Using the angle between the back-trajectories as a distance metric eliminates the effect of the velocity of the air. This effect, as we have seen in Figures 5.2 and 5.3, results in the generation of clusters having the same direction but different air velocities. The angle metric can correct for this effect.

In Figures 5.6 and 5.7, the clusters point in different directions (in the mean). Some of the clusters are similar to the ones obtained by the other techniques. Some are somewhat different.

One important characteristic of all the results is that the back-trajectories would overlap at the edges of the clusters if the clusters were plotted together. This is a characteristic of the back-trajectory clusters.
6. Final Remarks

Lack of time and space have made it impossible to mention many subjects related to
cluster analysis in this overview. The most important of these omitted subjects deals with
methods for selecting the number of clusters that one should retain. However, the reader can
gain fuller understanding of cluster analysis by consulting Anderberg (1973), Hartigan
(1975), Späth (1980) and Everitt (1980). The first three of these books include Fortran
programs to implement most of the techniques presented here. In addition, a short
introduction to cluster analysis can be found in Dillon and Goldstein (1984).

Most statistical software packages include functions or programs for the majority of
the techniques presented in this overview.

Finally, in cluster analysis, as in most other statistical fields, there is no technique or
method that can be used blindly for all data sets. In the present instance, it is the usefulness of
the clusters that determines whether a technique was appropriate. As in all statistical analyses,
the user of cluster analysis must apply his or her judgment and not accept results without due
consideration.
7. References


A Brief and Biased Overview of Time-Series Analysis or How to Find that Evasive Trend

Alain Sirois

Atmospheric Environment Service
Environment Canada
2121, Voie de Service Nord, Suite 500
Route Trans-canadienne,
Dorval, Quebec, Canada, H9P 1J3
Notice

This report is the background document for the presentation of the same name given at the WMO-EMEP Workshop on Advanced Statistical Methods and their Application to Air Quality Data Sets, Finnish Meteorological Institute, Helsinki, Finland, 14-18 September, 1998. The author could be reached at the following address: Dr. Alain Sirois (ARQI), Service de l'Environnement Atmosphérique, 2121, Voie de Service Nord, Suite 500, Route Trans-canadienne, Dorval, Quebec Canada, H9P 1J3. Tel: (514) 421-4674; Fax: (514) 421-2106; E-mail: Alain.Sirois@ec.gc.ca
Preface

The bias in the title of this presentation is a bias in favour of utility. In that spirit, the author has selected the techniques to be discussed here on the basis of two criteria:

- the technique has been used before by somebody;
- the technique might possibly be useful in the future.

In making selections, the author’s agreement with any particular technique was not a consideration as long as that technique was mathematically correct.

Another factor that may have affected the selection process is the simple fact that the author does not know everything. Therefore, some techniques could have been omitted, not because they are incorrect or uninteresting but because they were unknown to the author. Unfortunately, the time when someone could everything know that could possibly be known about a subject is now entirely gone.

A final factor has been the need for brevity in both the presentation and the written paper. As a result, the author has been able to consider only a limited number of techniques.

The author has tried in this overview not only to give a catalogue of both past and present techniques for time series analysis but also to present an overall framework for the subject. Time series analysis involves more than just applying statistical techniques blankly. It should be extended to a complete analysis of the data. Therefore, the author will also try in this overview to indicate how such an analysis, in his opinion, should be done. The views expressed in this overview reflect only the personal and biased opinions of the author, and he is the first to admit any shortcomings in his approach. However, the author feels that his methods are at least comprehensive and provide a useful guide to time series analysis.

The overview consists first of a description of techniques that may help in exploring the main characteristics of a given time series. Then, different techniques to test for the existence of or to extract the long-term trends and/or other temporal variations present in the time series are discussed. The last, technical, part of the overview examines techniques that permit the verification of the assumptions made when applying any of the tests or other techniques. This last point is a very important aspect of time series analysis.

An example illustrating the use of some of the techniques presented here is given at the end. Note, however, that many other examples of particular techniques or types of plots will be given in the sections dealing with those topics. The purpose of the final example is to provide an integrated illustration of the author’s approach to time series analysis.

Time for the preparation of this overview was limited. Although the author has tried as much as possible to make sure that the equations, examples, and descriptions are correct, errors could easily have crept in, especially in unifying the notations used by the different authors consulted. The author would therefore recommend that, before using the equations presented here, the reader verify in the references cited.
Contents

Preface iii

Contents v

1. Introduction 1

2. Exploring the Time Series 3
   2.1 Introduction 3
   2.2 Time-series Plots 4
   2.3 Missing Value Statistics 6
   2.4 Below-detection-limit Data 7
   2.5 Normal Probability Plots or Other Q-Q Plots 8
   2.6 Scatter Plots 10
   2.7 Spectral Analysis 11
   2.8 Smoothers and First Estimations 17
   2.9 Transforming the Data? 19
   2.10 Aggregating the Data into Longer Periods than the Original One? 21

3. Formulation of a General Time-series Model 23

4. Techniques of Time-series Analysis 25
   4.1 Introduction 25
   4.2 Nonparametric Techniques Based on Rank 25
      4.2.1 Introduction 25
      4.2.2 Testing for a Long-term Trend 26
         4.2.2.1 Mann-Kendall Test 26
         4.2.2.2 Farrell Test 27
         4.2.2.3 Seasonal Kendall Test 29
         4.2.2.4 Hirsch- Slack Test 31
         4.2.2.5 Dietz-Kil len Test 32
         4.2.2.6 van Belle-Hughes' Test 33
      4.2.3 Slope Estimation 35
         4.2.3.1 Sen's Slope Estimator 35
         4.2.3.2 Seasonal Kendall's Slope Estimator (Gilbert, 1987) 36
      4.2.4 Discussion 37
   4.3 Smoothers 37
      4.3.1 Introduction 37
      4.3.2 Running Mean Smoother 38
      4.3.3 Running Median Smoother 41
      4.3.4 Robust Locally Weighted Regression (Loess) Smoother 41
      4.3.5 Kernel Smoothers 44
      4.3.6 Smoothing Splines 50
      4.3.7 Periodic Smoothers 53
      4.3.8 Selecting the Smoothing Parameters 54
         4.3.8.1 Cross-validation 54
         4.3.8.2 Risk Estimation 55
      4.3.9 Discussion 56
1. Introduction

In atmospheric chemistry, an ever-increasing amount of data has been collected and archived in the past 20 to 30 years. Some of these data come from short-term field studies lasting from a week to a few months. Others, collected in the context of network monitoring over many years or decades, can be used to construct long time series. It is in this latter context that the question of temporal variation in time series arises.

The study of the temporal variation of time-series data is not only a matter of checking to see if a long-term trend is present and, if so, estimating it. There are many other characteristics of a time series that have to be analyzed. In addition to long-term variations, one has to consider long-term cycles, seasonal cycles, and others. The study of all these variations can add something to our knowledge of the data. It is therefore important to isolate and examine all these temporal variations in the time series under study. Another characteristic of time series data in atmospheric chemistry is that the data are not independent of one another. Consecutive samples are usually highly correlated. Some of the correlation can be explained by the presence of the long-term trend, long-term cycles, seasonal cycles, and other explainable variations. However, there usually remains a part of the between-samples correlation that cannot be explained by these factors. It is therefore important to study this remaining part also.

In statistical data analysis in general and in time series analysis in particular it is important to keep in mind two rules:

1. Know your data;
2. Know the statistical techniques you want to use.

The first of these rules implies a general knowledge of how the data were collected, their uncertainties, the presence of missing data, and so on. In addition, knowing the time series means that we have some idea of the kind of temporal variations that can be found in it. It is important to have some general idea of the types of temporal variations that are present in the time series before subjecting the data to any statistical analysis.

The second rule is related to the fact that there are hypotheses made in all statistical techniques. Therefore, it is important to know these hypotheses and also to know as much as possible about the consequences of breaking them. These assumptions or hypotheses may be quite general in some cases and quite specific in others. For example, one statistical technique may require only that the data be from a continuous distribution. Other techniques, however, may require that the data be from a normal distribution.

Once one knows the characteristics of the data set that one wants to study and the hypotheses made by the statistical techniques that one wants to use, it is possible to estimate whether these hypotheses apply to this particular set of data. If such an analysis is not made, one risks being the victim of the old saying, "Garbage in → Garbage out", without even knowing it. Any statistical technique, and time series analysis techniques in particular, should not be used as a black box where we put data through the In door and expect to get perfect results from the Out door. Such an approach produces generally random and sometimes embarrassing results and is not to be recommended.

We will also not recommend rejecting the results of a statistical analysis because the data do not completely respect the assumptions made by the techniques. The effects on the results of a statistical technique when data do not conform to its assumptions depend first on
the technique and secondly on the importance of the nonconformity. For example, the Student t-test is somewhat tolerant of data that may deviate from the normal distribution, but the F-test used in the analysis of variance technique is not nearly as tolerant. To take another example, a test that assumes that the data are independent may tolerate small correlations between the data but not medium or large ones. In such a case, it is important to know what is small for that test and what the correlation between the data is before being able to decide if the results of the test can be used. In summary, the user of statistical techniques has to judge the validity of the study. The only way, he or she can do that is to know the data and to understand the techniques that he or she is using.

This overview has been written to help the reader accomplish these objectives. It is based on a general framework developed by the author over the past 15 years, during which he has spent much of his time studying time series. This framework can be summarized by the following five procedural steps:

1. Explore the data.
2. Formulate a general conceptual time series model.
3. Select an appropriate statistical technique(s).
4. Applying the technique(s) to the data.
5. Verify the hypothesis made by the statistical technique(s).

We will discuss and describe these steps in the remaining sections of this overview. Note that steps 4 and 5 are related and will be discussed together in Section 4.

The data that constitute a time series may have been collected in many different ways. They may represent samples of the same length of time or not. They may have been collected at a regular frequency or not. In the present overview, we will only consider data that represent, in theory, sampling of the same length, although a small fraction of the data may differ slightly from that ideal. The data may or may not be collected with equal frequency. Examples of the type of data set that are considered here are: (1) hourly O₃ measurements, (2) 24-hours filter pack measurements of ambient air, and (3) daily sampling of precipitation. This last is an example of equal length samples that are not collected at a regular frequency.
2. Exploring the Time Series

2.1 Introduction

The first question that the reader might ask at this stage is: “What should I know about my data?” There is no universal answer to that question, but in the context of time series analysis any answer should address the following points:

- Are there any missing data? If yes, what fraction of the data is missing? Is there a pattern to the missing data?
- Are all or part of the data missing because they were collected irregularly? Precipitation data are an example of irregularly collected data.
- Are there any outliers? If yes, are they many or few? Are they very different from the other data?
- Are there any below-detection-limit values? If yes, how should we treat them?
- What seems to be the distribution of the data? Is it normal or log-normal? Is there any transformation that will help in normalizing the data?
- Is there heteroscedasticity in the data (i.e., does the variance of the data changes with time)? If yes, could the data be transformed to stabilize the variance? Usually the same transformation will help in normalizing the data and stabilizing the variance.
- Is there a long-term trend in the time series? If yes, is it decreasing monotonically, increasing monotonically, or does it exhibit some other pattern?
- Are there other variations (e.g., long-term cycles) with long periods?
- If the time series extends over more than a year, are there some seasonal cycles? If yes, are they characterized by a simple sine or cosine function or by more complex ones?
- If the data are sampled at periods shorter than a day, are there some diurnal cycles?
- Should I take into account the relationship of the studied variables with each other before looking for temporal variations? An example of such a relationship is that between ionic concentration in precipitation and the amount of precipitation.

This list of questions is not exhaustive, and the reader could easily add several other pertinent questions to it.

We will try in the remainder of this section to present and discuss some of the techniques that may help to answer these questions. The techniques presented here are ones that the author has found useful in the past to get a better understanding or “feeling” about the time series that he was trying to analyze. They are presented here in the hope that they may also be useful to the reader.
2.2 Time-series plots

This is the most basic type of plot. An example of plotting the data against time is presented in Figure 2.1. In it, the SO\textsubscript{2} ambient concentration at Chalk River, Ontario, is shown between 1979 and 1994 without any transformation of the data (see Appendix I for a map showing the location of the CAPMoN sites). Although interesting, this plot is not very informative. All that can be derived from this plot is that: (1) there are some missing data; (2) there seems to be some seasonal variation; and (3) the data are mainly concentrated between 0 and about 15 \text{ \mu g m}^{-3} with some very high values present. From this last characteristic, one can
arrive at the conclusion that the data are not normally distributed. A more informal way to plot these data is to use a logarithmic scale for the ordinate axis.

The result is illustrated in Figure 2.2. One gets a better idea of the data in this figure than in the preceding. One can see first that most of the data in the second half of 1988 are missing. This is the only long period with missing data. Secondly, there are seasonal cycles in the data. An idea of these cycles can be had by creating a plot like the one presented in Figure 2.3. In this plot, all the years have been merged into one, and well-defined seasonal cycles are evident.

A long-term trend is also evident in Figure 2.2. SO$_2$ ambient air concentration has decreased during the period. This decrease seems to be

Figure 2.3 Seasonal variation of ambient SO$_2$ concentration in air at Chalk River, Ontario, using a logarithmic axis.

Figure 2.4 Box-plots for ambient SO$_2$ concentrations in air at Chalk River, Ontario, using a logarithmic axis and grouped by year. In each box, the central bar is the median and the lower and upper limits are the first and third quartiles respectively. The lines extending vertically from the box indicate the spread of the distribution, with the length being 1.5 times the difference between the first and third quartiles. Observations falling beyond the limits of these lines are considered to be outliers and are indicated by dots.
monotonic and more important between 1979 and 1985 than after. This example illustrates that it is not enough to show just a time series of the data. One must also respect the distribution of the data if one wants to get useful information from them.

Another type of graph that may be useful in detecting the presence of a long-term trend is a box-plot, as shown in Figure 2.4. In this figure, the data are grouped by year and a long-term trend is evident. The corresponding figure for seasonal variation is presented in Figure 2.5. These plots are very useful for detecting a specific type of temporal variation. Similar plots could be done also for diurnal data, such as hourly O₃ mixing ratios.

2.3 Missing Value Statistics

As mentioned in the introduction, it is important before doing any time series analysis to check first for missing values, and, if they are present, to consider their number and their distribution across the time series.

Missing values in a time series may result from many different causes. On the one hand, they may be due to the fact that the data are not sampled regularly. This would be the case, for example, for ionic concentration in precipitation. Precipitation does not occur every day, and a period of a few days with precipitation can be followed by a few days without it. These dry days will appear in the data bank as a data gap. On the other hand, a short data gap may be due to a broken instrument, for example. In the case of monitoring that extends over many years, much longer data gaps may occur for a variety of reasons.

An interesting way to summarize information about missing values is illustrated in Figures 2.6 and 2.7. For these plots, the length of each missing period was calculated. The plots present the number of such periods in the time series. Figure 2.6 is for ambient SO₂ air concentrations at Chalk River, Ontario. Figure 2.7 is for the ionic concentrations of SO₄²⁻ in precipitation at the same site.
One notices first the large difference in the number of missing values between the two time series. They represent less than 2.5 % of the data in the case of ambient SO$_2$ air concentrations but about 20% of the data in the case of ionic concentrations of SO$_4^{2-}$ in precipitation. The shapes of the histograms are also quite different. For the SO$_2$ concentration in air, most of the missing periods are one day long and the number of cases decreases rapidly for longer periods. For SO$_4^{2-}$ concentration in precipitation, the decrease in the number of cases is more gradual.

Plots like Figures 2.6 and 2.7 help in getting a better understanding of the pattern of the missing data. This also helps in estimating their impacts on the results.

2.4 Below-detection-limit Data

In some data sets, the number of below-detection-limit values may cause some problems. In the techniques presented in this overview, there is no method for taking below-detection-limit data into consideration. If the detection limit is not known, these items are usually treated as missing values, although it might be more correct to replace them by a

Figure 2.7 Number of periods of missing data of different length for SO$_4^{2-}$ concentrations in precipitation at Chalk River, Ontario.

Figure 2.8 Kernel smoother estimate of the long-term trend for weekly mean H$^+$ concentrations at Alert, Canada.
small value. (Note that the use of 0 may create problems later in the analysis. The author personally uses the value corresponding to the number of digits reported; for example, he will use a value of 0.01 for data reported with two decimals after the decimal point.) If the detection limit is reported, one usually uses 1/2 or 2/3 of the detection limit. If the former fraction is used, it is assumed that data that are lower than the detection limit follow a rectangular distribution. If the latter fraction is used, they are assumed to follow a triangular distribution, with the hypotenuse going from zero to the detection limit. The author prefers using the second value.

The impact of below-detection-limit data on the results of any time series depends on the number and temporal distribution of such data. If they are equally spread across the time series, their main impact will be to shift any long-term trend downwards, although they will usually have no impact on the shape of the trend. If, in addition, they occur mainly in one season, they may influence the seasonal cycle.

Examples of the impact of a large number of below-detection-limit data on the long-term trend and on the seasonal cycle are given in Figures 2.8 and 2.9. The kernel smoother (see Section 4.3.5) was used to estimate the long-term trend and the seasonal cycle. The data are weekly high volume samples collected at Alert, Canada (see Sirois and Barrie, 1999). As can be seen in the two figures, the below-detection-limit values are well distributed across the years but occur mainly in summer. In Figure 2.8, although the shape of the long-term trend seems to be correct, the curve looks as if it was shifted downward. Its mean should be about 8-9 ngm$^{-3}$ and not about 2 ngm$^{-3}$. The impact of the presence of the below-detection-limit data is greatest in the estimation of the seasonal cycle (Figure 2.9) because most of the below-detection-limit data occur during the summer months, thus influencing mainly the estimated seasonal cycle during those months.

When the number of below-detection-limit values is large, it is important to estimate their impact on the results of the time series analysis. This can be done by redoing this analysis with different values for the below-detection-limit data. This technique usually gives some idea of the impact of the below-detection-limit data.

2.5 Normal Probability Plots or Other Q-Q Plots

Normal probability plots have been used for a long time to verify the possible normality of the distribution of data. Normal probability plots are only one example of a class of graphs called Q-Q plots (see Wilk and Gnanadesikan, 1968; Michael, 1983). The name indicates that the quartiles of one distribution is plotted against the quartiles of another. In
normal probability plots, the quartiles of the observed data are drawn against the quartiles of the normal distribution. That concept may be extended to other types of distributions.

In the context of time series analysis, these types of graphs can be used first to see if the data should be transformed before the analysis and also to test, in the case of some techniques, whether one or more of their hypotheses are valid for the data used.

An example of a normal probability plot is given in Figure 2.10. To create this graph, the n concentration observations, C₁, C₂, C₃, ..., Cₙ, are first ordered in increasing order to produce the order statistics, C_{(1)}, C_{(2)}, C_{(3)}, ..., C_{(n)}. Then the C_{(i)} (ordinate) are plotted against N⁻¹((i-1/2) (abscissa) where N⁻¹(.) is the quartile of the normal distribution. In Figure 2.10, although the quartiles are used, the percentile values are indicated to simplify the interpretation. If the data were normally distributed, they would form a straight line, which is not the case for the present data set.

A better fit to a straight line can be obtained if a log-normal distribution is assumed. One should note that for this example (Figure 2.11) the fit is not yet perfect but is better than if a normal distribution is assumed.

This type of graphs can easily be extended to other distributions, such as Gamma or Weibull distributions.

These plots may also point to the necessity of transforming the data before plotting or analyzing them. They may also help in determining the type of transformation to use. This is illustrated in Figure 2.12. The data, after being transformed by means of a power transformation, were plotted on normal probability paper. One sees that, except
**Figure 2.12** Normal probability plots after transforming the data by raising them to power 0.50 (a), 0.25 (b), and 0.20 (c). The ambient SO₂ concentrations in air at Chalk River, Ontario, were used.

for the lowest 0.5%, the data come quite close to falling along a straight line if they are raised either to the one fourth or one fifth power.

One should note that it may be best to use these transformations before applying the model, because part of the non-normality in the data may be due to the presence of a long-term trends, a seasonal cycles, or other such variations.

### 2.6 Scatter Plots

We have mentioned earlier that the time series that we want to analyze may depend on another time series. An example in atmospheric chemistry is the correlation between the ionic concentration in precipitation and the precipitation amount. As the precipitation amounts may have its own temporal variations that may be due to other factors than the ones we want to detect in the ionic concentrations, it is important to take that dependence into consideration. How we take that information into account depends on the technique we employ to analyze the time series.

**Figure 2.13** Scattergram of NO₃⁻ concentrations in precipitation versus precipitation amount at Chalk River, Ontario between 1979 and 1995.
Figure 2.14 Smoothed power spectrum for ambient $\text{SO}_4^{2-}$ air concentration at Kejimkujik, Nova Scotia, from 1979 to 1994.

Figure 2.15: Smoothed power spectrum for ambient $\text{SO}_4^{2-}$ air concentration at Kejimkujik, Nova Scotia, from 1979 to 1994, plotted logarithmically.

The dependence of the studied time series on others can be checked by calculating the regression coefficients or by plotting scattergrams. This last technique is illustrated in the case of $\text{NO}_3^-$ concentration in precipitation at Chalk River, Ontario (Figure 2.13). This figure shows that the concentrations decrease as the precipitation amount increases. This should be taken into account when studying the temporal variation of $\text{NO}_3^-$ concentration in precipitation.

2.7 Spectral Analysis

Until now we have only looked at the general characteristics of the time series, discussing, for example, how to deal with missing data or below-detection-limit data. Before
testing for or estimating some of the temporal variations present in the data, we would like to have some idea of which types of temporal variation or which characteristics of the time series we should take into account. In other words, we would like to know what type of temporal variations are present in the time series. A very powerful tool for answering that question is spectral analysis.

Programs that estimate the power spectrum of a time series can be found in any statistical computer package. However, these programs usually assume that there are no missing data in the time series, which is not often the case in atmospheric chemistry. Two techniques exist to solve that problem. The first technique is based on the technique developed by Dunsmuir (1984) to estimate the auto-correlation function in the presence of missing data. The theory is described in Sirois et al. (1995), and the necessary equations are given in Appendix II. In the second technique, the power is calculated directly. This technique was developed by Lomb (1976) and based on earlier work by Barning (1963) and Vaniček (1971). The equations for the power spectrum are given in Appendix III. This second technique may be more appropriate in the case of ionic concentrations, in precipitation, for example.

The basic relation behind the interpretation of the power spectrum is Parseval’s relation (see Priestley, 1981) which can be written as:

\[ \sigma^2 = 2\pi \int_0^{0.5} f(\omega) \, d\omega \]  

(2.1)

where \( \sigma^2 \) is the variance of the data, \( f(\omega) \) is the power spectrum, and 0.5 is the Nyquist limit. Parseval’s relation indicates that the power spectrum is, in fact, the decomposition of the total variance into frequency dependent components.

An example of a power spectrum is given in Figure 2.14 for ambient \( \text{SO}_4^{2-} \) air concentrations at Kejimkujik, Nova Scotia, from 1979 to 1994. Note that the data were transformed using a logarithmic transformation before the power spectrum was calculated. Although relation 2.1 can be used directly, this plot is difficult to interpret. It is usually more
Figure 2.17 Power spectra for auto-regressive model (relation 2.2) of order 1 with $a_i = 0.25$ (a and c) and 0.5 (b and d).

in Informative to plot the logarithm of the power spectrum against the logarithm of the frequency. This is illustrated in Figure 2.15. This latter graph is easier to interpret than the preceding one. The reader should remember that Parseval’s relation cannot be applied directly in this case. In this plot, the frequency of the seasonal cycle is about 0.0027, that of the first harmonic is about 0.0055, and that of the third harmonic, about 0.0082.

Instead of trying to deduce all the types of temporal variations that can be present in the time series from Figure 2.15, we will try to show the impact on the power spectrum of adding different types of temporal variations, starting with random noise and adding autocorrelation, a seasonal cycle, a long-term cycle, and a long-term trend in that order. To do that we will create an artificial 15-year time series of daily data.
Figure 2.18 Power spectra for time series with seasonal cycles and auto-regressive model of order 1 with \( a_1 = 0.5 \). Plots (a) and (c) incorporate one cycle. Plots (b) and (d) incorporate one cycle and one harmonic.

Figure 2.16 presents the power spectrum of normally distributed random errors, or white noise. In this figure, one sees that the power spectrum of random noise consists of random variation around a straight line. If one compares Figure 2.16 to Figures 2.14 and 2.15, one sees that the ambient SO\(_4^{2-}\) air concentrations at Kejimkujik do not constitute a random time series.

It is well known in atmospheric chemistry that consecutive data are usually not independent. This process can be mathematically described by the following auto-regressive model:
\[ C_i = \sum_{i=1}^{K} a_i C_{i-1} + \varepsilon_i \tag{2.2} \]

where \( a_i \) are coefficients, \( K \) is the order of the auto-regressive model, and \( \varepsilon_i \) is a random error. The power spectrum of a time series that follows such a model depends on the order of the auto-regressive model and on the values of the coefficients \( a_i \). In atmospheric chemistry, auto-regressive models of order 1 to 3 are usually found. Typically, the first auto-regressive coefficient is much larger than the other ones. Two examples of power spectra for an auto-regressive model of order 1 are shown in Figure 2.17.

In Figures 2.17a and 2.17c, \( a_1 \) is equal at 0.25, and in Figures 2.17b and 2.17d it is equal to 0.50. If one compares these figures to Figure 2.16, one can see that the presence of the auto-regressive model has modified the power spectrum and that the modification is more important when the auto-regressive coefficient is largest. The modifications result in a transfer of the variance or energy from the highest frequencies to the lowest. Note that the effect will be opposite if the auto-regressive coefficient is negative. When the order of the auto-regressive model is greater than 1, its effects on the power spectrum are more complex. However if the first coefficient is positive and much larger than the others, the effects will be similar to the results shown in Figure 2.17. From a comparison of Figure 2.17 with Figure 2.15, one can conclude that auto-correlations are present in the data. The first coefficient of the auto-regressive model will be positive and much larger than any other coefficients present.

The next step is to add a seasonal cycle. Such a cycle can usually be described by a relation such as:

**Figure 2.19**: Smoothed power spectra for an artificial time series containing a decreasing linear trend, seasonal cycles, an one-order auto-regression model, and normally distributed white noise. The plots correspond to different values for the slope of the linear trend.
\[ \sum_{m=1}^{M_s} \left( C_{sm}^S \sin \left( \frac{2\pi \ mt}{365.25} \right) + C_{cm}^S \cos \left( \frac{2\pi \ mt}{365.25} \right) \right) \]  

(2.3)

where \( M_s \) is the number of waves. If \( M_s \) is greater than 1, harmonics are present.

Figure 2.18 shows the power spectrum when seasonal cycles are added to the auto-regressive model of order 1 with \( a_1 = 0.50 \). In Figures 2.18a and 2.18c, a simple cosine wave was used to describe the seasonal cycle. In the remaining plots of the figure, the seasonal cycle was described by a cosine wave plus a sine wave with a 6-months period and half the amplitude. Figure 2.18 shows that it is more and more difficult to interpret the power spectrum if one does not plot the logarithm of the smoothed power against the logarithm of the frequency. The presence of the seasonal cycles is very clear in these plots. If now we return to Figure 2.15, it is easy for the reader to detect the presence of a seasonal cycle with one main cycle and two harmonics.

To complete our study, we will now add a decreasing linear trend to the log-transformed data. The power spectra for two linear decreasing trends are given in Figure 2.19 for two values of the slope. The slope in Figure 2.19b is twice that in Figure 2.19a. The main effect of adding the trend is to increase the energy or variance at frequencies lower than about 0.001 cycles day\(^{-1}\).

A comparison of the power spectra in Figure 2.19 with the spectrum in Figure 2.15 shows that all the spectra are very similar. Therefore, using a simple model (linear long-term trend, plus a seasonal cycle, plus a one-order auto-regressive model, plus normally distributed white noise), we have reproduced the general lines of the power spectrum of the observed data.

In practice, we have to do the inverse of what we have done here, namely, start from the power spectrum and deduce the main types of temporal variations that contribute to the
variance observed in the data. This may seem a difficult proposition for a beginner, but with time one becomes adept at interpreting these spectra. The author hopes that the short demonstration presented here may help the reader in acquiring a better understanding of power spectra.

Although the power spectra produced in this section appear realistic, readers may ask whether the artificial time series from which they were realistically simulates observed ambient air concentrations. The time series corresponding to the power spectrum of Figure 2.19a is shown in Figure 2.20. The line is the kernel smoother estimation (see Section 4.3.5) of the long-term trend. One notices first that the variability may be somewhat too large when compared with most 24-hour ambient air sampling data. This is due to the fact that too large a value was selected for the variance of the white noise. This does not affect the fact that the time series appears to be as complex as an observed one, although it is only the sum of a few different types of temporal variations.

We have shown in this section that spectral analysis can be effective in detecting the presence of different types of temporal variation. However, it does not help us in getting an estimate of the shape or form of these temporal variations. We will address that problem in the next section.

2.8 Smoothers and First Estimations

Although spectral analysis can help in determining which types of temporal variations, if any, are present in a time series, it cannot help in determining their characteristics. One tool that can give a first idea of those characteristics is smoothing. Different smoothers are available (see Section 4.3). As an example, we will use the kernel smoother (Section 4.3.5; Siros, 1993, 1997) and the daily ambient SO$_2$ concentration at the Experimental Lakes Area in Ontario.

The power spectrum is shown in Figure 2.21. We can deduce from it that a strong long-term trend is present. A seasonal cycle with a first harmonic at least is present. There are also some indications of auto-correlations between the data.

Figure 2.22 presents the kernel smoother estimation of the long-term trend. One sees that the concentrations decrease mainly between 1979 and 1985. They may also increase again slightly between 1987 and 1991. The SO$_2$ concentrations appear to have decreased after 1991. It is interesting to note that even the missing data for 1983 do not prevent the kernel smoother from giving a good approximation of the long-term trend. In summary, the SO$_2$
Figure 2.22 Kernel smoother estimate of long-term trend for daily ambient SO$_2$ concentrations at the Experimental Lakes Area, Ontario.

Concentrations at the Experimental Lakes Area have decreased almost monotonically since 1979 but not linearly.

After eliminating the long-term trend, one can estimate the seasonal cycle. Figure 2.23 presents the periodic kernel smoother estimate of the seasonal cycle. One can see that the seasonal cycle is more complex than a simple sinusoidal wave.

After eliminating the estimated seasonal cycle, we would like to test for auto-correlation. One easy visual way to do this is to plot the scattergram of $C_t$ versus $C_{t+1}$, $C_t$ versus $C_{t+2}$, etc. Some of these graphs are shown in Figure 2.24.

One sees a net relationship between the concentrations at time $t$ and time $t-1$. The relationship becomes less clear in the case of the concentrations at time $t$ and time $t-2$, and at time $t$ and time $t-3$. One should therefore try to fit the auto-regressive model given in relation 2.2. There are many ways to fit such a model. However, because there are missing data in the time series, one has to use a technique that will allow the presence of missing data. The easiest way is to use the regression technique (see Chatfield, 1984; Botteheim and Sirois, 1996). In that technique, one regresses the concentration at time $t$ against the concentration at time $t-1$, $t-2$, etc. For

Figure 2.23 Periodic kernel smoother estimate of seasonal cycles for daily ambient SO$_2$ concentrations at the Experimental Lakes Area, Ontario.
the present data, the results indicate that
an auto-regressive model of order 2 is
applicable, with $a_1 = 0.43$ and $a_2 = 0.06$.\footnote{The
meaning of $a_1$ and $a_2$ is explained in Section 3.3.}
One should keep in mind that, as in the
other results presented in this section,
the order of the auto-regressive model
and the coefficients obtained are only
first approximations.

The results presented in this
section are very useful for the remainder
of the study. However, they cannot be
taken as definitive because (1) some of
the hypotheses assumed for kernel
smoothing are not usually fulfilled, and
(2) there is no way to check the
statistical significance of the results. We
will discuss these points later in Section
4.4.

Those results are also very
useful for the next step in the statistical
study of temporal variations, namely,
the formulation of a conceptual general
time series model to describe the time
series one wants to study. Before
discussing that step, however, we will
discuss two other important subjects:
(1) the possible transformation of the
data and (2) their aggregation to form
longer sampling periods.

2.9 **Transforming the Data?**

Should I transform the data?\footnote{The
meaning of transformation is explained in
Section 3.1.}
This question is sometimes very difficult to
answer. The answer is usually related to
the technique that one wants
to use to study the data. In the case
of nonparametric techniques (Section 4.2),
transformation is not usually necessary.
However, it may be necessary to
transform the data, not to get them to
normality, but to make their variance
independent of time, since one of the
usual assumptions of nonparametric
techniques is that the variance is
independent of time. Therefore, most of the
time the data collected in atmospheric
chemistry will have to be transformed because they are log-normally distributed, which means that the

![Figure 2.24 Scattergrams of de-trend and de-seasonal SO$_2$ concentrations at time $t$ and $t-1$ (a), $t-2$ (b), and $t-3$ (c) for daily ambient SO$_2$ concentrations at the Experimental Lakes Area, Ontario.](#)
variance is linearly related to the concentration (see Siros, 1991). A log-transformation of the data eliminates that relationship. The same is true for techniques such as smoothing (Section 4.3). The data may have to be transformed to get them close to normality if techniques like the anova techniques (Section 4.5) and the maximum likelihood technique (Section 4.6) are used. Finally, it may be necessary in some cases to transform the data to make their distribution symmetrical.

In most cases, a simple power transformation will be enough to transform the data to normality and/or stabilize the variance, and/or get a symmetrical distribution. Such a transformation can be expressed as:

$$T_p(x) = \begin{cases} 
  x^p & (p > 0) \\
  \log(x) & (p = 0) \\
  -x^p & (p < 0)
\end{cases} \quad (2.4)$$

The most frequently used values are $p = 0$ and $p=1/2$. Power transformations are discussed in Emerson and Stoto (1983) and Emerson (1983).

In the case of logarithmic transformations, it is sometimes necessary to use a shifting parameter. The transformation then becomes:

$$T(x) = \log(x + \delta) \quad (2.5)$$

Techniques to estimate $\delta$ are given in Aitchison and Brown (1957), Crow and Shimizu (1988), and Wheeler (1980).

Another system of transformation to normality is the Johnson System (Johnson, 1949). This system consists of three families of distributions, which are generated by a transformation of the form:

$$T(x) = \gamma + \eta k_i(x; \lambda, \varepsilon) \quad (2.6)$$

where $T(x)$ is a standard normal variable, and the $k_i$ are selected to cover a wide range of shapes.

The three families suggested by Johnson (1949) are:

1) the $S_u$ distribution with $k_1(x; \lambda, \varepsilon) = \sinh^{-1}\left(\frac{x - \varepsilon}{\lambda}\right)$

2) the $S_b$ distribution with $k_2(x; \lambda, \varepsilon) = \ln\left(\frac{x - \varepsilon}{\lambda + \varepsilon - x}\right) \quad (2.7)$

3) the $S_l$ distribution with $k_3(x; \lambda, \varepsilon) = \ln\left(\frac{x - \varepsilon}{\lambda}\right)$

The $S_l$ distribution is, in essence, a three-parameter log-normal distribution, since the parameter $\lambda$ can be eliminated by letting $\gamma^* = \gamma - \eta \ln \lambda$, so that $T(x) = \gamma^* + \eta \ln(\lambda - \varepsilon)$. The $S_b$ is a distribution bounded on $(\varepsilon, \varepsilon + \lambda)$, and the $S_u$ is an unbounded distribution. The first step in using Johnson’s system of transformation is to select which of the three families to use. Then, the parameters of the family can be estimated. More details about Johnson’s system of transformations can be found in Johnson (1949, 1965), Draper (1952), Bukac (1972), Mage (1980), Slifcker and Shapiro (1980), and Wheeler (1980). Fortran programs to fit Johnson’s system can be found in Hill et al. (1976) and Hill (1976).
In summary, the important thing to remember is that one transforms the data not necessarily to normalize them but also to stabilize the variance or render the distribution symmetrical. Therefore, one sometimes has to transform the data, even if one wants to use nonparametric techniques to analyze the temporal variation in the data.

2.10 Aggregating the Data into Longer Periods than the Original One?

In the past, researchers often aggregated data to a longer period before analyzing them. This was sometimes done because one can usually get better normal approximations with longer sampling periods. Another reason was that one could reduce or eliminate the effects of missing data. Frequently, one would calculate aggregated data to eliminate autocorrelation and therefore simplify the analysis. However, when data gaps were present, this became a very complex problem because: (1) the aggregation technique could introduce a bias into the aggregated data estimate, and (2) the increase in the variance of the aggregated estimate would usually depend on parameters like the percentage of data missing. Therefore, the variance of the data would not be constant in time except if the number of data items missing was also constant in time.

As an example, Sirois (1990) has shown that, when data are missing, estimates of the monthly, seasonal, or annual precipitation-weighted-mean concentration of ions in precipitation are biased, and that the magnitude of the bias depends on the percentage of data missing and the percentage of the total precipitation associated with missing data. He has also shown that the variance of the estimation depends on those two parameters as well.

In summary, although it may seem that aggregating data to longer sampling periods will solve some of the difficulties usually associated with the analysis of temporal variation in atmospheric chemistry, the procedure introduces characteristics into the data (i.e., biases and variability that depend on parameters that vary with time) that invalidate most of the techniques presented in the present overview. Therefore, the author advises against aggregating the data, except if the amount of missing data is very small.
3. **Formulation of a General Time-series Model**

This step in the temporal analysis of a time series is a very important one for many reasons. First, this model embodies our knowledge of the characteristics of the time series. Secondly, this model also delimits the techniques that can be used to analyze the time series. In practice, the conceptual model and the technique to be used are selected at the same time, because behind all the techniques presented in the next section there is some conceptual model of the time series.

This conceptual model may be very simple or very complex; very general or very specific. An example of a very simple conceptual model that would apply to most of the smoothing techniques presented in Section 4.3 is the following:

\[ C_i = f(t_i) + \varepsilon_i \]  \hspace{1cm} (3.1)

where \( f(t) \) is a continuous function and the \( \varepsilon_i \) values are random noises from the same distribution with mean zero and constant variance. In most cases also, the existence of the first and second derivatives of \( f(t) \) is assumed.

Model 3.1 also applies in the case of nonparametric techniques if we add the condition that \( f(t) \) should be monotonic. This can be written as:

\[ f(t_i) \geq f(t_j) \quad \text{if} \quad t_i > t_j \]  \hspace{1cm} (3.2)

for a monotonically increasing function and

\[ f(t_i) \leq f(t_j) \quad \text{if} \quad t_i > t_j \]  \hspace{1cm} (3.3)

for a monotonically decreasing function.

For some of the techniques like the anova technique, model 3.1 is valid but the distribution of \( \varepsilon_i \) is assumed to be normal.

An example of a more complex conceptual model is one used by Sirois (1997). It can be written as:

\[ \log(C(t_i)) = C_0 + f^T(t_i) + f^C(t_i) + f^S(t_i) + f^D(t_i) \]  \hspace{1cm} (3.4)

where \( C_0 \) is a constant; \( f^T \), \( f^C \), \( f^S \) are general functions that describe the long-term trend, the long-term cycles, and the seasonal cycles respectively; and \( f^D \) is an auto-regressive model written as:

\[ f^D(t_i) = a_1 f^D(t_{i-1}) + a_2 f^D(t_{i-2}) + \ldots + a_N f^D(t_{i-N}) + \varepsilon_i \]  \hspace{1cm} (3.5)

where \( N \) is the order of the auto-regressive model; \( \varepsilon_i \) is the random noise component drawn from a normal distribution with mean zero and constant variance \( \sigma^2 \). This type of model is useful for techniques like those discussed in Section 4.6.

The important notion to keep in mind is that one should have a conceptual model and that it should be consistent with the technique used. It should also be consistent with what it is known about the data and especially with the results obtained in the preceding section. If the conceptual model do not describe the time series properly, there is a good chance that the method one wants to use does not apply to the data. One should always keep in mind that the results of any statistical analysis are valid only if the assumptions related to that technique are fulfilled by the data. As mentioned earlier, small deviations from these conditions may not
always invalidate the results, but they will increase the uncertainties, depending on the importance of the deviations.

Formulating the conceptual model also helps in the selection of the correct statistical techniques to use. One usually starts by extracting the main characteristics of the time series, employing the techniques described in the preceding section. A conceptual model is then constructed, and finally one or more of the statistical techniques compatible with the conceptual model is selected. *The reader should always remember that the data should drive the construction of the conceptual model and not the wishes of the researcher. We should always bear in mind that we do not control nature yet.*
4. Techniques of Time-series Analysis

4.1 Introduction

Many techniques have been developed during the past century or so to analyze the temporal variations of series of numbers or time series. These techniques range from the simple fitting of a straight line by least squares to a complex ARIMA (Auto-Regressive Integrated Moving-Average) model. In each of these techniques some hypothesis is made about the nature of the data and the characteristics of their temporal variations. These hypothesis can usually be summarized into a conceptual model of the data to which the technique applies. As we have seen in the preceding section, some of these models may be simple and others more complex. In either cases, the observed data are the important factor that determines the correct selection of the conceptual model and thus of the technique used to analyze them.

In this section, we will describe some of the techniques that have been used in atmospheric chemistry to analyze the temporal variation of different observed pollutants. For each of these techniques, we will not only describe how to apply them but will also explain the hypotheses about the data and the conceptual models that are implied when using them. We will also point out their limitations, if any.

The techniques have been grouped according to their main characteristics. We will first describe nonparametric techniques based on rank. We will then discuss smoothers that are nonparametric but that are associated with a somewhat different conceptual model. Next, we will mention the use of spectral analysis to test and estimate some of the temporal variations. We will finish by examining techniques that require the construction of more hypotheses but that are also more general in their application.

The present overview, although not exhaustive, should give a good idea of the techniques available for the analysis of temporal variation of observed data.

4.2 Nonparametric techniques based on rank

4.2.1 Introduction

The techniques presented in this section use basically the same conceptual model, which can be written as:

\[ C_i = f(t_i) + \varepsilon_i \]  

(4.1)

It is assumed that \( f(t) \) is a continuous monotonic increasing or decreasing function. The \( \varepsilon \) values are assumed to be from the same distribution with zero mean. It is therefore assumed that the dispersion (i.e., variance) of the distribution is constant in time.

In some of the techniques, the data are grouped to form a two-way or even three-way layout for the regional version of the tests. Usually, one of the layout indices will be the year; another will be either the week, month, or season; and in the cases of regional tests, the third will be the site. In these techniques, it is usually assumed that there is one observation for each layout cell, although it is possible in some cases to extend the technique to include more
than one observation per cell. We will indicate those cases and the changes that must be made to the equations.

Two types of analysis can be carried out using nonparametric techniques. First, one may test for the presence of a monotonic increasing or decreasing long-term trend, and secondly one may estimate the slope of a linear trend. We will address these two applications separately.

4.2.2 Testing for a Long-term Trend

Many nonparametric statistical tests exist. Since we cannot report all of them, we decided to limit ourselves to those that are most important or useful in atmospheric chemistry. Six tests are described here. Except for the first, all try to take the presence of seasonality into account, and assume monotonic trends for each season. Unless otherwise mentioned, the trends are assumed to be homogeneous across the seasons. The random errors are assumed to be independent, except in the case of the Hirsch-Slack test.

4.2.2.1 Mann-Kendall Test

Let be a sequence of observations, \(x_1, x_2, \ldots, x_n\), ordered by time. One wants to test the following hypothesis:

\[ H_0: \text{the observations are randomly ordered, i.e., } x_1, x_2, \ldots, x_n \text{ are independent identically distributed (or i.i.d.) samples,} \]

against the hypothesis:

\[ H_1: \text{there is a monotone trend over time, i.e., } F_{x_i}(x) \geq (\text{or } \leq) F_{x_j}(x) \text{ for all } i < j \text{ with at least one strict inequality,} \]

where \(F_{x_i}(x)\) is the cumulative distribution function of the random vector \(x_i\). Using conceptual model 4.1, the test can be described as \(H_0: f(t) = \text{constant}\) against \(H_1: f(t)\) a monotonic increasing or decreasing function.

Mann (1945) described a nonparametric techniques for applying such a test. Mann’s test is in fact a particular application of Kendall’s test for correlation (Kendall, 1938, 1975), commonly known as Kendall’s tau. The test statistic \(S\) is defined as

\[ S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \text{sgn}(x_j - x_k) \]  

where

\[ \text{sgn}(\theta) = \begin{cases} 
1 & \text{if } \theta > 0 \\
0 & \text{if } \theta = 0 \\
-1 & \text{if } \theta < 0 
\end{cases} 
\]

Note that the statistic \(T\) used by Mann (1945) is related to the statistic \(S\) used by Kendall (1938, 1975) by \(S = 2T - n(n-1)/2\). Mann showed that under \(H_0\) the distribution of \(T\) and
hence $S$ is symmetrical and is normal in the limit as $n \to \infty$. Under $H_0$, the mean and variance of $S$, if there is ties in $x$, are equal to:

$$E[S] = 0$$

$$\text{Var}[S] = \frac{n(n-1)(2n+5) - \sum t(t-1)(2t+5)}{18}$$

(4.4)

(4.5)

where $t$ is the extent of any given tie (number of $x$'s involved in a given tie) and the summation in relation 4.5 is over all ties. Both Mann and Kendall derive the exact distribution of $S$ for $n \leq 10$ and show that even for $n = 10$ the normal approximation is excellent, provided one uses a continuity correction of one unit. One computes the standard normal variate $Z$ by:

$$Z = \begin{cases} 
\frac{S - 1}{(\text{Var}(S))^{1/2}} & \text{if } S > 0 \\
0 & \text{if } S = 0 \\
\frac{S + 1}{(\text{Var}(S))^{1/2}} & \text{if } S > 0
\end{cases}$$

(4.6)

In a two-sided test, the hypothesis $H_0$ should be accepted if $|Z| \leq Z_{\alpha/2}$, where $F_N(Z_{\alpha/2}) = \alpha/2$, $F_N$ being the standard normal cumulative distribution function and $\alpha$ being the size of the significance level for the test. A positive value for $S$ indicates an increasing monotonic long-term trend and a negative $S$ a decreasing long-term trend.

To illustrate the use of the Mann-Kendall test, we will test for a monotonic trend using the 28-day mean $SO_4^{2-}$ concentration in precipitation at Montmorency, Quebec. The data are shown in Figure 4.1. Using relations 4.2 and 4.3, one obtains $S = -523$ and $\text{Var}(s) = 85083$. The P-value for the test using the normal approximation is equal to 0.074. Therefore, one cannot reject the hypothesis of no long-term trend at a 95% confidence level.

4.2.2.2 Farrell Test

In atmospheric chemistry, as in other fields, it often happens that a seasonal cycle is present in the data and that the long-term trend also varies with the season. Responding to that problem, Farrell (1980), following Sen (1968a), proposed the following test.

First subtract the monthly (or seasonal) average from each of the corresponding months (or seasons) in the $n$ years of data; i.e., calculate $x_{ij} - \bar{x}_j$ for year $i = 1$, 2, ..., $n$ and month $j = 1$, 2, ..., $m$.

![Figure 4.1 28-day precipitation-weighted-mean $SO_4^{2-}$ concentrations in precipitation at Montmorency, Quebec.](image)
where \( x_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \). Then, rank all the differences from 1 to \( nm \). This produces the matrix:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>\ldots</th>
<th>m</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( R_{11} )</td>
<td>( R_{12} )</td>
<td>( R_{13} )</td>
<td>\ldots</td>
<td>( R_{1m} )</td>
<td>( R_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( R_{21} )</td>
<td>( R_{22} )</td>
<td>( R_{23} )</td>
<td>\ldots</td>
<td>( R_{2m} )</td>
<td>( R_2 )</td>
</tr>
<tr>
<td>Year</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>( n )</td>
<td>( R_{n1} )</td>
<td>( R_{n2} )</td>
<td>( R_{n3} )</td>
<td>\ldots</td>
<td>( R_{nm} )</td>
<td>( R_n )</td>
</tr>
<tr>
<td>mean</td>
<td>( R_1 )</td>
<td>( R_2 )</td>
<td>( R_3 )</td>
<td>\ldots</td>
<td>( R_m )</td>
<td>( R_ )</td>
</tr>
</tbody>
</table>

where \( R_{ij} \) = rank of \( (x_{ij} - x_{ij}) \) among the \( nm \) differences. Ties are dealt with in the usual manner; to each of the \( t \) tied values, assign the average of the next \( t \) ranks. \( R_i \) is the average rank for each year, equal to \( \frac{1}{m} \sum_j R_{ij} \), and \( R_j \) is the average rank for each month, equal to \( \frac{1}{n} \sum_i R_{ij} \). We then calculate the statistic \( T \), defined as:

\[
T = \left( \frac{12m^2}{n(n+1) \sum_{i=1}^{n} \sum_{j=1}^{m} (R_{ij} - R_j)^2} \right)^{1/2} \left( \sum_{i=1}^{n} \left( i - \frac{n+1}{2} \right) \left( R_i - \frac{nm+1}{2} \right) \right) \tag{4.7}
\]

As \( m \) becomes larger, the distribution of \( T \) tends toward normality, with mean 0 (under the null hypothesis of no trend) and variance 1. The resulting statistical test is, therefore, straightforward: reject the hypothesis of no trend if \(|T| \) exceeds a prespecified percentile of the normal distribution. Some limited Monte Carlo studies indicate that the normal approximation is reasonable even for small samples.

The period described by the second index may be a month, week, or season. Note that this test does not allow for missing data. Farrell (1980) suggests filling in missing data with least squares estimates. That is, one estimates, via least squares, the parameter of the following model:

\[
x_{ij} = u + a_i + b_j + e_{ij} \tag{4.8}
\]

with \( i = 1, 2, \ldots, n \) and \( j = 1, 2, \ldots, m \) and

\[
\sum_i a_i = \sum_j b_j = 0 \tag{4.9}
\]

and then uses those estimates to calculate values for the missing \( x_{ij} \). Note that this procedure is not recommended by van Belle and Hughes (1984) if many values are missing, because the variance estimate implicit in relation 4.7 is then too small.
The reader should also note that this test assumes that the trends for the m months or seasons are homogeneous (i.e., the trend is the same in each month or season). Therefore, the test is invalid if the trends are heterogeneous.

4.2.2.3 Seasonal Kendall Test

Hirsch et al. (1982) have extended the Mann-Kendall test to take into account any seasonality in the data. They called this test the seasonal Kendall test.

In their test, the hypothesis to test is modified as follows. Let

\[ X = (X_1, X_2, \ldots, X_{12}) \] (4.10)

and

\[ X_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \] (4.11)

where \( X \) is the entire sample made up of subsamples \( X_1, X_2, \ldots, X_{12} \) (one for each month), and each subsample contains the ni annual values from month i. Note that ni may be different from nj (i \( \neq \) j). Therefore, missing values are allowed. We will discuss a modification to the test to allow more than one sample per month. Even if the test is presented here for 12 monthly values per year, it can also be used for 52 weekly values or 4 seasonal values or other combinations. The null hypothesis for the seasonal Kendall test is:

\[ H_0: \text{X is a sample of independent random variable (} x_{ij} \text{) and } X_i \text{ is a subsample of independent and identically distributed random variables } i = 1, 2, \ldots, 12. \]

and

\[ H_1: \text{for one or more months there is a monotonic increasing or decreasing trend.} \]

We define \( S_i \) as

\[ S_i = \sum_{k=1}^{n_i-1} \sum_{j=k+1}^{n_i} \text{sgn}(x_{ij} - x_{ik}) \] (4.12)

Under \( H_0 \), the subsample \( X_i \) satisfies the null hypothesis \( H_0 \) of the Mann-Kendall test. Therefore, one has (Section 4.2.2.1):

\[ E[S_i] = 0 \] (4.13)

\[ \text{Var}[S_i] = \frac{n_i(n_i - 1)(2n_i + 5) - \sum t_i(t_i - 1)(2t_i + 5)}{18} \] (4.14)

where \( t_i \) is the extent of a given tie in month i. The distribution of \( S_i \) is normal in the limit as \( n_i \rightarrow \infty \). We then define \( S' = \sum_{i=1}^{12} S_i \) and can derive its expectation, variance, and limit distribution.

\[ E[S'] = \sum_{i=1}^{12} E[S_i] = 0 \] (4.15)
\[ Var[S'] = \sum_{i=1}^{12} Var[S_i] + \sum_{i=1}^{12} \sum_{j=1, j \neq i}^{12} \text{cov}(S_i, S_j) \]  

(4.16)

Now \( S_i \) and \( S_j \) (\( i \neq j \)) are functions of independent random variables \( (S_i = f(X_i), S_j = f(X_j)) \), and \( X_i \cap X_j = \phi \) because \( X_i \) and \( X_j \) are the data from months \( i \) and \( j \) respectively, and all elements of \( X \) are independent, so \( \text{cov}(S_i, S_j) = 0 \). Note that this is only true if there is no autocorrelation in the time series. We will discuss an extension to the test that includes autocorrelation in the next section. Thus, we have:

\[ Var[S'] = \sum_{i=1}^{12} Var[S_i] = \sum_{i=1}^{12} \left\{ \frac{n_i(n_i - 1)(2n_i - 5) - \sum t_i(t_i - 1)(2t_i + 5)}{18} \right\} \]  

(4.17)

For using the normal approximation we define the standard normal deviate \( S' \) as:

\[ Z' = \begin{cases} 
\frac{S' - 1}{(Var(S'))^{1/2}} & \text{if } S' > 0 \\
0 & \text{if } S' = 0 \\
\frac{S' + 1}{(Var(S'))^{1/2}} & \text{if } S' > 0 
\end{cases} \]  

(4.18)

which is adequate for \( n_i \geq 3 \) for all \( i \).

When there are multiple observations for each month (or period), the preceding relation can be modified. The basic idea is to treat the data in the same month as tied observations with respect to their time of occurrence. Relation 4.12 becomes:

\[ S_i = \sum_{k=1}^{n_i - 1} \sum_{j=k+1}^{n_i} \sum_{l_1=1}^{n_{ij}} \sum_{l_2=1}^{n_{ij}} \text{sgn}(x_{ijl_1} - x_{ijl_2}) \]  

(4.19)

where \( x_{ijk} \) is observation \( k \) for month \( i \) and year \( j \) and \( n_{ij} \) is the number of observations in month \( i \) and year \( j \). Note that the \( n_{ij} \) can be different, so data can be missing without affecting the results. Because of the ties in time the variance of \( S_i \) must be modified. It is equal to (Kendall, 1975):

\[ Var[S_i] = \frac{n_i(n_i - 1)(2n_i + 5) - \sum t_i(t_i - 1)(2t_i + 5) - \sum_{k=1}^{n_i} n_{ik}(n_{ik} - 1)(2n_{ik} + 5)}{18} \]

\[ + \frac{\left\{ \sum t_i(t_i - 1)(t_i - 2) \right\} \left\{ \sum_{k=1}^{n_i} n_{ik}(n_{ik} - 1)(n_{ik} - 2) \right\}}{9n_i(n_i - 1)(n_i - 2)} \]

\[ + \frac{\left\{ \sum t_i(t_i - 1) \right\} \left\{ \sum_{k=1}^{n_i} n_{ik}(n_{ik} - 1) \right\}}{2n_i(n_i - 1)} \]  

(4.20)
$S'$ is still defined as $\sum_{i=1}^{12} S_i$ and $Var[S'] = \sum_{i=1}^{12} Var[S_i]$. Relation 4.18 can still be used to test the significance of the monotonic long-term trend.

Like the Farrell test, this test is not valid if the long-term trends in the different months or seasons are heterogeneous.

Using the data presented in Figure 4.1 and discussed in Section 4.2.2.1, the seasonal Kendall test gives us $S' = -84$ and $Var[S'] = 707.33$, and the P-value is equal to 0.0018. We can thus conclude that a decreasing monotonic trend that is statistically significant at a 95% confidence level is present in the data.

4.2.2.4 Hirsch- Slack Test

As mentioned in the preceding section, the seasonal Kendall test is incorrect when the data are correlated because the term $\sum_{i=1}^{12} \sum_{j=1}^{12} \text{cov}(S_i, S_j)$ in relation 4.16 was put equal to 0. Hirsch and Slack (1984) generalized the seasonal Kendall test to correlated data using a consistent estimator for $\text{cov}(S_r, S_j)$ developed by Dietz and Killeen (1981). It can be written as:

$$\text{cov}(S_g, S_h) = K_{gh}/3 + (n^3 - n)r_{gh}/9$$  \hspace{1cm} (4.21)

where

$$K_{gh} = \sum_{i<j} \text{sgn}[(x_{jg} - x_{ig})(x_{jh} - x_{ih})]$$  \hspace{1cm} (4.22)

$$r_{gh} = \frac{3}{n^3 - n} \sum_{i,j,k} \text{sgn}[(x_{jg} - x_{ig})(x_{jh} - x_{ih})]$$  \hspace{1cm} (4.23)

If there are no ties and no missing values, $r_{gh}$ is the Spearman’s correlation coefficient (Lehmann, 1975) for seasons or months $g$ and $h$. If there are no missing values, relation 4.21 reduces to:

$$\text{cov}(S_g, S_h) = \left[ K_{gh} + 4 \sum_{i=1}^{n} R_{ig} R_{ih} - n(n+1)^2 \right]/3$$  \hspace{1cm} (4.24)

where

$$R_{ig} = \left[ n + 1 + \sum_{i=1}^{n} \text{sgn}(x_{jg} - x_{ig}) \right]/2$$  \hspace{1cm} (4.25)

If missing data are present, the definition of the $\text{sgn}$ function in relation 4.3 is extended. We define $\text{sgn}(x_{jg} - x_{ig})$ to be zero if either $x_{jg}$ or $x_{ig}$ is missing. Relation 4.25 becomes:

$$R_{ig} = \left[ n_g + 1 + \sum_{i=1}^{n} \text{sgn}(x_{jg} - x_{ig}) \right]/2$$  \hspace{1cm} (4.26)

where $n_g$ is the number of nonmissing observations for season $g$. Now the ranks of the nonmissing observations are unchanged, and each missing value is assigned the average or
midrank value \((n_g+1)/2\). The Mann-Kendall test statistic \(S_g\) (relation 4.12) is unchanged and its variance remains the same, namely,

\[
Var[S_g] = \frac{n_g(n_g-1)(2n_g+5) - \sum t_g(t_g-1)(2t_g+5)}{18}
\]

(4.27)

In relation 4.24, \(K_{gh}\) remains unchanged, but \(r_{gh}\) takes on a new value so that

\[
\text{cov}(S_g, S_h) = \left[ K_{gh} + 4\sum_{i=1}^{n} R_{gi} R_{hi} - n(n_g+1)(n_h+1) \right]/3
\]

(4.28)

When there are below-detection-limit data, they may be replaced arbitrarily by a small value which is less than the detection limit, because the nonparametric tests are based on ranks rather than magnitudes. Therefore, all censored values (i.e., below-detection-limit values) may be viewed as sharing the same rank. Therefore, the problem of dealing with censored values reduces to the problem of dealing with ties. If there are ties but no missing values, \(S_g\) is computed using relation 4.12 and \(S'\) is the sum of the \(S_g\) values. The variance of \(S_g\) is given by relation 4.14. The formula for \(\text{cov}(S_g, S_h)\) remains the same, except that the midranks are used in assigning the values of \(R_{gi}\) for relation 4.24. Thus, if there are \(t_i\) censored values, they all have rank \(t_i(t_i-1)/2\).

When ties and missing values are both present, one must combine the modifications described for missing values and for ties. Note that, for this test there is only one observation per month (or season) of each year.

The original seasonal Kendall test (Section 4.2.2.3) is more powerful than the present test, but the significance level of the test can be seriously in error if there is serial correlation. The present test requires some sacrifice of power but offers a more nearly exact statement of significance for a wide variety of cases. Thus, choosing between the seasonal Kendall test and the Hirsch- Slack’s test involves a trade-off.

Like the original seasonal Kendall test, the Hirsch and Slack test assumes that the trends in the different months or seasons are homogeneous. Thus both are invalid if that hypothesis is not fulfilled.

### 4.2.2.5 Dietz-Killeen Test

The test proposed by Dietz and Killeen (1981) for monotonic long-term trends in multivariate data can also be adapted for use with monthly (or seasonal) data covering many years. As in the case of the seasonal trend, let

\[
X = (X_1, X_2, \cdots, X_m)
\]

(4.29)

and

\[
X_j = (x_{i1}, x_{i2}, \cdots, x_{in})
\]

(4.30)

\(X\) is the entire sample made up of subsamples \(X_1, X_2, \ldots, X_p\) (one for each period); \(x_{ij}\) is the observation for year \(i\) and period \(j\). Note that \(m\) equals 12 for monthly data. The null hypothesis is that the \(p\)-vectors are randomly ordered versus the alternative hypothesis that there is a monotonic trend in one or more of the \(m\) periods.

Let
\[ S = (S_1, \ldots, S_m)^T \]  
\[ \Sigma = (\sigma_{gh}) \]  
\[ S_g = \sum_{i<j} \text{sgn}(x_{ig} - x_{ij}) \quad g = 1, \ldots, m \]  
\[ \sigma_{gg} = n(n-1)(2n+5)/18 \quad g = 1, \ldots, m \]  
\[ \sigma_{gh} = \frac{1}{3} \left\{ \sum_{i<j} \text{sgn}[(x_{ig} - x_{ij})(x_{ig} - x_{ij})] + \sum_{(j,i,k)} \text{sgn}[(x_{ig} - x_{ij})(x_{ig} - x_{ik})] \right\} \]  
where \( g \neq h \). This last relation gives the estimated covariances of \( S_g \) and \( S_h \). The statistic \( T \) defined as:

\[ T = S^T \Sigma^{-1} S \]  
is asymptotically \( \chi_q^2 \) (i.e., chi-squared with \( q \) degrees of freedom), where \( \Sigma^{-1} \) is any generalized inverse of \( \Sigma \) and \( q \leq m \) is the rank of \( \Sigma \). In the presence of ties, relation 4.34 becomes

\[ \sigma_{gg} = \frac{n_i(n_i - 1)(2n_i + 5) - \sum_{t_i} t_i(t_i - 1)(2t_i + 5)}{18} \quad g = 1, \ldots, m \]  
and the remaining relations do not change. Therefore, the hypothesis of no monotonic trend in the data can be rejected if \( T \) is greater than \( \chi_q^2 \).

Note that this test is very conservative. Hirsch and Slack (1984) argued, on the basis of very limited Monte Carlo simulations, that it is probably applicable only to monthly data where at least 40 years of data are available.

Note that, unlike the seasonal tests presented earlier, this test is valid even if the trends in the different seasons are heterogeneous.

4.2.2.6 van Belle-Hughes’ Test

As we have seen in the descriptions of the Farrell, seasonal Kendall, and Hirsch-Slack tests, these tests are invalid if the trends in the different months or seasons are heterogeneous. Only the Dietz-Killeen test is valid in this situation, but it may only be applicable if a very long-time series is available. All the other tests of seasonal data mentioned above will be misleading if the trends are not homogeneous among months or seasons, especially when there are opposing trends in different seasons.

van Belle and Hughes (1984) developed a 2-way, anova-like, nonparametric test which can test first for the homogeneity of the trend at different locations and different seasons and then test for the presence of trends if the hypothesis of homogeneity cannot be rejected. For a single location, the test is constructed as follows.
Let the variable $Z_j$ be defined as

$$Z_j = \frac{S_j}{(\text{Var}(S_j))^{1/2}}$$

(4.38)

where $S_j$ and $\text{Var}(S_j)$ are defined in relation 4.12 and 4.14 respectively. Under the hypothesis of no trend in season $j$, $Z_j^2$ has, approximately, a chi-squared distribution with 1 degree of freedom (this follows from the asymptotic normality of $S_j$; see above). In addition, if the seasonal observations are far enough apart, then the $Z_j$ will be nearly independent, which is assumed in what follows. The overall statistic is

$$\chi^2_{\text{Total}} = \sum_{j=1}^{m} Z_j^2$$

(4.39)

which is, approximately, a $\chi^2_m$ under the null hypothesis of no trend in any of the seasons. Note that each season’s value is normalized before the total statistic is calculated. This normalization is equivalent to weighting each monthly or seasonal statistic with a weight that is inversely proportional to its variance. Therefore, the months or seasons with high variances contribute less to the total statistic.

A large value of $\chi^2_{\text{Total}}$ is not really meaningful because it still fails to distinguish heterogeneity between the individual $Z_j^2$'s from the overall trend. To do that, we will partition $\chi^2_{\text{Total}}$ in the following manner:

$$\chi^2_{\text{Total}} = \chi^2_{\text{homogeneous}} + \chi^2_{\text{trend}}$$

(4.40)

where $\chi^2_{\text{trend}} = \frac{m \bar{Z}^2}{\sum_{i=1}^{m} Z_i^2}$ and $\bar{Z} = \frac{1}{m} \sum_{i=1}^{m} Z_i$. $\chi^2_{\text{homogeneous}}$ is obtained by subtracting $\chi^2_{\text{trend}}$ from $\chi^2_{\text{Total}}$. Under the hypothesis that there is no trend in any of the months or seasons, $\chi^2_{\text{homogeneous}}$ and $\chi^2_{\text{trend}}$ have, respectively, chi-squared distributions, with m-1 and 1 degrees of freedom. Thus homogeneity of trend may be tested by comparing $\chi^2_{\text{homogeneous}}$ to tables of $\chi^2_{m-1}$ distribution. If $\chi^2_{\text{homogeneous}}$ is not significant, then a valid test for a common trend is possible by referring $\chi^2_{\text{trend}}$ to tables of the $\chi^2_1$ distribution. If $\chi^2_{\text{homogeneous}}$ is significant, then evaluation of $\chi^2_{\text{trend}}$ is not appropriate. Trend tests for each season may still be obtained from the individual $Z_j$ if desired.

In cases where the seasonal trends are different, it may be interesting to determine if there is a pattern to the heterogeneity. For example, one may want to know if the trends are similar for the winter months. To accomplish that, we use what is called in statistical literature a contrast. A contrast is defined by a linear expression of the form:

$$\sum_{j=1}^{m} c_j S_j = c_0$$

(4.41)

where the $c_j$, $j = 1, 2, ..., m$, are constants such that $\sum c_j = 0$. The variance of the contrast is given by:
\[ \text{Var}(\sum_{j=1}^{m} c_j S_j) = \sum_{j=1}^{m} c_j^2 \text{Var}(S_j) \] (4.42)

For example, to test whether the trend in December is different from the trends in January and February, one can define the contrast as \( S_{12} = 0.5*(S_1 + S_2) \). We then compare

\[
\frac{S_{12} - 0.5(S_1 + S_2)}{\text{Var}(S_{12}) + 0.25(\text{Var}(S_1) + \text{Var}(S_2))}
\] (4.43)

to the chi-squared distribution with 1 degree of freedom. If the contrast is suggested by the data, Fleiss (1981) recommends using two degrees of freedom. One can also use techniques such as the Newman-Keuls method (see Milliken and Johnson, 1984) to try to place the trends in homogeneous groups.

The presence of more than one value per season can be handled in the same way as in the seasonal Kendall test (see Section 4.2.2.3). Missing data can also be handled as they are in the seasonal Kendall test.

As mentioned earlier, the van Belle-Hughes test can be used where there are many sites. To test for trend, one must first test for homogeneity between the sites, secondly for homogeneity between seasons, and then for any possibility of interactions between sites and seasons. If we find that the trends are homogeneous across seasons and sites and no interactions are present, the test for the statistical significance for trend can be made. In these tests, it is assumed that the data are independent across seasons and sites. The reader is referred to van Belle and Hughes (1984) or Gilbert (1987) for more details.

### 4.2.3 Slope Estimation

In the preceding section, we have described some nonparametric tests for detecting monotonic trends. Once we have determined that a trend exists, we would like to be able to estimate that trend. For the two techniques presented here, it is assumed that \( f(t) \) in relation 4.1 is equal to

\[ f(t) = \alpha t + \beta \] (4.44)

Using these techniques, we will get nonparametric estimates of the coefficients of this linear relationship. The first estimator assumes that no seasonal cycle is present in the data, and the second allows for such cycles.

#### 4.2.3.1 Sen’s Slope Estimator

A consistent nonparametric estimator for the coefficients of a linear regression (relation 4.44) was proposed by Theil (1950). It is based on Kendall’s Tau (see Kendall, 1938, 1975) and was modified by Sen (1968b) to include the possibility of ties in the \( t_i \).

If one has \( n \) values for the pair \((t_i, C_i)\), the coefficients of the linear relationship (relation 4.44) can be estimated as follows:

1. The N values \( A_{[i]} = (C_i - C_j) / (t_i - t_j) \) for \( i = 1, 2, \ldots, n \) and \( j = 1, 2, \ldots, n \); 
   
   \( j > i \) and \( t_i \neq t_j \). The estimate of \( \alpha (a^5) \) is given by the median of the \( A_{[i]} \) defined as (after ordering the \( A_{[i]} \)):
\[
    a^s = \begin{cases} 
    A_{(N+1)/2} & \text{if } N \text{ is odd} \\
    \frac{1}{2} (A_{[N/2]} + A_{((N+2)/2)}) & \text{if } N \text{ is even}
    \end{cases}
\] (4.45)

2. To obtain an estimate of \( \beta \), we first calculate the \( n \) values \( C_i - a^s t_i \) (Theil, 1950). The median value gives an estimate of \( \beta(b^s) \).

A 100(1-\( \epsilon \))% two-sided confidence interval about the true slope for \( a^s \), which is valid if \( n \) is greater than 10 and if there are not many ties, can be obtained as follows (note that a nonparametric technique is given in Theil (1950) and Sen (1968b)):

1. Choose the desired confidence coefficient \( \epsilon \) and get \( Z_{1-\epsilon/2} \) where \( Z_p \) is the upper 100(1-\( p \))th percentile of the standard normal distribution.

2. Compute \( C_\epsilon = Z_{1-\epsilon/2} \sqrt{\text{Var}(S)} \), where \( \text{Var}(S) \) is computed using relations 4.5 or 4.20, depending on whether there are multiple observations for each time. Note that the latter relation must be modified to account for the fact that there is only one season in the present case.

3. Compute \( M_1 = (N - C_\epsilon) / 2 \) and \( M_2 = (N + C_\epsilon) / 2 \).

4. The lower and upper limits of the confidence interval are the \( M_i \)th largest and \( (M_i + 1) \)th largest of the \( N \) ordered \( A_{[ij]} \) respectively.

4.2.3.2 Seasonal Kendall’s Slope Estimator (Gilbert, 1987)

The Sen’s slope estimator can be extended to take into account the presence of seasonality as follows:

1. First, compute for each season \( k \) (\( k=1, 2, \ldots, m \)) as in the preceding section the \( N_k \) values \( A_{[ij]k} = (C_{ik} - C_{jk}) / (t_{ik} - t_{jk}) \), where \( C_{ik} \) is the concentration for year \( i \) and season \( k \).

2. Then rank the \( N = N_1 + N_2 + \ldots + N_m \) individual slope estimates and find their median. This median is the seasonal Kendall slope estimator.

Note that the slopes for the different seasons are assumed to be homogeneous. It is a straightforward exercise to extend this technique to cases in which there is more than one observation per season.

A 100(1-\( \epsilon \))% two-sided confidence interval about the true slope can be obtained as in the preceding section.

1. Choose the desired confidence coefficient \( \epsilon \) and get \( Z_{1-\epsilon/2} \), where \( Z_p \) is the upper 100(1-\( p \))th percentile of the standard normal distribution.

2. Compute \( C_\epsilon = Z_{1-\epsilon/2} \sqrt{\text{Var}(S')} \), where \( \text{Var}(S') \) is equal to \( \sum_{i=1}^{m} \text{Var}(S_i) \) and the \( S_i \) are computed as in the preceding section. Note that the data are assumed to be independent as in the seasonal Kendall’s test.

3. Compute \( M_1 = (N - C_\epsilon) / 2 \) and \( M_2 = (N + C_\epsilon) / 2 \).
4. The lower and upper limits of the confidence interval are the \( M_{th} \)th largest and \((M_2+1)th\) largest of the \( N \) ordered \( A_{[i]} \) respectively.

4.2.4 Discussion

The techniques described in this section all have the advantage that no assumption is made about the distribution of the residuals (relation 4.1), except that it does not change with time (i.e., the variance is constant). It is also assumed for most of these tests that the residuals are independent. As most observed data in atmospheric chemistry have seasonal cycles, the Mann-Kendall test is limited to annual data only.

Although not having to make assumptions about the shape of the distribution of the residuals is an advantage, the nonparametric tests still present many disadvantages. First, the trends have to be assumed monotonic for the tests or a straight line for the estimators. This limits their applicability, because in many cases these hypotheses cannot be fulfilled in atmospheric chemistry. We have also seen that, except for the Dietz-Killeen test, the seasonal tests all assume that the trends in the different seasons are homogeneous (i.e., the same) which is not necessarily true.

Therefore, the applicability of these nonparametric tests in atmospheric chemistry may be limited. However, they may be useful in some situations.

It should be noted that the nonparametric tests presented here are only a few of the many possible tests that can be used in special situations. An interesting example of the other tests that are available is the Jonckheere test (Jonckheere, 1954; Lehmann, 1975), which uses the Mann-Whitney statistic to test for step trends in data (see Hirsch and Gilroy, 1985; Blanchard \textit{et al.}, 1996).

To the knowledge of the author, only the seasonal Kendall test, of all the nonparametric tests presented here, has been used in atmospheric chemistry. Examples of its use can be found in Schertz and Hirsch (1985), Sisterton \textit{et al.} (1990), Pollack \textit{et al.} (1993), and Baier and Cohn (1993).

4.3 Smoothers

4.3.1 Introduction

We have seen in Section 2.8 how smoothers can be useful in the exploratory analysis of a time series. The problem in using smoothers to estimate trends is that, for most smoothing techniques, it is not possible to test the statistical significance of any estimated trend. Even for the few techniques for which confidence intervals for the estimated trend can be constructed, these intervals are difficult to estimate, valid only for large-samples, and subject to theoretical criticisms.

The basic conceptual model behind the smoothing techniques is the same as for the nonparametric techniques presented in the preceding section (relation 4.1) and is:

\[
C_i = f(t_i) + \varepsilon_i
\]

(4.46)

It is assumed that \( f(t) \) is a continuous function. The \( \varepsilon_i \)'s are unobserved random variables representing error terms. Unless otherwise specified, it is assumed that the error terms are
uncorrelated, with mean zero and constant variance, $\sigma^2$. The goal of the smoother techniques presented in this section is to obtain an estimate of $f(t)$ if $n$ values of $C_t$ are available.

In this overview, we will try to summarize some of the most popular smoothing techniques. Programs or functions for most of these techniques can be found in most statistical packages, such as SAS®, S-Plus®, and other similar programs. Where appropriate, we will discuss briefly the possibility of constructing confidence intervals for the estimated trends.

We will begin with two simple smoothing techniques, namely, (1) the running mean and (2) the running median. Then we will discuss more complex techniques like the loess smoother or regression, the kernel smoother or regression, and finally the spline smoother technique.

Basic to all these techniques, except the last, is the notion of span or bandwidth. Its definition can be given as follows. If we want to estimate $f(t)$ at time $t_0$, the span of the smoother, $h$, is defined so that only the observations in the interval $[t_0-h, t_0+h]$ are used in the calculation of $\hat{f}(t_0)$. We will denote the number of valid observations in that interval as $N_h(t_0)$. This is illustrated in Figure 4.2. All the data in the shaded area will be used in the calculation of the smoother estimate at $t_0$. Note that the span or bandwidth is sometimes specified as a fraction $\alpha (0 < \alpha < 1)$ of the total number of data items. In those cases, one first calculates a constant value for the number of data in the interval. This is equal to the integer part of $\alpha n$. Then, for each time $t_0$, a value of $h$ is found such that $N_h(t_0)$ is equal to that number. Note that if there are missing data and/or the data are not sampled at equal intervals, $h$ will be a function of $t_0$.

In most of these techniques, a problem occurs at both ends of the time series. This is called the border or edge effect. If we define $t_{\text{min}}$ and $t_{\text{max}}$ as the lowest and highest values for $t$, then the border effect occurs when $t_0 < t_{\text{min}} + h$ and $t_0 > t_{\text{max}} - h$. As $t_0$ gets closer to $t_{\text{min}}$ or $t_{\text{max}}$, the number of data available, $N_h(t_0)$, for the calculation will decrease and the uncertainties in the estimate, $\hat{f}(t_0)$, will increase. For most of the estimators there is no way to compensate for that problem. Where such a technique does exist, we will mention it. In most cases, the best solution is either not to estimate $f(t)$ in those regions or to indicate that large uncertainties exist in them.

4.3.2 Running Mean Smoother

The simplest of the smoothing techniques is the running mean, also called the window estimate or the moving average. The running mean estimate of $f(t)$ at $t_0$ is:
\[ \hat{f}_h(t_0) = \frac{1}{N_h(t_0)} \sum_{i \text{ such that } |t_i - t_0| \leq h} C_i \]

(4.47)

In other words, the estimate of \( \hat{f}(t) \) is the mean value of the point within the interval \([t_0-h, t_0+h]\). The smoothness of \( \hat{f}_h(t) \) depends on \( h \). Note that the choice of \( h \) must be large enough to keep \( N_h(t_0) \) from becoming too small. The choice of \( h \) is critical to the result obtained. This is illustrated in Figure 4.3, where different values of \( h \) are used to estimate the long-term trend of ambient SO\(_2\) air concentrations at Algoma. For \( h \) less than a year, the smother followed the seasonal cycles quite well. At \( h = 1 \) year, the seasonal cycles are smoothed out and one starts to detect the long-term trend. For \( h \) equal 2 and 3 years, the estimates of the long-term trend are very similar and they are smoother than for \( h = 1 \) year. One can conclude that in this case, a value of \( h \) equal to about 2 years should be the right value. Note that the border or edge effects seems to be small in this case.

Missing data do not present any problem, as we have only to ignore them. If they represent a large fraction of the total number of data, they may influence the selection of the span and therefore the smoothness of \( \hat{f}_h(t) \).

The running mean estimator, like the mean, is very sensitive to the presence of outliers. One technique that can be used to "robustify" it is to employ methods like the \( \alpha \)-trimmed estimator or the \( \alpha \)-Winzorized estimator of the means (see David, 1981; Barnett and Lewis, 1984). If \( n \) is the number of observations (= \( N_h(t_0) \) in the present context), let \( r \) equal the integer part of \( \alpha n \); then the \( \alpha \)-trimmed means is equal to:

Figure 4.3 Running mean smoother trend estimates for four different bandwidths for ambient SO\(_2\) air concentration at Algoma, Ontario.
\[ T_n(r) = \frac{1}{n-2r} \sum_{i=r+1}^{n-r} C[i] \quad (4.48) \]

where \( C[i] \) denotes the observations sorted in increasing order. The \( \alpha \)-Winzorized mean is defined as:

\[ W_n(r) = \frac{1}{n} \left( (r+1)(C_{r+1} + C_{n-r}) + \sum_{i=r+2}^{n-r-1} C[i] \right) \quad (4.49) \]

for \( 0 < r < \frac{1}{2}(n-1) \) and

\[ W_n(r) = C_{\left\lfloor \frac{(n+1)/2} \right\rfloor} \quad (4.50) \]

if \( n \) is odd and \( r = \frac{1}{2}(n-1) \). Although these equations are for symmetric \( \alpha \)-trimmed and \( \alpha \)-Winzorized mean estimators, they can be modified for trimming or Winzorizing only large values, which would be more appropriate in some situations (see Barnett and Lewis, 1984). Other robust estimators for the mean include L-estimator, M-estimator, and R-estimator (see Huber, 1981; Barnett and Lewis, 1984; Hampel et al., 1986).

Outliers may be due to many causes. They may only be apparent in the sense that the distribution of the \( \varepsilon \)'s in 4.46 may not be symmetrical but skewed to the left as in the log-normal or gamma distributions. In these cases, a simple transformation (see Section 2.9) before the use of the smoother may eliminate these outliers. The reader should keep in mind that use of the running mean is not recommended if the distribution of the residual is very skewed.

Border effects are very important in this technique and cannot be reduced.

Finally, there is no technique to fix the span or bandwidth. We are thus reduced to trying many values and selecting the value that seems to give a smooth curve without eliminating all variations.

**Figure 4.4**: Running median smoother trend estimates for four different bandwidths for ambient \( \text{SO}_2 \) air concentration at Algoma, Ontario.
4.3.3 Running Median Smoother

This technique is very similar to the preceding method. The only difference is that the median of the data in the window defined by the span is used in lieu of the mean. Therefore, we can define the estimate of \( f(t) \) at \( t_0 \) as:

\[
\hat{f}_h(t_0) = \text{med}_{\text{med of the observations } i \text{ such as } |t_i - t_0| \leq h} (C_i)
\] (4.51)

As in the case of the running mean smoother, the smoothness of \( \hat{f}_h(t) \) is a function of \( h \). This is illustrated in Figure 4.4. If one compares these results with those for the running mean smoother, it will be noticed that they are quite similar, especially with \( h \geq 1 \) year. Similar results cannot be expected in all cases, however, especially if the distribution of the data is skewed toward high values. As with the running mean smoother, a bandwidth of around 2 years seems to be right.

Unlike the running mean, however, the running median is quite resistant to the presence of outliers and can be used for skewed distributions. Missing values can be ignored, but they may influence the selection of the span if they represent a large fraction of the data.

Border effects are also important for this smoother, as with the running mean the selection of the span is arbitrary. Therefore, the user has to try many values and select one that gives a good compromise between smoothness and detail.

The great disadvantage of the running median compared with the running mean is that it is very computer calculation intensive, because for each value of \( t_0 \), the \( N_h(t_0) \) data have to be sorted in order before calculation of the median. This may take a long time if \( \hat{f}_h(t) \) is estimated at many values of \( t_0 \). Friedman and Stuetzle (1982) have presented an algorithm that improves the speed of running median smoothing.

4.3.4 Robust Locally Weighted Regression (Loess) Smoother

The idea of estimating \( \hat{f}_h(t_0) \) by the mean of the data in the interval \([t_0-h, t_0+h] \) can be extended by using the estimated value at \( t_0 \) of a polynomial of order \( d \) fitted to the data (within the interval) by least squares. This estimator can be written as:

\[
\hat{f}_h(t_0) = \beta_0 + \beta_1 t_0 + \beta_2 t_0^2 + \cdots + \beta_d t_0^d
\] (4.52)

where the coefficient \( \beta_i \) denotes the coefficients that minimize:

\[
\sum_{i=1}^{N_h(t_0)} \left( C(t_i) - \beta_0 - \beta_1 t_i - \beta_2 t_i^2 - \cdots - \beta_d t_i^d \right)^2
\] (4.53)

where the summation is on all the observations in the interval \([t_0-h, t_0+h] \). Note that the conceptual model 4.46 is also valid in this case. A better estimate can be obtained by using a weighted least squares (Cleveland, 1979) defined as:

\[
\sum_{i=1}^{N_h(t_0)} w_{t_0}(t_i) \left( C(t_i) - \beta_0 - \beta_1 t_i - \beta_2 t_i^2 - \cdots - \beta_d t_i^d \right)^2
\] (4.54)

where
\[ w_i(t_i) = W\left( \frac{t_i - t_0}{h} \right) \]  

(4.55)

where \( W(x) \) is a function such that

1. \( W(x) > 0 \) for \( |x| < 1 \)
2. \( W(-x) = W(x) \)
3. \( W(x) \) is a nonincreasing function for \( x \geq 0 \)
4. \( W(x) = 0 \) for \( x \geq 1 \)  

(4.56)

An example of \( W(x) \) is the tricube function given by:

\[ W(x) = \begin{cases} (1 - |x|^3)^3 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \geq 1 \end{cases} \]  

(4.57)

One problem with this smoother is that it is not resistant to the presence of outliers. It can be “robustified” in the following way (Cleveland, 1979).

Let's

\[ e_i = C(t_i) - \hat{f}_h(t_i) \]  

(4.58)

for \( i = 1, 2, \ldots, n \). Let \( s \) be the median of the \( |e_i| \). Define robust weights by:

\[ \delta_i = B(e_i / 6s) \]  

(4.59)

where \( B(.) \) is the bisquare weight function that is defined by:

\[ B(x) = \begin{cases} (1 - |x|^2)^2 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \geq 1 \end{cases} \]  

(4.60)

The next step in the analysis is to recalculate the coefficients of the weighted regression (relation 4.54) by replacing \( w_i(t_i) \) by \( \delta_i w_i(t_i) \) and then deriving a new estimate of \( \hat{f}_h(t_i) \) using relation 4.52. New values for \( \delta_i \) can be obtained and the process can be iterated \( m \) times.

The use of robust locally weighted regression smoothing techniques can be summarized in the following algorithm:

1. Put \( \delta_i = 1 \) for \( i = 1, 2, \ldots, n \).
2. For \( i = 1, 2, \ldots, n \), define the interval \([t_i-h,t_i+h]\) around \( t_i \) where \( h \) is the span.
   2.1 For each observation \( j \) in the interval \([t_i-h,t_i+h]\), calculate \( w_i(t_j) \) using relation 4.52.
   2.2 Estimate the correlation coefficient \( \beta_k \) by weighted least squares by minimizing the relation

\[
\sum_{j=1}^{N_i(t_i)} \delta_i w_i(t_j) \left( C(t_j) - \sum_{k=0}^{d} \beta_k t_i^k \right)^2
\]
2.3 Estimate \( \hat{f}_n(t_i) \) using relation 4.52.

3. Estimate the new \( \delta_i \) using relations 4.58 to 4.60.

4. Loop through steps 2 and 3 \( m_t \) times.

Four parameters, namely \( h \), \( d \), \( W \), and \( m_t \) have to be selected before using that algorithm. As with the other smoothers mentioned so far, the choice of \( h \) is somewhat arbitrary. Many values may have to be tried before an acceptable value can be found. For \( d \), Cleveland (1979) recommends 1, because for \( d \geq 2 \), the calculations get more complex and the need for computational convenience overrides the need for flexibility. \( W(x) \) can be any function that satisfies relation 4.56. In addition, it is desirable for it to decrease smoothly to 0 as \( x \) goes from 0 to 1. Finally, \( m_t \) could be selected indirectly by iterating until some convergence criterion is satisfied. On the basis of some experimentation, Cleveland (1979) concluded that this was unnecessary and that two or three iterations should be sufficient.

Examples of the use of loess smoothing are presented in Figure 4.5. As with the two preceding smoothers, the smoothness of the estimated trend improves as the bandwidth increases. Depending on the smoothness of the estimate, one should use a larger bandwidth (about 3 years) for this smoother than for the preceding two.

As with the running mean and running median smoothers, missing values can be ignored. As in the case of the running mean, it may also be useful to transform the data so that the residuals are symmetric around the assumed function \( f(t) \). Even though the loess smoother gives some protections against outliers, prior transformation of the data should give better results in situations where the distribution of the residuals is skewed toward high values.

![Figure 4.5: Loess smoother trend estimation for four different bandwidths for ambient \( \text{SO}_2 \) air concentration at Algoma, Ontario.](image)
An interesting property of the loess smoother is that the border effects are strongly reduced (see Hart, 1997). This is due to the fact that it implicitly includes the normalization that reduces those effects. They are, however, not totally eliminated as can be seen in Figure 4.5.

4.3.5 Kernel Smoothers

Kernel smoothers are a somewhat more sophisticated version of the running mean smoother. In these smoothers the mean value is replaced by a weighted mean. Usually, the weight is increased as the data points move closer to \( t_0 \). Several different versions of kernel smoothers exist. We will mention the three most important. The first is the Nadaraya-Watson estimator (Nadaraya, 1964 and Watson, 1964). It is defined as:

\[
\hat{f}_h^{NW}(t_0) = \frac{\sum_{i=1}^{n} C(t_i)K\left(\frac{t_i - t_0}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{t_i - t_0}{h}\right)}
\]

(4.61)

where \( K(.) \) is a function called, the kernel, and \( h \) is the span or bandwidth or smoothing parameter. We will discuss possible kernel functions later in this section. It is interesting to note that if we define the rectangular kernel, \( K_R \), as

\[
K_R(u) = \frac{1}{2} I_{(-1,1)}(u)
\]

(4.62)

where \( I_A \) denotes the indicator function for the set \( A \), i.e.,

\[
I_A(x) = \begin{cases} 
1 & x \in A \\
0 & x \notin A
\end{cases}
\]

(4.63)

\( \hat{f}_h^{NW}(t_0) \) is equal to the running mean.

The second type of kernel smoother is the Priestley and Chao (1972) estimator defined as:

\[
\hat{f}_h^{PC}(t_0) = \frac{1}{h} \sum_{i=1}^{n} (t_i - t_{i-1})C(t_i)K\left(\frac{t_i - t_0}{h}\right)
\]

(4.64)

The third type of kernel smoother is the Gasser-Müller (1979) estimator, defined as

\[
\hat{f}_h^{GM}(t_0) = \frac{1}{h} \sum_{i=1}^{n} C(t_i) \int_{s_{i-1}}^{s_i} K\left(\frac{t_0 - u}{h}\right) du
\]

(4.65)

where \( s_0 = 0, s_i = (t_i + t_{i+1})/2, i = 1, 2, \ldots, n-1, \) and \( s_n = 1 \). Another way of writing this estimator is:

\[
\hat{f}_h^{GM}(t_0) = \frac{1}{h} \int_{0}^{1} C_n(u)K\left(\frac{t_0 - u}{h}\right) du
\]

(4.66)

where \( C_n(.) \) is the piecewise constant function:
Table 4.1 Some of the most often used kernels.

<table>
<thead>
<tr>
<th>Kernel Type</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Rectangle</td>
<td>$K_{R}(u) = \frac{1}{2} I_{(-1,1)}(u)$</td>
</tr>
<tr>
<td>b) Triangular</td>
<td>$K_{T}(u) = (1-</td>
</tr>
<tr>
<td>c) Epanechnikov</td>
<td>$K_{E}(u) = \frac{3}{4} (1-u^2) I_{(-1,1)}(u)$</td>
</tr>
<tr>
<td>d) Quartic</td>
<td>$K_{Q}(u) = \frac{15}{16} (1-u^2)^2 I_{(-1,1)}(u)$</td>
</tr>
<tr>
<td>e) Rogosinski</td>
<td>$K_{R_{0}}(u) = (0.5 + \cos(\pi/5) \cos(\pi u) + \cos(2\pi/5) \cos(2\pi u)) I_{(-1,1)}(u)$</td>
</tr>
<tr>
<td>f) Gaussian</td>
<td>$K_{G}(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$</td>
</tr>
</tbody>
</table>

where

$A(x) = \begin{cases} 1 & x \in A \\ 0 & x \not\in A \end{cases}$

\[ C_n(u) = \sum_{i=1}^{n} C(t_i) I_{[x_i-x_j]}(u) \quad (4.67) \]

One should note the similarity of types between the last two smoothers. An interesting characteristic of the Gasser-Müller estimator is that it tends to the function $C_{n}(.)$ when the bandwidth tends to zero. By contrast, the Nadaraya-Watson estimate is not well defined for a sufficiently small $h$ when $K$ has finite support. As a result the latter is more unstable than the former for small values of $h$.

Chu and Marron (1991) have shown that when the observations are approximately evenly spaced, there is very little differences between the types of estimators. However, there may be great differences when the observations are not equally spaced.

One usually asks that the kernel function $K(.)$ be symmetric and have a unique maximum at 0. Another condition, that one wants to put on the kernel function, is that $\int K(u) du = 1$. A popular way of ensuring that these three conditions are satisfied is to take $K$ to be a probability density function (pdf) that is unimodal and symmetric about zero. Doing so also guarantees a positive regression estimate for positive data. This is an important property when it is known that $f(t) > 0$. However, there are some useful kernels that take negative values.

Table 4.1 lists some of the most often used kernel functions. The Rogosinski kernel differs from the others in that it can take negative values. The Gaussian Kernel is defined for
u ∈ (-∞, ∞). Kernels are typically designed either to reduce the variance or to optimize some property of the smoothers. More kernel functions are given in Gasser et al. (1985).

It is important to note that the three types of kernel smoothers presented here are biased in the sense that as n → ∞, h → 0 and nh → ∞, and \( \hat{f}_h(t) \) does not → f(t). The bias depends on many factors and is asymptotically proportional (see Jennen-Steinmetz and Gasser, 1988; Gasser and Engel, 1990) to \( h^2f''(t) \) for the Gasser-Müller smoother where \( f''(t) \) is the second derivative of f(t). In the case of the Nadaraya-Watson smoother, a second term of the same order of magnitude is added to the preceding term. This second term is proportional to \( h^2f''(t) \), where \( f''(t) \) is the first derivative of f(t), and to the distribution (i.e., design) of the observations, thus making the bias very complex. From these results, it might seem that one should make the bandwidth as small as possible to reduce the bias. However in reducing the bandwidth, the variance of the estimator is increased because the estimator variance for both types of smoothers is proportional to \((nh)^{-1}\) and to the variance of the data around f(t) and depends also on the observation design (see Jennen-Steinmetz and Gasser, 1988; Gasser and Engel, 1990). Therefore, the choice of bandwidth involves a compromise between decreasing the bias and keeping the variance of the estimator small. Note that for both the bias and the variance some of the constant terms entering in these relationships depend on factors like \( \int u^2 K(u) du \) and \( \int \{K(u)\}^2 du \) and thus on the kernel used. These relations could therefore be used in trying to develop kernels that would give the best compromise between bias and variance for a given value of the bandwidth (see Hart, 1997).

The fact that the bias for the Gasser-Müller smoother is independent of the distribution of the design should make it more useful than the Nadaraya-Watson smoother. However, in time series analysis the distribution of the observations, which is usually regular, should not have much influence on the bias of the latter method.

The most important facts to retain from this discussion are:

- Because the bias is proportional to \( f''(t) \), it will be highest when \( f(t) \) presents well-marked peaks or troughs. This can be compensated for by decreasing the bandwidth, but that will result in an increase in the variance of the smoother, rendering it more unstable. The effect of the bias on the shape of \( f(t) \) can be seen in Figure 4.2. When \( t_0 \) is near the peak in \( f(t) \) and the bandwidth is large, data at the peak and far from the peak will both be used in the calculation of \( \hat{f}_h(t_0) \), thus reducing its value. This problem can be corrected by reducing the bandwidth and thus increasing the variability of the estimator. It is interesting to note that the bias would be negative for a peak in \( f(t) \) and positive for a trough.
- The variance of the smoother is decreased when the amount of data increases, but the bias is not affected.
- The higher the variability of the data around \( f(t) \), the higher the variance of the smoother.
- Both bias and variance are dependent on the kernel used, although this effect is not usually important.

From the first point, we may conclude that when \( f(t) \) presents marked peaks and troughs it may be advantageous not to use a constant bandwidth but to vary it according to on the shape of \( f(t) \). We will discuss this point later in Section 4.3.8.
Figure 4.6 presents an example of the kernel smoother using the Nadaraya-Watson estimator. The same data as in the preceding examples were used. One should note that as for the running mean, running median, and loess smoothers, the smoothness of the estimation increases with the bandwidth. That means that, as we have seen, the bias will also increase, although $f^{(2)}(t)$ will always be small in the present example and will thus limit the bias. The increase in the bandwidth will reduce the variance of the estimator. In the present case a bandwidth of about 2 years should be appropriate.

Like the smoothers described in the preceding sections, the kernel smoother has edge effects. The variance of the smoother increases as to the number of points available decreases as one gets to the ends of the time series. Techniques have been developed to correct in part for these edge effects. We will discuss those for the Nadaraya-Watson and Gasser-Müller smoothers.

A correction for edge effects in the Nadaraya-Watson smoother was proposed by Rice (1984). The modification proposed by Rice is based on the fact that near the borders or edges of the time series the bias is proportional to $h$ and not to $h^2$ as we have seen before. To obtain an estimator that has a bias of the same proportionality relative to the bandwidth, one should subtract the term proportional to $h$ from the estimate $\hat{f}_h^{NW}(t_0)$ in 4.61. This new estimate can be written as:

$$\hat{f}_h^{NW}(t_0) = \hat{f}_h^{NW}(t_0) + \beta [\hat{f}_h^{NW}(t_0) - \hat{f}_h^{NW}(t_0)]$$  (4.68)

where
\[ \beta(\rho) = \frac{R(\rho)}{\alpha R(\rho/\alpha) - R(\rho)} \]  \hspace{1cm} (4.69)

with \( R(u) = K_i(u)/K_0(u) \) and

\[ K_i(u) = \int_{-1}^{u} v^i K(v) \, dv \]  \hspace{1cm} for \( i = 0, 1, 2 \).

In these equations, \( \rho \) is equal to \( t_0/h \). For the parameter \( \alpha \), Rice (1984), after discussing different possibilities, recommends making it equal to \( 2 - \rho \).

Relation 4.68 is applied only in the region \( t_0 \in [0, h] \) and \( [t_{\max} - h, t_{\max}] \). For other values of \( t_0 \), relation 4.61 is used. An example of the use of the modified Nadaraya-Watson smoother is presented in Figure 4.7. In that example, the differences between the modified and unmodified curves are small, but this is not always the case.

In the case of the Gasser-Müller smoother, Gasser and Müller (1979) proposed the use of boundary kernels. One way of constructing such kernels is to define \( K_0 \) for each \( \rho (= t/h) \in [0, 1) \) by

\[ K_\rho(u) = t_\rho(u) K(u) I_{(-1,0)}(u) \]  \hspace{1cm} (4.70)

where \( t_\rho \) is a function with the properties

\[ \int_{-1}^{t_\rho(u)} K(u) \, du = 1 \]

and

\[ \int_{-1}^{t_\rho(u)} u K(u) \, du = 0 \]  \hspace{1cm} (4.71)

At \( t_0 = \rho h \), the \( \hat{f}^{GM}(t_0) \) is estimated by

\[ \frac{1}{h} \sum_{i=1}^{n} C(t_i) \int_{-1}^{t_i} K_\rho \left( \frac{t_i - u}{h} \right) \, du \]  \hspace{1cm} (4.72)

A similar boundary adjustment can be done for \( t_0 \) near \( t_{\max} \) by simply writing \( \rho = (t_{\max} - t_0)/h \).

Figure 4.7 Modified kernel smoother trend estimates for four different band-widths for ambient SO\(_2\) air concentration at Algoma, Ontario.
A simple choice for \( t_\rho \) is

\[
t_\rho(u) = a_\rho + b_\rho u. \tag{4.73}
\]

To ensure that relations 4.71 are satisfied, one takes:

\[
a_\rho = \frac{I_{\rho,2}}{I_{\rho,2} I_{\rho,0} - I_{\rho,1}^2} \tag{4.74}
\]

and

\[
b_\rho = \frac{-I_{\rho,1}}{I_{\rho,2} I_{\rho,0} - I_{\rho,1}^2} \tag{4.75}
\]

where

\[
I_{\rho,i}(u) = \int_{-1}^{u} v^i K(v) \, dv \tag{4.76}
\]

for \( i = 0, 1, 2 \).

Another possible boundary kernel is (Müller, 1991):

\[
K_\rho(u) = 6(1 + u)(\rho - u) \frac{1}{(1 + \rho)^2} \times \left( 1 + 5 \left( \frac{1 - \rho}{1 + \rho} \right)^2 + 10 \frac{1 - \rho}{(1 + \rho)^2} u \right) I_{(-1,\rho)}(u) \tag{4.77}
\]

At \( \rho = 1 \), \( K_\rho(u) \) becomes simply the Epanechnikov kernel. To ensure a smooth transition, it would thus been sensible to use that kernel in the edge regions and the Epanechnikov kernel for the interior points.

When making estimates in statistical analysis, one would like to be able to create confidence intervals for the parameter estimated. For the estimation of \( f(t) \) by kernel smoothers, some large-sample confidence intervals have been proposed, either for pointwise confidence intervals or for simultaneous confidence bands for the entire function \( \hat{f}_h(t_0) \).

An asymptotically valid \( (1 - \alpha)100\% \) confidence interval for the Gasser-Müller smoother, \( \hat{f}_h^{GM}(t_0) \), is given by Hart (1997):

\[
\left( \hat{f}_h^{GM}(t_0) - (Z_{a/2} + \hat{B}_{C,n}) \sqrt{\sum_{i=1}^{n} \left( \frac{s_i}{s_{i-1}} K \left( \frac{t_0 - u}{h} \right) \, du \right)^2} \right),
\]

\[
\hat{f}_h^{GM}(t_0) - (Z_{a/2} + \hat{B}_{C,n}) \sqrt{\sum_{i=1}^{n} \left( \frac{s_i}{s_{i-1}} K \left( \frac{t_0 - u}{h} \right) \, du \right)^2} \right)^{1/2}, \tag{4.78}
\]

where \( \hat{\sigma} \) is an estimate of the variance of the residuals and

\[
\hat{B}_{C,n} = \frac{C^4 n^{-3/5} \sigma_K^2}{2 \sum_{i=1}^{n} \left( \frac{s_i}{s_{i-1}} K \left( \frac{n^{1/5} (t_0 - u)}{C} \right) \, du \right)^2} \hat{f}^{(2)}(t_0) \hat{\sigma}^{1/2}, \tag{4.79}
\]
where \( C \) is a constant that depends on the distribution of the observations, 
\[ \sigma_k^2 = \int u^2 K(u) \, du, \] 
and \( \hat{f}^{(2)}(t_0) \) is an estimate of the second derivative of \( f(t) \), which can be estimated using the following kernel estimator for the \( k \)th derivatives:
\[
\hat{f}^{GM}_{k, h}(t_0) = \frac{1}{h^{k+1}} \sum_{i=1}^{n} C(t_i) \int_{s_{i-1}}^{s_i} M\left(\frac{t_0 - u}{h}\right) \, du
\]
(4.80)

where \( M \) is a kernel with support \((-1, 1)\). This kernel is fundamentally different from the ones used in estimating \( \hat{f}^{GM}(t_0) \) (see Hart, 1997). An example of a valid kernel \( M \) for the second derivative is:
\[
M(u) = \frac{105}{16} (-5u^4 + 6u^2 - 1) I_{(-1,1)}(u)
\]
(4.81)

See Gasser et al. (1985) for more details about these types of kernels.

This technique has been criticized because one has to choose a completely new set of parameters for a smoother to estimate the uncertainty in the smoothed regression. Details for a bootstrap approach to obtaining a bias-adjusted interval as in relation 4.78 are provided by Härdle and Bowman (1988). One should note that this technique is complex and only gives asymptotically valid confidence intervals.

A number of methods have been proposed for constructing simultaneous confidence bands for the entire function \( f(t) \). As in the case of pointwise confidence intervals, one has to take into account the bias of the nonparametric smoother in order to guarantee the validity of the corresponding interval(s). Each of the techniques that has been proposed provides a means of dealing with this issue. The reader is referred to the following papers for more details: Knafl et al. (1985), Hall and Titterington (1988), Härdle and Bowman (1988), Li (1989), Härdle and Marron (1991), and finally Eubank and Speckman (1993).

Although we have mainly concentrated on the Gasser and Müller technique, corresponding relations can be derived for the Nadaraya-Watson and other kernel smoothers.

An interesting application of kernel smoothers is related to the use of probability bands to test the adequacy of a parametric model. For example, if one wants to test the adequacy of the straight line model for the regression of \( f(t) \), one can use the residual of a kernel smoother like the Gasser-Müller smoother to construct simulated confidence bands that can then be used to test the validity of the straight line model. See Hart (1997) and Hart and Wehrly (1992) for more details.

### 4.3.6 Smoothing Splines

A spline is a piecewise polynomial constructed in such a way that it is smooth at the observations, called knots, at which two polynomials are pieced together. To impose the smoothness criterion, one usually asks that the polynomials of order \( k \) be used and that their first \( k-1 \) derivatives be continuous. Most commonly, \( k \) equals 3. In that case, we speak of cubic splines. For example, a cubic spline is defined as:
\[
y = S(x) = P_j(x) = a_j + b_j x + c_j x^2 + d_j x^3
\]
(4.82)

with the conditions:
\[ \xi_{j-1} \leq x < \xi_j ; \quad (\xi_0 = -\infty ; \xi_{n+1} = \infty) \tag{4.83} \]

where the \( \xi_j \ (j = 1, 2, \ldots, n) \) are the knots, and
\[ P_j^{(k)}(\xi_j) = P_{j+1}^{(k)}(\xi_j) ; \quad k = 0, 1, 2; \ j = 1, 2, \ldots, n \tag{4.84} \]

where \( P^{(k)} \) is the kth derivative of \( P(x) \).

Splines can be used to approximate virtually any smooth function, at least if a sufficiently large number of knots is used.

As defined in the preceding paragraphs, the spline will pass through all the observations and therefore follow all the variations in the data. However, we would like to filter out the rapid variations to obtain a smooth estimate of \( \hat{f}(t) \), in other words, a smooth spline. A smooth spline, \( \hat{f}_\lambda(t) \), has the following properties:

1. It has knots at \( t_1, t_2, \ldots, t_n \).
2. It is a cubic polynomial on each interval \([t_i, t_{i+1}] \), \( i = 2, \ldots, n \).
3. It has two continuous derivatives.
4. Finally, it minimizes the function:
\[ E_\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} (C(t_i) - f(t_i))^2 + \lambda \int \left[f^{(2)}(u)\right]^2 du \tag{4.85} \]

where \( \lambda \) is a positive constant between 0 and \( \infty \).

In relation 4.85, the first term assures fidelity to the data in a least squares sense, and the second term is a roughness penalty imposed so that, for example, the curve \( \hat{f}_\lambda(t) \) will not have too many bends. The parameter \( \lambda \) adjusts the relative weighting given to the error sum of squares and the roughness penalty: a small \( \lambda \) means that fit is more important, whereas a large \( \lambda \) means that smoothness of function is more important than fit. It therefore controls the degree of smoothing.

For \( \lambda = 0 \), \( \hat{f}_\lambda(t) \) is the (unique) minimizer of \( \int \left[f^{(2)}(u)\right]^2 du \), subject to the constraint that \( f(t_i) = C_i, \ i = 1, 2, \ldots, n \). This spline is a so-called natural spline interpolant of \( C_1, C_2, \ldots, C_n \). This is the cubic spline mentioned earlier. When \( \lambda \to \infty \), \( \hat{f}_\lambda(t) \) becomes the least squares straight line fit to the data.

Note that there is a relationship between smoothing splines and kernel smoothers. Silverman (1984) showed that, asymptotically, minimizing relation 4.85 cause it to behaves like a variable-bandwidth kernel smoothing method.

The use of smoothing splines is illustrated in Figure 4.8 for four values of \( \lambda \). For small values of \( \lambda \) (Figure 4.8a), the smooth curve follows the seasonal cycles. As \( \lambda \) increases, these variations are filtered out and the smooth curve increasingly shows the long-term trend. However, as mentioned earlier, if \( \lambda \) becomes very large, the smooth curve becomes equal to the least squares straight-line fit to the data as illustrated in Figure 4.8d.

This example illustrates the importance of the selection of the smoothing parameter \( \lambda \). Different techniques have been proposed to help in the choice of \( \lambda \). The first is the use of
cross-validation. We will discuss the general principles of this technique later (Section 4.3.8); the reader is referred to that section for more details. Craven and Wahba (1979) have introduced a small variation on cross-validation, called generalized cross-validation, which minimizes the average squared error at the observation points. This technique is a very popular choice for selecting the smoothing parameter, and is used by the ISML package. Silverman (1985) has introduced an approximation to that technique which involves much less computation.

As mentioned earlier, one likes to be able to draw inferences (e.g., create confidence bands) in statistical data analysis. However, inferences about \( f(t) \) from relations 4.1 and 4.85 are not possible because relation 4.85 is a data-smoothing criterion; it does not define a statistical model. Therefore, this relation has to be rewritten to become a statistical model. This can be done using the Bayesian approach, in which we assume a prior distribution for \( f(t) \) and estimate a \( f(t) \) given that prior curve and the observed data. Wahba (1978, 1983) has shown that under certain conditions and fixed \( \lambda \), the posterior mean of \( f(t) \) is the smoothing spline that minimizes relation 4.85. An interval estimate can be obtained using the posterior variance. Silverman (1985), however, has put forward some qualifications regarding the use of that technique. Other techniques have been proposed by Wecker and Ansley (1983) and Silverman (1985).

As for the other smoothing techniques presented in this section, the missing data are simply ignored in the calculation.

In the case of time series, the possibility of serially correlated errors (i.e., the \( \epsilon_i \) values are not independent) may be strong.

Figure 4.8 Smoothing spline trend estimates for four different values of \( \lambda \) for ambient \( SO_2 \) air concentration at Chalk River, Ontario
Several ideas for extensions in this direction are given by Silverman (1985). Kohn and Ansley (1987) have also proposed a new algorithm for spline smoothing that deals with the possibility of serially correlated errors.

The reader who would like to investigate smoothing splines in more detail should refer to Wold (1974), de Boor (1978), Wegman and Wright (1983), Eubank (1984, 1988), and Silverman (1985). Computer algorithms to fit smoothing splines can be found in Herriot and Reinsch (1973) and Duris (1980).

4.3.7 Periodic Smoothers

As mentioned earlier, one difficulty in applying smoothing techniques is the problem of edge or border effects. Some corrective procedures have been mentioned for some of the smoothing techniques. A special case in which these effects can be easily corrected arises when $f(t)$ in relation 4.1 is assumed to be periodic, with a period equal to $t_p = t_{\text{max}} - t_{\text{min}}$. Mathematically this can be written as:

$$f(t + t_p) = f(t) \quad (4.86)$$

Such a data set might occur, for example, if observations were made over the course of a year and one wanted to estimate the seasonal cycle. A periodic data set can sometimes be created by folding the time series into a new series that extends from zero to $t_p$. An example of such a process is the folding of many years of observations into one annual period. To do that, there should be no long-term trend in the data, and the seasonal cycles should be the same for every years. Another example of such a folding would be the folding of hourly data for many days or weeks into a one-day period to study the diurnal cycles.

If $h$ is the bandwidth of a running mean, running median, loess, or kernel smoother, a periodic smoother can be created by first constructing a new data time series, $C^*(t)$, for $t \in [-h, t_p + h]$, as follows:

1. For $t \in [-h, 0)$, $C^*(t) = C(t + t_p)$.
2. For $t \in [0, t_p]$, $C^*(t) = C(t)$.

Figure 4.9: Creation of the new time series for periodic smoothing.

Figure 4.10 Seasonal cycles for detrended ambient SO$_2$ air concentrations at Chalk River, Ontario.
3. For \( t \in (t_p, t_p+h] \), \( C^*(t) = C(t - t_p) \).

Then, these data are analyzed for \( t \in [0, t_p] \). This insures that \( \hat{f}(t_0) \) will be equal to \( \hat{f}(0) \). This process is illustrated in Figure 4.9. The shaded areas are the regions where the data are extended. Note that the normal smoothers are used in periodic smoothing. As an example, Figure 4.10 presents the seasonal cycles of ambient SO\(_2\) air concentrations at Chalk River, Ontario, using a periodic kernel Nadaraya-Watson smoother. Note that the long-term trend has been eliminated using a kernel Nadaraya-Watson smoother.

4.3.8 Selecting the Smoothing Parameters.

We mentioned earlier that the result of smoothing techniques are highly dependent on the bandwidth in the case of running mean, running median, loess, and kernel smoothers and on the smoothing constant in the case of smoothing splines. The method used to select these parameters thus far has been very subjective, in the sense that several values of the bandwidth or the smoothing constant were used and the one giving the “best” result in some undefined sense was selected. If one wants only to explore the data, that technique should be enough. In other cases, the characteristics of the time series can help in fixing the bandwidth. For example, if one wants to estimate the long-term trends of a multiyear time series, a bandwidth of 3 to 4 years should filter out the seasonal cycle and any long-term cycles with periods between about 2 and 5 years. In other circumstances, one would like to have a less subjective method of fixing the bandwidth or the smoothing constant. We also mentioned earlier that for the running mean, running median, loess, and kernel smoothers, one would like to have a variable span or bandwidth, especially if there appear to be large peaks and troughs in \( f(t) \). That can be handled only if a non-subjective way of selecting the span or bandwidth is available. We will discuss next the two most important methods of doing this. The reader is referred to Hart (1997) and to the references given therein for other methods as well as more details on the methods discussed here.

4.3.8.1 Cross-validation

Let \( S \) designate any smoothing parameter. \( S \) can stand for the bandwidth or the smoothing parameter or any other parameter that has to be fixed. The cross-validation criterion \( CV(S) \) is defined as:

\[
CV(S) = \frac{1}{n} \sum_{i=1}^{n} (C(t_i) - \hat{f}_i(t_i; S))^2
\]  

(4.87)
where \( \hat{f}(t_i; S) \) is the estimated \( \hat{f}_S(t_i) \) obtained when the observation \( t_i \) has been left out of the calculation. The cross-validation smoothing parameter is that value of \( S \) that minimizes \( \text{CV}(S) \).

This technique can be used to determine either an overall value for \( S \) or a local value. In the latter case, the procedure has to be applied for each value \( t_0 \) at which \( f(t) \) has to be estimated by \( \hat{f}(t) \). An example of a smoother using local cross-validation is the so-called super-smoother of Friedman and Stuetzle (1982; see also Friedman, 1984). This smoother is a variable bandwidth loess smoother using a polynomial of order 1. An example of results produced by the super-smoother estimator is shown in Figure 4.11 for ambient SO\(_2\) air concentrations at Algoma, Ontario.

### 4.3.8.2 Risk Estimation

As an alternative to cross-validation as a way of selecting smoothing parameters, one can minimize a risk function. A risk function is a function of the smoothing parameters that describes some characteristic of the smoothing process, such as the expected shared errors. An example is the mean average square error (MASE), which is defined as

\[
MASE(S) = E\left[ \frac{1}{n} \sum_{i=1}^{n} (\hat{f}(t_i; S) - f(t))^2 \right] \tag{4.88}
\]

where \( \hat{f}(t_i; S) \) is the smoothed estimate and \( E[.] \) is the expected value of \( [.] \).

To illustrate the use of this technique, let us assume that the smoothing function, \( \hat{f}(t_i; S) \), can be written as

\[
\hat{f}(t_i; S) = \sum_{i=1}^{n} C(t_i) w_i(t; S) \tag{4.89}
\]

where the weights \( w_i(t; S) \) are constants for a fixed value of \( S \). This relation applies in the case of the running mean and kernel smoothers. The MASE of this estimator can be written as

\[
MASE(S) = E\left[ \frac{1}{n} \text{RSS}(S) \right] + 2\frac{\sigma^2}{n} \sum_{i=1}^{n} w_i(t_i; S) - \sigma^2 \tag{4.90}
\]

where \( \sigma^2 \) is the variance of the residuals and

\[
\text{RSS}(S) = \sum_{i=1}^{n} (C(t_i) - \hat{f}(t_i; S))^2 \tag{4.91}
\]

The quantity \( M(S) = MASE(S) + \sigma^2 \) has the same minimizer as \( MASE(S) \), and one can estimate \( M(S) \) as

\[
\hat{M}(S) = \frac{1}{n} \text{RSS}(S) + \frac{2\hat{\sigma}^2}{n} \sum_{i=1}^{n} w_i(t_i; S) \tag{4.92}
\]

where \( \hat{\sigma}^2 \) is a model free estimator of the variance. One can use, for example, a difference-based estimator such as
\[ \hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (C(t_i) - C(t_{i-1}))^2 \]  

(4.93)

The value of the smoothing parameter to use is the value of \( S \) that minimizes \( \hat{M}(S) \) and therefore the mean average squared error.

4.3.9 Discussion.

The smoothing techniques presented here are only examples of some of the different techniques that have been proposed. Other techniques include the use of polynomials fitted by least squares or the use of a finite series of functions or even a combination of the two (see Eubank and Speckman, 1993). Other forms of orthogonal functions, like Chebyshev’s polynomials, could also be used. The use of polynomial and/or trigonometric functions to fit a curve to data is related to fitting a model by least squares or by maximum likelihood, which we will discuss in Section 4.6.

In this overview, we have illustrated the use of smoothing techniques for time series. These techniques can also be used to estimate the relationship between two variables. For that purpose, relation 4.1 becomes:

\[ y(x) = f(x) + \varepsilon \]  

(4.94)

and \( n \) pairs of observations \((x_i, y_i)\) are available. An example of such an application would be the use of a smoothing technique to estimate the relationship between ionic concentration and precipitation amount in the case of ionic concentration in precipitation.

All of the techniques presented earlier can be used. The first step in the analysis is to order the pairs in increasing order for \( x \) such that \( x_1 \leq x_2 \leq \ldots \leq x_n \). Note that modifications may have to be made to some of the techniques (e.g., smoothing splines) if there are ties in the \( x \) variables.

All the smoothing techniques presented here are very useful, especially when they are used to get a feel for the data, as was illustrated in Section 2.8. However, all of them, except a special version of the smoothing spline technique, assume that the residuals are independent, which is an assumption that does not hold for most time series encountered in atmospheric chemistry. Therefore, if one constructs approximate confidence bands for the smoothing curves, these will only be indicative at best, as they incorporate the assumption that the residuals are not correlated. However, they may still be used to indicate the possible significance of the trend, and, especially in the case of long-term trends, whether the trend is a straight line or monotonic. This can help determine other techniques that could be used.

Smoothing techniques can also be very useful if the temporal variation of many ions at one or several sites has to be analyzed. They can be used to screen for those variations that are more important or significant and that are therefore good candidates for further analysis with more complex techniques.

4.4 Spectral Analysis

4.4.1 Introduction
We have seen in Section 2.7 that spectral analysis can be very useful in isolating the different components contributing to the temporal variations observed in a time series. Spectral analysis can also be very useful in two other ways. First, the smoothed power spectrum can be used to test which components are statistically significant. Secondly, spectral analysis can be used to filter out the high frequency variations in the time series, and thus isolating the long-term variations. We will discuss these two applications in this section.

4.4.2 Testing for the significance of temporal variation

The smoothed power spectrum of a time-series can be calculated by using the following relation (see Appendix I; Chatfield, 1984):

\[
\hat{f}(\omega) = \frac{1}{\pi} \left\{ \hat{\beta}_0 c_0^2 + 2 \sum_{k=1}^{M} \hat{\beta}_k c_k^2 \cos(2\pi \omega k) \right\}
\]

(4.95)

where \(\omega (0 \leq \omega \leq 0.5)\) is the frequency in cycles per hour, day, or other period, depending on the data observation frequency; \(c_k^2\) is the auto-covariance coefficient of lag \(k\); \(M (< N)\) is the truncation point that has to be chosen (the lower \(M\) is, the smoother the spectrum is); \(\hat{\beta}_k\) is the smoothing window; and \(N\) is the number of observations. Many smoothing windows exist, the most popular being the Tukey and the Parzen windows (see Jenkins and Watts, 1968; Bloomfield, 1976; Priestley, 1981; Chatfield, 1984; Fuller, 1996).

It can be shown that a \((1-\alpha)100\%\) approximate confidence band for \(\hat{f}(\omega)\) is (see Jenkins and Watts, 1968; Priestley, 1981; Chatfield, 1984; Brockwell and Davis, 1987):

Figure 4.12 Examples of power spectra with 95% confidence bands around the assumed spectrum.
where the number of degrees of freedom of the lag window, \( \nu \), is equal to:

\[
\nu = 2N \sum_{k=-N}^{N} \chi_k^2
\]  

(4.97)

and \( \chi^2(\nu) \) is the chi-squared distribution with \( \nu \) degrees of freedom.

Although one can use relation 4.96 to draw a confidence band around \( \hat{f}(\omega) \), this will not be very useful. A better use for that relation is as follows.

The power spectrum of random noise is given by (see Jenkins and Watts, 1968; Priestley, 1981; Chatfield, 1984; Brockwell and Davis, 1987; Fuller, 1996):

\[
f(\omega) = \frac{\sigma^2}{2\pi} \]  

(4.98a)

for \(-0.5 \leq \omega \leq 0.5\) or

\[
f(\omega) = \frac{\sigma^2}{\pi} \]  

(4.98b)

for \(0 \leq \omega \leq 0.5\), where \( \sigma^2 \) is the variance of the random noise.

If we assume that data are random noise, we can estimate the variance using Parseval's relation (Priestley, 1981) on the smoothed power spectrum of the observations or:

\[
\hat{\sigma}^2 = 2\pi \int_{0}^{0.5} \hat{f}(\omega) \, d\omega \]  

(4.99)

Figure 4.13 Examples of power spectra with 95% confidence intervals for the assumed spectrum. The observed spectrum is for ambient SO\textsubscript{2} air concentration at Chalk River, Ontario.

Using that estimate, one can create the power spectrum for a random process with the same variance as that observed in the data. A confidence band for that spectrum can be obtained
using relation 4.96 for the power spectrum of the random process. Figure 4.12a presents the results of such an analysis if the observations are from a normal random distribution. The constant solid line is the spectrum for the random process with the same variance as the data and the dashed lines are the 95% confidence bands. One can see that, overall, the spectrum for the observations is within the confidence band.

Figure 4.13a presents the spectrum for ambient SO₂ air concentrations at Chalk River, Ontario. On the same graph the power spectrum and the 95% confidence band of a white noise process with the same variance as the observed data are shown. One can clearly see that the observed spectrum is not from a random process. Significant long-term and seasonal variations are present. Also one notices that for high frequencies the observed power spectrum is lower than the lowest confidence limit. This indicates that auto-correlations are present in the observed data.

The analysis can now be continued by using the power spectrum of an auto-regressive models of different orders and coefficients to improve the fit between model and observations.

Let us write the auto-regressive model of order p as:

\[ y_i = \sum_{k=1}^{p} a_k y_{i-k} + \epsilon_i \]  \hspace{1cm} (4.100)

Then the power spectrum of this auto-regressive model is (see Jenkins and Watts, 1968; Priestley, 1981; Brockwell and Davis, 1987; Fuller, 1996):

\[ f(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{\left[ 1 - \sum_{j=1}^{p} a_j e^{i2\pi\omega j} \right] \left[ 1 - \sum_{j=1}^{p} a_j e^{-i2\pi\omega j} \right]} \]  \hspace{1cm} (4.101a)

for \(-0.5 \leq \omega \leq 0.5\) or

\[ f(\omega) = \frac{\sigma^2}{\pi} \frac{1}{\left[ 1 - \sum_{j=1}^{p} a_j e^{i2\pi\omega j} \right] \left[ 1 - \sum_{j=1}^{p} a_j e^{-i2\pi\omega j} \right]} \]  \hspace{1cm} (4.101b)

for \(0 \leq \omega \leq 0.5\) where \(\sigma^2\) is the variance of the \(\epsilon_i\) in relation 4.100.

For an auto-regressive model of order 1, 4.101b becomes:

\[ f(\omega) = \frac{\sigma^2}{\pi} \frac{1}{\left( 1 + a^2 - 2a \cos(2\pi\omega) \right)} \]  \hspace{1cm} (4.102)

It is interesting to note that for \(a > 0\), \(f(\omega)\) is a decreasing function of \(\omega\) and an increasing function if \(a > 0\).

For a second order auto-regressive model, relation 4.101b becomes:

\[ f(\omega) = \frac{\sigma^2}{\pi} \frac{1}{\left( 1 + a_1^2 + a_2^2 - 2a_1(1-a_2) \cos(2\pi\omega) - 2a_2 \cos(4\pi\omega) \right)} \]  \hspace{1cm} (4.103)
which is more complex than that for a first order auto-regressive model.

Using the same technique as was used for white noise, one can plot relation 4.102 using Parseval’s relation to estimate the variance. Figure 4.12b illustrates the power spectrum of a time series created with a first order auto-regressive model \(a = 0.4\) with normal errors. The theoretical power spectrum for an auto-regressive model of order 1 with the same variance as the “observed” data and a coefficient equal to 0.4 is also shown. The graph also includes the 95% confidence band. As in the case of white noise, the power spectrum for the “observed” data varies around the theoretical spectrum and remains generally within the confidence bands.

Figure 4.13b presents again the power spectrum for ambient SO\(_2\) air concentrations at Chalk River, Ontario. As a theoretical model, we have now used a first order auto-regressive model with coefficient equal to 0.25. One notices first that the shape of the theoretical power spectrum seems correct but is too high. The explanation for this is simple. The estimation of variance obtained using Parseval’s relation includes the variance due to all the different types of temporal variations present in the observed data. Therefore, it includes the contributions of the long-term trend and the seasonal cycle. Therefore, this estimate is too high if one uses only the auto-regressive model. An easy way to estimate only the variance due to the auto-regression in the data is to start the integration at an higher frequency to remove the influence of lower frequency temporal variations. The result obtained if one integrates for \(\omega\) between 0.02 and 0.5 is shown in Figure 4.14. One can see that the theoretical power spectrum is centered for the high frequencies in the middle of the observed power spectrum. This plot indicates that, in addition to auto-regression, there are statistically significant contributions to the total variance due to a seasonal cycle and longer period variations. These latter could include a long-term trend and long-term cycles.

This example illustrates how the power spectrum of the observed data can be used to test the statistical significance of the different temporal variations that may contribute to the total variance.

4.4.3 Linear Filters

A linear filter is defined by the relation

\[
\hat{C}_t = \sum_{k=\tau}^s g_k C_{t-k}
\]

(4.104)
where \( C_t \) is the observation at time \( t \) and \( \hat{C}_t \) is the filter data at time \( t \). The \( g_k \)'s are the coefficients of the filter. Usually the filter would be symmetric and \( r = -p \) and \( s = p \).

An interesting property of linear filters can be deduced if one assumes that the data are sinusoidal, i.e.,

\[
C_t = R \cos(\omega t + \phi) = \text{Re}\left( R e^{i(\omega t + \phi)} \right)
\]

where \( \text{Re}(\cdot) \) indicates the real part of \( \cdot \). If one replaces \( C_t \) in 4.104, one obtains:

\[
\hat{C}_t = \sum_{k=r}^{s} g_k \text{Re}\left( R e^{i(\alpha t + \phi)} \right) \\
= \text{Re}\left( R e^{i(\alpha t + \phi)} \sum_{k=r}^{s} g_k e^{-i\omega k} \right) \\
= \text{Re}\left( R e^{i(\omega t + \phi)} G(\omega) \right) \tag{4.105}
\]

where \( G(\omega) \) is defined as

\[
G(\omega) = \sum_{k=r}^{s} g_k e^{-i\omega k} \tag{4.106}
\]

The function \( G(\omega) \) is called the transfer function of the filter, since it indicates the way in which a sinusoid of frequency \( \omega \) is transferred by the linear filter. This is valid for all frequencies. If one knows the weights of the linear filter, the transfer function can be obtained. The squared magnitude of the transfer function, \( |G(\omega)|^2 \), is called the power transfer function. In the case of a symmetric filter, one gets

\[
G(\omega) = g_0 + 2 \sum_{k=1}^{p} g_k \cos(\omega k) \tag{4.107}
\]

If one multiplies each side of relation 4.106 by \( e^{i\omega j} \) and integrates \( \omega \) from \(-\pi\) to \( \pi\), one gets:

\[
\int_{-\pi}^{\pi} G(\omega) e^{i\omega j} d\omega = \int_{-\pi}^{\pi} \left( \sum_{k=r}^{s} g_k e^{-i\omega k} e^{i\omega j} \right) d\omega \\
= \sum_{k=r}^{s} g_k \left( \int_{-\pi}^{\pi} e^{-i\omega(k-j)} d\omega \right) \\
= \int_{-\pi}^{\pi} G(\omega) e^{i\omega j} d\omega = 2\pi \sum_{k=r}^{s} g_k \delta_{kj} \tag{4.108}
\]

where \( \delta_{kj} = \int_{-\pi}^{\pi} e^{-i\omega(k-j)} d\omega \) is the Kronecker delta and is defined as:

\[
\delta_{mn} = \begin{cases} 
0 & \text{if } m \neq n \\
1 & \text{if } m = n 
\end{cases} \tag{4.109}
\]

Therefore, relation 4.108 can be written as:
\[ g_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(\omega) e^{i\omega j} d\omega \quad (4.110) \]

Relations 4.106 and 4.110 show that the transfer function, \( G(\omega) \), and the constant \( g_k \) are Fourier transformed from one another. Thus, one can define a transfer function, and then using relation 4.110 obtain the coefficients \( g_k \) that can be used in relation 4.104, or, one can do the opposite.

An example of a symmetric linear filter is the moving-average obtained by writing:

\[ g_k = \frac{1}{2p+1} \text{ for } -p \leq k \leq p \quad (4.111) \]

Thus, the moving-average filter is:

\[ \hat{C}_r = \frac{1}{2p+1} \sum_{k=-p}^{p} C_{r-k} \quad (4.112) \]

and the transfer function, using relation 4.107, is:

\[
G(\omega) = \frac{1}{2p+1} + \frac{2}{2p+1} \sum_{k=1}^{p} \cos(\omega k)
\]

\[
= \frac{2}{2p+1} \left( \frac{1}{2} + \sum_{k=1}^{p} \cos(\omega k) \right)
\]

\[
= \frac{2}{2p+1} \frac{\sin((p+1/2)\omega)}{2\sin(\omega/2)}
\]

\[
= \frac{1}{2p+1} \frac{\sin((p+1/2)\omega)}{\sin(\omega/2)}
\]

\[ = D_{2p+1}(\omega) \quad (4.113) \]

The function \( D_n(\omega) \) is called the Dirichlet kernel.

The transfer function, \( G(\omega) \), of a moving average filter is illustrated in Figure 4.15 for different values of \( p \). From that figure, one can see: first, that the transfer function can be negative; second, that it is maximum at \( \omega = 0 \); third that it presents a decreasing oscillations around zero; and fourth, that the gain is more

Figure 4.15 Transfer functions for a moving average filter for three values of \( p \).
concentrated around zero as \( p \) is increased. This last characteristic shows that in choosing \( p \) large enough, one can isolate the low frequency contribution to the temporal variations in the time series. In other words, by using the moving average with a large value of \( p \) one can isolate the long-term trend.

In the preceding example, we first specified the weights, \( g_k \), and then we calculated the transfer function. As mentioned earlier, one can also do the opposite, first choosing a transfer function and then calculating the weight using relation 4.110.

We have seen in Section 2.7, that low frequencies are usually associated with long-term trends and high frequencies with random noise and autocorrelation. The seasonal and diurnal cycles are between the two. If the data have been gathered on a daily basis, the frequency associated with the seasonal cycle is equal to about 0.0027 cycles day\(^{-1}\). Thus if we want to isolate the long term variation, we would like to use a transfer function similar to the one in Figure 4.16 with \( \omega_c \) lower than 0.0027 cycles day\(^{-1}\). These types of filters are called low-pass filters.

The weights of the low-pass filter presented in Figure 4.16 can be calculated using relation 4.110. They are:

\[
g_k = \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} e^{i\omega k} d\omega
\]

\[
= \frac{1}{2\pi} \int_{0}^{\omega_c} (e^{i\omega k} - e^{-i\omega k}) d\omega
\]

\[
= \frac{1}{\pi} \int_{0}^{\omega_c} \cos(k\omega) d\omega
\]

The \( g_k \) are then:

\[
g_k = \frac{\sin(k\omega_c)}{\pi k} \quad |k| \geq 1 \quad (4.114)
\]

and

\[
g_0 = \frac{\omega_c}{\pi} \quad k = 0 \quad (4.115)
\]

Note that for this filter, \( r \) and \( s \) (relation 4.106) are respectively equal to \(-\infty\) and \(+\infty\). In practice, the summation has to be truncated to reasonable values for \( r \) and \( s \). If we do that, the gain function obtained by using those weights in relation 4.106 is different from the ideal
low-pass filter. Ideally, we would like to minimize the difference between the ideal transfer function and the latter. This requirement can be presented mathematically as follow:

Find the coefficient $g_k\{r \leq k \leq s\}$ such that

$$\int_{-\pi}^{\pi} |H(\omega)-G(\omega)|^2 d\omega = \int_{-\pi}^{\pi} |H(\omega)-\sum_{k=r}^{s} g_k e^{-i\omega k}|^2 d\omega$$  

(4.116)

is minimized. $H(\omega)$ is the ideal transfer function. The solution to the minimization problem is

$$g_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\omega) e^{i\omega k} d\omega \quad r \leq k \leq s$$  

(4.117)

which is identical to relation 4.110.

The transfer function associated with these weights is equal to:

$$G_{r,s}(\omega) = \sum_{k=r}^{s} g_k e^{-i\omega k}$$

$$= \sum_{k=r}^{s} \left( \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\lambda) e^{i\omega k} d\lambda \right) e^{-i\omega k}$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\lambda) \left( \sum_{k=r}^{s} e^{-i\omega k} e^{i\omega k} \right) d\lambda$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\lambda) \left( \sum_{k=r}^{s} e^{-i(\omega-\lambda)k} \right) d\lambda$$

$$= \frac{s-r+1}{2\pi} \int_{-\pi}^{\pi} H(\lambda) D_{r-s+1}(\omega-\lambda) e^{-i(r+s)(\omega-\lambda)} d\lambda$$  

(4.118)

where $D_n(.)$ is the Dirichlet kernel. If $H(\omega)$ is real and symmetric and we set $r = -p$ and $s = p$, we get $G_{r,s}(\omega)$ equal to:

$$G_p(\omega) = \frac{2p+1}{2\pi} \int_{-\pi}^{\pi} H(\lambda) D_{2p+1}(\omega-\lambda) d\lambda$$  

(4.119)

An example of the use of that relation is presented in Figure 4.17 for the transfer function presented in Figure 4.16. The two other curves, obtained using 4.119, are for $G_5(\omega)$ and $G_{20}(\omega)$ and $\omega_c = 0.3$. Both curves show pronounced ripples and one of them shows a pronounced overshoot on either sides of the cutoff frequency. This is known as Gibb’s phenomenon.
This overshoot and accompanying ripples may be greatly reduced as follows. It may be shown that the wavelength of the ripples is \( \delta = 4\pi/(2p+1) \). Therefore, the smoothed function
\[
\tilde{G}_s(\omega) = \frac{1}{\delta} \int_{-\delta/2}^{\omega+\delta/2} G_s(\lambda) d\lambda
\] (4.120)
would have far smaller ripples, since the integration is over one complete cycle.

Using relation 4.107, relation 4.120 can be written as:
\[
\tilde{G}_s(\omega) = \frac{1}{\delta} \left[ G_0 + 2 \sum_{k=1}^{p} g_k \cos(k\delta) \right] d\lambda
\]
\[
= g_0 + \frac{1}{\delta} \left[ 2 \sum_{k=1}^{p} g_k \int_{-\delta/2}^{\omega+\delta/2} \cos(k\lambda) d\lambda \right]
\]
\[
= g_0 + 2 \sum_{k=1}^{p} g_k \frac{\sin(k\delta/2)}{k\delta/2} \cos(k\omega) \tag{4.121}
\]
This corresponds to replacing \( g_k \) by:
\[
g_k \frac{\sin(k\delta/2)}{k\delta/2} = g_k \sin\left( \frac{2\pi k}{2p+1} \right) \tag{4.122}
\]
The multipliers
\[
\sigma_{p,k} = \sin\left( \frac{2\pi k}{2p+1} \right) \tag{4.123}
\]
are examples of converge factors.

Figure 4.18 presents \( \tilde{G}_s(\omega) \) and \( \tilde{G}_{20}(\omega) \). One sees that the ripples have been substantially reduced. The overshoots in \( \tilde{G}_{20}(\omega) \) have also been reduced.

It is usually desirable for a filter to pass a constant term (i.e., a component of frequency zero) without changing it. This requires that the transfer function be equal to 1 at \( \omega = 0 \). The \( g_0 \) can easily be modified to insure that the transfer function follows that requirement by simply normalizing them so that their sum equals 1.

**Figure 4.19:** Illustration of the transfer function in relation 4.122 for \( m = 2, 4, \) and 6.

**Figure 4.20:** Weights corresponding to the transfer function in relation 4.122 with \( m = 4 \) and \( \omega_c = 0.5 \).
An example of the use of such filters in atmospheric chemistry can be found in Li et al. (1993).

Another way to define a transfer function without ripples and overshoots is to use an exponential function like (see Thoning et al., 1989):

\[ G(\omega) = e^{-c_1\left(\frac{\omega}{\omega_c}\right)^m} \]  

(4.124)

The term \(c_1\) is usually equal to \(\ln(2)\), so that \(G(\omega) = 0.5\) at \(\omega = \omega_c\). The transfer function is symmetric around zero. This transfer function is illustrated for three values of \(m\) \((m = 2, 4, \text{and } 6)\) in Figure 4.19.

The weights associated with this transfer function can be obtained by using relation 4.110 and are given by:

\[ g_k = \frac{1}{\pi} \int_0^{\pi} e^{-c_1\left(\frac{\omega}{\omega_c}\right)^m} \cos(\omega k) \, d\omega \]  

(4.125)

An example of the weights for a transfer function with \(m = 4\) and \(\omega_c = 0.5\) is given in Figure 4.20. The highest weights are concentrated around \(k = 0\). The reader should note that the weights may be negative for some of the \(k\)'s.

Another way to create complex filters is to take a convenient set of simple filters, like moving averages (relation 4.112) of various lengths, and use them as building blocks to assemble a more complex filter. An example of such a filter is the Kolmogorov-Zubenko filter, \(KZ_{m,p}\), used by Rao and Zubenko (1994), Rao et al. (1995), Flaum et al. (1996), and Milanchuk et al. (1998).

The first step in creating the \(KZ_{m,p}\) filter is to apply a moving average to the data, defined as:

\[ \tilde{C}_i = \frac{1}{m} \sum_{j=-k}^{k} C_{i+j} \]  

(4.126)

where \(m = 2k+1\). The second step is to replace \(C_i\) by \(\tilde{C}_i\) in 4.126 and apply the filter a second time. The filtering operation is applied \(p\) times.

An adaptive version of the Kolmogorov-Zuchero filter has been developed to help in the identification of step changes in the data (see Zubenko et al., 1996). The filter is adaptive because the parameter \(m\) is modified locally to help identify the step changes. Another special adaptation of the Kolmogorov-Zuchero filter is the spatial version used by Zubenko et al. (1995).
Figure 4.22 Artificial data of Figure 2.20 and estimated long-term trend using a low-pass filter with $\omega_c = 0.002$ cycles day$^{-1}$.

The technique used to create the Kolmogorov-Zuchero filter is a standard one and thus can be used to define other complex filters that have been made from simple ones.

To illustrate the use of filters, a low-pass filter constructed using relations 4.112, 4.113, and 4.119 was applied to the artificial data of Figure 2.20. A cut-off frequency equal to 0.002 cycles day$^{-1}$ was used. The weights used for the low-pass filter are shown in Figure 4.21. The resulting estimated long-term trend is presented in Figure 4.22. The long-term trend was estimated only for the region outside the borders, defined so that no areas with missing data were used in relation 4.104. The reader should remember that the long-term trend used in creating the data was a straight line. If we assume that the small variation are not significant (some of the ripples could be eliminated by using a wider window for the filter), the low-pass filter is quite good at finding the long-term trend.

4.4.4 Discussion

We have shown in the preceding section that spectral analysis can be a useful tool for studying the temporal variations present in a time series. The testing techniques presented in Section 4.4.2 can be used to test for the significance of the long-term trend or other components, and the filters can be used to estimate their shape. These two techniques have to be used together to get a complete understanding of the temporal variation in a time series.

There are, however, some limitations to the use of spectral analysis, especially in the use of filters. Data gaps introduce problems in using the spectral analysis technique. Although the power spectrum of the time series can still be estimated using the techniques presented in Appendices II and III, the estimated power spectrum obtained from these techniques can be uncertain if the amount of missing data is large. That uncertainty increases as the number of
missing data items increases. The main problem occurs in the use of filters. In the description of the filter technique presented in Section 4.4.3, it was assumed that no missing data were present. When data are missing, the only solution is to ignore them when using relation 4.104, which is equivalent to assuming a value of zero for these data. Therefore, it is a good idea to center the time-series around zero before analyzing it by subtracting from the data either the mean or the median of the data. Another useful practice is to make sure that the data are also symmetric around the mean or zero. Thus, the data may have to be transformed using a logarithmic or other transformation before the analysis. If these two suggestions are followed, the effect of missing data on the filtering process may be reduced.

Another difficulty related to the missing data problem is the edge effect. This will occur when trying to use relation 4.104 with the index i less than r or greater than (n-s), where n is the total number of data items (note that missing data are considered as valid data in the calculation of n). If i is less than r, there would be (r-i) missing data to the left of i. This introduces an imbalance into the filter equation that may results in a large bias in the estimated trend. Unfortunately, there are no techniques, like those used to solve similar edge effect problems in kernel smoothing (Section 4.3.5), that can be used to compensate for these effects. The same problem occurs also if i is situated within the last s points at the end of the time series.

If the reader has compared relation 4.104 for the linear filters with relations 4.61, 4.64, or 4.65 for the different types of kernel smoothing, he or she will have noticed that these equations are the same, that is, they all have the following general form:

\[ \tilde{C}_i = \sum_{j=r}^{s} w_j C_j \]  

(4.127)

They differ only in the way that the weights \( w_i \) are defined. One can therefore consider linear filtering as kernel smoothing with the kernel defined by spectral analysis. An example of the correspondence between the two is the moving average filter, which is identical to a Nadaraya-Watson kernel smoother with a rectangular kernel (see Section 4.3.5). Note, however, that the two techniques differ when data are missing because they do not treat data gaps in the same ways. Although both techniques ignore missing data, the weights in the kernel smoothing are renormalized so that their sum for the non-missing data is equal to 1. This is not done in the moving average filter.

4.5 Anocov and Anova Techniques

4.5.1 Introduction

The techniques presented in this section have in common the fact that they group data into classes. We have already encountered such groupings in the nonparametric techniques presented in Section 4.2.2 (e.g., the seasonal Kendall test). In the case of the analysis of covariance (anocov) techniques, the grouping is used to define indicator variables that are used to create a linear model. In the analysis of variance (anova) techniques, the data are grouped into two-way tables and an analysis of variance is made. These models are fitted using standard statistical procedures, namely, linear regression in the case of the anocov model and anova technique in the case of the anova model. We will therefore concentrate on describing the models and evoking some of their principal characteristics.
These two techniques assume that the residuals after the elimination of the model are independent and are from a normal distribution with constant variance. This is quite different from the approach used by the techniques presented in the preceding sections. In addition, we have to make an assumption about the exact shape of the distribution of the residuals. Usually, if the amount of data is large, the exact distribution of the residuals may differ slightly from normality. The tolerance of the technique to such variations differs from method to method. We will return to this matter when discussing the two techniques.

### 4.5.2 Anocov Technique

The anocov technique has been used in recent years by Lynch et al. (1995, 1996) and Sirois et al. (1996). A complete description of the anocov techniques can be found in Neter and Wasserman (1974).

If s is the number of “seasons” used to describe the data, an analysis of covariance model can be written as:

\[
C_t = a + bt + \sum_{i=1}^{s-1} b_i I_i^S(t) + \varepsilon_t
\]  
(4.128)

where the residuals, \( \varepsilon_t \), are assumed to be independent random variables from a normal distribution with zero mean and constant variance, \( \sigma^2 \). The \( I_i^S(t) \) are indicator variables. At time t, if the observation is in “season” \( \ell \), we get, if \( i < s \):

\[
I_i^S(t) = \begin{cases} 
1 & i = \ell \\
0 & i \neq \ell 
\end{cases}
\]  
(4.129)

and if \( i = s \), all the \( I_i^S(t) \) are set equal to zero. The seasons could be the usual seasons (s = 4) or bi-monthly seasons (s = 6) or months (s = 12) or even weeks (s = 52).

In atmospheric chemistry, ion concentrations and other observed variables are often not normally distributed but log-normally distributed. An appropriate model in that case would be:

\[
\log(C_t) = a + bt + \sum_{i=1}^{s-1} b_i I_i^S(t) + \varepsilon_t
\]  
(4.130)

if s seasonal indicators are used. This is the model used by Lynch et al. (1995, 1996) on bi-monthly data (s = 6). In the case of ion concentrations in precipitation, a more appropriate model would be:

\[
\log(C_t) = a + bt + c \log(P(t)) + \sum_{i=1}^{s-1} b_i I_i^S(t) + \varepsilon_t
\]  
(4.131)

where P(t) is the precipitation amount associated with the sample at time t. Note that, depending on the length of the sampling period, the parameter c may not be statistically significant.

Although a straight line has been assumed thus far for the long-term trend in relations 4.128, 4.130, and 4.131, any polynomial and/or trigonometric function could be used to describe the long-term trend.
Figure 4.23: Fitting an anocov model (relation 4.129) with four seasons to 28-day \( \text{SO}_4^{2-} \) concentrations in precipitation at Penn State, Pennsylvania.

Once a model has been postulated, it is fitted to the data using the usual linear regression technique, and the statistical significance of the long-term trend is verified using Student-t tests. In the case of a straight line long-term trend, this means testing the hypothesis that \( b = 0 \) against the alternative that \( b \neq 0 \).

To illustrate the use of the anocov technique, the model described by relation 4.130 was fitted to 28-day \( \text{SO}_4^{2-} \) ionic concentration in precipitation from Penn State, Pennsylvania. The resulting long-term trends are shown in Figure 4.23. The P-value for the test of \( b = 0 \) against \( b \neq 0 \) is equal to 0.0725. Therefore, the hypothesis of no trend cannot be rejected at a 95% confidence level. In other words, the trends presented in Figure 4.23 are not statistically significant.

An interesting feature of the models described by relations 4.128, 4.130, and 4.131, is that they require us to fit not only one straight line to the data but \( s \) straight lines with the same slope but with different intercepts corresponding to the different seasons (see Figure 4.23). Note that some of these lines may not be statistically different. It is also assumed that the variance of the residuals is the same in each season.

These models can be extended to take into account the possibility that the variance and the slope in each season could be different by separately fitting a straight line in each season. If we assume that the slopes all have the same sign, we can test for the significance of the overall trend by testing whether the sum of the slopes is different from zero. Using the hypothesis that the slopes all have the same sign, their sum will be equal to zero only if they are all equal to zero. The description of this test can be found in Loftis et al. (1989).

A better way is to test the different slopes simultaneously for significance. This make it possible to see in which season the long-term trends are the most important. Techniques for multiple simultaneous tests can be found in Miller (1981) and Hochberg and Tamhane (1987).
One advantage of the anocov method is that missing data are easily treated because one can simply ignore them.

4.5.3 Anova Technique

To apply the anova technique, we group the data into a two-way table with year as the rows index and “season” as the columns index. If we let $C_{ijk}$ be the kth observation for year i, and “season” j, a general two-way table will look like this:

<table>
<thead>
<tr>
<th>Year</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$C_{111}$</td>
<td>$C_{112}$</td>
<td>$C_{121}$</td>
<td>$C_{122}$</td>
<td>$C_{1m1}$</td>
</tr>
<tr>
<td></td>
<td>$C_{1n1}$</td>
<td>$C_{1n2}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$C_{211}$</td>
<td>$C_{212}$</td>
<td>$C_{221}$</td>
<td>$C_{222}$</td>
<td>$C_{2m1}$</td>
</tr>
<tr>
<td></td>
<td>$C_{2n1}$</td>
<td>$C_{2n2}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>$C_{n11}$</td>
<td>$C_{n12}$</td>
<td>$C_{n21}$</td>
<td>$C_{n22}$</td>
<td>$C_{nm1}$</td>
</tr>
<tr>
<td></td>
<td>$C_{n1n}$</td>
<td>$C_{n2n}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where $n_{ij}$ is the number of data in year i and “season” j, n is the number of years, and m is the number of “seasons”. The definition of the “seasons” is arbitrary. They can be the 4 usual seasons, or 12 months, or even 52 weeks. Note that the $n_{ij}$ can be different.

This two-way table is quite general. The simplest case occurs when there is only one observation per cell. The two-way table then becomes:

<table>
<thead>
<tr>
<th>Year</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$C_{11}$</td>
<td>$C_{12}$</td>
<td></td>
<td></td>
<td>$C_{1m}$</td>
</tr>
<tr>
<td>2</td>
<td>$C_{21}$</td>
<td>$C_{22}$</td>
<td></td>
<td></td>
<td>$C_{2m}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>$C_{n1}$</td>
<td>$C_{n2}$</td>
<td></td>
<td></td>
<td>$C_{nm}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A possible linear model for these data would be:

\[ C_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]  

(4.132)

with the restrictions

\[ \sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0 \]  

(4.133)

where \( \mu \) is the overall mean, the \( \alpha_i \) (\( i = 1, 2, \ldots, I \)) are the year effects, the \( \beta_j \) (\( j = 1, 2, \ldots, J \)) are the seasons effects, and the \( \epsilon_{ij} \) are independent random errors from a normal distribution with equal variance \( \sigma^2 \). A more general model would include interactions between seasons and years and could be written as:

\[ C_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij} \]  

(4.134)

with the additional restrictions:

\[ \sum_{j=1}^{J} \gamma_{ij} = 0 \text{ for all } j, \quad \sum_{i=1}^{I} \gamma_{ij} = 0 \text{ for all } i \]  

(4.135)

where \( \gamma_{ij} \) is the interaction between year \( i \) and month \( j \).

These models can be easily fitted to the data using regular anova analysis (see Scheffé, 1959; Neter and Wasserman, 1974; Lehmann, 1986; and Miller, 1986). The parameters \( \mu, \alpha_i \) (\( i = 1, 2, \ldots, I \)) and \( \beta_j \) (\( j = 1, 2, \ldots, J \)) can be estimated. The variations of the \( \alpha_i \)'s are an estimate of the long-term trend and those of the \( \beta_j \)'s are an estimate of the seasonal cycle. The statistical significance of the long-term trend can be checked by testing the following hypothesis:

\[ H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_n = 0 \]  

(4.136)

against

\[ H_1: \text{any two of } \alpha_i \text{ are different.} \]  

(4.137)

The statistical significance of the seasonal cycle can also be verified using a similar test for the \( \beta_j \)'s. Note that these tests are resistant to non-normality in the distribution of the residual if the size of the design (i.e., \( IJ \)) is not too small (Miller, 1986).

A resistant and nonparametric way of estimating the parameters of model 4.132 is the median polish (see Emerson and Hoaglin, 1983). This is an iterative technique using the median. The reader is referred to Emerson and Hoaglin (1983) for more details. The technique is quite resistant to outliers, and missing data are allowed. The only disadvantage is that the statistical significance of the variation observed in the \( \alpha_i \)'s cannot be tested. An example of the use of median polish for long-term trend analysis can be found in Wu and Zidek (1989). The median polish technique can be extended for the model described by relation 4.134 when the interactive term can be written as proportional to the product of the two other terms, that is, when model 4.134 can be written as (see Emerson and Wong, 1985):

\[ C_{ij} = \mu + \alpha_i + \beta_j + \kappa \alpha_i \beta_j + \epsilon_{ij} \]  

(4.138)
The relations 4.132 and 4.134 are easily extended to the case where there are \( n \) observations (i.e., all \( n_{ij} = n \) in the general two-way table) for each season and year combination. Relation 4.132 becomes:

\[
C_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}
\]  

(4.139)

for \( k = 1, 2, \ldots, n \). The restrictions in relation 4.133 and 4.135 still apply. This model can be easily fitted to the data using the anova technique analysis (see Scheffé, 1959; Neter and Wasserman, 1974; Lehmann, 1986; and Miller, 1986), and hypotheses such as

\[
H_0: \gamma_{ij} = 0, \quad H_0: \alpha_i = 0 \quad \text{and/or} \quad H_0: \beta_j = 0
\]  

(4.140)

can easily be tested. As in the case where only one observation is available per cell, these tests are resistant to non-normality of the distribution of the residuals if the size of the design (i.e. \( n_{ij} \)) is not too small (Miller, 1986).

If the number of observations in each cell differs (i.e., the \( n_{ij} \) are not all equal), the analysis becomes more difficult. In most cases the best technique to use to fit relation 4.139 is to employ multiple regressions (see Draper and Smith, 1981, and Milliken and Johnson, 1984). Unfortunately, even if multiple regressions are used, the analysis is murkier than in the balanced case. One of the characteristics of such an analysis is that the sequence in which the hypotheses are tested in relation 4.140 makes a difference. The effect of non-normality of the distribution of the residuals on the tests is also more important in the case of an unbalanced design. These effects increase as the differences between the \( n_{ij} \) increase.

To illustrate the utilization of the anova technique, we will use the observed \( \text{SO}_4^{2-} \) ionic concentration in precipitation at Chalk River, Ontario, between 1979 and 1996. The time series is shown in Figure 4.24. One can detect on this plot a net decrease during the 18 years. A two-way table was generated by calculating the median concentration for each month. The two-way table thus obtained is given in Table 4.1. These data are also illustrated using a three dimensional plots in Figure 4.25. One can easily detect a net decrease in the
Table 4.1 Monthly median $\text{SO}_4^{2-}$ ionic concentrations in precipitation at Chalk River, Ontario.

<table>
<thead>
<tr>
<th>Year</th>
<th>J</th>
<th>F</th>
<th>M</th>
<th>A</th>
<th>M</th>
<th>J</th>
<th>J</th>
<th>A</th>
<th>S</th>
<th>O</th>
<th>N</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1979</td>
<td>27.1</td>
<td>42.7</td>
<td>87.5</td>
<td>47.9</td>
<td>104.1</td>
<td>121.8</td>
<td>176.9</td>
<td>152.0</td>
<td>127.0</td>
<td>72.9</td>
<td>60.4</td>
<td>53.1</td>
</tr>
<tr>
<td>1980</td>
<td>48.7</td>
<td>29.2</td>
<td>45.8</td>
<td>69.8</td>
<td>68.7</td>
<td>106.2</td>
<td>60.4</td>
<td>77.0</td>
<td>72.9</td>
<td>60.4</td>
<td>69.8</td>
<td>31.2</td>
</tr>
<tr>
<td>1981</td>
<td>86.4</td>
<td>66.6</td>
<td>41.6</td>
<td>97.9</td>
<td>81.2</td>
<td>137.4</td>
<td>143.7</td>
<td>93.7</td>
<td>90.6</td>
<td>43.7</td>
<td>54.1</td>
<td>53.1</td>
</tr>
<tr>
<td>1982</td>
<td>27.1</td>
<td>30.2</td>
<td>81.2</td>
<td>124.9</td>
<td>125.9</td>
<td>81.2</td>
<td>88.5</td>
<td>43.7</td>
<td>60.4</td>
<td>73.9</td>
<td>52.1</td>
<td>41.6</td>
</tr>
<tr>
<td>1983</td>
<td>28.1</td>
<td>18.7</td>
<td>136.2</td>
<td>35.2</td>
<td>54.1</td>
<td>58.1</td>
<td>77.4</td>
<td>44.1</td>
<td>71.1</td>
<td>38.3</td>
<td>33.5</td>
<td>27.5</td>
</tr>
<tr>
<td>1984</td>
<td>21.2</td>
<td>58.9</td>
<td>47.1</td>
<td>33.3</td>
<td>65.3</td>
<td>92.4</td>
<td>63.7</td>
<td>81.0</td>
<td>55.4</td>
<td>71.4</td>
<td>48.9</td>
<td>21.1</td>
</tr>
<tr>
<td>1985</td>
<td>10.2</td>
<td>18.9</td>
<td>44.0</td>
<td>54.1</td>
<td>68.1</td>
<td>46.7</td>
<td>48.5</td>
<td>60.2</td>
<td>51.4</td>
<td>41.9</td>
<td>29.4</td>
<td>17.8</td>
</tr>
<tr>
<td>1986</td>
<td>32.8</td>
<td>23.0</td>
<td>70.4</td>
<td>54.9</td>
<td>49.6</td>
<td>41.4</td>
<td>19.2</td>
<td>63.5</td>
<td>51.4</td>
<td>73.7</td>
<td>74.3</td>
<td>17.5</td>
</tr>
<tr>
<td>1987</td>
<td>13.0</td>
<td>4.8</td>
<td>51.2</td>
<td>40.2</td>
<td>80.9</td>
<td>76.7</td>
<td>60.2</td>
<td>47.1</td>
<td>45.8</td>
<td>47.2</td>
<td>20.8</td>
<td>19.0</td>
</tr>
<tr>
<td>1988</td>
<td>56.4</td>
<td>14.7</td>
<td>40.8</td>
<td>52.3</td>
<td>54.3</td>
<td>24.9</td>
<td>29.2</td>
<td>71.4</td>
<td>112.7</td>
<td>35.3</td>
<td>53.5</td>
<td>16.5</td>
</tr>
<tr>
<td>1989</td>
<td>29.9</td>
<td>3.8</td>
<td>27.3</td>
<td>74.4</td>
<td>40.8</td>
<td>62.3</td>
<td>49.8</td>
<td>71.8</td>
<td>43.9</td>
<td>27.3</td>
<td>22.9</td>
<td>9.8</td>
</tr>
<tr>
<td>1990</td>
<td>32.2</td>
<td>26.9</td>
<td>56.0</td>
<td>58.1</td>
<td>36.6</td>
<td>63.1</td>
<td>32.5</td>
<td>87.9</td>
<td>50.4</td>
<td>38.2</td>
<td>41.9</td>
<td>23.3</td>
</tr>
<tr>
<td>1991</td>
<td>12.7</td>
<td>18.1</td>
<td>28.7</td>
<td>34.0</td>
<td>72.7</td>
<td>102.1</td>
<td>41.4</td>
<td>17.1</td>
<td>33.6</td>
<td>40.0</td>
<td>38.2</td>
<td>13.6</td>
</tr>
<tr>
<td>1992</td>
<td>19.6</td>
<td>13.4</td>
<td>56.2</td>
<td>126.9</td>
<td>55.2</td>
<td>59.6</td>
<td>19.0</td>
<td>29.4</td>
<td>57.9</td>
<td>39.6</td>
<td>35.6</td>
<td>30.4</td>
</tr>
<tr>
<td>1993</td>
<td>11.5</td>
<td>17.9</td>
<td>27.6</td>
<td>82.7</td>
<td>59.3</td>
<td>42.7</td>
<td>30.8</td>
<td>43.7</td>
<td>27.1</td>
<td>59.3</td>
<td>35.2</td>
<td>26.6</td>
</tr>
<tr>
<td>1994</td>
<td>7.6</td>
<td>7.7</td>
<td>43.1</td>
<td>38.1</td>
<td>37.1</td>
<td>54.5</td>
<td>53.9</td>
<td>48.9</td>
<td>38.5</td>
<td>28.9</td>
<td>30.4</td>
<td>21.9</td>
</tr>
<tr>
<td>1995</td>
<td>11.5</td>
<td>22.7</td>
<td>67.1</td>
<td>71.0</td>
<td>41.9</td>
<td>86.5</td>
<td>24.6</td>
<td>36.7</td>
<td>33.3</td>
<td>16.9</td>
<td>14.8</td>
<td>15.2</td>
</tr>
<tr>
<td>1996</td>
<td>13.4</td>
<td>20.0</td>
<td>33.8</td>
<td>39.8</td>
<td>32.2</td>
<td>40.1</td>
<td>39.0</td>
<td>64.4</td>
<td>43.0</td>
<td>28.6</td>
<td>18.3</td>
<td>25.8</td>
</tr>
</tbody>
</table>

The median concentration with time in both Table 4.1 and in Figure 4.25. The concentrations are noticeably higher between 1979 and 1981 than between 1994 and 1996.

The model given by relation 4.132 was fitted to the data using both an anova analysis and the median polish technique. The anova analysis indicates statistically significant effects for both the monthly and annual factors. For both tests, P-values less than 0.00000 were obtained.

![Figure 4.25 Variation with month and year of the monthly median $\text{SO}_4^{2-}$ ionic concentration in precipitation at Chalk River, Ontario.](image-url)
The ANOVA and median polish estimates of the overall mean (parameter $\mu$ in relation 4.132) are 49.9 and 48.5 respectively. These two estimates are very close to each others.

The year effects are shown in Figure 4.24 for both the ANOVA analysis and the median polish analysis. The two results are again very close to each others. Overall the median polish estimates give a slightly lower reduction in the median concentration than the ANOVA analysis.

The month effects are shown for both techniques in Figure 4.27. As with the year effects, the results obtained using the two techniques are quite close to each other, although the median polish results are slightly lower than the ANOVA results.

This example shows that the ANOVA method can be used not only for testing for long-term trends in the data but also for estimating those long-term trends. Unlike the anocov technique discussed in the preceding section, the ANOVA technique does not require a hypothesis about the exact shape of the trend.

4.5.4 Discussion

As we mentioned earlier, the anocov techniques have been used in atmospheric chemistry by a number of authors. However, the assumptions made by these techniques are quite stringent. First, the long-term trends are assumed to be straight lines with the same slope. Even if one can assume that the long-term trends are straight lines, there is a high probability that their slopes will be different in the different seasons. Therefore, the reader should verify before using these techniques that there is a basis for assuming that the long-term trends for each season followed a straight line and that their slopes are the same. This can be done, for example, by using the smoothing techniques presented in Section 4.3. The reader should always remember that polynomials of higher degree than 1 can be used and that different long-term trends can be fitted for different seasons.
Three other hypotheses used in the anocov techniques should be checked before accepting the results of the analysis. The anocov technique is quite tolerant of non-normality in the residuals if the number of data items is large. Moderate to large variation across season and with time in the variance of the residuals will create problems, however. In such circumstances, it may be necessary to transform the data to stabilize the variance. The presence of large correlations between the residuals may also create problems. If such auto-correlations are present it may be prudent to use the iterative procedure described in Section 4.6.3.1.

To the knowledge of the author, the anova technique has not been used before to study data in atmospheric chemistry, although it may be a useful tool as the simple example presented here illustrates. On the one hand, the technique is versatile, as it is not necessary to assume a priori any shape for the long-term trend. On the other hand, there is no guarantee that the additive model will be appropriate for explaining the observations. However, one can always use an interaction term in the model to improve its fit to the data. The most important problem is the fact that in most cases the design will be strongly unbalanced (i.e., there will be large differences in the number of data items in the cells), and this will create problems with the testing and interpretation of the results.

As we have seen, the anova technique is quite resistant to the non-normality of the residuals if the size of the design is large. The only exception occurs when the design is strongly unbalanced. As in the case of the anocov technique, variability in the residual variance and/or serial correlations of the residuals will create problems if they are important. A small variability in the variance of the residuals will not drastically affect the result of the analysis. The same is true for serial correlations in the residuals.

An interesting extension of the anova technique involves using a three-way table in place of a two-way table. The third dimension could be used for sites, for example, if we wanted to study temporal variation on a regional scale. The median polish technique can also be extended. For the latter, the reader is referred to Cook (1985) for a complete description of the technique. This technique has been used in the past by Wu and Zidek (1989).

A nonparametric version of the anocov and anova techniques can be created by replacing the observations by their ranks (Conover, 1980). This technique can be used to test for the statistical significance of the trend but not to estimate it.

4.6 Fitting models by Least squares or Maximum Likelihood Techniques

4.6.1 Introduction

One of the most popular techniques for analyzing temporal variation in a time series involves fitting a more or less general linear model to the data using the least squares or maximum likelihood techniques. The most popular model can be written as:

\[ C_i = \alpha + \beta t + \epsilon_i \]  

or, when one wants to model the seasonal cycles,

\[ C_i = \alpha + \beta t + \gamma \sin(2\pi t + \phi) + \epsilon_i \]

where \( \epsilon_i \) are independent random errors from a normal distribution with mean zero and constant variance \( \sigma^2 \) (e.g., MAP3s/RAINE, 1982). This model is often slightly modified to
take into account the fact that the distributions of ionic concentrations are closer to log-normality than to normality. Relation 4.142 is then written as:

$$\log(C_i) = \alpha + \beta t + \gamma \sin(2\pi t + \phi) + \varepsilon_i$$  \hspace{1cm} (4.143)

or

$$\log(C_i) = \alpha + \beta t + \gamma_1 \sin(2\pi t) + \gamma_2 \cos(2\pi t) + \varepsilon_i$$  \hspace{1cm} (4.144)

(e.g., Dana and Easter, 1987). In the case of ionic concentration in precipitation, a further improvement is necessary to take into account the possible relationship between concentration and precipitation amount. This results in the following model:

$$\log(C_i) = \alpha + \delta \log(P_t) + \beta t + \gamma_1 \sin(2\pi t) + \gamma_2 \cos(2\pi t) + \varepsilon_i$$  \hspace{1cm} (4.145)

where $P_t$ is the precipitation amount associated with the sample at time $t$ (e.g., Berge, 1988). Buishand et al. (1988) have proposed the following model:

$$C_i = \alpha + \delta f(P_t) + \beta t + \gamma \cos(2\pi t - \phi) + \varepsilon_i$$  \hspace{1cm} (4.146)

where $f(P_t)$ could be either $\sqrt{P_t}$ or $1/\sqrt{P_t}$.

All these models are special cases of the following general model for hourly data (e.g., Sirois, 1993):

$$g(C_i) = C_0 + f^p(P_t) + f^T(t) + f^C(t) + f^S(t) + f^D(t) + f^H(t) + f^R(t)$$  \hspace{1cm} (4.147)

where the different terms are:

1. $g(C_i)$ is a transformation of the original data usually to normalize them and/or to eliminate variation in the variance in the error term. The most popular forms for this function are: (1) $g(C_i) = C_i$ and (2) $g(C_i) = \log(C_i)$. Note that other transformations could also be used.

2. $C_0$ is a constant or intercept.

3. The term $f^p(P_t)$ describes the relationship between ionic concentration and precipitation amount. It is usually written as $f^p(P_t) = C^p P_t$, $f^p(P_t) = C^p \sqrt{P_t}$ or $C^p/\sqrt{P_t}$ when $g(C_i) = C_i$ and $f^p(P_t) = C^p \log(P_t)$ when $g(C_i) = \log(C_i)$. Note that $f^p(P_t) = 0$ for other types of data than ionic concentrations in precipitation.

4. $f^T(t)$ describes the long-term variations or long-term trend. It will usually be written as a combination of a polynomial of order $N_T$ and a series of $M_T$ trigonometric functions with long (i.e., more than 6 to 10 years) periods, $\lambda_i$ (in years). In most cases, the best fit will be obtained using only a polynomial or a series of trigonometric functions. The general form of $f^T(t)$ can be written as:

$$f^T(t) = C_1^T t + C_2^T t^2 + \cdots + C_{N_T}^T t^{N_T} + \sum_{m=1}^{M_T} C_m^T \sin\left(\frac{2\pi t}{T\lambda_m}\right) + C_m^T \cos\left(\frac{2\pi t}{T\lambda_m}\right)$$  \hspace{1cm} (4.148)
where $T$ is equal to $24 \times 365.25$ for hourly data, $365.25$ for daily data, and so on, and $t$ is in hours from some original time before or at the beginning of the time series.

5. The function $f^C(t)$ describes the possible long-term cycles with periods greater than one year but less than about half the length of the time series (so they can be identified). Generally, $f^C(t)$ is written as a sum of $M_C$ trigonometric terms with periods equal to $\omega_i$ (in years) as follows:

$$f^C(t) = \sum_{n=1}^{M_C} \left[ C_{m}^{Cn} \sin \left( \frac{2\pi t}{T \omega_m} \right) + C_{m}^{Cn} \cos \left( \frac{2\pi t}{T \omega_m} \right) \right]$$  \hspace{1cm} (4.149)

6. The seasonal cycle is described by the term $f^S(t)$ and is usually approximated by a series of $M_S$ trigonometric functions as follows:

$$f^S(t) = \sum_{n=1}^{M_S} \left[ C_{m}^{Sn} \sin \left( \frac{2\pi m t}{T} \right) + C_{m}^{Sn} \cos \left( \frac{2\pi m t}{T} \right) \right]$$  \hspace{1cm} (4.150)

7. The term $f^D(t)$ describes the daily variations and can usually be approximated by a series of $M_D$ trigonometric functions with frequencies, $\tau_i$ (in days), as:

$$f^D(t) = \sum_{n=1}^{M_D} \left[ C_{m}^{Dn} \sin \left( \frac{2\pi m t}{24 \tau_m} \right) + C_{m}^{Dn} \cos \left( \frac{2\pi m t}{24 \tau_m} \right) \right]$$  \hspace{1cm} (4.151)

8. $f^H(t)$ is the diurnal cycle, described using a series of $M_H$ trigonometric functions as:

$$f^H(t) = \sum_{n=1}^{M_H} \left[ C_{m}^{Hn} \sin \left( \frac{2\pi m t}{24} \right) + C_{m}^{Hn} \cos \left( \frac{2\pi m t}{24} \right) \right]$$  \hspace{1cm} (4.152)

9. The last term in relation 4.147, $f^R(t)$, is the random error term. If the errors are assumed to be independent, this term will be equal to $f^R(t) = \varepsilon_i$, where the $\varepsilon_i$ are independent random errors from a normal distribution with mean zero and constant variance $\sigma^2$. If the errors are assumed to be correlated, an auto-regressive model of order $K$ is used to describe the error term. Its general form will be:

$$\varepsilon_i = a_1 \varepsilon_{i-1} + a_2 \varepsilon_{i-2} + \cdots + a_K \varepsilon_{i-K} + \eta_i$$  \hspace{1cm} (4.153)

where the $\eta_i$ are independent random errors from a normal distribution with mean zero and constant variance $\sigma^2$.

Note that for data with sampling periods greater than an hour, only the terms that are applicable will be used. For example, for daily data the term $f^D(t)$ and $f^H(t)$ will not be used. If the time series covers less than a year, the term $f^C(t)$ and $f^S(t)$ will not usually be used. In summary, this general model has to be adapted for the data under study.
The model given by relation 4.147 is the most general model and includes as special cases all the others mentioned earlier in this section. We will discuss how that model is fitted to the data later, but we must first mentioned how the parameters \( N_T, M_T, M_C, M_S, M_D, M_H \) and \( K \) can be selected. These parameters have to be chosen before trying to fit the model to the data. On the one hand, they can be selected when writing down the model and not modified later. On the other hand, as in Sirois (1993, 1997), many combinations of the parameters can be tried and the one for which the fit of the model to the data is the best would be selected. The latter method requires more work and time but reduces the number of a priori assumptions about the shape of the different temporal variations present in the time series.

One may often want to put some of the parameters (e.g., \( C_0, C_i^T, C_i^{Ts}, C_i^{Te}, C_i^{Ss}, \ldots \)) of the model equal to zero before trying to fit the model. This can be accomplished easily as we will see later. This permits more flexibility in the construction of the model.

When these arbitrary parameters have been selected, it is possible to fit the model to the observed data. The method used depends on whether the errors are assumed to be independent or serially correlated. In the former case, either the least squares technique or the maximum likelihood technique can be used, and both give the same solution. In the latter case, two different techniques can be used. We will describe them later. In all cases the first step in fitting the model to the data is to write down four matrices using the observations. We will summarize this process first before discussing the different techniques to fit the model.

If we assume that \( n \) observations, \( C_i \) (\( i = 1, 2, \ldots, n \)) are available at time \( t_i \) (\( i = 1, 2, \ldots, n \)), we can write the \((n \times 1)\) observation matrix, \( Y \), as:

\[
Y = \begin{pmatrix}
g(C_{t_1}) \\
g(C_{t_2}) \\
\vdots \\
g(C_{t_n}) 
\end{pmatrix}
\]

(4.154)

If we then write

\[
n_p = 1 + N_T + 2M_T + 2M_C + 2M_S + 2M_D + 2M_H - \text{(number of parameters put equal to zero)}
\]

(4.155)

the \((n \times 1)\) parameter matrix, \( b \), is:

\[
b = \begin{pmatrix}
C_0 \\
C_1^T \\
C_2^T \\
\vdots \\
C_1^{Ss} \\
\vdots \\
C_{M_H}^{He}
\end{pmatrix}
\]

(4.156)
Only the parameters not equalized to zero appear in \( b \). Finally, we defined the \((n \times n_p)\) matrix \( X \) as:

\[
X = \begin{pmatrix}
1 & t_1 & t_1^2 & \cdots & \sin\left(\frac{2\pi t_1}{T}\right) & \cos\left(\frac{2\pi t_1}{T}\right) & \cdots & \cos\left(\frac{2\pi M s t_1}{T}\right) & \cdots & \cos\left(\frac{2\pi M s t_1}{24}\right) \\
1 & t_2 & t_2^2 & \cdots & \sin\left(\frac{2\pi t_2}{T}\right) & \cos\left(\frac{2\pi t_2}{T}\right) & \cdots & \cos\left(\frac{2\pi M s t_2}{T}\right) & \cdots & \cos\left(\frac{2\pi M s t_2}{24}\right) \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
1 & t_n & t_n^2 & \cdots & \sin\left(\frac{2\pi t_n}{T}\right) & \cos\left(\frac{2\pi t_n}{T}\right) & \cdots & \cos\left(\frac{2\pi M s t_n}{T}\right) & \cdots & \cos\left(\frac{2\pi M s t_n}{24}\right)
\end{pmatrix}
\]

(4.157)

Note that the function of \( t \) associated with the parameters put equal to zero does not appear in the matrix \( X \). It is important that the matrices \( b \) and \( X \) be consistent. If one also writes the following \((n \times 1)\) matrix \( E \) as

\[
E = \begin{pmatrix} f^R(t_1) \\ f^R(t_2) \\ \vdots \\ f^R(t_n) \end{pmatrix}
\]

(4.158)

relation 4.147 can be written as a matrix equation:

\[
Y = X b + E
\]

(4.159)

To help the reader understand how these matrices are obtained, we will present two examples. The first one will be for the model described by relation 4.145. The different matrices are:

\[
Y = \begin{pmatrix} \log(C_1) \\ \log(C_2) \\ \vdots \\ \log(C_n) \end{pmatrix}, \quad b = \begin{pmatrix} \alpha \\ \delta \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad E = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}
\]

(4.160)

and

\[
X = \begin{pmatrix}
1 & \log(P_{n_1}) & t_1 & \sin(2\pi t_1) & \cos(2\pi t_1) \\
1 & \log(P_{n_2}) & t_2 & \sin(2\pi t_2) & \cos(2\pi t_2) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \log(P_{n_n}) & t_n & \sin(2\pi t_n) & \cos(2\pi t_n)
\end{pmatrix}
\]

(4.161)
As a second example, we will use the following more complex model for daily data (i.e., \( T = 365.25 \)):

\[
\log(C_t) = C_0 + C^P \log(P_t) + C_1^r t + C_3^r t^3 + C_1^{ts} \sin\left(\frac{2\pi t}{15T}\right) \\
+ C_1^{ce} \cos\left(\frac{2\pi t}{3T}\right) + C_1^{se} \sin\left(\frac{2\pi t}{T}\right) + C_2^{sc} \cos\left(\frac{4\pi t}{T}\right) + \varepsilon_t
\]  

(4.162)

The four matrices for that model are:

\[
Y = \begin{pmatrix}
\log(C_{t_1}) \\
\log(C_{t_2}) \\
\vdots \\
\log(C_{t_n})
\end{pmatrix}, \quad b = \begin{pmatrix}
C_0 \\
C^P \\
C_1^r \\
C_3^r \\
C_1^{ts} \\
C_1^{ce} \\
C_1^{se} \\
C_2^{sc}
\end{pmatrix}, \quad E = \begin{pmatrix}
\varepsilon_{t_1} \\
\varepsilon_{t_2} \\
\vdots \\
\varepsilon_{t_n}
\end{pmatrix}
\]  

(4.163)

and

\[
X = \begin{pmatrix}
1 & \log(P_{t_1}) & t_1 & t_1^3 & \sin\left(\frac{2\pi t_1}{15T}\right) & \cos\left(\frac{2\pi t_1}{15T}\right) & \frac{2\pi t_1}{3T} & \sin\left(\frac{2\pi t_1}{3T}\right) & \cos\left(\frac{4\pi t_1}{3T}\right) \\
1 & \log(P_{t_2}) & t_2 & t_2^3 & \sin\left(\frac{2\pi t_2}{15T}\right) & \cos\left(\frac{2\pi t_2}{15T}\right) & \frac{2\pi t_2}{3T} & \sin\left(\frac{2\pi t_2}{3T}\right) & \cos\left(\frac{4\pi t_2}{3T}\right) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & \log(P_{t_n}) & t_n & t_n^3 & \sin\left(\frac{2\pi t_n}{15T}\right) & \cos\left(\frac{2\pi t_n}{15T}\right) & \frac{2\pi t_n}{3T} & \sin\left(\frac{2\pi t_n}{3T}\right) & \cos\left(\frac{4\pi t_n}{3T}\right)
\end{pmatrix}
\]  

(4.164)

4.6.2 Without Auto-correlation in the Residuals

If the residuals are assumed to be independent random samples from a normal distribution with mean zero and constant variance \( \sigma^2 \), the least squares estimate of \( b \), \( \hat{b} \), is:

\[
\hat{b} = (X'X)^{-1}X'Y
\]  

(4.165)

where \( X' \) is the transpose of \( X \) and \( A^{-1} \) is the inverse of \( A \) (see Neter and Wasserman, 1974; Draper and Smith, 1981; Johnson and Wichern, 1982). An estimate of the variance \( \sigma^2 \), \( \hat{\sigma}^2 \), is:

\[
\hat{\sigma}^2 = \frac{1}{n-n_p}(Y - X\hat{b})'(Y - X\hat{b})
\]  

(4.166)

Note that the maximum likelihood estimate for these two parameters is also given by 4.165 and 4.166.
Figure 4.28: Estimated long-term trend for ionic concentration of SO$_4^{2-}$ in precipitation at Chalk River, Ontario

The estimate $\hat{b}$ is distributed following a multivariate distribution with mean, $\beta$ (i.e., the true values for the regression parameter), and standard deviation $\sigma^2(X'X)^{-1}$. It is therefore possible to test the statistical significance of the regression parameters or to create confidence intervals for their true values (see Neter and Wasserman, 1974; Draper and Smith, 1981; Johnson and Wichern, 1982).

As an example of the use of least squares regression where the residuals are assumed to be independent, we will present the results of such an analysis for the concentration of SO$_4^{2-}$ in precipitation at Chalk River, Ontario. The time series is shown in Figure 4.24. The following model was fitted to the data using the least squares technique:

$$\log(C_t) = C_0 + C_1^P \log(P_t) + C_1^T t + C_3^T t^3$$
$$+ C_1^SC \cos\left(\frac{2\pi t}{365.25}\right) + C_2^SC \cos\left(\frac{4\pi t}{365.25}\right) + C_2^SC \sin\left(\frac{4\pi t}{365.25}\right) + \varepsilon_t$$

(4.167)

Although this model may not be the best one for these data (as no efforts were made to find the best model for the present example), the long-term trend and the seasonal cycle should be close to the results that the best model would produce. The long-term trend is shown in Figure 4.28 and the seasonal cycle in Figure 4.29. The model explains about 28% of the total variance, with 2% explained by the relationship between ionic concentration and the precipitation amount, 10% by the long-term trend, and 16% by the seasonal cycle. A visual study of Figures 4.28 and 4.29 suggests that the fit between temporal model and observations is very good. Figure 4.28 indicates that most of the decrease occurred before 1990. Fitting a straight line to these data would give the wrong impression of the long-term variation in the data, especially for the last few years, because the mean ionic concentrations have not changed much in the last five years of the sampling period. The use of a simple sine/cosine function to describe the seasonal cycle would also be misleading, although it would be less so than the use of a straight line in the case of the long term trend (see Figure 4.28). If one considers the residuals, one will notice that serial correlations exist between them. One should therefore use the techniques described in the next section.
4.6.3 With Auto-correlation in the Residuals

Serial correlations or auto-correlations in the residuals result in part from the fact that some important temporal variations in the data are not included in the model used to describe the temporal variation in the observations. This can be seen if one fits a series of models of increasing complexity to the data. If one examines the auto-correlations after fitting each model, one will notice that they decrease after the inclusion of significant temporal variations in the model. In summary, part of auto-correlations in the residuals reflects the fact that not all significant temporal variations present in the observations have been explained by the model.

The main effect of the presence of auto-correlations in the residuals of the least squares estimates of the temporal model is either an under- or an over-estimation of the variance of the coefficients of the model, depending on the exact nature of the auto-corrrelation. This has an impact on the tests of statistical significance of these coefficients. It is important to note that the least-squares or maximum likelihood estimates of the regression coefficients are still unbiased. However, as mentioned earlier, the estimation of the variance can be quite inaccurate, thus rendering any test of significance meaningless.

It is therefore important to take auto-correlations into account. One way is to develop a comprehensive model for the temporal variation in the data. However, it is not always possible to do this. It is therefore necessary to find techniques to fit models described by relation 4.147 with a random term $f^R(t)$ as defined by relation 4.153 if we are to include the presence of serial-correlations in the data.

We will discuss two types of techniques that may be used to fit such a model. In the first, an ordinary least squares estimate is obtained. This is followed by fitting an auto-regressive model to the residuals and then, after correcting the data for the auto-regressive model, making a new least squares estimation of the regression coefficient. This process can be done iteratively until the coefficients of the temporal variation model and the auto-regressive model do not change. In the second technique, one writes for the observed data the maximum likelihood for the regression model with auto-correlations in the residuals and then
maximizes it. We will discuss these two types of techniques in more details in the remaining pages of this section.

4.6.3.1 Iterative Regression Method

We will start by discussing the simplest case of an auto-regressive model of order one. Relation 4.153 becomes:

\[ \epsilon_i = \phi \epsilon_{i-1} + \eta_i \]  \hspace{1cm} (4.168)

Let relation 4.147 be written as:

\[ Y = Xb + E \]  \hspace{1cm} (4.169)

where \( Y, b \) and \( X \) are defined in relations 4.154, 4.156, and 4.157 respectively. The vector \( E \) is defined as:

\[
E = \begin{pmatrix}
\epsilon_{t_1} \\
\epsilon_{t_2} \\
\vdots \\
\epsilon_{t_n}
\end{pmatrix}
\]  \hspace{1cm} (4.170)

with the \( \epsilon_i \) defined by relation 4.168.

If \( \phi \) is known, the generalized least squares estimates of \( b, \hat{b} \), is equal to (see Tiao et al., 1990; Seber and Wild, 1989):

\[ \hat{b} = (X'V^{-1}_\phi X)^{-1} X'V^{-1}_\phi Y \]  \hspace{1cm} (4.171)

with:

\[ V^{-1}_\phi = \frac{1}{1 - \phi^2} R_\phi R_\phi \]  \hspace{1cm} (4.172)

where \( R_\phi \) is a \((n \times n)\) matrix given by:

\[
R_\phi = \begin{pmatrix}
\sqrt{1 - \phi^2} & 0 & 0 & \cdots & 0 & 0 \\
-\phi & 1 & 0 & \cdots & 0 & 0 \\
0 & -\phi & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -\phi & 1
\end{pmatrix}
\]  \hspace{1cm} (4.173)

The covariance matrix of the \( \hat{b} \)'s is equal to

\[ \text{cov}(\hat{b}) = \frac{\sigma^2_\eta}{1 - \phi^2} (X'V^{-1}_\phi X)^{-1} \]  \hspace{1cm} (4.174)

where \( \sigma^2_\eta \) is the variance of the \( \eta_i \) in relation 4.168.
It is now necessary to estimate the auto-correlation coefficient. If the coefficient vector $\mathbf{b}$ is known, $\phi$ could be estimated using the following relation
\[
\hat{\phi} = \frac{\sum_{i=2}^{n} y_{i-1}^* y_i^*}{\sum_{i=2}^{n} (y_{i-1}^*)^2}
\] (4.175)

where $\mathbf{Y}^* = \mathbf{Y} - \mathbf{X}\mathbf{b}$, if the observations are distributed uniformly and there are no missing data. If there are observations missing, the technique developed by Dunsmuir (1984) could be used (see Appendix II). An estimation of $\phi$ is given by:
\[
\hat{\phi} = \frac{c_1^a}{c_0^a}
\] (4.176)

where the $c_i^a$ are calculated using relations A5 to A10 in Appendix II.

One can therefore see that to estimate the vector $\mathbf{b}$, one needs to know $\phi$, and to estimate the latter one needs to know the former. This problem can be solved by the following iterative procedure.

1. Assuming that $\phi = 0$, obtain a first estimation for $\mathbf{b}$ using ordinary least squares.
2. Calculate the residuals, i.e., $\mathbf{Y}^* = \mathbf{Y} - \mathbf{X}\mathbf{b}$.
3. Calculate an estimate of $\phi$ using relation 4.175 or 4.176.
4. Obtain a new estimate of $\mathbf{b}$ using the generalized least squares in relation 4.171.
5. Go back to step 2.
6. Loop through steps 2 to 5 until the values obtained for $\mathbf{b}$ and $\phi$ do not change.

Other techniques for solving that problem are presented in Seber and Wild (1989).

Seber and Wild (1989) present extensions of these techniques to auto-regressive models of order K. The reader is referred to this reference for more details. An example of the use of this technique can be found in Li et al. (1993).

4.6.3.2 Maximum Likelihood Method

The model to fit to the data can be written as:
\[
\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{E}
\] (4.177)

where $\mathbf{E}$ is defined by relation 4.170 with the $\varepsilon_i$ following an auto-regressive model of order K given by:
\[
\varepsilon_i = a_1 \varepsilon_{i-1} + a_2 \varepsilon_{i-2} + \cdots + a_K \varepsilon_{i-K} + \eta_i.
\] (4.178)

If n observations for y are available and the random errors $\eta_i$ are assumed to be independent and drawn from a normal distribution with zero mean and constant variance $\sigma^2$, the full log-likelihood function for this model is:
\[
\log \{L(y; a, b, \sigma^2)\} = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} n \log(\sigma^2) - \frac{1}{2} \log(|V|) - \frac{1}{2} \sigma^2 (Y - Xb)' V^{-1} (Y - Xb)
\]

(4.179)

where the \((n \times n)\) matrix \(V\) is defined as \(E(\varepsilon \varepsilon') = \sigma^2 V\) and thus \(V\) is a function of the parameters \(\alpha_i\). \(E(.)\) indicates the expected value of (.). The problem is to find the values of \(a, b\) and \(\sigma^2\) that maximize the log-likelihood function. The easiest way to solve this problem is to rewrite relation 4.176, using the state space formulation of the time series model (Fuller, 1996). As the equations become quite complex, we will not go into the details here but refer the interested reader to Harvey and Philips (1979) and Kohn and Ansley (1985) for a complete description. If there are no missing data, the algorithm of Ansley (1979) can be used to maximize the likelihood. If some of the observations are missing, one has to use the Kalman filter (see Kohn and Ansley, 1986).

As we have mentioned, the errors, \(\eta_i\), are assumed to be normally distributed. However, their distribution can deviate from normality when the amount of data is large. In that case, the technique gives unbiased estimates of the model parameters. This results from the central limit theorem (see Silvey, 1975). For large \(n\), it can be shown (Silvey, 1975) that the random vector \(\theta\) (i.e., \(\theta^t = (a, b, \sigma^2)\)) follows a multidimensional normal distribution with mean \(\theta\) and covariance \(\Theta\) equal to:

\[
\{\Theta\}_{ij} = \left\{ \frac{\partial^2 \log(L)}{\partial \theta_i \partial \theta_j} \right\}_{ij}^{-1}
\]

(4.180)

even if the distribution of the \(\eta_i\) is not normal. \(\theta^*\) is the vector for the parameters of the model that maximizes the log-likelihood. If one wants to test or create confidence intervals for more than one parameter, simultaneous inferences techniques, like Bonferroni's multiple testing technique, should be used. See Miller (1981) or Hochberg and Tamhane (1987) for more details.

Examples of the use of this technique can be found in Sirois (1997) and Sirois and

![Figure 4.30 Estimated long-term trend for ionic concentration of SO₄²⁻ in precipitation at Chalk River, Ontario](image)

86
Barrie (1999). Note that the S-Plus<sup>®</sup> function “arima.mle” which implements the method of Kohn and Ansley (1986) was used in these examples.

To illustrate this method, we will return to the example presented in Section 4.6.2. The model described by relation 4.167 was fitted to the data, but this time, the residuals were assumed to follow an auto-regressive model of order 1. The new model explained 33.2% of the total variance. The relative contribution of the relationship with the precipitation amount, of the long-term trend, and of the seasonal cycle did not change noticeably. All of the increase in the percentage of the total variance explained by the model is due to the inclusion of the auto-regressive model. The auto-regression coefficient is equal to 0.33. The long-term trend is shown in Figure 4.30 and the seasonal cycles in Figure 4.31. As we can easily see, there are no visible differences in this example between the results with the auto-regressive model and those without. Note that the auto-correlation was not very large. The coefficients in relation 4.167 have varied by less than 11%. Note also that the most important changes occurred in the seasonal cycles. We can therefore conclude that the presence of auto-correlations in the residuals did not strongly influence the estimations of the parameters.

4.6.3 Discussion

The techniques presented in this section are the most versatile and complete of all the techniques presented in this overview. They are versatile because any polynomial or trigonometric function can be used to describe temporal variations on any time scale. They are complete because the different components contributing to the total variation can be separated and modeled in the same general model. The only extra limiting assumption, compared with other techniques, is that the residuals should be normally distributed, and this is not as limiting as it seems when the number of data items is large. The other assumptions (i.e., independence of the residuals and constant variance) are common to all the techniques presented in this overview.
The only important drawbacks to the use of the techniques presented here are their complexity if serial correlations are present and the fact that the search for the best model can be labor intensive. In the case of this last restriction, the use of smoothers and spectral analysis can usually help accelerate the process of model selection. In the case of the first drawback, the use of statistical computer packages such as S-Plus® that includes programs that implement some or most of the techniques presented here is usually helpful.

One important consideration to keep in mind is that the presence of small serial correlations in the data will not usually have much of an influence on the estimation of the long-term trend or seasonal cycles if the number of observations is large. However, their presence will cause temporal variations that are not very strong to be considered not statistically significant. For example, if only a very weak long-term trend is present in the data, it may be rejected because of the presence of serial correlations. If these are taken into account, the long-term trend will then be judged significant. If the long-term trend is a strong one, however, the presence of serial correlations will only have a slight effect on the estimation of its shape. In conclusion, it is not always necessary to use models that include serial correlations but it is usually the cautious thing to do.

As we mentioned in the introduction of this section, it has been customary in the past to assume a priori that the long-term trend follows a straight line. That, in the opinion of the author, is a practice that should be avoided. Unless there are good physical reasons to assume any particular shape for the long-term trend, the reader should let the data tell him what the real shape of the long-term trend is. In all his work on time series analysis in atmospheric chemistry, the author has found only a few cases for which the straight line model was the appropriate one to describe the long-term variation. It is therefore recommended that other techniques such as smoothers be used to test the possible shape of the long-term trend before accepting any a priori conclusions. This remark also applies to any of the other temporal variations present in the data.

In this section, we have described techniques for including auto-regressive processes in the temporal model. These techniques can be easily extended to include MA (moving average) and ARMA (auto-regressive moving-average) models for the residuals. The reader is referred to Chatfield (1984) or Brockwell and Davis (1987) for details on these types of models. Descriptions of the techniques to use in these cases are given in Seber and Wild (1989) and Kohn and Ansley (1985, 1986).
5. Verification of the Hypothesis behind the Techniques Used

5.1 Introduction

In Section 4, we presented a number of techniques for estimating and/or testing the different temporal variations that can be present in a time series. In all of these techniques some hypotheses were made, either about the data themselves or about the residuals after the elimination of some estimated model. It is therefore important after fitting a model or testing one for statistical significance to verify whether these hypotheses have been fulfilled.

The first thing to check is the validity of the assumed model. For example, if one assumes a straight line model for the long-term trend, the residuals will not show any long-term trend if the model is correct. It is therefore important to verify that such is the case. This applies to all the components of the temporal variations that are assumed to be present in the time series.

The most common hypothesis made in the techniques presented in Section 4 is that the residuals are independent. It is therefore necessary to test that hypothesis. We will present different methods here that can be used to test for independence. It is also usually assumed that the variance of the distribution of the residuals is constant with time. We will therefore discuss how to determine if that hypothesis has been fulfilled.

In some of the techniques that we have discussed, it is assumed that the data or the residuals are normally distributed. Many techniques exist to test whether this is the case. We will summarize some of these tests here.

It is important not only to assess whether the hypotheses are fulfilled but also to determine the importance of the deviations from them when they are not fulfilled. As we have mentioned often in Section 4, some of the techniques are still approximately valid even if some of the hypotheses are not respected, as long as the deviation from them is not too large. It is therefore necessary to have some idea of the importance of any possible deviation.

Although verification of the hypotheses may seems unnecessary to some readers, this procedure is a very important part of any time series analysis. Not doing it is like walking blindfolded into an unknown place. We thus recommend that the reader always verifies the validity of the model and/or techniques used by employing the techniques presented here or others that are appropriate.

5.2 Validity of the Model

In some of the techniques presented earlier, the shape of the long-term trend, seasonal cycles, or other temporal variations has to be specified before the model is fitted. We have mentioned that it has been popular in the past to assume a straight line for the long-term trend and a simple trigonometric function for the seasonal cycle. The most current model to describe the temporal variation in the observed ionic concentrations is given by relation 4.145 or other similar models. If that model is the right one, the mean of the residuals (i.e., the \( e_i \)'s in relation 4.145) should be constant and equal to zero. Therefore, they should not present any indication of long-term trends, or seasonal cycles, or other temporal variations. One easy way to verify that this condition has been met is to plot the residual time series and use one of the smoothing techniques presented in Section 4.3 to estimate the temporal variations, if there are any.
Figure 5.1 Residual after eliminating temporal model described by relation 4.145 from the observed SO$_4^{2-}$ ionic concentrations in precipitation at Chalk River. The solid line is the estimated long-term trend using a kernel smoother.

To illustrate the technique, we have fitted the model given by relation 4.145 to the SO$_4^{2-}$ ionic concentrations in precipitation at Chalk River. We have seen in Section 4.6.2 that a good model is the one given in relation 4.167, with the $\varepsilon_i$'s following an auto-regressive model of order one.

The long-term trend in relation 4.167 is a third order polynomial. Thus, the residuals should present a long-term trend if a straight line is used to model it. This is illustrated in Figure 5.1, which presents the residuals and a kernel smoother estimate of the long-term trend of the residuals. As we can see, a noticeable long-term trend is still present in the data. The model underestimates the concentrations before 1983, overestimates them between 1984 and 1990, and is close to them after that date. This is not the case if a third order polynomial is used to describe the long-term trend as illustrated in Figure 5.3. Unlike Figure 5.1, the estimated long-term trend does not show any large variations and can therefore be assumed to be constant and equal to zero.

We will now consider the seasonal cycle. Figure 5.2 presents the seasonal variations of the residuals after eliminating the model in relation 4.145. The solid line presents the estimated seasonal cycle using a periodic kernel.

Figure 5.2 Residuals after eliminating temporal model described by relation 4.145 from the observed SO$_4^{2-}$ ionic concentrations in precipitation at Chalk River. The solid line shows the estimated seasonal cycle using a kernel smoother.
Figure 5.3 Residuals after eliminating temporal model described by relation 4.167 from the observed $\text{SO}_4^{2-}$ ionic concentrations in precipitation at Chalk River. The solid line is the estimated long-term trend using a kernel smoother.

smoother. It is easy to see that the simple trigonometric functions used by the model do not completely describe the seasonal cycle. If one now uses the seasonal cycle given in relation 4.167, one sees that the residuals do not present any seasonal cycle (Figure 5.4).

This example illustrates the importance of checking the validity of the assumed model. The reader should keep in mind that temporal variations that are not explained will be translated into serial correlations in the residuals that may invalidate any test of significance that is made. It is therefore essential that the model used to describe the temporal variations present in the data should be as comprehensive as possible.

5.3 Independence of the Residuals

As mentioned earlier one of the most universal assumptions made in time series analysis techniques is that the data or the residuals are independent. It is therefore very important to be able to test whether the residuals are independent. Many techniques have been developed in the past to do this, and we will review some of them here.

5.3.1 Testing for Independence

Figure 5.4 Residuals after eliminating temporal model described by relation 4.167 from the observed $\text{SO}_4^{2-}$ ionic concentrations in precipitation at Chalk River. The solid line shows the estimated seasonal cycle using a kernel smoother.
We will discuss here two possible tests for determining whether the residuals are independent. For these tests it is assumed that no long-term trend, seasonal cycles, or other temporal variations are present in the data. The first test is nonparametric, and the second assumes that the residuals are normally distributed. In what follows, $e_i$ ($i = 1, 2, \ldots, n$) will denote the residuals (i.e., $e_i = y_i - (X\hat{b})_i$).

5.3.1.1 Turning Points Test

The turning points test (also known as the peaks and troughs test) is one of the easiest tests of randomness to apply (see Kendall, 1976; Kendall et al., 1983). The idea behind this test is to count the number of peaks or troughs that the time series exhibits. A peak is defined as a value greater than its two neighbouring values and a trough as a value less than its two neighbouring values.

Define $X_i$ ($i = 1, 2, \ldots, n-2$) as:

$$X_i = \begin{cases} 1 & \text{if } e_i < e_{i+1} > e_{i+2} \text{ or } e_i > e_{i+1} < e_{i+2} \\ 0 & \text{otherwise} \end{cases} \quad (5.1)$$

The number of turning points (i.e., the sum of peaks and troughs) $p$ in the time series is:

$$p = \sum_{i=1}^{n-2} X_i \quad (5.2)$$

and we have:

$$E(p) = \frac{2}{3} (n-2) \quad (5.3)$$

and

$$\text{var}(p) = \frac{16n - 29}{90} \quad (5.4)$$

If $n$ is large, the variable $P$ is equal to:

$$P = \frac{p - E(p)}{\sqrt{\text{var}(p)}} \quad (5.5)$$

and follows a normal distribution with mean 0 and variance 1. If $P$ is statistically significantly different from zero, the residuals are not independent. A small value for $p$ (i.e., $P$ negative) indicates that the successive values are positively correlated. A large value for $p$ (i.e., $P$ positive) indicates rapid non-random oscillations in the time series. Note that no missing values are allowed. It is, however, possible to calculate $P_i$ independently for section $i$ of the time series without missing values. Assuming that these series are independent and that the length of each series is not too small, the variable $P^T = \sum_{i=1}^{n_s} P_i$ is normally distributed with mean zero and standard deviation equal to $1/\sqrt{n_s}$ where $n_s$ is the number of series.

5.3.1.2 Durbin-Watson Test

In this test, one calculates the statistic,
\[
D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}
\]

(5.6)

An exact test procedure is not available, but Durbin and Watson (1950, 1951) have obtained lower and upper bounds \(d_L\) and \(d_U\) such that a value of \(D\) outside these bounds leads to a definite decision. The decision rule for a test at level \(\alpha\), is as follows:

1. **One-sided test against alternatives, the auto-correlation coefficient, \(\rho\), greater than 0:**
   - If \(D > d_U\), conclude that \(\rho\) is not statistically different from zero.
   - If \(D < d_L\), conclude that \(\rho > 0\).
   - If \(d_L \leq D \leq d_U\), the test is inconclusive.

2. **One-sided test against alternatives, the auto-correlation coefficient, \(\rho\), less than 0:**
   - Repeat (1) using \((4 - D)\) in place of \(D\).

3. **Two-sided test against alternatives, the auto-correlation coefficient, \(\rho\), not equal to 0:**
   - If \(D < d_L\), or \(4 - D < d_L\), conclude that \(D\) is significant and reject the hypothesis of no auto-correlation at level 2\(\alpha\).
   - If \(D > d_U\), conclude that \(D\) is not significant and do not reject the hypothesis of no auto-correlation at level 2\(\alpha\).
   - If \(D > d_U\) and \(4 - D > d_U\), conclude that \(D\) is not significant and do not reject the hypothesis of no auto-correlation at level 2\(\alpha\).
   - Otherwise, the test is said to be inconclusive.

(5.7)

It has been discovered that, in many situations, treating the test as though \(d_L\) did not exist and \(d_U\) were the appropriate single critical value provides a very good approximation to the truth. A simplified approximate test procedure is as follows:

1. **Simplified one-sided test against alternatives \(\rho > 0\):**
   - If \(D < d_U\), conclude that \(\rho > 0\).
   - Otherwise, do not reject that \(\rho = 0\).

2. **Simplified one-sided test against alternatives, \(\rho < 0\):**
   - If \(4 - D < d_U\), conclude that \(\rho > 0\).
   - Otherwise, do not reject that \(\rho = 0\).

3. **Two-sided test against alternatives, \(\rho \neq 0\):**
   - If \(D < d_U\) or \(4 - D < d_U\), conclude that \(\rho \neq 0\) at level 2\(\alpha\).
Otherwise, do not reject that $\rho = 0.$

(5.8)

Tables of $d_U$ and $d_L$ for different numbers of $n$, for two levels of significance (0.05 and 0.01), and for various numbers of independent variables in the temporal model can be found in Neter and Wasserman (1974), Kendall (1976), or Draper and Smith (1981).

The present test assumes that there are no missing values. A variant of the test that takes the missing data into account was developed by Savin and White (1978). However, Bhargava (1989) has shown that there is no advantage in employing such a test. Using the usual Durbin-Watson test and ignoring the gaps in the data is just as effective. Savin and White (1978) have proposed other possible tests of independence for the residuals.

5.3.2 Plots of Auto-correlation and Integrated Periodogram

Many different types of graphics can be drawn to test the independence of the residuals visually or statistically. One of the simpler methods involves drawing scattergrams of $e_t$ versus $e_{t-1}$. We used these types of plots in Section 2.8 when trying to familiarize ourselves with the time series. This method is illustrated in Figure 5.5 for the residuals obtained by subtracting fitted models 4.145 or 4.167 from the observed $\text{SO}_4^{2-}$ ionic concentrations in precipitation at Chalk River. The straight lines represent least squares fits to the data. One can easily see that there are serial correlations in the residuals if model 4.145 is used but that this is not the case if model 4.167 (with a first order auto-regressive model to describe the error term) is used.

A second technique that can be used is the spectral analysis technique described in Section 4.4.2. We mentioned in that section that if the data are independent, their power spectrum should be a random variations around a straight line. A technique to test for

Figure 5.5 Scattergrams of $e_t$ versus $e_{t-1}$ for the residuals after fitting model 4.145(a) or 4.167(b) (with a first order auto-regressive model to model the error term) to the observed $\text{SO}_4^{2-}$ ionic concentrations in precipitation at Chalk River. The straight lines represent least squares fits to the data.
randomness was also presented in that section. The smoothed power spectra for the residuals after fitting model 4.145 or 4.167 (with a first order auto-regressive model to describe the error term) to the observed SO$_4^{2-}$ ionic concentrations in precipitation at Chalk River are shown in Figures 5.6a and 5.6b respectively. It is clear from the power spectrum that some part of the long-term trend and seasonal cycle remains if the model described by relation 4.145 is used. The presence of auto-correlations is also quite evident. However, if we use the model described by relation 4.167 with the error term following an auto-regressive model of order one, the smoothed power spectrum indicates that the residuals can be considered to be white noise or independent.

A simpler technique than using the smoothed power spectrum plots the auto-regression function between lag 1 and N where N is some predefined number. If there are no missing data in the time series, the auto-covariance coefficient at lag k is given by (Chatfield, 1984):

$$c_k = \frac{1}{n} \sum_{i=1}^{n-k} (e_i - \overline{e})(e_{i+k} - \overline{e})$$  \hspace{1cm} (5.9)

where $\overline{e}$ is the mean value of the residuals and therefore equal to zero. The auto-regression coefficient of lag k is equal to:

$$r_k = \frac{c_k}{c_0}$$ \hspace{1cm} (5.10)

If there are missing values, the technique developed by Dunmuir (1984) and presented in Appendix II can be used to estimate the auto-covariance function.

The auto-correlation coefficients have to be compared with the limits $\pm 1.96/\sqrt{n}$ (Chatfield, 1984).

Examples of the auto-correlation function for the residuals after fitting model 4.145 or 4.167 (with a first order auto-regressive model to model the error term) to the observed SO$_4^{2-}$ ionic concentrations in precipitation at Chalk River are presented in Figures 5.7a and 5.7b respectively. They were drawn up to a lag of 100. The auto-correlation function indicates the presence of a net auto-correlation in the case of the residuals of model 4.145 but not in the
case of the second model. This is consistent with the results obtained using the other techniques.

A final technique, which is related to the power spectrum technique, is the integrated or cumulative periodogram. It is defined as:

$$CP(\omega) = \frac{\int_{-\frac{\pi}{2}}^{\omega} f(\lambda) \, d\lambda}{\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f(\lambda) \, d\lambda} \quad 0 \leq \omega \leq 0.5 \quad (5.11)$$

where $f(\omega)$ is the power spectrum. For the integrated or cumulative periodogram plot, one draws CP(ω) versus ω. Notice that, by definition, CP(ω) varies between zero for ω = 0 and 1 for ω = 0.5. We have seen in Section 4.4.1 that the power spectrum of a random or white noise is a constant equal to $\sigma^2/\pi$ (relation 4.98b). Therefore, its cumulative periodogram is equal to:

$$CP(\omega) = 2\omega \quad (5.12)$$

The cumulative periodogram of random noise is therefore a straight line crossing the graph. As was mentioned in Section 4.4.2, the cumulative periodogram for a sample from a random noise would oscillate around that line. Confidence intervals for those oscillations can be obtained by using the Kolmogorov-Smirnov test, if the random noise is assumed to have been sampled from a normal distribution. The confidence band is defined by the following two lines (Brockwell and Davis, 1987):

$$\omega \pm \frac{k_\alpha}{\sqrt{n/2 - 1}} \quad (5.13)$$

where $k_{0.05} = 1.36$ and $k_{0.01} = 1.63$.

Figure 5.7 Auto-correlation function for the residuals after fitting model 4.145(a) or 4.167(b) (with a first order auto-regressive model to model the error term) to the observed SO4$^{2-}$ ionic concentrations in precipitation at Chalk River. The straight lines represent least squares fits to the data.

To illustrate the use of the integrated or cumulative periodogram, we will use the same data used to illustrate the other techniques presented in this section. Figure 5.8 presents the cumulative periodogram for the two sets of residuals. Relation 4.95 was used to estimate
the power spectrum. It is quite clear that the residuals of model 4.145 are not independent. As for the other set of residuals, although there is some deviations from independence, they are small and should not influence the validity of the model much.

We have reviewed four graphical techniques in this section that can be used to verify the independence of the residuals. In the example they all give the same results. Of the four techniques, the scattergram and plotting the auto-correlation function are the simplest to use. The smoothed power spectrum is more informative because it can indicate which types of temporal variation are present in the residuals. The cumulative periodogram can also be applied in this way, but it is more difficult to use.

5.4 Normality of the Residuals

In some of the techniques presented in Section 4, it was assumed that the residuals were normally distributed. It is therefore important when using these techniques to be able to assess whether that hypothesis has been fulfilled.

As many tests for normality exist, we will concentrate on the most frequently used, namely, the $W$ test of Shapiro and Wilk (1965), D’Agistino’s test (D’Agostino, 1971), the range test of David et al. (1954), and finally the Kolmogorov-Smirnov goodness of fit test (Kolmogorov, 1941; Smirnov, 1948). Other tests can be found in Sachs (1984) and Stephens (1970, 1974). The power of many of these tests is compared in this last paper.

Although it is useful to test for non-normality, it is sometimes more useful to know the importance of the deviation from normality. Normal and log-normal probability paper (or plots) are very useful in these cases. We will therefore described the uses of these tests here.
5.4.1 Testing for Normality

5.4.1.1 The \(W\) Test

This test was developed by Shapiro and Wilk (1965) to test the hypothesis:

\[ H_0: \text{The population has a normal distribution} \]  \hspace{1cm} (5.14)

against the hypothesis:

\[ H_1: \text{The population does not have a normal distribution} \] \hspace{1cm} (5.15)

It can therefore be used to test whether the residuals are from a normal distribution. Note that it can only be used for \(n \leq 50\). The test is as follows:

1. Calculate the parameter \(d\) given by:

\[
d = \sum_{i=1}^{n} (e_i - \overline{e})^2 = \sum_{i=1}^{n} e_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} e_i \right)^2 \] \hspace{1cm} (5.16)

2. Order the residuals in increasing order to obtain the order statistics \(e_{[1]} \leq e_{[2]} \leq \ldots \leq e_{[n]}\).

3. Compute \(k\) as:

\[
k = \begin{cases} 
\frac{n}{2} & \text{if } n \text{ is even} \\
\frac{n-1}{2} & \text{if } n \text{ is odd} 
\end{cases} \] \hspace{1cm} (5.17)

4. Turn to Table 5 in Shapiro and Wilk (1965) or Table A6 in Gilbert (1987) to get the coefficients \(a_1, a_2, \ldots, a_k\).

5. Compute \(W\), given by:

\[ W = \frac{1}{d} \left[ \sum_{i=1}^{k} a_i (e_{[n-i+1]} - e_{[i]}) \right]^2 \] \hspace{1cm} (5.18)

6. Reject \(H_0\) at the \(\alpha\) significance level if \(W\) is less than the quantile given in Table 6 of Shapiro and Wilk (1965) or Table A7 of Gilbert (1987).

This test can be used to test for log-normality when the data are all greater than zero by first using a logarithmic transformation to transform them.

5.4.1.2 D’Agostino’s Test

This test can be applied only if the number of samples is greater than or equal to 50. Thus, it may serve as a complement to the \(W\) test. D’Agostino’s test can be summarized as follows:

1. Order the residuals in increasing order to obtain the order statistics \(e_{[1]} \leq e_{[2]} \leq \ldots \leq e_{[n]}\).
2. Compute the statistic $D$ as:

$$D = \frac{\sum_{i=1}^{n} \left[ i - \frac{1}{2} (n + 1) \right] e_i}{n^2 s}$$  \hspace{1cm} (5.19)

where

$$s = \left[ \frac{1}{n} \sum_{i=1}^{n} (e_i - \bar{e})^2 \right]^{1/2}$$  \hspace{1cm} (5.20)

3. Calculate the statistic $Y$ given by:

$$Y = \frac{D - 0.28209479}{0.02998598 / \sqrt{n}}$$  \hspace{1cm} (5.21)

Note that one should aim for five-place numerical accuracy in computing $D$.

4. Reject at the $\alpha$ significance level the null hypothesis (relation 5.14) if $Y$ is less than the $\alpha/2$ quantile or greater than the $1 - \alpha/2$ quantile of the distribution of $Y$. These quantiles are given in D'Agostino (1971) and in Table A8 of Gilbert (1987).

5.4.1.3 Range Test of David et al. (1954)

A very rapid and simple method to test for normality is as follows.

1. Calculate the ratio:

$$\frac{\text{range}}{\text{standard deviation}} = \frac{R}{s}$$  \hspace{1cm} (5.22)

2. If the quotient does not lies between the tabulated critical values given in Table 3 of Pearson and Stephens (1964) or Table 72 of Sachs (1984), then the hypothesis of normality is rejected at the respective significance level.

It is quite apparent that this test is easy to use. Extensive tables for this procedure can be found in Pearson and Stephens (1964).

5.4.1.4 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test can be described by the following steps:

1. Estimate the mean and standard deviation:

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i$$  \hspace{1cm} (5.23)

and

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (e_i - \bar{e})^2}$$  \hspace{1cm} (5.24)
Note that in the case of the residuals $\bar{e} = 0$ by definition.

2. Order the residuals in increasing order to obtain the order statistics $e_{[1]} \leq e_{[2]} \leq \ldots \leq e_{[n]}$.

3. Calculate the variables $z_i$ equal to
   \[ z_i = \frac{e_{[i]} - \bar{e}}{s} \]  
   (5.25)

4. Evaluate the following statistics:
   \[ D^+ = \max((i/n) - z_i) \]  
   (5.26)
   \[ D^- = \max(z_i - (i - 1)/n) \]  
   (5.27)
   \[ D = \max(D^+, D^-) \]  
   (5.28)

5. Compare the value of $D$ with the critical values for the test (see Lilliefors, 1967, for tables). If $D > D_\alpha$, where $D_\alpha$ is the critical value for a test at significance level $\alpha$, the hypothesis that the sample is from a normal distribution can be rejected.

Note that $D_\alpha$ can be approximated by the relation (Stephens, 1974):
\[ D_\alpha \approx \frac{K_\alpha}{\sqrt{n - 0.01 + 0.85/\sqrt{n}}} \]  
(5.29)

where the $K_\alpha$ are equal to:

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0.15</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_\alpha$</td>
<td>0.775</td>
<td>0.819</td>
<td>0.895</td>
<td>0.955</td>
<td>1.035</td>
</tr>
</tbody>
</table>

The approximation becomes better as $n$ increases.

5.4.2 Normal Probability Plots

We have already mentioned normal probability plots and described some of their uses in Section 2.5. We have seen that they can be used to estimate the importance of any deviation from normality.

The normal probability plot is constructed by first ordering the observations so that $e_{[1]} \leq e_{[2]} \leq \ldots \leq e_{[n]}$ and then plotting the $e_{[i]}$ against $\Phi^{-1}\left(\frac{i - 0.5}{n}\right)$, where $\Phi$ is the cumulative distribution function for the standard normal distribution defined as:
\[ \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \, dt \]  
(5.30)
If the data are normally distributed, one will obtain a straight line. An $\alpha$ confidence band around that theoretical straight line, based on the Kolmogorov-Smirnov test, is given by the following two lines (see Michael, 1983):

$$y = \bar{e} + s \Phi^{-1}\left\{\Phi(x) \pm \left(D_\alpha - \frac{1}{2n}\right)\right\}$$

(5.31)

where $\bar{e}$ and $s$ are defined by relations 5.23 and 5.24 and relation 5.29 can be used to estimate $D_\alpha$. If the observed cumulative distribution gets out of this band, one can say that, with a level of confidence $\alpha$ the hypothesis that the sample is from a normal distribution can be rejected using the Kolmogorov-Smirnov test.

An example of such a plot is presented in Figure 5.9 for the residuals after fitting model 4.167 (with a first order auto-regressive model to model the error term) to the observed $SO_4^{2-}$ ionic concentrations in precipitation at Chalk River. One sees first that, using the Kolmogorov-Smirnov test, the hypothesis of normality can be rejected at both significance levels. However, one also sees that the deviation from normality is not extreme and therefore should not have much influence on the results of the analysis, as the number of observations available is large.

This simple example illustrates how useful these types of graphics can be to verify the normality of the distribution of sampled data.

### 5.5 Homoscedasticity of the Residuals

One of the basic hypotheses of all the techniques presented in Section 4 is that the variance of the residuals remains constant in time. It is therefore important to check if this hypothesis is fulfilled by the data studied.

The easiest way to check this hypothesis is to determine visually whether the spread in the residuals changes with time. If we consider the residuals in Figure 5.10, one notices that, apart from a few low values, the residuals seem to spread out equally around zero during all periods.

A better way to study the variation of the spread of the residuals with time is to consider the variations of the first and third quartiles with time. The difference between the two gives the variation of the range with time. This may help in getting a better idea of the variance changes with time. The variation of the quartiles may be estimated using a running quartile smoother. This estimator is similar to the running median smoother discussed in Section 4.3.3, except

![Figure 5.9 Normal probability distribution](image-url)
Figure 5.10 Residuals after fitting 4.167 (with a first order auto-regressive model to model the error term) to the observed $\text{SO}_4^{2-}$ ionic concentrations in precipitation at Chalk River

that the first or the third quartiles is estimated instead of the median. Figure 5.11 presents examples of the estimated running first and third quartiles for the residuals of Figure 5.10. These lines suggest that the variance may have increased slightly with time. The effect is very small, however, and can be neglected.

The running quartile smoother can also be used to verify that the distribution of the residuals is symmetric around zero. This is evidently true in the present example.

If we assume that the distribution of the residuals remains symmetric around zero, one

Figure 5.11 Same as Figure 5.8. The two lines are the running first and third quartile smoother estimates of the variations of those quartiles with time.
Figure 5.12 Absolute value of the residuals after fitting 4.167 (with a first order autoregressive model to model the error term) to the observed SO$_2^-$ ionic concentrations in precipitation at Chalk River. The line shows the running median smoother estimate of the variations of the median with time.

One way to test the statistical significance of this variation is to calculate either Spearman's correlation coefficient or Kendall's τ correlation coefficient between the absolute value of the residuals and time (see Hollander and Wolfe, 1973, for the definition of these correlation coefficients). Both correlation coefficients are nonparametric and based on rank. For the example studied in this section, Spearman's and Kendall's correlation coefficients are equal to 0.12 and 0.08 respectively, and both are significantly different from zero. This confirms the fact that a small increase in the variance is present in the residuals.

5.6 Discussion

In this section we have presented some of the techniques that can be used to verify the fulfillment of the hypothesis underlying various methods for analyzing the temporal variations of a time series. The reader should be aware that the list of techniques presented here is far from comprehensive and that many others exist.

The main goal of the author in this section was to illustrate the importance of testing the fulfillment of these hypotheses and indeed of knowing which hypotheses underlie the techniques one wants to use. Even if one accepts that some or all of these hypotheses will not
be fulfilled by the data at hand, it is still very important to have some idea of the effects of not conforming to them. These techniques can therefore be used to estimate if the techniques or models used are acceptable for a particular time series.
6. A Case Study: Particulate \( \text{SO}_4^{2-} \) Concentrations in Ambient Air at ELA, Ontario.

Although examples have been given of most of the techniques presented in this overview, it may be useful to present a complete example of the different steps involved in the study of a time series. We have selected as an example the variations of particulate \( \text{SO}_4^{2-} \) daily concentration in ambient air at the Experimental Lakes Area (ELA), Ontario. Because of lack of space and preparation time, we will not use all the techniques presented in this overview but rather a selection of the most important ones. Also, we will not try to interpret the results here in any detail.

The time series is presented in Figure 6.1. The first thing that one notices is that although some long-term trend may be present in the data no clear increase or decrease occurred during the sampling period. There is also some indications that a seasonal cycle is present in the time series. There is also evidence of missing data especially, in 1989. As Figure 6.2 illustrates, most of the missing periods are smaller than 3 days although there are a few very long ones.

The possible presence of a non-monotonic long-term trend is indicated by the annual box-plots presented in Figure 6.3. The presence of a marked seasonal cycle is confirmed by the monthly box-plots shown in Figure 6.4. The presence of both variations in the time series is confirmed by the smoothed power spectrum presented in Figure 6.5. The spectrum also shows that auto-correlations are present in the time series.

From Figure 6.6, we can conclude that a logarithmic transformation should be applied to the data before the analysis. One should, however, note that the distribution seems to deviate from the normal distribution (i.e., the observations do not follow a straight line). Part of that deviation could be due to the presence of the long-term trend, seasonal cycles, and

![Figure 6.1 Temporal variations of particulate \( \text{SO}_4^{2-} \) daily concentrations in ambient air at the Experimental Lakes Area, Ontario.](image-url)
possibly of other temporal variations. We will therefore assume that a logarithmic transformation will be appropriate for now.

A first approximation of the long-term trend was obtained using a Nadaraya-Watson kernel estimator. The result of the analysis is presented in Figure 6.7. Broadly speaking, the trend here is similar to the long-term trend obtained using the annual box-plots in Figure 6.3. After a decrease in the beginning of the 1980s, the concentration increased up to about the end of 1989 and the beginning of 1990 and decreased slowly thereafter.

Figure 6.3 Annual box-plots for the data presented in Figure 6.1. The squares are the annual means.
Figure 6.4: Monthly box-plots for the data presented in Figure 6.1. The squares are the monthly means.

Figure 6.5: Smoothed power spectrum for the data presented in Figure 6.1.

After elimination of the long-term trend, a periodic smoother was used to estimate the seasonal cycle. The resulting seasonal cycle is shown in Figure 6.8. The result is very similar to the one obtained using the monthly box-plot of Figure 6.4. The concentration reaches a maximum in March and a minimum in July and September.

After elimination of both the long-term trend and seasonal cycle estimates, scattergrams of the concentrations at time $t$ versus the concentrations at times $t-1$, $t-2$, and $t-3$ were constructed. They are shown in Figure 6.9. One can detect a net relationship between the concentration at time $t$ and the one at time $t-1$. The other two graphs are not as clear as this one. If one uses the regression technique described in Section 2.8, one obtains the following auto-regressive model:

$$e_t = 0.60e_{t-1} - 0.12e_{t-2} + 0.04e_{t-3} + \eta_t$$  \hspace{1cm} (6.1)

If one puts all these results together, we arrive at the following general temporal model to describe the time series:

1. A long-term trend that is neither a straight line nor monotonic.
2. A seasonal cycle that seems to be more complex than a simple sine/cosine wave.
3. An auto-regressive model of order two or three.
4. The smoothed power spectrum may indicate the presence of a long-term cycle with a period of about 3 years;

Figure 6.6: Log-normal probability plot for the data of Figure 6.1.
Table 6.1 Results of the seasonal Kendall test.

<table>
<thead>
<tr>
<th>Month</th>
<th>S</th>
<th>var(S)</th>
<th>Pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>9655</td>
<td>11895508</td>
<td>0.0051</td>
</tr>
<tr>
<td>February</td>
<td>981</td>
<td>9522241</td>
<td>0.7506</td>
</tr>
<tr>
<td>March</td>
<td>6385</td>
<td>12584630</td>
<td>0.0719</td>
</tr>
<tr>
<td>April</td>
<td>9084</td>
<td>10734322</td>
<td>0.0056</td>
</tr>
<tr>
<td>May</td>
<td>5211</td>
<td>11820522</td>
<td>0.1296</td>
</tr>
<tr>
<td>June</td>
<td>11022</td>
<td>9073361</td>
<td>0.0003</td>
</tr>
<tr>
<td>July</td>
<td>-4264</td>
<td>6578558</td>
<td>0.0964</td>
</tr>
<tr>
<td>August</td>
<td>1162</td>
<td>7313990</td>
<td>0.6674</td>
</tr>
<tr>
<td>September</td>
<td>8855</td>
<td>9200036</td>
<td>0.0035</td>
</tr>
<tr>
<td>October</td>
<td>6645</td>
<td>12352066</td>
<td>0.0587</td>
</tr>
<tr>
<td>November</td>
<td>5576</td>
<td>10113921</td>
<td>0.0796</td>
</tr>
<tr>
<td>December</td>
<td>2106</td>
<td>10113714</td>
<td>0.5078</td>
</tr>
<tr>
<td>Total</td>
<td>62418</td>
<td>121302864</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

5. A logarithmic transformation is necessary because the data seem to be close to a log-normal distribution.

From these results, we can conclude that the nonparametric techniques (Section 4.2) are not applicable because the long-term trend is not monotonic. If, however, the seasonal Kendall test (Section 4.2.2.3) is used, we obtain the results presented in Table 6.1. Although the test indicates a long-term trend at a 99.99% confidence level, the validity of the test is strongly in doubt, as the results differ widely from month to month. In fact, the test indicates that the trend may be significant for only three of the months. That problem would also be present if one used the seasonal Kendall’s slope estimator (Section 4.2.3.2).

Figure 6.7 Nadaraya-Watson kernel estimate of the long-term trend present in time series of Figure 6.1.
Figure 6.8 Periodic Nadaraya-Watson kernel estimate of the seasonal cycle present in time series in Figure 6.1.

The long-term trend obtained using the Nadaraya-Watson kernel is shown in Figure 6.7. A comparison of this estimate with the running mean, running median, loess, spline, and super smoother estimates is shown in Figure 6.10. The same span of 3 years was used for all the estimators. Except for the running median and loess estimators (dashed lines), all the estimators give about the same results. The shape of the running median is similar to the others but shifted slightly to larger values in comparison to the results of the other estimators. The loess estimate oscillates around the others, although a result similar to the others could be obtained if a span of about 4 years were used. In summary, all the smoothers, with the exception of the running median, produce very similar long-term trend estimates if the right span is used. In the case of the running median estimator, the mean long-term trend is slightly higher than that for the other smoothers. This may result from a slight asymmetry in the distribution of the observations, even after logarithmic transformation.

After elimination of the long-term trend using the Nadaraya-Watson estimate of Figure 6.7, one can look at the seasonal cycle. The periodic kernel estimation is shown in Figure 6.8. Results from three other smoothers, namely the running mean, running median,
Figure 6.10: Estimated long-term trend using the kernel, running mean, running median, loess, spline and super smoother techniques. A span of three years was used for all smoothers.

and the loess smoothers, are compared to the result of the kernel smoother in Figure 6.11. The estimates are quite close to one another. The one showing the biggest difference is the loess smoother (dashed line). As with for the long-term trend, the fit between the loess estimates and the others can be improved by slightly increasing the span.

As there are many data items missing, it is difficult to use a filter technique. Spectral analysis can still be used, however, to test the statistical significance of the different components of the temporal variations.

Figure 6.12 compares the smoothed power spectrum for the observations with the theoretical spectrum for white noise (i.e., random errors) and for a first order autoregressive process with an auto-correlation coefficient equal to 0.4. From this figure it is clear that an autoregressive process is present in the data and should be taken into account.

Figure 6.13a shows what happens when we eliminate from the data the long-term trend estimated with the kernel smoother. One can see that

Figure 6.11: Seasonal cycle estimates using the periodic kernel, running mean, running median, and loess periodic smoothers.
the spectrum associated with the long-term trend is no longer significant. The reader should note the significant long-term wave with a period of about 3 years.

The power spectrum after the periodic kernel estimates of the seasonal cycle have been eliminated from the data is shown in Figure 6.13b. One sees that most of the temporal variations present in the data have been explained, except perhaps for a long-term cycle with a period of 3 years. Note that the amplitude of this long-term cycle is small.

The monthly medians of the data are given in Table 6.2, and their temporal variations are shown in Figure 6.14. Unlike the example in Section 4.5.3, the present example shows no
Table 6.2 Monthly median $\text{SO}_4^{2-}$ ambient ionic concentrations in air at Experimental Lake Area, Ontario. Numbers in brackets have been interpolated.

<table>
<thead>
<tr>
<th>Year</th>
<th>J</th>
<th>F</th>
<th>M</th>
<th>A</th>
<th>M</th>
<th>J</th>
<th>A</th>
<th>S</th>
<th>O</th>
<th>N</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1979</td>
<td>1.11</td>
<td>1.40</td>
<td>2.51</td>
<td>1.77</td>
<td>0.70</td>
<td>0.54</td>
<td>1.08</td>
<td>0.82</td>
<td>0.55</td>
<td>0.65</td>
<td>0.81</td>
</tr>
<tr>
<td>1980</td>
<td>0.85</td>
<td>1.30</td>
<td>1.09</td>
<td>0.80</td>
<td>1.02</td>
<td>0.83</td>
<td>(1.00)</td>
<td>0.83</td>
<td>0.78</td>
<td>0.54</td>
<td>0.96</td>
</tr>
<tr>
<td>1981</td>
<td>1.57</td>
<td>1.45</td>
<td>0.95</td>
<td>1.29</td>
<td>1.12</td>
<td>0.60</td>
<td>0.83</td>
<td>0.97</td>
<td>0.43</td>
<td>0.77</td>
<td>0.79</td>
</tr>
<tr>
<td>1982</td>
<td>0.95</td>
<td>1.23</td>
<td>1.94</td>
<td>1.69</td>
<td>1.40</td>
<td>0.32</td>
<td>0.47</td>
<td>0.33</td>
<td>0.45</td>
<td>0.57</td>
<td>0.92</td>
</tr>
<tr>
<td>1983</td>
<td>0.85</td>
<td>2.07</td>
<td>2.61</td>
<td>0.99</td>
<td>0.68</td>
<td>0.76</td>
<td>0.98</td>
<td>1.17</td>
<td>0.67</td>
<td>0.82</td>
<td>0.40</td>
</tr>
<tr>
<td>1984</td>
<td>1.76</td>
<td>2.18</td>
<td>2.53</td>
<td>1.71</td>
<td>1.41</td>
<td>1.19</td>
<td>0.70</td>
<td>1.16</td>
<td>0.57</td>
<td>1.41</td>
<td>1.43</td>
</tr>
<tr>
<td>1985</td>
<td>0.97</td>
<td>1.22</td>
<td>0.93</td>
<td>1.21</td>
<td>1.07</td>
<td>0.92</td>
<td>0.62</td>
<td>0.66</td>
<td>0.36</td>
<td>0.70</td>
<td>1.10</td>
</tr>
<tr>
<td>1986</td>
<td>0.99</td>
<td>1.40</td>
<td>1.49</td>
<td>0.89</td>
<td>1.51</td>
<td>0.67</td>
<td>0.61</td>
<td>0.53</td>
<td>0.47</td>
<td>1.04</td>
<td>1.00</td>
</tr>
<tr>
<td>1987</td>
<td>1.12</td>
<td>1.87</td>
<td>2.03</td>
<td>1.41</td>
<td>1.04</td>
<td>0.97</td>
<td>0.55</td>
<td>0.83</td>
<td>0.75</td>
<td>0.77</td>
<td>0.93</td>
</tr>
<tr>
<td>1988</td>
<td>1.30</td>
<td>1.14</td>
<td>1.53</td>
<td>1.45</td>
<td>1.32</td>
<td>0.94</td>
<td>0.66</td>
<td>0.94</td>
<td>0.69</td>
<td>0.73</td>
<td>1.58</td>
</tr>
<tr>
<td>1989</td>
<td>1.67</td>
<td>1.83</td>
<td>2.30</td>
<td>1.86</td>
<td>1.25</td>
<td>0.82</td>
<td>(0.70)</td>
<td>0.61</td>
<td>0.85</td>
<td>0.94</td>
<td>0.99</td>
</tr>
<tr>
<td>1990</td>
<td>1.64</td>
<td>1.62</td>
<td>1.84</td>
<td>1.94</td>
<td>1.77</td>
<td>0.89</td>
<td>0.74</td>
<td>1.01</td>
<td>0.47</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td>1991</td>
<td>1.60</td>
<td>1.13</td>
<td>1.64</td>
<td>1.13</td>
<td>0.96</td>
<td>0.79</td>
<td>0.62</td>
<td>0.75</td>
<td>0.76</td>
<td>0.78</td>
<td>0.84</td>
</tr>
<tr>
<td>1992</td>
<td>1.70</td>
<td>1.77</td>
<td>1.95</td>
<td>1.70</td>
<td>1.00</td>
<td>0.70</td>
<td>0.80</td>
<td>0.95</td>
<td>1.15</td>
<td>0.93</td>
<td>1.11</td>
</tr>
<tr>
<td>1993</td>
<td>1.01</td>
<td>1.13</td>
<td>2.09</td>
<td>1.95</td>
<td>1.24</td>
<td>0.96</td>
<td>0.68</td>
<td>1.07</td>
<td>0.55</td>
<td>0.69</td>
<td>1.32</td>
</tr>
<tr>
<td>1994</td>
<td>0.81</td>
<td>1.63</td>
<td>1.77</td>
<td>1.44</td>
<td>0.84</td>
<td>1.08</td>
<td>0.58</td>
<td>0.44</td>
<td>1.18</td>
<td>1.01</td>
<td>1.16</td>
</tr>
</tbody>
</table>

net decrease. However, there seems to be a minimum in the middle of the 1980s and a maximum at the beginning of the 1990s. The long-term trend and the seasonal cycles estimates using the median polish and anova techniques are shown in Figures 6.15 and 6.16. The anova technique indicates that both the yearly and monthly variations are statistically significant. A simple additive model was used.

The seasonal cycles are similar to those obtained with the smoothers (Figures 6.15 and
Figure 6.15 Estimated long-term trend using anova (continuous line) and median-polish (dashed line). The overall mean was added to the yearly effects.

Figure 6.16 Estimated seasonal cycle using anova (continuous line) and median-polish (dashed line). The overall mean was added to the monthly effects.

It is interesting to note that the long-term trends obtained here are similar to those obtained with the loess smoother. They differ from the results of the other smoothers because of the presence of a local maximum in 1984. Note that the other smoothers will also show this maximum if they use a span of less than three years. One should finally note that the results of the median polish and anova techniques are very similar. The values of both the long-term trend and the seasonal cycle are smaller in the former. This is due, in part, to the fact that the estimate of the overall mean (which was added to the yearly or monthly effects to produced these two graphs) is smaller for the median polish method.

Figure 6.17 Estimated long-term trend using the maximum likelihood technique of Section 4.6.3.2
Table 6.3 Percentile contribution of model components to the total variance observed in the data.

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentile Contribution to Total Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long-term Trend</td>
<td>0.79</td>
</tr>
<tr>
<td>Long-term cycle (period = 3 years)</td>
<td>0.75</td>
</tr>
<tr>
<td>Seasonal Cycle</td>
<td>12.27</td>
</tr>
<tr>
<td>Auto-regressive model of order 2</td>
<td>19.10</td>
</tr>
<tr>
<td>$a_1 = 0.487$</td>
<td></td>
</tr>
<tr>
<td>$a_2 = -0.044$</td>
<td></td>
</tr>
</tbody>
</table>

The last technique to be illustrated fits a comprehensive model (see relation 4.147) which includes a long-term trend, long-term cycles, a seasonal cycle, and an auto-regressive model. The maximum likelihood technique described in Section 4.6.3.2 was used.

The model fitted to the data explains about 33% of the total variance observed. The components of the model fitted to the observations are listed in Table 6.3, with the fraction of the total variance they explain. The long-term trend and the seasonal cycle are shown in Figures 6.17 and 6.18. Although statistically significant, the long-term trend and the long-term-cycle represent less than 1% of the total variance. The presence of this long-term cycle confirms our earlier analysis of the power spectrum. The highest contributors to the total variance are the seasonal cycle and the second order auto-regressive component, which contributed 12 and 19% respectively. Although statistically significant, the second component of the auto-regression model is very small.

It is interesting to compare this long-term trend with those obtained using the smoothers. Such a comparison can be seen in Figure 6.19. One can see that the differences

![Image of Figure 6.18 Estimated seasonal cycle using the maximum likelihood technique of Section 4.6.3.2](image-url)
Figure 6.19: A comparison between the long-term trend estimates obtained using the maximum likelihood technique (continuous line) and the kernel, running mean, super and splines smoothers (dashed lines).

between the various estimates of the long-term trend are small, and that, overall, they give the same picture of the long-term trend, a small decrease in the first few years followed by an increase until the beginning of the 1990s, followed again by a decrease.

The correspondence between the maximum likelihood estimate and the results

Figure 6.20: A comparison between the seasonal cycle estimates obtained using the maximum likelihood technique (continuous line) and the kernel and running mean smoothers (dashed lines).

Figure 6.21: Smoothed power spectrum of the residuals. The dashed lines indicate the 95% confidence band for a random process with the same variance. The correspondence obtained using the smoothers holds for the
seasonal cycle. This is illustrated in Figure 6.20. One sees that, overall, the fit between the different estimators is quite good.

To illustrate the different methods presented in Section 5 for verifying the validity of models and techniques, we will employ some of these methods to analyze the residuals obtained when using the maximum likelihood technique to fit a comprehensive model.

The smoothed power spectrum of the residuals (Figure 6.21) indicates that the hypothesis that they are random cannot be rejected. We can arrive at the same conclusion by considering the auto-correlation coefficient up to lag 100 as shown in Figure 6.22.

The distribution of the residuals seems to deviate noticeably from the normal...
distribution, as Figure 6.23 shows. Although the deviation of the distribution is important, this should not influence the results drastically as a large number of observations are available. One should also note that the variance of the distribution may have changed slightly with time, as Figure 6.24 shows. The variation is small, however, as the differences between the third and the first quartiles change only slightly with time.

In summary, all of these analyses have shown that the ambient $SO_4^{2-}$ air concentrations at the Experimental Lakes Area have varied between 1979 and 1994, but with no net decrease or increase. The long-term trend cannot be approximated by a straight line and it is not monotonic. Thus, the nonparametric techniques of Section 4.2 should not be applied in this case.

A long-term cycle with a period equal to about 3 years is also present in the data. The amplitude of the cycle is small, and it contributes only a small fraction of the total variance. However, this contribution is about the same order of magnitude as the long-term trend.

All the analyses presented in this section indicate that the seasonal cycle is a complex and cannot be described by a simple sine/cosine wave. Therefore, a more complex description using a combination of trigonometric functions has to be used.

The presence of auto-correlations in the observations is well established. Although an auto-regressive model of order 2 was used, a model of order 1 should be enough to describe the auto-correlations.
7. Final Remarks

In this overview of time series analysis, the author has tried to give the reader more than just a catalogue of techniques. He has tried, in particular, to present a framework that can be used in the study of the temporal variations of time series. This is not the only possible framework, but it has the advantage of reducing the number of a priori hypotheses as much as possible. The main driving force behind this framework is the belief that the data should guide the analysis and the choice of conceptual model as much as possible.

The main points of this framework are the following:

1. The study of the temporal variations in a series of numbers should not be done using statistical techniques as black-boxes.

2. Users should have a good knowledge of their data. To acquire that knowledge, they have to explore the data first, giving special attention to its distribution and temporal characteristics.

3. They must also have a good knowledge of the technique(s), they want to use.

4. They must check that the technique selected can be applied to the data. It very often happens that one or many of the assumptions that are made when using a technique are not fulfilled by the data.

5. It is generally good practice to use more than one technique to explore and analyze the data.

6. Finally, one should test or verify the hypotheses made after using a particular technique.

Naturally, the use of the complete framework is not possible in some circumstances; such as when we have to analyze a large number of time series. In those situations, the use of a combination of techniques can be very useful. For example, spectral analysis and smoothing techniques can easily be used to analyze a large number of time series. The calculation and drawing of a number of smoothed power spectra and smoother estimates is not such a difficult and/or lengthy task. These techniques can be employed to isolate the time series with the more interesting characteristics which can be studied later with more demanding techniques such as the fitting of a general model by maximum likelihood.

Throughout this overview, the author has tried to present the advantages and disadvantages of the different techniques discussed. As a final review, it may be useful to give some general comments on the different types of techniques.

Nonparametric tests and slope estimators may seem to have a net advantage in that they do not assume any specific form for the distribution of the error term. However, they in fact present many disadvantages that outweigh this one advantage. First, the long-term trends are assumed to be monotonic or even to be a straight line in the case of the estimators. This is a very stringent hypothesis. Therefore, the author does not recommend the use of these techniques, except in special circumstances.

The smoothing techniques are very useful for exploring the data, and in most cases when the quantity of data is large will give very good estimates for some temporal variations, such as the long-term trend or seasonal cycle, even if some of the hypotheses (such as the independence of the residuals) are not fulfilled. We have seen an example of that in the case study presented in Section 6.
Spectral analysis is another very useful and general technique for exploring temporal variations in a time series. This technique can also be used to determine the statistical significance of those variations. On the other hand, the use of filters is quite limited because the filters have difficulties in taking missing data into account. If the number of missing data items is small, filters can still be employed, but their use becomes problematic if large amounts of data are missing.

As the reader may already have discerned, the author's preferred technique is the fitting of a very general model to the data by maximum likelihood (Section 4.6). The model used, however, should not be restricted to long-term trends that are straight lines. The more general the model, the better it can fit the data without doing injury to them.

The author would like to conclude by pointing out that the job of data analysis is not to fit the data to an *a priori* concept of what they should be but to extract what they have to say.
Appendix I : Map of the CAPMoN Network

The Canadian Air and Precipitation Monitoring Network as it was in November 1997.
Appendix II: Estimation of Smoothed Power Spectrum When Observations Are Missing

The smoothed power spectrum is obtained from the auto-covariance function using the relation (Chatfield, 1984):

\[
f(\omega) = \frac{1}{\pi} \left\{ \lambda_0 c_0^2 + 2 \sum_{k=1}^{M} \lambda_k c_k^2 \cos(2\pi\omega k) \right\}
\]  
(A1)

where \( \omega (0 \leq \omega \leq 0.5) \) is the frequency in cycles per hour, day, or other period, depending on the data observation frequency; \( c_k^2 \) is the auto-covariance coefficient of lag \( k \); \( M (\leq N) \) is the truncation point that has to be chosen (the lower the value of \( M \), the smoother the spectrum); \( \lambda_k \) is the smoothing window; and \( N \) is the number of hourly, daily, or others periods between the first and last data items. One of the possible windows is the Parzen window (Chatfield, 1984) given by:

\[
\lambda_k = \begin{cases} 
1 - 6 \left( \frac{k}{M} \right)^2 + 6 \left( \frac{k}{M} \right)^3, & 0 \leq k \leq M/2 \\
2 \left( 1 - \frac{k}{M} \right)^3, & M/2 \leq k \leq M 
\end{cases}
\]  
(A2)

Other possible windows are described by Bloomfield (1976) and Chatfield (1984).

Because of the possibility of missing data, a special technique has to be used to estimate the auto-covariance function, \( c_k^2 \). The technique developed by Dunsmuir (1984) is used. If \( X_1, X_2, \ldots, X_N \) is a sample in which some of the \( X_i \) can be missing, one first defines the indicator variable \( a_i \) as:

\[
a_i = \begin{cases} 
0 & \text{if data } i \text{ is missing} \\
1 & \text{otherwise}
\end{cases}
\]  
(A3)

\( i = 1, 2, \ldots, N \), where \( N \) is the total number of possible data points; then a new variable defined as

\[
Y_i = a_i X_i
\]  
(A4)

is calculated. To estimate the auto-covariance function; one first calculates

\[
\bar{a} = \frac{1}{N} \sum_{n=1}^{N} a_n
\]  
(A5)

\[
\bar{Y} = \frac{1}{N} \sum_{n=1}^{N} Y_n
\]  
(A6)

\[
\bar{\mu} = \bar{Y} / \bar{a}
\]  
(A7)

and then defines

\[
c_k = \frac{1}{N} \sum_{n=1}^{N-k} a_n a_{n+k}
\]  
(A8)
\[ c_k^y = \frac{1}{N} \sum_{n=1}^{N-k} a_n a_{n+k} \{X_n - \bar{\mu}\} \{X_{n+k} - \bar{\mu}\} \]  \hspace{1cm} (A9)

Finally the auto-covariance \( c_k^a \) is equal to

\[ c_k^a = \frac{c_k^y}{c_k} \quad \text{if} \quad c_k \neq 0 \]  \hspace{1cm} (A10)
Appendix III: Lomb’s Technique for Estimating Power Spectrum with Missing Data

Suppose that there are N data points \( c_i \equiv c(t_i), i = 1, 2, \ldots, N \). Then, the mean and variance of the data are found in the usual way:

\[
\bar{c} = \frac{1}{N} \sum_{i=1}^{N} c_i \quad (A11)
\]

\[
s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (c_i - \bar{c})^2 \quad (A12)
\]

Now, the Lomb normalized periodogram is defined by:

\[
P_N(\omega) = \frac{1}{2s^2} \left\{ \frac{\left[ \sum_j (c_j - \bar{c}) \cos \omega(t_j - \tau) \right]^2}{\sum_j \cos^2 \omega(t_j - \tau)} + \frac{\left[ \sum_j (c_j - \bar{c}) \sin \omega(t_j - \tau) \right]^2}{\sum_j \sin^2 \omega(t_j - \tau)} \right\} \quad (A13)
\]

where \( \tau \) is defined by:

\[
\tan(2\omega \tau) = \frac{\sum_j \sin 2\omega t_j}{\sum_j \cos 2\omega t_j} \quad (A14)
\]

The constant \( \tau \) is a kind of offset that makes \( P_N(\omega) \) completely independent of shifting all the \( t_j \)’s by any constant. See Lomb (1976) and Scargle (1982) for more details. Computer programs in Fortran and C can be found in Press et al. (1992a and b).
REFERENCES


7. Fourth Analysis on Reference Precipitation Samples by the Participating World Meteorological Organization Laboratories by Robert L. Lampe and John C. Puzak, December 1981*

8. Review of the Chemical Composition of Precipitation as Measured by the WMO BAPMoN by Prof. Dr. Hans-Walter Georgii, February 1982


11. Summary Report on the Status of the WMO Background Air Pollution Monitoring Network as at May 1982

12. Report on the Mount Kenya Baseline Station Feasibility Study edited by Dr. Russell C. Schnell


14. Effects of Sulphur Compounds and Other Pollutants on Visibility by Dr. R.F. Pueschel, April 1983

15. Provisional Daily Atmospheric Carbon Dioxide Concentrations as Measured at BAPMoN Sites for the Year 1981, May 1983


17. General Consideration and Examples of Data Evaluation and Quality Assurance Procedures Applicable to BAPMoN Precipitation Chemistry Observations by Dr. Charles Hakkarinen, July 1983

19. Forecasting of Air Pollution with Emphasis on Research in the USSR by M.E. Berlyand. August 1983

20. Extended Abstracts of Papers to be Presented at the WMO Technical Conference on Observation and Measurement of Atmospheric Contaminants (TECOMAC), Vienna, 17-21 October 1983


22. Report of the Fifth Session of the WMO Executive Council Panel of Experts on Environmental Pollution, Garmisch-Partenkirchen, Federal Republic of Germany, 30 April - 4 May 1984 (TD No. 10)

23. Provisional Daily Atmospheric Carbon Dioxide Concentrations as Measured at BAPMoN Sites for the Year 1982. November 1984 (TD No. 12)


26. Sulphur and Nitrogen in Precipitation: An Attempt to Use BAPMoN and Other Data to Show Regional and Global Distribution by Dr. C.C. Wallén. April 1986 (TD No. 103)


29. Recommendations on Sunphotometer Measurements in BAPMoN Based on the Experience of a Dust Transport Study in Africa by Dr. Guillaume A. d’Almeida. September 1985 (TD No. 67)


35. Provisional Daily Atmospheric CO₂ Concentrations as Measured at BAPMoN Sites for the Year 1983. December 1985 (TD No. 77)


43. Recent progress in sunphotometry (determination of the aerosol optical depth). November 1986


46. Provisional Daily Atmospheric Carbon Dioxide Concentrations as Measured at BAPMoN Sites for the Year 1984. December 1986 (TD No. 158)

47. Procedures and Methods for Integrated Global Background Monitoring of Environmental Pollution by F.Ya. Rovinsky, USSR and G.B. Wiersma, USA. August 1987 (TD No. 178)


50. Provisional Daily Atmospheric Carbon Dioxide Concentrations as Measured at BAPMoN Sites for the Year 1985. December 1987 (TD No. 198)


53. WMO Meeting of Experts on Strategy for the Monitoring of Suspended Particulate Matter in BAPMoN - Reports and papers presented at the meeting, Xiamen, China, 13-17 October 1986. October 1988

54. Global Atmospheric Background Monitoring for Selected Environmental Parameters. BAPMoN Data for 1983, Volume II: Precipitation chemistry, continuous atmospheric carbon dioxide and suspended particulate matter (TD No. 283)

55. Summary Report on the Status of the WMO Background Air Pollution Monitoring Network as at 31 December 1987 (TD No. 284)


58. Provisional Daily Atmospheric Carbon Dioxide Concentrations as measured at BAPMoN sites for the years 1986 and 1987 (TD No. 306)


62. Provisional Daily Atmospheric Carbon Dioxide Concentrations as measured at BAPMoN sites for the year 1988 (TD No. 355)


64. Report of the consultation to consider desirable locations and observational practices for BAPMoN stations of global importance, Bermuda Research Station, 27-30 November 1989


68. Global Atmospheric Background Monitoring for Selected Environmental Parameters. BAPMoN Data For 1989, Volume I: Atmospheric Aerosol Optical Depth

69. Provisional Daily Atmospheric Carbon Dioxide Concentrations as measured at Global Atmosphere Watch (GAW)-BAPMoN sites for the year 1989 (TD No. 400)


72. Integrated Background Monitoring of Environmental Pollution in Mid-Latitude Eurasia by Yu.A. Izrael and F.Ya. Rovinsky, USSR (TD No. 434)

73. Report of the Experts Meeting on Global Aerosol Data System (GADS), Hampton, Virginia, 11-12 September 1990 (TD No. 438)

75. Provisional Daily Atmospheric Carbon Dioxide Concentrations as measured at Global Atmosphere Watch (GAW)-BAPMoN sites for the year 1990 (TD No. 447)

76. The International Global Aerosol Programme (IGAP) Plan: Overview (TD No. 445)

77. Report of the WMO Meeting of Experts on Carbon Dioxide Concentration and Isotopic Measurement Techniques, Lake Arrowhead, California, 14-19 October 1990

78. Global Atmospheric Background Monitoring for Selected Environmental Parameters BAPMoN Data for 1990, Volume I: Atmospheric Aerosol Optical Depth (TD No. 446)


83. Report on the Global Precipitation Chemistry Programme of BAPMoN (TD No. 526)

84. Provisional Daily Atmospheric Carbon Dioxide Concentrations as measured at GAW-BAPMoN sites for the year 1991 (TD No. 543)

85. Chemical Analysis of Precipitation for GAW: Laboratory Analytical Methods and Sample Collection Standards by Dr Jaroslav Santoch (TD No. 550)


89. 4th International Conference on CO₂ (Carqueiranne, France, 13-17 September 1993) (TD No. 561)


91. Extended Abstracts of Papers Presented at the WMO Region VI Conference on the Measurement and Modelling of Atmospheric Composition Changes Including Pollution Transport, Sofia, 4-8 October 1993 (TD No. 563)


94. Report on the Measurements of Atmospheric Turbidity in BAPMoN (TD No. 603)

95. Report of the WMO Meeting of Experts on UV-B Measurements, Data Quality and Standardization
of UV Indices, Les Diablerets, Switzerland, 25-28 July 1994 (TD No. 625)

96. Global Atmospheric Background Monitoring for Selected Environmental Parameters WMO GAW Data for 1993, Volume I: Atmospheric Aerosol Optical Depth

97. Quality Assurance Project Plan (QAPjP) for Continuous Ground Based Ozone Measurements (TD No. 634)


99. Status of the WMO Global Atmosphere Watch Programme as at 31 December 1993 (TD No. 636)


104. Report of the Fourth WMO Meeting of Experts on the Quality Assurance/Science Activity Centres (QA/SACs) of the Global Atmosphere Watch, jointly held with the First Meeting of the Coordinating Committees of IGAC-GLONET and IGAC-ACE, Garmisch-Partenkirchen, Germany, 13-17 March 1995 (TD No. 689)

105. Report of the Fourth Session of the EC Panel of Experts/CAS Working Group on Environmental Pollution and Atmospheric Chemistry (Garmisch, Germany, 6-11 March 1995) (TD No. 718)


107. Extended Abstracts of Papers Presented at the WMO-IGAC Conference on the Measurement and Assessment of Atmospheric Composition Change (Beijing, China, 9-14 October 1995) (TD No. 710)


110. Report of the WMO-NOAA Expert Meeting on GAW Data Acquisition and Archiving (Asheville, NC, USA, 4-8 November 1995) (TD No. 755)


113. The Strategic Plan of the Global Atmosphere Watch (GAW) (TD No. 802)

114. Report of the Fifth WMO Meeting of Experts on the Quality Assurance/Science Activity Centres (QA/SACs) of the Global Atmosphere Watch, jointly held with the Second Meeting of the
Coordinating Committees of IGAC-GLONET and IGAC-ACE\textsuperscript{Ed}, Garmisch-Partenkirchen, Germany, 15-19 July 1996 (TD No. 787)


120. WMO-UMAP Workshop on Broad-Band UV Radiometers (Garmisch-Partenkirchen, Germany, 22-23 April 1996) (TD No. 894)


124. Fifth Session of the EC Panel of Experts/CAS Working Group on Environmental Pollution and Atmospheric Chemistry, (Geneva, Switzerland, 7-10 April 1997) (TD No. 898)

125. Instruments to Measure Solar Ultraviolet Radiation

126. Guidelines for Site Quality Control of UV Monitoring (lead author A.R. Webb) (TD No. 884)


128. The Fourth Biennial WMO Consultation on Brewer Ozone and UV Spectrophotometer Operation, Calibration and Data Reporting, (Rome, Italy, 22-25 September 1996) (TD No. 918)

129. Guidelines for Atmospheric Trace Gas Data Management (Ken Masarie and Pieter Tans), 1998 (TD No. 907)

130. Jülich Ozone Sonde Intercomparison Experiment (JOSIE, 5 February to 8 March 1996), (H.G.J. Smit and D. Kley) (TD No. 926)

131. WMO Workshop on Regional Transboundary Smoke and Haze in Southeast Asia (Singapore, 2-5 June 1998) (Gregory R. Carmichael). Two volumes


133. WMO/EMEP Workshop on Advanced Statistical Methods and their Application to Air Quality Data Sets (Helsinki, 14-18 September 1998) (TD No.956)